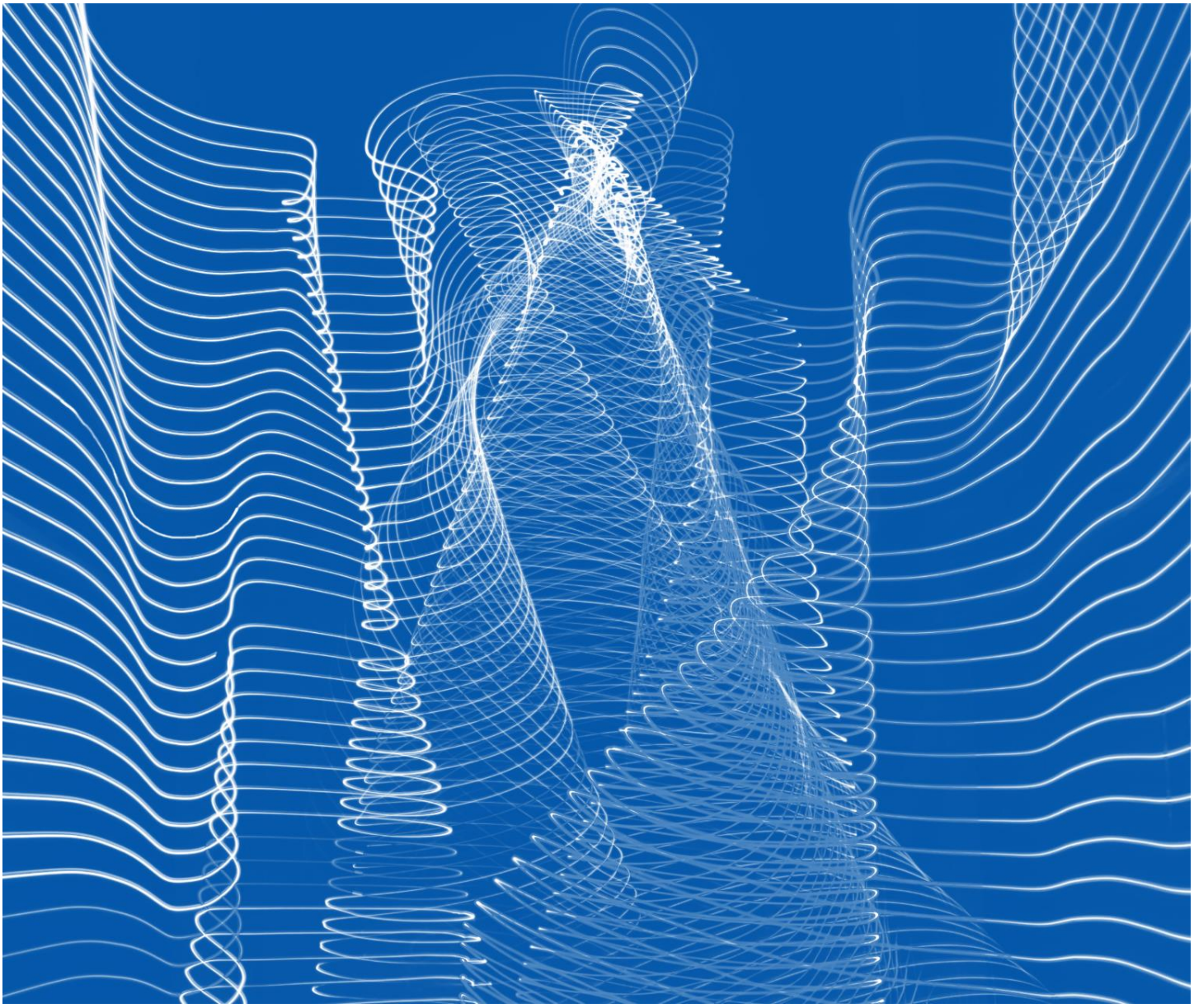
Public Disclosure Authorized



Digital Pulse

An exploration of non-traditional data for entrepreneurship ecosystem diagnostics

Credit: Photo by Zak Sakata on Unsplash (License allows free use of images).

Rights and Permissions

Table of Contents

List of Tables

List of Figures

Acknowledgements

This publication was funded by 'Digital Industries and Skills Development Sharing: The Korean Experience' grant funded by the Korean Trust Fund for ICT4D at the World Bank and implemented under the Digital Entrepreneurship Program (DEP), Global Knowledge and Learning project.

The Innovation Policy Platform (IPP – www.innovationpolicyplatform.org), was developed by the World Bank Group (WBG) and the Organization for Economic Co-operation and Development (OECD) as a global resource for knowledge, learning, indicators/data, and communities of practice on the design, implementation, and evaluation of innovation policies around the world. It was one of the core external knowledge products offered by the Finance, Competitiveness & Innovation (FCI) global practice and served to raise awareness of FCI's product portfolio, facilitate global engagement and advocacy, and build staff skillset. The platform was retired in 2019.

Several people provided input and contributed to the report. Najy Benhassine and Denis Medvedev provided overall guidance. The project was led by Prasanna Lal Das with support from Adela Antic. Ma. Regina Paz Saquido Onglao is the principal author of the report, with contributions from Prasanna Lal Das. Alberto Sanchez Rodelgo (IMF) and Romulo Cabeza (ILO) were the peer reviewers.

1 Executive Summary

The ongoing data revolution has made a prodigious amount of new data and analytical tools available to researchers. This data is available at higher frequency and at a much more granular level that traditional data collected through field work and surveys. This creates new opportunities for research but also raises significant questions about the usefulness, reliability, and quality of such data. Proponents of big data research have called for the development of new analytical techniques and tools to take advantage of new opportunities while others have cautioned against its seductive power.

In the current note we provide an initial examination of the usefulness of such data in the context of entrepreneurship ecosystem diagnostics. The World Bank is currently updating the methodology it uses to assess entrepreneurship ecosystems, in particular the Digital Entrepreneurship Ecosystem Diagnostic (DEED) framework.[1] One of the features of the new methodology is an 'all data' approach that seeks to blend standard data sources (surveys, official data) with online data (open or proprietary).

Entrepreneurship ecosystems are fluid environments containing complex interactions and relationships between entities. Most of the current ecosystem assessments rely on secondary sources of data that are generally based on small samples and often don't include information about entity relationships and networks[2]. Such methodologies are also generally expensive to repeat. This leads to significant data gaps, including coverage and timeliness. Primary data, when collected, is seldom global and generally infrequently gathered. Almost no assessment methodologies utilize so called 'big data' sources and very few reuse or combine data from non-traditional sources like online platforms. And most methodologies focus on a specific set of actors within the entrepreneurship ecosystem – either the firms or investors or government agencies or intermediaries, but almost never all of them. This means that the findings of such ecosystem assessments are often high-level, not comparable over time and geographies, and not necessarily actionable.

---

[1] The DEED toolkit relies on a framework used to assess 1) the current environment, 2) strengths & successes, 3) weaknesses & barriers, and 4) opportunities for growth across the six domains of an entrepreneurship ecosystem identified by the Babson Entrepreneurship Ecosystem Project: policy, financial capital, markets, culture, human capital, and supports.

[2] Exceptions include the following reports: (i) the World Bank Ecosystem Connections Mapping Project in collaboration with GERN, Endeavor, and other institutions which maps connections between key actors within startup ecosystems around the world; (ii) Endeavor Insight's report on "The Power of Entrepreneur Networks" focusing on how founder networks have accelerated New York City's tech sector growth, and (iii) Startup Genome's Global Startup Ecosystem Report which uses company and founder data to generate Local and Global Connectedness index measures.

The assessment challenges have been further exacerbated in the digital economy in which economic activity, including entrepreneurship, is difficult to measure fully using traditional indicators. Digital entrepreneurs, whether they be start-ups or within incumbent firms, face several new and different challenges (and opportunities) than 'traditional' entrepreneurs do. Digital businesses also generate new types of data whose exhaust can be a powerful way to measure conditions that are unique to the digital entrepreneurship ecosystems.

The approach described in the current note tries to address the challenges in measuring entrepreneurship ecosystem assessments using alternative data and related techniques. The note describes –

- New data sources and data collection techniques covering –
    - Basic definitions
    - Review of related literature for exploring new data sources
    - Identification and prioritization of data sources
    - Examples of scraped data and their presentation
- Natural language processing (NLP), visualization, and machine learning techniques including -
    - Named entity recognition and classification for entity extraction
    - Sentiment analysis
    - Topic modeling
    - Thematic classification
    - Network visualization
- Data quality issues and mitigations covering issues such as –
    - Linkages between different sources
    - Data management and storage
    - Biases in data and quality testing
    - Limitations of web scraping
- A brief discussion of the ethics of web scraping

The examples used in the note are drawn from developing countries such as Senegal, Kenya, and Indonesia which tend to be 'data poor' and where the proposed approaches may both have the greatest potential but also face the most significant challenges given their relatively low level of digital development.

Please note that the current report is designed as a data science practitioner guide and assumes a degree of technical familiarity with the subject matter. We should also clarify that we do not propose that alternative data should replace or is 'superior' to other data sources – the purpose of the current note is to provide a technical examination of specific data sources and the tools available to utilize them.

It is also important to consider the sustainability and reproducibility of web scraping when incorporating such data into the research methodology. Many sites have begun to close themselves off to scrapers and while not widespread this may apply more forcefully to some projects than others.

The code associated with the work below is available at https://github.com/mrpsonglao/Machine-Learning-Pilots. Note that we have scrubbed the code of any personally identifiable information, to make it fit for public use.

2 Exploration of alternative data sources

The modern world is awash in data. As a recent world Bank report on data driven development[3] pointed out, just in one second people send 2.7 million emails, watch 75,000 videos on YouTube, and transmit almost 60,000 gigabytes of data. In that one second, many individual airplanes generate 10 GB of data and connected cars gather even more GBs of data about everything ranging from weather and traffic conditions to every driving action and the response of other vehicles on the road. This data, as has been documented by the World Bank[4] and others, is an important source of economic growth and to deliver public services.

The research community, after initial skepticism, has gradually warmed to the benefits of such data. In the US, for instance, the Bureau of Labor Statistics increasingly uses 'big data' to track the economy[5]. Examples of such data include apparel prices gathered directly from big departmental stores, vehicle prices gathered directly from private sector aggregators, and drug prices sourced from pharmacy chains. In Canada government statisticians have started collecting price data online. Statistical agencies in New Zealand, Norway and the Netherlands also gather sales data through checkout scanners in stores. Similarly, the Billion Prices project, seeded at MIT, began by scraping data from online sellers at scale.

The drivers for such work include the need for more frequent, cheaper, and timely data that are relevant for policymakers and allow researchers to fill data gaps. Access to such data at a large scale also lets researchers test non-probability sampling methods (***Section 4.3 Adjusting for Biases & Testing Quality of alternative data sources*** provides more details on non-probability sampling methods).

In this section we provide a broad definition of alternative data and describe tools and techniques we employed to scrape data from online sources to support entrepreneurship ecosystem assessments in Senegal, Kenya, and Indonesia. The steps included –

1. Identify and prioritize data sources for web scraping;

---

[3] Harnessing data technologies for development https://openknowledge.worldbank.org/handle/10986/30437

[4] Internet of Things – the new government to business platform http://documents.worldbank.org/curated/en/610081509689089303/Internet-of-things-the-new-government-to-business-platform-a-review-of-opportunities-practices-and-challenges

[5] Government economists turn to big data to track the economy https://www.wsj.com/articles/government-economists-turn-to-big-data-in-estimating-inflation-11556622001

2. Extraction of a list of ecosystem actors (e.g., companies, accelerators, incubators) from private sector big data sources or public websites;
3. Extraction of actor-level metadata (e.g., year founded, address, number of employees, connections with other actors) from private sector big data sources or public websites.

The compilation of all sample outputs for this section can be accessed in a shared view-only Google Drive folder.[6]

2.1 What is alternative data and web scraping?

In the current note, we use the term 'alternative data' to refer to online, digital data either published in consumable open data format or otherwise available for scraping. Such data includes open data sets such as the ones published by the World Bank at http://data.worldbank.org, social data on platforms such as Twitter, general online content such as on http://worldbank.org, or data made available by proprietary resources such as https://www.telegeography.com/.

Web scraping or data scraping refers to automated 'copying' of the content of a website into a database, typically through a bot or a web crawler.

2.2 Prior work on alternative data collection methods

There are several examples of previous successful work which leverage both traditional and alternative data sources, as documented by Blumenstock, Cadamuro, and On (2015); Olson (1996, 1999); Beskow, Sandler, and Weinberger (2006); and Ginsberg et al. (2009). Other examples include the combination of Facebook and survey data Burke and Kraut (2014) and the research by Ansolabehere and Eitan Hersh (2012) on US voting patterns using proprietary data.

The table below, inspired by Salganik 2017, highlights the strengths and weaknesses of traditional and alternative data sources.

Table 1. Strengths and Weaknesses of Traditional and Alternative data sources

	Strengths	Weaknesses
Traditional data (e.g., surveys, interviews)	• Custom-made for the research problem • In depth • Good for opinion and perception-related questions	• Usually narrow in scope • Usually expensive and suffers from funding and/or time constraints • Infrequent or not timely • Lack of scale with respect to geographic coverage • Lack of coverage with respect to ecosystem actors • Publicly-available data usually lacks granularity

[6] The view-only Google Drive folder can be accessed here: https://drive.google.com/drive/folders/1VW07ZUisEhcH1yQnt9Vg7fQlvfU1XJjr?usp=sharing

	Strengths	Weaknesses
Alternative data	• Big, which allows minimizing of random error or noise during modelling • Provides real-time estimates • Substantially cheaper • Usually provides more granular data	• Digital biases such as non-representative and systemic biases • Sparse or incomplete data • Possible drifting especially for social media platforms, such as population drift (change in user base), behavioral drift (change in how users use the platform), and system drift (change in the system itself) • Algorithmically confounded, that is, user behavior is affected by engineering goals of the systems • Some may contain sensitive data, which is a potential risk for data ownership and legal use issues

The animating idea behind such work is that combining traditional and alternative data allows researchers to produce a larger, richer, and more complete database than using one or the other. Using both traditional and alternative methods affords researchers the benefits of both types of data sources - the in-depth and custom-made nature of traditional data sources together with the scale, speed, and granularity of alternative data sources. It also mitigates the weaknesses of each one.

- Alternative data sources and traditional data can complement each other by filling each one's data gaps.
- Alternative data sources can augment the sampling frame for traditional surveys by providing a potential list of respondents.
- Traditional data can supplement alternative data sources by providing representative data with which to check or triangulate the alternative data against..

Salganik 2017 provides more detail about the ideas above and introduces the concept of "enriched asking".

2.3 Identification and prioritization of data sources for web scraping

Alternative data comes in many shapes and formats and from a variety of sources including social media, websites, IoT, and others. Depending on the research question, the first step is to develop a list of criteria to prioritize certain sources and data types over others. For the work on entrepreneurship ecosystem diagnostics, the team decided to focus specifically on online data and shortlisted the list of sources based on the following criteria:

- Accessibility – Listed both public and private/proprietary datasets

- Scope – Listed both global and country-specific data sources for Indonesia, Vietnam, Kenya, Senegal, and Nigeria
- Granularity – Listed both country-level and company-level data sources

Metadata, such as the following, per data source[7] was also noted down for analysis and prioritization purposes:

- Goal or intention for web scraping (e.g., for extracting lists of companies, for extracting ecosystem actor-level metadata)
- Potential extraction methodology, data type, and notes
- Important notes and potential issues when extracting data from the source

To identify the data sources for the current demonstration project, the team conducted desk research including online searches (Google) and parsing through relevant entrepreneurship-related documents and toolkits to shortlist potential data sources to scrape.

The team employed an iterative and test-heavy approach in extracting data from these sources. To strategically test web scraping across these sources, the team further shortlisted data sources for initial web scraping based on two main criteria -- extraction priority and extraction difficulty. The shortlisting criteria and sub-criteria are further detailed in the table below.

