Face Distance: Unpacking the Role of Ethnic Ties in Venture Capital Investment

by

Jane Y. Wu

B.Com., Queen's University, 2012
B.A., Queen's University, 2012

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

Master of Science in Management Research

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2018

Signature redacted

Signature of Author _____
Sloan School of Management
May 1, 2018

Signature redacted

Certified by _____
Scott Stern
David Sarnoff Professor of Management
Thesis Supervisor

Signature redacted

Accepted by _____
Catherine Tucker
Sloan Distinguished Professor of Management
Chair, Sloan PhD Program

Face Distance: Unpacking the Role of Ethnic Ties in Venture Capital Investment

by

Jane Y. Wu

Submitted to the Sloan School of Management on May 1, 2018
in partial fulfillment of the requirements for the degree of
Master of Science in Management Research

Abstract

Venture capitalists have been shown to be more likely to invest in entrepreneurs of the same ethnicity. At the same time, this result rests on assumptions about how shared ethnicity is defined both theoretically and empirically. Current measurement of ethnic ties is problematic due to misclassifications, mixed heritage individuals, and variation in accuracy by ethnicity. This paper overcomes these limitations by taking advantage of a novel source of data – face photographs – and by applying advanced machine learning techniques to compute the facial similarity between investors and entrepreneurs in a large scale dataset of realized and potential investments. Results suggest that previous work has vastly underestimated the relationship between ethnic ties and investment. Moreover, this relationship is more nuanced than previously documented, varies with the stage of investment and the type of investors involved, and is associated with a lower likelihood of securing follow-on funding or achieving an exit.

Thesis Supervisor: Scott Stern
Title: David Sarnoff Professor of Management

> "The same people, living the same lives, and having the same experiences making largely the same decisions...For every idea we fund, how many great ideas don't even get a sounding board because we can't relate to the problem or the entrepreneur?"
>
> *Chamath Palihapitiya, venture capitalist*

1 Introduction

Venture capitalists adjudicate ideas and entrepreneurs for society. By determining which high uncertainty, high potential startups are worthwhile of funding and support (Gompers and Lerner, 2001; Kaplan and Stromberg, 2001; Puri and Zarutskie, 2012), venture capital is essentially a selection mechanism for who gets to participate in the innovation economy.[1] At the same time, venture capital is allocated by humans with imperfect information, and therefore is susceptible to subjectivity.[2] Unpacking what factors influence investment is therefore crucial for understanding potential sources of frictions in the allocation of venture capital, and whether there should be concern over the lack of diversity among venture capital investors.

One dimension that past research has shown to positively influence the likelihood of investment is shared ethnicity between investors and entrepreneurs (Hegde and Tumlinson, 2014; Bengtsson and Hsu, 2015). This result however, hinges on assumptions about what shared ethnicity represents theoretically and how to capture it empirically. In the past, theory has not always survived improvements in measurement, and there are reasons to believe that the current approach to measuring shared ethnicity is highly imperfect. These

[1]The venture capital sector has an outsized impact on economic growth. Venture capital has been shown to generate a disproportionate share of IPOs (Gompers and Lerner, 2001; Gornall and Strebulaev, 2015), technological innovation (Florida and Kenney, 1988; Kortum and Lerner, 2000; Hellmann and Puri, 2000), and positive spillovers to auxiliary firms and industries (Samila and Sorenson, 2011).

[2]Research has shown that in many different industries and regions, venture capital investment is influenced by overlap between investors and entrepreneurs in gender (Verheul and Thurik, 2001; Coleman and Robb, 2009; Brooks et al., 2014), education (Bengtsson and Hsu, 2015), geography (Stuart and Sorenson, 2003; Chen et al., 2010), nationality (Bottazzi et al., 2016) and ethnicity (Hegde and Tumlinson, 2014; Bengtsson and Hsu, 2015; Zhang et al., 2016)

limitations can be summarized as misclassifications arising from name-based techniques, the increasingly relevant issue of individuals with mixed ethnic heritage, and variation in the accuracy of ethnic taxonomies by ethnicity. As a result, there could be both type I and II errors in assessing shared ethnicity, which suggests previous findings could be over- or underestimating the relationship between shared ethnicity and investment.

The objective of this paper is to overcome these limitations by leveraging a novel, untapped source of data – face photographs – and by applying facial recognition techniques based on machine learning to measure the similarity – "face distance" – between investors and entrepreneurs' physical appearances. Using this novel approach to measure shared ethnicity within a fine-grained dataset of early-stage venture capital investors and entrepreneurs, this paper finds evidence that face distance can be used not only to correct for misclassifications arising from extant techniques, but also to empirically explore potential mechanisms previously bundled within shared ethnicity. The results of the paper suggest that previous findings have underestimated the relationship between shared ethnicity and investment. Furthermore, the relationship is more nuanced than previously documented: face distance is less relevant as more information becomes available in later stages of financing, and is less relevant among younger investors. Investing on the basis of close face distance is also found to be associated with a lower likelihood of both securing subsequent venture capital funding and achieving an exit through an acquisition or IPO.

The paper begins in Section 2 with a discussion of the related literature, highlighting a key gap in measurement for research on shared ethnicity. Section 3 proposes a novel measurement approach to augment existing techniques, and outlines the sample, methods and data used to apply this new approach to venture capital investment data. Section 4 presents the main results, explores heterogeneous effects and explores alternative explanations. Section 5 concludes and discusses directions for future research.

2 Related Literature

2.1 Theoretical and Empirical Challenges for Shared Ethnicity

People's affinity for those similar to them has been observed since at least as early as Aristotle and Plato (McPherson et al., 2001). Commonly referred to as homophily (Lazarsfeld et al., 1954) or as a subset of discrimination (Bertrand and Duflo, 2017), this phenomenon has been documented in a variety of contexts ranging from close personal ties such as marriages and friendships, to hiring and workplace teams, to more informal acquaintance networks (McPherson et al., 2001). Race is a particularly well documented factor of homophily, that is robust across different age groups[3] and racial settings.[4] However, unpacking the microfoundations of racial homophily has not been straightforward. Sociology, economics and psychology scholars have developed numerous theories on the mechanisms of homophily typically categorized as bias in preferences[5] or bias in opportunities[6] (Currarini et al., 2009). As an example, observed racial homophily between an investor and entrepreneur could reflect a taste for working with someone of the same race, or shared circumstances that make this relationship the least costly to form (e.g. both individuals are otherwise discriminated in the venture capital market, both selected into living in the same city). Teasing apart these theories however, is difficult as real relationships are difficult to emulate in the lab, while data on naturally occurring social ties only allow researchers to observe part of the picture.[7]

[3]E.g. Shrum et al. (1988) observe racial homophily in schoolchildren friendships, Hallinan and Williams (1989) finds this in high school, and Mollica et al. (2003) in MBA students

[4]E.g. Yakubovich (2005) identifies homophily in the labor market or Samara, Russia, Jacquemet and Yannelis (2012) do so in Chicago, US)

[5]I.e. in-group favoritism (Tajfel, 1970; Tajfel and Turner, 1979), taste-based discrimination (Becker, 1957), unconscious association (Greenwald and Banaji, 1995), choice homophily (McPherson and Smith-Lovin, 1987)

[6]I.e. induced homophily due to required effort (Zipf, 1949) from extant physical (Kalmijn, 1998) or social distance (Lieberson, 1980) as interactions are required to form and sustain relationships (Carley, 1991); and less exposure limits signal extraction ability (Aigner and Cain, 1977)

[7]For instance, in many of these cited studies, the strength of racial homophily can in part be explained away by the fact that race subsumes other factors of homophily such as status (e.g. education, income) and values (e.g. religion) (Lazarsfeld et al., 1954)

Whereas novel methods have mitigated some of these issues,[8] a separate but relatively understudied obstacle for research progress is measurement accuracy. This is important as there is a tight interdependence between theory and measurement in understanding homophily. For instance, whereas much of the early work on racial homophily operated under the assumption that race was the key divisor, the ability to extract information from large databases (e.g. government birth records and death certificates) facilitated classification of individuals at more granular ethnic levels. This led to the realization that observed racial homophily was driven by "ethnic levels of categorical differentiation that are nested in racial categories" (Kao and Joyner, 2004; Wimmer and Lewis, 2010) – in other words, the concept of racial homophily was in fact *ethnic* homophily. Therefore, efforts to improve measurement have the potential to advance both the theoretical and empirical understanding of the role of homophily at an ethnic or potentially even more granular level. This can also improve our understanding of the underlying mechanisms driving observed homophily, and facilitate the exploration of settings where ethnic composition is diverse, but racial categories are relatively homogenous.

One setting that benefited from the ability to measure not just race, but also ethnicity, is venture capital. Specifically, recent research has found shared ethnicity between investors and entrepreneurs to be an influential factor in investment decisions: Hegde and Tumlinson (2014) use the VentureXpert database to compute the average ethnicity of an investment firm, and find that venture capital firms are more likely to invest in entrepreneurs that share an ethnicity, and that this is particularly pronounced when there are less clear signals of entrepreneurial quality; Bengtsson and Hsu (2010, 2015) employ a similar approach using VentureEconomics data and find that even with additional information on school affiliations, gender, education level and professional positions, the strongest predictor of investment be-

[8]E.g. audit studies (Jacquemet and Yannelis, 2012), randomized control trials or closely replicating the field (Greenberg and Mollick, 2017)

tween an investor and entrepreneur is shared ethnicity. These findings contribute a potential mechanism for the ethnic "clusters" of entrepeneurship and investment networks observed by Saxenian (1999) in Silicon Valley, and are consistent with observed investor behavior in other related settings such as syndication decisions (Gompers et al., 2016) and angel investments (Venugopal, 2017).[9]

These studies however, are constrained by data and measurement in two ways. First, commonly used sources of venture capital investment data typically report deals at the fund level so researchers have to assume that all of the investors involved in a fund have equal decision-making power on each deal.[10] To see why this is problematic, consider an observed investment between a given fund and a startup. Suppose the fund has three active general partners, two of which are of German descent and one of Chinese descent, and a German entrepreneur. If the entrepreneur met with a German partner, who then had enough conviction in the deal to persuade her other partners to make the investment, the average (or a binary indicator, Mahalanobis distance, or other normalizing measure) of the venture funds' ethnicity would correctly identify this investment as occuring between individuals with shared ethnicity. If instead, the Chinese partner led support for the deal, it would be erroneously classified and findings based on this data would be biased.

Second, even assuming that data at the decision-maker level is available, a separate challenge arises from the measurement of shared ethnicity itself. The standard approach of using names data to deduce ethnicity not only runs into straightforward misclassification issues, but also is limited by the informational content of names themselves (See Appendix C for a more detailed discussion of name-based ethnic classification).

[9]This is related to a separate, rich literature in finance on trust as a key input into investment decisions. Perhaps the most relevant paper is Bottazzi et al. (2016) which finds evidence that the degree of trust between the nationalities of investors and entrepreneurs influences investment in the European venture capital market. This accords with findings of shared ethnicity as ethnicity and nationality share substantial overlap.

[10]Or in cases where board membership is available, they can extrapolate that those investors took the lead on those deals.

2.2 Measurement Challenges for Shared Ethnicity

To make the limitations of name-based ethnic classification more concrete, suppose that investors are the decision-makers, with a discrete choice of whether to invest in an entrepreneur or not given the available information:

$$y_{i,e} = \alpha_0 + \beta E_{i,e} + \alpha_1 \mathbf{X}_{i,e} + \delta_t + \epsilon_{i,e}$$

where $y_{i,e}$ is an indicator for whether investor i invests in entrepreneur e; $E_{i,e}$ is the shared ethnicity between that investor-entrepreneur dyad; $\mathbf{X}_{i,e}$ contains dyad-level covariates that predict investment between investors and entrepreneurs; and δ_t represent dummies for financing year. Now suppose that the extant method of using names to measure shared ethnicity generates a noisy measure, $\tilde{E}_{i,e}$, with three specific, potential errors.

First, the measure could contain errors, aggregated as $\mu^1_{i,e}$, because individuals are wrongly classified. For instance, an individual named Anna Lee would be classified as English on the basis of her full name, or Chinese on the basis of her surname, but she could be Norwegian. This leads to type I errors for all pairs where Anna is deemed to share an ethnicity with an English person, and type II errors for those with Norwegians. Second, names may only capture part of the ethnic composition of individuals with mixed heritage. Consider, for instance, an individual named Joseph Weber with a Filipino mother and German father who is classified as fully German by name-based classification. This could lead to type II errors, $\mu^2_{i,e}$, where dyads that share ethnic overlap are wrongly classified as being of different ethnicities. Third, ethnic taxonomy could vary in accuracy for individuals of different ethnic groups. For some dyads, shared ethnicity could be conflating heterogeneous sub-ethnic relationships, leading to type I errors $\mu^3_{i,e}$. As a stark example, suppose there are three individuals with the same surname, Peter Ma and David Ma who are of Hui Chinese origin, and John Ma who is of Cantonese Chinese origin. The measure would treat the shared ethnicity between

these three as homogenous even though the overlap is stronger between one dyad relative to the other.[11] Putting this together, this suggests that using names to elicit whether two individuals share an ethnicity actually measures $\tilde{E}_{i,e} = E_{i,e} \pm \mu^1_{i,e} + \mu^2_{i,e} - \mu^3_{i,e} = E_{i,e} + \boldsymbol{\mu}_{i,e}$. Making the simplifying assumption that the errors have mean zero, and are uncorrelated with $Y, E_{i,e}, \mathbf{X}_{i,e}, \epsilon$, then regression estimates of the relationship between shared ethnicity and investment will be biased:

$$\text{plim } \hat{\beta} = \frac{cov(\tilde{E}, Y)}{var(\tilde{E})} = \frac{\sigma^2_E}{\sigma^2_E + \sigma^2_{\boldsymbol{\mu}}}\beta$$

Given that $\tilde{E}_{i,e}$ could be positively or negatively correlated with $\mu^1_{i,e}$, is positively correlated with $\mu^2_{i,e}$, and is negatively correlated with $\mu^3_{i,e}$, it is not possible to deduce if prior findings using names as a measure of shared ethnicity are overestimating or underestimating the effect. This raises the questions: How can we improve measurement of shared ethnicity, and if we do so, how does the relationship between shared ethnicity and venture capital investment change?

3 Measurement, Methods and Data

To explore these research questions, this paper proposes a new approach to measuring shared ethnicity using face photographs that can augment existing name-based approaches. This measure is then applied to a sample of venture capital investors and entrepreneurs, with a dual aim to test whether previous findings hold, and to explore if there are novel insights that can be gleaned from better measurement.

[11] This example would be consistent with work by Maurer-Fazio (2012) which finds evidence of discrimination against non-Han candidates on a Chinese job board.

3.1 A new measure of shared ethnicity

Face photographs are a rich yet relatively untapped pool of data resulting from the growth of social networking platforms. Often called "profile pictures," these photographs are advantageous as they are self-selected by individuals to represent the best version of themselves to potential employers, partners and friends. Therefore, face photographs on professional networking platforms are reliable proxies for how an individual might appear in an interview or a pitch meeting.

Face photographs can be used to identify demographic variables (e.g. gender, age range) and physical appearance (e.g. eye color, hairstyle), a process which has improved in accuracy and in scale with advances in deep learning. One application that can be leveraged to measure shared ethnicity is the "face distance" between two face photographs. Developed largely for the purposes of facial recognition where an unknown face is identified from a set of known faces on the basis of lowest face distance, face distance is a quantitative measure of how similar two individuals are in physical facial features. Facial recognition typically works by setting a particular threshold where photographs with a sufficiently low face distance are considered to be a match. The underlying measure of face distance however, can also be informative when leveraged to determine the likelihood that two individuals share an overlap in ancestral heritage, or in other words, shared ethnicity (see Appendix D for more details).

Face photographs and face distance allow researchers to observe physical traits and similarities between individuals, previously not possible with names alone. Returning to the issues laid out in Section 2.2, this face-based approach to measuring shared ethnicity can potentially be used to correct those dyads where one or both of the individuals are misclassified ($\mu^1_{i,e}$), to identify physical similarities between individuals (that extend beyond names) to mitigate issues of multiple ethnicities ($\mu^2_{i,e}$), and to explore whether there is residual similarity at a level more granular than ethnic categories ($\mu^3_{i,e}$).

