Data Analysis Case Studies

Patrick Boily[1,2,3]

Abstract

In Data Science and Data Analysis, as in most technical or quantitative fields of inquiry, there is an important distinction between understanding the theoretical underpinnings of the methods and knowing how and when to best apply them to practical situations.

The successful transition from clean pedagogical toy examples to messy situations can be complicated by a misunderstanding of what a useful and insightful solution looks like in a non-academic context.

In this report, we provide examples of data analysis and quantitative methods applied to "real-life" problems. We emphasize qualitative aspects of the projects as well as significant results and conclusions, rather than explain the algorithms or focus on theoretical matters.

Keywords

case studies, data science, machine learning, data analysis, statistical analysis, quantitative methods

[1]Centre for Quantitative Analysis and Decision Support, Carleton University, Ottawa
[2]Department of Mathematics and Statistics, University of Ottawa, Ottawa
[3]Idlewyld Analytics and Consulting Services, Wakefield, Canada
Email: patrick.boily@carleton.ca

Contents

Data Analysis Case Studies

The case studies were selected primarily to showcase a wide breadth of analytical methods, and are not meant to represent a complete picture of the data analysis landscape. In some instances, the results were published in peer-reviewed journals or presented at conferences. In each case, we provide the:

- project title and citation references;
- author(s) and publication date;
- sponsors (if there were any), and
- methods that were used.

Depending on the case study, some of the following items are also provided:

- objective;
- methodology;
- advantages or disadvantages of specific methods;
- procedures and results;
- evaluation and validation;
- project summary, and
- challenges and pitfalls, etc.

Since the various sponsoring organizations have not always allowed the dissemination of specific results (for a variety of reasons), we have opted to follow their lead; when such results are available, the interested reader can consult them in the appropriate publications or presentations.

1. Classification: Tax Audits

Large gaps between revenue owed (in theory) and revenue collected (in practice) are problematic for governments. Revenue agencies implement various fraud detection strategies (such as audit reviews) to bridge that gap.

Since business audits are rather costly, there is a definite need for algorithms that can predict whether an audit is likely to be successful or a waste of resources.

Title Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue [1]

Authors Kuo-Wei Hsu, Nishith Pathak, Jaideep Srivastava, Greg Tschida, Eric Bjorklund

Date 2015

Sponsor Minnesota Department of Revenue (DOR)

Methods classification, data mining

Objective The U.S. Internal Revenue Service (IRS) estimated that there were huge gaps between revenue owed and revenue collected for 2001 and for 2006. The project's goals were to increase efficiency in the audit selection process and reduce the gap between revenue owed and revenue collected.

Methodology

1. *Data selection and separation:* experts selected several hundred cases to audit and divided them into training, testing and validating sets.
2. *Classification modeling* using MultiBoosting, Naïve Bayes, C4.5 decision trees, multilayer perceptrons, support vector machines, etc.
3. *Evaluation of all models* was achieved by testing the model on the testing set. Models originally performed poorly on the testing set until it was realized that the size of the business being audited had an effect of the model accuracy: the task was split in two parts to model large businesses and smaller business separately.
4. *Model selection and validation* was done by comparing the estimated accuracy between different classification model predictions and the actual field audits. Ultimately, MultiBoosting with Naïve Bayes was selected as the final model; the combination also suggested some improvements to increase audit efficiency.

Data The data consisted of selected tax audit cases from 2004 to 2007, collected by the audit experts, which were split into training, testing and validation sets:

- the **training data** set consisted of *Audit Plan General* (APGEN) *Use Tax* audits and their results for the years 2004-2006;
- the **testing data** consisted of APGEN Use Tax audits conducted in 2007 and was used to test or evaluate models (for Large and Smaller businesses) built on the training dataset,
- while **validation** was assessed by actually conducting field audits on predictions made by models built on 2007 Use Tax return data processed in 2008.

None of the sets had records in common (see Figure 1).

Strengths and Limitations of Algorithms

- The Naïve Bayes classification scheme assumes independence of the features, which rarely occurs in real-world situations. Furthermore, this approach tends to introduce bias to classification schemes. In spite of this, classification models built using Naïve Bayes have a successfully track record.
- MultiBoosting is an ensemble technique that uses forms a committees (i.e. groups of classification models) and group wisdom to make a prediction; unlike other ensemble techniques, it also uses a committee of sub-committee It is different from other ensemble techniques in the sense that it forms a committee of sub-committees (i.e. a group of groups of classification models), which has a tendency to reduce both bias and variance of predictions.

Procedures Classification schemes need a response variable for prediction: audits which yielded more than $500 per year in revenues during the audit period were *Good*; the others were *Bad*. The various models were tested and evaluated by comparing the performances of the manual audit (which yield the actual revenue) and the classification models (the predicted classification).

The procedure for manual audit selection in the early stages of the study required:

1. Department of Revenue (DOR) experts selecting several thousand potential cases through a query;
2. DOR experts further selecting several hundreds of these cases to audit;
3. DOR auditors actually auditing the cases, and
4. calculating audit accuracy and return on investment (ROI) using the audits results.

Once the ROIs were available, data mining started in earnest. The steps involved were:

1. **Splitting the data** into training, testing, and validating sets.
2. **Cleaning the training data** by removing inadequate cases.
3. **Building** (and revising) **classification models** on the training dataset. The first iteration of this step introduced a separation of models for larger businesses and relatively smaller businesses according to their **average annual withholding amounts** (the threshold value that was used is not revealed in [1]).