Table 2. Criteria for Prioritizing Data Sources for Web Scraping

Criteria	Possible values	Rationale
Extraction Priority (1=highest to 5=lowest)	Assessed priority of extracting the data for this specific dataset. General rule of thumb: • 1 (highest) = Public, global datasets • 2 = Public, region/country-specific datasets with ecosystem actor-level data • 3 = Public, region/country-specific datasets with country-level data • 4 = Relevant but proprietary sources • 5 (lowest) = Other data sources for exploration The sub-criteria used is as follows:	
	Source Accessibility (Public vs Proprietary)	There are terms of use limitations for proprietary sources, which usually require a fee or partnership to access the granular data. Public datasets allow us to freely download the data and use for research/analysis.

[7] For more details, please refer to this comment-only Google spreadsheet for the list of data sources considered and their corresponding attributes based on the criteria above: https://docs.google.com/spreadsheets/d/1nKkgaueEdiym1fmXUYdUy2Ctmh4_YDk-eVpG1PkQC4U/edit#gid=739685063

Criteria	Possible values	Rationale
	Geographic Scope (Global vs Country-specific)	A good mix is best for web scraping, since: • Global websites are preferred since this allows web scraping to be scalable, since the same code can be used to get data for more countries and companies within the same website domain (given that company/country pages usually have the same HTML structure within the same website). These can be used as the baseline or starting point of any web scraping activity. • On the other hand, country-specific websites tend to be localized and usually contain more unknown company data, so scraping these can augment and widen the scope of the scraped global websites.
	Granularity (Country-level vs Actor-level)	Ecosystem actor-level is preferred, since there are a lot of existing resources and datasets already for country-level data. On the other hand, ecosystem actor-level data is hard to find while allowing us to generate interesting in-depth insights.
Extraction Difficulty (1=Easy to 5=Hard)	Assessed difficulty of implementing the extraction for this specific dataset. General rule of thumb: • 1 (Easy) = Minimal coding required. Data is already in a table-structured, easily-parsable format (e.g., single JSON endpoint, CSV or Excel file). • 2 = Some coding required. Data can be pulled via API that is structured well, which allows creation of reusable code that is applicable across different countries/companies. • 3 (Intermediate) = Intermediate coding required. Scripts specific to the websites with somewhat structured data need to be written to extract data from page source. • 4 = Intensive coding required. Need to set up web crawlers to get slightly-structured website data, or would need to extract data regularly from PDFs. • 5 (Hard) = Method for extracting data is unclear / for exploration. Or, data pull is disallowed due to owner's recent decisions (e.g., Facebook). The sub-criteria used is as follows:	
	Extraction data type	Data type of the data to be extracted (e.g., JSON, HTML) affects ease in extraction. For example, PDFs are harder data types to extract data from when compared to CSV files or JSON endpoints.
	Extraction method	Extraction methods available (e.g., API, data download, website pull limits) affect ease in extraction since this will determine the difficulty and complexity of the scraping code required.
	Notes / Potential Issues	API rate limits and other notes will affect the scalability and frequency of usage of web scraping code.

To illustrate the results of data collection through web scraping, here is a sample company profile of an Indonesian startup through scraping diverse sources. Note that

the team has not implemented data quality checks for the sample profile below. For more details on data quality, please refer to ***Section 4 Data quality and mitigation*** below**.**

Table 3. Illustrative example: Indonesian startup profile using online data

Basic Profile	
Name	**Bukalapak**
Actor Type	startup
HQ Location	Indonesia
Description	Situs Jual Beli Online Mudah Dan Terpercaya
Description (detailed)	Bukalapak - Place of selling / buying the most comfortable & safe online with Payment System which ensures buyers and sellers 100% risk free online scams. Bukalapak.com, Sell Buy Easy & Reliable.
Founding Date	September 2011
Estimated Number of Employees	1,500
Industry	Marketplaces, E-Commerce
Company website	https://bukalapak.com/
Related Articles	http://endeavorindonesia.org/id/bukalapak-raih-penghargaan-bergengsi-dari-jokowi/

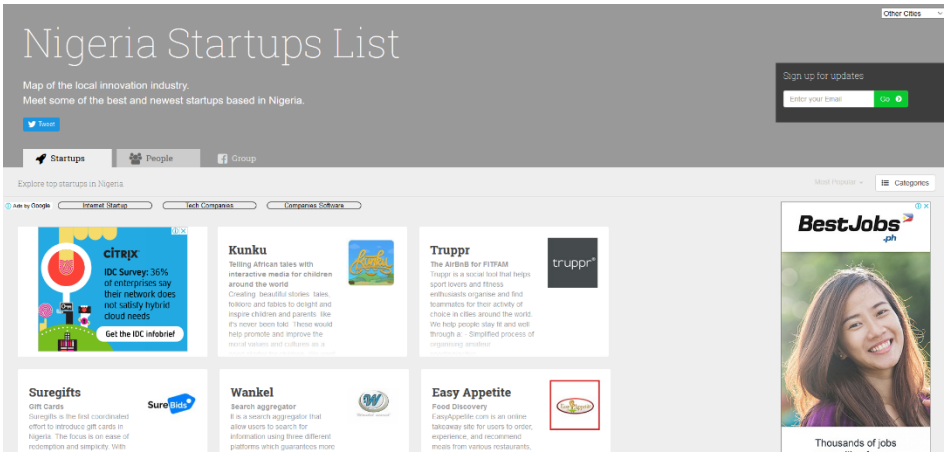
Founder data	
Co-founder & CTO	Nugroho Herucahyono
Co-founder & CEO	Achmad Zaky
Investor data	
Venture (November 2017)	Undisclosed
Series B (February 2015)	Emtek Queensbridge Venture Partners 500 Startups
Series A (September 2012)	Gree Ventures
Social Media accounts and statistics	
Facebook	https://www.facebook.com/bukalapak
Instagram	https://www.instagram.com/bukalapak
LinkedIn	https://www.linkedin.com/company/pt-bukalapak-com
Twitter	https://twitter.com/bukalapak

2.4 Extraction of data from private sector big data sources and public websites

Data extraction begins after the data sources have been identified and tested. For the current demonstration, the team extracted sample lists of companies, accelerators, and incubators from a diverse set of sources to show the kind of company data that can be extracted per source type.

Here are the results of the sample web scraping for lists of companies, accelerators, and incubators.

2.4.1 A data collector or directory website

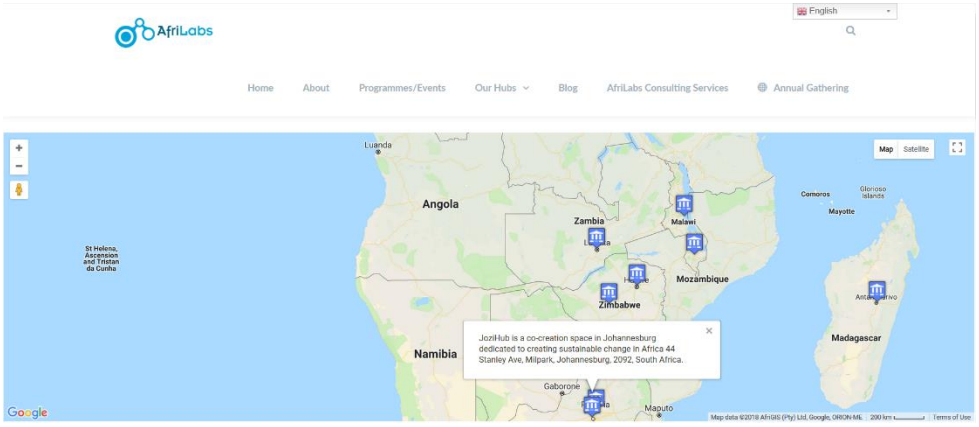
Data source	Startups List, a website which contains global country-specific listings (http://nigeria.startups-list.com/). Here's how the website looks like: 
Implementation & Results	We were able to extract data on 251 startups located in Nigeria using Python scripts on the website's page source.
Relevant extracted fields	Startup name, description, website, logo (link to image), and keywords.
Strategic use of scraped data	Having the startup website allows us to build another scraping layer by extracting data (e.g., text, contact details, images, etc.) from the startups' respective websites.

Here is what the sample output looks like. For the full output, please see `2018-07-10 - Nigeria Startups List.csv`.

	Startup Name	Description	URL	Keywords	Logo URL
1	Truppr	The AirBnB for FITFAM Truppr is a social tool that helps sport lovers and fitness enthusiasts organise and find teammates for their activity of choice in cities around the world. We help people stay fit and well through a: - Simplified process of organising amateur sporting/active ...	https://www.truppr.com/	Truppr - fitness personal health corporate wellness active lifestyle	https://d1qb2nb5cznatu.cloudfront.net/startups/i/390218-64243e8459fac2a0764c593ddcdc9608-thumb_jpg.jpg?buster=1398884019
2	RubiQube	Location-based app recommendation RubiQubeÂ® is a (cloud-based) mobile applications discovery and aggregator that seeks to connect locally developed mobile apps (HTML 5 apps) with their target market using a location based app	http://www.therubiqube.com	RubiQube - cloud computing android application platforms app stores	https://d1qb2nb5cznatu.cloudfront.net/startups/i/311337-3dabab1d944fc169b7529aa64f974de6-thumb_jpg.jpg?buster=1387363971

		recommendation system in the app store. The application is available …			
3	ChopUp	Mobile Social Gaming for Africa Chopup is a social platform that allows mobile game players to interact based on in-game achievements. The following are features of the platform: - Targeted exclusively at mobile devices (not excluding feature phones) - Social profiles for each user - Realtime …	http://www.chopup.me	ChopUp - social games social media platforms mobile games virtual currency	https://d1qb2nb5cznatu.cloudfront.net/startups/i/90872-aae89861281498b4fc2ba9ca37847637-thumb_jpg.jpg?buster=1371473075

2.4.2 An embedded map

Data source	Afrilabs – has a map of African accelerators / hubs (http://www.afrilabs.com/afrilabs-passport/). Here's a screenshot of how the interactive map looks like: 
Implementation & Results	We were able to extract data on 57 accelerators, incubators, or hubs in Africa using Python scripts in the map's underlying code.
Relevant extracted fields	Company address, description, geocoordinates, city, state, country, and postal code.
Strategic use of scraped data	Having the startup geocoordinates gives us flexibility in conducting geospatial analysis on the scraped company metadata. This will be a potential analysis dimension when doing network analysis.

Here is what the sample output looks like, filtered down to a few columns since the original dataset has many columns. For the full output, please see `2018-07-10 - Afrilabs List of Accelerators or Hubs.csv`.

Startup Name	Address	Description	Latitude	Longitude	City	State	Country

ActivSpaces	Cefam Rd, Buea, Cameroon	ActivSpaces is an open collaboration space, innovation hub and startup incubator for African techies. Established in 2009, ActivSpaces was one of the earliest African coworking spaces to provide free and open access to members actively pursuing technology-based ventures. Based in Buea, Cameroon.	4.1515548	9.2327857	Buea	Southwest	Cameroon
AkiraChix	Kenyatta Avenue, Nakuru, Kenya	AkiraChix is a not for profit organisation that aims to inspire and develop a successful force of women in technology who will change Africaâ€™s future.	-0.2849853	36.0693113	Nakuru	Nakuru County	Kenya

2.4.3 Google Search via search nearby places

Data source	Google Places API, "Nearby Search" endpoint (API documentation here: https://developers.google.com/places/web-service/search)
Implementation & Results	We were able to extract data on 60 establishments located near Nigeria. Specifically, we scraped all establishments on Google Places API which is within a 50km-radius from Nigeria's capital, Abuja.
Relevant extracted fields	Establishment name, geocoordinates, Google place_id, opening hours, photo (link), Google user rating, type of establishment (e.g., hotel, restaurant, lodging)
Strategic use of scraped data	Aside from providing accurate geocoordinates, using Google Places API allows us to expand the diversity in company types as well as data types extracted by including photos, user ratings, and opening hours in the mix. The "`place_id`" field also allows us to pull greater detail on the business/establishment using another Google API endpoint.

Here is what the sample output looks like, filtered down to a few columns since the original dataset has many columns. For the full output, please see `2018-07-10 - Nigeria Google Places API - Nearby Places endpoint.csv`.

Name	Geocoordinates	Google Place ID	Google Places rating	Entity type	vicinity
World Bank	{'location': {'lat': 9.0428389, 'lng': 7.523834399999998}, 'viewport': {'northeast': {'lat': 9.043978930291502, 'lng': 7.525221030291501}, 'southwest': {'lat': 9.041280969708497, 'lng': 7.522523069708496}}}	ChIJpS5lz-ELThAR7Hh_IdCw1HI	4.7	['bank', 'finance', 'point_of_interest', 'establishment']	102 Yakubu Gowon Crescent, Abuja
Ecobank	{'location': {'lat': 9.0756033, 'lng': 7.478640400000001}, 'viewport': {'northeast': {'lat': 9.0769522802915, 'lng': 7.479989380291503}, 'southwest': {'lat': 9.074254319708496, 'lng': 7.477291419708498}}}	ChIJbe1WOPkKThARbRjyUqnO7xo	5	['bank', 'finance', 'point_of_interest', 'establishment']	Ademola Adetokunbo Crescent, Abuja
Intercontinental Bank	{'location': {'lat': 9.0580188, 'lng': 7.486050700000001}, 'viewport': {'northeast': {'lat': 9.059367780291502, 'lng': 7.487399680291503}, 'southwest': {'lat': 9.056669819708498, 'lng': 7.484701719708498}}}	ChIJM8qNp6gLThARdPRXvFMXWRo		['bank', 'finance', 'point_of_interest', 'establishment']	Abuja

2.4.4 Google Search via text search on places

Data source	Google Places API, "Text Search" endpoint (API documentation here: https://developers.google.com/places/web-service/search)
Implementation & Results	We were able to extract data on 24 startups/accelerators/hubs located in Indonesia and 53 of which in Kenya. We did this by specifying a keyword list and querying Google Places API using those keywords while limiting the results to a specific country (e.g., Indonesia, Kenya). The keyword list used is: ["accelerator", "hub", "startup", "business", "company", "incubator"]
Relevant extracted fields	Similar to #3 above, but with the added metadata of keyword used for source, and country used to restrict search results.
Strategic use of scraped data	Having the keyword mapped to each result allows us to easily group results together as an additional analysis dimension.

Here is what the sample output looks like, filtered down to a few columns since the original dataset has many columns. For the full output, please see `2018-07-10 - Consolidated Google Places API - Text Search endpoint.csv`, `2018-07-10 - Kenya Google Places API - Text Search endpoint.csv`, and

2018-07-10 - Indonesia Google Places API - Text Search endpoint.csv.

Name of Entity	Country	Address	Geocoordinates	Keyword	Google Place ID	Google Places Rating	Entity Type
Kenya Methodist University	Kenya	Kemu Hub, Koinange St, Nairobi	{'location': {'lat': -1.2813276, 'lng': 36.8177021}, 'viewport': {'northeast': {'lat': -1.280026070107278, 'lng': 36.81895152989272}, 'southwest': {'lat': -1.282725729892722, 'lng': 36.81625187010728}}}	hub	ChIJTeyBBdMQLxgRxKyG9h36RbY	4.3	['university', 'point_of_interest', 'establishment']
Meru Institute Of Business Studies	Kenya	Njuri Ncheke Street, Meru	{'location': {'lat': 0.0466226, 'lng': 37.6554979}, 'viewport': {'northeast': {'lat': 0.04797242989272221, 'lng': 37.65684772989272}, 'southwest': {'lat': 0.04527277010727779, 'lng': 37.65414807010728}}}	business	ChIJ3axayeQhiBcRkkvyRGtL65w	5	['university', 'point_of_interest', 'establishment']
Murang'a University of Technology	Kenya	Murang'a University College, Muranga, MURANGA TOWN (fomer Fort Hall)	{'location': {'lat': -0.7163028, 'lng': 37.1476829}, 'viewport': {'northeast': {'lat': -0.7142872701072778, 'lng': 37.14864597989273}, 'southwest': {'lat': -0.7169869298927222, 'lng': 37.14594632010728}}}	business	ChIJ7RmbKXOYKBgRBhWoq4lTZ5c	4	['university', 'point_of_interest', 'establishment']

2.5 Extraction of company-level metadata from private sector sources or public websites

As proof of concept, the extracted company-level metadata from a diverse set of sources -- ranging from Google search results to social media data -- to concretely illustrate the kind of data we can extract per source type.

