1

The Boundaries of Fintech: Data-Driven Classification and Domain Delimitation

CLAIRE INGRAM BOGUSZ AND JONAS VALBJØRN ANDERSEN

I. INTRODUCTION

Although fintech has been of considerable interest for researchers,[1] policymakers[2] and practitioners,[3] issues remain around how to define what activities and which firms should be considered 'fintech' and which should not. This issue stems partly from a question of whether fintech, or financial technology, is just an instance of digital technologies being used to deliver (new) financial services, or if there is something more to the phenomenon.

At its core, this is a question of boundaries: which services and firms should be included in fintech and how should one decide. Where these boundaries are placed has implications not only for researchers that seek to understand this emerging phenomenon, but also for policymakers – for instance when trying to establish the size and economic importance of fintech, and for regulators when trying to assess whether existing rules apply to fintech organisations, and whether new ones are needed. For authorities interested in competition, the identification of industry boundaries, and the usefulness of existing data for delimiting these boundaries, can help them better define markets or assess the effects of future policies. Boundaries are also of importance when it comes to

[1] See, eg, P Gomber et al, 'On the Fintech Revolution: Interpreting the Forces of Innovation, Disruption, and Transformation in Financial Services' (2018) 35 *Journal of Management Information Systems* 220; and R Teigland et al, *The Rise and Development of Fintech: Accounts of Disruption from Sweden and Beyond* (Routledge, 2018).

[2] See, eg, M Obschonka and DB Audretsch, 'Artificial Intelligence and Big Data in Entrepreneurship: A New Era Has Begun' (2020) 55 *Small Business Economics* 529; D Audretsch et al, 'Innovative Start-ups and Policy Initiatives' (2020) 49 *Research Policy* 104027.

[3] eg, Deloitte, 'Fintech: On the Brink of Further Disruption' (2020), available at: www2.deloitte.com/content/dam/Deloitte/nl/Documents/financial-services/deloitte-nl-fsi-fintech-report-1.pdf; PWC, 'Blurred Lines: How Fintech Is Shaping Financial Services' (2016), available at: www.pwc.com/il/en/home/assets/pwc_fintech_global_report.pdf.

deciding whether policies are needed and what effect they might have. Analysis reliant on registry and panel data is thus done ex ante when considering policy.

How, then, should one distinguish between fintech and adjacent industries like finance and IT? In this chapter, we take an empirical approach to answering that question. Based on a sample of 356 already identified fintech firms in Sweden, we use a supervised machine learning algorithm to (a) derive a dictionary that will allow us to identify 'missing' fintech firms in the Swedish Companies Registry; (b) cluster the resulting firms according to how they describe themselves in order to derive sub-categories or fintech domains; and (c) then compare the resulting fintech firms and their sub-categories against the classification codes used by the Swedish Registries Office, which are built on international standards. This third step is taken to see to what extent existing data can be used to reliably identify fintech firms. Sweden represents a suitable case at it has a considerable fintech ecosystem and follows (European Union) EU data standards, making the method generalisable to at least other EU countries and countries following a similar standard.

Sweden is a good site for a study of this kind, for several reasons. First, Swedish registry data are used frequently in academic and industry research, suggesting that they are extensive and reliable. Second, the country, in addition to an agency tasked with collecting data, Statistics Sweden, has a dedicated agency tasked with conducting analysis for the purposes of guiding policy and facilitating impact and growth assessments, the Swedish Agency for Growth Policy Analysis (Tillväxtanalys). Finally, the country regularly ranks highly in international assessments of its fintech firms, suggesting that there is a population of firms that can be identified in the data.

In so doing, we treat fintech as a phenomenon that spans classifications, specifically finance and information technology, or IT, classifications. Classification-spanning firms and industries present a challenge for policymakers in general because they are poorly understood and hard to identify.[4] For any single area of classification-spanning economic activity, it is hard to identify which firms to include and which to include when conducting analyses – and downstream policymaking. The inability to identify classification-spanning forms calls into question the usefulness of existing data for understanding these new forms, including their impact on productivity and inequality.[5]

II. BACKGROUND: FINTECH AS BOTH A FINANCIAL AND TECHNOLOGICAL PHENOMENON

There are many, though not always compatible, definitions of what fintech is, and thus which firms should be included in a resulting classification. In general,

[4] T Ciarli et al, 'Digital Technologies, Innovation, and Skills: Emerging Trajectories and Challenges' (2021) 50 *Research Policy* 104289.

[5] ibid.

definitions of fintech include two elements: finance, and technology, although they differ in their understandings of which are involved. This has downstream consequences for understanding the different domains or areas of activity within fintech.

Some studies treat fintech as a primarily financial phenomenon empowered by digital technologies,[6] while others – notably including industry analysts – see fintech as something that spans these two classifications, or at least comprises elements of both.[7] Those studies that see fintech as an extension of finance point to the fact that finance and technology have co-evolved:

> [Finance is] … a social technology, based on a system of recording assets and liabilities (credits and debits), which has developed through a series of innovations from coins, through to bills of exchange, double-entry book-keeping, insurance and central banking, all the way to financial derivatives and high-frequency algorithmic trading.[8]

Within this understanding, there is also the observation that technologies 'support and enable' the delivery of financial services,[9] but that while fintech firms are often start-ups, it may also be the case that fintech services are delivered by incumbent actors like banks.

This perspective seems to be consistent with the roots of the term 'fintech'. Schueffel, in 2015, points to the fact that the word 'fintech' was used as early as 1972 to 'stand for financial technology, combining bank expertise with modern management science techniques and the computer'.[10] However, given the advances in technology since then – and in particular the argument that digital technologies have fundamentally changed digital entrepreneurship by decentralising control and making agency unclear,[11] it is entirely possible that the ordinary understanding of the portmanteau may have evolved.

One possible understanding is to emphasise that fintech firms are technology firms first, but that they happen to provide financial services or services to the financial industry. Studies that highlight the importance of the technology in fintech emphasise that 'products and services provided by the industry are financial in nature, the processes and tools are mostly from the technology

[6] eg, E Knight and D Wójcik, 'Fintech, Economy and Space: Introduction to the Special Issue' (2020) 52 *Environment and Planning A* 1490; B Nicoletti, *The Future of Fintech: Integrating Finance and Technology in Financial Services* (Palgrave Macmillan, 2017).

