The Black Founder Dataset: Bridging Computer Vision and Self-Identification to Map Entrepreneurial Diversity

Emmanuel Yimfor*

September 24, 2025

Preliminary & Incomplete — Please Do Not Circulate

Abstract

We introduce the Black Founder Dataset, a comprehensive demographic identification infrastructure addressing a critical gap in entrepreneurship research: the absence of systematic racial data. While Crunchbase Diversity Spotlight, launched in 2020, attempts to provide diversity data through voluntary self-reporting, we demonstrate that such approaches suffer from severe selection bias that undermines research validity. Building on our methodology from Cook, Marx and Yimfor (2023), we classify the race of over 160,000 U.S. founders using computer vision and clerical review, achieving correlations of 0.92 with perceived race. By comparing Black founders identified through our comprehensive methodology with those self-reporting in Crunchbase Diversity Spotlight, we quantify the selection bias inherent in voluntary disclosure platforms. Our analysis reveals that only 3.47% of venture-funded founders are Black, substantially below widely cited industry statistics. The dataset will enable researchers to move beyond biased self-reported data and provide media outlets with accurate demographic statistics to replace the current reliance on incomplete Crunchbase figures.

Keywords:
JEL Classification:

1. Introduction

Studies of venture capital funding for high-growth startups typically lack demographic data, particularly regarding race. As noted in our forthcoming Journal of Finance paper,

*Columbia Business School, Columbia University; emmanuel.yimfor@columbia.edu

"databases of high-potential ventures typically do not report the founders' race or ethnicity" (Cook et al., 2023). This data gap has led researchers and media outlets to rely on incomplete and potentially biased sources for demographic statistics in entrepreneurship.

The most prominent attempt to address this gap is Crunchbase Diversity Spotlight, launched in August 2020, which enables company representatives to self-report diversity characteristics of their leadership teams. However, this voluntary self-reporting mechanism introduces systematic selection bias that might invalidates much existing research and media coverage on minority entrepreneurship.

2. The Problem with Self-Reported Demographic Data

2.1. Crunchbase Diversity Spotlight: Promise and Limitations

Crunchbase Diversity Spotlight represents the industry's primary source for diversity statistics, widely cited in media reports claiming Black entrepreneurs receive only 1.2% of venture funding. The platform allows verified company representatives to add diversity tags indicating if founders or CEOs identify with underrepresented racial, ethnic, or gender groups. As of 2021, only 1,393 contributors had added diversity data, updating 23,604 profiles—a fraction of the hundreds of thousands of startups in the Crunchbase database.

The fundamental limitation of this approach is selection bias. Companies choosing to self-report diversity characteristics differ systematically from those that do not. Our preliminary analysis suggests that founders who publicly disclose racial identity through Diversity Spotlight are:

- Disproportionately concentrated in social impact sectors where diversity is valued
- More likely to have elite educational credentials that offset potential discrimination
- Located in geographic markets with stronger diversity initiatives
- At funding stages where diversity disclosure provides strategic advantages

2.2. The Consequences of Selection Bias

This selection bias has profound implications for research and policy. Studies relying on Crunchbase Diversity Spotlight data are not examining Black entrepreneurship comprehensively but rather a specific subset of Black entrepreneurs who calculate that public racial identification provides net benefits. Media reports citing these statistics systematically undercount Black founders who choose not to self-identify, while potentially overrepresenting those in diversity-friendly sectors or geographies.

3. Our Methodological Approach

To address these limitations, we construct a comprehensive dataset classifying the race of over 160,000 unique founders and investors. Our methodology, detailed in Cook et al. (2023), employs:

3.1. Comprehensive Identification

Rather than relying on voluntary self-reporting, we systematically classify all founders in our sample through:

- Collection of profile photos from LinkedIn, company websites, and social media (coverage increasing from 84% to over 95% during our sample period)
- Computer vision algorithms combined with clerical review
- Validation showing 0.92 correlation with perceived race versus 0.86 with self-reported race

This approach captures both founders who would self-identify publicly and those who would not, eliminating the selection bias inherent in voluntary disclosure systems.