What was not explored in this proof of concept is *indirect data source discovery* by leveraging existing knowledge graphs such as Wikipedia which link related entities and individuals with one another by design. For example, we can use Wikipedia pages (or LinkedIn) specific to an entrepreneur and use the links in each page to identify firms related to this entrepreneur (e.g., he/she may be a founder for Company A, employee for Company B, co-founder with Individual C, etc.). Leveraging these sites allow us to easily connect firms and individuals in the entrepreneurship ecosystem. We suggest exploring this idea for future proof of concepts.

Here are the results of the sample web scraping for company-level metadata. For this analysis, the team focused on accelerators/incubators in Indonesia identified in the previous section, specifically "**GnB Accelerator**".

2.5.1 Google Places API

Data source	Google Places API, "Place Details" endpoint (API documentation here: https://developers.google.com/places/web-service/details)
Implementation & Results	We were able to extract metadata on the 3 accelerators/hubs located in Indonesia, by inputting their unique place_id which was extracted through Google Places API (see above section)
Relevant extracted fields	This provides more details compared to the Google Place API endpoints used in extracting company listings above. Additional fields include: contact details / phone number, website link, and more details in Google user reviews
Strategic use of scraped data	Getting company contact details will help when contacting the company for an interview or FGD. Having the startup website allows us to build another scraping layer by extracting data (e.g., text, contact details, images, etc.) from the startups' respective websites.

Here is what the sample output looks like, filtered down to one row and a few columns of the original dataset since the original dataset has many columns. For the full output, please see `2018-07-10 - Indonesia Google Places API - Place Details endpoint.csv`.

Entity Name	GnB Accelerator
Address	Metropolitan Tower, Jl. R. A. Kartini, Kav. 14, RT.10/RW.4, Cilandak Bar., Cilandak, Kota Jakarta Selatan, Daerah Khusus Ibukota Jakarta 12310, Indonesia
Geocoordinates	{'location': {'lat': -6.292881999999999, 'lng': 106.784808}, 'viewport': {'northeast': {'lat': -6.291533019708496, 'lng': 106.7861569802915}, 'southwest': {'lat': -6.294230980291501, 'lng': 106.7834590197085}}}
Google Place ID	ChIJPZLvV9rxaS4R6uvDTPMej_Y
Google Places Reviews	[{'author_name': 'Yeli Risna', 'author_url': 'https://www.google.com/maps/contrib/102120318279663282381/reviews', 'profile_photo_url': 'https://lh4.googleusercontent.com/-Ynu7tWTlj3I/AAAAAAAAAAI/AAAAAAAAAAA/AAnnY7ogOyKkvKtEuI3N3wuvuOChCuI-yg/s128-c0x00000000-cc-rp-mo/photo.jpg', 'rating': 2, 'relative_time_description': '7 months ago', 'text': '', 'time': 1512631153}]
Entity Types	['point_of_interest', 'establishment']
Website	https://gnb.ac/

2.5.2 Google Search

Data source	Google Custom Search API (API documentation here: https://developers.google.com/custom-search/json-api/v1/using_rest)
Implementation & Results	We were able to extract detailed data on the top 10 Google search results for "GnB Accelerator" and "MAD Incubator", respectively, with the results restricted to the location of Indonesia for localized search results.
Relevant extracted fields	Google search result title, text snippet, link, and rich snippet information such as sublinks and images (see here for more details: https://developers.google.com/custom-search/docs/snippets). The number of Google search results for that keyword is also returned.
Strategic use of scraped data	This easily gives us additional data sources specific to the company by feeding the Google search result links into the extraction pipeline. The number of Google search results for that keyword can also serve as a proxy indicator for company digital presence.

Here is what the sample output looks like, filtered down to a few columns of the original dataset since the original dataset has many columns. For the full output, please see `2018-07-10 - Indonesia Mad Incubator Google Search API.csv` and `2018-07-10 - Indonesia GnB Accelerator Google Search API.csv`.

Search Snippet Title	Search Result Snippet Text	Search Result Link URL
GnB Accelerator â€" Local Identity, Global Opportunity	innovative technology companies. GnB is a collaborative program between Japanese IT company Infocom Corporation and Fenox Venture Capital from SiliconÂ ...	https://gnb.ac/
6 Startup Indonesia di GnB Accelerator Batch Ketiga 2017	6 Sep 2017 ... Untuk penyelenggaraan kali ini, GnB Accelerator gaet 6 startup dari berbagai latar belakang, masing-masing dengan keunikan model bisnisÂ ...	https://id.techinasia.com/6-startup-di-gnb-accelerator-batch-3
GnB Accelerator Batch Ketiga Umumkan Enam Startup Terpilih ...	5 Sep 2017 ... Program GnB Accelerator mengumumkan enam startup terpilih menjadi peserta batch ketiga dan berhak mengikuti program selama tiga bulanÂ ...	https://dailysocial.id/post/gnb-accelerator-batch-ketiga-umumkan-enam-startup-terpilih
GnB Accelerator - Home | Facebook	GnB Accelerator, South Jakarta. 1738 likes Â· 15 talking about this Â· 30 were here. We're a startup accelerator in Jakarta, Indonesia. We offer...	https://www.facebook.com/gnbaccelerator/

GnB Accelerator | LinkedIn	Learn about working at GnB Accelerator. Join LinkedIn today for free. See who you know at GnB Accelerator, leverage your professional network, and get hired.	https://www.linkedin.com/company/gnb-accelerator

2.5.3 Social Media - Twitter

Data source	Twitter API (API documentation here: https://developer.twitter.com/en/docs)
Implementation & Results	We were able to extract detailed Twitter status data on the company "GnB Accelerator" using: keyword search for the phrase "GnB Accelerator" which pulls relevant tweets from the past 7 days containing that keyword direct pull of tweets from GnB Accelerator's public user timeline
Relevant extracted fields	We are able to extract data on the user who posted the status, as well as status-level fields such as created time, text, location, interactions (replies to user / statuses), retweets, URLs, user mentions, hashtags used, place/coordinates used, contributors.
Strategic use of scraped data	Twitter data, and social media data in general, is ripe for natural language processing analysis such as sentiment analysis and topic modelling. This allows us to extract online intelligence not commonly found in traditional datasets.

Here is what the sample output looks like, filtered down to one row and a few columns of the original dataset since the original dataset has many columns. For the full output, please see `2018-07-10 - Indonesia GnB Accelerator Twitter API.csv`

Created timestamp	Tue Aug 23 08:26:10 +0000 2016
Tweet Text	RT @VCInsiderNews: Why Japanâ€™s Leading IT Firm Decides to Invest in Indonesia https://t.co/FkC5qlTQNB @GnBAccelerator #startup https://t.coâ€¦
Was retweeted by GnB's followers?	FALSE
Source URL	Twitter Web Client
User	{"created_at": "Fri Mar 04 07:42:38 +0000 2016", "favourites_count": 1, "followers_count": 46, "friends_count": 115, "id": 705659703041208321, "id_str": "705659703041208321", "lang": "en", "listed_count": 1, "location": "Jakarta Capital Region", "name": "GnBAccelerator", "profile_background_color": "000000", "profile_background_image_url": "http://abs.twimg.com/images/themes/theme1/bg.png", "profile_background_image_url_https": "https://abs.twimg.com/images/themes/theme1/bg.png", "profile_banner_url":

	"https://pbs.twimg.com/profile_banners/705659703041208321/1475842390", "profile_image_url": "http://pbs.twimg.com/profile_images/784366082509254660/zWOhEKJg_normal.jpg", "profile_image_url_https": "https://pbs.twimg.com/profile_images/784366082509254660/zWOhEKJg_normal.jpg", "profile_link_color": "7FDBB6", "profile_sidebar_border_color": "000000", "profile_sidebar_fill_color": "000000", "profile_text_color": "000000", "screen_name": "GnBAccelerator", "statuses_count": 5, "url": "https://t.co/uiBJe1D2V7"}
URLs linked in Tweet body	[URL(URL=https://t.co/FkC5qlTQNB, ExpandedURL=http://goo.gl/mMolHm)]
Who are the users mentioned?	[User(ID=732181521235247105, ScreenName=vcinsidernews), User(ID=705659703041208321, ScreenName=GnBAccelerator)]
What were the hashtags used?	[Hashtag(Text='startup')]
Retweeted history	{"created_at": "Mon Aug 22 05:03:51 +0000 2016", "favorite_count": 1, "hashtags": [{"text": "startup"}], "id": 767588069935505408, "id_str": "767588069935505408", "lang": "en", "media": [{"display_url": "pic.twitter.com/aSvbDRCrPQ", "expanded_url": "https://twitter.com/VCInsiderNews/status/767588069935505408/photo/1", "id": 767587950859137026, "media_url": "http://pbs.twimg.com/media/CqcFLKjUMAIftJu.jpg", "media_url_https": "https://pbs.twimg.com/media/CqcFLKjUMAIftJu.jpg", "sizes": {"large": {"h": 1536, "resize": "fit", "w": 2048}, "medium": {"h": 900, "resize": "fit", "w": 1200}, "small": {"h": 510, "resize": "fit", "w": 680}, "thumb": {"h": 150, "resize": "crop", "w": 150}}, "type": "photo", "url": "https://t.co/aSvbDRCrPQ"}], "retweet_count": 1, "source": "Twitter Web Client", "text": "Why Japan\u2019s Leading IT Firm Decides to Invest in Indonesia https://t.co/FkC5qlTQNB @GnBAccelerator #startup https://t.co/aSvbDRCrPQ", "urls": [{"expanded_url": "http://goo.gl/mMolHm", "url": "https://t.co/FkC5qlTQNB"}], "user": {"created_at": "Mon May 16 12:10:52 +0000 2016", "description": "We are an online magazine featuring in-depth stories of today\u2019s investors & entrepreneurs.", "favourites_count": 9, "followers_count": 370, "friends_count": 180, "geo_enabled": true, "id": 732181521235247105, "id_str": "732181521235247105", "lang": "en", "listed_count": 44, "location": "Kuala Lumpur City", "name": "VC Insider News", "profile_background_color": "000000", "profile_background_image_url": "http://abs.twimg.com/images/themes/theme1/bg.png", "profile_background_image_url_https": "https://abs.twimg.com/images/themes/theme1/bg.png", "profile_banner_url": "https://pbs.twimg.com/profile_banners/732181521235247105/1513671620", "profile_image_url": "http://pbs.twimg.com/profile_images/932920167624990720/0hkeqo0Q_normal.jpg", "profile_image_url_https": "https://pbs.twimg.com/profile_images/932920167624990720/0hkeqo0Q_normal.jpg", "profile_link_color": "000000", "profile_sidebar_border_color": "000000", "profile_sidebar_fill_color": "000000", "profile_text_color": "000000", "screen_name": "vcinsidernews", "statuses_count": 245, "url": "https://t.co/pTErhHsfzy"}, "user_mentions": [{"id": 705659703041208321, "id_str": "705659703041208321", "name": "GnBAccelerator", "screen_name": "GnBAccelerator"}]}

2.5.4 Company website page source and visible text

Data source	Company's own website page source, visible text, and language
Implementation & Results	We were able to pull the entire page source of GnB Accelerator's website using Python scripts. We also extracted the visible text of each Indonesian website as a pre-processing step for Natural Language Processing (NLP) and other machine

	learning techniques. To do this, we removed all HTML tags and trailing/internal whitespaces. Last, we also identified the language used for each website. This will help us adjust the natural language processing techniques used on the text to account for language differences.
Strategic use of scraped data	Pulling the websites' entire page source allows us to extract all relevant text and links within the page in an automated fashion. Website data is ripe for natural language processing analysis such as Named entity recognition (NER) to get lists of other company names, startups, partners associated with each company, as well as topic modelling to identify topics/themes associated with each company. This allows us to extract online intelligence not commonly found in traditional datasets.

Here is what the sample text snippet looks like from the visible text of the company website of GnB accelerator. For the full output, please see `2018-07-10 - Indonesia GnB Accelerator Website data.html` or `2018-07-12 - Indonesia Companies Website data - Visible Text.csv`.

> *Home Startups Whatâ€™s On Contact Us Apply Accelerator First global acccelerator in Indonesia dedicated to progress and innovation that brings together the people, the funding, and the partners that drive business velocity We invest in talented and passionate early stage startups of all backgrounds, helping them to create innovative technology companies. GnB is a collaborative program between Japanese IT company Infocom Corporation and Fenox Venture Capital from Silicon Valley. It is a global network dedicated to progress and innovation that brings together the people, the funding, and the partners that drive business velocity. World Experts Join Forces INFOCOM CORPORATION Infocom Corporation, a subsidiary of Teijin, is a leader in IT systems and operation management services that provide diverse IT solutions and healthcare IT for dozens of pharmaceutical companies and thousands of hospitals. …*

3 Data analysis and Machine learning

As described above, alternative data tends to be larger and more heterogenous than data available through typical official statistical channels or gathered through surveys or fieldwork. A different set of analytical tools has thus been recently developed to derive value from such datasets.

As proof of concept, the team focused on machine learning techniques and tools to demonstrate how to concretely derive intelligence and insights relevant to digital entrepreneurship from various types of scraped data. Note that for this proof of concept, the team implemented the following:

- Implemented a pre-trained model, such as Named Entity Recognition (see Section 3.1 below) and VADER (Valence Aware Dictionary for sEntiment Reasoning) for sentiment analysis (see Section 3.2 below), for the purposes of illustrating this technique;
- Built a model using the available data such as Latent Dirichlet allocation (LDA) for topic modelling (see Section 3.3 below), and then applied the generative LDA model for thematic classification (see Section 3.4 below); or
- Visualized the collected data through network visualization (see Section 3.5 below).

The use cases below demonstrate how alternative data can complement standard data sources, if used carefully and in the appropriate context.

Table 4. Summary of Machine Learning Use cases and Potential value-add for this report

Use case: Machine Learning technique or tool	*Potential value-add*
Section 3.1 Named Entity Recognition and Classification for Entity Extraction. • Use case: Extract related companies/entities per accelerator from their website data • Data source: Raw website page source (from respective company websites)	**Productivity and speed gains.** Named Entity Recognition (NER) can be used to extract relevant entities from website data, which leads to productivity and speed gains when parsing through large chunks of text for relevant data. Standard data sources can then be used to check the quality of the data extracted.
Section 3.2 Sentiment Analysis on social media data. • Use case: Determine polarity (positive/negative/neutral) of tweets related to each company • Data source: Social media data (Twitter)	**New metrics.** A potentially useful new metric is general sentiment or "pulse" regarding a certain topic or entity, which we can derive using sentiment analysis to determine the polarity (i.e., positive/negative/neutral) of a given text.

Use case: Machine Learning technique or tool	*Potential value-add*
	We can then check if this new metric strongly correlates with any of the existing standard metrics, and derive insights from patterns uncovered.
Section 3.3 Topic modelling. • Use case: Extract general topics for startups in Nigeria **Section 3.4 Thematic classification.** • Use case: Group companies into clusters based on their topic association scores Data source: Extracted company metadata from an online data collector / directory	**Knowledge discovery and compact representation.** We can use topic modelling to automatically extract topics (represented through relevant word clusters) from various texts. We can then group entities into clusters based on their topic association scores through thematic classification. Subject matter experts can then be tapped and consulted to verify if the resulting topics and entity clusters make sense.
Section 3.5 Network Visualization. • Use case: Map various entrepreneurship ecosystem actors with one another based on relationships or connections (e.g., investor-investee connection). • Data source: All data sources used above (including dummy data)	**New data.** By collecting relationship data between ecosystem actors (such as investor-investee relationships), we can leverage this new data to create network visualizations which allow us to map various entrepreneurship ecosystem actors with one another and look for patterns (e.g., how central an actor is, if there are clustering present in the ecosystem). We can then check if these patterns are aligned with our knowledge of the entrepreneurship ecosystem based on the standard DEED framework.