[7] eg, Deloitte, 'Fintech: On the Brink of Further Disruption' (n 3); Gomber et al (n 1); N Wesley-James et al, 'Stockholm Fintech: An Overview of the Fintech Sector in the Greater Stockholm Region' (2015), available at: www.hhs.se/contentassets/b5823453b8fe4290828fcc81189b6561/stockholm-fintech---june-2015.pdf.

[8] Knight and Wójcik (n 6) 1490.

[9] Nicoletti (n 6) 12.

[10] P Schueffel, 'Taming the Beast: A Scientific Definition of Fintech' (2016) 4 *Journal of Innovation Management* 32, 33.

[11] S Nambisan, 'Digital Entrepreneurship: Toward a Digital Technology Perspective of Entrepreneurship' (2017) 41 *Entrepreneurship Theory and Practice* 1029.

industry'.[12] Understandings of fintech within this category build not on which services a firm provides, but rather on the technologies that it uses to provide them – although they do acknowledge that the services themselves need to be offered to customers or other firms in the finance industry. Examples include defining blockchain as primarily a fintech technology[13] and defining wearables that offer payment interfaces, for instance Apple Pay, as also being fintech.[14]

Other, complementary, characteristics of fintech transcend the question of whether finance or technology is more prominent. Instead, they emphasise things like agility,[15] novelty and innovativeness,[16] and the observation that fintech has the potential to, and in the case of for instance cryptocurrencies already does, dissolve physical and geographic boundaries.[17]

A. Domains within Fintech

A significant part of defining what fintech is, therefore, might be thought of in terms of the sub-classifications, or domains, in the larger classification. Again, there are two approaches: one loosely along financial service lines and the other along technology lines.

The approach to defining within-fintech domains along product lines (ie, classify fintech firms in terms of the services that they provide) is taken by Knewtson and Rosenbaum. They argue that fintech can be divided up into four sub-categories (see Figure 1), namely Monetary Alternatives, Capital Intermediation, InvestTech, and Infrastructure. However, one problem with this classification is that it defines financial services very broadly to include not only insurance (InsurTech), but also regulatory services in finance (Financial RegTech). While there are other systems that also have this 'big tent' approach to understanding fintech,[18] others have argued that these are separate areas of economic activity entirely.[19]

[12] HS Knewtson and ZA Rosenbaum, 'Toward Understanding Fintech and Its Industry' (2020) 46 *Managerial Finance* 1043, 1044.

[13] K Leong, 'Fintech (Financial Technology): What is It and How to Use Technologies to Create Business Value in Fintech Way?' (2018) 9 *International Journal of Innovation, Management and Technology* 74.

[14] MA Chen, Q Wuand and B Yang, 'How Valuable Is fintech Innovation?' (2019) 32 *Review of Financial Studies* 2062.

[15] Knewtson and Rosenbaum (n 12).

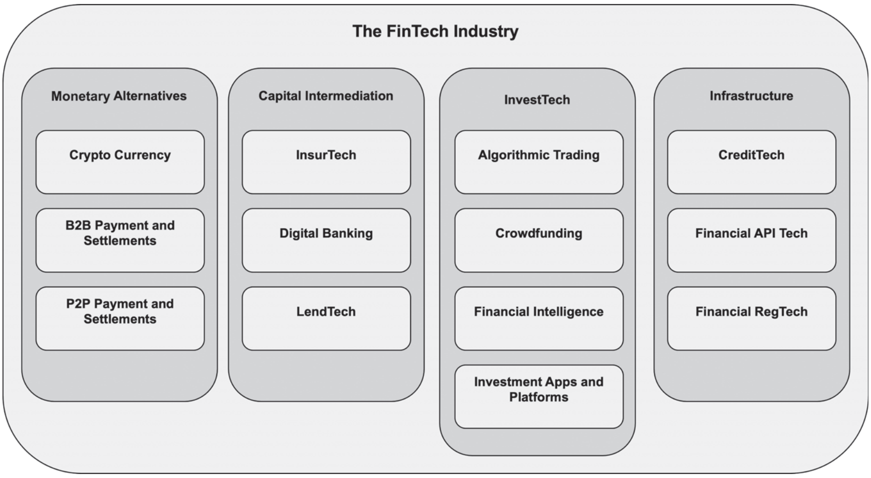
[16] Chen et al (n 14).

[17] Knight and Wójcik (n 6).

[18] eg, Deloitte, 'Closing the Gap in Fintech Collaboration: Overcoming Obstacles to a Symbiotic Relationship' (2018), available at: www2.deloitte.com/content/dam/Deloitte/us/Documents/financial-services/us-fsi-dcfs-fintech-collaboration.pdf; Deloitte, 'Fintech: On the Brink of Further Disruption' (n 3); and PWC (n 3).

[19] eg, Findexable, *The Global Fintech Index 2020* (2020), available at: findexable.com/wp-content/uploads/2019/12/Findexable_Global-Fintech-Rankings-2020exSFA.pdf.

Figure 1 A financial services-derived classification of fintech domains[20]



Another alternative is to classify fintech domains along technology lines (ie, classify fintech firms according to the technologies they use). One study that considers how to define fintech along technology lines looks at patent data in order to assess the value of fintech, rooted in an understanding that the technologies are central to the wider phenomenon.[21] Again, this classification of fintech according to key technologies (see Table 1) defines fintech very broadly and includes broader phenomena, for instance, big data, machine learning, and smart devices as being within the ambit of fintech.

Table 1 Definitions and examples of fintech domains in a technology-centric understanding[22]

Domain and definition	Key (digital) technologies	Examples
Cybersecurity: Hardware or software used to protect financial privacy or safeguard against electronic theft or fraud	Encryption, tokenisation, authentication, biometrics	Iris-scanning ATM, Biometric Cards
Mobile transactions: Technologies that facilitate payments via mobile devices, eg, smartphones, tablets, and wearables	Smartphone wallets, digital wallets, near-field communication	Apple Pay, Android Pay, PayPal Venmo

(continued)

[20] Adapted from Knewtson and Rosenbaum (n 12).
[21] Chen et al (n 14).
[22] Adapted from Chen et al (n 14).