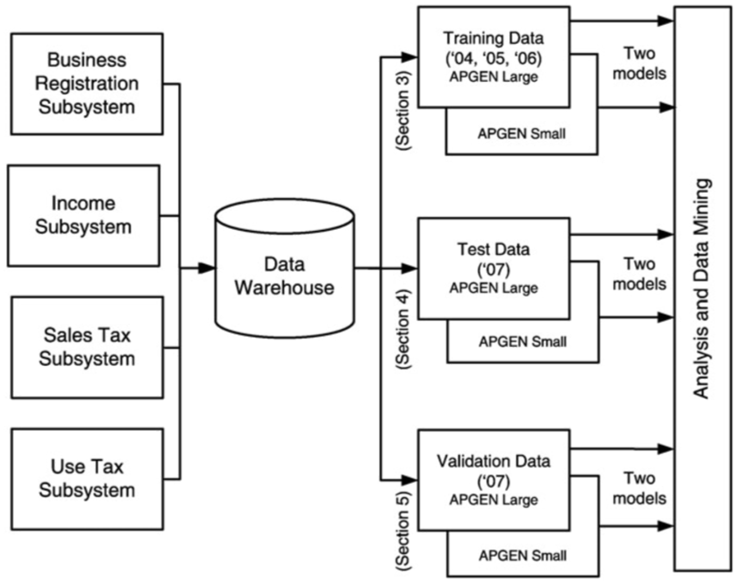


Figure 1. Data sources for APGEN mining [1]. Note the 6 final sets which feed the Data Analysis component.

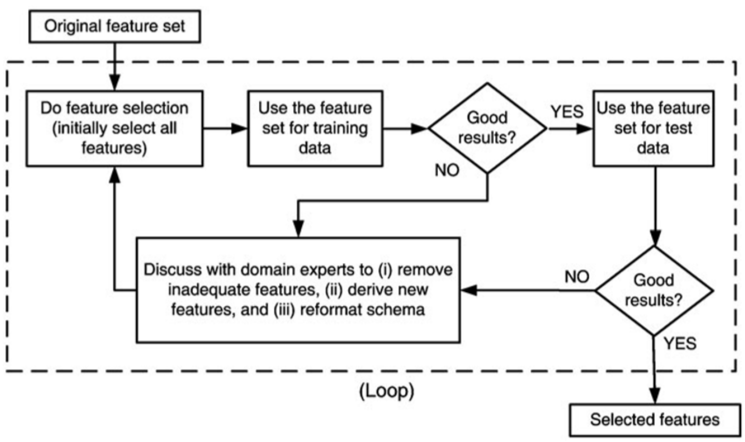


Figure 2. The feature selection process [1]. Note the involvement of domain experts.

4. **Selecting separate modeling features** for the APGEN Large and Small training sets. The feature selection process is shown in Figure 2.
5. **Building classification models** on the training dataset for the two separate class of business (using C4.5, Naïve Bayes, multilayer perceptron, support vector machines, etc.), and assessing the classifiers using **precision** and **recall** with improved estimated ROI:

$$\text{Efficiency} = \text{ROI} = \frac{\text{Total revenue generated}}{\text{Total collection cost}} \quad (1)$$

Results, Evaluation and Validation The models that were eventually selected were combinations of MultiBoosting and Naïve Bayes (C4.5 produced interpretable results, but its performance was shaky).

For APGEN Large (2007), experts had put forward 878 cases for audit (495 of which proved successful), while the classification model suggested 534 audits (386 of which proved successful). The theoretical best process would find 495 successful audits in 495 audits performed, while the manual audit selection process needed 878 audits in order to reach the same number of successful audits.

For APGEN Small (2007), 473 cases were recommended for audit by experts (only 99 of which proved successful); in contrast, 47 out of the 140 cases selected by the classification model were successful. The theoretical best process would find 99 successful audits in 99 audits performed, while the manual audit selection process needed 473 audits in order to reach the same number of successful audits.

In both cases, the classification model improves on the manual audit process: roughly 685 data mining audits would be required to reach 495 successful audits of APGEN Large (2007), and 295 would be required to reach 99 successful audits for APGEN Small (2007), as can be seen in Figure 3.

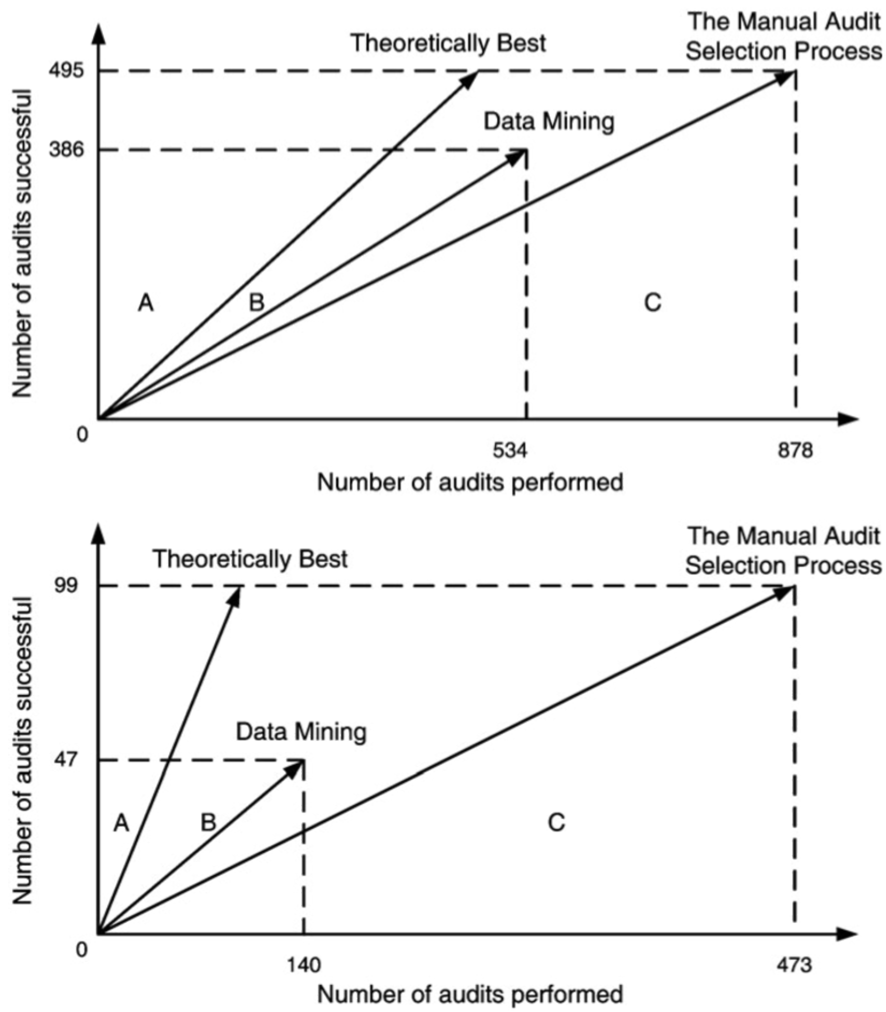


Figure 3. Audit resource deployment efficiency [1]. Top: APGEN Large (2007). Bottom: APGEN Small (2007). In both cases, the Data Mining approach was more efficient (the slope of the Data Mining vector is "closer" to the Theoretical Best vector than is the Manual Audit vector).

