The use of AI for company data gathering

Finding and monitoring fintechs in Germany and France

Guillaume Belly, Banque de France
Andy Bosyi, Neusinger
Elisabeth Devys, Banque de France
Ulf von Kalckreuth, Deutsche Bundesbank

1. Introduction

Fintech happens where innovation takes place in the financial sector, where new methods and products emerge, are tested and made ready for the market. The results are shaping the financial industry as a whole. Central banks need to identify, describe and understand fintech activities.

In dealing with fintechs, we have to do without the cornerstones of traditional statistics. There are few if any standardised reporting requirements, no developed taxonomy and no established set of quantitative measures. By definition, innovation involves new activities, and this is intrinsically difficult for traditional statistics, which need stable classifications. Company registers are mostly useless here. The business environment and market structures are changing rapidly. The segment is characterised by a high rate of "metabolism": entries, mergers and acquisitions, exits. Any list of fintech companies is rapidly outdated. To make things even worse, there is not even an accepted general definition of what "fintech" means.[1]

In order to prepare a multi-purpose fintech monitoring system for statistics, regulation and financial stability, we need to find new ways of collecting data and new tools to identify this type of firm. This should involve mostly publicly available information, plus information that is shared voluntarily. The project described here concentrates on the aspect of gathering information on the activity of fintech firms and identifying both known and unknown firms. The task is perfectly general – the toolbox developed for this project will thus be well-suited for monitoring fintechs all over the world, and any other rapidly evolving sector.

2. Project objective

We present a project in its infancy. However, we think that the approach is straightforward and generic. The project aims to create AI tools for gathering the master data needed for fintech monitoring in the first place, and for monitoring them over time. The task has two major aspects: (a) finding new fintechs and characterising their activity, and (b) recording major changes in activity among known market participants. As a co-operative effort of most of the major central banks in the world, the Irving Fisher Committee (IFC) Working Group on Fintech Data Issues, with the close co-operation of the BIS, has reviewed the state of

[1] Schueffel (2016), after a painstaking search through the available literature, arrives at three common features: 'technology', 'innovation' and 'finance'. However, what 'technology' and 'innovation' mean will always depend on time and context. Von Kalckreuth and Wilson (2020) argue that these terms cannot be part of an operational statistical classification system.

affairs and outlined a targeted roadmap for constructing fintech statistics[2] (see footnote for a link to the report). Developing non-administrative AI methods is part of this roadmap. Activities in the fintech industry are notoriously cross-border, and purely national or regional data collection is of limited value. Data collection on the fintech industry should be co-ordinated worldwide. A solution for gathering master data on fintechs that can be implemented globally would be an important first step.

3. Problem statement

By definition, innovation involves new activities. The business environment and market structures are changing rapidly. This is intrinsically difficult for traditional statistics, which need stable classifications. In addition, there are hardly any reporting requirements. Essentially, this will also remain the case in the future, because reporting requirements can only be legislated for activities which have been known for quite some time, not for innovative activities which are as yet unknown.

New and innovative methods are required in order to consolidate data and find new fintech entities, to characterise their activity, and to record major changes in the activities of existing market participants. The market activity of fintech companies is almost exclusively web-based, and it is well documented on their websites and in the dedicated online economic media. We can make use of this fact in several ways. Web scraping in connection with AI methods (text mining and graph theory) make it possible to find new fintechs and track the activities of known companies. With an operational list of fintechs at hand, we can use supervised or unsupervised learning on a large set of websites of IT-related companies. Concerning external information, we first concentrate on data from specialised information services. Later on, the websites of known fintechs may be added, as well as information from official registers, the websites of fintech associations, and consulting, venture capital and recruiting companies. Finally, internal data within central banks may further improve data quality. Charting the links between websites can yield a machine-based representation of conglomerates and company networks. This type of methodology is used at the Organisation for Economic Co-operation and Development (OECD) for its Analytical Database on Individual Multinationals and Affiliates (ADIMA), but not yet by central banks.

Using AI-based methods of probabilistic matching, it is possible to map websites to universal registers such as the ESCB Register of Institutions and Affiliates Database (RIAD) or the registers of statistical firms. This links information on economic activity to information on legal entities.