Table 1 *(Continued)*

Domain and definition	Key (digital) technologies	Examples
Data analytics: Technologies and algorithms that facilitate transactions data or consumer financial data analysis	Big data, cloud computing, artificial intelligence, machine learning	Credit scoring, sentiment analysis
Blockchain: Distributed ledger technologies used mainly in financial services	Cryptocurrencies, smart contracts	Bitcoin, Ripple, JPM coin
Peer-to-peer (P2P): Software, systems, or platforms that facilitate direct financial transactions between consumers	Crowdfunding, P2P lending, customer-to-customer payments	GoFundMe, Kickstarter, Lending Club
Robo-advising: Computer systems or programmes that provide automated financial advice to customers or portfolio managers	Artificial intelligence, machine learning	Automated investment advice, portfolio placement recommendations
Internet of things (IoT): Technologies relating to smart devices that gather data in real time and communicate via the internet	Smart devices, near-field communication, wireless sensor networks	Smart home sensors, vehicle sensors

These understandings of both what fintech is and which domains are within fintech formed the backdrop for our own empirical study. There are various reasons for individual firms to register as either a financial service or a technology provider, including lower regulatory oversight for technology versus financial service firms, organisational culture and history, and strategic trajectory rather than actual output. Consequently, we used the financial services and technology categories to delineate the population of fintech firms, but opted not to define fintech as being either financial service-first or technology-first. Instead, we defined fintech as a class of firms delivering services that are qualitatively distinct and thus emerge from both categories without necessarily including all firms registered in each category. Therefore, the distinct characteristics of fintech firms is visible in their self-descriptions of their activities rather than in their specific register category. Taking this approach allowed us to build on the understanding echoed in previous studies that fintech combines elements of both finance and digital technologies.

III. RESEARCH DESIGN

There has been considerable enthusiasm from management scholars (and also from other disciplines such as law) in using new, digital methods to advance empirical research, and entrepreneurship research in particular.[23] In the academic and policy realms, much of this analysis is done in order to better understand a phenomenon, for instance to make predictions about sector growth, to infer consumer or investor sentiment, or to develop and text complex models in a data-first way. However, in legal scholarship it has also been suggested that digital empirical methods are not just a method for understanding phenomena, but that they also identify and evaluate unlawful conduct – perhaps even in real-time.[24] Indeed, both private actors and governmental agencies across the globe are creating roles like 'Chief Data Officer' and 'Chief Information Officer' not only to ensure compliance with data privacy laws like the General Data Protection Regulation,[25] but to pioneer and advance data-driven strategies, which typically use advanced analytics like machine learning.[26]

The premise upon which this enthusiasm lies is, first, in the belief that there is an abundance of so-called 'big data'[27] available for complex data-first analysis.[28] Second, proponents highlight that these data have opened the possibility of computational inductive methods[29] and computational theory development.[30]

While it has also been argued that proponents of using these methods of data analysis may misunderstand or oversimplify the processes involved,[31] ignore existing state of the art discussions on quantitative rigour[32] and overestimate the usefulness of the data available for complex analyses, we nevertheless show that, despite these limitations, registry data in conjunction with machine learning methods can provide a useful tool to identify and analyse classification-spanning entrepreneurial firms and associated within-classification economic domains.

[23] eg, M Maula and W Stam, 'Enhancing Rigor in Quantitative Entrepreneurship Research' (2020) 44 *Entrepreneurship Theory and Practice* 1059; Audretsch et al (n 2); Obschonka and Audretsch (n 2).

[24] eg, C Coglianese and A Lai, 'Antitrust by Algorithm' (2022), available at: scholarship.law.upenn.edu/faculty_scholarship/2755.

[25] Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and the free movement of such data [2016] OJ L119/1.

[26] In Sweden, for instance, the Swedish Tax Authority boasts an innovation team tasked with exploring how automated and data-driven analyses can improve their service and better detect tax-related crimes. In the United Kingdom, an agency for Government Digital Service was established in 2011, tasked with developing platforms and data-driven methods for improving service delivery.

[27] See R Kitchin, 'Big Data, New Epistemologies and Paradigm Shifts' (2014) 1 *Big Data & Society* 1 for a comprehensive discussion of what big data is, and is not.

[28] After more than a little cleaning, eg, V Mayer-Schönberger and K Cukier, *Big Data: A Revolution That Will Transform How We Live, Work, and Think* (Houghton Mifflin Harcourt, 2013).

[29] eg S Debortoli et al, 'Text Mining for Information Systems Researchers: An Annotated Topic Modeling Tutorial' (2016) 39 *Communications of the Association for Information Systems* 7.

[30] N Berente et al, 'Research Commentary – Data-Driven Computationally Intensive Theory Development' (2019) 30 *Information Systems Research* 50.

[31] D Carter and D Sholler, 'Data Science on the Ground: Hype, Criticism, and Everyday Work' (2016) 67 *Journal of the Association for Information Science and Technology* 2309.

[32] Maula and Stam (n 23).

This work builds on the arguments that (a) using the wealth of data that have become available for social science research should allow researchers to uncover previously complex insights not easily accessible using human intelligence;[33] (b) this should allow researchers to conduct studies on a population, rather than a sample, level;[34] and (c) using digital methods rooted in data could make studies more objective.[35] This enthusiasm extends both to using so-called 'big data' in entrepreneurship research,[36] and to the use of machine learning and artificial intelligence.[37]

A. Context

The choice of Sweden as our reference country is justified on the grounds of being a highly developed economy and one where standardised data is readily available. Sweden typically ranks among the best in the world when it comes to both innovation and good environment for doing business, including a well-established financial centre in Stockholm. According to the Swedish Bankers' Association, the financial industry accounted for 3.8 per cent of total output in Sweden in 2019 and employed around 95,000 people. At the same time, according to Statistics Sweden, 88,200 people, or around 2 per cent of the workforce were employed in finance and 191,100 people, or around 4 per cent of the workforce, were employed in ICT in 2019.