Table 1 presents the confusion matrices for the classification model on both the APGEN Large (2007) and APGEN Small (2007) data. Columns and rows represent predicted and actual results, respectively. The revenue R and collection cost C entries can be read as follows: the 47 successful audits which were correctly identified by the model for APGEN Small (2007) correspond to cases consuming 9.9% of collection costs but generating 42.5% of the revenues. Similarly, the 281 bad audits correctly predicted by the model represent notable collection cost savings. These are associated with 59.4% of collection costs but they generate only 11.1% of the revenues.

Once the testing phase of the study was conpleted, the DOR validated the data mining-based approach by using the models to select cases for actual field audits in a real audit project. The prior success rate of audits for APGEN Use tax data was 39% while the model was predicting a success rate of 56%; the actual field success rate was 51%.

Take-Aways A substantial number of models were churned out before the team made a final selection. Past performance of a specific model family in a previous project can be used as a guide, but it provides no guarantee regarding its performance on the current data – remember the *No Free Lunch (NFL) Theorem* [2]: nothing works best all the time!.

There is a definite iterative feel to this project: the feature selection process could very well require a number of visits to domain experts before the feature set yields promising results. This is a valuable reminder that the data analysis team should seek out individuals with a good understand of both data and context. Another consequence of the NFL is that domain-specific knowledge has to be integrated in the model in order to beat random classifiers, on average [3].

Finally, this project provides an excellent illustration that even slight improvements over the current approach can find a useful place in an organization – data science is not solely about Big Data and disruption!

	Predicted as good	Predicted as bad
Actually good	386 (Use tax collected) R = $5,577,431 (83.6 %) C = $177,560 (44 %)	109 (Use tax lost) R = $925,293 (13.9 %) C = $50,140 (12.4 %)
Actually bad	148 (costs wasted) R = $72,744 (1.1 %) C = $68,080 (16.9 %)	235 (costs saved) R = $98,105 (1.4 %) C = $108,100 (26.7 %)

	Predicted as good	Predicted as bad
Actually good	47 (Use tax collected) R = $263,706 (42.5 %) C = $21,620 (9.9 %)	52 (Use tax lost) R = $264,101 (42.5 %) C = $23,920 (11 %)
Actually bad	93 (costs wasted) R = $24,441 (3.9 %) C = $42,780 (19.7 %)	281 (costs saved) R = $68,818 (11.1 %) C = $129,260 (59.4 %)

Table 1. Confusion matrices for audit evaluation [1]. Top: APGEN Large (2007). Bottom: APGEN Small (2007). R stands for revenues, C for collection costs.

References

[1] Hsu, K.W., Pathak, N., Srivastava, J., Tschida, G., Bjorklund, E. [2015] *Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue*, in: Abou-Nasr, M., Lessmann, S., Stahlbock, R., Weiss, G. (eds) *Real World Data Mining Applications*, Annals of Information Systems, v.17, Springer. doi:10.1007/978-3-319-07812-0_12

[2] Wolpert, D.H. [1996] The Lack of a priori distinctions between learning algorithms, *Neural Computation*, v.8, n.7, pp.1341-1390, MIT Press. doi:10.1162/neco.1996.8.7.1341

[3] Wolpert, D.H., Macready, W.G. [2005] Coevolutionary free lunches, *IEEE Transactions on Evolutionary Computation*, v.9, n.6, pp.721-735, IEEE Press. doi:10.1109/TEVC.2005.856205

2. Sentiment Analysis: BOTUS and Trump & Dump

In 2013, the BBC reported on various ways in which social media giant Twitter was changing the world, detailing specific instances in the fields of business, politics, journalism, sports, entertainment, activism, arts, and law [9].

It is not always clear what influence Twitter users have, if any, on world events or business and cultural trends; it was once thought (perhaps without appropriate evidence) that entertainers, athletes, and celebrities, that is to say, users with extremely high followers/following ratios, wielded more "influence" on the platform than world leaders [1].

Certainly, such users continue to be among the most popular – as of September 13, 2017, Twitter's 40 most-followed accounts tend to belong to entertainers, celebrities, and athletes, with a few exceptions [15].

One account has recently bridged the gap between celebrity and politics in an explosive manner: @realDonaldTrump, which belongs to the 45th President of the United States of America, has maintained a very strong presence on Twitter.

As of September 13, 2017, the account had 38,205,766 followers, and it was the 26th most-followed account on the planet, producing 35,755 tweets since it was activated in March 2009 [15].

Titles BOTUS [5], Trump & Dump Bot [16]

Authors Tradeworx (BOTUS), T3 (Trump & Dump)

Date 2017

Sponsor NPR's podcast *Planet Money* (BOTUS)

Methods sentiment analysis, social media monitoring, AI, real-time analysis, simulations

Objective There is some evidence to suggest that tweets from the 45th POTUS may have an effect on the stock market [8, 10]. Can sentiment analysis and AI be used to take real-time advantage of the tweets' unpredictable nature? Let's take a look at bots built for that purpose by NPR's *Planet Money* and by T3 (an Austin advertising agency).

Methodology Tradeworx followed these steps:

1. *Data collection*: tweets from @realDonaldTrump are collected for analysis.
2. *Sentiment analysis of tweets*: each tweet is given a sentiment score on the positive/negative axis.
3. *Validation*: the sentiment analysis scoring must be validated by observers: are human-identified positive or negative tweets correctly identified as such by BOTUS?
4. *Identification of the company in a tweet*: is the tweet even about a company? If so, which one?
5. *Determining the trading universe*: are there companies that should be excluded from the bot's trading algorithms?
6. *Classifying tweets as "applicable" or "unapplicable"*: is a tweet's sentiment strong enough for BOTUS to engage the trading strategy?
7. *Determining a trading strategy*: how soon after a flagged tweet does BOTUS buy a company's stock, and how long does it hold it for?
8. *Testing the trading strategy on past data*: how would BOTUS have fared from the U.S. Presidential Election to April 2017? What are BOTUS' limitations?

