Mapping the Shape of the U.S. Economy: A Topological Data Analysis Approach with BallMapper

Ryan Johnson[*]

[?] Department of Mathematics, University of Alaska Anchorage, Anchorage, AK

Student: *johnson.ryan1019@gmail.com*[*]
Mentor: *scook25@alaska.edu*

KEYWORDS

Topological Data Analysis; BallMapper; Data Science; Economics; Macroeconomics; Topology;

ABSTRACT

Topological Data Analysis is (TDA) is a new data analysis method which gained popularity starting in the early 21st century. Currently, a large body of TDA research utilizes the traditional Mapper algorithm. We aim to fill an entry level gap into TDA as well as expand the body of literature on BallMapper, a new, Mapper adjacent algorithm. Using widely used U.S. macroeconomic data from the Bureau of Economic Analysis (BEA), the Federal Reserve Bank (FRB), and Bureau of Labor Statistics (BLS), we do this is three parts. We show an example of BallMapper's utility in exploratory data analysis which also provides an example of how to interpret BallMapper's output; analyze our topological graphs to notable historic economic events; and test the stability of BallMapper's output and the topological graph's features. Results show. [...]

INTRODUCTION

Topological Data Analysis (TDA) is a new and emerging field of data analysis that is increasing in popularity. Broadly, traditional TDA applications use two tools: an algorithm called Mapper that produces a network-like graph and and analysis technique called Persistence Homology. Mapper can be thought of the exploratory data analysis (EDA) and visual display of statistical analysis. That is, we may see patterns in our data or need a visual tool to help explain the results of an analysis. Persistence Homology on the other hand is more akin to statistical models like the Generalized Linear Model or Bayesian Statistics. They provide a mathematical framework and used as a tool to give a researcher confidence in where to look further or what they might be seeing in their data. Moreover, the combination of Mapper and Persistence Homology is what forms the central argument for TDA: data contains an underlying shape.[1],[2]

For this analysis we will be focusing on the graph creation portion of TDA. Specifically, we are examining various macroeconomic indicators of the United States of America (U.S.) using a Mapper-adjacent algorithm called BallMapper(BM).[3] BallMapper is of particular interest to us because it significantly reduces the parameters to create a topological map. It simplifies traditional Mapper by reducing the need for the user to pass data through multiple functions, each with their own parameters.[4] One benefit to this reduction in parameters is that it removes some of the barriers to learning Mapper (or Mapper-like algorithms), and more generally, TDA. Those who have a mathematical background, introductory course in Topology, or conceptual knowledge of data science methods and algorithms will fare much easier. However, if you are not equipped with any of the aforementioned tools, TDA might seem unnecessarily complicated for data analysis.

Beyond the main objectives of this analysis, we hope that this paper is serve as an on-ramp for anyone interested in TDA but feels inundated with jargon upon early stage researching. Out objectives are to add to the small body of literature whose main focus is applications with BallMapper.[5] We go about this by first showing BallMapper's use in exploratory data analysis (EDA). Using our

[1] https://arxiv.org/pdf/2504.09042
[2] Chazal F and Michel B (2021) An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists. Front. Artif. Intell. 4:667963. doi: 10.3389/frai.2021.667963
[3] https://arxiv.org/pdf/1901.07410
[4] add reference
[5] add Mapper survey paper refrence here

results from EDA, we then will compare BallMapper's outputs (topological graphs) to notable historic economic events [and maybe less known?]. Lastly we will test the stability and consistency of BallMapper's output. The conclusion will talk broadly about results and other considerations of note, and the discussion will elaborate on future work and questions for research.

and is increasingly being applied to From the small body of literature we have reviewed, TDABM has removed many of the parameters that are required for more traditional methods of TDA. Generally speaking, traditional TDA contains four broad steps, each with multiple user-defined parameters. Conversely, TDABM has reduced this process to an the user selecting their data, a coloring variable and an epsilon value – details on this follow below. What should be noted, although TDABM reduces the number of steps needed to produce similar outcomes of traditional TDA methods, we lose control of being able fine tune out outputs. We also find that interpreting results becomes more difficult. However, TDABM is still in its infancy, so there is not a large body of research on interpretation. (Dłotko 2019) One major motivation for this paper is a paper written by Dłotko, Rudkin, and Qiu (2019) that applied TDABM to a global macroeconomic dataset to compare how countries have evolved over time, their transformation from the Great Depression Era, and various views on wealth and inequality. We could not find other TDA literature found specifically focused on the macroeconomic economy of singular countries. Hence, our topic of choice.