International rankings suggest that Stockholm is the biggest fintech hub per capita in Europe[38] and that Sweden, depending on the definition of fintech and associated metrics, is either seventh in the world[39] or third in the world.[40] At the same time, Swedish state agencies collect extensive data around registered firms, including not only their performance data, but free-text registered descriptions of firms, and registered classifications. We therefore chose to conduct an inductive study of fintech based on Swedish data, with the good quality data making computational analysis viable, and the international rankings indicating that the Swedish population of fintech firms might be considered representative.

[33] ibid.

[34] HJ Miller, 'The Data Avalanche Is Here. Shouldn't We Be Digging?' (2010) 50 *Journal of Regional Science* 181.

[35] D Boyd and K Crawford, 'Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon' (2012) 15 *Information, Communication & Society* 662.

[36] A Schwab and Z Zhang, 'A New Methodological Frontier in Entrepreneurship Research: Big Data Studies' (2019) 43 *Entrepreneurship Theory and Practice* 843.

[37] M Lévesque et al, 'Pursuing Impactful Entrepreneurship Research Using Artificial Intelligence' (2022) 46 *Entrepreneurship Theory and Practice* 803.

[38] Teigland et al (n 1).

[39] Findexable (n 19).

[40] Lucerne University, *FinTech Made in Switzerland: Clouds on the Horizon* (2020), available at: www.hslu.ch/en/lucerne-university-of-applied-sciences-and-arts/about-us/media/medienmitteilungen/2021/03/03/fintech-study-2021.

Swedish data are widely used in entrepreneurship and policy research,[41] making them a credible source of data for an attempt of this kind.

Sweden makes use of the Swedish Standard Industrial Classification (SIC) to classify firms and workplaces according to the industrial activities they carry out. This is based on the EU's recommended standards.[42] As such, then the results of our study can readily be applied to other EU countries and can with few adaptations be applied to other jurisdictions that follow similar standards. To our knowledge, no similar study has been done with SIC classifications. However, a recent study of US patents also made use of machine learning to identify and classify patents that could be considered to be artificial intelligence patents.[43]

B. Data

Our dataset consisted of companies' registry data maintained by the Swedish Companies Registration Office (Bolagsverket, BV) and related data from the Swedish Tax Authority (Skatteverket, SV) for the years from 2002 to 2018. We chose this dataset because it is widely used by policymakers and researchers. Registry data have long been the go-to data for following firms and firms in a particular industry over time, through a large number of repeated measures across different levels of analysis. This allows scholars not only to track firms and industries, but to draw causal inferences and employ multilevel research methods.[44]

As mentioned, Sweden applies the Swedish Standard Industrial Classification (SIC) to classify firms according to their activities. SIC codes are assigned through firm self-selection from a predefined array upon registration. Firms are legally obliged to update them if their industry of operation changes, and these SIC codes are a key way to delineate and classify firms, including entrepreneurial ones. Despite this legal obligation, however, self-identification is not without its problems. The most obvious of these problems is the subjectivity of self-assessment. A further problem, identified through informal conversations with experts, is the suspicion that many firms either register their firms in the broadest possible category to avoid having to change their registered classification later (at a cost), or forget to update their registered category despite the firm

[41] eg, T Ejdemo and D Örtqvist, 'Related Variety as a Driver of Regional Innovation and Entrepreneurship: A Moderated and Mediated Model with Non-linear Effects' (2020) 49 *Research Policy* 104073; or M Grillitsch et al, 'Knowledge Base Combinations and Firm Growth' (2019) 48 *Research Policy* 234.

[42] NACE Rev.2. SNI 2007.

[43] M Miric et al, 'Using Supervised Machine Learning for Large-Scale Classification in Management Research: The Case for Identifying Artificial Intelligence Patents' (2023) 44 *Strategic Management Journal* 491.

[44] K Wennberg, 'Entrepreneurship Research Through Databases: Measurement and Design Issues' (2005) 8(2) *New England Journal of Entrepreneurship* 9.

swinging into a new industry or product, which is commonplace among fintech firms. As becomes obvious below (see section V, Analysis of Results) this may affect the findings.

Some of these problems are, however, mitigated through the use of a second data source. Swedish firms also have the possibility of describing their area of operations in a free text format, usually at the time of registration. These data were also used to create a dictionary for the identification of fintech firms within the broad categories of finance and technology, limiting the effects of outdated or broad SIC category registration.

We delineated our analysis by focusing on firms that had self-selected both of the SIC categories: Technology and Finance. Working from this assumption, we employed a kind of machine learning algorithm, known as natural language processing (NLP), to identify, categorise and analyse patterns of fintech firms.

C. Method and Analysis

Our analysis was conducted in three phases: (i) distinguishing between fintech and non-fintech firms; and then (ii) identifying and categorising fintech firms based on both their free text descriptions and registered description of their activities. Based on this identification and categorisation we then (iii) explored patterns in registry-derived classifications and our NLP-derived classifications to understand if there was a relationship between the two.

i. Defining and Identifying the Fintech Firm Population

We first tried to identify fintech firms from self-reported descriptions in the entire company registry data using neural networks.[45] However, the free-text firm descriptions were too short to yield meaningful categories across the full registry data.