T3's Trump and Dump uses a similar process (see Figure 4).

Figure 4. T3's Trump and Dump process [16].

Data The data consists of:

- tweets by @realDonaldTrump (from around Election Day 2016 through the end of March 2017 for BOTUS; no details are given for T3) (see Figure 5 for sample);
- a database of publicly traded companies, such as can be found at [3, 14, 17], although which of these were used, if any, is not specified (no explicit mention is made for BOTUS), and
- stock market data for real-time pricing (Google Finance for T3) and backcasting simulation (for BOTUS, source unknown).

It is not publicly known whether the bots are upgrading their algorithms by including new data as time passes.

Strengths and Limitations of Algorithms and Procedure

- In sentiment analysis, an algorithm analyzes documenta in an attempt to identify the attitude they express or the emotional response they seek. It presents numerous challenges, mostly related to the richness and flexibility of human languages and their syntax variations, the context-dependent meaning of words and lexemes, the use of sarcasm and figures of speech, and the lack of perfect inter-rater reliability among humans [12]. As it happens, @realDonaldTrump is not much of an ironic tweeter – when he uses "sad", "bad" and "great", he usually means "sad", "bad" and "great" in their most general sense. This greatly simplifies the analysis.
- The bots have to learn to recognize whether a tweet is directed at a publicly traded company or not. In certain cases, the ambiguity can be resolved relatively easily with an appropriate training set (Apple the company vs. apple the food-item, say), but no easy solutions were found in others (Tiffany the company vs. Tiffany the daughter, for example). Rather than have humans step in and instruct BOTUS when it faces uncertainty (which would go against the purpose of the exercise), a decision was made to exclude these cases from the trading universe. What T3's bot does is not known.
- Once the bot knows how to rate @realDonaldTrump's tweets and to identify when he tweets about publicly-traded companies, the next question is to determine what the trading strategy should be. If the tweet's sentiment is negative enough T3 shorts the company's

Figure 5. Examples of @realDonaldTrump tweets involving Delta, Toyota Motor, L.L.Bean, Ford, Boeing, Nordstrom.

stock.[1] Of course, this requires first purchasing the stock (so that it can be shorted). Planet Money's decision was similar: buy once the tweet is flagged, and sell right away... but what does "right away" mean in this context? There is a risk involved: if the stock goes back up before BOTUS has had a chance to purchase the low-priced stock, it will lose money. To answer that question, Tradeworx simulated the stock market over the last few months, introducing the tweets, and trying out different trading strategies. It turns out that, in this specific analysis, "right away" can be taken to be 30 minutes after the tweet.

Results, Evaluation and Validation For a trading bot, the validation is in the pudding, as they say – do they make money? T3's president says that their bot is profitable (they donate the proceeds to the ASPCA) [16]: for instance, they netted a return of 4.47% on @realDonaldTrump's Delta tweet (see Figure 5); however, he declined to provide specific numbers (and made vague statements about providing monthly reports, which I have not been able to locate) [11].

The BOTUS process was more transparent, and we can point to Planet Money's transcript for a discussion on sentiment analysis validation (comparing BOTUS's sentiment rankings with those provided by human observers, or running multiple simulations to determine the best trading scenario) [5] – but it suffers from a serious impediment: as of roughly 4 months after going online, it still had not made a single trade [4]!

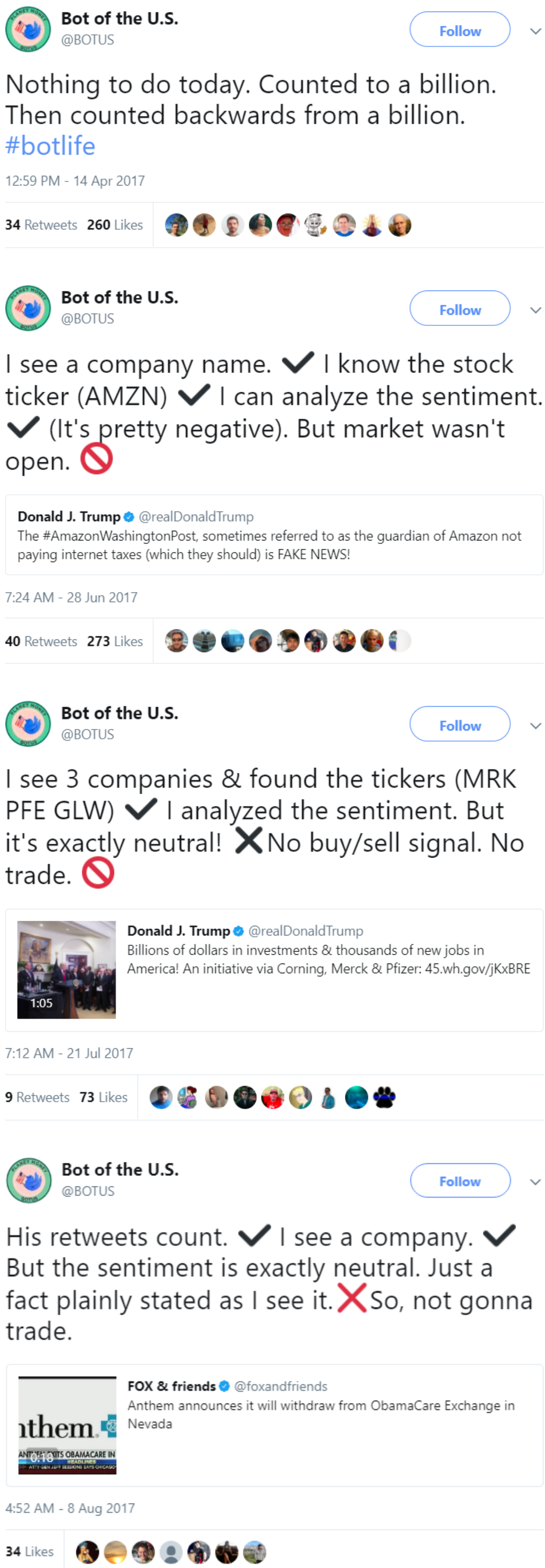
The reasons are varied (see Figures 6 and 7), but the most important setback was that @realDonaldTrump had not made a single valid tweet about a public company whose stock BOTUS could trade during the stock market business hours. Undeterred, Planet Money relaxed its trading strategy: if @realDonaldTrump tweets during off-hours, BOTUS will short the stock at the market's opening bell.

