The Invisible Majority: Selection Bias in Self-Reported Data

Emmanuel Yimfor, Columbia Business School

NBER I3 Technical Working Group Meeting, Fall 2025

Self-reported diversity data are increasingly central to research and policy

- **Research**: Survey-based studies, voluntary disclosure databases
  - Kauffman Firm Survey: racial funding gaps (Fairlie et al. 2022, 2023; Morazzoni & Sy 2022)
  - Small Business Credit Survey: credit access (Barkley & Schweitzer 2022)
  - Annual Business Survey: innovation and VC (Wojan 2024)
  - Crunchbase Diversity Spotlight: “Black founders receive less than 1% of VC”

Self-reported diversity data are increasingly central to research and policy

- **Research**: Survey-based studies, voluntary disclosure databases
  - Kauffman Firm Survey: racial funding gaps (Fairlie et al. 2022, 2023; Morazzoni & Sy 2022)
  - Small Business Credit Survey: credit access (Barkley & Schweitzer 2022)
  - Annual Business Survey: innovation and VC (Wojan 2024)
  - Crunchbase Diversity Spotlight: “Black founders receive less than 1% of VC”

- **Policy**: California SB 54 (signed October 2023)
  - First law requiring VCs to collect founder diversity data
  - Founder participation is **explicitly voluntary**
  - VCs cannot incentivize or influence responses
  - First reports due **April 2026**

Self-reported diversity data are increasingly central to research and policy

- **Research**: Survey-based studies, voluntary disclosure databases
  - Kauffman Firm Survey: racial funding gaps (Fairlie et al. 2022, 2023; Morazzoni & Sy 2022)
  - Small Business Credit Survey: credit access (Barkley & Schweitzer 2022)
  - Annual Business Survey: innovation and VC (Wojan 2024)
  - Crunchbase Diversity Spotlight: "Black founders receive less than 1% of VC"

- **Policy**: California SB 54 (signed October 2023)
  - First law requiring VCs to collect founder diversity data
  - Founder participation is **explicitly voluntary**
  - VCs cannot incentivize or influence responses
  - First reports due **April 2026**

- **The problem**: Selection into self-reporting is unlikely to be random
  - If founders who opt in differ systematically, conclusions may be biased
  - Direction not obvious: Garcia & Darity (2022) find Black PPP borrowers who disclosed race received **52% less** funding (negative selection)

This paper: First systematic analysis of selection bias in diversity data

- **Setting**: Crunchbase Diversity Spotlight
  - Largest voluntary diversity database for startups
  - Launched August 2020, weeks after George Floyd's murder
  - Source of "Black founders receive 1% of VC" statistic

- **Key advantage**: I can observe the **full population**
  - Algorithmic classification identifies 8,564 Black founders
  - Only 1,077 (12.6%) appear in Diversity Spotlight
  - Compare self-reporters to non-self-reporters

This paper: First systematic analysis of selection bias in diversity data

- **Setting**: Crunchbase Diversity Spotlight
  - Largest voluntary diversity database for startups
  - Launched August 2020, weeks after George Floyd's murder
  - Source of "Black founders receive 1% of VC" statistic

- **Key advantage**: I can observe the **full population**
  - Algorithmic classification identifies 8,564 Black founders
  - Only 1,077 (12.6%) appear in Diversity Spotlight
  - Compare self-reporters to non-self-reporters

- **Questions**:
  - Who selects into self-reporting?
  - How does selection bias affect funding gap estimates?
  - What does this imply for policy (SB 54) and research?

Crunchbase Diversity Spotlight: Institutional origins matter

- Launched **August 11, 2020**, nine weeks after George Floyd's murder
- Official launch partners: **Backstage Capital, Harlem Capital, BLCK VC, All Raise, Precursor Ventures**
  - These diversity-focused VCs encouraged portfolio companies to register
- Participation is entirely **voluntary** and requires verified employee to add tags
- Coverage: Only **71,793 U.S. companies** (2-3% of database) have diversity tags
- Crunchbase acknowledges: "Voluntary or selective reporting bias... may paint a rosier picture"

