



# MICRO VENTURES

---

PREDICTING SUCCESSFUL STARTUPS  
FOR MICRO INVESTORS

# CONTENTS

**01**

What is Venture Capital?

**02**

Our Project

**03**

Data Source Used

**04**

Data Collection and Cleaning

**05**

How We Use Machine Learning?

**06**

Project Execution Plan

# What Is Venture Capital ?

## Keep the tech world going

- Venture Capital is the money provided by investors to startups that have potential to reshape markets and grow very fast.
- The money deployed by VC Firm usually comes from institutional investor, corporation or wealthy individuals.
- The cash invested is not liquid and is invested at high risk with intention of getting high rewards in the future.

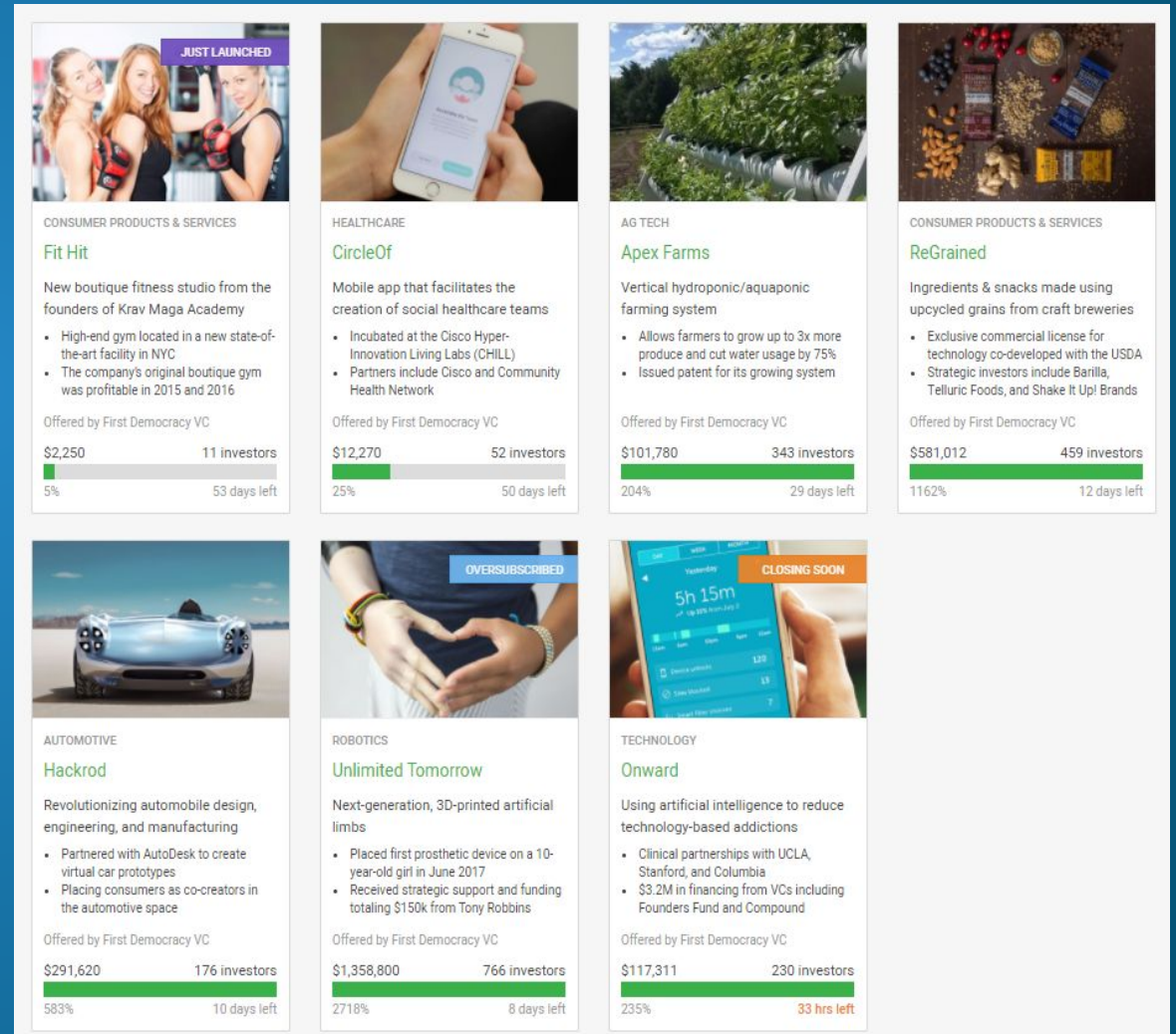


