Database, Methodological Tools, and Research Opportunities: Creative Destruction Lab and Early-Stage Technology Ventures

Preliminary & Incomplete
Please Do Not Distribute

Amir Sariri Evgenia Gatov Kyle Robinson

Sonia Sennik Michael Vertolli Avi Goldfarb*

December 6, 2024

*Sariri is with Purdue University. Gatov, Robinson, Sennik, and Vertolli are with Creative Destruction Lab. Glodfarb (PI) is with Rotman School of Management & NBER. We thank Ajay Agrawal, Kevin Bryan, Joshua Gans. Financial support from the Government of Canada's Strategic Innovation Fund is acknowledged. We thank Creative Destruction Lab staff for supporting the creation of this database.

1 Introduction

The early stages of startup formation remain one of the least understood aspects of firm development and growth. Decisions made during these formative periods often have irreversible consequences, shaping the trajectory and potential success of new ventures (Howell, 2017; Eisenhardt and Schoonhoven, 1990). However, these formative periods are also inherently the most challenging to observe in a systematic way that allows rigorous empirical analysis (Guzman et al., 2019; Howell, 2020). Startup programs offer a distinctive window into this process, providing detailed and structured records of early firm development (Hallen et al., 2020).

In this paper, we introduce a new dataset built from one such program called Creative Destruction Lab (CDL), a global non-profit mentoring program for early-stage high-technology startups. We believe that the nature of this program and the data we collected from its operations are particularly suited for investigating open questions in the economics and management of advice, entrepreneurial strategy, entrepreneurial finance, and the complex process of technology transfer.

Since its inception in 2012 through the 2023/2024 academic year, 3,565 startups participated in CDL's 9-month program that, as of September 2024, have collectively raised USD$9.3 billion and are valued at approximately USD$30 billion. The program is designed as a series of meetings over one full day every eight weeks in which mentors provide feedback to founders on prioritizing measurable business objectives. The mentors then select specific startups to support in achieving those objectives.

The database we construct from this program combines three critical features for conducting rigorous empirical research. First, the sample size is large and spans several years, containing approximately 15,000 applicants, 6,600 founders among admitted startups, and 1,900 mentors. Second, academic research opportunities for this data were recognized from the beginning of the program, and so the organization has kept accurate records and built enterprise IT infrastructure that links over 200 datasets, including a dataset containing 170 codified characteristics at the startup-level, a dataset of structured longitudinal data on the evolution of firm development and operations, and a dataset of unstructured verbatim transcripts of the moderated discussions between mentors and founders during mentoring sessions. Third, data cover 28 technological domains ranging from therapeutics to quantum computing,

addressing 22 sectors at the two-digit NAICS level from 67 countries.

In developing this database and making it widely available, our objective is to create a resource that enables deeper insights into the innovation process by advancing our knowledge about how startups navigate the crucial early stages of building a business. We believe that the scholarship opportunities created will expand the foundations of research in entrepreneurship, technology commercialization, and organizational economics. In this paper, we provide an overview of the setting from which data is collected, a high-level description of available data, and information on how to access these data for research.

2 The CDL Program: An Overview

Founded at the Rotman School of Management at the University of Toronto in 2012, Creative Destruction Lab includes sites in 15 universities located in Canada, the United States, Europe, and Australia across 28 specialized technology streams including health and life sciences, artificial intelligence, digital technology, financial technology, quantum computing, and energy. The stated mission of the program is "to enhance the commercialization of science for the betterment of humankind" by addressing what it terms as a "failure in the market for judgment." The design philosophy of the program rests on the hypothesis that the uneven distribution of business judgment on how to build a business explains much of the disparity in entrepreneurial success rates across regional economies–regions that are otherwise quite competitive in the production of ideas.

On the demand side of the market for judgment, inexperienced entrepreneurs need knowledge and support in how to build a business. On the supply side, mentors such as venture capitalists and angel investors with significant industry experience can provide this knowledge because, in part, doing so enhances the pipeline of investable companies (Miller and Bound, 2011). A market failure occurs to the extent that there is not a developed market for the exchange of advice, other than as a value-added service of attracting external investors (Gans, 2018).