4. The two aspects of the task

The system needs a robust infrastructure to organise information. Once a sufficient amount of granular information on fintechs is gathered, classification methods (e.g. based on cluster analysis or graph mining to structure, partition and chart interactions within this ecosystem) can help provide a better understanding of the fintech landscape. The challenge has two dimensions, each interesting in its own right: monitoring known fintechs and finding new, hitherto unknown fintechs.

4.1. Monitoring known fintechs

Known fintechs can be monitored by tracking information brokers, lists and web fora on fintech. Websites can be monitored for relevant changes in activity in order to aid classification. This can help identify events

[2] See Irving Fischer Committee (2020), the final report of the working group.

such as mergers, acquisitions, insolvencies, and exits for a large number of firms. The downloading of balance sheets and other financial statements can be partly automatised. Key positions can be fed directly into information systems. This is very much analogous to existing "know your client" solutions on the market.

4.2. Finding new fintechs

New fintechs can be found by periodically running classification algorithms over large databases of private companies. Additionally, data from specialised information brokers, lists and web fora on fintech can be fed into the system. A rather free approach that does not rely on specific data structures is to scrape websites from companies in a larger list of IT companies in order to detect patterns that are characteristic of fintech companies.

5. Towards a non-administrative solution

The solution envisaged by the Bundesbank and the Banque de France is non-administrative and does not rely on the cooperation of the entities being monitored. It mostly uses publicly available data,[3] which is both an advantage and a limitation. Both central banks are performing preparatory pilot studies concerning feasibility and the prerequisites of the information structure, and are working together with fintechs. Thus far, the teams have concentrated on the classification problem, i.e. how to tell fintechs and non-fintechs apart. There is no closed-form and time-invariant definition of fintech, so the concept of fintech we use is implicit. We start from lists of firms which are considered to be fintechs based on the fintech monitoring in both central banks. This list embodies (implicit) information on the characteristics of innovative firms and technology oriented firms which offer or enable financial services that are of interest in the various areas of central bank work: supervision, financial stability, payments, etc.

The Banque de France has developed preliminary AI algorithms to identify fintechs in a large database of French firms, using a training data set of firms which the Banque de France has classified as fintechs. This is a well understood, fast and reproducible type of data collection, with natural connections to traditional statistics. The Bundesbank is exploring a graph approach that allows a rather open type of data gathering. Exploiting news data or the websites of companies, graphs of named entities (locations, organisations and persons) are created and enriched with additional information. Graph information can be processed by NLP techniques (such as named entity recognition, bag of words, word distances, etc.), and firms are classified by the nature of their links to the nodes. This has the advantage of being very up-to-date, as it draws directly on real-time information in news media or on websites.[4] Potentially, the approach may actually recognise new types of fintechs hitherto unknown to classifiers.

The authors believe that these two approaches are complementary – both are needed to keep track of developments in fast evolving, innovative markets.

---

[3] Most of the data are publicly available or "readable", but sometimes restrictions concerning use or the ability to apply scraping techniques – for example with regard to data from free online newspapers, blogs or social networks – make them almost proprietary. This is the case if, for example, fees are charged to collect them, depending on the business model and volume. In the case of using a third-party data broker, the data become proprietary even though most of the data are publicly available, since the collecting, aggregating, cleaning and formatting steps are charged.

[4] In order to do so, the Deutsche Bundesbank collaborates with Neusinger, an independent IT consultant firm specialised in AI and machine learning. Neusinger was one of the winners of a Bundesbank innovation challenge conducted in 2020 and 2021.

6. A graph-based approach

The first and principal step of the approach followed by the Bundesbank team is to extract relevant text information from publicly accessible websites or news databases for classification purposes. The name of a company to be classified is assumed to be known. In the following, we emphasise web scraping, as in the case of structured news information databases the task is similar, but simpler. Two tools are needed, a search engine and an article extraction engine.

The task of the search engine consists in searching for a term in the web or news database and return relevant information. The team used the Google search engine to search for terms. This service is provided by Rapid Search API. Returned results will depend on the search term, but search engines typically also allow the specification of search parameters, such as regular websites, news, images or places. The proper use of the parameters that would provide the best resulting data distribution is itself a matter for extensive research and analysis.

The article extraction engine is needed for properly extracting content from a web page. Modern HTML language provides different structural components (such as the <article> tag) for specifying the exact location of an article within the page. Still, this task is complex since:

- not all pages are structured conveniently – even if there is an article within the page, it may not be separated from the rest of the page;
- not all pages contain proper articles – the relevant content may be placed as pieces of text on different parts of the page;
- a lot of extra HTML code may have to be cleaned up and the parts containing relevant text have to be preserved, which needs to be done carefully.