3.2 Venture capital sample

To test this new measure of shared ethnicity on a sample comparable to prior work, this paper focuses on venture capital investments completed in the United States in the post-dot-com period of 2001 to 2017.[12] This sample is further restricted to focus on the early stages of financing where prior work has found shared ethnicity to be most salient: Seed, Series A and Series B deals when there is the least amount of information available to investors and the probability of success for the startup is most uncertain relative to subsequent stages.[13]

3.3 Methods

An ideal experiment designed to estimate the causal effect of shared ethnicity on investment would randomly allocate entrepreneurs (with perfectly observable quality and characteristics) to pitch meetings with investors.[14] This would allow for a straightforward analysis of whether investors were, holding all else equal, more likely to invest in entrepreneurs of the same ethnicity. This type of experiment however is rarely implemented in practice due to the challenge of securing participation from actual entrepreneurs and investors, and relatedly, the limited size such an experiment can be implemented at. This study therefore relies on observational data of U.S. early-stage venture capital investments. Since only completed investments are observed in the data, a set of counterfactual dyads that were plausibly at risk of investment given observables has to be constructed (Section 3.3 explains counterfactual

[12]This avoids the issue of aberrant IPO activity during the dot-com bubble (Ljungqvist and Wilhelm, 2003).

[13]As each round of financing generates more observable information about the quality of a startup (e.g. research and development efforts, actual revenues, user growth), investors typically shift their emphasis from entrepreneurs' characteristics towards more objective information in later rounds. This general intuition is confirmed by Hegde and Tumlinson (2014) which find that while shared ethnicity matters for venture capital, it has no statistically significant effect on later stage financing. This early stage focus also has the added advantage that there is less concern about investors investing with the intent to replace the CEO in earlier rounds of investment which could dilute the importance of overlap between investors and entrepreneurs (Ewens and Marx, 2017).

[14]See Bertrand and Duflo (2017) for a related review of audit and experimental studies of bias in evaluation.

construction in more detail). Pooling together actual and counterfactual dyads allows for an estimation of a simple regression:

$$y_{i,e} = \alpha_0 + \beta_1 E^f_{i,e} + \beta_2 E_{i,e} + \alpha_1 \mathbf{X}_{i,e} + \delta_e + \delta_i + \delta_t + \epsilon_{i,e}$$

where $y_{i,e}$ is an indicator variable for deals where an investment is realized between investor i and entrepreneur e and equals zero for all counterfactual pairs; $E^f_{i,e}$ is shared ethnicity between that investor and the entrepreneur based on this novel measure using face distance; $E_{i,e}$ is the shared ethnicity as measured using extant name-based techniques; $\mathbf{X}_{i,e}$ contains dyad-level covariates that predict investment between investors and entrepreneurs; δ_e and δ_i represent dummies for entrepreneur and investor characteristics respectively; and δ_t are dummies for year of financing. This is also estimated using a logit model where the dependent variable is $P(y_{i,e} = 1)$. A similar model is estimated on the dyads that realized an investment replacing the dependent variable with $s_{i,e}|y_{i,e} = 1$ and $P(s_{i,e} = 1|y_{i,e} = 1)$ for the linear and logistic regression respectively, where $s_{i,e}$ is a binary outcome variable of startup success.

This approach however, poses challenges for a causal interpretation, as it hinges on the conditional independence assumption where treatment is orthogonal to outcomes after conditioning on observable covariates:

$$y_{i,e} \perp E_{i,e}|\mathbf{X}_{i,e}, \delta_e, \delta_i, \delta_t$$

In other words, the dyads with and without shared ethnicity are assumed to be comparable once the other covariates are partialled out – an assumption which cannot be tested. As a result, there is a potential risk for omitting variables that affect both investment and at least one of the independent variables (e.g. unobserved dimensions of entrepreneurial quality, degree of complementarity with the fund's existing portfolio). Relatedly, without data on the precise start-ups each investor considered, this approach assumes the conditions that

determine the choice set investors face are homogenous (e.g. an Anglo-Celtic and a Chinese investor are all else equal assumed to see and consider the same set of start-ups). Furthermore, given that counterfactual dyads are constructed from the pool of entrepreneurs with at least one observed investment, this sample excludes the start-ups that tried to fundraise but did not receive any funding. While some of these start-ups may be excluded because they are of lower quality, there could also be entrepreneurs who face such significant homophilic barriers that they are excluded from venture capital entirely. If the latter exists, then findings from this sample will underestimate the true effect of shared ethnicity on access to venture capital.

3.4 Data Construction

3.4.1 Investor-Entrepreneur Dyads

The primary data source for this study is Preqin, a commercial vendor of alternative assets data. A key feature is that in addition to standard information on completed deals between venture funds and startup firms,[15] Preqin also identifies the *individual* investor within each fund responsible for leading an investment.[16] This data was supplemented using AOL's Crunchbase, a startup information database that combines crowdsourced contributions with data sourced from news articles, press releases, LinkedIn, Twitter and other private sources. By matching startup firms from Preqin to Crunchbase using the firm name, website domain, and geographic location, the founding CEO of each startup was identified. In situations where the founding CEO could not be identified in this manner, data from LinkedIn and press releases on the financing was hand-collected to identify (in order of preference) the founding President, the founder Chairman or the CEO active at the time of fundraise. This

[15] Available in the commonly used Venture Source and VentureXpert data sources

[16] In other words, this identifies the individual making the investment decision which improves upon the prior literature which has had to assign equal probability weights across all general partners and the measurement issue discussed in Section 2.2.

gave a sample of 7,208 investor-entrepreneur dyads where an actual investment took place.

As mentioned above, given that only completed investments are observed, the study relies on analyzing a data set with all 7,208 actual investor-entrepreneur dyads between 3,262 investors and 3,885 entrepreneurs, along with many counterfactual pairs. If all investors are assumed to be at risk of investing in any randomly drawn startup in the sample, there are over 12.6 million hypothetical pairs that form the counterfactual. Beyond being computationally expensive however, given that investors have limited time for meeting and evaluating entrepreneurs, the actual dyads at risk of investment are likely a subset of this larger pool where startup characteristics align with investor preferences. To mitigate concerns that the counterfactual construction could influence results, several different approaches to building a control group were explored. Results hold across these different approaches and are presented in the Appendix.

The main analysis of the paper uses a conservative approach that restricts the full sample on four key dimensions. First, to account for potential heterogeneity in investment activity driven by fund lifecycles,[17] an investor is considered to be actively investing only in the years where an observed investment occurs. Second, observed investment activity is used to determine the financing stage preferences of each investor. This is used to restrict the set of counterfactual dyads to those where the startup's financing stage accords with investor preferences. This is important as whereas all investors in the data are early stage investors, there is heterogeneity in the degree of stage-specificity (e.g. investors at First Round Capital and Founder Collective primarily consider seed stage startups[18]). Third, observed invest-

[17]For instance, if a venture capital fund is seven years into a ten year life cycle, they may be managing their portfolio rather than actively investing in new startups.

[18]"We only invest at the seed stage and don't follow-on which means when it comes time to raise the next round we are fully-aligned with the founding team" - Founder Collective (http://www.foundercollective.com/).
"Question: I'm raising my Series B or Series C – should I contact you? Answer: Unfortunately, we're named First Round for a reason. If you're raising your second, third or fourth round, consider one of the great VCs on this list of peer-ranked firms." - First Round Capital (http://firstround.com/faq/?question=610)

ment activity and Preqin-classified industry categories[19] are used to identify the industry preferences of each investor. This is used to further restrict the set of counterfactual dyads to those where the startup firm is operating in an industry that accords with investor preferences. Whereas some venture funds are relatively industry agnostic, many focus on only a select number of industries (e.g. Orbimed primarily invests in biotechnology, health and medical devices[20]). Fourth, startup headquarter locations were used to determine the geographic preferences of investors at a state-level. This is important as investors typically have implicit (e.g. deal sourcing arises from primarily localized social networks) or explicit (e.g. New England Venture Capital Association invests only in New England startups[21]) geographic areas of focus.

This gives a data set consisting of 7,208 actual investor-entrepreneur pairs and 265,566 counterfactual dyads that were plausibly at risk of an investment given the observed startup financing year, industry, stage, and location. In other words, this assumes that an investor saw on average, 37 startups before making one investment. This accords with numbers from venture capital investors, where they fund around 2% of the startups that they meet with.[22]

[19]Startup firms are classified by Preqin into 15 industries: business services, clean technologies, consumer discretionary, energy and utilities, food and agriculture, health care, industrials, infrastructure, internet, materials, other IT, real estate, semiconductors and electronics, software and related, and telecommunications.

[20]"From biopharmaceuticals to medical devices, diagnostics, and healthcare services, OrbiMed is scouting the globe for innovations that will help ensure humanity lives healthier, longer and more productive lives." - Orbimed (http://www.orbimed.com/en/about-us)

[21]"Connecting, strengthening and advocating for New England's innovation economy." - NEVCA (https://newenglandvc.org/)

[22]Note that this is the startups that are of sufficient quality to warrant a meeting with. While there is limited research on venture capital deal flow, practitioner numbers include for instance statistics from Emergence Capital estimating that they fund 10 deals out of 500 startups they meet with (https://venturebeat.com/2014/04/19/heres-a-look-inside-a-typical-vcs-pipeline-a-must-read-for-entrepreneurs/) and Homebrew which funded 9 deals out of 399 that they met with (https://venturegeneratedcontent.com/2014/01/09/homebrews-1-the-vc-metrics-behind-investing-in-one-of-every-100-companies-we-meet/).

3.4.2 Novel Image-Based Measure of Shared Ethnicity

To implement the proposed novel approach to measuring shared ethnic heritage between a given investor and entrepreneur pair, face photographs of each individual were collected using a combination of sources. First, the full names and employer name were used to collect images of individuals from Crunchbase, which aggregates profile photos from LinkedIn and Twitter. For individuals that either could not be found in the Crunchbase database, did not have a face photograph on Crunchbase, or had an unusable image on Crunchbase,[23] face photographs were collected using the Bing Image Search API,[24] or hand-collected from LinkedIn, Twitter, AngelList or directly from company websites.

To obtain a separate, novel measure of the main covariate of interest, machine learning was used to calculate the distance between each dyad. Specifically, this face distance measure was constructed using OpenFace, an open source pre-trained deep neural net, to calculate the L2 distance between the two faces (transformed into 128-unit hypersphere representations) in each investor-entrepreneur pair. This generated a score ranging from 0.0 being the same image to a maximum distance of 4.0 for each dyad (Amos et al., 2016). For ease of interpretation, indicator variables were then constructed for dyads with face distances in the 5th, 10th and 25th percentile across the full sample (i.e. the most similar pairs with the closest face distance in the sample).

To make this more concrete, Figure 1, 2 and 3 illustrate this face distance measure on different sets of face photographs. Figure 1 shows that face distance can be used to detect similarities between individuals of the same and of different gender. Figure 2 illustrates that face distance is able to detect similarities even when there are changes in physical

[23]For instance due to a small image size, an image with multiple individuals, a logo or avatar in place of a photo. In this dataset, because it was largely a professional platform with integrations into LinkedIn, only 2.4% of photos created processing issues. See Appendix 1 for further details.

[24]The full name, position and firm name of the individual were used as inputs for a Bing Image Search using the API. This returned the top five image results which were processed to detect if there was a face in the photographs, and if the photographs were of the same person. If all five images were considered a match then that was considered to be an image of the individual.

appearance: Individuals B and D are detected to have very close face distance even though one individual is wearing glasses[25] and Individual E is detected to have close face distance with C even though they have relatively different hairstyles (as compared to Individual F, G or H). Figure 3 illustrates that face distance is effective in identifying shared ethnicity as the scores are significantly lower for individuals of the same ethnicity (as compared to the other figures). Face distance however, can also detect residual heterogeneity in similarity within ethnicity. That is, face distance can distinguish between more versus less similar individuals of the same ethnicity. See Appendix D for further details on the face distance measure.

3.4.3 Dyad-Level Covariates

Name-based Ethnicity - In addition to face distance, the extant approach to measuring shared ethnicity using names data was used.[26] Specifically, the names of investors and entrepreneurs were used to predict their ethnicities using NamePrism, a classifier trained on 74MM labeled names from 118 countries (Ye et al., 2017). This is an open, academic alternative to commercial vendors of name-based ethnic classification intended for targeted marketing but commonly used in research (such as in Kerr (2007); Kerr and Lincoln (2010); Hegde and Tumlinson (2014)). NamePrism assigns names to one of 39 ethnicities (Figure 6 provides an example of the ethnic classification output from NamePrism for a given set of surnames). Ethnicity was predicted for each individual twice: first, using only the surname (e.g. Gil which is classified as Spanish ethnicity), and then second, using the full name (e.g. Elad Gil is classified as Jewish, Martim Gil as Portuguese and Nathaniel Gil as Filipino). As shown in Table C1, for some individuals, the surname-only approach is more effective at detecting ethnicity whereas for others, the full name is more accurate. The output from this name-based classification was then used to generate indicator variables for whether an investor and entrepreneur dyad share the same ethnicity.

[25] In fact these are photographs of the same individual.

[26] Another existing approach is to use surveys to collect self-reported data (e.g. US Census).

Gender - A significant gap in venture capital investment exists between male and female entrepreneurs (Verheul and Thurik, 2001; Coleman and Robb, 2009). Lab and field experiments have found that this is driven in part by gender bias (Brooks et al., 2014; Lee and Huang, 2018). To control for gender effects on investment, the gender of investors and entrepreneurs in the sample was collected. Investor gender was identified from the prefix variable in the Preqin data. For investors with a unisex prefix such as Dr. or Prof., data from web searches and LinkedIn was used to hand-classify the individual. For entrepreneurs, first names were used to predict gender using Genderize.io, an open source classifier trained on 200,000 labelled first names. This was then verified against the predicted gender of an individual generated from running their face photograph through the Microsoft Azure Face API. This data was then used to generate an indicator variable for investor-entrepreneur dyads that share the same gender.

Age - Estimated age was collected using the Microsoft Azure Face API on investor and entrepreneur face photographs. This was used to create age quartiles for investors and entrepreneurs, and a dummy for a dyad being of the same age quartile.

Physical Appearance - Face photographs were also used to predict whether a given individual is bald, wears reading glasses, or has a beard. Although these factors can be easily manipulated, because these images are self-selected for LinkedIn, it is assumed that this revealed "best" version is the version put forth in pitch meetings. This allowed for creation of a dummy for overlap in these features of physical appearance.

Physical Attractiveness - Aesthetic attractiveness has been shown to affect the perceived competence and favorability of an individual, i.e. such individuals experience a "beauty premium" (Hamermesh, 2011). Brooks et al. (2014) find direct evidence of this in the lab where attractive young males were assessed more positively when pitching an identical idea to other less attractive individuals. To control at least in part for this, a coarse measure of attractiveness was used by building a classifier on the OpenFace neural net. A set of images

labeled with the mean level of attractiveness as scored by human evaluators on a scale of 1 to 10 from the MIT 10K US Adult Faces database was used to train this predictor. Because the training set is relatively small, the images were classified into terciles as being of low, medium, or high attractiveness. Figure 5 has an example of this attractiveness variable. This was used to generate an indicator variable for when the entrepreneur is more attractive than the investor, i.e. where the beauty premium is expected to be most salient.

Image-based Ethnicity - To tease apart how much of same ethnicity is driven by misclassified ethnicities on the basis of names, a racial (i.e. broad ethnic) classifier was built using a set of labeled images composed from the University of Massachusetts Labeled Faces in the Wild, Columbia University FaceTracer, and Wikipedia List of Americans databases to classify individuals into one of seven broad ethnic categories. This was used to generate an indicator for clearly mismatched individuals where their assigned name-based ethnicity did not match up to their image-based race. For instance, it would be considered a misclassification if an individual is classified as Anglo-Celtic on the basis of her name, and as East Asian on the basis of her face photograph. This approach however cannot capture less pronounced misclassifications, for instance if an individual is classified as Indian based on her name but is actually Pakistani.

Image Quality In order to assess the sensitivity of face distance as a measure to image quality, the Azure Face API was used to identify face photographs that were blurry and those with overexposure. An additional measure of file size was also collected. Figure 4 summarizes the different variables extracted from the face photographs.

