A Longitudinal Panel of GitHub Engineering Velocity for Venture-Backed Startups: Dataset and Early Observations

Maryan Kindrat

signals@gitdealflow.com

https://orcid.org/0009-0002-2222-4112

Research Article

Keywords: venture capital, alternative data, GitHub, open source, engineering velocity, startup analytics, panel data

Posted Date: August 22nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10745326/v1

Additional Declarations: The authors declare potential competing interests as follows: The corresponding author operates the VC Deal Flow Signal product and declares this commercial interest. No other competing interests exist.

A Longitudinal Panel of GitHub Engineering Velocity for Venture-Backed Startups

Dataset and Early Observations

Author: Maryan Kindrat, VC Deal Flow Signal (signal@gitdealflow.com)

Date: 2026-04-19 **Revised:** 2026-08-18

Version: 1.0.1

Dataset DOI: 10.5281/zenodo.19650920

Abstract

We release a quarterly longitudinal panel of GitHub engineering-velocity signals across 55 venture-backed startups in 20 sectors, spanning five quarters from Q2 2025 through Q2 2026 (219 startup-period observations). For each observation we record commit velocity over a rolling 14-day window, unique-contributor count, new-repository creation, and a deterministic classification into one of four acceleration patterns: framework migration, engineering hiring burst, infrastructure buildout, and deploy frequency spike. We describe the data-collection methodology, report descriptive statistics across sectors and geographies, and document known limitations, most importantly the absence of linked funding-event labels in this release. Observed 14-day commit velocity has a median of 71, a mean of 173, and a 90th percentile of 392; quarter-over-quarter velocity change ranges from -94% to +1,647% with 49% of observations showing positive growth. Framework migration dominates the signal distribution at 75% of observations. The dataset is distributed under CC BY 4.0 at https://signals.gitdealflow.com/api/signals.csv and mirrored on Kaggle, Data.world, and Zenodo. We invite replication studies joining the panel against external funding-event sources.

Keywords: venture capital; alternative data; GitHub; open source; engineering velocity; startup analytics; panel data.

JEL codes: G24; L26; O32; C81.

1. Introduction

Venture capital deal sourcing is shifting from network-mediated referral flow toward alternative-data-driven inbound. Investor surveys consistently rank proprietary data and network signals above cold inbound as sources of the highest-quality deals (Gompers, Gornall, Kaplan, and Strebulaev, 2020). Among the signals most promising for early detection of breakout startups is the public engineering footprint on GitHub: commit activity, contributor growth, and repository creation. These signals are observable before most fundraise announcements and, in aggregate, are harder to game than self-reported metrics.

The use of GitHub data as a leading indicator of company outcomes is an area of growing but still thinly documented practice. Existing research focuses on individual-developer productivity, on project-level health metrics for open-source governance, or on post-IPO firm-level studies. Public longitudinal panels linking organizational-level GitHub activity to venture-capital outcomes are rare.

This paper releases one such panel. We describe:

1. The sector-stratified seed list of 55 venture-backed startup GitHub organizations.
2. A deterministic collection pipeline built over the GitHub REST API that captures rolling 14-day commit velocity, contributor counts, and repository-creation events per organization per quarter.
3. A classification scheme that labels each observation into one of four human-readable acceleration patterns.
4. Descriptive statistics across sectors, geographies, and periods.
5. Known limitations and a roadmap for linking observations to funding events in subsequent data releases.

We intend this release as a public baseline. It is distributed under CC BY 4.0 and is designed for replication, extension, and joining against external ground-truth sources such as Crunchbase, PitchBook, or Dealroom.

2. Related work

The literature most directly related to this dataset sits at the intersection of three strands.

Venture-capital decision-making. Kaplan and Strömberg (2004) document the heuristics VCs apply when screening early-stage opportunities; Gompers et al. (2020) update this evidence with a large survey of practicing investors and highlight the role of proprietary sourcing as the dominant channel for highest-quality deal flow. Alternative data (signals collected outside traditional filings and founder self-reports) has become the mechanism by which many sourcing teams attempt to industrialize that proprietary edge.

Open-source development dynamics. A large body of empirical software engineering research characterizes commit activity, contributor growth, and repository structure as observables with predictable temporal dynamics (Mockus, Fielding, and Herbsleb, 2002; Vasilescu, Filkov, and Serebrenik, 2013). This work establishes that organizational-level GitHub activity is noisy at daily timescales but stable at two-week or longer windows, a finding that directly informs our choice of a rolling 14-day observation window.

Alternative data in finance. A growing literature examines the information content of non-traditional signals, web traffic, app downloads, satellite imagery, hiring postings, for firm-level outcomes. Public-GitHub activity has received attention as a proxy for engineering hiring and product velocity, though systematic longitudinal panels at venture scale are not, to our knowledge, publicly available.