# What Is Micro Venture - Crowdfunding ?

## Crowdfunding in the contemporary world

- Crowdfunding enables large group of investors to invest small amount of investments in startups in exchange of equity.
- A larger pool of investors decreasing risk, generate large amounts of funding for startups.

**What if we can predict the success of a startup with machine learning?**



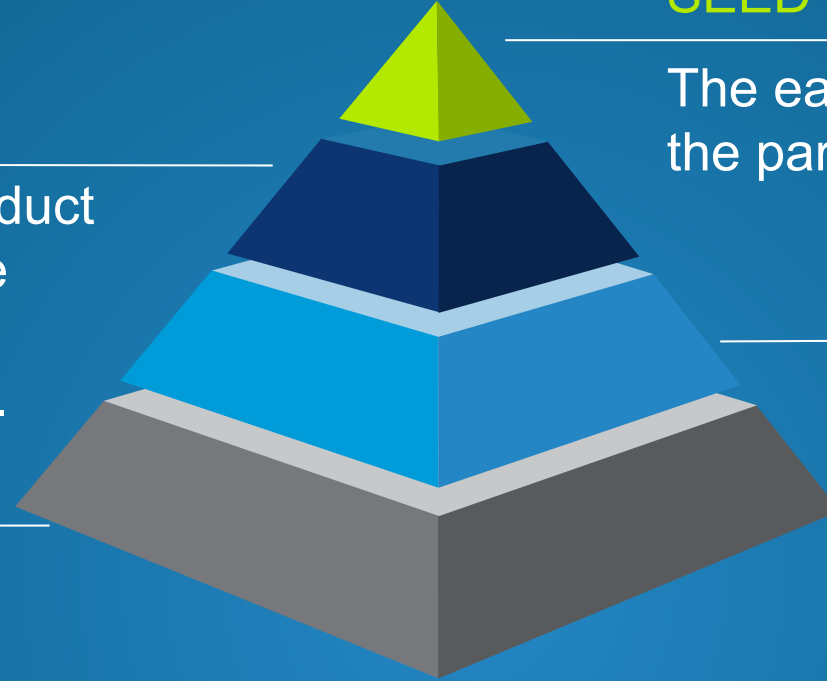
# Startup Journey And Funding Stages

## SERIES A

Company has established product and market fit, started to make some serious buzz and its customer base is growing fast.

## SERIES C

company has grown up and is likely operating on a global scale. Ready for IPO or acquisition.



## SEED FUNDING

The earliest stage of funding to get the party going

## SERIES B

Company has started to make considerable revenues in selected markets and is looking to expand operations.