Methodology: Race Classification

Methodology → Preview → Results → Mechanism

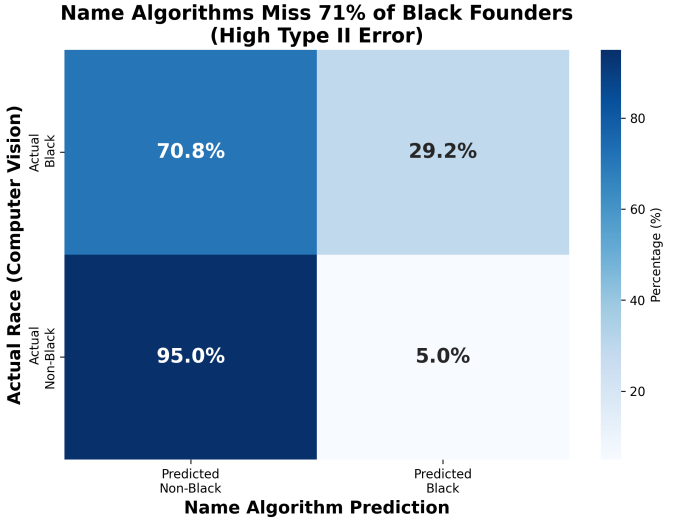
What is the goal? Perceived race, not self-reported race

- In some financing contexts (mortgage lending), race is explicit via self-report
- Probably **not** the case in startup funding
  - Investors likely respond to perceived race, not self-reported race

What is the goal? Perceived race, not self-reported race

- In some financing contexts (mortgage lending), race is explicit via self-report
- Probably **not** the case in startup funding
  - Investors likely respond to perceived race, not self-reported race
- Crunchbase Diversity Spotlight
  - Firm-level data crowdsourced from venture investors + others
  - Outreach methodology, response rates unknown
  - Even given perfect methods, measures **self-report** not **perceived** race
- **Our approach**: Determine perceived race of founders using images

Name classification algorithms alone have high Type II error rates



Data construction: Combine algorithms with clerical review

1. **Step 1**: DeepFace + NamePrism algorithmic classification
   - Images from LinkedIn profiles linked in Crunchbase
   - Initial algorithmic assignment for 174,347 founders
   - Algorithm identifies ~6,400 potential Black founders

Data construction: Combine algorithms with clerical review

1. **Step 1**: DeepFace + NamePrism algorithmic classification
   - Images from LinkedIn profiles linked in Crunchbase
   - Initial algorithmic assignment for 174,347 founders
   - Algorithm identifies ~6,400 potential Black founders
2. **Step 2**: Manual review of all algorithmically-classified Black founders

Data construction: Combine algorithms with clerical review

1. **Step 1**: DeepFace + NamePrism algorithmic classification
   - Images from LinkedIn profiles linked in Crunchbase
   - Initial algorithmic assignment for 174,347 founders
   - Algorithm identifies ~6,400 potential Black founders
2. **Step 2**: Manual review of all algorithmically-classified Black founders
3. **Step 3**: Manual review of non-Black founders for false negatives (~2,100 additional)
   - Affinity groups (Nigerian Leadership Initiative), news reports
   - Crowd-sourced lists of Black founders, HBCU attendance

Data construction: Combine algorithms with clerical review

1. **Step 1**: DeepFace + NamePrism algorithmic classification
   - Images from LinkedIn profiles linked in Crunchbase
   - Initial algorithmic assignment for 174,347 founders
   - Algorithm identifies ~6,400 potential Black founders
2. **Step 2**: Manual review of all algorithmically-classified Black founders
3. **Step 3**: Manual review of non-Black founders for false negatives (~2,100 additional)
   - Affinity groups (Nigerian Leadership Initiative), news reports
   - Crowd-sourced lists of Black founders, HBCU attendance
4. **Final sample**: 8,564 Black founders identified
   - ~75% from algorithmic classification + clerical review
   - ~25% from manual search for false negatives

Step 1: Algorithmic classification via images

- DeepFace assigns each image to highest probability folder: **Black, White, Asian, Hispanic**

Example 1: 100% Black

Example 2: 32% Black

Initially in *White folder* $\rightarrow$ **Moved to Black** after

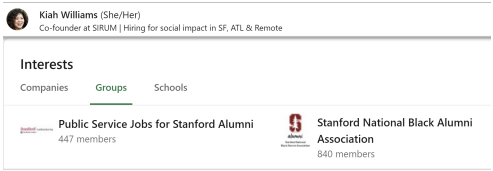
Step 3: Profile-based reclassification catches false negatives

- Systematic search for founders **missed by algorithm** (~2,100 additional Black founders)

Kiah Williams

Algorithm classified as *Asian*

LinkedIn Profile Evidence



Stanford National Black Alumni Association → Reclassified as Black

Does our method correlate with perceived or self-reported race?