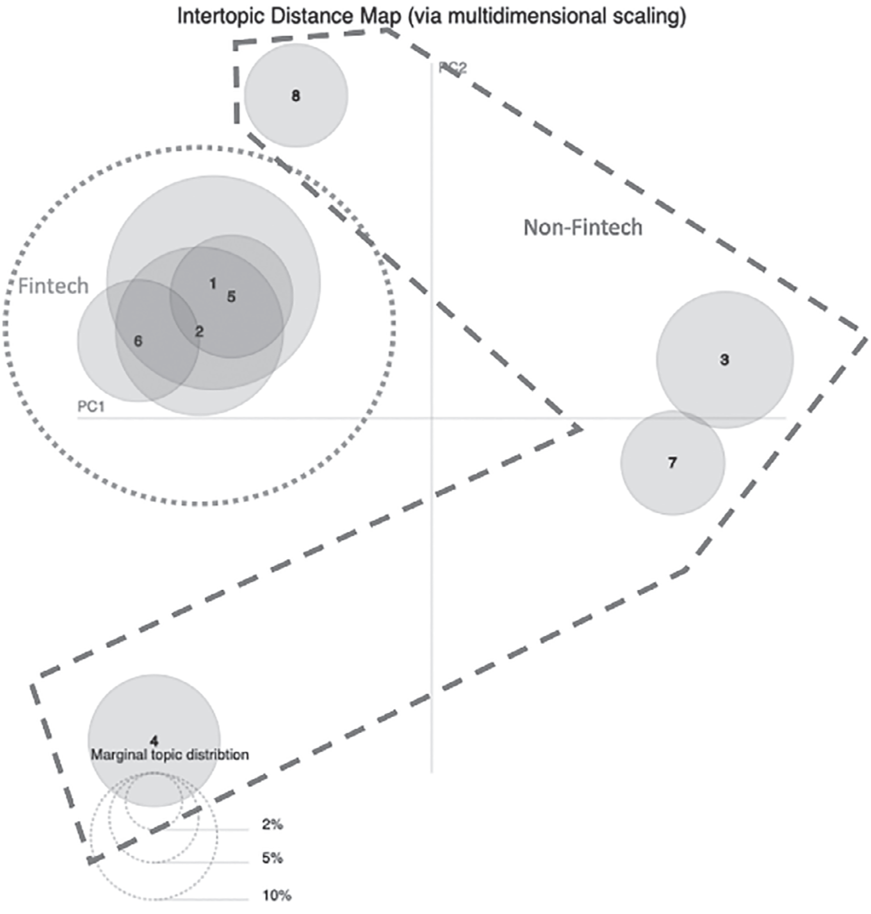
Another approach might be to train a machine learning algorithm, such as a neural network or support vector machine, on some test data to identify fintech firms in the entire population. Having obtained a list of 356 self-identified fintech firms, these test data proved insufficient to train an unsupervised algorithm and obtain useful classifications.

We therefore turned to training an NLP algorithm using a training dataset of 356 confirmed fintech companies provided by the Swedish Agency for Growth Policy Analysis (Tillväxtanalys, TVA). The test data was used to derive vocabularies relating to fintech that we could then apply to identifying fintech firms from the entire population of companies.

[45] CM Bishop, *Pattern Recognition and Machine Learning* (Springer, 2006).

Specifically, we ran the NLP topic modelling algorithm latent dirichlet allocation (LDA)[46] on the most recent free text descriptions of the confirmed fintech firms. LDA determines categories in corpuses of text, in this case specifically firm descriptions, based on term frequency, ie, how many times words appear in the same descriptions. In tuning the LDA algorithm on the training data, we took specific care to determine the right setting of the lambda parameter, which determines the exclusivity of words that are categorised within the same topic. High lambda allows for more topic overlap, and low lambda is more discriminatory and excludes terms that are also prevalent in other categories.[47]

Figure 2 Inter-topic distance map showing fintech related topic cluster



[46] Debortoli et al (n 29); DM Blei, 'Probabilistic Topic Models' (2012) 55 *Communications of the ACM* 77; and DM Blei et al 'Latent Dirichlet Allocation' (2003) 3 *Journal of Machine Learning Research* 993.

[47] Blei (n 46); Blei et al (n 46).

The LDA model revealed clusters of topics with some overlap. Semantically, topic clusters indicated whether a topic was related to either finance or technology. Figure 2 illustrates in a two-dimensional principal component analysis how topics relating to fintech firms are semantically distinguishable as a distinct cluster that separates them from non-fintech topics.

Using this method, we identified two distinct syntactic vocabularies that consistently related to either Technology or Finance. The stemmed terms included in each vocabulary are presented in Appendix A. To ensure the validity of our vocabularies, we manually inspected specific descriptions for prevalence of the selected terms and made minor adjustments.

Having done this, we ran a search algorithm to filter all firms using both Finance and Technology terminology in their free text description. We limited this search to those firms which had registered as being in SIC industries of Finance, Technology, Professional Services, and Other, which included Administrative Services.

We ensured validation of the results in terms of model specification and vocabulary relevance through three steps. First, we fitted the LDA model to the entire dataset and validated the results against the test set to make sure we did not miss any companies (ie, validated for false negatives). Second, based on the initial results, we updated both the vocabularies (ie, lists of words associated with finance and technology) and the LDA model parameters and re-ran the analysis until we had eliminated false negatives. Finally, based on the results from our updated model, we ensured face validity of the results by manually going through the descriptions of identified fintech firms with low frequencies of terms associated with finance and technology to ensure validity in terms of false positives in the included companies (ie, to ensure that we did not include companies that were not fintech). These steps were then also repeated for each category to ensure the validity of fintech firm identification.

In this way, we identified a total of 509 fintech firms through their own descriptions of their operations from the relevant SIC industry codes within the entire Swedish company registry.

ii. Categorising Fintech Firms

The second step involved reapplying LDA to the population of identified firms to discern if there were distinct categories within fintech, both in order to support nuanced policymaking and to relate these categories back to SIC codes. The initial results were manually validated to ensure a meaningful number of categories (represented by LDA parameter k), that each category was distinct and meaningful (LDA parameter λ), and to align the top-level label for each category (but not the content or delineation) with industry nomenclature.[48]

[48] Using Deloitte, 'Fintech: On the Brink of Further Disruption' (n 3) as a baseline.

Thereafter a new LDA analysis was repeated with updated parameter settings. From the resulting topic clusters, we derived distinct vocabularies for each category using a similar method as in step one. By using the topic distribution for each description (θn) resulting from the LDA model, we assigned each firm a score of how strongly it related to each of the categories based on the frequency of word usage associated with each vocabulary divided by the length of each specific vocabulary.