---

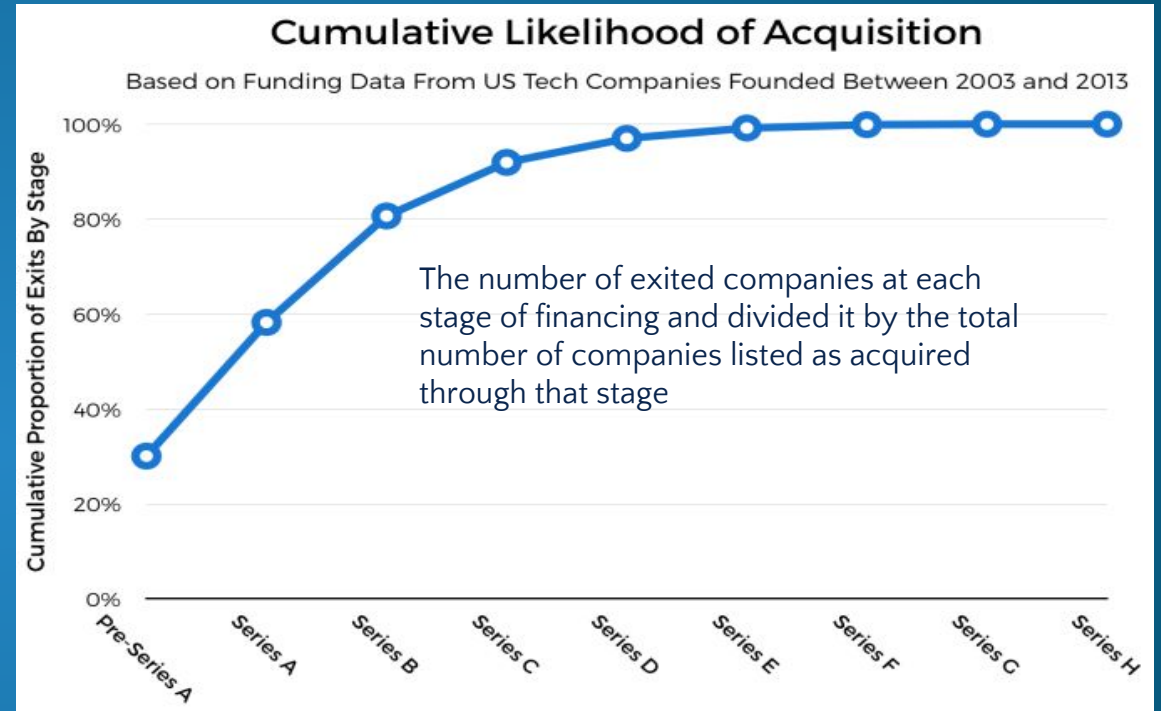
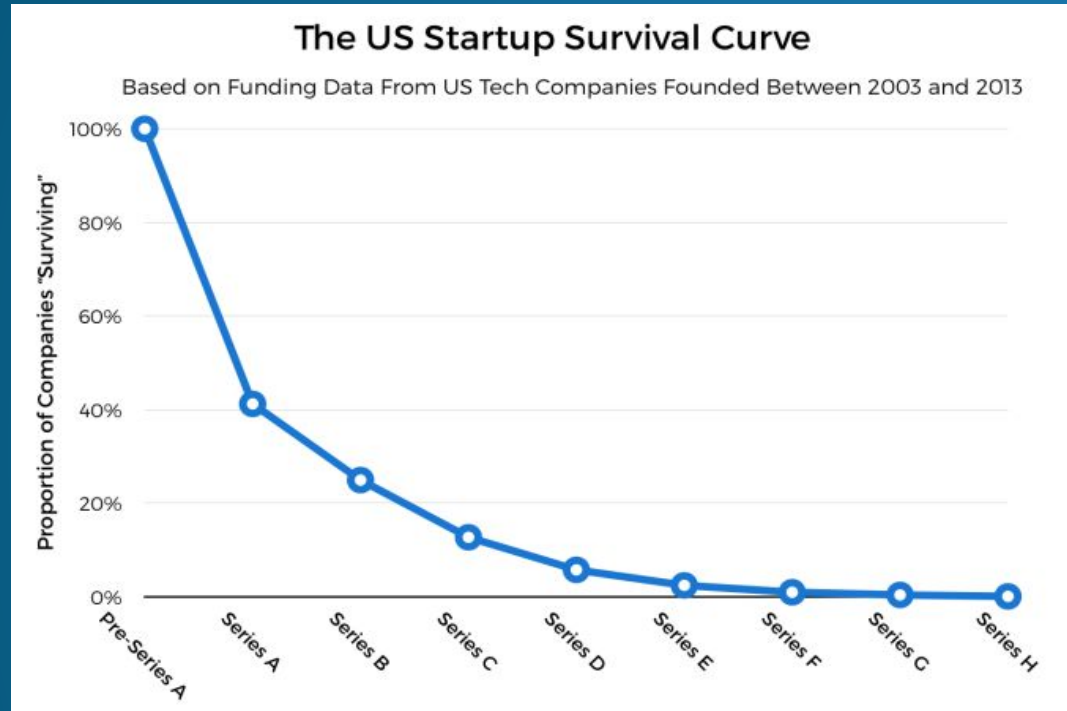
A startup exits this funding stage only when one of the following occurs