3. Data collection

3.1 Seed list

The 55 organizations in the dataset were selected to satisfy three criteria: (a) a primary GitHub presence with at least one public repository receiving commits in the observation year; (b) independent, venture-backed status, public-company and non-VC-backed open-source projects were excluded; (c) sector coverage, we balance the sample across 20 sectors from AI/ML through

Gaming, Robotics, and Legal Tech to avoid over-weighting the developer-tools category that dominates most GitHub panels.

The sector taxonomy is ours; full definitions and sector membership are available at https://signals.gitdealflow.com/methodology.

3.2 Collection pipeline

For each organization and each quarterly period, we:

1. Enumerate the organization's public repositories via the GitHub REST API `/orgs/{org}/repos` endpoint.
2. Select the most active repository by commits in the trailing 14-day window, defined as the window ending on the first day of the quarter.
3. Pull commit activity via `/repos/{owner}/{repo}/commits` and unique-contributor counts via `/repos/{owner}/{repo}/contributors`.
4. Record new-repo creation events from the organization-level `created_at` metadata.
5. Compute percent changes against the preceding 14-day window.
6. Apply a deterministic classifier (described in §3.3) to label the observation.

3.3 Signal classification

Each observation is classified into one of four acceleration patterns using deterministic heuristics over the collected variables:

- **Framework migration.** A concentrated burst of commits to dependency manifests and build-configuration files without a corresponding jump in unique-contributor count. Empirically the modal pattern (165 of 219 observations, 75%). Often precedes a product rewrite.
- **Engineering hiring burst.** Unique-contributor count rises faster than commit volume; new accounts appear in merge-commit authorship. (20 of 219, 9%.) Consistent with pre-announcement team expansion.
- **Infrastructure buildout.** New public repos spawning under the organization; disproportionate activity in infrastructure repos relative to the primary product repo. (8 of 219, 4%.) Often precedes a platform pivot or enterprise-tier launch.
- **Deploy frequency spike.** Commit velocity rises sharply without proportional contributor growth. (26 of 219, 12%.) Consistent with a small team sprinting toward a milestone.

The classifier is deterministic and documented at https://signals.gitdealflow.com/methodology. It is intentionally simple and interpretable; future releases may evaluate ML-based classifications against this baseline.

4. Dataset description

4.1 Structure

The dataset is distributed as three CSV files with a Frictionless Data schema:

- `startup_signals.csv`, primary observation table, 219 rows. Primary key `(period, startup_name)`. Columns: period, sector_slug, sector_name, startup_name, stage, geography, commit_velocity_14d, commit_velocity_change_pct, contributors, contributor_growth_pct, new_repos, signal_type, github_url.
- `sector_aggregates.csv`, sector-level aggregates per quarter, 72 rows. Primary key `(period, sector_slug)`.
- `signal_type_timeseries.csv`, distribution of signal types per period, 15 rows.

4.2 Summary statistics

Velocity distribution. Across 219 observations, 14-day commit velocity has a median of 71, a mean of 173, and a 90th percentile of 392. The distribution is right-skewed with a heavy upper tail driven by a handful of high-throughput infrastructure projects.

Velocity change. Quarter-over-quarter commit-velocity change ranges from –94% to +1,647% with a median near zero (–1%). 49% of observations show positive velocity growth. Extreme positive outliers concentrate in two sectors, Gaming and Space Tech, which also contain the two highest-velocity-change observations in the most recent period (castle-engine at +344%, orbiternassp at +329%).

Signal type distribution. Framework migration dominates (75%); deploy frequency spike (12%), engineering hiring burst (9%), and infrastructure buildout (4%) together account for the remainder. Signal mix is stable across periods, with the share of framework-migration classifications varying by less than five percentage points period to period.

Geography. Where geography is identified (108 of 219 observations, 49%), US dominates with 60 observations, followed by EU (24), LATAM (12), APAC (8), and Canada (4). The remaining 111 observations are classified Unknown; this reflects the practical difficulty of attributing geography from GitHub metadata alone and is a known limitation.

Sectors. 20 distinct sectors are represented. Sample size per sector ranges from 1 (Legal Tech, HR Tech's smaller bucket) to 8 (Data Infrastructure, Cybersecurity), reflecting real-world heterogeneity in the density of venture-backed, open-source-first startups.

4.3 Examples of heterogeneity worth studying

The dataset supports several cross-sectional questions out of the box:

1. **Do hiring-burst signals lead or lag framework-migration signals?** Under a fundraise-preceding-hypothesis, hiring bursts would concentrate in the 14-day windows immediately preceding announcement-heavy periods; framework migrations would be more uniformly distributed.
2. **Is velocity change sector-mean-reverting?** The panel structure permits fixed-effect regressions of velocity change on its own lag, controlling for sector.
3. **Geography × signal-type interactions.** Preliminary inspection suggests US observations skew toward hiring-burst and deploy-frequency-spike classifications, while EU observations

skew toward framework-migration, but the sample is too small for strong claims.