METHODS AND PROCEDURES

Data Selection & Preparation

This paper relied on three publicly available data from US government sources. The Bureau of Economic Analysis (BEA), The Federal Reserve Board (FRB), and The Bureau of Labor Statistics (BLS). The data were gathered using R using two application programming interfaces (APIs): one for data from the BEA, and the other from the Federal Reserve Economic Data (FRED) API. FRED aggregates data from national and international sources, as well as public and private sources. We additionally used recession dates based on business cycle contractions and expansions provided by the National Bureau of Economic Research (NBER). These agencies were selected because of they are authoritative sources for U.S. economic data. Their widespread use in both the private and public sector gives us high confidence in the accuracy and integrity of the data.[6]

Data Series	Series Abbreviation	Year Start	Year End	Data Frequency	Data Type	API
Gross Domestic Product	GDP	1930	2024	Annual	Flow	BEA
Personal Income & Its Disposition	PID	1948	2024	Annual	Flow	BEA
Foreign Direct Investment	FDI	2014	2024	Annual	Flow	BEA
Federal Funds Rate	FFR	1955	2024	Annual	Rate	FRED
Employment Cost Index	ECI	2001	2025	Quarterly	Index	FRED
Consumer Price Index	CPI	1947	2025	Monthly	Index	FRED
New Privately-Owned Housing Units Started	Housing Starts	1959	2025	Monthly	Flow*	FRED
Producer Price Index - All Commodities	PPI	1913	2025	Monthly	Index	FRED
Unemployment Rate	Unrate	1948	2025	Monthly	Rate	FRED

Table 1: *Housing Starts is considered a flow because it is provided as a Seasonally Adjusted Annual Rate (SAAR).

Table 1 shows all the data series we considered for this analysis. We excluded Employment Cost Index (ECI) and Foreign Direct Investment (FDI) due to their limited availability of years. Of the remaining data series, New Privately-Owned Housing Units Started (Housing Starts) has the smallest range and will provide the lower end for the years used in our analysis (1960-2024). Table 2 shows the final data series we will be using. These specific series were chosen because they represent different aspects of the macroeconomy, and are described in *Functional Description*. Additionally, we are focusing on an annual time frame for this analysis so some data transformation was needed. The final data series transformations follow.

[6]Hughes-Cromwick, Ellen, and Julia Coronado. 2019. "The Value of US Government Data to US Business Decisions." Journal of Economic Perspectives 33 (1): 131201346. DOI: 10.1257/jep.33.1.131

Series Abbreviation	Economic Role	Business Cycle Timing	Functional Descriptioin	Transformation Applied
GDP	Growth Composition	Concurrent	Consumption, Investment, Gov., and Trade	None (Source in % Change)
PID	Income Structure	Concurrent	Wages, Entrepreneurship, Gov. Transfers	Real Adjustment & Log Difference
UnRate	Labor Market Dynamics	Lagging	Labor Distress Level and Trend	None (Annual Rate) & Simple Difference
FFR	Monetary Policy	Policy/Reactive	Cost of Capital and Borrowing Conditions	None (Annual Rate) & Simple Difference
PPI	Supply-Side Signal	Leading	Producer Costs & Measures Cost-Push Inflation	12-month Avg. & Log Difference
CPI	Demand-Side Signal	Lagging	Cost of Living & Measures Demand-Pull Inflation	12-month Avg. & Log Difference
Housing Starts	Leading Indicator	Leading	Physical Residential Production	12-month Avg. & Log Difference

Table 2: All years for analysis are 1960-2024.

The first transformation was Personal Income and Its Disposition (Personal Income).[7] Personal Income is reported in nominal dollars, so it does not account for inflation. Thus, we first adjusted it using the Consumer Price Index (CPI) to get Personal Income into Real Dollars.[8]

$$\text{Real Dollars} = \frac{\text{Nominal Dollars}}{\text{CPI}} * 100 \tag{1}$$

We then took the the log difference (Equation 2) from the output from Equation 1 to get a percentage change from the preceding year. This log difference helps to make our data more linear which aides BallMapper's use of the standard Euclidean Distance between each row of data. [If BM did not use the Euclidian distance between points, we might not need to take log differences. However topic is for another paper.] [9] [10]

$$\Delta \ln(\text{Level}) = [\ln(\text{Level}_t) - \ln(\text{Level}_{t-1})] * 100 \tag{2}$$

Our next transformation was on Housing Starts, Producer Price Index - All Commodities (PPI)[11], and CPI. All three of these sources were only provided on a monthly time frame, so to get an annual value we used a simple arithmetic mean. Once these data are in annual form, we then found the percent change from the previous year using Equation 2 for their final data in our analysis.