- 1) Startup fails
- 2) Company gets acquired
- 3) Company gets IPO

We are interested in companies that have the potential to **either get IPO or get acquired.**



## Funding data from around 15,600 U.S.-based technology companies founded (2003 – 2013)



- Around **60 percent** of companies that raise **Pre-Series A** funding fail to make it to **Series A** or beyond
- It shows that, of the companies in our data set that were acquired and have raised venture financing, **around 92 percent** of those raised through **Series C**.

# What Will Our **Machine Learning** Model Do ?

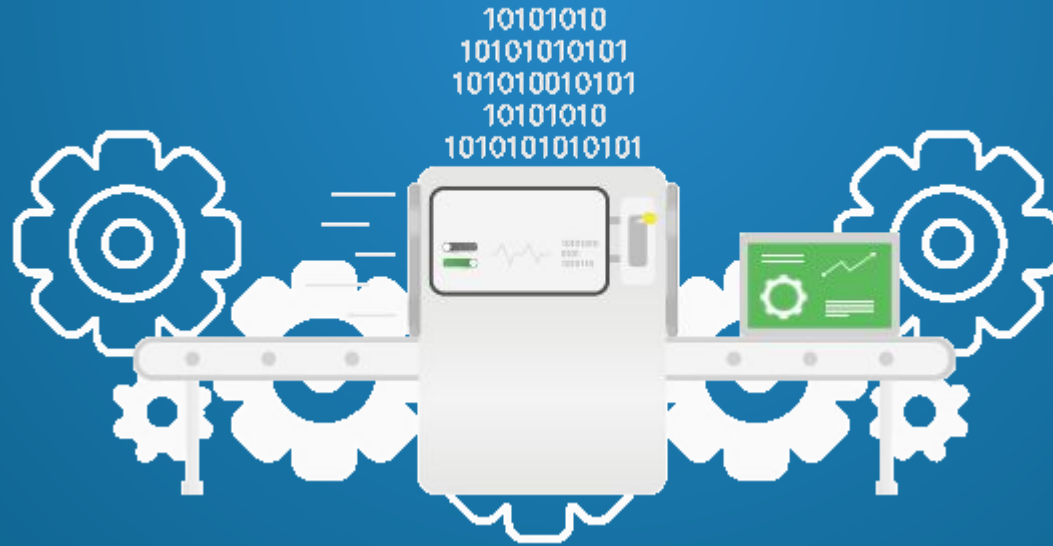
**Digital Presence** of business **correlates** to the **success** of a company

Accenture - **Growing the Digital Business**  
**Analysis on digital presence and profitability of business**

**Growth**  
+  
**Digital Presence**



 **SimilarWeb**



We predict whether a startup will **cross the Series C** type funding.

# OUR DATA SOURCES



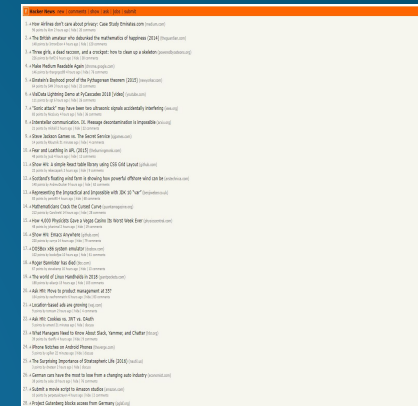


Hacker News is a social news website focusing on computer science and entrepreneurship run by startup incubator YCombinator.

## How we get the data ? Data Since 2006 18 Million Stories - 3.1 GB Stories Data - 4 GB Comments

Public data available on Google big query containing all the comments and stories.