We deliberately do not pre-report statistical tests on these questions; they belong in derivative work using this release as input.

5. Limitations

No linked funding-event labels. This release describes the left-hand-side signals only. Linking observations to subsequent fundraise announcements, IPOs, or acquisitions requires joining against a funding-event source (Crunchbase, PitchBook, Dealroom). We view such a join as the most valuable immediate extension of this dataset and invite replication.

Selection bias toward open-source-active orgs. The dataset over-represents sectors where open-source work is conventional (developer tools, data infrastructure, AI/ML) and under-represents sectors where it is not (most consumer apps, many fintechs). Cross-sector comparisons must be interpreted with this in mind.

14-day observation windows. Shorter windows would capture sprint dynamics but introduce noise from weekends and holidays; longer windows would smooth out short-term acceleration. The 14-day choice is a pragmatic tradeoff with precedent in empirical software engineering literature, but it is not optimized against any downstream objective.

Survivorship. Our seed list is derived from currently-active GitHub organizations; orgs that have shut down or been acquired and dismantled during the observation window are not retroactively added.

Organization-to-startup mapping. Some GitHub organizations host multiple products; we score the most active public repo, which can misrepresent activity on secondary product lines.

6. Future work and how to contribute

We plan three extensions in upcoming releases:

1. **Funding-event joins.** Crunchbase- or Dealroom-derived funding timestamps per startup, enabling event-study analyses.
2. **Daily-granularity snapshots.** Supplementary daily snapshots alongside the current quarterly panel, for researchers who need higher-frequency structure.
3. **Extended sector coverage.** Addition of an additional 100 organizations across currently thin sectors.

Contributions are welcome: issues, methodology critiques, and pull requests to the sector taxonomy can be submitted through https://gitdealflow.com. Replication studies are particularly welcome, and we are happy to co-author with researchers who join the panel against external funding-event data.

7. Data availability

The dataset is distributed under CC BY 4.0:

- Canonical live endpoint (CSV): https://signals.gitdealflow.com/api/signals.csv

- Canonical live endpoint (JSON): https://signals.gitdealflow.com/api/signals.json
- Static snapshot mirror on Kaggle (DOI pending): [URL pending upload]
- Static snapshot mirror on Data.world: [URL pending upload]
- Static snapshot mirror on Zenodo (DOI assigned on upload): [DOI pending]

Machine-readable citation metadata is available in the bundle's `CITATION.cff` file.

8. Citation

> VC Deal Flow Signal. (2026). *A Longitudinal Panel of GitHub Engineering Velocity for Venture-Backed Startups: Dataset and Early Observations* (v1.0.1). https://gitdealflow.com

References

Gompers, P. A., Gornall, W., Kaplan, S. N., and Strebulaev, I. A. (2020). How do venture capitalists make decisions? *Journal of Financial Economics*, 135(1), 169-190.

Kaplan, S. N., and Strömberg, P. (2004). Characteristics, Contracts, and Actions: Evidence from Venture Capitalist Analyses. *Journal of Finance*, 59(5), 2177-2210.

Mockus, A., Fielding, R. T., and Herbsleb, J. D. (2002). Two case studies of open source software development: Apache and Mozilla. *ACM Transactions on Software Engineering and Methodology*, 11(3), 309-346.

Vasilescu, B., Filkov, V., and Serebrenik, A. (2013). StackOverflow and GitHub: Associations between software development and crowdsourced knowledge. In *2013 International Conference on Social Computing*, 188-195.

9. Declarations

Artificial intelligence use disclosure. In accordance with the SocArXiv AI policy (https://socopen.org/ai-policy/), the author discloses the following uses of AI-assisted tools in the preparation of this work:

- **Machine-assisted data analysis.** Large language models were used as coding and analysis assistants for parts of the collection and classification pipeline that produced the dataset. The pipeline is deterministic, version-controlled, and public (https://github.com/kindrat86/gitdealflow-signal-classifier), and its outputs were verified by the author against the live panel at https://signals.gitdealflow.com/api/signals.csv.
- **Copy-editing and formatting.** AI-assisted tools were used to copy edit and format the manuscript.
- **Pre-writing assistance.** AI-assisted tools were used for literature searches and organization of ideas. This manuscript reports on original research conducted by the author.

The author attests to thorough human supervision of all AI-assisted work. Every cited source was verified by the author to exist and is accurately characterized. The dataset, statistics, and results reported here were produced from the author's own data collection via the public GitHub API; they

are not simulated or fabricated. All AI-assisted output was thoroughly reviewed, edited, and confirmed by the author before inclusion, and the author accepts full responsibility for the content of this manuscript. No AI tool or language model is an author or co-author of this work.

Funding. This research received no external funding.

Competing interests. The author has a commercial interest in the VC Deal Flow Signal product, which is built on the data described here. No other competing interests are declared.

Correspondence. signal@gitdealflow.com