2.1 Program Structure

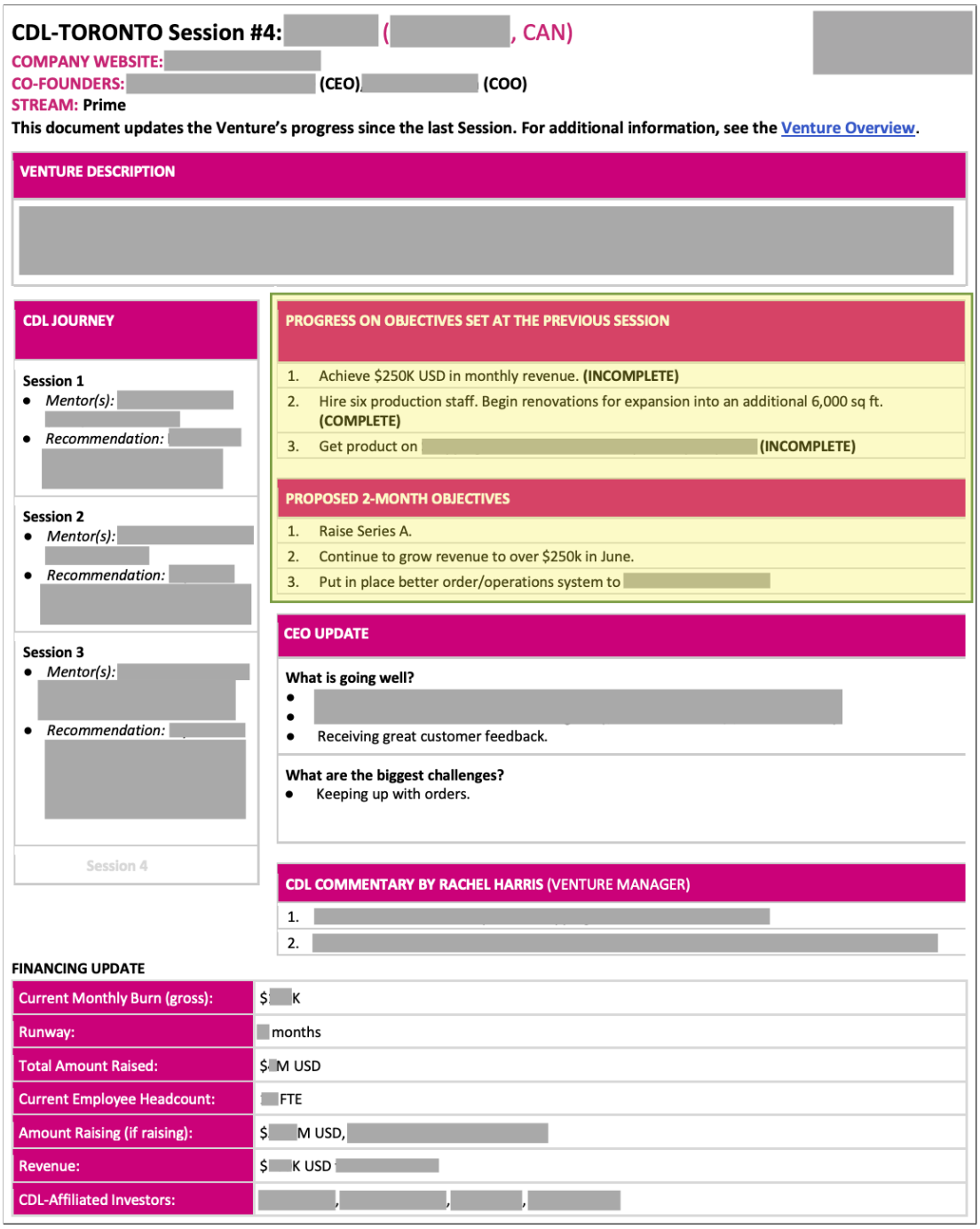
Admission into the program is competitive and open to startups from around the world. Candidates are subjected to an evaluation process that includes submitting a detailed application and participating in business and technical assessment interviews. Finalists are offered admission to a specific geographic site and technology stream (e.g. Toronto-Artificial Intelligence or Berlin-Health), a program track where mentors with relevant domain expertise are assembled.

The core of the program involves four full-day sessions approximately every eight weeks to help founders prioritize three measurable business objectives. During each session, mentors focus on helping entrepreneurs identify business objectives to prioritize. They also assess ventures' progress. Approximately one week before each session, mentors receive brief dossiers like the one in Figure 1 on every startup in their track, outlining the founders' proposed objectives for the upcoming eight weeks, progress on previously-set objectives, relevant updates and challenges, updated financial metrics, and other details ranging from target customer and core technology to founders' educational background. Mentors are asked to familiarize themselves with each venture's progress and formulate their feedback on the proposed objectives in preparation for the session day.

On the morning of session days, founders meet privately with 4-6 mentors from their cohort in small group meetings (SGMs) to receive private feedback on their proposed objectives. Two of the SGM mentors are responsible for suggesting revisions to the proposed objectives, guided by the venture dossiers and additional details discussed in the meeting. In the afternoon, each track's mentors and founders convene in a separate classroom like the one in Figure 2 for the large room discussions (LRD). During LRDs, the SGM mentors assigned to critique founders' proposed objectives describe their views, then a faculty member moderates a broader discussion with the founders and remaining other mentors to reconcile different advice and finalize the startups' objectives for the next eight weeks.