Based on this score, we then assigned each firm a category based on the highest relevance score. To ensure face validity, we again manually inspected the categorisation of specific firms (a) to confirm category fit, especially when there was an identical or similar score in two categories; and (b) to filter out false positives within each category. The result was then validated against the list of fintech firms identified in step 1. To ensure external validity, we asked an external panel of fintech experts to scrutinise our identification and categorisation of the fintech firms. The panel consisted of regulators and industry experts selected by the Swedish Agency for Growth Policy Analysis for their considerable experience in the fintech industry. Each expert was shown the results of the initial LDA topic model and asked to scrutinise the results of the algorithmic categorisation.[49] The panellists' feedback was used to refine and delineate the resulting categories. Following this methodology, we identified 10 categories of fintech firms within the identified 509 fintech firms.

Based on this identification, we also unpacked the firms' year of first registration to see if young (and thus entrepreneurial) firms were overrepresented. We then compared our classification system to the classification system currently in use in Sweden, which corresponds to international standards, to examine how useful existing classification systems are in identifying fintech firms. We then further compared the identified fintech firms and their domains of activity against existing industrial classification codes (SICs) to see the usefulness of these codes in identifying fintech firms. In what follows, we discuss these results.

IV. RESULTS: FINTECH AND ASSOCIATED DOMAINS

In order to derive a definition of fintech based on how fintech firms identified themselves, rather than definitions from academics or policymakers, we relied on the methods and data described above. As fintech represents a new class of industrial activities that are rapidly evolving, transforming, and merging with other classes of industrial activity, general classifications like the one outlined in Figure 1 are never fully up to date nor do they account for the idiosyncrasies of specific jurisdictions at specific points in time. Therefore, more accurate classifications of the fintech must be derived from activities as undertaken by the specific population of fintech firms through self-characterisation or records of

[49] This correspondence is on record with the authors.

business transactions. The task of producing a useful classification of fintech firms therefore involves identification of the population of fintech firms, classification of different sub-categories within the population, and quantitative analysis to identify patterns between sub-categories over time.

Using machine learning techniques, we first identified keywords that characterised how fintech firms described themselves in their registered free-text descriptions. The dictionary of descriptors is contained in Appendix A. For readers who do not speak Swedish we point out that the descriptions of finance included words like 'invest', 'credit', 'market', 'pension', 'transaction', 'advice' and 'pay'. Where they occurred together with IT descriptors like 'application', 'data', 'digital', 'internet', 'software', 'solution' and 'online', we considered those to be good candidates for fintech firms.

We then searched for the firms that contained combinations of both vocabularies and identified 509 firms (including the 356 we used to derive the dictionary). Based on this, we then derived the 10 domains of fintech contained in Table 2 and based on the dictionary of words contained in Appendix B.

Interestingly, the definitions of the 10 categories are very inclusive: they include obvious financial services like credit, payment services and financial management, but also adjacent finance-like services like data and analytics, RegTech (regulation technology), InsurTech (insurance technology) and even digital infrastructures.

There was little or no mention of specific technologies, with the exception of blockchain technologies. These firms were spread across several domains, but were most prominent in the infrastructure domain. This may be a practical choice from blockchain firms, in that they feel that blockchain does not adequately describe their operations. However, it might equally be a strategic decision – to avoid the scrutiny of regulators and similar, given the scepticism with which blockchain has historically been regarded.

Table 2 Descriptions of identified categories, stemmed vocabularies in Appendix B

Category	Count	% cent of total	Category description
Credit	78	15.4	Credit, loans and savings products, including crowdfunding and sales of invoices
Financial management	46	9.0	Financial management services directed towards individuals
Data	23	4.5	Data and analytics
Infrastructure	78	15.2	Technical services sold as products to other firms (typically B2B) to enable financial and fintech activities. Includes security, ERP and some blockchain firms

(continued)

Table 2 *(Continued)*

Category	Count	% cent of total	Category description
Insurance	21	4.1	Applications of fintech specifically within Insurance, includes both insurance firms and firms supporting insurance
Consulting	48	9.4	Consultant firms providing bespoke services (eg, to-order development) within fintech
Payments	77	15.2	Firms offering payments, transactions and remittance services
RegTech	10	1.9	Firms offering compliance and legal (tech) services
Wealth management	62	12.1	Firms offering investment and other wealth management services
Other	67	13.3	Firms that are fintech but not clearly in one of the above categories

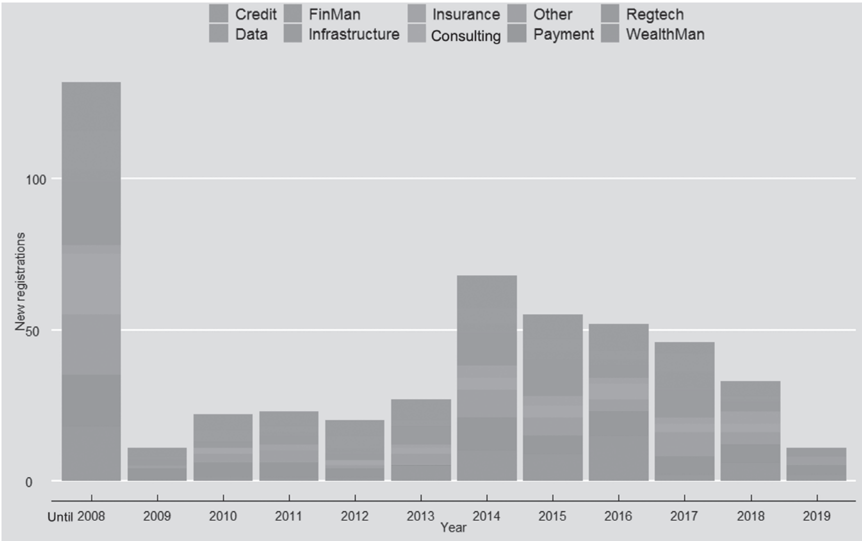
A. Grey Areas

Overall, there were considerable grey areas in this analysis. In addition to the 509 fintech firms, we identified a list of 2,247 firms which did not use technology in finance, but rather engaged in both technology *and* finance activities, for instance, by doing both software development and investing in listed and unlisted firms. This suggests that there is considerable untapped potential in the Swedish fintech market given the high number of firms with a good understanding of both finance and technology.