[Need to add parts more about the data itself?]

METHODS

BallMapper

In depth descriptions of Ball Mapper's theory have been covered in various papers. As the aim of our paper is to provided a introductory foray into BallMapper and TDA in general, we will omit some of the more technical details below but provide reference to more in-depth explanations.

Mathematically, the concept of Ball Mapper supposes we are given a dataset X in K dimensions with N observations. Then for some point $x \in X$ and given $\epsilon > 0$, we create a ball, $b(x, \epsilon)$, centered around x with radius ϵ. Our aim to create set of balls, B, such that $B = \bigcup_{i=0}^{n} b(x, \epsilon)$ for all $x \in B$.

In practice, though, we call our dataset X a Point Cloud with N rows. Our dimensions, K, are each column of our data and usually our point cloud has dimensions $K > 2$. When we are selecting a point, $x_i \in X$, to draw a ball, we are randomly selecting a row of our point

[7] Table 2.1 from the BEA's Interactive Data, National Data section.
[8] https://www.dallasfed.org/research/basics/nominal
[9] https://econbrowser.com/archives/2014/02/use-of-logarithms-in-economics
[10] https://www.numberanalytics.com/blog/log-transform-econ-how-to-guide
[11] PPI tracks only physical goods such as Farm Products and Feed, Industrial Commodidities, and Other Commodidites such as Aircraft, Steel, and Lumber.

cloud and then taking he Euclidean Distance in K dimensions such that $i \neq j$ and $x_i, x_j \in X$ to every other x_j in our point cloud. Then with our given epsilon, $\epsilon > 0$, we assign any points where the distance is less to epsilon to a ball with center x_i (Equation 3).

$$d(x_i, x_j) = \sqrt{(x_{i_1} - x_{j_1})^2 + (x_{i_2} - x_{j_2})^2 + ... + (x_{i_k} - x_{j_k})^2} < \epsilon \tag{3}$$

Algorithmically, Ball Mapper is as follows:

1. Select a random point x_i from point cloud X.
2. Given $\epsilon > 0$, construct a ball $b(x_i, \epsilon)$ with a center x_i by associating all other points where $d(x_i, x_j) < \epsilon$ and $i \neq j$.[12]
3. Place $b_i, \epsilon)$ in a set of balls B.
4. Repeat steps 1-3 until all $x \in X$ belong to some $b(x, \epsilon)$.
5. Draw an edge between $b_i, b_j \in B$ if they contain the same x, weighting the edge based on number distinct x values b_i and b_j both contain.

Analysis

As we saw above we need three parts for BallMapper: a Point Cloud, a variable to color by, and an epsilon. For this analysis, our point cloud was made up of 16 variables (dimensions) which can be see in Table 3. It is important to note two dimensions in our analysis were created: *Unemployment Change* and *Fed Rate Change*. All other variables where transformed but no new data created from them. Albeit, Unemployment is a lagging indicator and the Federal Funds Rate is a policy or reactive choice, we thought these two measures were important. Since Unemployment is a flow, is it useful to see if the hose of unemployed is expanding or shrinking. Similarly, it is informative to see changes in the Federal Funds Rate to see the policy choices made are working or not.

Data Series	Analysis Name
All	Year
Producer Price Index - All Commodities	PPI Change
Personal Income & Its Disposition	Personal Income
Personal Income & Its Disposition	Compensation
Personal Income & Its Disposition	Entrepreneurship
Personal Income & Its Disposition	Transfers
Gross Domestic Product	Consumption
Gross Domestic Product	Domestic Investment
Gross Domestic Product	Government Spending
Gross Domestic Product	Exports
Gross Domestic Product	Imports
Consumer Price Index	Inflation
Unemployment Rate	Unemployment
Unemployment Rate	Unemployment Change
Federal Funds Rate	Fed Rate
Federal Funds Rate	Fed Rate Change
New Privately-Owned Housing Units Started	Housing Change

Table 3: Variables inculde in BallMapper Point Cloud

To our knowledge there is no research about finding epsilon values for BallMapper. So to find a BM graph which represents our data well we go about finding an epsilon in an iterative manner. We first start by finding the upper and lower bounds of our map, trying small and large epsilon values, respectively. What we look for in our lower bound is a map such that the epsilon is small enough to make most all, or all depending on the dataset size, data points singletons, giving an image of many disconnected points. Conversely, our upper bound is an epsilon so large that it includes most every point and produces a map of two or less nodes. With this idea of finding a lower and upper bound, in an iterative manner aim to create around 100 maps at even intervals, review the maps, continue to narrow our interval to a map that shows our data "in focus", and presents what we might think is a good or interesting representation of our data.