Socioeconomic Status - With an aim to separate out socioeconomic status from ethnicity, full name and employer state data for investors and entrepreneurs was used to collect the ZIP code of each individual using the White Pages and Intelius. These ZIP codes were then matched with the 2010 US IRS Individual Income Tax ZIP Code data to collect proxy measures of socioeconomic status from the number and types of tax returns. Data was

collected on all individuals, but only those for whom their name and state returned one unique individual are included in the analysis. For this subset, a set of dummies was created for investors and entrepreneurs living in ZIP codes with above the median socioeconomic status measures.[27]

Education - Educational experience has been shown to be a factor of homophily in venture capital(Gompers et al., 2016; Bengtsson and Hsu, 2010). This raises the question of whether school admissions already plays an influential role in selecting people of similar backgrounds (and thus face distances), and whether shared ethnicity is therefore driven by shared education experience. To explore this, data on schooling was collected for a 10% subset was randomly drawn from the actual investor-entrepreneur dyads (721 actual dyads between 625 investors representing 19.2% of investors in full sample, and 671 entrepreneurs representing 17.3% of entrepreneurs). Data on the education institution, degree, major of study, and year of undergraduate completion were hand-collected from LinkedIn and Bloomberg Markets. This was used to create dummies for investor and entrepreneur dyads that attended the same school, studied the same major, and completed the same degree.

3.5 Descriptive Statistics

Descriptive statistics on investors and entrepreneurs are presented in Tables 1, 2, A2, and A1. The mean investor is a male of Anglo-Celtic ethnicity, 44.2 years of age, and based in a startup hub region (California, Massachusetts or New York). He was actively investing between 2010 and 2014, investing in 3.8 startup firms. Based on education data for the randomly drawn subset, he likely holds an MBA, and attended either Harvard University, Stanford University or University of Pennsylvania.[28]

[27]Defined as the median of the sample rather than the general population as investors and entrepreneurs are all in the right tail of the general US income distribution.

[28]In the subsample 54.65% have an MBA degree, 28.16% studied engineering at either the undergraduate or graduate level, 6.52% hold a PhD, and 6.66% have a JD. In terms of schools, in this subsample the three most commonly attended schools are Harvard University (19.83 % of investors attended for either

The mean entrepreneur is a male of Anglo-Celtic ethnicity, 41.0 years of age, also located in a startup hub state and he is slightly more attractive than the average investor. He was actively fundraising between 2012 and 2013 for a startup in the software industry, and raised funding from 2.4 venture capital investors over the observed period. There is greater heterogeneity in the education background of entrepreneurs relative to investors.[29]

Dyad-level descriptive statistics are presented in Table 2. On average, investor and entrepreneur pairs that realize an investment have lower face distances, are more likely to be of the same ethnicity (as classified by names) and estimated age range, and to be from the same city than those pairs that do not.

4 Results

The results consist of three main steps. First, evidence of a positive relationship between shared ethnicity and investment is presented in Section 4.1. Specifically, the analysis finds the effect of shared ethnicity as measured by name-based ethnic classification to be consistent with past findings, but that shared ethnicity as measured by close face distance has an independently positive relationship with venture capital investment. Second, Section 4.2 provides support that face distance captures more than just misclassifications, and finds suggestive evidence that residual similarity arises from multi-ethnic individuals. Third, analysis using this new measure of face distance in Section 4.3 shows that the relationship between shared ethnicity and investment is more complex than previously documented. Rather than being a static relationship, shared ethnicity is documented to be less relevant as more information becomes available in later funding rounds and to vary across investors.

undergraduate or graduate), Stanford University (19.14%) and the University of Pennsylvania (8.74%).

[29]For instance 6.52% of entrepreneurs did not complete undergraduate versus 1.39% of investors. Stanford University, which is the most commonly shared educational institution, was attended by 7.07% of entrepreneurs versus 19.85% which attended the most common university among investors.

4.1 Face Distance is Significant and Separate

Table A3 and Figure 9 (a) and (b) present the main results. A simple OLS regression shows that being of the same race (i.e. broad ethnic category such as "European" or "East Asian") is associated with a positive, significant 18.5% increase in the likelihood of investment between an investor-entrepreneur pair (Column 1). Consistent with the prior literature, when a surname-based measure of shared ethnicity is added, the effect of racial homophily is completely absorbed by ethnic homophily. Moreover, shared ethnicity is able to capture a stronger relationship with investment (24.6%) than what shared race could observe (Column 2). Adding in an additional name-based measure of shared ethnicity (using full names) in Column 3 absorbs part of the effect of surname-based shared ethnicity. This is because full names and surnames have significant overlap, but each approach also captures additional dyads with shared ethnicity the other cannot.[30]

Adding the novel face distance measure of shared ethnicity captures a positive, significant relationship (18.5%) with investment. Yet notably, the coefficients for the name-based measures of shared ethnicity (and the other covariates) are stable to this addition. This implies that face distance is measuring an independent dimension of similarity associated with venture capital investment, previously unobservable using name-based classification alone.

For ease of interpretation, close face distance in Column 4 is an indicator variable for whether the face distance between a given investor-entrepreneur dyad is in the 10th percentile of face distances. This additive, positive relationship with investment is robust to defining close face distance as dyads in the 25th percentile or 5th percentile of face distances (Table A3, Columns 4 and 5).[31] This also holds when the raw face distance score is used as a measure of shared ethnicity, where each one unit increase in face distance reduces the likelihood of

[30] In this sample, both approaches agree on the shared ethnicity measure for 78.2% of the dyads. For 17.9% of the dyads, surname-only classification deems them as of the same ethnicity whereas full name classification does not; the reverse is true for the remaining 3.9%.

[31] As expected, the relationship is relatively stronger when close face distance is an indicator for the 5th percentile (20.8%) and weaker for the 25th percentile (10.6%).

investment by 44.3% (Table A3 Column 6). Table A presents evidence on a subsample of investors and entrepreneurs with ZIP code information, finding that this result is robust to controlling for ZIP code level socioeconomic conditions (Columns 3-5), the racial diversity of the state (Column 6), and the political leaning of the state (Column 7).

4.2 Face Distance as Shared Ethnicity

The lack of overlap between close face distance and shared ethnicity raises the question of what is exactly captured by this new measure?[32] In other words, does use of face distance correct for misclassifications ($\mu^1_{i,e}$), measure multiple ethnic heritages ($\mu^2_{i,e}$) and/or observe within-ethnicity heterogeneity in social proximity ($\mu^3_{i,e}$)?

To identify $\mu^1_{i,e}$, it was necessary to determine what share of dyads with close face distance were essentially correcting for misclassifications. To do so, face photographs were used to measure race (discussed in Section 3.4.3), which was then used to identify individuals with name-based ethnicities that differed from their image-based race. These were hand-verified and found to be a conservative measure of misclassifications, identifying only individuals that were clearly misclassified.[33] This was then used to create an indicator variable for dyads with close face distance that included at least one individual that was misclassified by name-based techniques. Including this variable in the analysis found that although correcting for misclassifications make up a significant component of close face distance, it is not solely driven by it (Table A4 Column 2). Furthermore, separating the remaining dyads into those of different ethnicities (consistent with $\mu^2_{i,e}$) and those of the same ethnicity (consistent with

[32]It should be noted that although face distance is partially explained by overlap in education (primarily obtaining the same degree and pursuing the same field of study) (Table A10 shows that for a 10% randomly drawn subsample) and image size (Table A11 Column 2), the remaining, unexplained variation in face distance remains).

[33]I.e. where the predicted race was an obvious mismatch for the name-based ethnicity. This method was found to be effective as 92% of the full name misclassifications and 90% of surname misclassifications were validated by hand. Misclassifications were driven by erroneous classifications of Asian individuals with Anglicized names or African-Americans as Anglo-Celtic.

$\mu^3_{i,e}$) finds that close face distance has a significant, positive effect in both cases (Table A4 Column 3).

Next, to explore $\mu^2_{i,e}$ and $\mu^3_{i,e}$ further, two tests were conducted focusing on dyads classified as being of different ethnicities and the same ethnicity, respectively. The first test, presented in Table A5, uses ancestry data from the 2000 U.S. Census to identify the (top quartile) ethnicities with the highest probability of being of two or more ethnicities.[34] This allowed dyads to be separated into those where at least one individual is likely to be multi-ethnic, from those where both individuals are relatively less likely to have more than one ethnic heritage. Columns 2 and 3 show that the residual effect of close face distance (after partialling out misclassifications) is driven by individuals with mixed heritage, which is suggestive evidence of the existence of $\mu^2_{i,e}$. The second test focuses on a subsample of dyads where both the investor and entrepreneur are classified as being Chinese. This subsample was chosen because the romanization of Chinese surnames occurred at heterogeneous times across different regions and ethnic groups, making it possible to leverage these differences to classify ethnicities on a more granular, sub-ethnic level. For example, 蔡 is romanized as Choi or Choy among Cantonese-speaking Chinese from Guangdong, Hong Kong and Macau; Chai among Hakka-speaking Chinese primarily from Fujian; Cai among Mandarin-speaking Chinese from mainland China; and Tsai among Mandarin-speaking Taiwanese. As shown in Table A6, there is a positive relationship between close face distance and investment among Chinese dyads, however this is not significant given the small size of the subsample (Column 1). Once surnames are used to split Chinese investor-Chinese entrepreneur pairs into those where their surnames are of the same sub-ethnic group (e.g. Taiwanese-Taiwanese) or different sub-ethnic groups (e.g. Taiwanese-Hong Kong), the effect of close face distance on

[34]Specifically, answers to to the *First, Second, and Total Responses to the Ancestry Question by Detailed Ancestry Code* were used to calculate for each ethnicity in the U.S., the proportion of individuals with a second ethnicity. Taking the full distribution of those individuals, the top quartile was classified as being relatively diverse ethnicities, and the rest were considered relatively homogenous ethnicities in the U.S.

investment is contained entirely in the latter. In other words, within sufficiently granular definitions of ethnicity, face distance does not offer additional information on familiarity among shared sub-ethnic pairs. In contrast, it is strongest where there is potential co-ethnic ties across sub-ethnic backgrounds that names are unable to capture. With the caveat that this is a limited sample size, this evidence is in support of mixed sub-ethnic heritage (in the same vein as $\mu^2_{i,e}$) as the underlying mechanism rather than sub-ethnic homophily ($\mu^3_{i,e}$).

4.3 Shared Ethnicity Across Investors

Given that close face distance is shown to augment name-based measures of shared ethnicity in at least two ways – correcting misclassifications and capturing similarities for multi-ethnic individuals – this raises the question of whether this improved measurement can be used to glean novel insights. Namely, is the relationship between shared ethnicity and investment stable?

On the one hand, if shared ethnicity confers an advantage to investors and entrepreneurs (e.g. because it allows investors and entrepreneurs to communicate more effectively, or allows them better interpret each others' signals of quality), then it should remain relatively unchanged across different financing stages, perhaps declining a little as additional team members become more important in subsequent stages. On the other hand, if shared ethnicity is a more subjective heuristic that investors rely on when confronted with uncertainty, then as more information becomes available and startups can be better ranked, the effect of co-ethnic ties should decrease. To explore this, Table 4 separates pairs by financing stage, finding that the magnitude of the relationship decreases as more information becomes available, moving from 38.7% in the Seed stage to 14.7% at Series A to 13.05% in Series B (Columns 1-3). One potential concern is that there is more syndication activity at the Series A and Series B stages of financing. In other words, in later rounds of investment, some investors may make their investment decisions based on trust in the syndicate lead rather

than on direct evaluation of the entrepreneur (Hochberg et al., 2007; Gompers et al., 2016). To test this, assuming that investors are relatively more likely to meet and directly evaluate local entrepreneurs, co-located dyads were separated (Table A7 Columns 1-3) from those in different states (Columns 4-7). The trend of close face distance declining in magnitude across stages persists in this subsample of co-located dyads, moving from 44.6% at seed to 21.5% in Series A and an insignificant 14.2% at Series B. Furthermore, although evidence of this behavior is stronger for investors located in startup hub states (California, New York and Massachusetts), the drop in magnitude between Seed versus Series A and B also exists in non-hub states (Table A8). Although not fully conclusive, this provides some support that shared ethnicity is not a persistent influence on investment, and instead is negatively correlated with the amount of information available to investors.

At the same time, this analysis highlights that close face distance does not have the same magnitude of effect across all investors. This raises the question of whether close face distance is more or less relevant for investors with greater experience? To explore this, investors were separated into four quartiles of age. Table 5 shows that close face distance matters less for younger investors (Columns 1 and 2), and is largely driven by investors that are above the median in age (Column 3 and 4). These older investors were then separated into those that belong to a top venture capital firm versus a non-top firm.[35] Doing so finds that the relationship between close face distance and investment is quantitatively positive for old investors at both top firms and at non-top venture capital firms, and is only statistically significant for the latter (Table 6 Column 4). On one hand, this is consistent with the idea that investors become more ingrained in their investment preferences over time, i.e. they become better at pattern-matching. Under this logic, young investors are

[35]Top venture capital firms are based on InvestorRank, which uses the syndication networks of VC investors to quantify the degree of influence a VC firm has. The specific venture capital firms categorized as top firms are Andreessen Horowitz, Sequoia Capital, Accel, Benchmark Capital, Union Square Ventures, General Catalyst Partners, New Enterprise Associates, Kleiner Perkins Claufield & Byers, Kholsa Ventures and Greylock Partners.

more exploratory and open to working with different types of entrepreneurs, while older investors have identified an "ideal" entrepreneurial profile for entrepreneurs they work well with through experience. If this proven profile involves shared ethnicity, then it would be expected that older investors (and particularly those with a stable deal flow pipeline) will exhibit a stronger relationship between shared ethnicity and investment. On the other hand, it could be that investors at non-top venture capital firms have a worse ability at evaluating entrepreneurial quality, or that older investors apply less effort in evaluating deals as they have less at stake in the late stages of their careers. In these situations, investors could be defaulting to shared ethnicity as a heuristic in making investment decisions. In order to tease these explanations apart, it is necessary to understand if investing on shared ethnicity is positive or negative. Looking only at investor and entrepreneur dyads where an actual deal took place, close face distance has a negative, statistically significant relationship with follow-on investment (Table 7 Column 1) and achieving an exit through acquisition (Column 2);[36] and a negative, insignificant relationship with the startup reaching an IPO.[37] Given the rare event nature of startup exits, this should be considered as tenuous evidence that investing on the basis of close face distance is linked to worse, subsequent performance outcomes, i.e. investing on shared ethnicity comes at a cost.

5 Conclusion

By combining a novel source of data – face photographs – with a new, machine learning-based measure of shared ethnicity – face distance – this paper finds that ethnic ties not only matter for venture capital investment, but they matter much more than previously thought. This new measure of face distance is shown to augment extant methods of measuring shared ethnicity in two key ways: first, it can correct for individuals with misclassified ethnicities;

[36]Follow-on funding and acquisition data was obtained from Preqin.
[37]IPOs were collected from Preqin and Crunchbase.

and second, it can capture similarity between multi-ethnic individuals that names otherwise cannot. Furthermore, applying this novel measurement approach to a fine-grained data set of investors and entrepreneurs finds that the relationship between shared ethnicity and investment varies across investors, and specifically, is less relevant for younger investors. Within old investors, close face distance has a more significant relationship with investment for investors at non-top venture capital firms. Among realized investments, shared ethnicity is associated with a lower likelihood of raising subsequent venture capital funding, and achieving an exit through an acquisition or IPO. Taken together, this evidence suggests that investing on close face distance comes at a cost. These findings raise several questions for future research.

First, in the setting of U.S. early stage venture capital, this paper finds that shared ethnicity is strongest at the seed stage, and declines as more information is available at the Series A and Series B stages of investment. Extrapolating from these results suggests that investments at the pre-seed or angel round, where there is even greater uncertainty available, could exhibit a stronger effect. This could be further exacerbated as an increasing number of early-stage investments move to online platforms (Agrawal et al., 2016), where investors have even less information and often heavily weight observable characteristics of the entrepreneur in forming their investment decisions (Zhang and Liu, 2012; Bernstein et al., 2017). Exploration of how digitization of the investment process affects investor behavior, and specifically shared ethnicity (as measured by both face distance and names), could be a fruitful path of inquiry.

Second, variation across investors by age and quality suggests that there may not be a direct link between increasing the diversity of venture capital investors and an increase in the diversity of venture-backed entrepreneurs. This suggests that further work could unpack whether other factors of homophily (e.g. gender) also vary by investor characteristics. Panel data, including face photographs over an extended period of time, could also be used to

understand if shared ethnicity is a persistent factor that influences investors' investment decisions, or if it has additional dynamics that evolve over time and experience.

Third, this paper is limited by observational data and opens the opportunity for using randomized experiments to identify causal links and unpack whether shared ethnicity is a behavioral preference among investors (independent of opportunity). Experiments can also test whether face similarity can be used to influence investor behavior, as it has in other settings such as voting (Bailenson et al. (2006) finds that face similarity with a political candidate influences voter perceptions of the candidate) and consumer purchases (Xiao and Ding (2014) find that morphing the facial similarity of people in a marketing print advertisement can influence purchasing decisions by up to 15%).