[https://bigquery.cloud.google.com/dataset/bigquery-public-data:hacker\\_news](https://bigquery.cloud.google.com/dataset/bigquery-public-data:hacker_news)



### What attributes we intend to generate from the data ?

We need the data mentioned below in three time ranges that is **inception to seed funding**, **Seed Funding to Series A** and **Series A to Series B**.

- 1) Number of stories related to startups.
- 2) Number of comments related to startups.
- 3) Cumulative up votes and down votes on the stories about the startups.
- 4) Number of job postings.

by	STRING	NULLABLE	Username of commenter or submitter
score	INTEGER	NULLABLE	Story score
time	INTEGER	NULLABLE	Unix time
title	STRING	NULLABLE	Story title
type	STRING	NULLABLE	Type of details (comment, comment_ranking, poll, story, job, pollopt)
url	STRING	NULLABLE	Story url
text	STRING	NULLABLE	Story or comment text
parent	INTEGER	NULLABLE	Parent comment ID
deleted	BOOLEAN	NULLABLE	Is deleted?
dead	BOOLEAN	NULLABLE	Is dead?
descendants	INTEGER	NULLABLE	Number of story or poll descendants
id	INTEGER	NULLABLE	Unique type ID
ranking	INTEGER	NULLABLE	Comment ranking

Big Query Data Schema

**Traction** the startup is generating on **social media**



**Facebook – page feeds and total likes.** The page feeds are extracted in three different ranges that is inception to Series A, Series A to Series B and Series B to Series C.

**How we get the data? Graph API**

Graph API Explorer

Access Token: EAACEdEose0cBAEFYVrxZB9PctpuKYNBwDLVYs3WcQNrVxHHg1Cn8lqeBvS9cZAFigQvSyQjnRihp7QLh7gqrS7J2tsZCn8K9bFaK82ZBK

GET → /v2.12 /visier/feed

Edge: visier/feed

+ Search for a field

```
{
  "data": [
    {
      "created_time": "2018-03-02T18:13:03+0000",
      "message": "Meet People Analytics experts for lunch in Toronto on April 5 and get answers from HR Tech expert Lexy Martin.",
      "id": "199188963462452_1720790554635611"
    },
    {
      "created_time": "2018-03-01T20:20:01+0000",
      "message": "Dave Weisbeck talked to CIO Magazine about how data & analytics can help companies identify if they have a tox",
      "id": "199188963462452_1719754898072510"
    },
    {
      "created_time": "2018-03-01T17:01:39+0000",
      "message": "Discover how Visier People, the leading people analytics & workforce planning solution, enables HR to connect",
      "id": "199188963462452_171958595809404"
    },
    {
      "created_time": "2018-03-01T00:58:18+0000",
      "message": "Be sure to check out Visier CEO John Schwanz's article, \"Which Comes First - Management Or Analytics Software",
      "story": "Visier shared HR.com's post.",
      "id": "199188963462452_1718907828157217"
    },
    {
      "created_time": "2018-03-01T00:56:44+0000",
      "story": "Visier updated their cover photo.",
      "id": "199188963462452_1718906848157315"
    },
    {
      "created_time": "2018-02-28T23:13:05+0000",
      "message": "Goodbye sleep, hello baby! I",
      "story": "Visier added a new photo.",
      "id": "199188963462452_1718837271497606"
    },
    {
      "created_time": "2018-02-28T23:13:04+0000",
      "message": "Goodbye sleep, hello baby! I",
      "story": "Visier shared a post.",
      "id": "199188963462452_1718837234830943"
    }
  ]
}
```

Response received in 241 ms

Copy Debug Information </> Get Code Save Session

News Feed

Graph API Explorer

Access Token: EAACEdEose0cBAEFYVrxZB9PctpuKYNBwDLVYs3WcQNrVxHHg1Cn8lqeBvS9cZAFigQvSyQjnRihp7QLh7gqrS7J2tsZCn8K9bFaK82ZBK