3.2. Comparative Analysis Framework

By the time of conference presentation, we will have completed a systematic comparison of:

- Black founders identified through our comprehensive methodology
- Black founders self-reporting through Crunchbase Diversity Spotlight
- Characteristics differentiating these populations

This comparison will quantify the magnitude and direction of selection bias in self-reported data, providing researchers with correction factors for existing studies.

4. Dataset Scope and Applications

The Black Founder Dataset encompasses:

- Over 160,000 founders with racial classification (2000-2020)
- Integration with rich resume data including educational background and professional experience
- Network scores measuring professional and educational overlap with VC partners
- Temporal coverage enabling longitudinal analysis of diversity trends

4.1. Replacing Biased Industry Statistics

Our primary contribution is providing unbiased demographic statistics to replace the widely cited but flawed Crunchbase figures. We find that only 3.47% of founders who raise VC funding are Black—a fourfold underrepresentation relative to working adults. This figure, based on comprehensive identification rather than self-selection, provides a more accurate baseline for policy discussions.

4.2. Enabling Rigorous Research

The dataset enables researchers to:

- Study the true extent of racial disparities in venture funding without selection bias
- Analyze strategic patterns in racial disclosure across sectors and geographies
- Examine how funding outcomes vary between founders who self-identify versus those who do not
- Track temporal changes in diversity without confounding changes in disclosure propensity

5. Long-Term Impact and Dissemination

5.1. Publishing Comprehensive Industry Statistics

Our long-term goal is to publish regular reports on fundraising statistics by race and investment firm, providing media outlets and policymakers with accurate, comprehensive data. These reports will:

- Replace reliance on incomplete Crunchbase Diversity Spotlight figures
- Track progress (or lack thereof) in diversity initiatives over time
- Identify investment firms and sectors leading or lagging in diversity
- Provide granular geographic and temporal breakdowns

5.2. Open Science Commitment

Following the December 2025 meeting, the dataset will be publicly available, enabling:

- Replication and extension of our findings
- Integration with other innovation datasets such as LinkedIn
- Validation and improvement of identification methodologies

6. Preliminary Findings

Our analysis reveals several patterns obscured by selection bias in self-reported data:

6.1. The True Funding Gap

Black-owned startups raise approximately half as much venture capital as other startups formed in the same year, industry, and location. This gap persists after controlling for observable founder and startup characteristics, suggesting gaps beyond differences in human capital or geography.

6.2. Strategic Non-Disclosure

A substantial fraction of Black founders identified through our methodology do not appear in Crunchbase Diversity Spotlight, suggesting strategic non-disclosure. These "invisible founders" are:

- More likely to be in traditional technology sectors
- Less likely to have elite educational credentials
- More geographically dispersed
- At earlier funding stages

7. Conclusion

The Black Founder Dataset addresses a fundamental limitation in entrepreneurship research: the absence of comprehensive demographic data. By moving beyond voluntary self-reporting systems like Crunchbase Diversity Spotlight, we eliminate selection bias that has distorted our understanding of diversity in venture capital. Our finding that only 3.47% of venture-funded founders are Black—based on systematic identification rather than self-selection—provides a more accurate foundation for research and policy.

As we prepare to compare our comprehensive identification with Crunchbase Diversity Spotlight data, we anticipate revealing the magnitude of selection bias in current industry statistics. This comparison will demonstrate why media outlets and researchers must move beyond self-reported data to understand true patterns of diversity and capital allocation in entrepreneurship. By publishing regular, comprehensive demographic statistics, we aim to replace the flawed but widely cited Crunchbase figures with rigorous, unbiased measurement of funding disparities in venture capital.

References

Cook, Lisa D, Matt Marx, and Emmanuel Yimfor, "Funding Black High-Growth Startups," 2023. Accepted, Journal of Finance.