Finally, this paper takes a first step in using face images to improve the measurement of shared ethnicity in social science research, highlighting that face distance can be used to capture ethnic ties that names would otherwise miss. This suggests that it could be worthwhile to revisit settings where shared ethnicity was found to have weak effects to validate findings with this more granular measure. More broadly, this paper highlights that face photographs themselves are a rich source of data that have been relatively underutilized, despite becoming readily available and low cost to process at scale in recent years. Although certainly applicable for social sciences research in general, in the entrepreneurship and innovation context, face photographs are promising avenues for future research to explore social proximity (including shared ethnicity) in team formation, funding and grant decisions, and other decisions where there may be frictions in evaluating people (and their ideas) under uncertainty.

References

Agrawal, A., C. Catalini, and A. Goldfarb (2016). Are syndicates the killer app of equity crowdfunding? *California Management Review 58*(2), 111–124.

Aigner, D. J. and G. G. Cain (1977). Statistical theories of discrimination in labor markets. *ILR Review 30*(2), 175–187.

Ambekar, A., C. Ward, J. Mohammed, S. Male, and S. Skiena (2009). Name-ethnicity classification from open sources. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pp. 49–58. ACM.

Amos, B., B. Ludwiczuk, and M. Satyanarayanan (2016). Openface: A general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118, CMU School of Computer Science.

Bailenson, J. N., P. Garland, S. Iyengar, and N. Yee (2006). Transformed facial similarity as a political cue: A preliminary investigation. *Political Psychology 27*(3), 373–385.

Bainbridge, W. A., P. Isola, and A. Oliva (2013). The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General 142*(4), 1323.

Becker, G. S. (1957). *The Economics of discrimination.* Chicago: University of Chicago Press.

Bengtsson, O. and D. H. Hsu (2010). How do venture capital partners match with startup founders?

Bengtsson, O. and D. H. Hsu (2015). Ethnic matching in the us venture capital market. *Journal of Business Venturing 30*(2), 338–354.

Bernstein, S., A. Korteweg, and K. Laws (2017). Attracting early-stage investors: Evidence from a randomized field experiment. *The Journal of Finance 72*(2), 509–538.

Bertrand, M. and E. Duflo (2017). Field experiments on discrimination. *Handbook of Economic Field Experiments 1*, 309–393.

Bitouk, D., N. Kumar, S. Dhillon, P. Belhumeur, and S. K. Nayar (2008). Face swapping: automatically replacing faces in photographs. *ACM Transactions on Graphics (TOG) 27*(3), 39.

Bottazzi, L., M. Da Rin, and T. Hellmann (2016). The importance of trust for investment: Evidence from venture capital. *Review of Financial Studies 29*, 2283–2318.

Brooks, A. W., L. Huang, S. W. Kearney, and F. E. Murray (2014). Investors prefer entrepreneurial ventures pitched by attractive men. *Proceedings of the National Academy of Sciences 111*(12), 4427–4431.

Carley, K. (1991). A theory of group stability. *American Sociological Review*, 331–354.

Chen, H., P. Gompers, A. Kovner, and J. Lerner (2010). Buy local? the geography of venture capital. *Journal of Urban Economics 67*(1), 90–102.

Coldman, A. J., T. Braun, and R. P. Gallagher (1988). The classification of ethnic status using name information. *Journal of Epidemiology and Community Health 42*(4), 390–395.

Coleman, S. and A. Robb (2009). A comparison of new firm financing by gender: evidence from the kauffman firm survey data. *Small Business Economics 33*(4), 397.

Currarini, S., M. O. Jackson, and P. Pin (2009). An economic model of friendship: Homophily, minorities, and segregation. *Econometrica 77*(4), 1003–1045.

Eisenthal, Y., G. Dror, and E. Ruppin (2006). Facial attractiveness: Beauty and the machine. *Neural Computation 18*(1), 119–142.

Ewens, M. and M. Marx (2017). Founder replacement and startup performance. *The Review of Financial Studies 31*(4), 1532–1565.

Fiscella, K. and A. M. Fremont (2006). Use of geocoding and surname analysis to estimate race and ethnicity. *Health Services Research 41*(4p1), 1482–1500.

Florida, R. L. and M. Kenney (1988). Venture capital-financed innovation and technological change in the usa. *Research Policy 17*(3), 119–137.

Fu, S., H. He, and Z.-G. Hou (2014). Learning race from face: A survey. *IEEE transactions on pattern analysis and machine intelligence 36*(12), 2483–2509.

Goldin, C. and M. Shim (2004). Making a name: Women's surnames at marriage and beyond. *The Journal of Economic Perspectives 18*(2), 143–160.

Gompers, P. and J. Lerner (2001). The venture capital revolution. *The Journal of Economic Perspectives 15*(2), 145–168.

Gompers, P. A., V. Mukharlyamov, and Y. Xuan (2016). The cost of friendship. *Journal of Financial Economics 119*(3), 626–644.

Gornall, W. and I. A. Strebulaev (2015). The economic impact of venture capital: Evidence from public companies.

Greenberg, J. and E. Mollick (2017). Activist choice homophily and the crowdfunding of female founders. *Administrative Science Quarterly 62*(2), 341–374.

Greenwald, A. G. and M. R. Banaji (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological Review 102*(1), 4.

Hallinan, M. T. and R. A. Williams (1989). Interracial friendship choices in secondary schools. *American Sociological Review*, 67–78.

Hamermesh, D. S. (2011). *Beauty pays: Why attractive people are more successful.* Princeton University Press.

Hegde, D. and J. Tumlinson (2014). Does social proximity enhance business partnerships? theory and evidence from ethnicity's role in us venture capital. *Management Science 60*(9), 2355–2380.

Hellmann, T. and M. Puri (2000). The interaction between product market and financing strategy: The role of venture capital. *Review of Financial Studies 13*(4), 959–984.

Hochberg, Y. V., A. Ljungqvist, and Y. Lu (2007). Whom you know matters: Venture capital networks and investment performance. *The Journal of Finance 62*(1), 251–301.

Huang, G. B., M. Ramesh, T. Berg, and E. Learned-Miller (2007). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst.

Jacquemet, N. and C. Yannelis (2012). Indiscriminate discrimination: A correspondence test for ethnic homophily in the chicago labor market. *Labour Economics 19*(6), 824–832.

Kalmijn, M. (1998). Intermarriage and homogamy: Causes, patterns, trends. *Annual Review of Sociology 24*(1), 395–421.

Kao, G. and K. Joyner (2004). Do race and ethnicity matter among friends? activities among interracial, interethnic, and intraethnic adolescent friends. *The Sociological Quarterly 45*(3), 557–573.

Kaplan, S. N. and P. Stromberg (2001). Venture capitals as principals: contracting, screening, and monitoring. *American Economic Review 91*(2), 426–430.

Kazemi, V. and J. Sullivan (2014). One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1867–1874.

Kerr, W. R. (2007). The ethnic composition of us inventors: Evidence building from ethnic names in us patents. *Harvard Business School Working Paper*.

Kerr, W. R. and W. F. Lincoln (2010). The supply side of innovation: H-1b visa reforms and us ethnic invention. *Journal of Labor Economics 28*(3), 473–508.

Khosla, A., W. A. Bainbridge, A. Torralba, and A. Oliva (2013). Modifying the memorability of face photographs. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3200–3207.

King, D. E. (2009). Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research 10*(Jul), 1755–1758.

Kortum, S. and J. Lerner (2000). Assessing the contribution of venture capital to innovation. *The RAND Journal of Economics 31*(4), 674–692.

Lazarsfeld, P. F., R. K. Merton, et al. (1954). Friendship as a social process: A substantive and methodological analysis. *Freedom and control in modern society 18*(1), 18–66.

Lee, M. and L. Huang (2018). Gender bias, social impact framing, and evaluation of entrepreneurial ventures. *Organization Science 29*(1), 1–16.

Lerner, J. and U. Malmendier (2013). With a little help from my (random) friends: Success and failure in post-business school entrepreneurship. *Review of Financial Studies 26*(10), 2411–2452.

Lieberson, S. (1980). *A piece of the pie: Blacks and white immigrants since 1880.* Univ of California Press.

Ljungqvist, A. and W. J. Wilhelm (2003). Ipo pricing in the dot-com bubble. *The Journal of Finance 58*(2), 723–752.

Mateos, P. (2007). A review of name-based ethnicity classification methods and their potential in population studies. *Population, Space and Place 13*(4), 243–263.

Mateos, P., R. Webber, and P. Longley (2007). The cultural, ethnic and linguistic classification of populations and neighbourhoods using personal names. *UCL (University College London), Centre for Advanced Spatial Analysis (UCL).*

Maurer-Fazio, M. (2012). Ethnic discrimination in china's internet job board labor market. *IZA Journal of Migration 1*(1), 12.

McEvoy, B. and D. G. Bradley (2006). Y-chromosomes and the extent of patrilineal ancestry in irish surnames. *Human Genetics 119*(1-2), 212–219.

McPherson, J. M. and L. Smith-Lovin (1987). Homophily in voluntary organizations: Status distance and the composition of face-to-face groups. *American Sociological Review*, 370–379.

McPherson, M., L. Smith-Lovin, and J. M. Cook (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology 27*(1), 415–444.

Mollica, K. A., B. Gray, and L. K. Treviño (2003). Racial homophily and its persistence in newcomers' social networks. *Organization Science 14*(2), 123–136.

Ng, H.-W. and S. Winkler (2014). A data-driven approach to cleaning large face datasets. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pp. 343–347. IEEE.

Puri, M. and R. Zarutskie (2012). On the life cycle dynamics of venture-capital-and non-venture-capital-financed firms. *The Journal of Finance 67*(6), 2247–2293.

Rampell, C. (2013, May). U.s. women on the rise as family breadwinner. Published online 2013/05/30.

Roberts, S. (2010, August). New life in u.s. no longer means new name. Published online 2010/08/26.

Samila, S. and O. Sorenson (2011). Venture capital, entrepreneurship, and economic growth. *The Review of Economics and Statistics 93*(1), 338–349.

Saxenian, A. (1999). *Silicon Valley's New Immigrant Entrepreneurs*. Public Policy Institute of California.

Schroff, F., D. Kalenichenko, and J. Philbin (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815–823.

Shrum, W., N. H. Cheek Jr, and S. MacD (1988). Friendship in school: Gender and racial homophily. *Sociology of Education*, 227–239.

Shue, K. (2013). Executive networks and firm policies: Evidence from the random assignment of mba peers. *The Review of Financial Studies 26*(6), 1401–1442.

Stuart, T. and O. Sorenson (2003). The geography of opportunity: spatial heterogeneity in founding rates and the performance of biotechnology firms. *Research Policy 32*(2), 229–253.

Tajfel, H. (1970). Experiments in intergroup discrimination. *Scientific American 223*(5), 96–103.

Tajfel, H. and J. C. Turner (1979). An integrative theory of intergroup conflict. *The Social Psychology of Intergroup Relations 33*(47), 74.

Venugopal, B. (2017). Homophily, information asymmetry and performance in the angels market.

Verheul, I. and R. Thurik (2001). Start-up capital:" does gender matter?". *Small Business Economics 16*(4), 329–346.

Wang, W. (2015, June). Interracial marriage: Who is "marrying out"? Published online 2013/06/12.

Wimmer, A. and K. Lewis (2010). Beyond and below racial homophily: Erg models of a friendship network documented on facebook. *American Journal of Sociology 116*(2), 583–642.

Xiao, L. and M. Ding (2014). Just the faces: Exploring the effects of facial features in print advertising. *Marketing Science 33*(3), 338–352.

Yakubovich, V. (2005). Weak ties, information, and influence: How workers find jobs in a local russian labor market. *American Sociological Review 70*(3), 408–421.

Ye, J., S. Han, Y. Hu, B. Coskun, M. Liu, H. Qin, and S. Skiena (2017). Nationality classification using name embeddings. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 1897–1906. ACM.

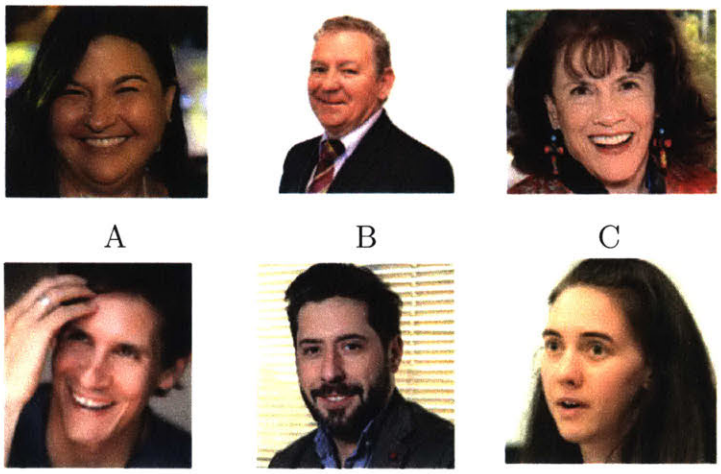
Zhang, J. and P. Liu (2012). Rational herding in microloan markets. *Management Science 58*(5), 892–912.

Zhang, J., P. K. Wong, and Y. P. Ho (2016). Ethnic enclave and entrepreneurial financing: Asian venture capitalists in silicon valley. *Strategic Entrepreneurship Journal 10*(3), 318–335.

Zipf, G. (1949). *Human Behaviour and the Principle of Least-Effort.* Addison-Wesley.

Figures

Figure 1: Face distance scores are the distance between two faces predicted by *OpenFace*.



	Face Distance Scores					
Image	A	B	C	D	E	F
A	0.00	1.833	1.608	2.058	2.692	2.265
B	–	0.00	2.035	2.469	1.624	1.963
C	–	–	0.00	1.198	2.330	1.553
D	–	–	–	0.00	2.131	1.756
E	–	–	–	–	0.00	1.480

Notes: Above scores are squared L2 distance between (the 128-dimensional unit hypersphere representations of) two faces predicted by *OpenFace*. This can detect similarities in faces within gender (e.g. C and F are more similar than C and A as seen in purple) and across gender (e.g. C and D are more similar than C and B as seen in blue). Given copyrights, these images are not from the actual study sample. Instead they are for illustrative purposes only and obtained from Flickr.com under the Creative Commons Public Domain license.

Figure 2: Face distance detects similarities across ethnicity and different appearances.

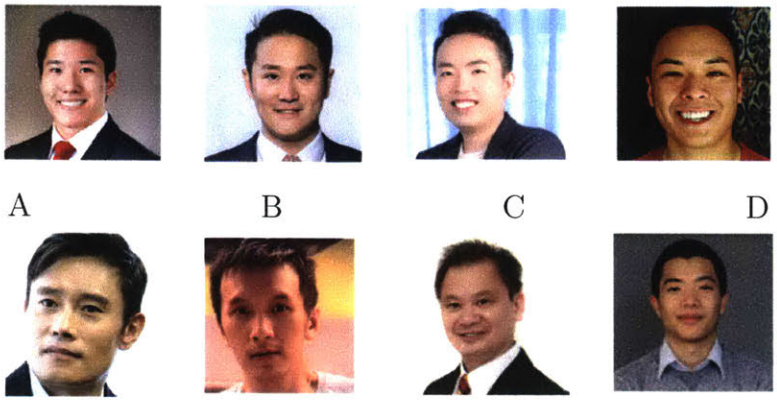
A B C D

E F G H

Image	Face Distance Scores							
	A	B	C	D	E	F	G	H
A	0.00	2.016	1.167	1.667	1.397	1.214	1.809	1.770
B	–	0.00	1.801	0.317	1.052	2.016	1.624	1.850
C	–	–	0.00	1.785	0.639	1.618	2.067	1.824
D	–	–	–	0.00	1.261	1.749	1.770	2.294
E	–	–	–	–	0.00	1.796	1.320	1.704
F	–	–	–	–	–	0.00	1.708	1.380
G	–	–	–	–	–	–	0.00	1.556

Notes: Above are face distance scores output from a face comparison classifier built on the OpenFace neural net. This example illustrates the ability for face distance to detect similarities between individuals that are classified as different ethnicities (e.g. Individuals A and C are more similar to each other, as seen in blue, than to other individuals in the comparison set), and can detect similarities between individuals even when one is wearing glasses or is in a different light (e.g. Individuals B and D as seen in purple). Given copyrights, these images are not from the actual study sample. Instead they are for illustrative purposes only and obtained from Flickr.com under the Creative Commons Public Domain license.

Figure 3: Face distance can detect relative similarities within the same ethnic category.