- Validate using Chicago Face Database (Ma, Correll, Wittenbrink 2015)
  - 197 Black and 183 White faces with both actual and perceived race
  - Perceived race collected from ~43 participants per image

- We classify CFD images and correlate vs. Perceived & Self-reported race:

	Black Image	White Image
DeepFace correlation with **perceived** race	0.94	0.88
DeepFace correlation with **self-reported** race	0.86	0.79

- Our method captures **perceived race** better than self-reported race

Preview of Findings

Methodology → Preview → Results → Mechanism

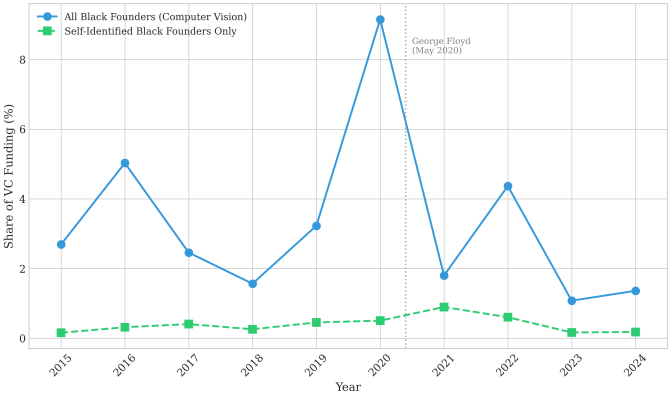
Preview: Self-reported data severely misrepresent Black founder funding

1. Self-reported data miss **87% of Black founders**
   - Only 1,077 of 8,564 Black founders appear in Diversity Spotlight
2. Black founders receive **2.4% of VC funding**, not the 0.9% self-reported data suggest
3. The funding gap is **larger** than self-reported data imply
   - Extensive margin: 8.7pp reduction in funding probability (vs. +19pp with self-ID data)
   - Intensive margin: 70% less funding conditional on raising (vs. 23%)
4. Self-identified founders **underperform** controlling for funding raised
5. **Why?** Selection is driven by **investor networks**
   - Launch partners encouraged portfolio companies to register

Results: Selection Bias in Funding Data

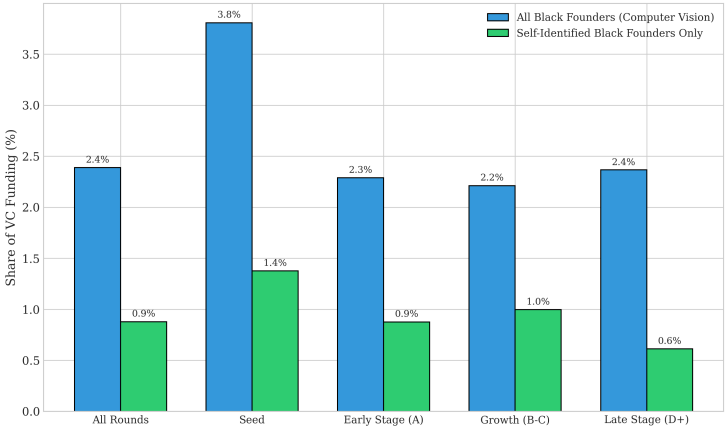
Methodology → Preview → Results → Mechanism

Black founders' share of funding varies substantially over time



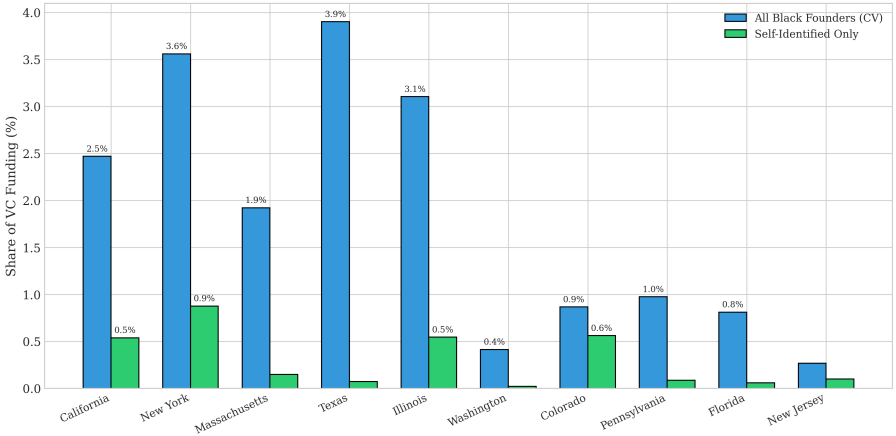
- Self-reported data would capture **less than one-fifth** of Black founder funding in most years

