Data scraping

Searching for contact information of CEOs, founders and lower managers of VCs in LA

⚠ **WARNING!**

NDA: some names and data in the sample report have been hidden or changed to comply with NDAs.

A	B	C	D	E	F	G	I	J	K	L	M	N
name	founded	investr	exits	about	portfolio	stages	website	focus	ROLE	NAME	EMAILS	LinkedIn
3L Capital	2017	39	4	3L is a multi-stage	Place, Polly, TerraTrue, Getta	Seed, Series A, Se	https://www.3lc	PropTech, Software, Enterprise, InsurTech, SaaS, E-Commerce	CEO	Mark Souvenir	mark@3l-capital.cor marksouvenir@gma msouvenir@3l-capit	https://www.linkedin.com
Act One Ventures	2016	61	2	Act One Ventures i	BLAZE, AuditBoard, Narrativ, Veryfi	Seed	http://www.acto	Hardware, RetailTech, Software, SaaS, Advertising (AdTech), Big Data, Analytics, Marketing (MarTech), E-Commerce, FinTech, Apps, Developer Tools, Enterprise, Artificial Intelligence & Machine Learning (AI/ML)				
Karlin Ventures	2013	51	14	Karlin Ventures, a (		Pre-Seed, Seed, S	http://www.karl	Software, SaaS, Consumer, FinTech				
Goal Ventures	2018	6	0	Goal Ventures is a		Seed, Series A, Se	https://goalvent	Advertising (AdTech), Artificial Intelligence, Machine Learning (AI/ML), E-Commerce, FinTech				
Flux Capital	2020	12	0	Flux Capital is a ver	SpaceX, Carta	Seed, Series A, Se	http://www.flux	SaaS, FinTech, Consumer, E-Commerce				
Emerging Ventures	2019	7	0	Emerging Ventures		Seed, Series A, Se	https://emergin	Artificial Intelligence, Machine Learning (AI/ML), ClimateTech, CleanTech, IoT (Internet of Things), Robotics, SaaS				
Embedded Ventures	2020	5	0	Skunkworks of dee	Inversion, KittyCAD, Slingshot Aerospace	Pre-Seed, Seed	http://embedde	ClimateTech & CleanTech, Robotics, DeepTech				
Fika Ventures	2016	63	3	Fika Ventures is a s	Edge Impulse, Berbix, Openpath, Elementary Robotics, WeeCare, First Resonance, Spinwheel, Chatdesk, HeyRenee, PathSpot, Bowery Valuation, Specright, PullRequest	Seed	http://www.fika	Artificial Intelligence & Machine Learning (AI/ML), Apps, Developer Tools, Cybersecurity, SaaS, Software, Hardware, PropTech, IoT (Internet of Things), Robotics, EdTech, FinTech, Food and Beverage				

Table of content

Introduction: conditions, goals and review	*Initial client task and analysis of possibilities to complete it*	3
Sources, Methods and Options	*What options (sources,methods and final dataset format) are available for the client*	4
Milestones	*Our steps and their order*	5
Final Database	*How it looks and what is a result*	6

Introduction: conditions, goals and review

Context, Input data and Task

> We are looking for a detail-oriented researcher to compile a comprehensive list of 100 venture capital firms located in Los Angeles. The final deliverable should include the following information for each firm:
>
> - Company name
> - Owners' or CEOs' full names
> - Email addresses and other relevant contact information
> - Official website URL
>
> This project requires a high level of accuracy and attention to detail to ensure the data collected is current and reliable. The ideal candidate will have experience in data research and a solid understanding of the venture capital landscape.

Task review

- Venture capital firms are quite open companies, and the names of top executives are usually not hidden: this part of the task is easily feasible
- The public pages of these people are also usually available on Facebook or LinkedIn, because presence on these platforms is part of the funds' work
- Personal contacts of top executives of venture capital funds in the US can mostly be found in various sources (page 4)
- The specified number of funds also exists

Solution: the task is feasible.