	Face Distance Scores							
Image	A	B	C	D	E	F	G	H
A	0.00	0.160	0.202	0.764	0.286	0.747	0.675	0.337
B	–	0.00	0.351	0.745	0.185	0.859	0.526	0.412
C	–	–	0.00	0.638	0.365	0.842	0.891	0.316
D	–	–	–	0.00	0.690	0.586	0.830	0.557
E	–	–	–	–	0.00	0.735	0.822	0.575
F	–	–	–	–	–	0.00	0.902	1.001
G	–	–	–	–	–	–	0.00	0.644

Notes: Above are face distance scores that show (1) scores are much lower (closer) in general relative to the scores for cross-ethnic face distances, making it possible to detect shared ethnicity; and (2) there is heterogeneity in the face distance scores that can detect more or less similarity *within* an ethnicity. For example, Individual B is more similar to Individuals A and E, as seen in blue, than to other individuals in the comparison set. Individuals F and H are the least similar as seen in purple. Given copyrights, these images are not from the actual study sample. Images are obtained from Flickr.com and are under the Creative Commons license with some in the public domain, and some with attribution to SoCal Photo Design, Luke Ma, Chris Marchant, and Josh Liba.

Figure 4: Variables extracted from face images using predictive algorithms.

A

B

C

D

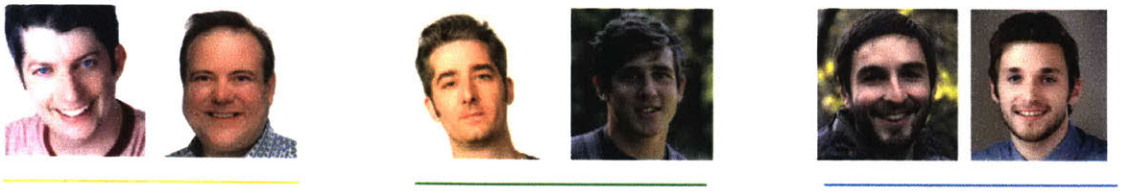
E

F

Image	Age	Broad Ethnicity	Pr(Bald)	Pr(Beard)	Glasses	Gender	Blurry	Exposure
A	26.4	East Asian (0.98)	0.01	0.0	No	Female	Low (0.16)	Over Exp. (0.9)
B	32.0	South Asian (0.55)	1.00	0.2	No	Male	Low (0.03)	Good (0.7)
C	38.2	Black (0.93)	0.29	0.0	No	Female	Low (0.12)	Over Exp. (0.75)
D	43.4	White (0.70)	0.01	0.8	Yes	Male	Low (0.00)	Good (0.66)
E	64.5	White (0.76)	0.10	0.0	No	Male	Low (0.04)	Over Exp. (0.85)
F	29.2	White (0.54)	0.03	0.0	Yes	Female	Low (0.09)	Good (0.68)

Notes: Broad ethnicity is predicted using a classifier built on OpenFace with labeled data from University of Massachusetts' *Labelled Faces in the Wild* database, Columbia University's *FaceTracer* database, and Wikipedia. Age, gender, probability of being bald, probability of having a beard, and a binary predictor of whether an individual is wearing glasses are predicted using the Microsoft Azure Face API. Given copyrights, these images are not from the actual study sample. Images are obtained from Flickr.com and are under the Creative Commons Public Domain license.

Figure 5: Example of the output from the attractiveness classifier.



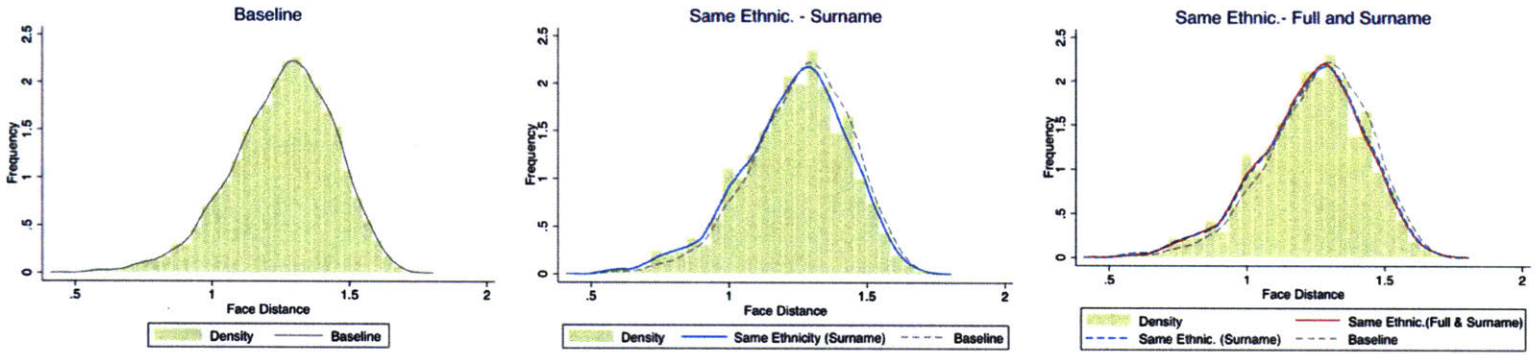
Notes: The attractiveness classifier was built using the OpenFace neural net and the images from the MIT 10K US Adult Faces database scored by human evaluators for attractiveness as a training set. These images are arranged from relatively least attractive (left) to average (middle) to relatively most attractive (right). Given copyrights, these images are not from the actual study sample. Images are obtained from Flickr.com and are under the Creative Commons license with attribution to moedym, Bre Pettis, James Darpinian, Scott Parker, Brad Carroll, Citrix Partners.

Figure 6: Ethnic classification using surnames and full names.

Name	Predicted Broad Ethnicity	Predicted Ethnicity	Prediction Confidence
Kumar	South Asian	South Asian	0.9978
Angelos	European	Greek	0.9910
Okamoto	East Asian	Japanese	0.9907
Ludwig	European	German	0.9739
O'Keefe	English	Anglo-Celtic	0.9281
Yang	East Asian	Chinese	0.8573
Ciambella	European	Italian	0.8314
Lashkari	Arab	Persian	0.7982
Yasar	Arab	Turkic	0.6245
Christensen	Nordic	Danish	0.546
Jennifer Lee	English	Anglo-Celtic	0.6739
Yimin Lee	East Asian	Chinese	0.7833
Soo Lee	East Asian	Korean	0.8414
Lars Hansen	Nordic	Danish	0.8058
Barbara Moretti	European	Italian	0.7423

Notes: Surname and full names are used to predict the ethnicity that an individual is of using Stony Brook University's *NamePrism*, a leading open-source name-based tool using a labeled data set of 74 million names with 118 countries of origin. These countries of origin are then mapped to a taxonomy of ethnic/nationality categories that has 10 broad categories (e.g. European) and 26 more granular categories (e.g. German). As shown above, this classifier can be used for both surnames only and for full names and has varying degrees of prediction confidence depending on how ethnically unique the name is based on Wikipedia entries.

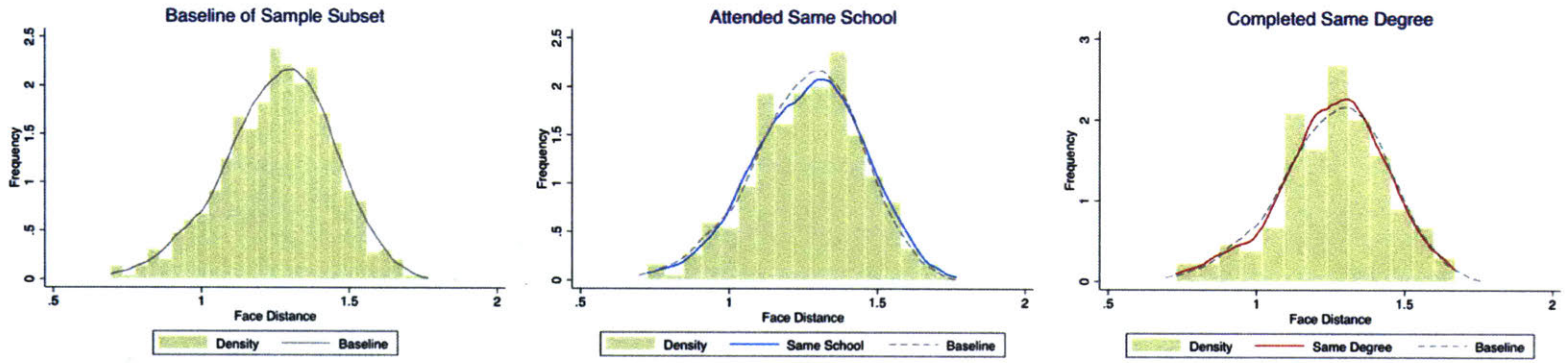
Figure 7: Face distance distribution persists among same ethnicity dyads.



(a) Sample Baseline (b) Same ethnicity on surname basis (c) Same ethnicity on full name basis

Notes: Raw face distance scores are plotted by histograms and kernel density. Graph (a) begins with a plot of the baseline distribution of face distance between the investor-entrepreneur dyads; (b) shows that the distribution shifts only slightly lower (i.e. close in face distance) when restricted to dyads of the same ethnicity as classified by surnames; (c) shows that this is essentially unchanged once further restricted to dyads of the same ethnicity as classified by full names.

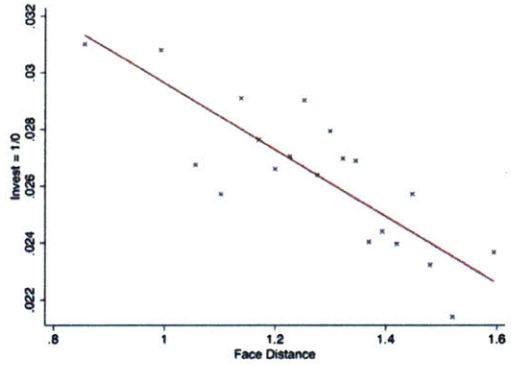
Figure 8: Face distance distribution persists among dyads with similar education experience.



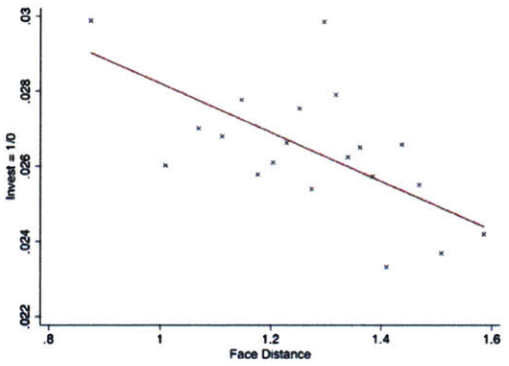
(a) Sample Baseline (b) Same ethnicity on surname basis (c) Same ethnicity on full name basis

Notes: Schooling data was collected for a 10% randomly drawn sample of investor-entrepreneur dyads. Graph (a) plots the baseline distribution of face distance between the dyads in this subset. Graph (b) shows that this distribution remains qualitatively similar when restricted to dyads that attended the same university. Graph (c) shows this similarity for dyads that completed the same major of study at either the undergraduate (e.g. BSc Engineering) or graduate level (e.g. MBA).

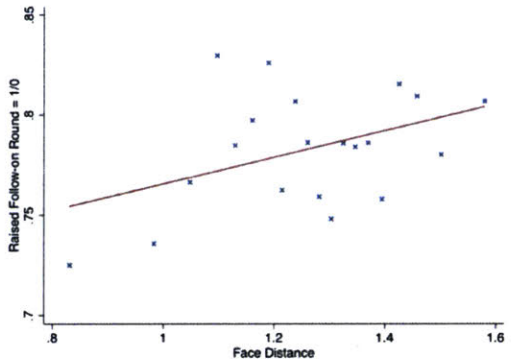
Figure 9: Face Distance and Investment, Startup Outcomes



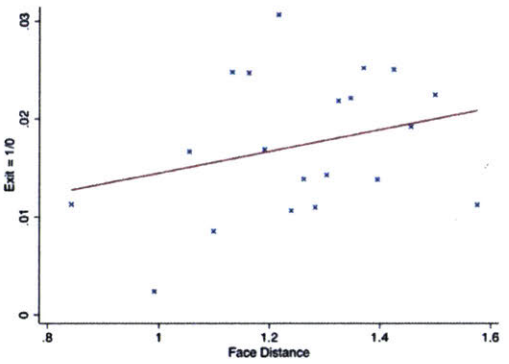
(a) Likelihood of Investment - No Controls



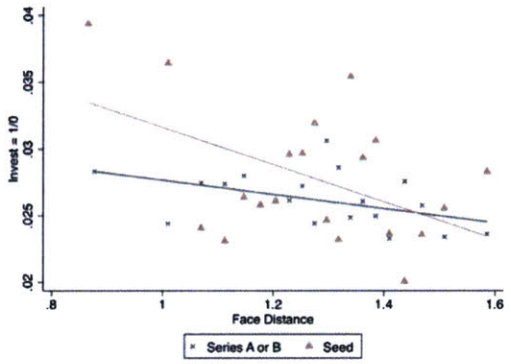
(b) Likelihood of Investment - With Controls



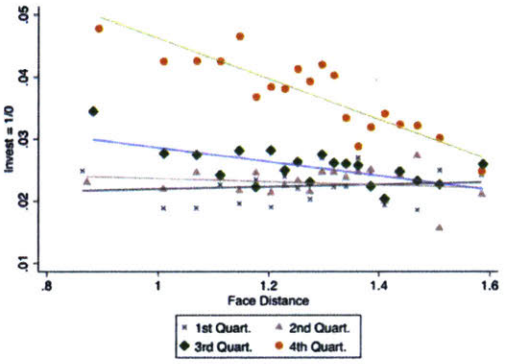
(c) Raising Follow-on Financing



(d) Acquired or Completed IPO



(e) Strongest in Seed Stage



(f) Driven by Oldest VCs

Notes: Binned scatterplots of face distance against investment and startup outcomes with the OLS best linear fit line overlaid. Graph (a) shows that as face distance between a dyad increases, the likelihood of investment decreases. Graph (b) shows that this relationship holds with the addition of controls for dyadic overlap in gender, age, ethnicity based on both surnames only and on full names, physical features (e.g. glasses, beard, bald), state and city. Graph (c) and (d) show that of those dyads that received investment, as face distance between a dyad decreases, the likelihood of raising a subsequent funding round or achieving an exit through acquisition or IPO decreases. Graph (e) separates (b) by deals at the seed level, and those at Series A or B showing that the negative relationship between face distance and investment likelihood is stronger at the earliest stage of financing. Graph (f) separates (b) by dyads with investors by quartile of investor age, showing this negative relationship is significantly stronger in the oldest group of investors.

Tables

Table 1: Descriptive Statistics: Individual Variables

Variable	Mean	Std. Dev.	Min.	Max.	Mean	Std. Dev.	Min.	Max.
		Investor				*Entrepreneur*		
Estimated Age	44.167	9.710	22.8	87.3	41.037	9.114	18	80.400
Male	0.922	0.268	0	1	0.915	0.278	0	1
Hub State	0.673	0.469	0	1	0.643	0.479	0	1
Attract. Score	4.548	2.573	1	10	4.95	2.66	1	10
Year of Entry	2010	4.647	2001	2017	2012	3.952	2001	2017
Year of Exit	2014	3.643	2001	2017	2013	3.941	2001	2017
N	3,262				3,885			

Notes: Ages are estimated from face photographs using the Azure Face API. Gender is obtained from Preqin prefixes for investors, and from first names using the Genderize.io API for entrepreneurs, and are hand-verified where not found or prediction confidence is less than 0.50. Hub states are defined as the top three states with the highest number and dollar amount of investments: California, Massachusetts and New York. Attractiveness is predicted using a classifier built on the OpenFace neural net and trained using a set of human-evaluated attractiveness photos from the 10k US Adult Faces database. Years are rounded for ease of comprehension.