Sessions conclude in the early evening with deliberations, when founders leave the room, and the moderator asks mentors to "raise their hands," one startup at a time, to commit four hours of their personal time to support founders in achieving their finalized objectives. These are high-stakes decisions for ventures, as those

Figure 1: Example of Startup Dossier



CDL-TORONTO Session #4: [redacted] ([redacted], CAN)

COMPANY WEBSITE: [redacted]

CO-FOUNDERS: [redacted] (CEO), [redacted] (COO)

STREAM: Prime

This document updates the Venture's progress since the last Session. For additional information, see the Venture Overview.

VENTURE DESCRIPTION

[redacted]

CDL JOURNEY

Session 1
- *Mentor(s):* [redacted]
- *Recommendation:* [redacted]

Session 2
- *Mentor(s):* [redacted]
- *Recommendation:* [redacted]

Session 3
- *Mentor(s):* [redacted]
- *Recommendation:* [redacted]

Session 4

PROGRESS ON OBJECTIVES SET AT THE PREVIOUS SESSION

1. Achieve $250K USD in monthly revenue. **(INCOMPLETE)**
2. Hire six production staff. Begin renovations for expansion into an additional 6,000 sq ft. **(COMPLETE)**
3. Get product on [redacted] **(INCOMPLETE)**

PROPOSED 2-MONTH OBJECTIVES

1. Raise Series A.
2. Continue to grow revenue to over $250k in June.
3. Put in place better order/operations system to [redacted]

CEO UPDATE

What is going well?
- [redacted]
- Receiving great customer feedback.

What are the biggest challenges?
- Keeping up with orders.

CDL COMMENTARY BY RACHEL HARRIS (VENTURE MANAGER)

1. [redacted]
2. [redacted]

FINANCING UPDATE

Current Monthly Burn (gross):	$[redacted]K
Runway:	[redacted] months
Total Amount Raised:	$[redacted]M USD
Current Employee Headcount:	[redacted] FTE
Amount Raising (if raising):	$[redacted]M USD, [redacted]
Revenue:	$[redacted]K USD [redacted]
CDL-Affiliated Investors:	[redacted], [redacted], [redacted], [redacted]

Notes: This figure shows a sample startup dossier distributed to mentors before sessions. It includes updated objectives, a status update from the CEO, commentary by the CDL manager responsible for the startup, the latest financial information, and a history of main mentor recommendations from prior sessions. Portions that may reveal the identity of the startup or CDL are redacted.

Figure 2: Finalizing Objectives via a Moderated Debate

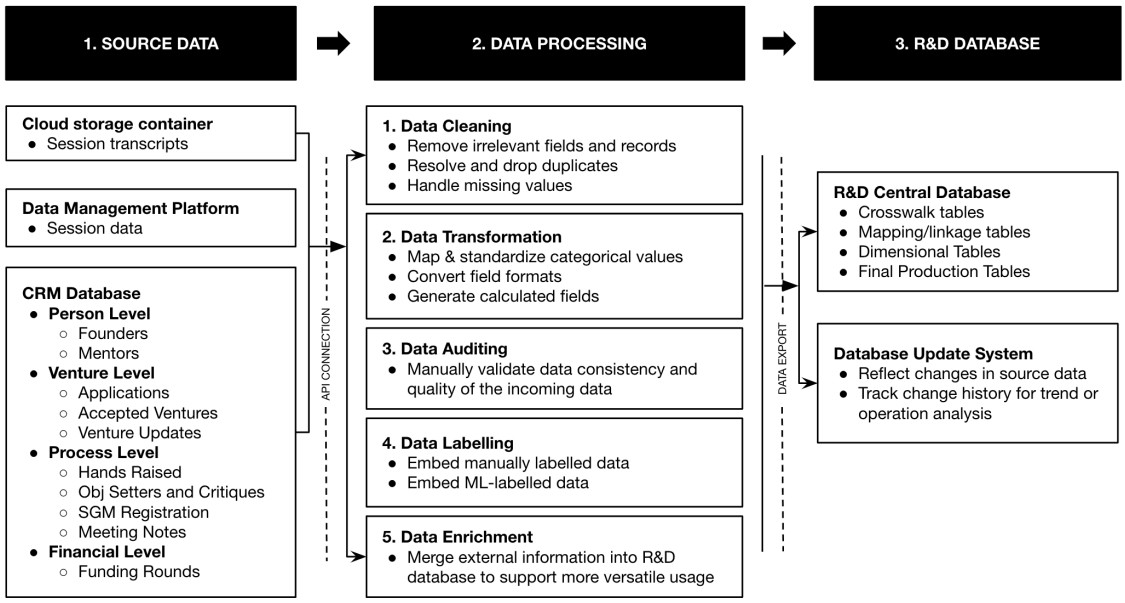
Notes: This image shows an in-progress discussion in the large room. Founders and mentors engage in a debate moderated by a business school professor (hidden behind the founder) to arrive at a finalized set of objectives for the next eight weeks.

without formal time commitment from a mentor are dropped from subsequent sessions. For startups that receive formal support, designated staff are responsible for connecting founders and mentors shortly after the session and facilitating the setting up of off-cycle meetings. They also liaise with founders throughout the eight-week cycle to document progress on objectives and track mentors' honoring of their time commitment. The cycle ends with founders submitting a draft dossier for the next session.

