# Datasheet for Primary Data

## Motivation for Dataset Creation

- Created to consolidate features in one single document related to venture financing of an array of startups.
- The data can also be used to assess the performance of the startup ecosystem across various regions considering the ‘status’ of each company as a target variable.

## Dataset Composition

- The dataset features essential company information encoded such as name, industry, relevant dates, funding amount, investment rounds, etc.
- Relations between instances are majorly explicit.
- Instances consist of feature data with no specific target label. For our project, we have extracted everything and parsed funding rounds, dates of fundings, and founding to specify a list of columns.

## Data Collection Process

- The dataset is likely to be sourced from publicly declared data for each company on Crunchbase and media outlets.
- The sampling done in this dataset would be deterministic to select only companies in the USA and in health and tech-related industries.
- The larger dataset would be all startups involved in venture funding globally, that have published information in Crunchbase prior to 2015.

## Data Preprocessing

- For our use case, we would use only USA startups that were funded on and after the year 2004 by converting arguments of dates to DateTime format.
- All instances with no founding year were dropped.
- We added new features related to time components such as the number of years until funding was received.
- The dataset already had undergone pre-processing and we didn’t get access/confirmation on the original location of the raw dataset.

## Dataset Distribution

- The dataset is distributed via a public repository on Github.
- The dataset is likely governed by CrunchBase Terms of Service and Licensing Policy.

## Dataset Maintenance

- The open-source data set was created by an individual contributor and hasn’t been updated since.
- Once the dataset becomes obsolete, it would lose engagement on the platform or the same is likely to be communicated in the comment section.

## Legal & Ethical Considerations

- The dataset does not relate to any people in particular and is made publicly available by the companies in question via Crunchbase and/or media outlets.
- Since no personal data is stored, GDPR does not apply.
- A notable exception would be one-person companies if any present in the dataset.
- An ethical constraint must be to verify the accuracy of data about companies on a third-party website. The most accurate information is always on the company's web page itself. Inflation or deflation of funding amounts for companies can be deceptive to analysts/investors.