[12] See @eq-euclideanDistance for $d(x_i, x_j)$

For our data we started off by using an interval of $[0.3 - 1.4]$ and then narrowed our interval to $[0.4 - 0.75]$. When reviewing our maps, we look for connected components (connected nodes) appearing or disappearing, significant changes in coloration in the graph, or the any patterns in size or color of the nodes. We finally decided on the value 0.511 because it presented a little of all the aforementioned features we look for: coloration and size patterning and connected components.

For our initial map we used *Year* as our coloring variable. This is because Year is a metric that is not dependent on any economic variables and is a good marker for large macroeconomic events such as the dot com bubble of the 2000s. *Year* is not included in the construction of our Point Cloud. This is an important point because we show later that being able to create metrics to color by adds a depth the Ball Mapper's usefulness, even if the metrics themselves are not included in our Point Cloud.

Instead of using the graph output provided by BallMapper, we created graph output using other R packages to bring clarity to our analysis. We have not changed the core structure of the BallMapper output, but rather the presentation. The largest changes of note to the graphs are: spread out nodes to avoid overlapping; display the edge weights between nodes, representing the number number of shared values between two nodes; custom color scale for accessibility.

CONCLUSION

Exploratory Data Analysis (EDA) Application

For our exploratory data analysis exploration we colored our graph by the *Year*, Figure 1. The coloration of each node represents the average of all the years within in a node. The size of our node is commensurate with the number of years it houses. Between two nodes, th edge thickness is determined the number of same data points in each endpoint.Looking at Figure 1, we first observe the overall structure of our graph. There is one large connected component, one small component, and the rest singleton nodes.[13] For the largest connected component (C1), we immediately are interested in the three large nodes (4, 24, and 26) because they are the largest in size for C1 and the same for their edge thickness (edge strength). Additionally, we also notice that each of the three large nodes have at least one flare (node connected only to themselves) as well as sharing others. What this structure tells us is, there are three distinct groups and each flare representing similar data to that specific group but perhaps is distinctly different to the other flares.

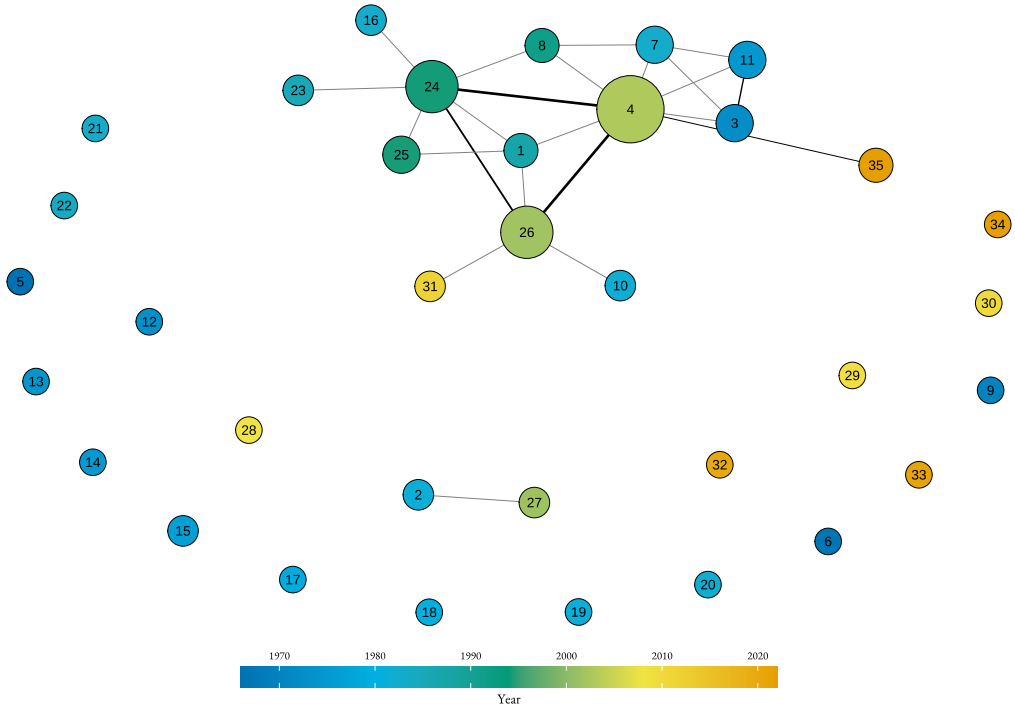

Figure 1: Ball Mapper Graph colored by Year with Epsilon set to 0.511.