3 Data Collection

Data is collected throughout the year by specialized staff from the Research & Development team who are trained in gathering and storing research-grade data, supported by liaison staff at the sites and supervised by the principal investigator. The participating founders and mentors consent to the collection and use of information from program operations for facilitating program participation, internal research and analysis, and academic research (see Privacy Policy). In particular, all participants acknowledge this policy as a part of their annual onboarding process that outlines the type of data that the program will store and the potential use of de-identified data for research purposes. An opt-out choice is available for participants that do not wish their data to be collected, or elect to delete their information retroactively. By the end of 2024, eight startups requested their data be deleted.

Figure 3: Data Collection Pipeline



Notes: This figure shows an overview of the main steps of the data collection procedure.

Figure 3 provides an overview of the data collection procedure. Data is initially ingested directly from the enterprise CRM infrastructure used by various parts of the organization to run the program and interact with stakeholders. Site staff use this infrastructure to, for instance, create new dossiers and record new financing outcomes for past alumni. Next, a downstream data processing pipeline validates the initial intake based on various quality and reliability tests. During this stage, we also perform additional automatic, semi-automatic, and manual codifications as required. This pipeline entails filtered data ingestion from multiple CRM objects, additive and destructive updates to crosswalk tables for ventures and individuals through deterministic and probabilistic linkage, followed by identifying and resolving duplicates and assigning unique person and venture identifiers.

As the program has evolved over the years, some procedures have changed, most notably the questions asked on applications. To maximize continuity, we reconcile conflicting variable types across cohorts. Standardized and calculated variables are then added to enhance the dataset. We also run data quality audits at least once a

year. A benefit of these audits is the introduction of new solutions to standardize data collection at the source and minimize the need for manual corrections. The database is securely stored as a relational database on an enterprise cloud storage service. Currently, there are approximately 200 data tables and 20 mapping tables. The primary unique identifiers linking the tables are those of ventures, individuals (founders and mentors), and sites, streams, program years, and sessions.

3.1 Data Elements

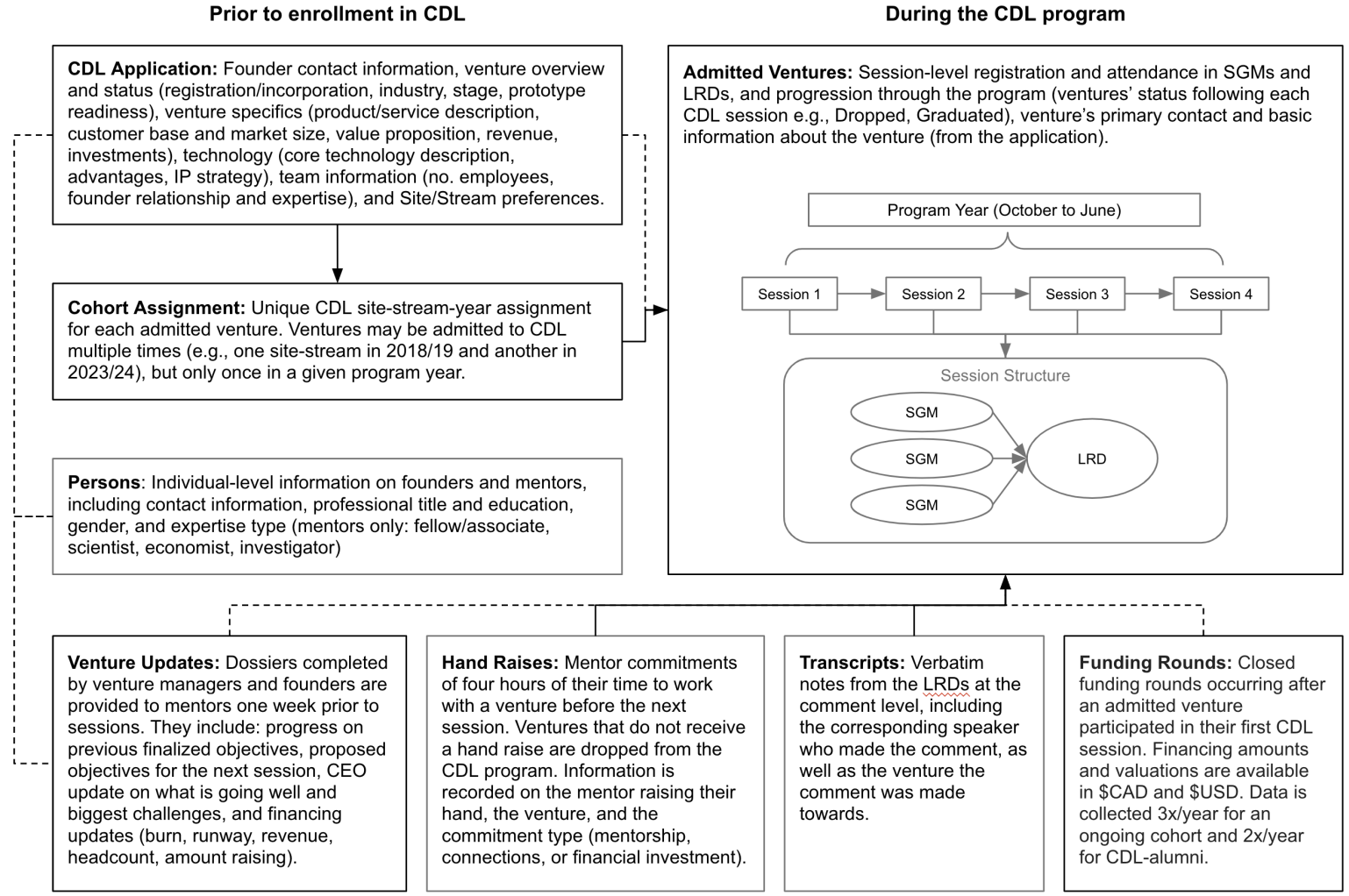
Figure 4 provides an overview of the types of information tracked. Data spans all aspects of the program from applications to market performance years after venture participation. From applications alone, we codify 40 variables' venture-level characteristics, including intellectual property strategy and financial metrics such as grants and existing funding. For admitted startups, the majority of these covariates can be observed in panels due to updates made to the dossiers. Dossiers also reveal new panel data, notably the proposed objectives as well as progress on the finalized objectives from the previous session. At the individual level, we codify demographic, educational, and professional information. Among other session-level information, we record verbatim transcripts of the discussions between mentors and startups. We believe these data are particularly useful in identifying the patterns of product market strategic choices made in the early stages of firm formation, especially in light of new AI tools that have streamlined unstructured data analyses.