In the subsection below, we describe the methodology used, data source used, results of the analysis, possible next steps or improvements for each case, as well as the strategic benefit of implementing the chosen methodology.

The compilation of all sample outputs for this section can be accessed in a shared view-only Google Drive folder.[8]

3.1 Named Entity Recognition and Classification for Entity Extraction

Data source used. We used Python scripts to pull the homepage source from 3 websites. Note that the company name and website were extracted via text search on Google Place API, with search restricted to Indonesia.

- 'GnB Accelerator': 'https://gnb.ac/'

[8] The view-only Google Drive folder can be accessed here: https://drive.google.com/open?id=1YXA218oOwUvB65JBGlH_h4PcMMh9KIsZ

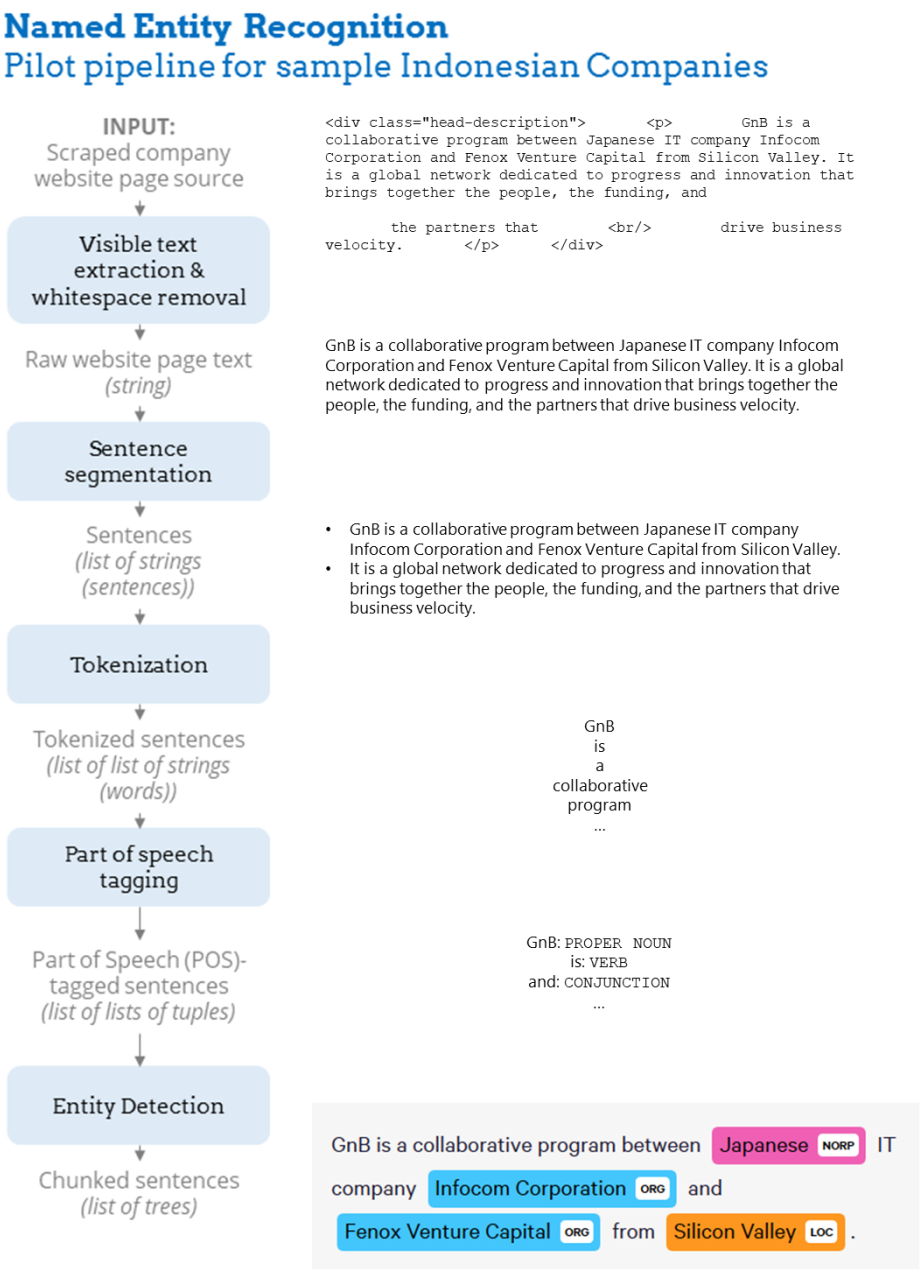
- 'Mad Incubator': 'http://www.incubator.com.my/'
- 'The Accelerator': 'http://www.accelerator.co.id/'

See ***Subsection 2.5.4 Company website page source and visible text*** above for a snippet of the scraped data.

3.1.1 Methodology

Here's a visual representation of the methodology used to extract entities (such as persons, organizations, locations) from websites' page source for the sample Indonesian companies:

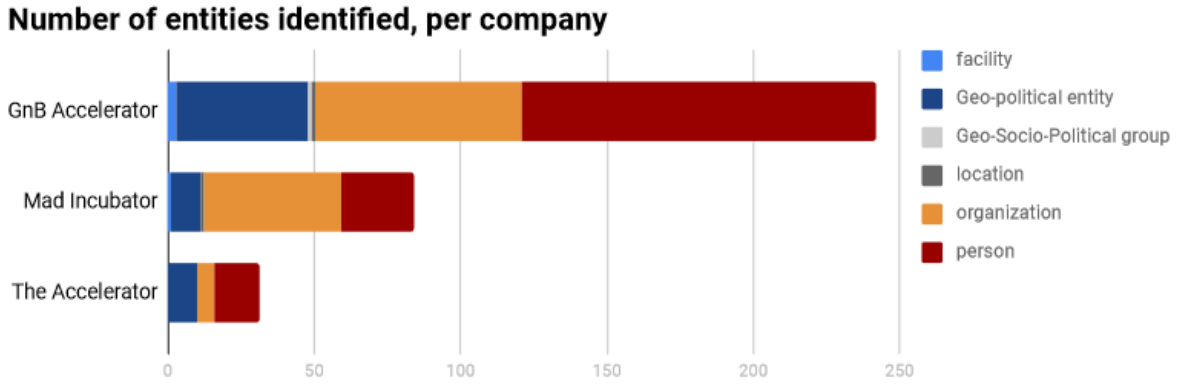
Figure 1. Pilot pipeline for Named Entity Recognition



3.1.2 Results

We extracted a total of 357 related entities from the website text of the 3 Indonesian companies. The distribution per entity type is as follows:

Figure 2. Number of entities identified per company using NER



For example, here are some examples found per entity type for the Indonesian company "GnB Accelerator":

Table 5. Example entities found on GnB Accelerator's website

Entity Type	*Examples for "GnB Accelerator"*
Facility	The Bridge, Wall Street, Y Combinator
Geo-political entity (GPE)	Indonesia, Japan, Asia, Jakarta, Singapore, China
Geo-Social-Political group (GSP)	US
Location	Southeast Asia
Organization	Fenox Venture Capital, Infocom
Person	Joshua Kevin, Adamas Belva Syah Devara CEO

Here's a snippet of the output dataframe from NER for GnB accelerator:

Table 6. Example NER output for GnB Accelerator

entity	label
Adamas Belva Syah Devara CEO	PERSON
Alfatih Timur CEO	PERSON
Bridestory	GPE
CEO Appsocially Willson Cuaca	ORGANIZATION
CEO Bridestory Katsuhiro Okamura	ORGANIZATION
CEO Bridestory Kevin Mintaraga	ORGANIZATION

CEO Fenox Venture Capital Anis Uzzaman	ORGANIZATION
CEO Fenox Venture Capital Kentaro Hashimoto	ORGANIZATION
CEO Intangible Communications Peter	ORGANIZATION
CEO Intangible Communications Toshihisa Wanami	ORGANIZATION

3.1.3 Next steps

Here are some immediate next steps (beyond the scope of this initial proof of concept) to improve model performance as well as the insights extracted:

- Further refine the model to remove false positives from identified entities by creating an ensemble model which combines the initial results (generated using NLTK[9]) with other NER open-source libraries such as StanfordNERTagger[10] and Polyglot[11], etc.
- Enable NER with multi-language support using polyglot (especially since not all websites are in English).
- Make semi-automated process to tag subtypes per entities and their relationship with the company (e.g., partner, mentor, founder, etc.).
- Mix NER with supervised/semi-supervised machine learning techniques, which could improve its outputs particularly for entrepreneurship-related text.

3.1.4 Strategic benefit of implementing this methodology

NER allows us to automate the extraction of persons, organizations, and other entities related to each company, ultimately allowing us to build a network or ecosystem of actors surrounding each company. These small networks - wherein one company is at its center - can then be merged to generate a bigger, area- or country-wide ecosystem mapping of entities.

This compiled network can then be used to:

- Augment the entity data collected via surveys;
- Implement more robust network analysis, since company nodes already have metadata extracted through web scraping (see Task B above). This additional metadata can be used for thematic clustering and other network analysis techniques to supplement the network analysis conducted through other assessments.

[9] NLTK means "Natural Language Toolkit", which is the dominant Python package used for natural language processing. For more details, please see https://www.nltk.org/

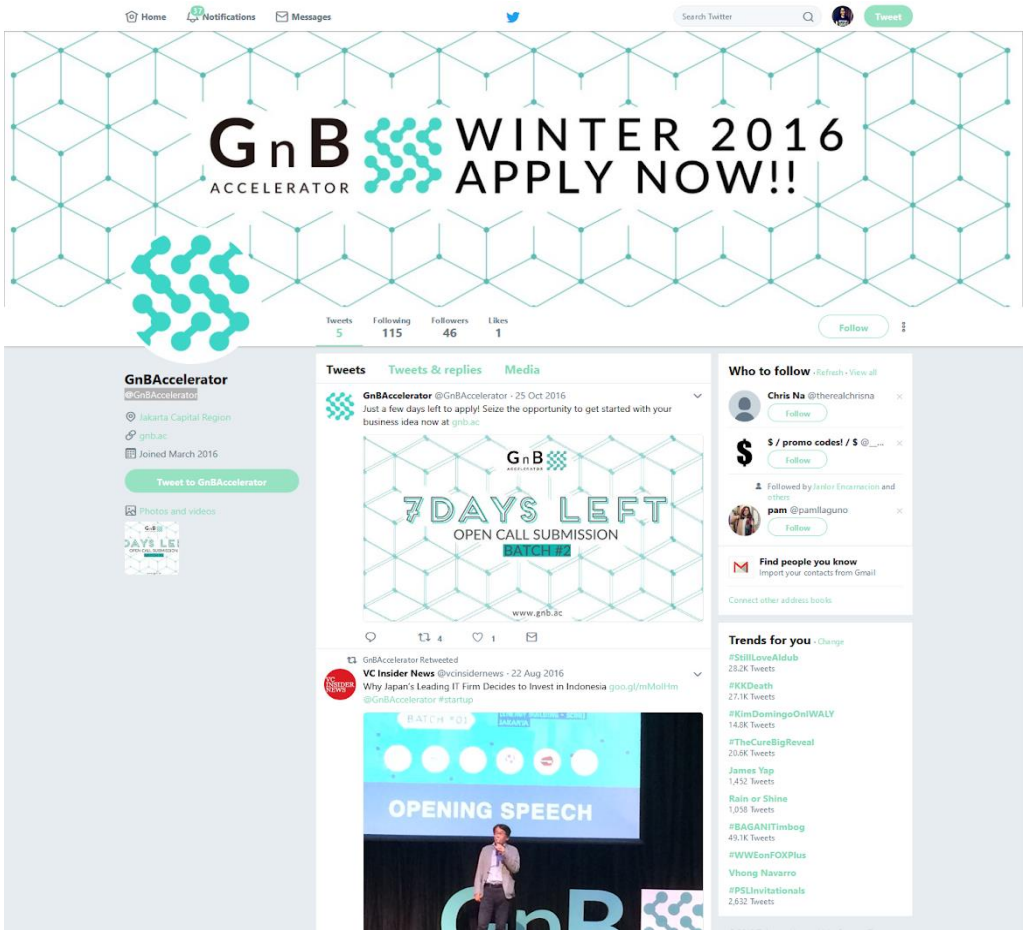
[10] https://nlp.stanford.edu/software/CRF-NER.html#Download

[11] http://polyglot.readthedocs.io/en/latest/Installation.html

3.2 Sentiment Analysis on social media data

Data source used. We used the Twitter API to pull detailed Twitter status data on the Indonesian company "GnB Accelerator" using:

- keyword search for the phrase "GnB Accelerator" which pulls relevant tweets from the past 7 days containing that keyword
- direct pull of tweets from GnB Accelerator's public user timeline with Twitter username @GnBAccelerator (https://twitter.com/GnBAccelerator). Here's how its Twitter public timeline looks like:



For a snippet of the scraped Twitter data, please see ***subsection 2.5.3 Social Media - Twitter*** above.

3.2.1 Methodology

We used the VADER (Valence Aware Dictionary for sEntiment Reasoning) model[12], which is a well-known and often-used model for implementing sentiment analysis on social media text created by C.J. Hutto and Eric Gilbert from Georgia Institute of Technology.

The VADER model takes sentences as an input, and outputs four sentiment metrics for each sentence. Let's take for example the sentence "The food is good and the atmosphere is nice."

Table 7. Sample output using VADER for Sentiment Analysis

VADER sentiment metric	*Definition*	*Score for example sentence*
Positive (pos)	Proportion of the sentence/text that falls under positive lexicon	45%
Neutral (neu)	Proportion of the sentence/text that falls under neutral lexicon	55%
Negative (neg)	Proportion of the sentence/text that falls under negative lexicon	0%
Compound	Sum of all lexicon ratings, standardized to range between -1 and 1	69%

The VADER model works well with social media text since it also considers slang or informal speech such as multiple punctuation marks, acronyms, emoticons, capitalization, and word context. Each word in the lexicon is assigned a sentiment rating such that positive words have a positive value, and negative words have a negative value. Note that "more positive" words have a higher rating, as seen when you compare "great" (3.1) to "good" (1.9).

To further illustrate the VADER model, here are examples of its usage which showcases sentences with slang words and emoticons:

Table 8. Sample output using VADER for Sentiment Analysis using slang words and emoticons

Example sentence	**Compound**	**Negative**	**Neutral**	**Positive**
:) and :D	79%	0%	12%	88%
<BLANKLINE>	0%	0%	0%	0%

[12] For the original research paper on the VADER model, please see http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf. For an easy-to-understand introduction to VADER, please see http://t-redactyl.io/blog/2017/04/using-vader-to-handle-sentiment-analysis-with-social-media-text.html.