This is a risky approach, and so far it has not proven very effective: a single trade of Facebook's trade, on August 23rd, which resulted in a loss of 0.30$ (see Figure 7).

Take-Aways As a text analysis and scenario analysis project, both BOTUS and Trump & Dump are successful – they present well-executed sentiment analyses, and a simulation process that finds an optimal trading strategy. As predictive tools, they are sub-par (as far as we can tell), but for reasons that (seem to) have little to do with data analysis *per se*.

Unfortunately, this is not an atypical feature of descriptive data analysis: we can explain what has happened (or what is happening), but the modeling assumptions are not always applicable to the predictive domain.

[1] It sells the stock when the price is high, that is, *before* the tweet has had the chance to bring the stock down, and it repurchases it once the price has been lowered by the tweet, but before the stock has had the chance to recover.



Bot of the U.S.
@BOTUS
Follow

Nothing to do today. Counted to a billion. Then counted backwards from a billion. #botlife

12:59 PM - 14 Apr 2017

34 Retweets 260 Likes

Bot of the U.S.
@BOTUS
Follow

I see a company name. ✔ I know the stock ticker (AMZN) ✔ I can analyze the sentiment. ✔ (It's pretty negative). But market wasn't open. 🚫

Donald J. Trump @realDonaldTrump
The #AmazonWashingtonPost, sometimes referred to as the guardian of Amazon not paying internet taxes (which they should) is FAKE NEWS!

7:24 AM - 28 Jun 2017

40 Retweets 273 Likes

Bot of the U.S.
@BOTUS
Follow

I see 3 companies & found the tickers (MRK PFE GLW) ✔ I analyzed the sentiment. But it's exactly neutral! ✖ No buy/sell signal. No trade. 🚫

Donald J. Trump @realDonaldTrump
Billions of dollars in investments & thousands of new jobs in America! An initiative via Corning, Merck & Pfizer: 45.wh.gov/jKxBRE

1:05

7:12 AM - 21 Jul 2017

9 Retweets 73 Likes

Bot of the U.S.
@BOTUS
Follow

His retweets count. ✔ I see a company. ✔ But the sentiment is exactly neutral. Just a fact plainly stated as I see it. ❌ So, not gonna trade.

FOX & friends @foxandfriends
Anthem announces it will withdraw from ObamaCare Exchange in Nevada

0:18

4:52 AM - 8 Aug 2017

34 Likes

Figure 6. BOTUS reporting on its trades.

Bot of the U.S.
@BOTUS
Follow

2/ I see a company. But also another company! ❌ With two once, I can't tell which is the smart trade. Safer not to guess. No trade. 🚫

Donald J. Trump @realDonaldTrump
Toyota & Mazda to build a new $1.6B plant here in the U.S.A. and create 4K new American jobs. A great investment in American manufacturing!

5:49 AM - 4 Aug 2017

6 Retweets 67 Likes

Bot of the U.S.
@BOTUS
Follow

3/3 I see a company. Found the ticker. ✔ But I can't fetch the price because it's in Taiwan market & I only do U.S. markets. No trade. 🚫

Donald J. Trump @realDonaldTrump
....and don't forget that Foxconn will be spending up to 10 billion dollars on a top of the line plant/plants in Wisconsin.

5:59 AM - 4 Aug 2017

38 Likes

Bot of the U.S.
@BOTUS
Follow

Replying to @realDonaldTrump

.@realdonaldtrump tweeted about Facebook, Inc. I shorted the stock at $168.67 and lost $0.30.

Donald J. Trump @realDonaldTrump
Thank you Arizona. Beautiful turnout of 15,000 in Phoenix tonight! Full coverage of rally via my Facebook at: facebook.com/DonaldTrump/vi...

1:44

7:01 AM - 23 Aug 2017

26 Retweets 169 Likes

Bot of the U.S.
@BOTUS
Follow

One trade in total so far. Down less than 1%. (Happens to the best of bots, right?) I'm gonna keep on hanging in here. You with me?

3:43 PM - 15 Sep 2017

12 Retweets 460 Likes

Figure 7. BOTUS reporting on its trades.

References

[1] Arthur, C. [2010], **The big bang visualisation of the top 140 Twitter influencers**, retrived from **The Guardian.com** on September 13, 2017.

[2] Basu, T. [2017], **NPR's Fascinating Plan to Use A.I. on Trump's Tweets**, retrieved from **inverse.com** on September 12, 2017.

[3] **Company List: NASDAQ, NYSE, & AMEX Companies**, retrieved on **NASDAQ.com** on September 15, 2017.

[4] Dieker, N. [2017], **Planet Money's BOTUS Bot Has Yet to Make a Single Stock Trade**, retrieved from **Medium.com's The Billfold** on September 12, 2017.

[5] Goldmark, A. [2017], **Episode 763: BOTUS**, Planet Money podcast, retrieved from **NPR.org's Planet Money** on September 12, 2017.

[6] Green, S. [2017], **Trump Tweets: Separate Positive and Negative Tweets**, retrieved from **Green Analytics** on September 15, 2017.

[7] Greenstone, S. [2017], **When Trump Tweets, This Bot Makes Money**, retrieved from **NPR.org** on September 12, 2017.

[8] Ingram, M. [2017], **Here's What a Trump Tweet Does to a Company's Share Price**, retrieved from **Fortune.com** on September 15, 2017.

[9] Lee, D. [2013], **How Twitter Changed the World, Hashtag-by-Hastag**, retrieved from **BBC.com** on September 13, 2017.