Table 2: Descriptive Statistics: Investor-Entrepreneur Dyadic Variables

Variable	Mean	Std. Dev.	Min.	Max.	Mean	Std. Dev.	Min.	Max.
		Invest = 1				*Invest = 0*		
Face Distance (Raw)	1.258	0.187	0.41	1.804	1.273	0.183	0.333	1.792
Face Distance (5th pctile)	0.059	0.235	0	1	0.05	0.217	0	1
Face Distance (10th pctile)	0.117	0.321	0	1	0.1	0.299	0	1
Face Distance (25th pctile)	0.271	0.445	0	1	0.249	0.433	0	1
Same Ethnicity (Full Name)	0.434	0.496	0	1	0.34	0.474	0	1
Same Broad Ethnic. (Full Name)	0.451	0.498	0	1	0.358	0.479	0	1
Same Ethnicity	0.266	0.442	0	1	0.200	0.400	0	1
Same Broad Ethnic. (Surname)	0.316	0.465	0	1	0.254	0.435	0	1
Same Gender	0.873	0.333	0	1	0.858	0.349	0	1
Same Age Range	0.300	0.458	0	1	0.264	0.441	0	1
Same City	0.168	0.374	0	1	0.119	0.324	0	1
Same State	0.494	0.5	0	1	0.542	0.498	0	1
Both in Hub Region	0.578	0.494	0	1	0.84	0.366	0	1
Same Attractiveness	0.306	0.461	0	1	0.254	0.435	0	1
Both Wear Glasses	0.042	0.200	0	1	0.032	0.176	0	1
Both Bald	0.019	0.136	0	1	0.015	0.121	0	1
Both Have Beard	0.002	0.042	0	1	0.002	0.048	0	1
N	7,208				265,566			

Notes: Counterfactual investor-entrepreneur dyadic pairs (Invest = 0) are constructed by restricting the full cross-product of 12.6MM possible dyads (between 3262 investors and 3885 entrepreneurs) in 4 ways: (1) Investor made a deal in that financing stage in the given year, (2) Investor made a deal in that financing stage between 2001-2017, (3) Investor made a deal in that industry in that period, and (4) Investor made a deal in that state at some point between 2001-2017 . This gives an average of 36-37 counterfactual deals for each observed deal.

Table 3: Face distance matters for investment and is separate from shared ethnicity

	(1) Broad Ethnicity	(2) + Ethnicity (Surname)	(3) + Ethnicity (Full Name)	(4) + Face Distance (10th pctile)
Dependent Variable		*Invest = 1/0*		
Same Race	0.0049	-0.0002	-0.0000	-0.0003
	(0.0009)***	(0.0015)	(0.0015)	(0.0015)
Same Ethnicity (Surname Only)		0.0065	0.0032	0.0033
		(0.0019)***	(0.0020)	(0.0019)*
Same Ethnicity (Full name)			0.0054	0.0054
			(0.0010)***	(0.0010)***
Close Face Distance				0.0049
				(0.0012)***
Same Gender	0.0030	0.0029	0.0029	0.0024
	(0.0014)**	(0.0013)**	(0.0013)**	(0.0013)*
Same Features	0.0025	0.0025	0.0027	0.0024
	(0.0020)	(0.0020)	(0.0020)	(0.0020)
Ent. More Attractive	-0.0003	-0.0003	-0.0002	-0.0001
	(0.0014)	(0.0014)	(0.0014)	(0.0014)
Same State	0.0026	0.0026	0.0027	0.0027
	(0.0015)*	(0.0015)*	(0.0015)*	(0.0015)*
Same City	0.0170	0.0170	0.0169	0.0170
	(0.0018)***	(0.0018)***	(0.0018)***	(0.0018)***
Constant	0.3967	0.3965	0.3937	0.3933
	(0.0520)***	(0.0519)***	(0.0519)***	(0.0519)***
R-squared	0.0459	0.0460	0.0462	0.0462

Notes: N = 272,774. Mean of invest is 0.0264. All regressions include dummies for entrepreneurial age (by quartile), investor state, startup firm industry and financing year. Robust standard errors, clustered at the investor level, are reported in parentheses below each coefficient estimate. For ease of interpretation, an indicator variable was created for whether the face distance between a given investor-entrepreneur dyad is in the 10th percentile of face distances (i.e. the most similar which is shorthanded as "close face distance"). OLS regressions show in Column (1) that there is a positive, statistically significant relationship between investment and being of the same broad ethnic (i.e. "racial") category (e.g. European, South Asian). In Column (2) shows that consistent with prior literature, this is entirely driven by a more granular shared ethnicity as measured by individuals' surnames (e.g. German, Indian). In Column (3), when shared ethnicity as measured by individuals' full names is added, it absorbs part of the effect of surname-based shared ethnicity. When an indicator variable for a dyad being of close face distance (defined as being in the 10th percentile of face distance scores) is added, it is statistically significant and positive. Most importantly it is of a similar magnitude but it is a separate concept from what is captured by shared ethnicity on the basis of names.
$* : p < 0.10; ** : p < 0.05; *** p < 0.01$

Table 4: Face distance is less relevant as more information is available

	(1)	(2)	(3)	(4)	(5)	(6)
		OLS			Logit	
	Seed	Series A	Series B	Seed	Series A	Series B
Dependent Variable		*Invest = 1/0*			*Invest = 1/0*	
Close Face Distance	0.0110	0.0034	0.0038	0.3631	0.1350	0.1198
	(0.0034)***	(0.0016)**	(0.0018)**	(0.1006)***	(0.0668)**	(0.0636)*
Same Ethnicity	0.0063	0.0038	0.0029	0.2273	0.1214	0.0839
	(0.0046)	(0.0024)	(0.0028)	(0.1872)	(0.1129)	(0.1016)
Mean of Invest	0.0284	0.0231	0.0291			
N	35890	118282	118602	35859	118273	118595
R-squared	0.0652	0.0498	0.0503			

Notes: N = 272,774. Control variables used in Column (4) of Table A3 are included in all regressions (not reported). Fixed effects for investor age quartile, investor state, startup fundraising year, and startup industry are included in the above regressions. Robust standard errors, clustered at the investor level, are reported in parentheses below each coefficient estimate.
$* : p < 0.10; ** : p < 0.05; *** p < 0.01$

Table 5: Older investors more likely to invest on face distance

	(1)	(2)	(3)	(4)	(5)	(6)
			OLS		Logit	
Investor Age	Quartile 1	Quart. 2	Quart. 3	Quart 4	Below Median	Above Median
Dependent Variable			*Invest = 1/0*		*Invest = 1/0*	
Close Face Distance	0.0024	0.0009	0.0078	0.0077	0.0796	0.2425
	(0.0019)	(0.0019)	(0.0023)***	(0.0038)**	(0.0631)	(0.0587)***
Mean of Invest						
N	68801	66934	79044	57995	135730	137025
R-squared	0.0375	0.0376	0.0538	0.0879		

Notes: OLS regressions on data split by investor age quartile. Control variables used in Column (4) of Table A3. Fixed effects for investor state, startup fundraising year, and startup industry are included in the above regressions. Robust s.e. clustered at the investor level are reported. $*: p < 0.10; **: p < 0.05; ***p < 0.01$

Table 6: Older investors at non-top VC firms more likely to invest on face distance

	(1)	(2)	(3)	(4)	(5)	(6)
			OLS		Logit	
Investor Age	Below Median		Above Median		Above Median	
VC Firm Quality	Top	Non-Top	Top	Non-Top	Top	Non-Top
Dependent Variable			*Invest = 1/0*		*Invest = 1/0*	
Close Face Distance	0.0003	0.0020	0.0086	0.0083	0.3447	0.2375
	(0.0027)	(0.0016)	(0.0056)	(0.0022)***	(0.2056)*	(0.0607)***
Mean of Invest	0.0173	0.0235	0.0194	0.0316		
N	19438	116297	15879	121160	15879	121146
R-squared	0.0090	0.0367	0.0252	0.0700		

Notes: OLS regressions on data split by investor age quartile. Control variables used in Column (4) of Table A3. Fixed effects for investor state, fundraising year, and startup industry are included in the above regressions. Robust s.e., clustered at the investor level, are reported in parentheses. $*: p < 0.10; **: p < 0.05; ***p < 0.01$

Table 7: Face distance is associated with lower likelihood of follow-on funding and exit

	(1)	(2)	(3)	(4)	(5)
		OLS		Logit	
Dependent Var.	Follow-On	Acquisition	IPO	Follow-On	Exit
Close Face Distance	-0.0452	-0.0066	-0.0048	-0.2531	-0.8166
	(0.0224)**	(0.0028)**	(0.0073)	(0.1373)*	(0.4058)**
Same Ethnicity	0.0312	-0.0096	-0.0200	0.2086	-0.7533
	(0.0360)	(0.0082)	(0.0125)	(0.2281)	(0.4393)*
Mean	0.7826	0.0103	0.0465		
R-squared	0.1279	0.0394	0.1020		

Notes: Dependent variables for are indicator variables for whether a startup observed to raise Seed or Series A funding is observed to raise Series A or Series B subsequently; and whether a startup is observed to have exited via acquisition by Preqin anytime up to 2018 or to have gone public by Preqin or Crunchbase anytime up to 2018. Control variables used in Column (4) of Table A3 are included in all regressions (not reported). Fixed effects for investor age quartile, investor state, startup fundraising year, and startup industry are included in the above regressions. Logit odds ratios for close face distance under Column (4) is 0.7764 and under Column (5) is 0.4419.
$*: p < 0.10; **: p < 0.05; ***p < 0.01$

Appendix

A Supplementary Tables

Table A1: Frequency of startup industry.

Startup Industry	% Share Investments	% Share Firms
Software & Related	29.66	30.63
Internet	18.99	19.51
Health Care	18.66	16.55
Telecoms	10.53	10.81
Other IT	9.39	9.32
Business Services	2.62	2.86
Consumer Discretionary	2.51	2.45
Clean Technology	2.39	2.32
Semic. & Electronics	2.05	2.16
Food & Agriculture	1.33	1.44
Industrials	1.33	1.31
Other*	0.45	0.65

Notes: "Other" includes Energy & Utilities, Materials, Real Estate and Infrastructure which represent 0.37, 0.1, 0.04, and 0.01 % of investments and 0.39, 0.15, 0.08, 0.03 % of startup firms respectively. Reported are % shares of the 7208 total investments and 3885 unique startup firms. Industry definitions are from Preqin, and are similar to S&P 500 sector definitions.

Table A2: Frequency of name-based predictions of ethnicity.

Ethnicity	Investor % Share	Entrepreneur % Share	Ethnicity	Investor % Share	Entrepreneur % Share
Anglo-Celtic	48.04	43.58	Greek	0.55	0.59
German	13.24	13.05	East African	0.52	0.57
South Asian	7.69	9.52	Swedish	0.46	0.21
French	5.64	6.90	Danish	0.43	0.49
Chinese	4.35	3.29	Norwegian	0.43	0.33
Italian	3.53	2.60	East European	0.31	0.49
Spanish	2.67	3.50	Romanian	0.25	0.28
Portuguese	1.53	1.93	South Slavic	0.25	0.49
Indonesian	1.35	1.54	Turkish	0.18	0.31
Korean	1.20	1.06	Pakistani	0.15	0.33
West African	1.13	1.16	Maghreb	0.12	0.08
Jewish	1.04	1.13	South African	0.09	0.18
Vietnamese	0.95	0.69	Malaysian	0.09	0.15
Japanese	0.83	0.57	Finnish	0.06	0.13
Filipino	0.80	1.03	Arabic	0.06	0.08
Russian	0.71	0.75	Myanmar	0.03	0.05
Persian	0.71	1.34	Bangladeshi	0.03	0.39
Nubian	0.58	1.16	Baltics	0.00	0.05

Notes: These show the percentage share of the 3262 investors and 3885 entrepreneurs classified under each ethnicity based on surnames using Stony Brook University's *NamePrism*. See the note for Figure 6 for more details on the tool.

Table A3: Results are robust to different definitions of "Close Face Distance"

	(1) Broad Ethnicity	(2) + Ethnicity (Surname)	(3) + Ethnicity (Full Name)	(4) + Face Distance (25th pctile)	(5) + Face Distance (5th pctile)	(6) Raw
Dependent Variable		*Invest = 1/0*				
Same Race	0.0049	-0.0002	-0.0000	-0.0002	-0.0003	-0.0005
	(0.0009)***	(0.0015)	(0.0015)	(0.0015)	(0.0015)	(0.0015)
Same Ethnicity (Surname Only)		0.0065	0.0032	0.0033	0.0033	0.0034
		(0.0019)***	(0.0020)	(0.0020)*	(0.0019)*	(0.0019)*
Same Ethnicity (Full name)			0.0054	0.0053	0.0054	0.0052
			(0.0010)***	(0.0010)***	(0.0010)***	(0.0010)***
Face Distance (25th pctile)				0.0028		
				(0.0008)***		
Face Distance (5th pctile)					0.0055	
					(0.0017)***	
Face Distance (Raw)						-0.0117
						(0.0021)***
Same Gender	0.0030	0.0029	0.0029	0.0023	0.0026	0.0013
	(0.0014)**	(0.0013)**	(0.0013)**	(0.0014)*	(0.0013)*	(0.0014)
Same Features	0.0025	0.0025	0.0027	0.0024	0.0025	0.0021
	(0.0020)	(0.0020)	(0.0020)	(0.0020)	(0.0020)	(0.0020)
Ent. More Attractive	-0.0003	-0.0003	-0.0002	-0.0001	-0.0002	0.0000
	(0.0014)	(0.0014)	(0.0014)	(0.0014)	(0.0014)	(0.0014)
Same State	0.0026	0.0026	0.0027	0.0027	0.0027	0.0027
	(0.0015)*	(0.0015)*	(0.0015)*	(0.0015)*	(0.0015)*	(0.0015)*
Same City	0.0170	0.0170	0.0169	0.0170	0.0170	0.0170
	(0.0018)***	(0.0018)***	(0.0018)***	(0.0018)***	(0.0018)***	(0.0018)***
Constant	0.3967	0.3965	0.3937	0.3934	0.3936	0.4098
	(0.0520)***	(0.0519)***	(0.0519)***	(0.0518)***	(0.0520)***	(0.0520)***
R-squared	0.0459	0.0460	0.0462	0.0462	0.0462	0.0463

Notes: N = 272,774. Mean of invest is 0.0264. All regressions include dummies for entrepreneurial age (by quartile), investor state, startup firm industry and financing year. Robust standard errors, clustered at the investor level, are reported in parentheses below each coefficient estimate. The results in Table A3 are robust to alternative definitions of close face distance, using the 25th and 5th percentile of face distance, and to using the raw face distance score. As expected the result is relatively stronger when the threshold for close face distance is more strict (5th distance) and weaker when it is more relaxed (25th percentile).
$* : p < 0.10; ** : p < 0.05; *** p < 0.01$

Table A4: Close face distance is not solely misclassifications

	(1)	(2)	(3)
Dependent Variable		*Invest = 1*	
Close Face Distance	0.0049	0.0040	0.0028
	(0.0012)***	(0.0012)***	(0.0013)**
Misclassified		0.0087	0.0099
		(0.0043)**	(0.0043)**
Close Face, Same Ethnic.			0.0055
			(0.0033)*
R-squared	0.0462	0.0463	0.0463

Notes: N = 272,774. Mean of invest is 0.0264. All regressions include the covariates used in Column (4) of Table A3 and dummies for entrepreneurial age (by quartile), investor state, startup firm industry and financing year. Robust standard errors, clustered at the investor level, are reported in parentheses below each coefficient estimate. For ease of interpretation, Close Face Distance is an indicator variable for whether the face distance between a given investor-entrepreneur dyad is in the 10th percentile of face distances. Misclassified is an indicator variables for dyads with close face distance that included at least one individual who was misclassified by name-based ethnic prediction where the former. This was generated using a supervised classifier built on the OpenFace neural net to predict the race of the individuals and identify individuals with assigned ethnicities (on the basis on names) that differ from the predicted race (on the basis of face photographs).
$*: p < 0.10; **: p < 0.05; ***p < 0.01$

Table A5: Stronger among individuals with higher likelihood of mixed heritage

	(1) Different Ethnicity	(2) Likely Mixed Ethnicity	(3) Likely Single Ethnicity
Dependent Variable		*Invest = 1/0*	
Close Face Distance	0.0028	0.0024	0.0048
	(0.0013)**	(0.0014)*	(0.0030)
Misclassified	0.0106	0.0117	0.0032
	(0.0042)**	(0.0045)**	(0.0115)
Mean	0.0243	0.0249	0.0188
N	228300	206757	21543
R-squared	0.0412	0.0428	0.0327

Notes: All regressions include the covariates used in Column (4) of Table A3 and dummies for entrepreneurial age (by quartile), investor state, startup firm industry and financing year. Robust standard errors, clustered at the investor level, are reported in parentheses below each coefficient estimate. For ease of interpretation, Close Face Distance is an indicator variable for whether the face distance between a given investor-entrepreneur dyad is in the 10th percentile of face distances. Misclassified is an indicator variables for dyads with close face distance that included at least one individual who was misclassified by name-based ethnic prediction where the former. Likely Multiple Ethnicity are those dyads where at least one individual belongs to an ethnicity that is in the top quartile of ethnicites most likely to be of mixed heritage (as per the 2000 U.S. Census).
$* : p < 0.10; ** : p < 0.05; *** p < 0.01$