The gap between algorithmic and self-ID is largest at late stages



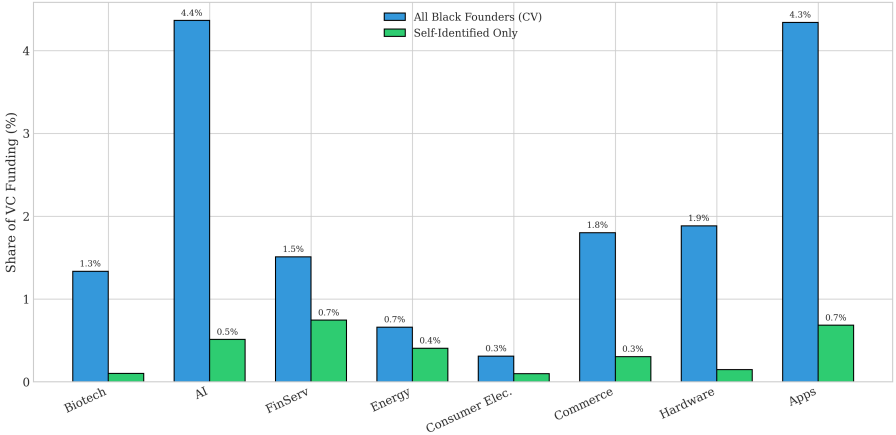
- Black founders receive their highest share at seed (3.8%), declining at later stages
- Self-reported data captures only **26% of Black founder funding** at Series D+

Geographic variation in Black founder funding shares



- Texas (3.9%), New York (3.6%), and Illinois (3.1%) show highest Black founder shares
- Self-reported data captures as little as **3% of Black founder funding** in Texas

Industry variation reveals systematic undercounting



- Black founders receive highest share in AI (4.4%), Hardware (1.9%), Commerce (1.8%)
- Gap is particularly large in Biotech (1.3% vs. 0.1%) and Hardware (1.9% vs. 0.2%)

Self-identified founders are positively selected on observables

	Self-ID N=1,077	Non-Self-ID N=7,487	t-stat
Panel A. Founder Characteristics			
I(Bachelor's Degree)	0.75	0.66	5.36***
I(MBA)	0.18	0.13	3.71***
I(Attended Top School)	0.33	0.19	8.66***
Prior Startups	0.53	0.32	7.05***
Panel B. Startup Outcomes			
I(Raised Any Funding)	0.73	0.25	30.45***
Total Raised ($M)	11.73	1.94	17.23***
I(Good Exit)	0.01	0.01	2.01**

- Self-identified Black founders are:
  - More likely to have **bachelor's degree** (75% vs. 66%)
  - More likely to have **attended top school** (33% vs. 19%)
  - More likely to be **serial founders** (0.53 vs. 0.32 prior startups)
- They also raise **6x more funding** ($11.7M vs. $1.9M)

What is the Black funding gap? Algorithmic vs. Self-ID

$$\mathbf{Y} = \beta_1 \text{P(Black)} + \beta_2 \text{Controls} + \lambda_{\text{state}} + \gamma_{\text{industry}} + \eta_{\text{year}}$$

- **Y**: I(Raised VC) or Ln(VC Funding)
- **P(Black)**: Algorithmic classification OR Diversity Spotlight indicator
- **Controls**: Education, experience, serial founder status, age
- **Key comparison**: Same regression, different race measurement
- **Sample**: 174,347 founders, 2000–2024

Self-reported data produce *wrong-signed* extensive margin estimates

	Algorithmic		*Self-ID Only*	
	I(Raised VC)	Ln(Funding)	I(Raised VC)	Ln(Funding)
P(Black)	-0.087***	-1.205***	**0.193*** **	-0.266**
	(0.005)	(0.075)	(0.020)	(0.117)
Obs.	174,347	46,431	174,347	46,431
State, Year, Industry FE	Yes	Yes	Yes	Yes