3.2 Ventures

There are three primary sources of venture data: self-reported pre-program information from venture applications, in-program information from venture participation, and market-based post-program outcomes such as financing.

Pre-Program Applications: Applications contain a great deal of information about the companies. With the program evolution over the years, the application fields have also evolved. We reconcile these cases in our downstream pipeline to the extent possible to maximize data continuity. These are some of the main structured

Figure 4: The Main Data Elements Collected



Notes: provides an overview of the main types of information tracked.

and unstructured information captured from applications:

- Founder information: A range of numerical and categorical indicators for demographic, educational, and professional background, including gender, educational degrees (including level, specialization, and institution), work experience, and level of involvement in the current venture (number of hours worked, part-time vs. full-time).

- Venture overview: These fields include unstructured responses to a number of questions, including: "Briefly describe your venture", "Please describe the product or service you are building/have built, and how this product or service works/will work from the perspective of a potential customer", "Who is your first (or potential first) customer, and what is the value proposition of your technology/product to this customer?", and "What is the long-term vision of the venture?".

- Incorporation: An indicator equal to 1 if the venture is incorporated, in which case additional fields record the date and jurisdiction of incorporation.

- Prototype readiness: An indicator equal to 1 for a positive response to the question "Do you have a working prototype that can be demonstrated? A prototype can be any version of your technology that has generated data to demonstrate the potential to add value to a customer."

- Market: An indicator equal to 1 if having any paid customers or unpaid pilots, in which case additional unstructured response exists for describing the customers and current relationships with them. Two more textual fields ask "Share any estimations of your target market size." and "Who currently sells to your target customers? What makes your product better than the competition?".

- Revenue: An indicator equal to 1 if the venture has generated revenue, in which case revenue amount, currency, and type (e.g., recurring revenues, paid pilot single, etc.) are also reported.

- Funding: As with revenues, with additional data fields for non-dilutive (e.g., government grants) and other sources of capital, plus details on currency and

the amount of these funding amounts.

- Core Technology: Founder description of the core technology in non-technical terms with additional questions such as "the unique element of your core innovation that enables value creation for your customers. E.g., if your value prop is speed, what element of your core technology enables or unlocks speed?" An indicator equals 1 if the venture's core technology was developed at, or with the support of, a research institute.

- IP strategy: A categorical variable from a multi-select field asking "Which of the following describes the current IP strategy?" with options "The venture has filed patents", "The venture is keeping its algorithms/core technology as trade secret", "A freedom to operate search has been completed and no conflicting IP found", "The founders have disclosed the invention to their affiliated academic institution", "The founders are in the process of having the IP assigned to the venture", "The venture has access to proprietary data", "The venture has not yet decided on an appropriate IP strategy".

- Team Information: Founding team-level data corresponding to application questions such as "How long have the founders known one another?", "How did the founders meet and why are you working on this project?", "Explain why you are the right person, or team, to tackle the problem(s) you have set out to solve. Highlight any extraordinary challenges or learnings that may not be captured by academic or career achievements." Self-reported number of full-time employees is also collected.