Table A6: Chinese subsample effect is driven by different sub-ethnic dyads

	(1) Both Chinese	(2) Same Sub-Chinese	(3) Diff Sub-Chinese
Dependent Variable		*Invest = 1/0*	
Close Face Distance	0.0015	-0.0494	0.0108
	(0.0177)	(0.0318)	(0.0141)
Mean	0.0321	0.0357	0.0264
N	468	112	227
R-squared	0.16163	0.39831	0.46245

Notes: All regressions include the covariates used in Column (4) of Table A3 and dummies for entrepreneurial age (by quartile), investor state, startup firm industry and financing year. Robust standard errors, clustered at the investor level, are reported in parentheses below each coefficient estimate. This subsample is limited to Chinese-Chinese dyads where surnames could be used to classify individuals' sub-ethnicities (e.g. Taiwanese, Cantonese, Hui, etc.). Dyads were then separated into those of the same sub-ethnicity and those of different sub-ethnicities, where the effect of close face distance in the aggregate for Chinese dyads is driven by the latter.
$* : p < 0.10; ** : p < 0.05; *** p < 0.01$

Table A7: Stronger among geographically proximate firms

	(1)	(2)	(3)	(4)	(5)	(6)
		Co-located			Not Co-located	
	Seed	Series A	Series B	Seed	Series A	Series B
Dependent Variable		*Invest*			*Invest*	
Close Face Distance	0.0133	0.0046	0.0036	0.0081	0.0012	0.0042
	(0.0047)***	(0.0021)**	(0.0023)	(0.0048)*	(0.0024)	(0.0029)
Mean of Invest	0.0298	0.0213	0.0253	0.0269	0.0253	0.0338
N	18426	64450	64744	17464	53832	53858
R-squared	0.13343	0.08102	0.06965	0.05294	0.05753	0.06173

Notes: Colocated is defined as being situated within the same state. Control variables used in Table 1 Column (7) are included in all regressions (not reported). Fixed effects for investor age quartile, investor state, startup fundraising year, and startup industry are included in the above regressions.
$* : p < 0.10; ** : p < 0.05; * * * p < 0.01$

Table A8: Stronger within hubs, but pattern persists in non-hubs

	(1)	(2)	(3)	(4)	(5)	(6)
		Hub			Not Hub	
	Seed	Series A	Series B	Seed	Series A	Series B
Dependent Variable		*Invest*			*Invest*	
Close Face Distance	0.0095	0.0029	0.0025	0.0198	0.0052	0.0117
	(0.0034)***	(0.0015)*	(0.0017)	(0.0110)*	(0.0059)	(0.0068)*
Mean of Invest	0.0204	0.0154	0.0206	0.0667	0.0613	0.0730
N	29679	98343	99340	6211	19939	19262
R-squared	0.02010	0.01981	0.02785	0.15566	0.11856	0.11672

Notes: Hub regions are defined as being in the top 3 states based on investment and deal amount: California, New York, and Massachusetts. Control variables used in Table 1 Column (7) are included in all regressions (not reported). Fixed effects for investor age quartile, investor state, startup fundraising year, and startup industry are included in the above regressions.
$* : p < 0.10; ** : p < 0.05; * * * p < 0.01$

Table A9: Robust to controlling for micro-geographic SES conditions

	(1) Baseline	(2) ZIP Subsample	(3) + Income	(4) + Unemployment	(5) + Children	(6) + State Diversity	(7) + State Poli.
Dependent Variable				*Invest = 1/0*			
Close Face Distance	0.0049	0.0038	0.0039	0.0039	0.0040	0.0040	0.0041
	(0.0012)***	(0.0020)*	(0.0020)*	(0.0020)*	(0.0020)**	(0.0020)**	(0.0020)**
Both AM orddiv			-0.0012	-0.0008	-0.0002	-0.0002	-0.0002
			(0.0058)	(0.0058)	(0.0057)	(0.0057)	(0.0058)
Both AM qualdiv			-0.0011	0.0009	-0.0012	-0.0012	-0.0010
			(0.0058)	(0.0058)	(0.0057)	(0.0057)	(0.0058)
Both AM businc			-0.0047	-0.0061	-0.0063	-0.0063	-0.0064
			(0.0021)**	(0.0022)***	(0.0021)***	(0.0021)***	(0.0022)***
Both AM unemp				-0.0066	-0.0072	-0.0072	-0.0069
				(0.0016)***	(0.0016)***	(0.0016)***	(0.0016)***
Both AM eitc				0.0193	0.0188	0.0187	0.0182
				(0.0041)***	(0.0041)***	(0.0041)***	(0.0041)***
Both AM child					0.0055	0.0056	0.0055
					(0.0014)***	(0.0014)***	(0.0014)***
Same State Diversity						0.0020	0.0032
						(0.0043)	(0.0043)
AM White State							0.1220
							(0.0404)***
Same State Poli. Lean							-0.0349
							(0.0107)***
Mean	0.0264	0.0254	0.0254	0.0254	0.0254	0.0254	0.0254
N	272774	81280	81280	81280	81280	81280	81280
R-squared	0.04625	0.04657	0.04686	0.04743	0.04771	0.04771	0.04873

Notes: All regressions include the covariates used in Column (4) of Table A3 and dummies for entrepreneurial age (by quartile), investor state, startup firm industry and financing year. A 30% subsample of the data had unique ZIP code data available, which was linked to IRS data on income tax filings to construct proxy measures for relative socioeconomic status. For these dyads, the effect of close face distance on investment is robust to controlling for being from ZIP codes with above the median (within the venture capital sample, not the U.S. population) ordinary and qualified dividends, and business income (Column 3); above the median unemployment or EITC filings; and above the median number of children per household. This was largely unaffected by adding in controls for the investor and entrepreneur being from states with the same ethnic diversity and political lean.

$* : p < 0.10; ** : p < 0.05; *** p < 0.01$

Table A10: Variation in face distance persists to controlling for education

	(1) Baseline	(2) Education Subsample	(3) + Same Degree	(4) + Same School
Dependent Variable			*Face Distance*	
Same Race	-0.0453	-0.0457	-0.0458	-0.0459
	(0.0055)***	(0.0105)***	(0.0104)***	(0.0104)***
Same Ethnicity (Surname Only)	0.0176	0.0109	0.0115	0.0115
	(0.0058)***	(0.0108)	(0.0106)	(0.0106)
Same Ethnicity (Full name)	-0.0188	-0.0235	-0.0258	-0.0258
	(0.0035)***	(0.0066)***	(0.0066)***	(0.0066)***
Same Gender	-0.1362	-0.1379	-0.1349	-0.1349
	(0.0045)***	(0.0077)***	(0.0077)***	(0.0077)***
Ent. More Attractive	0.0208	0.0272	0.0270	0.0272
	(0.0036)***	(0.0072)***	(0.0071)***	(0.0071)***
Same Features	-0.0557	-0.0587	-0.0584	-0.0581
	(0.0043)***	(0.0086)***	(0.0087)***	(0.0087)***
Same City	0.0040	0.0040	0.0036	0.0037
	(0.0029)	(0.0052)	(0.0051)	(0.0051)
Same State	-0.0034	0.0014	0.0012	0.0014
	(0.0027)	(0.0053)	(0.0054)	(0.0053)
Same Bach. Deg.			-0.0146	-0.0146
			(0.0053)***	(0.0053)***
Both have MBA			0.0121	0.0113
			(0.0065)*	(0.0066)*
Both study STEM			-0.0198	-0.0201
			(0.0057)***	(0.0058)***
Both have JD			-0.0584	-0.0595
			(0.0259)**	(0.0259)**
Both have PhD			-0.0458	-0.0462
			(0.0196)**	(0.0196)**
Same School				-0.0030
				(0.0045)
Same College				-0.0047
				(0.0156)
Mean	1.2728	1.2676	1.2676	1.2676
N	272774	25571	25571	25571
R-squared	0.10511	0.10515	0.10906	0.10913

Notes: Education data was collected for a 10% randomly drawn subsample of the data. This was used to generate indicator variables for attending the same school, same alma mater, studying a STEM major, and obtaining the same degree. Adding data on overlap degrees of study increases the R-squared of the model, after which adding addition covariates on school adds only a slight additional increase. Individuals that pursue the same undergraduate degree, study a STEM subject, pursue a JD or pursue a PhD are found to have a lower face distance (i.e. look more similar), which could be due to admissions selecting on the basis of a particular profile, or perhaps self-selection into certain fields of study that is correlated with some physical trait. $* : p < 0.10; ** : p < 0.05; * * * p < 0.01$

Table A11: Variation in face distance persists to controlling for image quality

	(1) Baseline	(2) + Image Size Issues	(3) + Image Quality Issues
Dependent Variable		*Face Distance*	
Same Race	-0.0457	-0.0458	-0.0455
	(0.0105)***	(0.0104)***	(0.0105)***
Same Ethnicity (Surname Only)	0.0109	0.0112	0.0110
	(0.0108)	(0.0107)	(0.0108)
Same Ethnicity (Full name)	-0.0235	-0.0236	-0.0235
	(0.0066)***	(0.0066)***	(0.0067)***
Same Gender	-0.1379	-0.1390	-0.1392
	(0.0077)***	(0.0076)***	(0.0076)***
Ent. More Attractive	0.0272	0.0281	0.0269
	(0.0072)***	(0.0072)***	(0.0074)***
Same Features	-0.0587	-0.0575	-0.0577
	(0.0086)***	(0.0086)***	(0.0086)***
Same City	0.0040	0.0047	0.0047
	(0.0052)	(0.0052)	(0.0052)
Same State	0.0014	0.0010	0.0013
	(0.0053)	(0.0053)	(0.0053)
Inv Image Small		0.0019	0.0015
		(0.0193)	(0.0192)
Ent Image Small		0.0219	0.0217
		(0.0047)***	(0.0047)***
Inv Image Blurry			0.0110
			(0.0087)
Ent Image Blurry			0.0030
			(0.0054)
Mean of Face Distance	1.2676	1.2676	1.2676
N	25571	25571	25570
R-squared	0.1052	0.1062	0.1065

Notes: Image size was collected based on the file size information. Image blurriness and exposure was calculated using the Microsoft Azure Face API and classified as blurry if the predicted image blur was high. None of the images in the sample were considered overexposed. Column 3 excludes any images where part of the face was obstructed by sunglasses.
$* : p < 0.10; ** : p < 0.05; * * * p < 0.01$

B Counterfactual Construction

Given that the data used in this study only observes actual investments between investors and entrepreneurs, counterfactual dyads had to be constructed to form the control group. In order to assuage the concern that such counterfactual construction is influencing the findings, four different sets of control dyads were developed and used to replicated the main analysis:

Counterfactual A This is the least restrictive set of controls. First, an investor was assumed to be actively investing in the period beginning in the year of his first observed investment (i.e. the maximum of 2001 and year of first investment - 1) and ending in the year after his last (minimum of 2017 and year of last investment +1). Second, observed investment activity was used to determine the financing stage preferences of each investor. This is used to restrict the set of counterfactual dyads to those where the startup financing stage accords with investor preferences. Third, observed investment activity and Preqin-classified industry categories[38] were used to identify the industry preferences of each investor. All investors that had observed deals in more than three different industries were classified as industry-agnostic. This is used to further restrict the set of counterfactual dyads to those where the startup firm is operating in an industry that accords with investor preferences. Fourth, observed investment activity was used to determine the geographic preferences of each investor at the state level. If an investor was observed to invest in more than three different states, then the investor was deemed location-agnostic. This resulted in a set of 1,800,574 control dyads for the 7,208 dyads with actual investments.

Counterfactual B This is the counterfactual group used in the paper and discussed in more detail in Section 3.4.1. It is more restrictive than set A as investors were defined

[38]Startup firms are classified by Preqin into 15 industries: business services, clean technologies, consumer discretionary, energy and utilities, food and agriculture, health care, industrials, infrastructure, internet, materials, other IT, real estate, semiconductors and electronics, software and related, and telecommunications.

as only active in the years where an investment was observed to occur. For instance, if a given investor made an investment in 2010 and another in 2016, rather than defining the active period as 2009 to 2017, this approach defined the investor as active only in years 2010 and 2016. Furthermore, this approach did not assume that investing in more than three industries or states meant that investors were agnostic on those factors. This resulted in a set of 265,566 control dyads for 7,208 dyads with actual investments.

Counterfactual C This approach used subindustries data from Preqin to further restrict the set of controls in B to those where investor *subindustry* preferences accorded with startup subindustries. As a concrete example, startups classified as being in the Internet industry, are further separated into subindustries that include the Search Engines, e-Financial, e-Commerce, Web Development, Multimedia and Graphics, Mobile Applications, and Email subindustry. This resulted in a set of 118,798 control dyads for 7,208 dyads with actual investments.

Counterfactual D This approach used city data collected from Crunchbase and hand-collected from web searches to further restrict the set of controls in B to those where investor geographic preferences at a *city* level matched startup locations. For instance, if an investor was observed to invest only in Austin, TX then startups based in Houston, Austin and San Antonio would be excluded from the control set. This resulted in a set of 106,566 control dyads for 7,208 dyads with actual investments.

Replicating the main analysis finds that the results are qualitatively similar across these different control groups. OLS regressions in Table B1 show that close face distance has a positive, statistically significant relationship with the likelihood of investment with a magnitude of 20.0% under A, 18.5% under B, 15.1% under C, and 14.7% under D (Columns 1-4). Table B2 shows that consistent with the findings in Table 4, close face distance becomes less

relevant as more information becomes available in later financing stages. Table B3 presents evidence consistent with Table 5 where older (above the median in age) investors are driving the effect of close face distance on investment. This is less pronounced under A (Column 1-2), relative to B (Table 5), C (Column 3-4) and D (Column 5-6). Finally, focusing only on these older investors in finds that this holds across investors at top and non-top venture capital firms, but has greater significance in the latter.

Table B1: Results hold across different control group definitions

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Counterfactual	A	B	C	D	A	B	C	D
Dependent Variable	*Invest = 1/0*				*Invest=1/0*			
Face Dist. (10th pctile)	0.0008	0.0049	0.0081	0.0093				
	(0.0002)***	(0.0012)***	(0.0024)***	(0.0027)***				
Face Distance (Raw)					-0.0021	-0.0117	-0.0188	-0.0201
					(0.0003)***	(0.0021)***	(0.0044)***	(0.0049)***
Same Ethnic. (Surname)	0.0003	0.0033	0.0053	0.0048	0.0003	0.0034	0.0054	0.0050
	(0.0003)	(0.0019)*	(0.0041)	(0.0048)	(0.0003)	(0.0019)*	(0.0041)	(0.0048)
Same Ethnic. (Full Name)	0.0006	0.0054	0.0110	0.0113	0.0006	0.0052	0.0108	0.0111
	(0.0001)***	(0.0010)***	(0.0022)***	(0.0028)***	(0.0001)***	(0.0010)***	(0.0022)***	(0.0028)***
Same Gender	-0.0000	0.0024	0.0047	0.0013	-0.0002	0.0013	0.0028	-0.0006
	(0.0002)	(0.0013)*	(0.0030)	(0.0035)	(0.0002)	(0.0014)	(0.0030)	(0.0036)
Same Features	0.0004	0.0024	0.0024	0.0083	0.0003	0.0021	0.0018	0.0078
	(0.0003)	(0.0020)	(0.0039)	(0.0051)	(0.0003)	(0.0020)	(0.0039)	(0.0051)
Ent. More Attractive	0.0001	-0.0001	-0.0044	-0.0110	0.0002	0.0000	-0.0042	-0.0108
	(0.0002)	(0.0014)	(0.0030)	(0.0044)**	(0.0002)	(0.0014)	(0.0030)	(0.0044)**
Same State	0.0050	0.0027	0.0067	0.0136	0.0050	0.0027	0.0067	0.0135
	(0.0002)***	(0.0015)*	(0.0030)**	(0.0030)***	(0.0002)***	(0.0015)*	(0.0030)**	(0.0030)***
Same City	0.0070	0.0170	0.0262	-0.0169	0.0070	0.0170	0.0263	-0.0168
	(0.0006)***	(0.0018)***	(0.0035)***	(0.0032)***	(0.0006)***	(0.0018)***	(0.0035)***	(0.0032)***
Mean	0.0040	0.0264	0.0538	0.0634	0.0040	0.0264	0.0538	0.0634
N	1807782	272774	126006	113774	1807782	272774	126006	113774
R-squared	0.00335	0.04625	0.12675	0.14228	0.00337	0.04632	0.12685	0.14236