B. Firm Ages

We used the year of first SIC registration as a proxy for year of first registration (as firms register their SIC codes on registration). We can see that 132 firms were registered in 2008 or earlier, and that firm registrations have increased consistently year on year, reaching a peak of about 68 in 2014 (Figure 3). With around 25 per cent of the firms more than 14 years old at the time of writing, this suggests that fintech is by no means a phenomenon that is only being pioneered by entrepreneurs, or new firms.

Figure 3 Number of fintech firms by year and domain



C. Fintech Domains and Industry Classification (SIC) Codes

Based on the identified fintech firms and the resulting fintech domain categories, we now turn to discussing the usefulness of SIC codes in identifying fintech firms. Our hope was that the SIC codes would have some predictive value, given their importance for policymakers and researchers to track industries, draft supportive policies[50] and broadly encourage entrepreneurship.[51]

In particular, our hope was that there would be a relationship between the SIC codes, firm registered descriptions and categorisation. In such a case, a machine learning method like this could then be used to identify other kinds of cross-classification firms, for instance those in AgTech (agriculture), PropTech (property) or similar. Moreover, automated identification and classification of firms could considerably streamline a larger automated process in which analyses of industries and/or industrial sectors could be made. Moreover, such classifications could be used as part of a larger toolbox in ensuring that firms have the correct licences to, for instance, offer credit or financial advice.

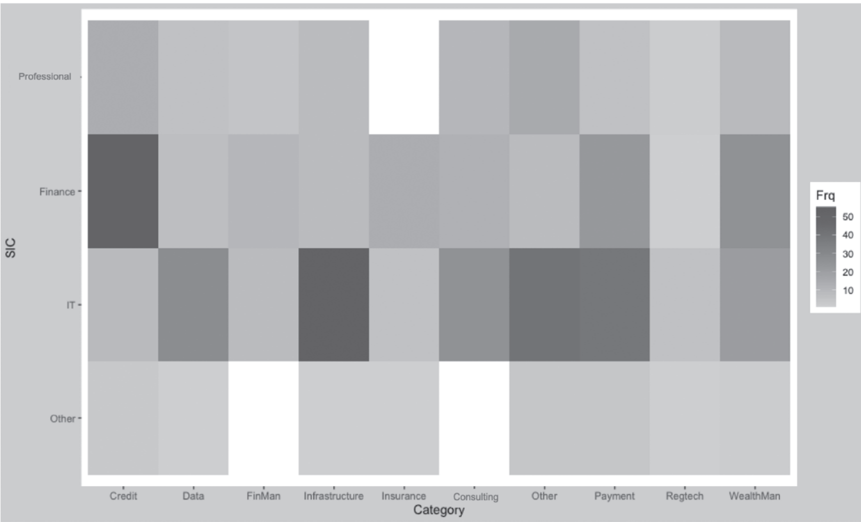
Using the most recent SIC codes of the 509 firms, we explored which SIC codes they used to classify themselves. Interestingly, 239 of these (47 per cent)

[50] Audretsch et al (n 2).

[51] Z Acs et al, 'Public Policy to Promote Entrepreneurship: A Call to Arms' (2016) 47 *Small Business Economics* 35; Z Acs et al, 'National Systems of Entrepreneurship: Measurement Issues and Policy Implications' (2014) 43 *Research Policy* 476.

defined themselves as being Tech (IT) companies (SIC group J, 58–63), and only 162 of them (31.8 per cent) described themselves as being primarily finance (SIC group K, 64–68). Almost 15 per cent (14.7 per cent) classified themselves as doing professional work (75 firms, SIC group M, 69–75), while just 33 (6.5 per cent) defined themselves as doing something else (all other SIC codes, including administration and other). A heat map of the number of firms in each fintech category across SIC codes is contained in Figure 4.

Figure 4 Heat map of SIC codes by fintech domain



V. ANALYSIS OF RESULTS

In what follows, we discuss some of the key take-aways of this data-first approach to understanding fintech. In particular, we point to how some areas of fintech are more finance oriented (eg, credit), and others more technology oriented (eg, infrastructure), but that broadly fintech is larger than even just finance and technology.

A. Fintech is Broader than Just Finance and Technology

While SIC codes are somewhat limited when it comes to identifying fintech firms in general, they are a better predictor for categories within fintech. Fintech firms that operate in heavily regulated areas of finance, like credit, classify themselves as being financial actors. However, those that operate in tech-heavy areas or which choose to signal that they are technology firms, rather than financial ones, instead choose technology classifications.

Only 73.4 per cent of the firms had a finance or technology SIC code as their primary classification; the rest classified themselves as something different. There are several possible explanations for this.

One possible explanation is that the phenomenon itself is broader than just finance and technology. This idea is supported by the emergence of categories like 'RegTech' and 'consulting' in the analysis. These are not a priori obvious categories in Financial Technology. However, the inclusion of finance-adjacent activities in the definition of fintech is not without precedent: RegTech itself is explicitly included in the definition of fintech by at least one producer of industry reports.[52]

Yet another explanation is that the SIC codes and the free text descriptions do not line up, either with each other, or with the firm's current activities. This might be because either the registered SIC code or the free text description are out of date, or just very broad. Indeed, when manually inspecting the free text descriptions, we noticed that many of them were very broad. For instance, one firm building an international payments network described their firm thus: 'The company will engage in software development, consulting services within IT, own shares in other companies, and related activities'[53] (translated by the authors from Swedish). This is clearly much broader than the scope of their day-to-day activities, although not inaccurate. It also makes strategic sense from the firm's point of view to describe their activities broadly rather than narrowly in order to limit how often they legally have to change their firm's description.

There may also be strategic reasons to prefer one SIC classification over another. For instance, firms may opt for an IT classification when their operations span two classifications, for the simple reason that they are less likely to attract the attention of regulators than in the more heavily regulated realm of finance.

B. Ambiguity and Boundaries in Classifications

Although data like company registry data are thought to present objective and consistent classification and quantification over time, the fact that both free text descriptions and SIC code registrations are self-selections on the part of the firms involved introduces ambiguity, both in the production of unstructured data points such as firm descriptions and in its analysis and interpretation.