Today sux	-36%	71%	29%	0%
Today kinda sux! But I'll get by, lol	22%	20%	53%	27%
Very bad movie.	-58%	66%	25%	0%
VERY BAD movie!	-76%	74%	27%	0%

3.2.2 Results

We were able to compute the VADER sentiment metrics of the 15 sample tweets related to GnB accelerator. Here's a snippet of the output dataframe:

Table 9. Sample output for VADER Sentiment Analysis on scraped data

Tweet text	Negative	Neutral	Positive	Compound
Applications for GnBAccelerator, SE Asiaâ€™s first multinational startup accelerator, are now available online at https://t.co/ZZABui0Dxq .	0%	100%	0%	0%
I'm giving out shout out to @ahlijasa as #StartupWorldCupChampion #INDONESIA regional finale!	0%	80%	20%	40%
Just a few days left to apply! Seize the opportunity to get started with your business idea now atâ€¦ https://t.co/G7QwHNCBi8	0%	85%	15%	48%
KILAS INFO @tabloidpulsa EDISI 391 | Huawei - 3 - GnB Accelerator - Asia IoT Bussines Platform | Cc.â€¦ https://t.co/TDta9NJrMq	0%	100%	0%	0%
RT @VCInsiderNews: Why Japanâ€™s Leading IT Firm Decides to Invest in Indonesia https://t.co/FkC5qlTQNB @GnBAccelerator #startup https://t.coâ€¦	0%	100%	0%	0%
RT @VentureShire: Why Japanâ€™s leading IT firm decides to invest in Indonesia @GnBAccelerator @FenoxVC #Infocom https://t.co/R3jt8s6ltx	0%	100%	0%	0%

In general, GnB Accelerator's tweets are mostly neutral or slightly positive. There are no negative tweets found in the sample generated. Its high score in neutrality makes sense since most of the sampled tweets are retweets of GnB Accelerator-related articles/posts by news agencies, which tend to have neutral-sounding headlines.

3.2.3 Next steps

Here are some immediate next steps (beyond the scope of this initial proof of concept) to improve model performance as well as the insights extracted:

- The current code only works with English language tweets. We need to implement sentiment analysis with multi-language support using polyglot or other solutions.

- Identify themes, entities, or keywords which generate high sentiment scores. To do this, we can segment the tweets into highly positive and highly negative tweets, and then identify the top keywords prominent in each segment.
- Possibly derive a company-level indicator for social media sentiment using the aggregated scores of tweets related to each company.
- Compare the various sentiment analysis methods and how well they perform on entrepreneurship-related text given the methods' pros and cons.

3.2.4 Strategic benefit of implementing this methodology

Evaluating the general sentiment of companies, entities, and topics on social media would be an interesting new dimension of analysis when it comes to digital entrepreneurship. This can also possibly lead to the development of new indicators which can augment some of the intangible DEED domains, such as Culture (e.g., attitudes).

3.3 Topic modelling

Data source used. We used the extracted data on 251 startups based in Nigeria from the website http://nigeria.startups-list.com. Specifically, we focused on analyzing the brief descriptions of all startups.

Topic modelling works best on a large set of same-language text data with multiple rows or entries. For simplicity, we chose the largest scraped English language dataset from the previous section, which is the Nigeria startups list dataset.

See `2018-07-10 - Nigeria Startups List.csv` for the scraped data on the Nigeria startups. For a snippet of the scraped data, please refer ***to subsection 2.4.1 A data collector or directory website*** above.

3.3.1 Methodology

For this note, we used one of the most oft-used models for topic modelling – the Latent Dirichlet Allocation (LDA) model[13]. LDA allows us to extract N topics from a set of documents, wherein each topic is defined by a set of keywords which are strongly associated with that topic.

Note that this method requires some interpretation on the part of the analyst with the help of subject matter experts, since the model requires N as input – that is, the analyst will be the one to set the number of topics (N) that the LDA model will look for. For the purposes of this pilot, we picked N = 3 by manually checking the diversity of the topics generated using N = 3, 4, 5. For this specific dataset, N = 3 seems to work the best.

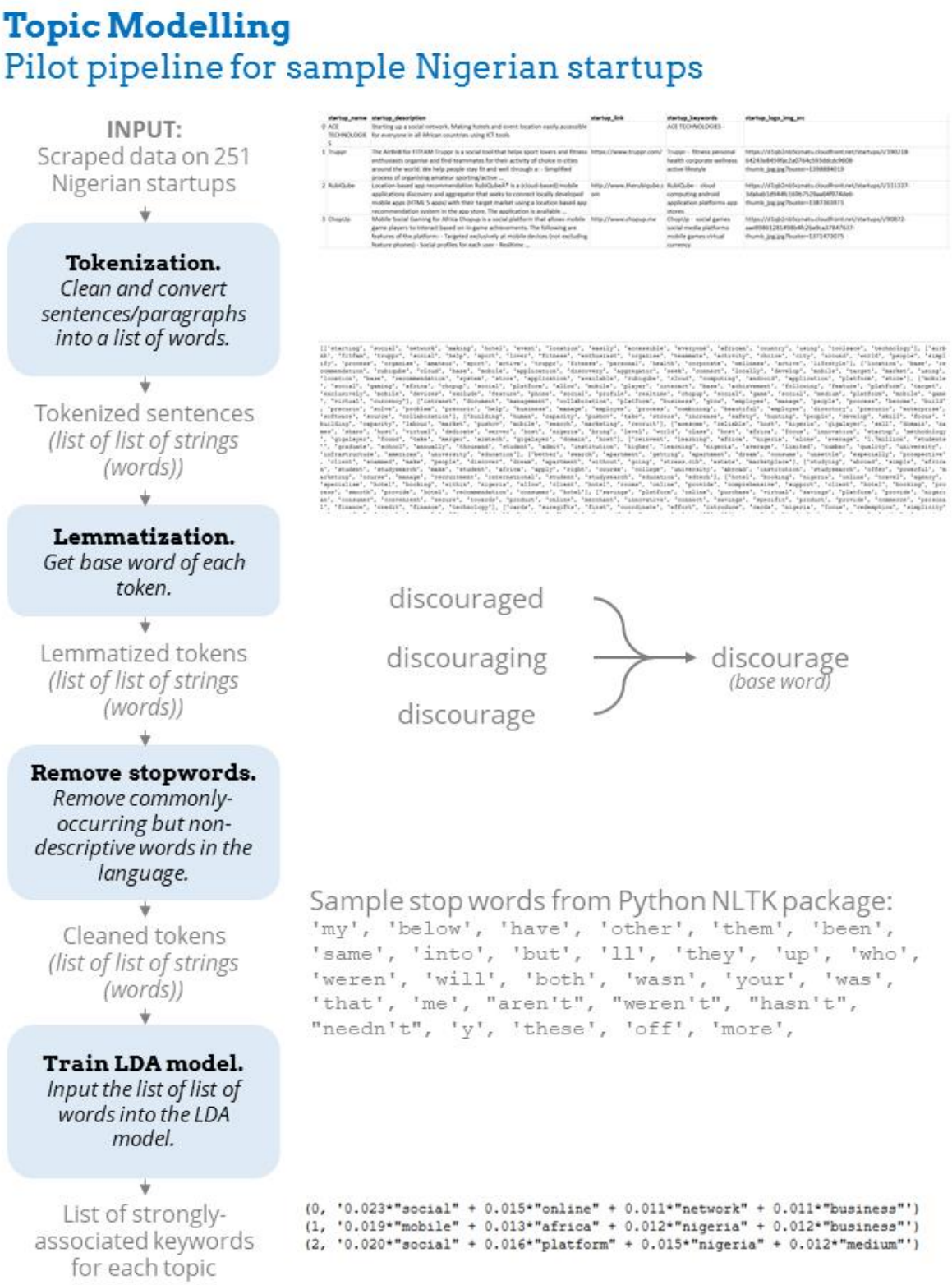
[13] For more information, please see https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation

The lambda (λ) parameter is important to tune when building the LDA model. When calculating the relevance or importance of a word in a topic, $0 \leq \lambda \leq 1$ can be interpreted as the reverse weight given to the overall frequency of a given word in the corpus. That is, if $\lambda = 1$, then we don't care about how rare the word is in the corpus. Alternatively, if $\lambda = 0$, the relevance of each word is inversely proportional to its overall frequency in the corpus.

Also, topic modelling requires a lot of heavy text pre-processing before the data can inputted to train the model, to account for multiple versions of the same word/idea (e.g., "discourage" vs "discouraging") and commonly-occurring but non-descriptive words in the language (e.g., "a, the, an" for English).

Here is a visual representation of the methodology implemented:

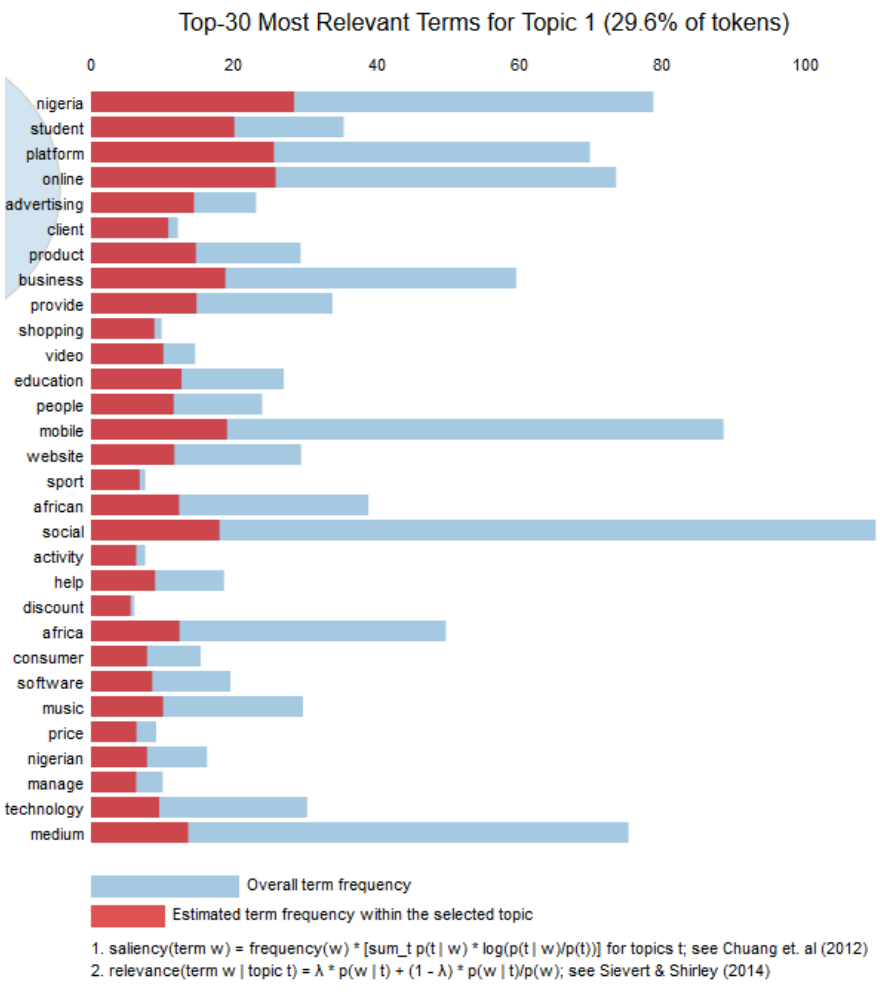
Figure 3. Pilot pipeline for Topic Modelling

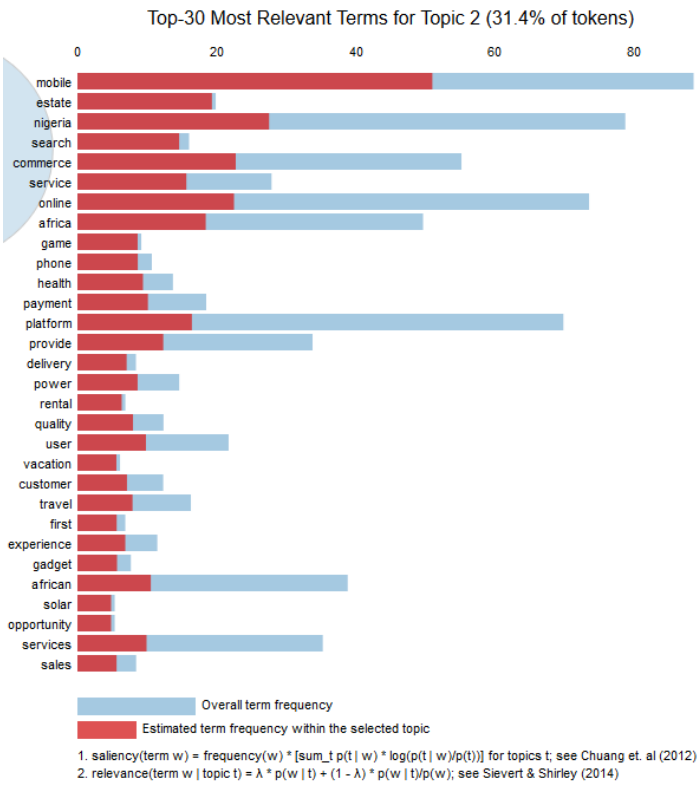
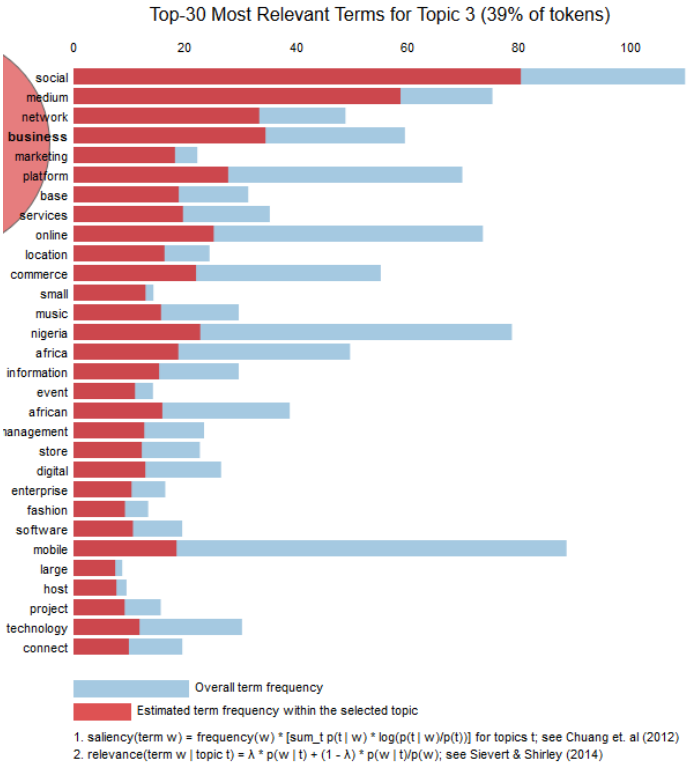


3.3.2 Results

We were able to extract 3 topics using the company descriptions of the 251 scraped Nigerian startups. The 3 identified topics are as follows, with their corresponding top 30 most relevant keywords. Note that the keywords were derived by setting$\lambda = 0.6$, as suggested by Sievert and Shirley in their paper "LDAvis: A method for visualizing and interpreting topics".