[10] McNaney, B. [2017], **A Negative Trump Tweet About Your Company Is An Eye Opener, Not A Crisis**, retrieved from **The Buzz Bin** on September 22, 2017.

[11] Mettler, K. [2017], **'Trump and Dump': When POTUS tweets and stocks fall, this animal charity benefits**, retrieved from the **Washington Post** on September 19, 2017.

[12] Ogneva, M. [2010], **How Companies Can Use Sentiment Analysis to Improve Their Business**, retrieved from **Mashable** on September 17, 2017.

[13] Popper, N. [2017] **A Little Birdie Told Me: Playing the Market on Trump Tweets**, retrieved from **The New York Times** on September 22, 2017.

[14] **List of Publicly Traded Companies**, retrieved on **InvestorGuide.com** on September 15, 2017.

[15] **Top 100 Most Followed Twitter Accounts**, retrieved from **Twitter Counter** on September 13, 2017.

[16] **Trump & Dump Bot: Analyzes Tweets, Short Stocks**, retrieved from **T3** on September 14, 2017.

[17] **The World's Biggest Public Companies List - Forbes 2000**, retrieved on **Forbes.com** on September 15, 2017.

3. Clustering: The Livehoods Project

When we think of similarity at the urban level, we typically think in terms of neighbourhoods. Is there some other way to identify similar parts of a city?

Title The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City [2]

Authors Justin Cranshaw, Raz Schwartz, Jason I. Hong, Norman Sadeh

Date 2012

Sponsors National Science Foundation, Carnegie Mellon's CyLab, Army Research Office, Alfred P. Sloan Foundation, CMU/Portugal ICTI, with additional support from Google, Nokia, and Pitney Bowes.

Methods spectral clustering, social dynamics

Objective The project aims to draw the boundaries of **livehoods**, areas of similar character within a city, by using clustering models. Unlike static administrative neighborhoods, the livehoods are defined based on the habits of people who live there.

Methodology The case study introduces a spectral clustering model (the method will be described later) to discover the distinct geographic areas of the city based on its inhabitants' collective movement patterns. Semi-structured interviews are used to explore, label and validate the resulting clusters, as well as the urban dynamics that shape them.

Livehood clusters are built and defined using the following methodology:

1. a geographic distance is computed based on pairs of check-in venues' coordinates;
2. social similarity between each pair of venues is computed using cosine measurements;
3. spectral clustering produces candidate livehoods clusters;
4. interviews are conducted with residents in order to validate the clusters discovered by the algorithm.

Data The data comes from two sources, combining approximately 11 million **Foursquare** (a recommendation site for venues based on users' experiences) check-ins from the dataset of Chen et al. [1] and a new dataset of 7 million Twitter check-ins downloaded between June and December of 2011. For each check-in, the data consists of the user ID, the time, the latitude and longitude, the name of the venue, and its category.

In this case study, livehood clusters from Pittsburgh, Pennsylvania, are examined using 42,787 check-ins of 3840 users at 5349 venues.

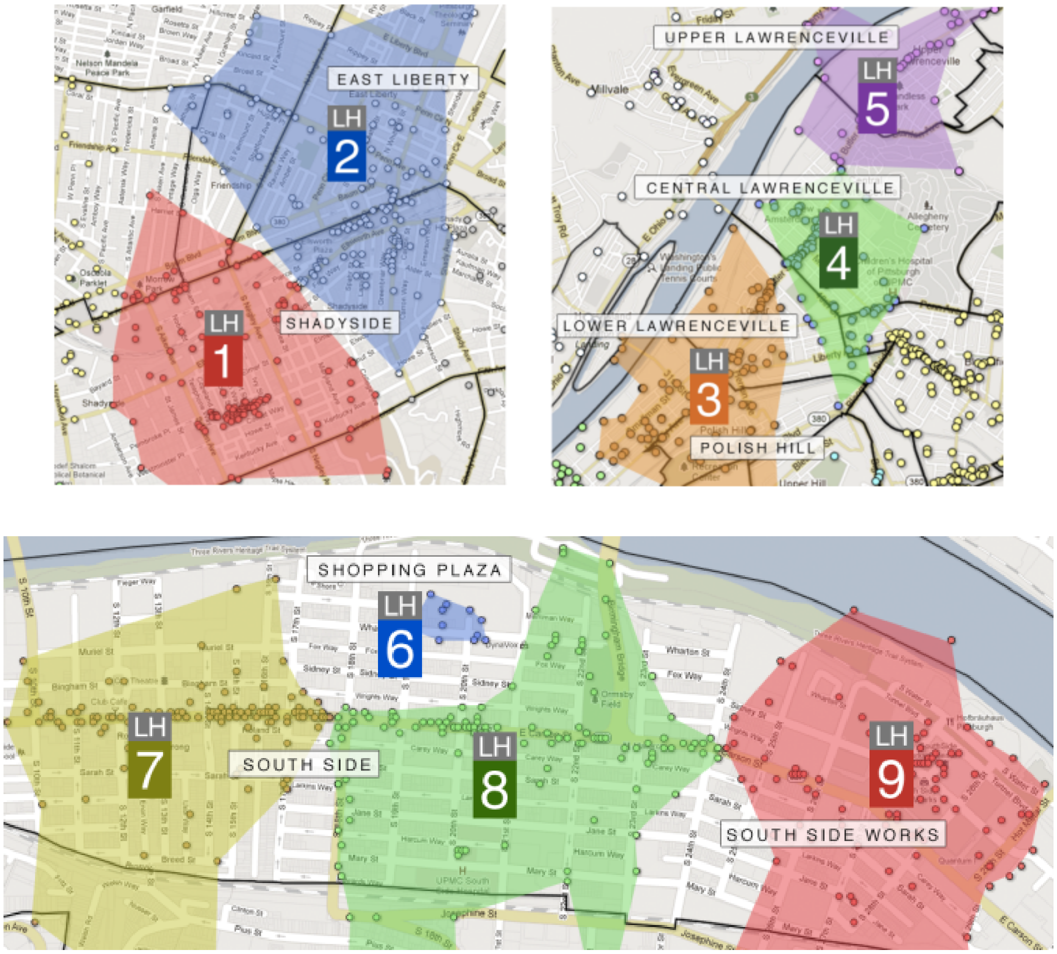


Figure 8. Some livehoods in metropolitan Pittsburgh, PA: in Shadyside/East Liberty, Lawrenceville/Polish Hill, and South Side. Municipal borders are shown in black.