[52] eg, Deloitte, 'Fintech: On the Brink of Further Disruption' (n 3); and Deloitte, 'Closing the Gap in Fintech Collaboration' (n 18).

[53] These free text descriptions are publicly available data; also available on request from the authors.

When it comes to the boundaries between finance and tech, it makes sense that as finance becomes more and more technical, it becomes a de facto area of applied IT, in which an IT classification makes the most sense for the firms involved. This is supported by recruitment data that shows that banks are increasingly developing new capabilities and expanding their software portfolios.[54]

When it comes to the boundaries of sub-categories within fintech, our machine learning-generated 'score' which allowed us to place boundaries between different types of fintech also highlighted that the boundaries between the different categories are not sharp. Instead, the vast majority of the firms identified had a dominant or top score, and scores in multiple other areas. However, only eight of the 509 firms had the same score in multiple categories[55] – suggesting that at least a firm's primary area of business is relatively distinct.

One further area of ambiguity lies in the distinction between tech *in* finance (or finance in tech) and finance *and* tech: as mentioned above, in addition to the 509 fintech firms, we identified a list of 2,247 firms which did not use technology in finance, but rather engaged in both technology *and* finance activities. For instance, by doing both software development and investing in listed and unlisted firms. Although registry data are said to track formal developments in economic and industrial activity,[56] it is problematic that this ambiguity exists when it comes to classifications, not least when a classification should be binary.

C. Policy Initiatives to Improve Data

If policymakers at national as well as EU levels want to launch policies that foster and regulate emerging digital entrepreneurship in classification spanning industries, such as fintech, they must first be able to identify and classify firms that participate in these classifications. While it is often the case that legislation itself specifies the kinds of firms to which it applies qualitatively, ex ante analyses of certain industries and industrial sectors are done on the basis of registry data. Registry data are thus used, among others, to conduct ex ante risk and impact assessments.

The data we relied upon in this case were publicly available. This means that an analysis such as this one might not only be used by state agencies interested in understanding new and existing industries, but also by private actors – for

[54] eg R Hendrikse et al, 'The Appleization of Finance: Charting Incumbent Finance's Embrace of Fintech' (2018) 4 *Finance and Society* 159.

[55] Their category was then confirmed by manual inspection.

[56] Acs et al, 'National Systems of Entrepreneurship' (n 51).

instance, companies may use such data to identify competitors. Initiatives by individuals can also benefit, for instance, for a jobseeker identifying potential employers. However, Sweden has a long history of public access to data, and has invested significant resources into collecting and verifying such data. These risks are therefore not new risks, but rather allow for the identification of firms and their classification in new, and perhaps less laborious, ways.

Our large-scale analysis combining supervised natural language processing analysis of known fintech firms and a similar analysis of the company registration database of firms in Sweden confirms that existing industry categories as represented by SIC codes are insufficient to identify fintech firms, and it provides a detailed sub-categorisation of fintech in Sweden as well as details of its economic development in several key dimensions. We believe this method has the potential to serve as a reliable tool for identifying and categorising classification spanning entrepreneurship and their economic impact for both technology and non-technology entrepreneurial activities. As the method is not conditional on the type of entrepreneurial activity, we are confident that it can be applied to a variety of emergent entrepreneurial phenomena including digital and social entrepreneurship as well as emerging ecosystems within, for instance, 'GreenTech' (Green Technology), 'AgTech' (Agricultural Technology), 'SpaceTech' (Space Technology) and others.

VI. CONCLUSIONS

This chapter describes how such classification spanning entrepreneurial firms can be identified and categorised by leveraging existing company registry data. This also provides insights into the structure of the classification-spanning entrepreneurship and its relation to existing industries that is useful for strategically nurturing and regulating these forms of entrepreneurship.

As digital technologies permeate existing industry categories, fintech firms are just one of many classes of new firms that span existing industry classifications. The rise in these kinds of entrepreneurship come against a backdrop of advances in understanding digital entrepreneurship, which has been described as blurring organisational and field boundaries,[57] but which is still emergent. Blurred boundaries are at the core of entrepreneurship, and classification-defying forms of entrepreneurship are a consequence, with associated challenges. Fintech is one kind of new classification-spanning portmanteau.

We hope the method presented in this research note will inspire researchers to apply and validate the method in classification-spanning entrepreneurship beyond fintech and that government agencies and regulators can implement it as a means of identifying, nourishing and regulating emerging entrepreneurial

[57] Nambisan (n 11).

categories. Such implementations may either apply the method in its current form, or with slight adaptations by adding additional textual data sources from a firm's public websites or social media profiles to provide more current and fine-grained classifications. This will not only enhance insights into entrepreneurial activities, but also provide a crucial point of reference for nourishing and integrating firms better with the surrounding economy, thus enhancing the impact and value of emergent entrepreneurship for industry and society at large.

APPENDICES

Appendix A: Stemmed Vocabularies Used to Classify Firms as Fintech/ Not Fintech

Finance	Tech
bank	analys
betal	applikation
bokför	data
crowdfunding	digital
försäkring	finansindustri
invest	hård
kredit	information
lån	internet
marknad	lösning
pension	mjuk
råd	online
räkning	produkt
transaktion	programmering
värdepapper	social
	system
	teknisk
	teknologi
	utveckla
	webb

Appendix B: Stemmed Vocabularies (in Swedish) of Fintech Domains

Insurance	Consulting	Investment	Credit	Payments	Data	Regtech	Infrastructure	Financial management	Other
försäkring	konsult	investera	instrument	betal	data	compliance	identi	privat	inspektion
ansvar	ledning	onoterade	kredit	mobil	analys	juridiska råd	system	översikt	
lösning	planering	rådgivning	lån	program	information	efterlevnad	licens	portfölj	
2004	coaching	försäkring	faktur	betaltjänst	visualis	avtal	säker	jämför	
	strategisk	värdepapper	spar		underlätta	kontrakt	stöd	folkbildning	
	utveckling	valuta			beslut	regel	block	rådgivning	
						signering	valuta	växling	
							växling		