[13] A Connected Component is a subset of nodes in our graph such that each point in the subset has a path to all other points in the subset.

The small component (C2), nodes 2 and 27, are of interest because it tells us us that these related nodes are somehow distinctly different from C1 and all other nodes. Seeing a small components like this sometimes can indicate outliers in data. In our case, this could be years where large economic events happened such as a recessions or extraordinary growth. However, we also can see this behavior in singleton nodes much like we have Figure 1. When we look at the coloration of the singleton nodes, we see indication of outlier events such as The Great Recession and the COVID-19 Pandemic.

Focusing out attention back to C1, we look closer

Small Economic Analysis

Node 26 in the large component seem to represent recovery years after big economic shocks. Some years were "Non-Labor Recovery" years but generally things were working as they were suppose to.

Node 31 seems to be a variation of recovery because rates where so low already the gov was a bit "cornered" with monetary policy tools.

Node 10 seems to be another variation on reconvery because there was a lot of economic stimulus and help. Lots of wanting to recover provided by the government.

Node 24 was "pressure cooker" economy. Everything firing on all cylinders.

Small Comparative Analysis

- "New Deal"
- 08 financial crisis
- Covid
- Flash Crash 87
- Tech Boom
- AI Boom
- Oil Embargo/ high unemployment

Robustness and Stabilization

- Shuffle Dataset –> Run BallMapper
  - Look at:
    *

      ## nodes

      * coloring consistency
      * look at financne application and brexit vote

DISCUSSION

REFERENCES

SUPPLEMENTARY

Year	PPI_Change	Personal_Income	Compensation	Entrepreneurship	Transfers	GDP	Consumption
1963	-0.23724804	3.731312	4.087260	0.8372622	4.5584131	4.4	4.1
1964	0.18457487	5.627684	5.729020	3.4204277	2.6284659	5.8	6.0
1994	1.28818409	2.407677	2.267551	3.9193375	1.8741264	4.0	3.9
1995	3.51438036	2.837265	1.912013	2.4681910	3.5996948	2.7	2.9
1996	2.31747076	3.020134	2.187463	9.3354433	2.1543086	3.8	3.5
1997	-0.06528694	3.626595	4.088269	4.8170322	0.4135061	4.4	3.8
1999	0.83364383	2.845324	4.163589	6.2261586	2.0363770	4.8	5.4
2003	5.20715602	1.289027	1.227378	0.8257153	2.4423212	2.8	3.2
2004	6.00601825	2.847558	2.966330	4.3995092	2.7135075	3.8	3.8
2005	7.05710812	1.848092	1.719347	-1.6373993	3.1929878	3.5	3.5
2006	4.56388374	3.751035	2.516709	3.8985395	3.0340329	2.8	2.9
2014	0.93380498	3.356754	2.989498	-0.2632255	3.1185974	2.5	2.8
2015	-7.52075668	4.438330	4.620931	-1.7580072	5.3847529	2.9	3.4
2016	-2.69651355	1.380372	1.453018	-1.1487527	2.0952864	1.8	2.5
2017	4.30689625	2.654343	2.386592	3.6106069	0.6842140	2.5	2.6
2018	4.25703791	2.653720	2.577156	2.1531535	1.7263802	3.0	2.7
2019	-1.04947557	2.856578	2.570816	2.1656075	3.7838706	2.6	2.1
2024	-0.45876015	2.540428	2.700120	1.1835512	3.8735863	2.8	2.9