Options could be combined

1. List(s) of VC firms (data sources)

- **Provided by Client:** if the client knows which firms they are interested in, I can skip searching for them and go straight to collecting the necessary information
- **Parsed from catalogs:** if there is no initial list, I can compile one
- **Searched manually:** in some cases, manual search for source data is required, but this is rare

2. Sources and Methods

- **Lists of sources:** one or more and their degree of complexity - these are all options that can be chosen
- **Parsing:** automatic collection of information using parsers
- **Manual collection:** works when hidden or unpopular information is needed

3. Data format and file import preparation

- **Google Spreadsheet (by default):** Preference for buying in physical stores, but with growing adaptation to online orders.
- **File import preparation:** data format, file extension, readiness of the file for easy import
- **Import setup:** if you don't want to import yourself, I can also help with importing into your system if I'm familiar with it and you provide access

Milestones

Checking an availability of needed data, acceptation of error share

After discussing the task details, it is necessary to check how accessible the contacts are and how open people are in this field. Look for niche contact databases, if they exist, and estimate how hard and with what amount of errors or missing rows the final dataset will be, and agree with the client.

Definition of needed file format

It is important to do this before starting work, as the collection will initially be carried out taking into account the required format. The data will be collected in a Google Spreadsheet, but the client may need to process or import the data. Therefore, if necessary, the file can be prepared.
For example:

- .csv, contact numbers without extra characters, only one email per line

or

- any format available for import into HubSpot

In case the import errors, I will help fix them and deliver the data to your CRM

Acceptation of data sources and primary data amount

You will send a list of sources, companies or people you need, or I will find them myself. In the latter case, I will show you the sources, and we will agree if they are suitable for you. It could be, for example, a niche catalog with 543 firms - and you will respond that it is suitable. The final dataset will have ±543 firms (there may be discrepancies: empty catalog pages, liquidated companies, etc)

[WORKING] Data parsing, scraping and preparation

Acceptance of work

Fixes and data format changes if needed. DONE!

Result

Parsing of source data (catalog)

The client required essential data: company name, website, full name of at least one key person, and their email or alternative contact. The source catalog provided *additional information*, which was also included in the dataset: founding year, investment and exit counts, description, portfolio, focus areas, and preferred partnership stages.

Contact collection (manual)

For each firm, data on key company personnel was collected: names, their positions, email, LinkedIn and other information that I considered significant. The final file is presented as a Google Sheet and a .csv with only names and emails, prepared for import into the contacts.

A	B	C	D	E	F	G	I	J	K	L	M	N
name	founded	investr	exits	about	portfolio	stages	website	focus	ROLE	NAME	EMAILS	LinkedIn
3L Capital	2017	39	4	3L is a multi-stage	Place, Polly, TerraTrue, Getta	Seed, Series A, Se	https://www.3lc	PropTech, Software, Enterprise, InsurTech, SaaS, E-Commerce	CEO	Mark Souvenir	mark@3l-capital.cor marksouvenir@gma msouvenir@3l-capit	https://www.linkedin.com
Act One Ventures	2016	61	2	Act One Ventures i	BLAZE, AuditBoard, Narrativ, Veryfi	Seed	http://www.acto	Hardware, RetailTech, Software, SaaS, Advertising (AdTech), Big Data, Analytics, Marketing (MarTech), E-Commerce, FinTech, Apps, Developer Tools, Enterprise, Artificial Intelligence & Machine Learning (AI/ML)				
Karlin Ventures	2013	51	14	Karlin Ventures, a		Pre-Seed, Seed, S	http://www.karl	Software, SaaS, Consumer, FinTech				
Goal Ventures	2018	6	0	Goal Ventures is a		Seed, Series A, Se	https://goalvent	Advertising (AdTech), Artificial Intelligence, Machine Learning (AI/ML), E-Commerce, FinTech				
Flux Capital	2020	12	0	Flux Capital is a ver	SpaceX, Carta	Seed, Series A, Se	http://www.flux	SaaS, FinTech, Consumer, E-Commerce				
Emerging Ventures	2019	7	0	Emerging Ventures		Seed, Series A, Se	https://emergin	Artificial Intelligence, Machine Learning (AI/ML), ClimateTech, CleanTech, IoT (Internet of Things), Robotics, SaaS				
Embedded Ventures	2020	5	0	Skunkworks of dee	Inversion, KittyCAD, Slingshot Aerospace	Pre-Seed, Seed	http://embedde	ClimateTech & CleanTech, Robotics, DeepTech				
Fika Ventures	2016	63	3	Fika Ventures is a s	Edge Impulse, Berbix, Openpath, Elementary Robotics, WeeCare, First Resonance, Spinwheel, Chatdesk, HeyRenee, PathSpot, Bowery Valuation, Specright, PullRequest	Seed	http://www.fika	Artificial Intelligence & Machine Learning (AI/ML), Apps, Developer Tools, Cybersecurity, SaaS, Software, Hardware, PropTech, IoT (Internet of Things), Robotics, EdTech, FinTech, Food and Beverage				