GET → /v2.12 /visier/likes

Edge: visier/likes

+ Search for a field

```
{
  "name": "Human Capital Management Institute",
  "id": "158135087563646"
},
{
  "name": "BranchOut",
  "id": "194620133889100"
},
{
  "name": "Michelleharris.social",
  "id": "286402439379843"
},
{
  "name": "STETrevisions",
  "id": "95395235501"
},
{
  "name": "Sage",
  "id": "147313841946705"
},
{
  "name": "Talentnet",
  "id": "131228530245152"
},
{
  "name": "SecureSheet Online Spreadsheet",
  "id": "178883342145225"
},
{
  "name": "DriveThru HR",
  "id": "104663489565177"
},
{
  "name": "Halogen Software",
  "id": "156967857652009"
},
{
  "name": "IMPACT PEOPLE PRACTICES",
  "id": "180218101428"
},
{
  "name": "Impact99",
  "id": "172996122757258"
}
```

Response received in 241 ms

Copy Debug Information </> Get Code Save Session

Like

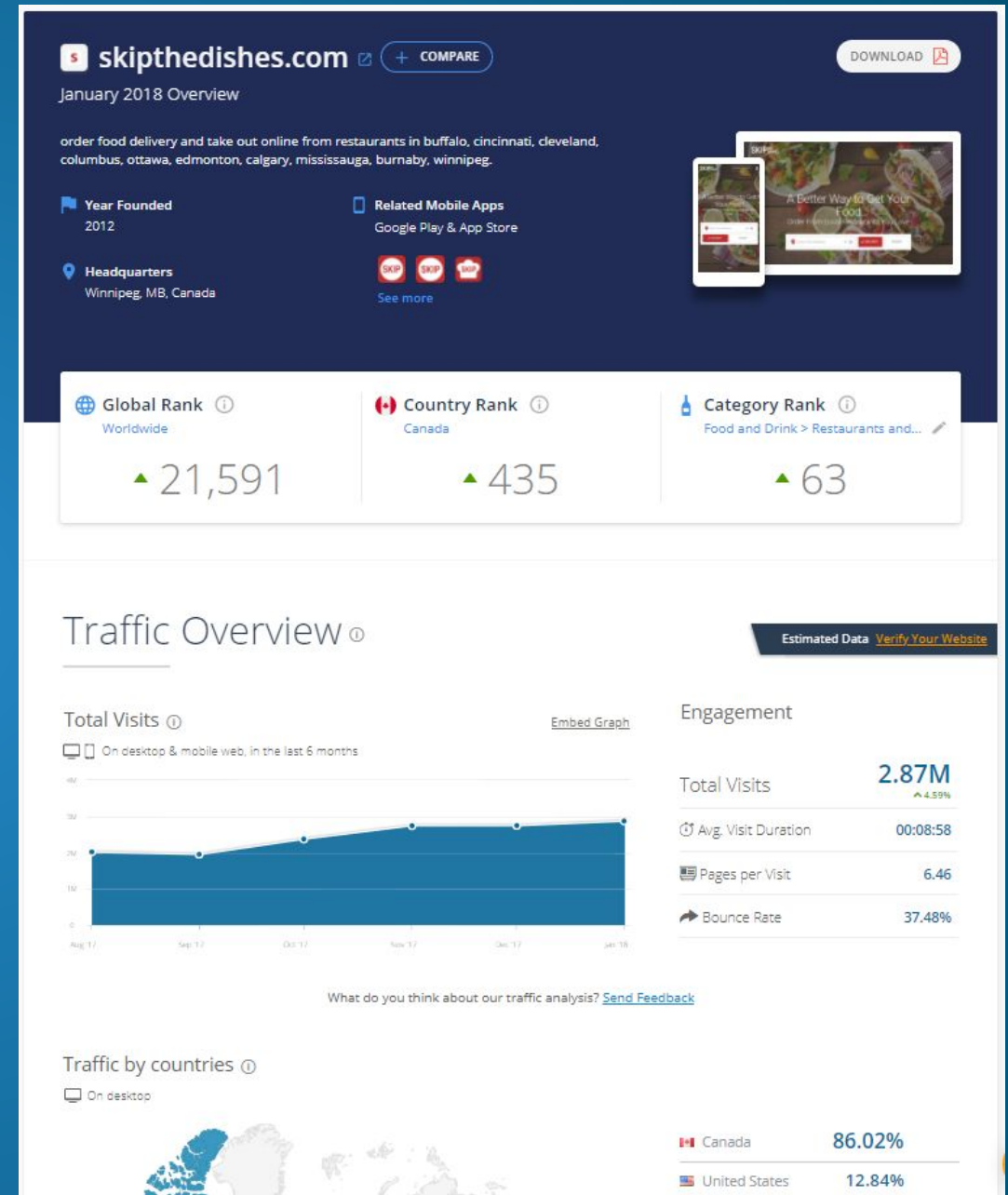
**Limitation:** We need to access the page insights API in order to get the page likes and other metrics related to the startup page which requires admin access to page, hence we are using the data available.