Setting up an infrastructure for web scraping is a complex and time-consuming task. The team is using a readily available tool called Zyte (formerly ScrapingHub). When provided with a link to an article, it is capable of returning much useful information: article body, article HTML, article author list and main author, publication date, etc.

The data collection process for a single search term is summarised in Figure 1. Below we provide a short description of the steps.

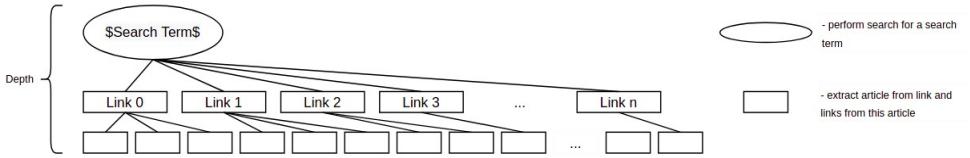


Figure 1: Data collection process

6.1. Performing search

The process starts with a search on a specific API, specifying the search terms. The initial search term for each company should be its name. Search term optimisation may be used here. For example, adding keywords like “company”, “fintech”, “finances”, etc. may improve the relevance of results. Using search query punctuation may lead to certain unwanted results being filtered out. Again, the modalities of such optimisation is a matter of research.

6.2. Processing links

For each link, the same actions are performed:

- An article or articles are scraped from the link using Zyte or a similar tool. The returned information is stored for future use and analysis.
- All named entities are extracted using Spacy, an article extraction tool. Although other entities may be used for modelling, the named entities with the highest potential are PERsons, LOCations and other ORGanisations.
- Links within the extracted article are extracted using tools for identifying regular expressions.

This process leads to a tree of links. The results need to be stored properly, preserving the information on hierarchical relations.

6.3. Proceeding for greater depth

The procedure on links described above is iterated until a desired depth level is reached. Judging from experiments, a depth of two levels (search links and next level links) may be sufficient.

6.4. Extending the graph

If more information is required, one may use extracted named entities as search terms and perform the same procedure from the beginning. In order to limit the number of links to be processed, it may be advisable not to go beyond one additional level of depth (search results links) for collected named entities.

6.5. Data collection results

To test the approach, the team started with a list of 1,190 companies, 390 of which were classified as fintech companies in the Bundesbank statistics. We performed a search for each one of them without search term optimisation, but querying for both regular google search results and specific news items (50 for each). This yielded 39K links on the first level and 518K links on the second level. A total of 39GB of data was retrieved, with 6.3M named entities in total. Among those, 1.1M were different, unique named entities.

6.5. Graph embeddings and DNN results

In order to transform the collected graph to a statistical model for predicting whether a firm is fintech or not, it is required to decrease the amount of data by cleaning procedures. Filtering nodes by the number of connections reduces the number of graph nodes to 1/60 of the original and filtering by the signal content of nodes cut the number of nodes by another $4/5^{th}$. The latter was achieved by retaining those nodes that had either a high or a low fraction of edges leading to fintech entities, thus being informative. At the same time, the cleaning procedure reduces the number of edges, and thus the graph size, from 54M to 260K.