- Affiliations and advisors: Indicators equal to 1 if the venture has any affiliations with other startup programs, incubators or accelerators, either currently or in the past, and an indicator for whether the venture has existing advisors.

- Site and Stream preferences: Ventures' preference rankings for different sites and streams. This information is used later for matching admitted ventures to tracks.

- Venture stage: Categorical variable with options i) concept (idea generation, market exploration), ii) technology (developing core product infrastructure),

iii) prototype (initial product version creation), iv) validation (testing market fit and demand), v) early-revenue (generating initial sales or income), vi) and profitable (sustaining operations with positive cash flow).

- Geography: Variables for the city, province/state, and country.
- Expected value: Since 2022/23, a series of indicators codified from the response to the question "How can Creative Destruction Lab provide the most value to your venture? (Select all that apply)", with an indicator equal to 1 when each of the following is selected: "customer research, go-to-market planning", "financial, IP, sales, or regulatory planning", "product, technology roadmap", "product market fit validation", "technology/ regulatory validation", "market selection, sales processes", "sales and marketing", "organization building", "tech dev., approval, launch", "hire employees, advisors, co-founders", "license, patent IP, obtain data for machine learning", and "raise money". The business function categories are derived from the classification system developed by Sariri (2024).

In-Program Activity: The following is a subset of data elements captured from the participation of the admitted ventures in the program, corresponding to information gathered before, during and after each session:

- Year and track assignments: Categorical variables for the program-year, site, and stream of venture participation.
- Session attendance: Numerical variables for the number and date of the sessions attended, and an indicator for whether the session was in-person or virtual (virtual started in 2020).
- Schedules: Codified SGM schedules indicating which mentors met which founders at each session. Additional variables indicate mentor and startup attendance at each session.
- Mentoring decisions: Variable containing the list of mentor identifiers who raised their hand for mentoring a startup at a given session.

- Objectives: String and categorical variables pertaining to the raw text and the business function type of the proposed and finalized objectives at each session. The business function categories are derived from the classification system developed by Sariri (2024). For finalized objectives from the previous session, there is a categorical variable indicating whether the objective was complete, partially complete, or incomplete.

- Startup trajectory: String variable containing the CEO report on what is going well, biggest challenges.

- Financial updates: Numerical and categorical variables for current monthly cash burn rate, revenues and revenue type (e.g., annual-recurring, monthly-recurring, year-to-date), total amount raised to date, amount currently being raised, cash runway (time before running out of cash, and the current number of full-time and part-time employees.

Post-program Financing Data: Designated staff survey commercial private equity databases, founders, and investors to record the latest financing events of alumni companies using a specialized tracker in the organizational CRM. Direct sourcing of information from founders allows for recording highly detailed financing terms. The timing of the data collection is based on a predetermined schedule requiring two surveys a year for the alumni and three surveys a year for current startups. Only closed funding rounds are included in reporting. The organization and sites' leadership are responsible for the accuracy of this information, so utmost care is taken in tracking detailed information. This is because alumni funding events constitute a key success metric of the CDL program. These data include:

- Date round closed: The date the funding round was closed. Financing events prior to the venture's attendance are excluded.

- Financing stage: Categorical variable with values pre-seed and seed, Series A, B, and so on, and Exit (IPO, Merger, Acquisition, or Private Offering).

- Instrument used: Categorical variable with values Equity (shares in the company are exchanged for investment immediately), Debt (a loan that must be

repaid), Convertible Note (allows investors to convert the money "loaned" to the company into shares with certain rights and restrictions), SAFE (allows the investor the right to receive equity of the company on certain triggering events), and EXIT (selling of the entire company to another investor via stock swap, cash, or both).

- Terms: Various variables for the financing terms, including equity ownership, voting rights, liquidation preferences, and other key provisions. The availability of these fields depends on the instrument used. Missing values exist if founders or investors were not willing to share this information.

- Amount Raised: Numerical variable for the amount raised, a categorical variable for the currency, and a numerical variable for the conversion rate to Canadian dollars as of the closing date.

- Pre-Money Valuation (PMV): Numerical variable for the pre-money valuation at the given financing round. When founder ownership is unknown, valuation is imputed as four times the financing amount and the imputation is flagged as a separate variable.

3.3 Mentors

These data include substantive information on individual mentor background, their participation records in the program, history of mentoring commitments made to startups, as well as other structured and unstructured data pertaining to the specific advice conferred to startups at a given session. The primary source for this information is administrative records from program operations, including

- Mentor type: Categorical variable with values identifying business mentors (exited entrepreneurs, angel investors, and venture capitalists), and non-business mentors who provide technical expertise (typically PhD scientists and faculty).

- Session attendance: Indicators for actual mentor attendance at a session. The main source is administrative registration (RSVP) records, though last minute cancellation and RSVP's exist. These cases are rectified using additional day-of scheduling and hands-raised information.