Notes: All regressions include dummies for entrepreneurial age (by quartile), investor state, startup firm industry and financing year. Robust standard errors, clustered at the investor level, are reported in parentheses below each coefficient estimate.
$* : p < 0.10; ** : p < 0.05; * * * p < 0.01$

Table B2: Less relevant as more information available across different controls

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Seed	Seed	Seed	Series A	Series A	Series A	Series B	Series B	Series B
Counterfactual	A	A	A	C	C	C	D	D	D
Dependent Variable		*Invest = 1/0*			*Invest=1/0*				
Face Distance (10th pctile)	0.0029	0.0007	0.0006	0.0154	0.0056	0.0075	0.0171	0.0068	0.0087
	(0.0008)***	(0.0002)***	(0.0003)**	(0.0063)**	(0.0032)*	(0.0037)**	(0.0064)***	(0.0037)*	(0.0045)*
Mean of Invest	0.0058	0.0034	0.0042	0.0544	0.0470	0.0608	0.0555	0.0544	0.0765
N	176496	809458	821828	17804	55087	53115	18329	50310	45135
R-squared	0.00860	0.00369	0.00279	0.10605	0.12167	0.15217	0.12245	0.12848	0.17690

Notes: All regressions include the covariates used in Column (4) of Table A3 and dummies for entrepreneurial age (by quartile), investor state, startup firm industry and financing year. Robust standard errors, clustered at the investor level, are reported in parentheses below each coefficient estimate.
$* : p < 0.10; ** : p < 0.05; * * * p < 0.01$

Table B3: Variation by age generally holds across different controls

	(1)	(2)	(3)	(4)	(5)	(6)
Age	Below Med.	Above Med.	BM	AM	BM	AM
Counterfactual	A	A	C	C	D	D
Dependent Variable			*Invest = 1/0*			
Face Distance (10th pctile)	0.0007	0.0008	0.0023	0.0158	0.0021	0.0161
	(0.0003)**	(0.0002)***	(0.0027)	(0.0044)***	(0.0032)	(0.0047)***
Mean of Invest	0.0040	0.0040	0.0450	0.0631	0.0533	0.0736
N	761655	1046127	64620	61386	57591	56183
R-squared	0.00251	0.00460	0.09011	0.16694	0.09705	0.18728

Notes: All regressions include the covariates used in Column (4) of Table A3 and dummies for entrepreneurial age (by quartile), investor state, startup firm industry and financing year. Robust standard errors, clustered at the investor level, are reported in parentheses below each coefficient estimate.
$* : p < 0.10; ** : p < 0.05; *** p < 0.01$

Table B4: Variation by quality among older investors holds across different controls

	(1)	(2)	(3)	(4)	(5)	(6)
Quality	Top	Non-Top	T	NT	T	NT
Counterfactual	A	A	C	C	D	D
Dependent Variable			*Invest = 1/0*			
Face Distance (10th pctile)	0.0006	0.0009	0.0287	0.0148	0.0191	0.0164
	(0.0007)	(0.0003)***	(0.0126)**	(0.0045)***	(0.0110)*	(0.0050)***
Mean of Invest	0.0030	0.0041	0.0387	0.0664	0.0408	0.0787
N	104108	942019	7339	54047	7554	48629
R-squared	0.0017	0.0050	0.0860	0.1760	0.1045	0.1982

Notes: All regressions include the covariates used in Column (4) of Table A3 and dummies for entrepreneurial age (by quartile), investor state, startup firm industry and financing year. Robust standard errors, clustered at the investor level, are reported in parentheses below each coefficient estimate.
$* : p < 0.10; ** : p < 0.05; *** p < 0.01$

C Name-based Classification

Current race and ethnic classification methods in the social sciences generally fall into two categories: (1) self-reported measures, and (2) name-based classification methods.[39]

Self-reported measures usually come from government or researcher administered surveys, relying on respondents to label themselves as belong to one or more defined racial or ethnic groups. For example, the U.S. Census asks individuals to self-identify from five broad racial categories (White, Black or African American, American Indian or Alaska Native, Asian, Native Hawaiian or Other Pacific Islander) or to specify Other, and only since 2000, allow respondents to choose more than one category.[40] The National Institutes of Health follow the same standard with the requirement that research participants retain the right to indicate Unknown/Not Reported.[41] Similarly, medical records data also rely on self-identification by patients. For example, the U.S. National Statistics System suggests birth certificates include mothers' and where applicable, fathers' self-reported race from 15 race categories (although there is no uniform standard and local vital records offices have discretion).[42][43] Such self-reported data is limited by administration costs, lack of precision and data omissions. For example, the U.S. Census definition of Asian encompasses all of the Indian subcontinent and Southeast Asia. In doing so, it abstracts away from the heterogeneity within Asians, for example, by identifying those of Indian origin as the same race as those of Japanese origin. The U.S. Census has also had a growing percentage of citizens, including Hispanic, Latino, Arab and mixed race respondents, choose Other or not answer altogether.[44]

[39]It bears mention that a third, though nascent, approach is analyzing genetic markers and genetic clustering to determine race although this is currently costly and limited to health, epidemiology and broader biology research. The National Human Genome Research Institute and Howard University have an ongoing project that uses similarity in alleles to determine common genetic origins.

[40]The U.S. Census follows the definitions of ethnicity set by the federal Office of Management and Budget. https://census.gov/topics/population/race/about.html

[41]https://grants.nih.gov/grants/funding/women_min/race_ethnicity_qa.htm#3735

[42]https://www.cdc.gov/nchs/data/dvs/birth_edit_specifications.pdf

[43]https://oig.hhs.gov/oei/reports/oai-02-86-00001.pdf

[44]http://www.pewresearch.org/fact-tank/2014/03/14/u-s-census-looking-at-big-changes-in-how-it-asks-

Name-based classification methods originated in public health as a means to classify in the absence of self-reported data. The standard approach is to (1) aggregate a dataset of names labeled with known races/ethnicities to develop a frequency count, and apply a certain set of rules to assign the highest frequencies to corresponding races to produce a reference list; and (2) use this reference list to train a probabilistic classifier (Mateos (2007) provides a comprehensive review of the development of the methodology). Labeled datasets have traditionally come from administrative or health records such as death records (Coldman et al. (1988)) or electoral records (Mateos et al. (2007)). A more recent effort by Ambekar et al. (2009) and Ye et al. (2017) uses all the tagged entries of people on Wikipedia to as a training set to develop a classifier with improved precision and granularity relative to extant methods, which at the time of writing, is one of only few non-commercial APIs available.[45]

Limitations of name-based classification arise from outdated or insufficient training data sets and perhaps more importantly, the informational content of names themselves is heterogeneous across different regions, ethnic groups and time (Mateos (2007)). For example, in the United States, the challenge is disentangling white and black Americans that often share many common surnames (e.g. Smith, Johnson, Brown, Jones were both in the Top 5 of most common surnames for blacks and whites in the 2000 U.S. Census) which some scholars have attempted to deal with through the use of geocoding though this is most effective only in settings with high neighborhood segregation (Fiscella and Fremont (2006)). Furthermore, the propensity for immigrants to Anglicize or abbreviate first and surnames differs by origin and destination country, as well as time (Roberts, 2010)

Multi-ethnic names also result in classification errors either because of similar English translations or etymologies. Consider for example, the surname Lee which has origins in old English (*Leah*), Irish/Gaelic (*Laodaigh*), Korean, Chinese, and Norwegian (*Lie*). This

about-race-and-ethnicity/

[45]Available for open use here: http://www.name-prism.com/

is salient in surnames derived from Latin as they have spread through Romance languages, many of which were or continue to be *lingua francas*. Martin for example, a derivative of the Latin name *Martinus*, is a popular surname in Romantic language countries France, Spain and Portugal. Due to conquests and colonialism it is also a common surname in the United Kingdom, Canada, United States (where it is common for both white and African Americans) and Latin America. Moreover as social and cultural name conventions evolve, the relationship between name and ethnicity becomes increasingly less clear. For example, historically in many western Anglophone countries surname data was a relatively precise signal for men (McEvoy and Bradley (2006)), as women adopted their spouses' surnames upon marriage. Goldin and Shim (2004) found that this name convention is changing, experiencing sharp declines in popularity in the 1970s and 1980s followed by an increase in the 1990s. Additionally, with the rise in females as the primary earners in U.S. households (Rampell, 2013) and the legalization of same sex marriage, this convention is evolving further. The rise in global immigration and interracial marriage (Wang, 2015), and the prevalence of intercountry adoptions has further complicated the signal value of names.

To illustrate the limitations of the name-based ethnic classification, Table C1 presents within the sample of this study, a summary table that documents the share of investors and entrepreneurs that are clearly misclassified.

Table C1: Share of Individuals with Misclassified Ethnicity

		Surname Only	
		Correct	Wrong
Full name	Correct	93.29%	2.88%
	Wrong	2.36%	1.47%

(a) Investors

		Surname Only	
		Correct	Wrong
Full name	Correct	92.38%	3.24%
	Wrong	1.60%	2.78%

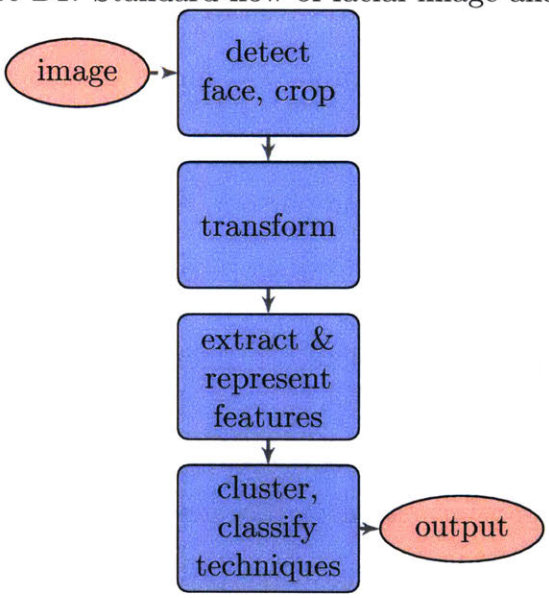
(b) Entrepreneurs

Notes: Individuals with misclassified ethnicity are identified using an ethnic classifier built on the OpenFace neural net and a training set of face photographs from Wikipedia, University of Massachusetts *Labeled Faces in the Wild* database, and the Columbia University FaceTracer Database labeled in six racial/broad ethnic categories (Europe, Black, American Indian and Alaskan Native, East Asian, South Asian, Arab and Hispanic/Latino). This was used to identify clear misclassifications where the prediction confidence was greater than or equal to 0.80, and the labelled ethnicity (based on names) differed from the assigned race (based on face photographs). This was then hand-verified.

D Extracting data from faces

Facial recognition technology in its current state can be thought of as two components: (1) face detection, and (2) identity verification. The former is a subset of object detection in computer vision and involves identifying human faces in images. It is fairly standard technology in photography (e.g. cameras, image processing) and video (e.g surveillance, video chat, security), and accessible with open source face detection software *dlib*[46] and *OpenFace.*[47] Once faces are detected, the latter part of face recognition is focused on linking known identities to faces by extracting and quantitatively representing facial attributes, and comparing it against known images. This is crucial to a range of applications such as detecting a user in their friend's photograph on social media, improving search results for images of a given person in an online search, identifying a target of interest on video surveillance or verifying that an individual at a digital border crossing matches the identity on a passport.

Figure D1: Standard flow of facial image analysis



[46]http://dlib.net/
[47]http://OpenFace.org/

Figure D1 depicts an example of the standard flow for face image analysis: raw images are processed using a face detector to crop the image to around the face and identify the fiducial points (e.g. positioning of the corners of eyes, mouth), images are then often transformed to a low-dimensional grayscale and normalized, attributes of the face are extracted and represented quantitatively using a neural net, and then depending on the type of analysis a clustering, classification or comparison algorithm is applied. The frontier of research in this space is focused on increasing the precision and speed at which images can be linked to identities, but the underlying data generated by such processes, such as the facial distance scores between two faces, is already a potentially valuable source of data.

D.1 Calculating Face Distance

The OpenFace deep neural net was used to implement a facial resemblance algorithm that compares the similarity between two images. OpenFace's neural net is built with Torch, Lua and luajit and trained on a publicly available database of 494,414 images of 10,575 individuals and FaceScrub (Ng and Winkler, 2014) database of 100,000 images of 530 individuals. With this trained neural net, each preprocessed image of a face is embedded into a 128-dimensional representation such that the distance between two face embeddings determines the probability that they are of the same person. Amos et al. (2016) provides a thorough description with further details but the basic approach that OpenFace takes is:

- (1) Detect a face using the face detector algorithm in *dlib* (King, 2009) that is pre-trained with high accuracy.
- (2) Conduct a 2-D transformation of the face into a standard format by using the face landmark detector in *dlib*'s real-time pose estimation (King, 2009; Kazemi and Sullivan, 2014) which identifies 68 feature positions on a face. The transformation rotates the face in the image to be frontal, crops it to just the face, normalizes the

position of the eyes and mouth to be uniform across all images, and resizes it to 96x96 pixels.

- (3) Use a deep neural network modeled after FaceNet which uses a triplet loss function to evaluate how accurately the network is in matching an image to its identity. A triplet loss function is also used in Google's FaceNet (Schroff et al., 2015). At a basic level, a triplet loss assesses accuracy by looking at a triplet of images: an image of a known person (anchor), an image of the same person (positive), and a image of a different person (negative). This triplet is then used to calculate Euclidean distance between the positive and anchor plus an additional threshold requirement, φ, is less than that between the positive and negative. From (Schroff et al. (2015) as reported in Amos et al. (2016)) the triplet loss function $\mathcal{L}$ is thus defined as:

$$\mathcal{L}(a, p, n) = ||f_\theta(a) - f_\theta(p)||_2^2 + \varphi - \max\{0, ||f_\theta(a) - f_\theta(n)||_2^2\}$$

where f_θ is the neural net θ representation of the image, and $||\cdot||_2^2$ is the squared Euclidean norm.

As above, in notation the algorithm computes $||f_\theta(a) - f_\theta(b)||_2^2$ where $f_\theta(a)$ is the representation of image a generated by a neural net f parameterized by θ. This generates an estimated score of facial distance between image a and b in the range of 0-4.0, where a lower score closer to 0 means the images are more similar. This score is typically used to match identities to images but it also allows for a comparison of facial resemblance.

D.2 Additional Face-based Variables

Ethnic Classifier The OpenFace pre-trained deep neural net is also used to train a racial/broad ethnic classification model. To build the training set for the race, a dataset

of racially labeled face images was compiled from three publicly available sources. First, the collection of *Lists of Americans* on Wikipedia[48], which lists notable Americans (e.g. military, scientists, writers, inventors) by race/ethnicity, was downloaded. To be conservative, images were only included for people where both parents are known and of the same race/ethnicity.[49] Second, a subset of the University of Massachusetts *Labeled Faces in the Wild* database (Huang et al., 2007)[50] was manually classified using Wikipedia and Google Search to verify the ethnicity of subjects. Third, the Columbia University FaceTracer Database was used to download real-world face images with race labels manually assigned by a group of researchers (Bitouk et al., 2008).[51] The result was 670 high-resolution images labeled as Europe, Black, American Indian and Alaskan Native, East Asian, South Asian, Arab and Hispanic/Latino. These were used to train a classifier that could predict the racial/ethnic category and produce a score of confidence in the estimate.

Attractiveness Classifier Physical attractiveness scores were collected using a similar approach (i.e. essentially a simplified implementation of Eisenthal et al. (2006)). To build the training set for attractiveness, the authors of the MIT *10k U.S. Adult Faces Database* (Bainbridge et al., 2013; Khosla et al., 2013) provided access to 2,222 natural face photographs rated attractive on a 1 (least attractive) to 5 (most attractive) scale by 12 participants of different age, race and gender per face. Mean attractiveness scores are aggregated and split into deciles according to the percentile distribution (i.e. 10th percentile in attractiveness, 20th percentile, and so on). These 10 percentile groups were used to train a classifier built on the OpenFace pre-trained neural net. This classifier was then used to estimate which percentile in attractiveness a face falls.

[48] https://en.wikipedia.org/wiki/Lists_of_Americans
[49] Multi-ethnic racial identification is a frontier topic. See Fu et al. (2014) for more details.
[50] Available at http://vis-www.cs.umass.edu/lfw/
[51] http://www.cs.columbia.edu/CAVE/databases/facetracer/