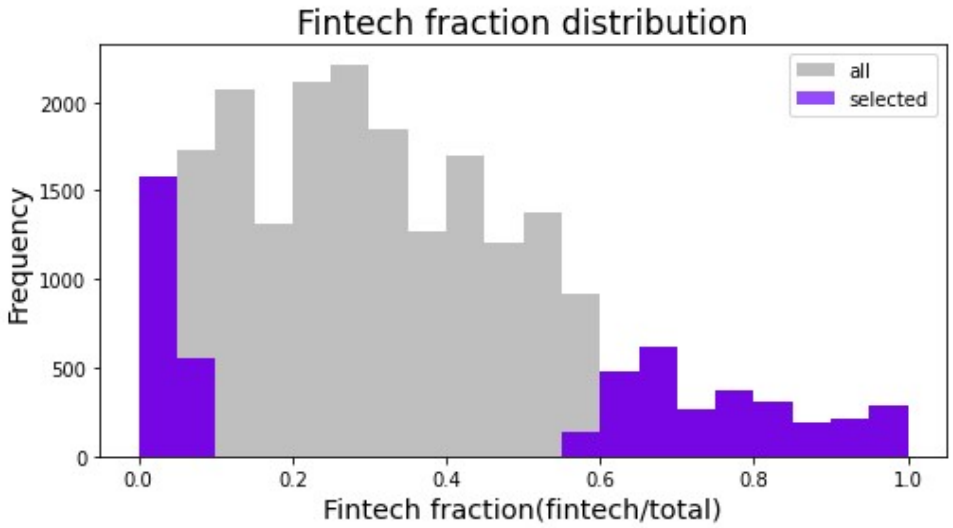


Figure 2: Selecting nodes based on their connections to fintechs and non-fintechs

Transformation from graph structure to a flat table is called node embedding and is carried out using the node2vec approach. Based on the deepwalk algorithm, the routine learns the graph structure distribution and trains a skip-gram encoder-decoder that can predict surrounding nodes from the characteristics of a given node. Taking the encoder hidden layer yields a trained neural network that can convert every node to a vector in the latent z-space. Figure 3 is a visual representation of the embedding process.

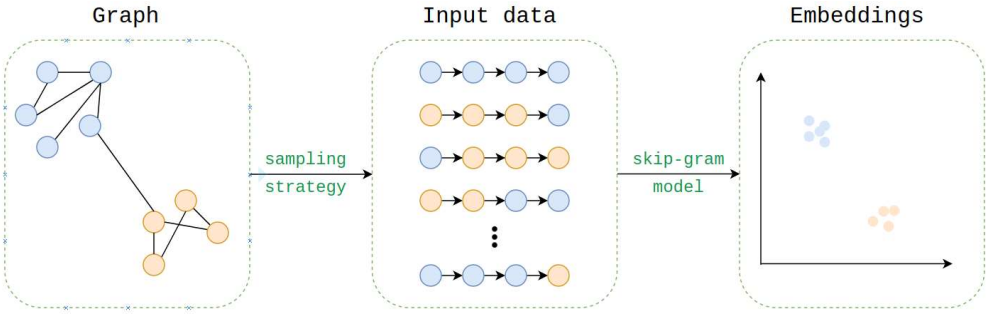


Figure 3: Embedding

The result in form of a 64 feature vector for every company is passed to a multilayer perceptron (DNN) of 64x16x1 layers with sigmoid activation function to determine the probability of the company being a fintech. Figure 3 is a t-SNE visualisation of data points by subset and label.

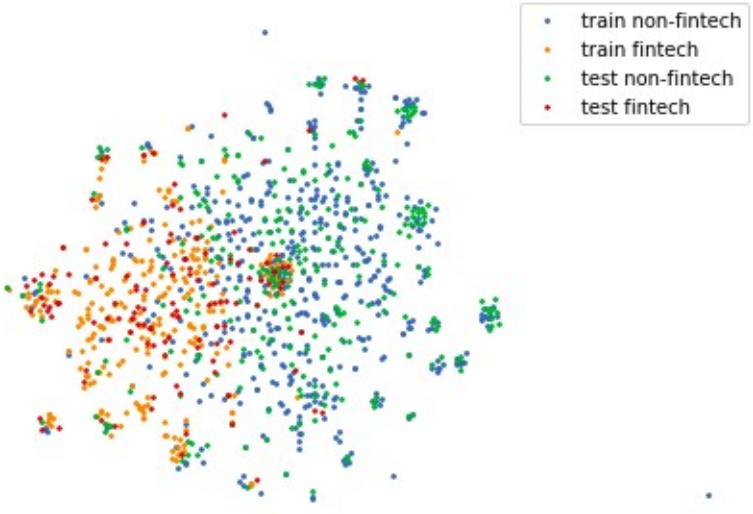


Figure 4: Visualisation of data point by subset and label

The evaluation of the model is carried out using k-fold cross validation. The dataset is split in three parts. In each round, the entire set of data is used for modelling, but only two of the three parts with the true label. In each turn, one of the three parts was used for evaluation, the other two for training. Ultimately, the metrics are averaged in the result table.

rithm performance

Table 1

uracy	ıll	cision
.87	3	.88

The experiment has shown that public web data (news, webpages, articles) on companies can be used to classify them as fintech or non-fintech with relatively good accuracy. More graph and model tuning will improve the results to the production level.