- Algorithmic: Black founders are **8.7pp less likely** to raise VC

Self-reported data produce *wrong-signed* extensive margin estimates

	Algorithmic		*Self-ID Only*	
	I(Raised VC)	Ln(Funding)	I(Raised VC)	Ln(Funding)
P(Black)	-0.087***	-1.205***	**0.193*** **	-0.266**
	(0.005)	(0.075)	(0.020)	(0.117)
Obs.	174,347	46,431	174,347	46,431
State, Year, Industry FE	Yes	Yes	Yes	Yes

- Algorithmic: Black founders are **8.7pp less likely** to raise VC
- Self-ID data: Black founders appear **19pp MORE likely** to raise VC
  - The **sign flips** because self-ID founders are positively selected

Self-reported data produce *wrong-signed* extensive margin estimates

	Algorithmic		*Self-ID Only*	
	I(Raised VC)	Ln(Funding)	I(Raised VC)	Ln(Funding)
P(Black)	-0.087***	-1.205***	**0.193*** **	-0.266**
	(0.005)	(0.075)	(0.020)	(0.117)
Obs.	174,347	46,431	174,347	46,431
State, Year, Industry FE	Yes	Yes	Yes	Yes

- Algorithmic: Black founders are **8.7pp less likely** to raise VC
- Self-ID data: Black founders appear **19pp MORE likely** to raise VC
  - The **sign flips** because self-ID founders are positively selected
- Intensive margin: Both show negative gap, but self-ID understates by **78%**

The funding gap varies by stage

	Seed		*Early*		*Later*	
	I(Raised)	Ln(Amt)	I(Raised)	Ln(Amt)	I(Raised)	Ln(Amt)
Panel A: Algorithmic Classification						
P(Black)	-0.019***	-0.861***	-0.056***	-0.444***	-0.014***	0.037
Panel B: Self-ID Only						
P(Black)	0.290***	-0.198**	0.047***	-0.275**	0.001	-0.049

- Algorithmic shows funding gap **persists across all stages**
- Self-ID shows **wrong-signed estimates** at seed and early stages
- The bias is most severe where most Black-founded startups operate (seed stage)

Do self-identified founders perform better? Outcome test

	Good Exit		*Any Exit*		*IPO*
	(1)	(2)	(3)	(4)	(5)
I(Self-ID)	0.002	**-0.018*** **	0.001	**-0.022*** **	**-0.015*** **
	(0.003)	(0.007)	(0.004)	(0.007)	(0.005)
Ln(Total Raised)		0.012***		0.015***	0.015***
Controls for Funding	No	Yes	No	Yes	Yes

- Without controls: No significant difference in exit rates

Do self-identified founders perform better? Outcome test

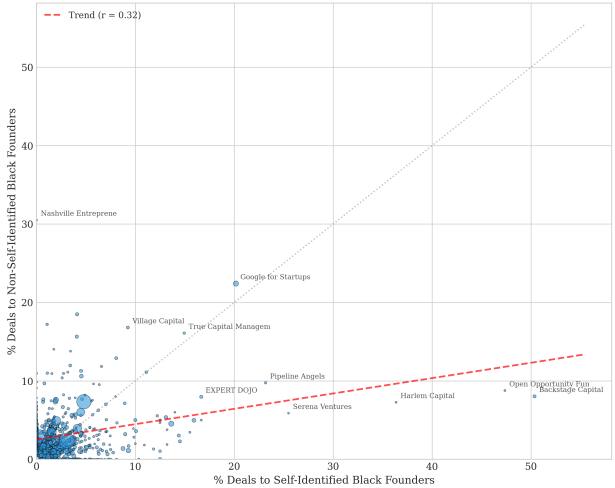
	Good Exit		*Any Exit*		*IPO*
	(1)	(2)	(3)	(4)	(5)
I(Self-ID)	0.002	**-0.018*****	0.001	**-0.022*****	**-0.015*****
	(0.003)	(0.007)	(0.004)	(0.007)	(0.005)
Ln(Total Raised)		0.012***		0.015***	0.015***
Controls for Funding	No	Yes	No	Yes	Yes

- Without controls: No significant difference in exit rates
- Controlling for funding raised: Self-identified founders are **1.8pp less likely** to achieve good exit
- Interpretation: Self-identified founders **underperform** relative to capital deployed
- Suggests “visible” are overfunded relative to the “invisible” majority

Mechanism: Why Selection Bias?

Methodology → Preview → Results → Mechanism

Why does self-reported data exhibit such severe selection bias?