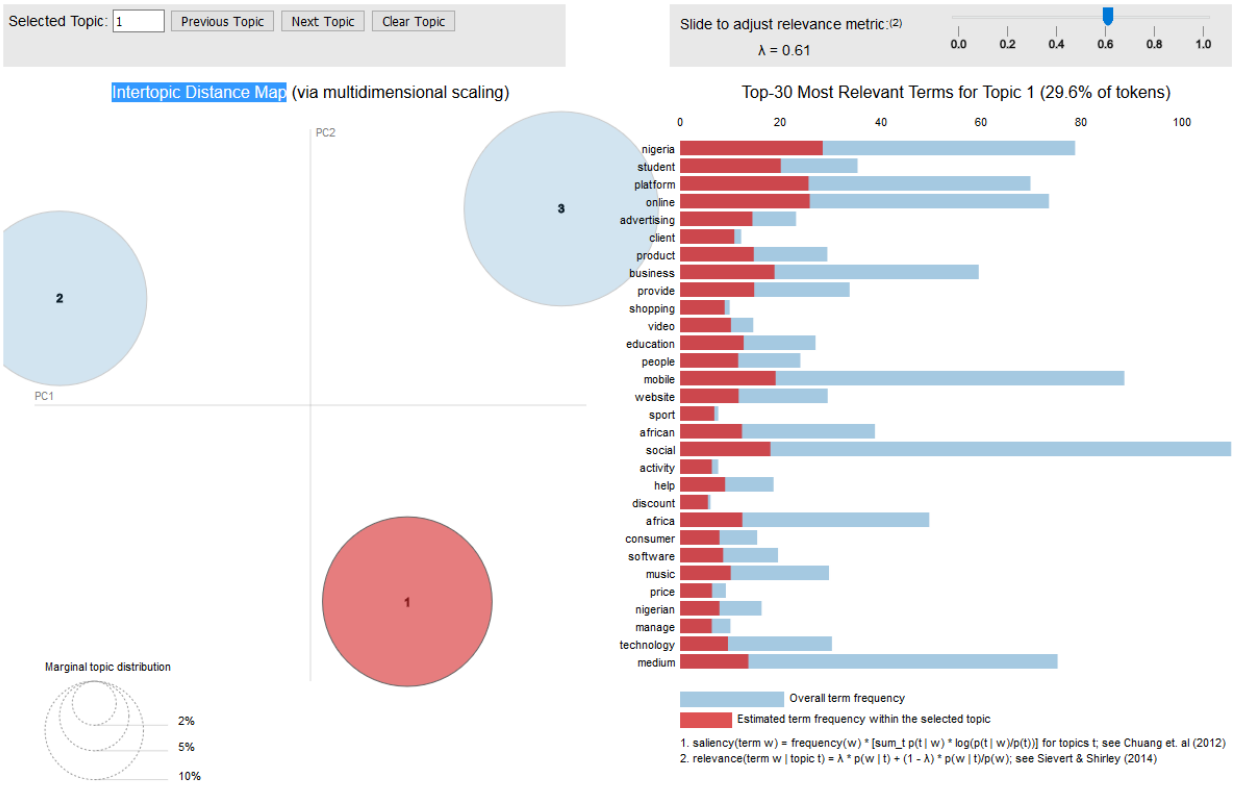
Table 10. Outputs for Topic Modelling on scraped data.

Topic	*Top 30 Most Relevant Keywords per topic (using* $\lambda = 0.6$*)*
Topic 1: Online student / education products	

Topic	*Top 30 Most Relevant Keywords per topic (using $\lambda = 0.6$)*
Topic 2: Mobile-based services for estate, commerce, search, etc.	
Topic 3: Social media network and marketing	

We also built an interactive tool[14] for interpreting the results of the trained LDA model on the Nigeria startups data (see Figure below). One of the interesting features of this tool is that it can visualize the size and possible overlaps among topics (notice the bubbles to the left of the figure below).

Figure 4. Screenshot of Interactive Tool for visualizing LDA results



3.3.3 Next steps

Here are some next steps (beyond the scope of this initial proof of concept) to improve model performance as well as the insights extracted:

- The current code only works with English language text. We need to implement topic modelling and keyword extraction with multi-language support using polyglot or other solutions.
- To get more robust results, we would need more text as inputs for the Latent Dirichlet Allocation (LDA) model. This can be achieved by pulling the website text for all startups in the list and then applying topic modelling to this bigger text base.

[14] For more details, refer to the paper by Sievert and Shirley entitled "LDAvis: A method for visualizing and interpreting topics".

- Experiment with changing the value of N (number of topics) as well as lambda (λ) for extracting top relevant keywords for each topic.

3.3.4 Strategic benefit of implementing this methodology

Using this methodology properly allows us to automatically and efficiently describe a vast set of text by grouping its elements into topics as well as extracting relevant keywords per topic. This can be easily extended to open-ended survey responses and other qualitative data, whose textual content is usually analyzed through manual methods.

Also, this methodology easily lends itself to thematic classification, by bucketing the inputted startup data into the topics extracted using this method.

3.4 Thematic classification

As a follow-up analysis, we used the same data source as the previous methodology (topic modelling) as well as the results of topic modelling.

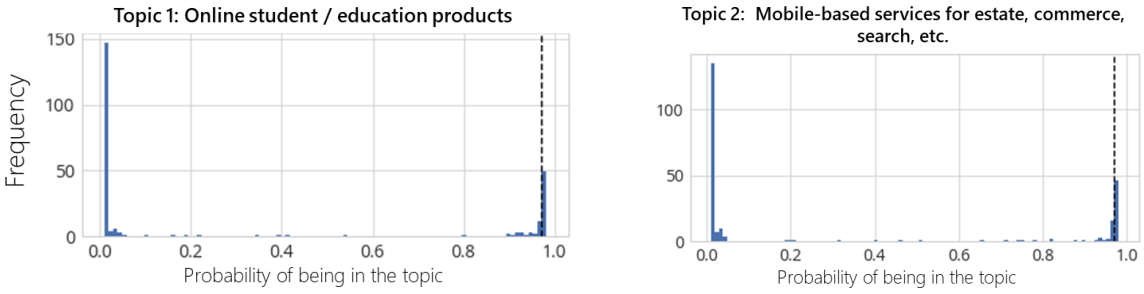
Specifically, we applied the trained Latent Dirichlet Allocation (LDA) model on the entire Nigerian startup list dataset, and derived scores for each startup as to which topic they are most closely assigned to.

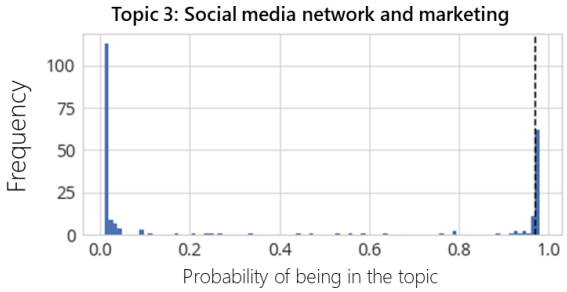
3.4.1 Methodology

We ran the trained LDA model (see previous methodology section) on each company description per Nigerian startup. Running the model outputs three scores for each startup – that is, one score for each topic – wherein the sum of all 3 scores is 1.00 per startup. In other words, it returns the probability distribution of a startup belonging to each of the topics.

We then classified the startups per topic by assigning them to the topic such that their score for that topic is beyond a certain threshold. For the purposes of this note, we used the threshold of 0.97 based on histograms of the topic scores (see plots below). Notice the clean cutoff for each topic histogram at approximately 0.97 (denoted by the black dotted line in the histograms below).

Figure 5. Probability Distribution of a startup belonging to each of the topics





3.4.2 Results

Using this method, we were able to classify the Nigerian startups into the three topics, with the following distribution as shown in the table below. Note that out of the 251 startups, 89 startups were not assigned to any cluster since their topic scores are all below our threshold of 0.97 – that is, their topic association wasn't high enough to merit being assigned to any topic.

Table 11. Examples of Entities per cluster derived from Thematic Classification on scraped data

Cluster	*No. of companies*	*Examples of companies per cluster*
Cluster 1: Online student / education products	53	**studyscholarshipng** **Supporting students into higher education** Non -merit studyscholarship website. Applicants purchase studyscholarship scratch cards to register online. Winners are randomly selected online from the various entries received in each category at the **UniSmart** **Gamified blogging for Nigerian students** UniSmart is a social community for Nigerian university students. Our platform allows student to share and discover interesting content (gist) from people within their university network and verified students can enjoy added perks.
Cluster 2: Mobile-based services for estate, commerce, search, etc.	46	**Abuja Health Pages** **Organizing Nigerian healthcare in the cloud** Our online health portal is for patients accessing healthcare in Nigeria. It catalogs all healthcare services to help patients select the most appropriate location to present for healthcare. It also **Maliyo Games** **We share the experiences of everyday African through games, published for web and mobile** Maliyo Games developed casual online and mobile games based on localised narratives, characters, environment and sound. Our concepts originate by
Cluster 3: Social media network and marketing	60	**Qeeub** **Quick personalized sharing with your clique across the websphere via SMS.** Qeeub Inc is a start company bringing together organization, friends,family and group on a platform making information sharing quick and easier via SMS. Our consumers are **Contactly** **Dropbox for contacts** Contactly is a re-definition of your address book. It's a mobile application that allows you to share your contact details securely with people you meet. It works by creating a profile giving your a personalized ID (Contactly ID) that links all your contact

The startup names, descriptions, and respective topic-level scores can all be seen in the file `2018-07-12 - Nigeria Startups - Thematic classification.csv`. Here's a snippet of the output:

Table 12. Output for Thematic Classification for scraped data

Startup Name	Description	Assigned Topic (based on threshold of 97%)	Probability of being in...		
			Topic 1	Topic 2	Topic 3
Friendite	African Dating Site FrienDite.com helps African connect with loved ones, helps you mingle, find your soul mate and fall in love easily. We help Africans improve a better marriage and a better love connection.Friendite - social media online dating social network media match making	Topic 3	1%	1%	97%
Estatenode	Search for Real-estate listings around you We provide a more convenient and effective way for property seekers to discover their desired property through the up-to-date property information available on our database, available for free, accessible 24 hours a day to anyone with web access and far more complete ...Estatenode - mobile real estate	Topic 2	1%	97%	1%
Educandlab	Learn | Learn smarter | Learn better Educandlab provides access to education with personalized experience based on the future ambition of our student through : 1)a video lecture platform, 2)simulation of key concepts in a field and 3) easy access to books.Educandlab - education edutainment k 12 education	Topic 1	97%	1%	1%

3.4.3 Next steps

Here are some next steps (beyond the scope of this initial proof of concept) to improve model performance as well as the insights extracted:

- The current code only works with English language text. We need to implement topic modelling and keyword extraction with multi-language support using polyglot or other solutions.
- Confirm or improve on topic modelling and thematic classification model accuracy by comparing the results against a manually-generated clustering/grouping of the same set of startups.

3.4.4 Strategic benefit of implementing this methodology

We can use this methodology to easily find thematic groupings among entities with lots of textual data associated with them. Keeping the trained model can allow us to have comparative data for longitudinal surveys. This is helpful for future survey assessments which may include new, unseen startups, as reusing the trained model allows us to classify these new startups to the topics extracted in the baseline survey.

3.5 Network Visualization

3.5.1 Methodology

We scraped online data on startups, investors, incubators, accelerators, associations, and mentors in Indonesia and generated an interactive network visualization from this. When interpreting these visualizations, please consider the following caveats:

- Only scraped online data was used to generate these visualizations for simplicity. Other data sources (e.g., survey data, proprietary data, official data) were not used for these visualizations (beyond the scope of this initial proof of concept).
- Due to time limitations, we were not able to check the online scraped data for bias, nor were we able to recalibrate the data to reflect more accurate estimates (beyond the scope of this initial proof of concept).
- For illustrative purposes, we filled in a few columns with dummy information to generate the visualizations. Some examples of our use of dummy data include initializing reasonable[15] random[16] values for variables with missing data, including:
    - org/company size (bubble size for the network diagrams)
    - investment amount (connection line width for the network diagrams)
    - geocoordinates (for the map), etc.

Here are some biases to consider when interpreting the following charts:

- Bias for sources for inferable relationship data - e.g., program/accelerator pages which explicitly state "mentor" and "members", directories which connect investors to startups
- Bias for investor data linked to startups - Investor data primarily seeded from startup data

[15] Note that what we mean by "reasonable" differs for each context (e.g., for geocoordinates, these should be found on the Indonesia land mass).

[16] We deliberately chose to generate random dummy data to fill in missing data (rather than imputation) with the purpose of quickly showing how the network could possibly look like given differing values across different ecosystem actors.

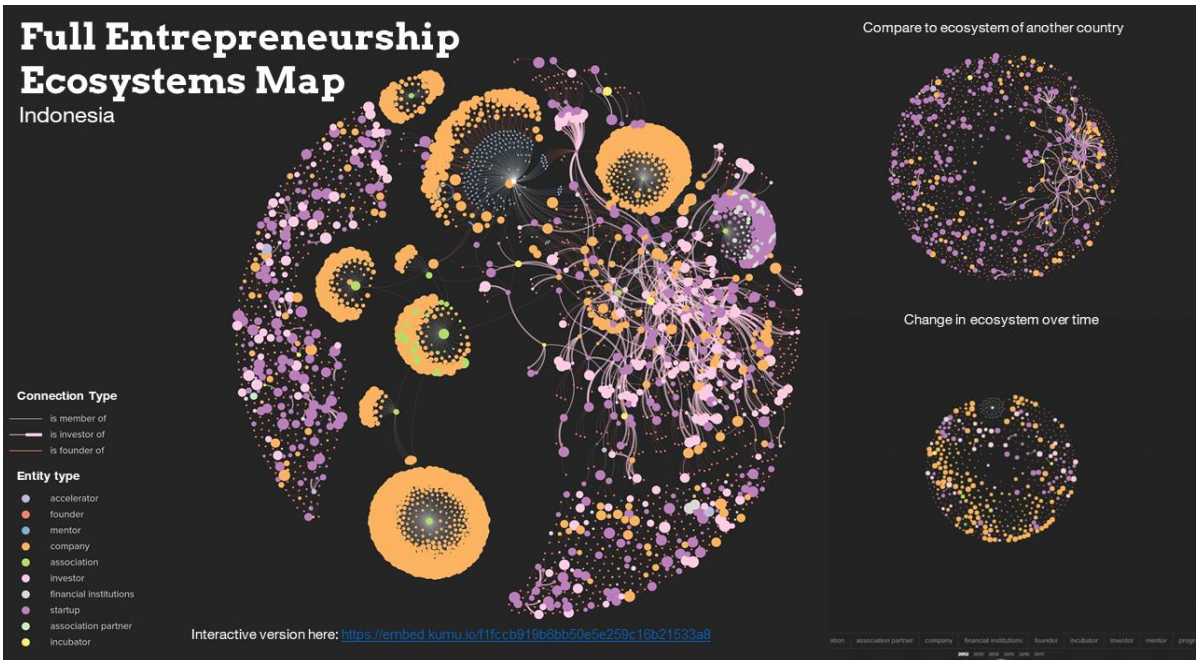
- Disconnected circles - we expect to add more connections as we scrape secondary sources of data (company websites, articles, social media accounts)
- Deduplication not yet done (beyond the scope of this initial proof of concept) so there's a small redundant set here (but not so much – the majority are well known startups/companies)

3.5.2 Results

Here are the resulting ecosystem network visualizations based on the scraped data and dummy data as described above. We were able to generate three network visualizations:

- Full Entrepreneurship Ecosystem
- Zooming in - Investment Flow Ecosystem Map
- Geographic ecosystems map

Figure 6. Network Visualization for Indonesia: Full Entrepreneurship Ecosystem[17]



Here are a few observations we can make from the sample visualization (again, filled in by dummy data), guided by the DEED framework for digital entrepreneurship:

[17] The interactive version can be viewed here: https://embed.kumu.io/f1fccb919b6bb50e5e259c16b21533a8.