**How do we get the data?** We simply scrape it for now there is a paid API as well.

## Attributes interested in

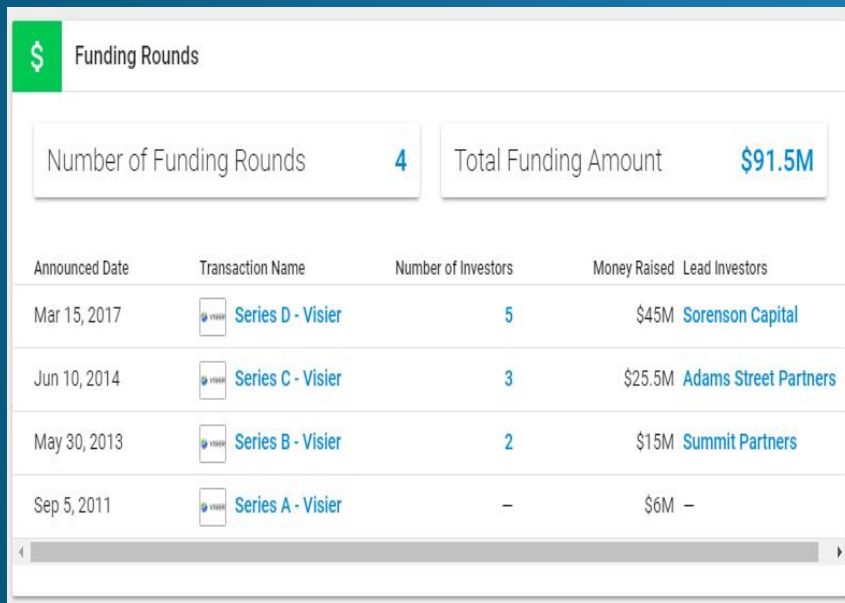
- Global Rank
- Country Rank
- Category Rank
- Total Visits
- Average Visit Duration
- Bounce Rate



## How we get the data ?

- Currently we are scrapping data from CruchBase for our prototype.
- In future we can use the enterprise license to get the data we need

## Data we are interested in!



The screenshot shows the 'Funding Rounds' section of a startup's profile on CrunchBase. It includes summary statistics and a detailed table of funding rounds.

Announced Date	Transaction Name	Number of Investors	Money Raised	Lead Investors
Mar 15, 2017	Series D - Visier	5	\$45M	Sorenson Capital
Jun 10, 2014	Series C - Visier	3	\$25.5M	Adams Street Partners
May 30, 2013	Series B - Visier	2	\$15M	Summit Partners
Sep 5, 2011	Series A - Visier	—	\$6M	—

- Number of funding rounds
- Total Funding Raised
- Lead investors at each stage. (They are a couple 100 investing most of the startups at initial stages)
- Date of funding stages. (This is basically how we know the time period of three stages for each startup)
- Last funding stage reached - help us get our label