- Positive correlation (r=0.32) between backing self-ID and non-self-ID Black founders
- Investors far to the right (Backstage, Harlem Capital) were **Diversity Spotlight launch partners**
  - High self-ID rates reflect program participation
- Investors above 45-degree line back Black founders regardless of self-ID status

Selection is driven by investor networks, not founder quality

Dep. Var.:	I(Self-ID in Diversity Spotlight)		
	(1)	(2)	(3)
I(Launch Partner Backed)	0.675***	0.251***	0.255***
	(0.031)	(0.043)	(0.046)
I(Diversity Investor Backed)		0.443***	0.273***
		(0.032)	(0.035)
Sample	All	All	Funded
Founder Controls	No	Yes	Yes

- Launch partner backing increases self-ID probability by **67.5pp**
  - Baseline rate: 11.9%
  - A **six-fold increase**

Selection is driven by investor networks, not founder quality

Dep. Var.:	I(Self-ID in Diversity Spotlight)		
	(1)	(2)	(3)
I(Launch Partner Backed)	0.675***	0.251***	0.255***
	(0.031)	(0.043)	(0.046)
I(Diversity Investor Backed)		0.443***	0.273***
		(0.032)	(0.035)
Sample	All	All	Funded
Founder Controls	No	Yes	Yes

- Launch partner backing increases self-ID probability by **67.5pp**
  - Baseline rate: 11.9%
  - A **six-fold increase**
- Selection is driven by **investor networks**, not founder quality
- Self-ID founders are overfunded because they're backed by connected VCs

Conclusion

Conclusion: Self-reported data severely undercount Black founders

Conclusion: Self-reported data severely undercount Black founders

- Self-reported data miss **87% of Black founders** in the startup ecosystem
  - Black funding share understated by factor of 2.7 (0.9% vs. 2.4%)
  - Extensive margin estimates have the **wrong sign**
- Selection into self-reporting is **strongly positive** (opposite of PPP loans)
  - Self-identified founders have better education, more experience, raise more capital
  - But they **underperform** controlling for funding raised
- Implications for research and policy:
  - Self-reported diversity data may not reliably measure true funding gaps
  - Algorithmic methods may provide more accurate estimates

Data and Code Availability

- **Race classification methodology and data**:
  - Cook, Marx, and Yimfor (2022): “Funding Black High-Growth Startups”
  - Forthcoming, *Journal of Finance*
  - Code: `https://github.com/eyimfor/race_classifier_fbhgs`

- **Crunchbase Diversity Spotlight**:
  - `https://www.crunchbase.com/discover/diversity-spotlight`
  - Self-reported diversity tags available via Crunchbase Pro

- **DeepFace library**:
  - Serengil and Ozpinar (2020)
  - `https://github.com/serengil/deepface`

emmanuel.yimfor@columbia.edu

Appendix: Launch partners drive selection into Diversity Spotlight

Dependent Variable:	**I(Self-Identified in Diversity Spotlight)**				
Sample:	*All Black-Founded Startups*				*Funded Only*
	(1)	(2)	(3)	(4)	(5)
I(Launch Partner Backed)	0.675***	0.667***	0.246***	0.251***	0.255***
	(0.031)	(0.031)	(0.044)	(0.043)	(0.046)
I(Diversity Investor Backed)			0.456***	0.443***	0.273***
			(0.032)	(0.032)	(0.035)
Observations	6,607	6,607	6,607	6,607	2,574
State, Year, Industry FE	Yes	Yes	Yes	Yes	Yes
Founder Controls	No	Yes	No	Yes	Yes

- Launch partner backing increases self-ID probability by **67.5pp** (baseline: 11.9%)
- Selection is driven by **investor networks**, not founder characteristics

Appendix: Investment activity of self-identified Black-led VC firms

		Black-Founded Deals			
Investor	**Total Deals**	**Total**	**Self-ID**	**Non-Self-ID**	**% Self-ID**
Precursor Ventures	352	285	35	250	12.3%
Kapor Capital	188	172	25	147	14.5%
MaC Venture Capital	182	163	23	140	14.1%
Backstage Capital	144	133	63	70	47.4%
Harlem Capital Partners	64	59	18	41	30.5%
Serena Ventures	60	54	12	42	22.2%
Total (All 113 Investors)	4,937	4,305	410	3,895	9.5%

- Even among self-identified Black-led VCs, only **9.5%** of their Black-founded deals are with self-ID startups
- Backstage Capital (47%) and Harlem Capital (31%) have highest rates; they were launch partners