Figure 7. Observations based on the Network Visualization of Indonesia

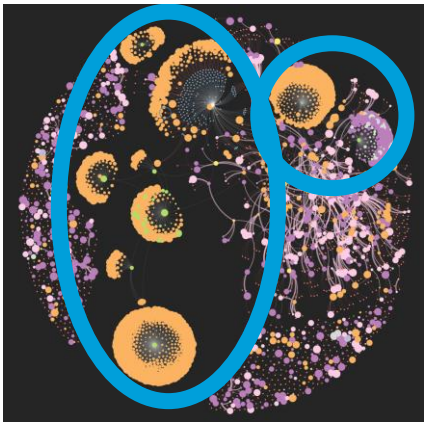
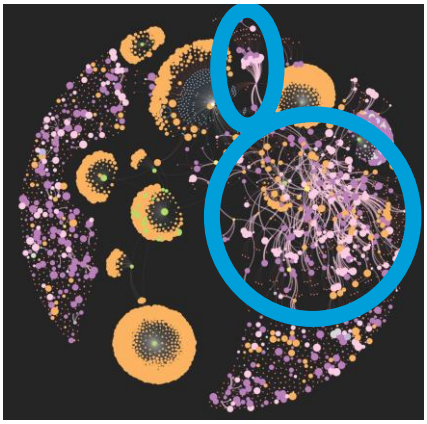
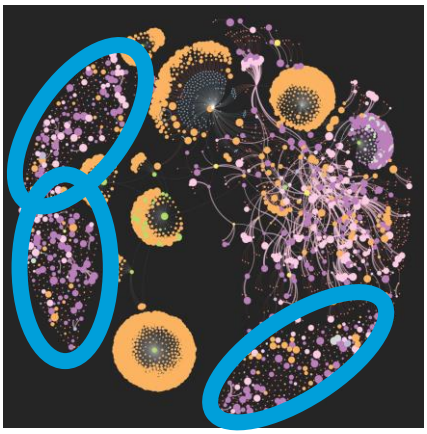
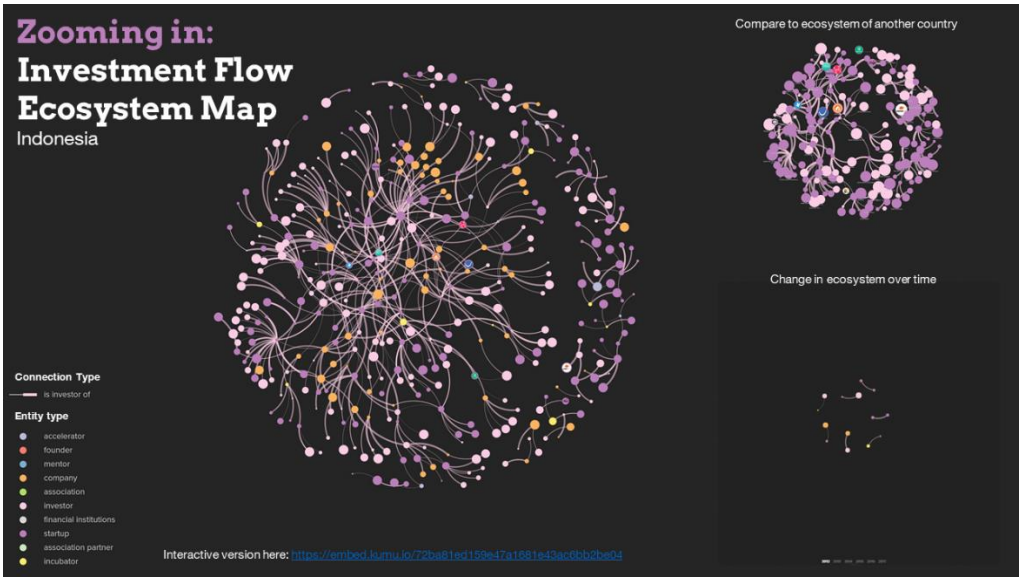
Visualization	Observation
	**Some Clustering present.** There is some clustering visually present in the Indonesia ecosystem, primarily composed of association members or startup program participants. The rest of the ecosystem does not display much clustering.
	**Sub-ecosystem with low Accumulation/Allocation Barriers.** This sub-ecosystem with relatively higher density and degree of connections is composed of investors, incubators, accelerators, and firms. Notice that these are primarily the key institutions for access to finance and social capital.
	**Sub-ecosystem with high Accumulation/Allocation Barriers.** The sub-ecosystem with relatively lower density and degree of connections is composed of unconnected firms. This implies that these firms are not part of an association, nor do they have investors, accelerators, or incubators who are mentoring or investing in them.

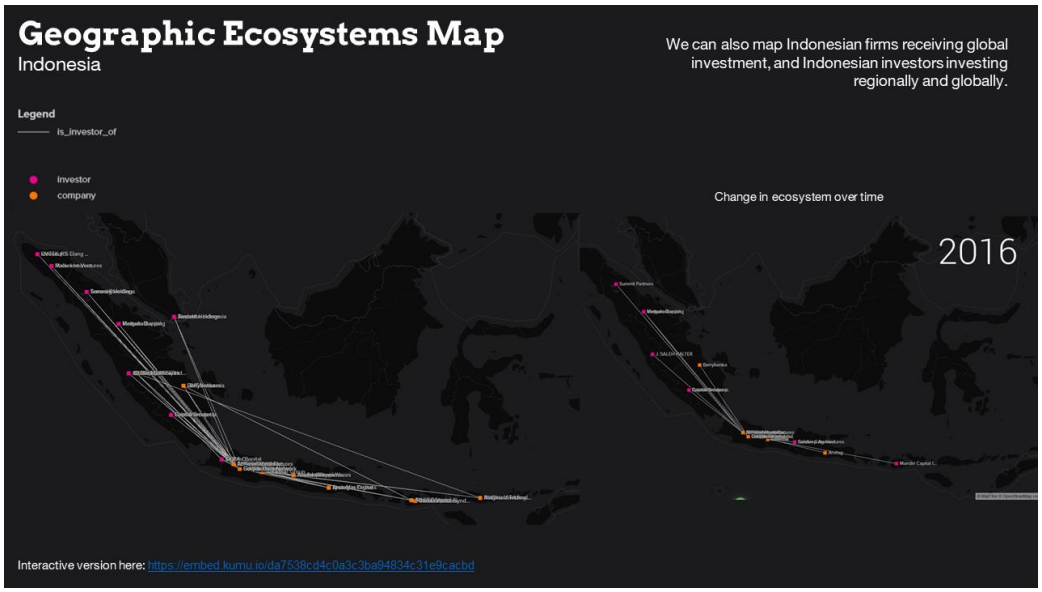
Figure 8. Network Visualization for Indonesia: Investment Flow Ecosystem Map[18]



A few interesting things to notice in the diagram above:

- Some startups/companies tend to attract more investors (spaghetti mass in middle).
- At the fringes, there are some startups/companies with relatively few investors connected.

Figure 9. Network Visualization for Indonesia: Geographic Ecosystems Map[19]



[18] You can view the interactive version here: https://embed.kumu.io/72ba81ed159e47a1681e43ac6bb2be04

[19] Interactive version here: https://embed.kumu.io/da7538cd4c0a3c3ba94834c31e9cacbd. Caveat: Only 100 of companies/investor data is here due to rendering difficulties.

3.5.3 Next Steps

Here are some next steps (beyond the scope of this initial proof of concept) to improve on the insights extracted:

- Clustering of actor types in a region
- Investment flow patterns (line thickness) specific to locations
- Change in clustering and flow patterns over time
- Indonesian firm vis-a-vis global market (foreign investors to Indonesian firms, or Indonesia firms with international market)
- Easily compare ecosystems with other countries (for instance, we can make "ecosystem typologies" based on patterns across different countries)
- See change in investment flow and ecosystem over time

It is also important to ensure that our network measures and analysis are robust and sensitive to missing data by selecting appropriate centrality measures based on suggestions from existing research in this area. For instance, there have been empirical tests which look at the correlation between calculated centrality measures and the actual centrality measures by simulating missing data. They have found that there are some centrality measures – such as in-degree centrality and simple eigenvector centrality – whose resulting measures are relatively stable despite having a low sampling level such as 50% missing data (Costenbader & Valente 2003).

4 Data quality and mitigation

The quality of alternative data is a source of significant concern for researchers. Almost none of this data is gathered for research purposes, is representative in any way, follows any international standards, is consistent with other online sources, or provides any assurance about quality. It is thus critical to apply a high bar for quality when using such resources.

The following section describes suggested mitigation approaches to:

- record linkage between the diverse sets of sources,
- manage data and its storage, and
- adjust for biases and test the quality of alternative data sources.

4.1 Recording Linkage between Data Sources

One of the key challenges in combining diverse sets of traditional and alternative data sources is the problem of record linkage, which has two main subproblems: (a) maintaining a data structure which can accept different data from different sources and (b) matching records about the same actor from different data sources.

Maintaining a data structure which can accept different data from different sources. This is a common problem and is typically addressed by using a NoSQL Database, which is a flexible data structure which can accept generally any document structure (compared to the frequently-used SQL database which requires a certain structure before accepting data).

Matching records about the same actor from different data sources. To do this, we can implement several techniques and checks, such as:

- Data processing involving fuzzy matching, which allows us to approximately detect matches across records from different data sources;
- Triangulation of indicator data collected across different data sources;
- Handling data discrepancies through a combination of semi-automated checks guided by internally-defined criteria (such as source reliability as defined by subject matter experts, recency of data collected, and frequency of value among all sources considered) and manually checking a random sample of the records to ensure proper handling of edge cases; and
- Exploration of probabilistic models for record linkage (such as fastLink) which allows a mixed approach where the user provides input to update the model.

To illustrate the record linkage process, see the schematic below based on earlier research by Ansolabehere and Eitan Hersh (2012) which also combined traditional and alternative data sources (the figure is from Salganik 2017).

Figure 10. Record Linkage for Traditional and Alternative data sources (Source: Salganik 2017)

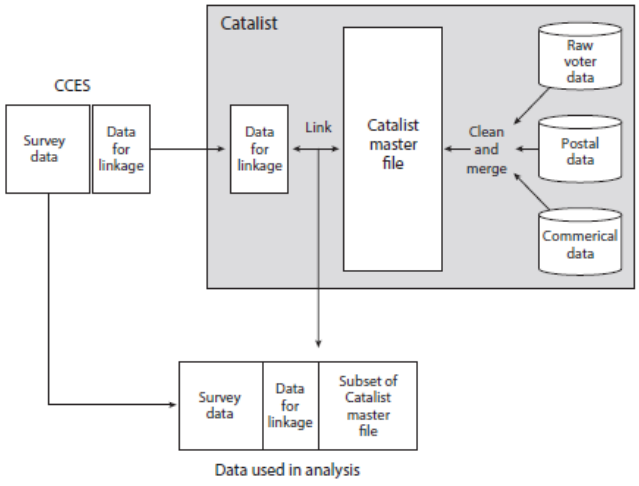


Figure 3.13: Schematic of Ansolabehere and Hersh's (2012) study. To create the master datafile, Catalist combines and harmonizes information from many different sources. This process of merging, no matter how careful, will propagate errors in the original data sources and will introduce new errors. A second source of errors is the record linkage between the survey data and the master datafile. If every person had a stable, unique identifier in both data sources, then linkage would be trivial. But, Catalist had to do the linkage using imperfect identifiers, in this case name, gender, birth year, and home address. Unfortunately, for many cases, there could be incomplete or inaccurate information; a voter named Homer Simpson might appear as Homer Jay Simpson, Homie J Simpson, or even Homer Sampsin. Despite the potential for errors in the Catalist master datafile and errors in the record linkage, Ansolabehere and Hersh were able to build confidence in their estimates through several different types of checks.

Using the data collection process to our advantage. We can also structure the data collection process in such a way that it will be easier for us to do record linkage later on. In particular, we can start with some "seed sources" which contains certain data on ecosystem actors and institutions such as their name, website (if any), social media accounts, and the like. Typically, these seed sources are online data directories which aggregate information from various sources.

This initial round provides us with URLs and keywords for search engines which can feed into the next round of web scraping, while ensuring that some of the data scraped are definitely linked to that particular actor or institution.

To illustrate the benefits of this method, notice that the data we have found for Bukalapak in the illustrative example above comes from the first round of web scraping. We can then use the links from the "Company Website", "Facebook", "Instagram", Twitter", and "LinkedIn" fields to scrape more information which can be linked to Bukalapak.

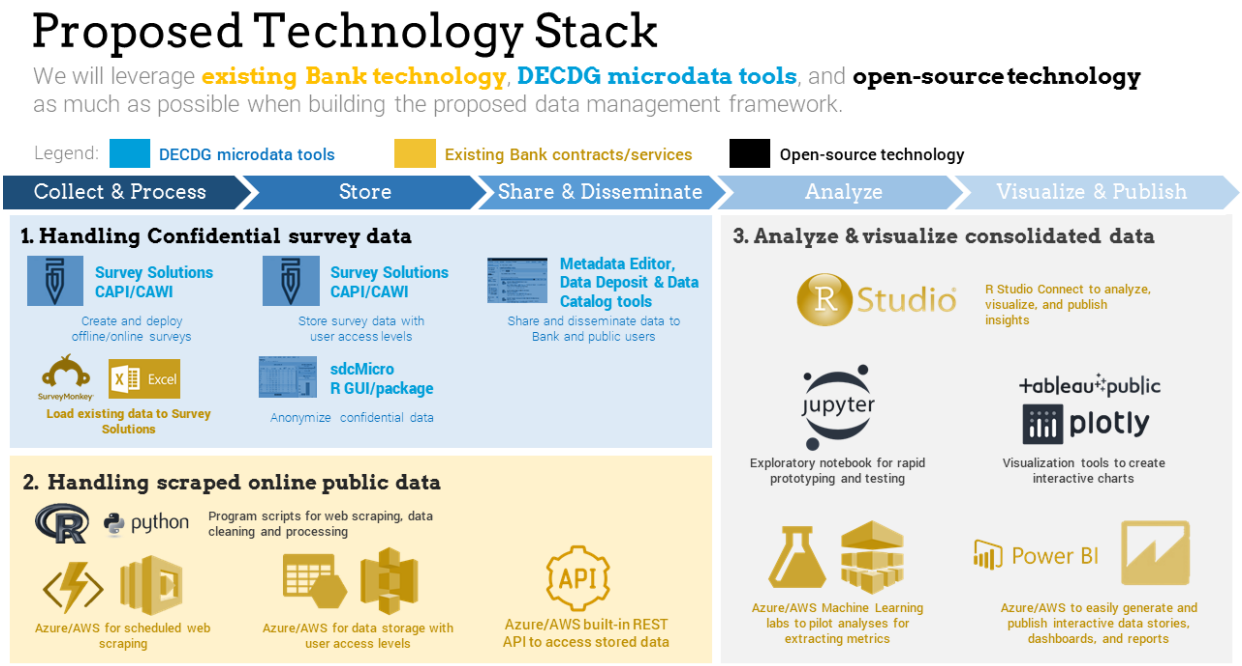
We can also add whitelisted or blacklisted sources for the scraping process, to filter dubious or less credible links and avoid scraping them.

4.2 Data Management and Storage

Once the web scraping exercise reaches a point of wide-scale implementation, it is important to support this with an appropriate data management and storage technology stack to build a long-term data asset which will consolidate all collected data and metrics. This data asset will potentially grow more valuable over time, as more features, countries, and data sources are collected consistently over time. One potential is that this asset unlocks powerful analyses and comparisons of the same indicators across different countries and over time.

It is important to leverage easy-to-use and flexible templates and tools for data analysis, visualization, and dissemination to enable ease in data sharing with both internal and external stakeholders. The example below shows a combination of World Bank Group and open-source tools / technology.

Figure 11. Potential technology stack for Data Management and Storage (within the World Bank environment)



Here are a few notes which may be useful in developing the data management framework:

- **Build a minimum viable product (MVP).** While this data asset may be a critical output, there are a lot of design choices which will be discovered along the way. Hence, it is good practice to implement a lean, agile methodology when developing this infrastructure by starting with a lean prototype with low investment and iterating on this based on regular stakeholder feedback.
- **Leverage existing organization tools as much as possible.** If the data management infrastructure is within the context of an organization, leveraging existing tools and

services is key to the long-term sustainability of the data management framework. This will avoid redundancy with and allow piggybacking on existing organization tools and processes at a lower cost.

- **Closely collaborate with key organization units.** In different organizations this might include groups as disparate as research, technology, information security, policy and others. Close collaboration with all of them can help ensure that the proposed data management framework can be easily integrated within existing organization tools and services.