Strengths and Limitations of the Approach

- The technique used in this study is **agnostic** towards the particular source of the data: it is not dependent on meta-knowledge about the data.
- The algorithm may be prone to "majority" bias, consequently misrepresenting or hiding minority behaviours.
- The data are based on a limited sample of check-ins shared on Twitter and are therefore biased towards the types of places that people typically want to share publicly.
- Tuning the clusters is non-trivial: experimenter bias may combine with "confirmation bias" of the interviewees in the validation stage.

Procedures The Livehoods project uses a **spectral clustering model** to provide structure for local urban areas (UAs), grouping close Foursquare venues into clusters based on both the **spatial proximity** between venues and the **social proximity** which is derive from the distribution of people that check-in to them.

The guiding principle of the model is that the "character" of an UA is defined both by the types of venues it contains and by the people frequent them as part of their daily activities. These clusters are referred to as **Livehoods**, by analogy with more traditional neighbourhoods.

Let V be a list of Foursquare venues, A the associated **affinity matrix** representing a measure of similarity between each venue, and $G(A)$ be the graph obtained from the A by linking each venue to its nearest m neighbours. Spectral clustering is implemented by the following algorithm:

1. Compute the diagonal degree matrix $D_{ii} = \sum_j A_{ij}$;
2. Set the Laplacian matrix $L = D - A$ and

$$L_{\text{norm}} = D^{-1/2} L D^{-1/2};$$

3. Find the k smallest eigenvalues of L_{norm}, where k is the index which provides the biggest jump in successive eigenvalues of eigenvalues of L_{norm}, in increasing order;
4. Find the eigenvectors $e_1, \dots e_k$ of L corresponding to the k smallest eigenvalues;

5. Construct the matrix E with the eigenvectors $e_1, ...e_k$ as columns;
6. Denote the rows of E by $y_1, ..., y_n$, and cluster them into k clusters $C_1, ..., C_k$ using k-means. This will induce a clustering $A_1, ..., A_k$ defined by $A_i = \{j | y_j \in C_i\}$.
7. For each A_i, let $G(A_i)$ be the subgraph of $G(A)$ induced by vertex A_i. Split $G(A_i)$ into connected components. Add each component as a new cluster to the list of clusters, and remove the subgraph $G(A_i)$ from the list.
8. Let b be the area of bounding box containing coordinates in the set of venues V, and b_i be the area of the box containing A_i. If $\frac{b_i}{b} > \tau$, delete cluster A_i, and redistribute each of its venues $v \in A_i$ to the closest A_j under the distance measurement.

Results, Evaluation and Validation The parameters used for the clustering were $m = 10$, $k_{\text{min}} = 30$, $k_{\text{max}} = 45$, and $\tau = 0.4$. The results for three areas of the city are shown in Figure 8. In total, 9 livehoods have been identified and validated by 27 Pittsburgh residents (see Figure 8; the original report has more information on the interview process).

- **Municipal Neighborhoods Borders:** livehoods are dynamic, and evolve as people's behaviours change, unlike the fixed neighbourhood borders set by the city government.
- **Demographics:** the interview displayed strong evidence that the demographics of the residents and visitors of an area often play a strong role in explaining the divisions between livehoods.
- **Development and Resources:** economic development can affect the character of an area. Similarly, the resources (or lack there of) provided by a region has a strong influence on the people that visit it, and hence its resulting character. This is assumed to be reflected in the livehoods.
- **Geography and Architecture:** the movements of people through a certain area is presumably shaped by its geography and architecture; livehoods can reveal this influence and the effects it has over visiting patterns.

Take-Away k—means is not the sole clustering algorithm in applications!

References

[1] Chen, Z., Caverlee, J., Lee, K., Su, D.Z. [2011], *Exploring millions of footprints in location sharing services*, ICWSM.

[2] Cranshaw, J., Schwartz, R., Hong, J.I., Sadeh, N. [2012], *The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City*, International AAAI Conference on Weblogs and Social Media, p.58.

4. Association Rules: Danish Medical Data

Title Temporal disease trajectories condensed from population wide registry data covering 6.2 million patients

Authors Anders Boeck Jensen, Pope L. Moseley, Tudor I. Oprea, Sabrina Gade Ellesøe, Robert Eriksson, Henriette Schmock, Peter Bjødstrup Jensen, Lars Juhl Jensen, and Søren Brunak

Date 2014

Sponsor Danish National Patient Registry

Methods association rules mining, clustering

Objective Estimating disease progression (trajectories) from current patient state is a crucial notion in medical studies. Trajectories have so far only been analyzed for a small number of diseases or using large-scale approaches without consideration for time exceeding a few years.

Using data from the Danish National Patient Registry (an extensive, long-term data collection effort by Denmark), this study finds connections between different diagnoses and how the presence of a diagnosis at some point in time might allow for the prediction of another diagnosis at a later point in time.

Methodology The following methodological steps were taken:

1. compute strength of correlation for pairs of diagnoses over a 5 year interval (on a representative subset of the data);
2. test diagnoses pairs for directionality (one diagnosis repeatedly occurring before the other);
3. determine reasonable diagnosis trajectories (thoroughfares) by combining smaller (but frequent) trajectories with overlapping diagnoses;
4. validate the trajectories by comparison with non-Danish data;
5. cluster the thoroughfares to identify a small number of central medical conditions (key diagnoses) around which disease progression is organized.

Data The Danish National Patient Registry is an electronic health registry containing administrative information and diagnoses, covering the whole population of Denmark, including private and public hospital visits of all types: inpatient (overnight stay), outpatient (no overnight stay) and emergency. The data set covers 14.9 years from January '96 to November '10 and consists of 68 million records for 6.2 million patients.

Challenges and Pitfalls

- Access to the National Patient Registry is protected and could only be granted after approval by the Danish Data Registration Agency the National Board of Health.