7. Semi-supervised AI database filtering

Complementary to the Bundesbank work, the Banque de France approach has focussed on classifying/detecting companies based on their characteristics, predicting whether a company is a fintech or not. Those characteristics were obtained by means of a company information database provided by a third-party data broker and a dedicated semi-supervised classification algorithm that was built and trained on it.

The use case that the algorithm has to solve is: "Among a very large set of companies, filter out the few of them that can be considered as fintechs and extract the features that contribute the most."

7.1 Data quality and featuring

The third-party database contains mostly publicly available data such as web-scraped data from social networks, legal data from public registers, news and economic press articles as well as financial data. This

database covers only French companies and contains more than 10 million referenced companies and more than 1,000 features.

As a proof of concept, it has been decided to limit our investigations to 10,000+ companies extracted from the database. To keep the strong imbalanced structure of the problem, the 10,000+ set of data has been built as follows:

- Around 350 companies have been selected and pre-tagged by the Banque de France experts as "potential fintechs".
- 10,000 "non-fintechs" extracted randomly from the database.

Data types are mostly categorical (e.g. economic sector, kind of newspaper, etc.), text-based (company description, title or text of news articles, employees job titles, etc.), numerical (e.g. turnover, number of employees, etc.) or dates (company establishment date, article publication date) and as a matter of consequence fully semi-structured.

As for most of these "big data" sources, a preliminary study has been conducted to get rid of the sparsity of the initial features (missing values), to select both the densest and the most business-oriented initial features. This has led to an initial draft set of 147 features including non-financial and financial data. In addition to this preliminary study, the iterative process of maximising the performance of the classification algorithm helped us to reduce this number to 84 features. Most of the financial features have been discarded because of their high level of sparsity, which creates too many false positives or false negatives in the classification results.

Once selected, the initial features were processed in the form of numerical categorical vectors (for categorical features), word embeddings or vectors of TF-IDFs, combining term frequency with inverse document frequencies (for text features) and scaled numerical features (for numerical features and dates) to be learnt or predicted by the algorithm.

7.2 Algorithm training and performance

The algorithm devised to classify an entity as a fintech or a non-fintech is semi-supervised. It undertakes one-sided learning of fintechs' characteristics, depicting non-fintechs as anomalies to the fintechs, and is based on a technique called "isolation forest".[5] The algorithm has been trained on a subset of the fintechs tagged by the experts of the Banque de France. The remaining fintechs and the non-fintechs have been kept for performance (prediction) evaluation, as could be done in real use. The semi-supervised nature of the algorithm helps to deal with the strong imbalance of the data (fintechs vs. non-fintechs). However, to promote better training, performance computation and thus model selection (involving the optimisation of hyperparameters), additional "synthetic fintechs" based on the "real" fintechs have been generated.

The best model obtained for our algorithm exhibits very low false negative and false positive rates. Accuracy, recall and precision are given in Table 2.

ase filtering: Algorithm performance

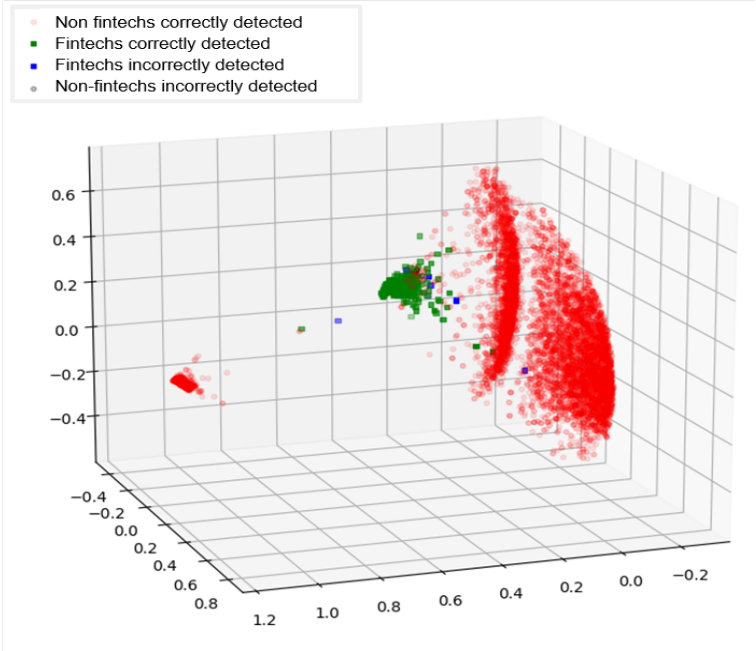
Table 2

uracy	ıll	ːision
995	)1	980

[5] Liu, F.T., Ting, K.M., Zhou, Z.-H. (2008)

Another way to visualise the result is to plot and colour the data points as functions of the algorithm results. Using a nonlinear kernel-based transformation (cosine based), the data can be projected and drawn in a 3D space. Each of the data points in this space corresponds to a company. The 10,000+ set of companies are plotted in Figure 5.