In addition, here are some additional criteria to select the components and tools of the technology stack:

Table 13. Proposed Initial criteria for selecting components of the potential technology stack

Suggested Criteria	*Rationale*
Leverages existing organization data assets and software licenses as much as possible	To be consistent with the organization tool suite and best practices, and to minimize infrastructure and data tool costs where possible.
Enabling & Flexible. Allows ease and flexibility of use for both technical and non-technical users	Flexibility is key to adoption. Pinning this down at an early stage is particularly crucial, since one of the hardest issues when introducing new data tools is adoption of these by the target users (considering their technical skills, comfort with new tools, etc.).
Free and/or open-source	As much as possible, data tools should be preferably free and/or open-source to ensure flexibility and avoids long-term funding commitments.
Cloud-based. (Especially for the data storage and management tools)	Security and data backup are outsourced to industry-standard tools. Industry-standard tools can ensure that the data is fault-tolerant, requires low maintenance, is always accessible, and durable & distributed geographically.
Web browser-based.	Ensures that the data tool is always up-to-date. Increases chances of user adoption since minimal/no installation steps required. (Potential downside: increased reliance on good internet connection to upload, pull, analyze data.)
Has user login / authentication enabled	Survey/interview/FGD data also has some confidential aspects, so this can't be publicly published. Different internal teams and external stakeholders should have different access levels to the consolidated database.
Allows user collaboration	Members within survey teams usually need to collaborate to finalize reports and outputs.
Easily allow pulling in external data via data upload or API pull	For instance, the DEED methodology has special emphasis on sourcing / exploring data on TCdata360. Most sample surveys support their results using data from WBG or other external institutions.

Suggested Criteria	*Rationale*
Leverage off-the-shelf APIs of data tools when possible	Survey Monkey, AWS, and Microsoft Azure technology stacks are some tools which have off-the-shelf APIs which allow ease in sharing data across data tools.
Flexible data analysis and storytelling tools which can generate interactive, shareable data stories/visualizations	Offer a customizable tool that can be used to assess a particular ecosystem with ease at any point in time and to respond to specific client requests.
Conditional access for public access	Instead of having the whole tool password protected, there could be tiers of the tool that are open to the public, or where the public is even encouraged to contribute and modify directly the content.
Third party access/ crowdsourcing/ wikis for the tool.	In the overall data architecture, there could be value in having some layers not only open to third party providers, but also explicitly adopting a crowdsourcing/wiki approach.

4.3 Adjusting for Biases & Testing Quality of alternative data sources

While the proposed approaches have a lot of potential upside, they also contain several limitations and weaknesses, such as handling bias and data ownership issues. It is thus crucial to compare the extracted indicator data from alternative methods against official data sources (e.g., census data, household survey microdata) to (a) check for data quality and credibility and (b) test and adjust for biases.

The following are the overarching guidelines to consider:

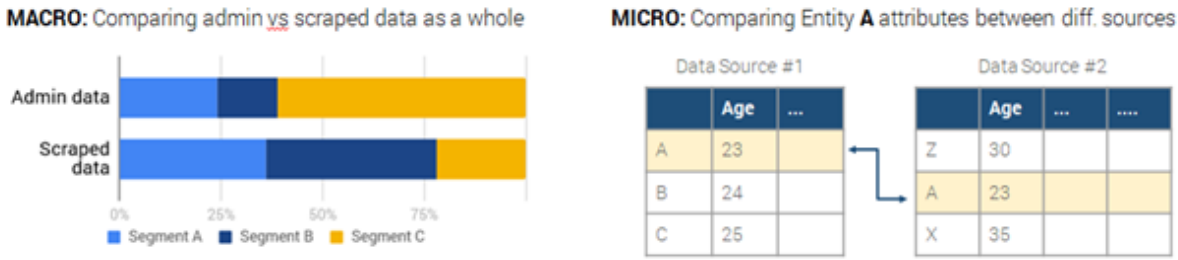
- **Greater value can be obtained by combining traditional and alternative data sources.** Traditional data plays a pivotal role in assessing and recalibrating the quality, validity, and accuracy of the scraped data collected (and show possible biases), for all steps of the process.
- **Leverage existing domain knowledge to ensure relevance and actionability.** We will have close coordination with subject matter experts and country survey teams every step of the way to get feedback and ensure relevance and applicability of the results to policymakers.
- **Transparency in data collection and analysis.** Showing how the data was collected and the metrics extracted can foster feedback, research replicability, and further interest and investigation.
- **Continuously check data and re-calibrate algorithms even in production.** Algorithms are never really "done" since the entrepreneurship ecosystem as well as the digital ecosystem dynamically changes over time. Continuous recalibration and

updating of the algorithm vis-a-vis latest traditional data will keep the data and metrics relevant, accurate, and of quality.

4.3.1 Checks for data quality and credibility

Comparing traditional and alternative data sources will help spot glaring differences, possible biases, and observe underlying patterns for these biases. We can check for quality along two levels, namely (1) macro or aggregated data and (2) micro or on the actor level.

Figure 12. Macro and Micro checks for data quality



We can implement the following quality checks at the micro-level and macro-level:

- Whitelist or blacklist certain online data sources based on credibility and advice from subject matter experts;
- Identify "gold standard" data sources among the available sources (typically census data or household survey data) to serve as the "ground truth" for the data comparisons.
- Triangulate indicator data (either granular or aggregated data) collected across different data sources. For example, triangulate the results of sites with unsure credibility against those which are identified as credible, and check the overlap or similarity of results returned.
- Identify data discrepancies among the data compared, and handle these through a combination of semi-automated checks guided by internally-defined criteria (such as source reliability as defined by subject matter experts, recency of data collected, and frequency of value among all sources considered) and manually checking a random sample of the records to ensure proper handling of edge cases.
- Work with legal teams to confirm and clarify the terms of use of public sources of data before proceeding with wide and long-term data scraping.

4.3.2 Testing and adjusting for biases

Note that alternative data sources commonly suffer from nonrepresentative and digital bias, and you cannot expect these sources to be accurate at the onset. It is thus important to leverage existing traditional data sources and use these to calibrate and adjust the collected

data from alternative sources. Traditional data and subject matter expertise will play a pivotal role for this process.

To merge the two data sources, the alternative data sources need to be checked for bias and adjusted using methods suitable for non-probability samples (which is often the case for alternative data sources and methods) such as:

- **Post-stratification** using auxiliary information about population strata (which are assumed to be mutually exclusive and exhaustive groups). This requires fulfillment of *"homogeneous response propensities within groups"* assumption wherein there should be little variation in the response propensity and outcome among the homogeneous groups formed.
- **Multi-level regression** wherein we estimate outcomes per group without enough (or zero) respondents by pooling together estimates from people in very similar groups.
- **Other methods to handle non-probability samples include:**
    - Sample matching (Ansolabehere and Rivers 2013; Bethlehem 2015)
    - Propensity score weighting (Lee 2006; Schonlau et al. 2009)
    - Calibration (Lee and Valliant 2009)
- **Some specific methodologies for adjusting for different types of biases:**
    - Adjust for population bias via reweighting by population segment (e.g., segmented by location, industry, firm size, firm age)
    - Adjust for selection bias via propensity score matching between survey and non-survey data (test: propensity of subject being in the survey data)
    - Adjust for activity bias (esp. for social media datasets, search datasets) by clustering data based on participant activity (e.g., recency, frequency)

We can then calibrate and reweight data from alternative data sources to adjust for bias. To test the reliability of the adjusted metrics, we can compare the adjusted data against corresponding traditional data as baseline (if available) and/or feedback from subject matter experts.

Figure 13. Depiction of data calibration

4.4 Limitations of the approach

Here are a few caveats for and limitations of this approach:

- **Care must be taken when interpreting results gathered from non-traditional sources.** Data gathered from online sources tend to suffer from some bias, especially depending on the data collection methods of that online source. For instance, global sources such as Pitchbook and Crunchbase may have incomplete data on African countries compared to their American counterparts. This may lead to overrepresenting some subset of the digital entrepreneurship ecosystem "population" whereas underrepresenting another subset.
    - To mitigate this, we can explore a mix of global and local data sources to complement one another, and to use triangulation to check for discrepancies in data collected among the different data sources.
- **The richness of the results greatly depends on the available non-traditional sources per country.** The quality of the data collected largely depends on the quality of the data from the non-traditional sources, so the results must always be taken with a grain of salt". Also, it is possible that there are data-poor countries which will have inadequate data sources to implement this approach.
- **Refinement of the methodology and data collected requires some manual checking by subject matter experts.** The quality of the data and the robustness of the methodology can be developed and further refined over time through feedback from subject matter experts.

5 The ethics of web scraping

While the techniques described above have great potential for research, questions inevitably arise about the propriety of scraping data without permission from website users and from entities described on such websites. Typical concerns include the following –

- **Technical**. Web scraping can place undue demands on websites and slow down their performance
- **Permission**. Some sites often explicitly prohibit scraping but do not have the technical resources to enforce it (see this related court ruling)
- **Deception**. Very often web scrapers do not identify themselves correctly to sites they are scraping from
- **Reuse**. Scrapers may not always have the permission to reuse the data they harvest
- **Awareness**. It is sometimes the case that web owners are unaware of the technical possibility of scraping and may be giving away their data out of ignorance

We propose the following mitigations –

- **Technical**. Scrapers must take care to not over-burden websites; scraping should ideally be infrequent or at off-peak hours and respect the technical infrastructure limitations of the source sites
- **Technical.** Scrapers must use the website API if the source website provides it
- **Permission.** Scrapers must first carefully review the terms and conditions of all websites they plan to scrape and not scrape content from websites that prohibit it even if they don't possess the technical means to enforce it. Some sites present a clear robots.txt message; others do not but state their objections through terms and conditions
- **Identification**. Scrapers must always identify themselves clearly and honestly. Inserting such information in code headers is easy and standardized. The information should also ideally include contact information
- **Reuse**. Scrapers must refer to the terms and conditions and respect the conditions for reuse. Intellectual property and trademark laws typically dictate how a website's content may be used; in any case scrapers should credit all information and as much as possible use it in a non-rivalrous fashion

It is also important to consider the sustainability and reproducibility of web scraping when incorporating such data into the research methodology. Many sites have begun to close themselves off to scrapers and while not widespread this may apply more forcefully to some projects than others. There are also cases wherein public APIs have been closed off for public use (e.g., Facebook, Instagram) or have had changes in access rights (e.g.,

AngelList), deprecation of API methods, rate limit changes, monetization strategies, among others. It is therefore important to keep in mind long-term sustainability and reproducibility when identifying which sources and techniques to implement at scale when establishing good initial foundations for the methodology, while being aware of the potential future deprecation and changes in data accessibility.

6 Conclusion and Looking ahead

The note provides a description of tools to both gather and analyze data from alternative, digital sources and apply them to answer some of the research and measurement questions related to entrepreneurship ecosystem assessments. The description above shows the value of such resources but also describes their limitations and a few mitigation approaches.

In general, the report demonstrates that such data can be a powerful complement to standard data sources, if used carefully and in the appropriate context, such as the following applications explored in this report:

- *Productivity and speed gains.* Techniques such as Named Entity Recognition (NER) can be used to extract relevant entities from website data, which leads to productivity and speed gains when parsing through large chunks of text for relevant data. Standard data sources can then be used to check the quality of the data extracted.
- *Knowledge discovery and compact representation.* Techniques such as topic modelling can be used to automatically extract topics (represented through relevant word clusters) from various texts. We can then group entities into clusters based on their topic association scores through thematic classification. Subject matter experts can then be tapped and consulted to verify if the resulting topics and entity clusters make sense.
- *New metrics.* A potentially useful new metric is general sentiment or "pulse" regarding a certain topic or entity, which we can derive using sentiment analysis to determine the polarity (i.e., positive/negative/neutral) of a given text. We can then check if this new metric strongly correlates with any of the existing standard metrics, and derive insights from patterns uncovered.
- *New data.* By collecting relationship data between ecosystem actors (such as investor-investee relationships), we can create network visualizations which allow us to map various entrepreneurship ecosystem actors with one another and look for patterns (e.g., how central an actor is, if there are clustering present in the ecosystem). We can then check if these patterns are aligned with our knowledge of the entrepreneurship ecosystem based on the standard DEED framework.

It is important for researchers to also consider a few additional issues and caveats if they would like to include alternative data in their methodology. These include –

- *Data and computational infrastructure.* Alternative data sources require sophisticated data and computational infrastructure to be scaled beyond small pilots. Projects or organizations thus need to make appropriate investments in their infrastructure.

- *Policies and guidelines*. Many organizations still do not have appropriate policies or guidelines in place for the use of alternative data. Recent experience has highlighted the numerous ethical, social, and other challenges associated with the gathering and use of such data. It is thus important for organizations to develop appropriate mechanisms and policies governing some of the techniques discussed.
- *Partnerships*. As the volume and variety of alternative data sources grows, it is impossible for most organizations to develop either the infrastructure or the skills to gather and manage such data. Data partnerships or collaboratives can offer a way forward in such situations.
- *Skills*. Data science is a fast-developing area and organizations should consider programs to develop and nurture the capacity of staff to use the techniques described above. Otherwise organizations face the risk of a wall between their data science teams and subject matter experts.
- *Sustainability and long-term reproducibility*. Changes and deprecation of API and general data access over time have been observed across various data sources such as Facebook, Instagram, AngelList, and the like. To mitigate this risk, it is important to establish good initial foundations for any methodology involving alternative data.

Bibliography

Ansolabehere, Stephen, & Hersh, Eitan. (2012). "Validation: What Big Data Reveal About Survey Misreporting and the Real Electorate." *Political Analysis* 20 (4): 437–59. doi:10.1093/pan/mps023.

Beskow, Laura M., Sandler, Robert S., & Weinberger, Morris. (2006). "Research Recruitment Through US Central Cancer Registries: Balancing Privacy and Scientific Issues." American *Journal of Public Health* 96 (11): 1920–26. doi:10.2105/AJPH.2004.061556.

Blumenstock, Joshua E., Cadamuro, Gabriel, and On, Robert. (2015). "Predicting Poverty and Wealth from Mobile Phone Metadata." *Science* 350 (6264): 1073–6. doi:10.1126/science.aac4420.

Costenbader, E., & Valente, T. W. (2003). The stability of centrality measures when networks are sampled. Elsevier B.V. Retrieved from https://www.bebr.ufl.edu/sites/default/files/Costenbader%20and%20Valente%20-%202003%20-%20The%20stability%20of%20centrality%20measures%20when%20networks.pdf

Endeavor Insight. 2014. *The Power of Entrepreneur Networks: How New York City Became the Role Model for Other Urban Tech Hubs.* http://www.nyctechmap.com/nycTechReport.pdf.

Ginsberg, Jeremy, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. (2009). "Detecting Influenza Epidemics Using Search Engine Query Data." *Nature* 457 (7232): 1012–14. doi:10.1038/nature07634.

Groves, Robert M. (2004). *Survey Errors and Survey Costs*. Hoboken, NJ: Wiley.

———. (2006). "Nonresponse Rates and Nonresponse Bias in Household Surveys." *Public Opinion Quarterly* 70 (5): 646–75. doi:10.1093/poq/nfl033.

———. (2011). "Three Eras of Survey Research." *Public Opinion Quarterly* 75 (5): 861–71. doi:10.1093/poq/nfr057.

Judson, D. H. (2007). "Information Integration for Constructing Social Statistics: History, Theory and Ideas Towards a Research Programme." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 170 (2): 483–501. doi:10.1111/j.1467-985X.2007.00472.x.

Olson, Janice A. (1996). "The Health and Retirement Study: The New Retirement Survey." *Social Security Bulletin* 59: 85.

http://heinonline.org/HOL/Page?handle=hein.journals/ssbul59&id=87&div=13&collection=journals.

Olson, Janice A. (1999). "Linkages with Data from Social Security Administrative Records in the Health and Retirement Study." *Social Security Bulletin* 62: 73. http://heinonline.org/HOL/Page?handle=hein.journals/ssbul62&id=207&div=25&collection=journals

Salganik, Matthew J. (2017). *Bit by Bit: Social Research in the Digital Age.* Princeton, NJ: Princeton University Press.

Startup Genome LLC. (2018). *Global Startup Ecosystem Report 2018: Succeeding in the New Era of Technology*. Retrieved from https://startupgenome.com/download-report/?file=2018