- Gender-specific differences in diagnostic trends are clearly identifiable (pregnancy and testicular cancer do not have much cross-appeal). But many diagnoses were found to exclusively (or at least, predominantly) be made in different sites (inpatient, outpatient, emergency ward), which suggests the importance of stratifying by site as well as by gender.
- In the process of forming small diagnoses chains, it became necessary to compute the correlation using large groups for each pair of diagnoses. To compensate for multiple testing for close to 1 millions pairs and obtain a significant p—value, more than 80 million samples would have been required for each pair. This would have translated to a few thousand years' worth of computer running time. In order to avoid this pitfall, a pre-filtering step was included. Pairs included in the trajectories were eventually validated using the full sampling procedure, however.

Project Summaries and Results The dataset was reduced to 1,171 significant trajectories. These thoroughfares were clustered into patterns centred on 5 key diagnoses central to disease progression: *diabetes*, *chronic obstructive pulmonary disease* (COPD), *cancer*, *arthritis*, and *cerebrovascular disease*.

Early diagnoses for these central factors can help reduce the risk of adverse outcome linked to future diagnoses of other conditions. Three author quotes illustrate the importance of these results:

> The research could yield tangible health benefits as we move beyond one-size-fits-all medicine.
> — L.J.Jensen

> The sooner a health risk pattern is identified, the better we can prevent and treat critical diseases.
> — S.Brunak

> Instead of looking at each disease in isolation, you can talk about a complex system with many different interacting factors. By looking at the order in which different diseases appear, you can start to draw patterns and see complex correlations outlining the direction for each individual person.
> — L.J.Jensen

Among the specific results, the following "surprising" insights were found:

- a diagnosis of anemia is typically followed months later by the discovery of colon cancer;
- gout was identified as a step on the path toward cardiovascular disease, and
- COPD is under-diagnosed and under-treated.

The disease trajectories clusters for two key diagnoses are shown in Figure 9.

References

[1] Jensen, A.B., Moseley, P.L., Oprea, T.I., Ellesoe, S.G., Eriksson, R., Schmock, H., Jensen, P.B., Jensen, L.J., Brunak, S. [2014], Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients, *Nature Communications*.

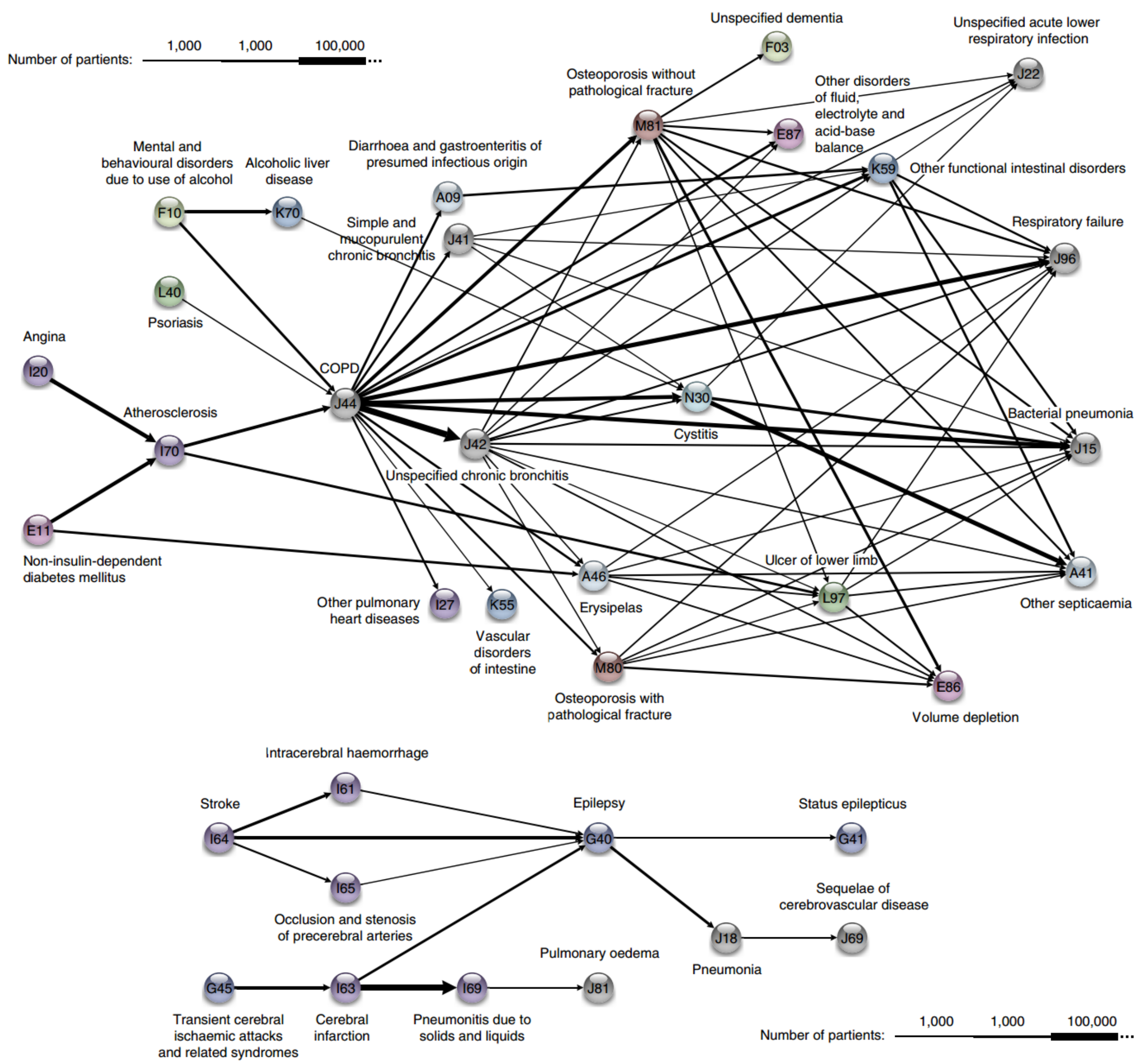


Figure 9. The COPD cluster showing five preceding diagnoses leading to COPD and some of the possible outcomes; Cerebrovascular cluster with epilepsy as key diagnosis.