It is important to note the dense ball-like structure defined by the data points representing fintechs, which contrasts with the data points of non-fintechs, which inhabit a more elongated space. In addition to this, the distance between the fintechs and the non-fintech clusters is large enough to limit any overlapping between the two domains, improving the quality of partitioning and, as a consequence, the quality of the classification as fintech or non-fintech. These two observations highlight the relevance of the data preparation (the numerical featurisation step and the appropriate quality level of the data for our classification task).



Non-fintechs correctly detected: true negatives

Fintechs correctly detected: true positives

Fintechs incorrectly detected: false positives

Non-Fintechs incorrectly detected: false negatives

Figure 5: Fintechs and non-fintechs plot (3D embedding axis)

7.3 Explainability of the results and the importance of features importance

The ten features that globally contribute the most to the decision to classify an entity as a fintech or a non-fintech include:

- news articles: topic, source of a paper/journal;
- administrative: activity code, description, description of goods/services;
- job titles of several employees in the companies;
- name of the executives/funds/people on the boards;
- sector of registered trade marks.

7.4 Lessons learnt and way forward

From our work, we have proven the ability to detect fintechs with a very high level of confidence in a time snapshot and subset of French companies.

The main challenge has been to deal with the large data volume, as well as the variety and quality of data. This is a very time-consuming task, requiring a lot of expertise and a dedicated infrastructure to accommodate big data, known as a data lake. Scaling up from 10,000+ companies to 10 million (French perimeter), and from 10 million to several hundred million (worldwide perimeter) is a challenge even for third-party database providers.

To go further, a first step would be to confirm the ability of the algorithm to scale up to the full French perimeter and detect new fintechs over time and to monitor the efficiency over time over learning cycles

A second step would be to scale up this algorithm to at least the European perimeter (or even larger) and check the volume, variety, quality and availability of the data at this scale. As part of this second step, it could be assessed how and to what extent the graph-based approach and "algorithm training" could be combined to optimize the overall performance of an identification and monitoring framework.

A third step would be to assess the typology of the fintechs, figuring out whether or not some of them could be grouped into distinct classes.

Conclusion

In the process of preparing a multi-purpose fintech monitoring system for statistics, regulation and financial stability, the Banque de France and the Deutsche Bundesbank have started to investigate two complementary approaches based on AI. Both of them consider a wide range of non-structured or semi-structured data directly crawled from the web or in a more traditional way from third-party data brokers – "big data like" company information databases. While the Bundesbank approach is focussed on data gathering and extracting the properties of the subsequent graph to find unknown fintechs, the Banque de France approach concentrates on learning the characteristics of fintechs in a semi-supervised way to find the unknown fintechs from the whole pool of firms within a country.

In both approaches, the main challenges are to address the scaling up of the AI tools, and to address the amount of data to gather, process and feature at the level of a country, an economic area (eg the euro area) or worldwide with a robust and dedicated "big data like" infrastructure/data lake.

Literature

Irving Fisher Committee on Central Bank Statistics, Towards monitoring financial innovation in central bank statistics, IFC Report 12, July 2020

Schueffel, P., "Taming the beast: a scientific definition of fintech, *Journal of Innovation Management,* Vol. 4 (4) (2016), pp. 32-54.

von Kalckreuth, U., Wilson N., *Fintech and statistics – the challenge of classifying something that hasn't existed before*, in Irving Fisher Committee on Central Bank Statistics, Towards monitoring financial innovation in central bank statistics, IFC Report 12, July 2020, pp. 126-136.

Liu, F.T., Ting, K.M., Zhou, Z.-H. *Isolation Forest*, 2008 Eighth IEEE International Conference on Data Mining, IEEE, January 2009, 978-0-7695-3502-9.