Table 4: Component C1, Node 4 Members

Year	PPI_Change	Personal_Income	Compensation	Entrepreneurship	Transfers	GDP	Consumption
1986	-2.93487742	3.819274	4.355013	4.278550864	4.04803495	3.5	4.1
1987	2.60309480	2.102181	3.338292	7.542323766	0.09155415	3.5	3.4
1988	3.94171012	4.086906	3.942467	8.762714220	2.05634064	4.2	4.2
1989	4.83707704	2.985721	1.617455	0.002884136	4.30784088	3.7	2.9
1994	1.28818409	2.407677	2.267551	3.919337479	1.87412645	4.0	3.9
1995	3.51438036	2.837265	1.912013	2.468190951	3.59969483	2.7	2.9
1996	2.31747076	3.020134	2.187463	9.335443284	2.15430862	3.8	3.5
1997	-0.06528694	3.626595	4.088269	4.817032211	0.41350613	4.4	3.8
2004	6.00601825	2.847558	2.966330	4.399509244	2.71350749	3.8	3.8
2006	4.56388374	3.751035	2.516709	3.898539532	3.03403293	2.8	2.9

Table 5: Component C1, Node 24 Members

Year	PPI_Change	Personal_Income	Compensation	Entrepreneurship	Transfers	GDP	Consumption
1992	0.5633416	3.2409853	3.122708	9.2134780	8.298746	3.5	3.7
1993	1.4542117	1.3265049	1.056303	3.7969739	2.909281	2.7	3.5
1994	1.2881841	2.4076768	2.267551	3.9193375	1.874126	4.0	3.9
1995	3.5143804	2.8372652	1.912013	2.4681910	3.599695	2.7	2.9
1996	2.3174708	3.0201336	2.187463	9.3354433	2.154309	3.8	3.5
2003	5.2071560	1.2890269	1.227378	0.8257153	2.442321	2.8	3.2
2004	6.0060183	2.8475578	2.966330	4.3995092	2.713507	3.8	3.8
2012	0.5497585	2.4164836	2.011891	3.6123112	-1.868500	2.3	1.4
2013	0.6164234	-0.3764019	1.621224	2.4534348	1.107013	2.1	1.7
2014	0.9338050	3.3567538	2.989498	-0.2632255	3.118597	2.5	2.8

Table 6: Component C1, Node 26 Members

Old Paper

Discussion

This paper serves as a proof-of-concept of the usefulness of TDABM as a methodology for exploratory data analysis. By no means is this a totally comprehensive method to gain insight into data, but instead TDA BallMapper proves itself to be useful tool wehen employed in addition to other traditional analyses.

In our section we briefly went over two ways to interpret TDABM graphs. As seen in many other, longer papers written by Dlotko and colleagues, there is a lot of room to expound our TDABM graph here,though that is beyond the scope of this paper. In general, though, we can see that TDBM can give insight into our data that we might not see otherwise. Seeing that there is data that is not like the others gives an indication that there is something to investigate. Consider our comparison of the 08' - 09' financial crisis and the 2020 COVID Pandemic: We see learned that the years following the financial crisis were similar economically. Another way to view this is, clustering revels that the financial crisis was a singular problem that spurred the collapse of the financial system. Conversely, the COVID pandemic was a singular issue but exposed many different weaknesses in our current world.

For the future development of TDABM, and TDA in general, there are many possibilities of future research and one glaring downside. The downside to TDA is that it is a high barrier to entry to understand the methodologies inner workings. Those with a technical background will have an easier time, but nonetheless a higher barrier than methods such as linear regressions. Subject wise, TDA has an advantage in some of the social sciences because it can be somewhat of a bridge between qualitative and quantitative analyses. TDABM has been used to analyze the Brexit Vote data, seeking to understand quantitative and qualitative motivations behind its outcome Rudkin et al. (2023). TDA also has the advantages of dimension reduction which could be useful to the life-sciences. Fields such as genetics, chemistry, and biology have many fields of data, and being able to consolidate it into an understandable form could benefit the body of knowledge greatly.

Data Source

- Bureau of Economic Analysis
  - Real Exports and Imports
  - New Foreign Direct Investment in the United States
  - Real GDP
  - Personal Income and Its Disposition
- Federal Reserve Bank
  - Selected Interests Rates

- Bureau of Labor Statistics
  - CPI
  - Employment Cost Index
  - PPI
  - Housing Starts
  - Unemployment Rate

Dłotko, Paweł. 2019. "Ball Mapper: A Shape Summary for Topological Data Analysis." https://arxiv.org/abs/1901.07410.

Dłotko, Paweł, Simon Rudkin, and Wanling Qiu. 2019. "Topologically Mapping the Macroeconomy." https://arxiv.org/abs/1911.10476.

Rudkin, Simon, Lucy Barros, Paweł Dłotko, and Wanling Qiu. 2023. "An Economic Topology of the Brexit Vote." *Regional Studies* 58 (3): 601–18. https://doi.org/10.1080/00343404.2023.2204123.