- Objective setter and critiquer: Indicators at the mentor-startup-session level for whether the SGM mentor who met with the startup was designated as the Objective Setter (responsible for posting the first-round revisions to proposed objectives) or the Critiquer (second-round revisions to proposed objectives).

- Mentor advice: String variable corresponding to the verbatim text of mentor advice as spoken during LRD meetings. These strings are parsed at the comment level and are defined at the mentor-startup-session level.

4 Descriptive Statistics

From the initial 2012/13 cohort through 2023/24, CDL has operated in 15 sites and 28 streams, with over 750 sessions completed. Through 2023/24, CDL data includes approximately 14,473 applicant ventures, and 6,523 founders of 3,565 admitted ventures. On the mentor side, approximately 1,873 individuals have participated in the program. In the rest of this section, we provide select statistics to describe the overall characteristics of the venture and mentoring data.

Table 1 describes the startups in sample by showing select characteristics. Overall, startups are early-stage, with roughly 40% claiming to have generated revenue by the time they applied. Companies are also early in terms of funding stage. Prior to participation, 48% claimed to have raised external capital of greater than $500,000. A little over one-third of ventures were concept, prototype, or technology stage, with 25% in product-market validation stage. Startups are predominantly from the high-technology sector. More than half of the founding teams have at least one PhD founder and on average 1.6 PhD founders, conditional on having one.

The sample of startups is diverse, spanning 28 technological domains. At the NAICS supersector level, nearly half of the ventures are in the Professional and Business Services sector, followed by 31% of the startups in Goods Producing industries, including agriculture, mining, oil and gas extraction, construction, and utilities. 17% of the startups are in the artificial intelligence stream, the largest and oldest specialized stream. The more recent streams include Cancer and Minerals, launched in 2023/24, with 3% of the startups admitted in each, respectively. In terms of geographic diversity, 20% are from Europe, and roughly 70% are from Canada and the

Table 1: Select Characteristics of Admitted Startups
Program Years 2012/13 to 2023/24

$N = 3,565$	
Country	
Canada	1488 (41.7%)
USA	1120 (31.4%)
Europe	727 (20.4%)
Asia	105 (2.9%)
Vertical Groups	
Sustainability & Environment	978 (27.4%)
Digital Innovation	761 (21.3%)
Healthcare & Life Sciences	682 (19.1%)
Industry & Manufacturing	427 (12.0%)
Finance, Commerce, & Insurance	205 (5.8%)
2-digit NAICS	
Professional and Business Services	1622 (45.5%)
Goods Producing Industries	1106 (31.0%)
Other Services and Government	209 (5.9%)
Stage	
Concept Stage	30 (0.8%)
Prototype Stage	881 (24.7%)
Technology Stage	443 (12.4%)
Validation Stage	996 (27.9%)
Early-Revenue Stage	909 (25.5%)
Profitable	108 (3.0%)
Founders	
Num. Founders (SD)	2.35 (1.16)
Has Ph.D.	1815 (50.9%)
Mentoring	
Sessions Attended (SD)	2.87 (1.26)
Mean Hours Mentored (SD)	22.2 (19.32)

Notes: This table shows select characteristics of the startups admitted to CDL in program years from 2012/13 to 2023/24.

United States. The high fraction of North American startups is expected due to the high concentration of technology startups seeking venture funding in this region.

For most ventures, CDL is not their first mentoring program, with 69% reporting having participated in other startup programs. Furthermore, firms report a variety of different reasons for applying to the program, including to raise money (77%), receive help with sales and marketing (43%), and support with search and planning activities such as customer research, go-to-market strategies, financial, IP, sales, or regulatory planning (40%). In terms of program participation, admitted ventures attended 2.9 sessions on average. On average, startups receive formal support from mentors 5.5 times, which amounts to approximately 22 hours of mentors' personal time spent with founders between sessions.

Table 2 reports some information on post-program funding activity. One in five startups that participated in the program with 1.5% experiencing an exit via an IPO, merger, or acquisition. Among ventures that raised capital after the program, the average amount raised was %9.6 million with the average time to the first raise equal to 11 months after joining the program. The most common form of instrument was equity financing and the most frequent funding stage reported was at the seed stage.

5 Access to the CDL Data

To access the de-identified CDL data, it is necessary to submit an application. The application involves four main steps: (1) a research proposal submission, (2) ethics approval from the home institution of the Principal Investigator, (3) a Data Transfer Agreement (between the Principal Investigator and CDL), and (4) Non-Disclosure Agreements for all members of the research team accessing the data. The research proposal submission is accessible via this link. Additional information about the process can be obtained by clicking this link.