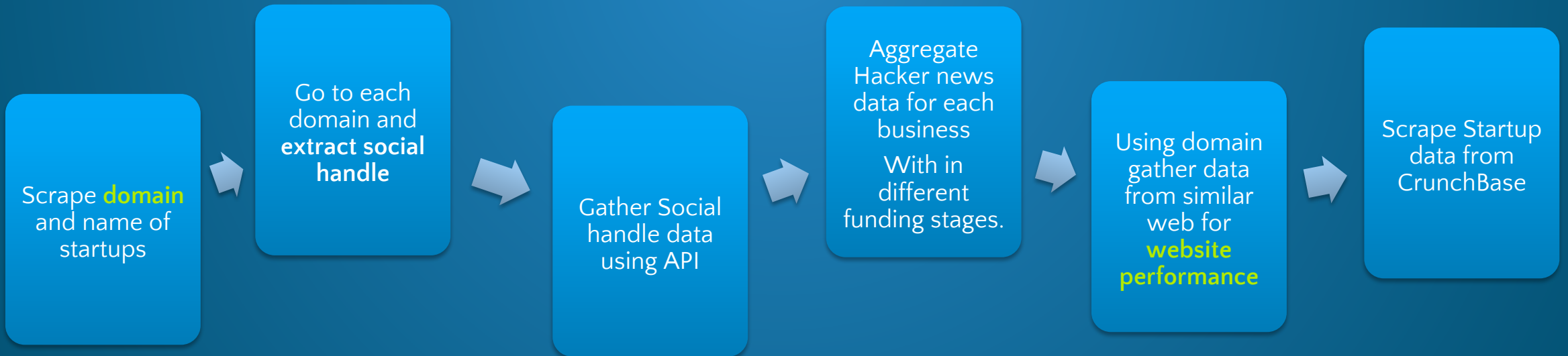


## Finally we also need a list of startups as well.

To get all the digital presence data mentioned above we just need one thing about a startup its **domain**.

We have scrapped information  
(website, Facebook page, twitter  
handle) around  
**42000** startups from different web  
sources.

### Data Collection and Integration Process





# How will Our **Model Predict** a **Successful Startup**?



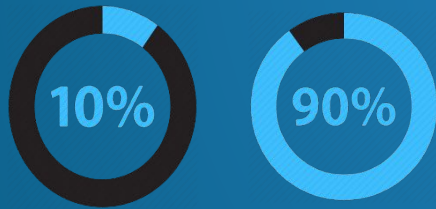
We want to predict if a **startup** which is in Series A or B stage will succeed



Train our model on **previously failed and successful companies**, different models for success prediction at different funding stages



**Binary classification problem** which we intend to solve using **appropriate machine learning algorithm**



The model returns the **confidence, likelihood** of the startup to reach Series C funding

## Challenges to Overcome

- Extract time information from hacker news articles by using change patterns in data aggregation over time ranges for data sources that does not have time information and only the current cumulative sum.  
**As we need data aggregation over the time range of startup funding stages**
- Twitter has changed its API policy for batch historical data and currently only provide last 30 days tweets for a handle. **We have ignored twitter for now only using facebook feeds.**
- Use API for *Similar web* for more data, such as Country Rank based on time. **This problem can be easily solved once we start using API** (however, It's paid)

## Future Improvements

- For the machine learning model generate features from hacker news articles, twitter and Facebook posts using natural language processing techniques.
- Predicting Growth of company such as the size of company and revenue.
- Analyzing Startup based on the team previous experience.

# Project Execution Plan And Progress

(February 12, 2018 – April 8, 2018)

## PROBLEM DEFINITION AND SOLUTION



Research and analyzing  
what data is available to  
us and how we can use it.

15 days

12<sup>th</sup> February – 26<sup>th</sup> February



Complete

## DATA COLLECTION AND INTEGRATION



Scraping data, cleaning  
and transformation of  
data

14 days

27<sup>th</sup> February – 12<sup>th</sup> March



Inprogress (70%)

## MODEL AND TUNING



Using various machine  
learning model for  
classification, comparing  
accuracy and tuning

11 days

12<sup>th</sup> March – 22<sup>nd</sup> March



Not Started

## PRODUCT PROTOTYPE

Web Frontend with d3.js  
visualization running on  
backend api based on  
Django framework

16 days

23<sup>rd</sup> March – 8<sup>th</sup> April



Not Started



# Questions & Feedback