In each proposal, only data pertinent to the study objectives are shared. All identifiable information, including, but not limited to names of companies and individuals, email addresses, company billing addresses, social media accounts such as the LinkedIn URL, and phone numbers, is stripped prior to data transfer. Additionally, any open-text fields containing person or company names are masked using auto-

Table 2: Post-Program Funding Activity
Startups that Raised from Program Years 2012/13 to 2023/24

$N = 789$	
Financing Activity in $USD Million	
Raise Mean (SD)	9.6 (25.3)
Valuation Mean (SD)	36.2 (93.6)
Times Raised Mean (SD)	1.66 (1.00)
Months to Raise Mean (SD)	11.17 (10.60)
Highest Raise Event - Instrument	
Equity	544 (68.9%)
SAFE	98 (12.4%)
Convertible Note	107 (13.6%)
Debt	7 (0.9%)
Highest Raise Event - Stage	
Pre-seed	115 (14.6%)
Seed	391 (49.6%)
Bridge Round	27 (3.4%)
Series A	176 (22.3%)
Series B	44 (5.6%)

Notes: This table shows summary of the funding activity for 789 of the admitted startups that raised capital after joining the program.

mated scripts that replace these values with that of the venture's or person's unique identifiers. Once the de-identification process is complete, data tables are shared with researchers via a secure file transfer link.

6 Concluding Remarks

In this paper, we outline a novel dataset containing a broad range of characteristics pertaining to early-stage, science-based ventures, and the mentorship they receive at a global entrepreneurship accelerator. We believe that CDL data offers a new significant opportunity to the academic community to make contributions to the economics of innovation and entrepreneurial strategy.

To date, a number of academic research projects have leveraged these data, focusing on an array of entrepreneurship and innovation research areas. For example, Sariri (2024) examined the causal effect of mentoring on startup success, and uses detailed data on business objectives to characterize the nature and provision of entrepreneurial advice. Bryan et al. (2022) propose a novel mechanism that reduces information frictions for workers applying to startup jobs, thus offering a novel mechanism for high-quality startups to tackle the challenge of hiring workers early on. Parra and Winter (2022) leverage CDL data to develop a theory of early-stage venture financing, focusing on ventures' choice between issuing equity or a "SAFE", which gives investors the right to a number of shares to be determined by a future equity price. Davidsson et al. (2021) develop an instrument for venture idea assessment that decouples the assessment of venture ideas from the people who pursue these ideas (e.g., team members, investors, mentors, employees, and the general public). Most recently, Heydari Nejad (2024) explored the causal role of mentorship in improving entrepreneurial success. Using a dynamic structural model that incorporates rich mentorship interaction data from CDL, the study quantifies the value of mentorship in reducing uncertainty and improving startup financing performance, incorporating a thorough classification of mentorship advice by applying machine learning approaches to open-text data.

We believe that by making CDL data available to the larger research community, scholars can find and address diverse questions centered around early-stage

entrepreneurship, questions that were previously difficult to tackle due to data availability and measurement problems. Particularly given the crucial role of early strategic decisions on the future development path of companies, we hope that our efforts in expanding and distributing rich data on the early-stage of startup formation will make meaningful contributions to our understanding of how startups succeed.

References

Bryan, Kevin A, Mitchell Hoffman, and Amir Sariri, "Information Frictions and Employee Sorting Between Startups," *NBER Working Paper Series*, 2022.

Davidsson, Per, Denis A Grégoire, and Maike Lex, "Venture Idea Assessment (VIA): Development of a needed concept, measure, and research agenda," *Journal of Business Venturing*, 2021, *36* (5), 106130.

Eisenhardt, Kathleen M. and Claudia Bird Schoonhoven, "Organizational Growth: Linking Founding Team, Strategy, Environment, and Growth Among U.S. Semiconductor Ventures, 1978-1988," *Administrative science quarterly*, 1990, *35* (3), 504–529.

Gans, Joshua S, "A Better Way to Bring Science to Market," 2018.

Guzman, Jorge, Scott Stern, Antoinette Schoar, Javier Miranda, John Haltiwanger, and Erik Hurst, "Nowcasting and Placecasting Entrepreneurial Quality and Performance," in "Measuring Entrepreneurial Businesses," Vol. 75, Chicago: University of Chicago Press, 2019, pp. 63–110.

Hallen, Benjamin L, Susan L Cohen, and Christopher B Bingham, "Do Accelerators Work? If So, How?," *Organization science (Providence, R.I.)*, 2020, *31* (2), 378–414.

Howell, Sabrina T., "Financing Innovation: Evidence from R&D Grants," *The American economic review*, 2017, *107* (4), 1136–1164.

_ , "Reducing information frictions in venture capital: The role of new venture competitions," *Journal of financial economics*, 2020, *136* (3), 676–694.

Miller, Paul and Kirsten Bound, "The startup factories," *NESTA*, 2011.

Nejad, Mohaddeseh Heydari, "A structural model of mentorship in startup accelerators: Matching, learning, and value creation," *Univerity of Toronto Working Paper*, 2024.

Parra, Alvaro and Ralph A. Winter, "Early-stage venture financing," *Journal of corporate finance (Amsterdam, Netherlands)*, 2022, *77*, 102291–.

Sariri, Amir, "Economics of Advice: Evidence from Entrepreneurial Mentoring," *Purdue University Working Paper*, 2024.