Stanford
Computational
Antitrust

ARTICLE

Gleaning Insight from Antitrust Cases Using Machine Learning

Giovanna Massarotto* & Ashwin Ittoo**

Abstract. The application of AI and Machine Learning (ML) techniques is becoming a primary issue of investigation in the legal and regulatory domains. Antitrust agencies are in the spotlight because they represent the first arm of government regulation in that they reach new markets before Congress has had time to draft a more specific regulatory scheme. A question the antitrust community is asking is whether antitrust agencies are equipped with the appropriate tools and powers to face today's increasingly dynamic markets. Our study aims to tackle this question by building and testing an antitrust machine learning (AML) application based on an unsupervised approach, devoid of any human intervention. It shows how a relatively simple algorithm can, in an autonomous manner, discover underlying patterns from past antitrust cases by computing the similarity between these cases based on their measurable characteristics. Our results, achieved with simple algorithms, show much promise from the use of AI for antitrust applications. AI, in its current form, cannot replace antitrust agencies such as the FTC. Instead, it is a valuable tool that antitrust agencies can exploit for efficiency, with the potential to aid in preliminary screening, analysis of cases, or ultimate decision-making. Our contribution aims to pave the way for future AI applications in market regulation, starting with antitrust regulation. Government adoption of emerging technologies, such as AI, appears to be key for ensuring consumer welfare and market efficiency in the age of AI and big data.

* Adjunct Professor University of Iowa, Research Associate UCL CBT.
** Associate Professor, University of Liege.

I. Introduction

Big data has become a game-changer, and Artificial Intelligence (AI) models the best way to fully exploit such large amounts of data. This paper builds on the growing interest in the application of Machine Learning (ML) techniques to the legal and regulatory domains. The main innovation of our research is that it explores the application of AI to antitrust enforcement.

In recent years, the ability of AI to generate anticompetitive behavior, such as the phenomenon of algorithmic collusion and price discrimination, has been deeply investigated.[1] In contrast to these earlier studies of AI and antitrust, we adopt a different perspective. In particular, we are concerned with whether AI can assist antitrust enforcers in addressing the critical need for both accelerating and harmonizing globally the enforcement of antitrust principles.[2]

The main question our paper aims to answer is: *Can AI be usefully employed by antitrust enforcement agencies?* This overarching question can be decomposed into a number of sub-questions:

> 1) Are there pertinent characteristics that can be extracted from past antitrust cases?
> 2) Are there underlying patterns characterizing antitrust cases?
> 3) Are the issues surrounding the adoption of AI in other branches of law, such as bias in criminal law, relevant to antitrust? In other words, is antitrust a better testing ground for the adoption of AI in the legal domain?

To the best of our knowledge, our study is the first to address these questions. They are at the heart of the Assistant Attorney General's recent speech emphasizing the need to understand the potential of cutting-edge technologies like AI and ML to advance the antitrust field.[3]

To address our research questions, we first created a dataset of seventy-two past antitrust cases, spanning from 2005 to 2019. Because antitrust law is jurisdiction-based, we focused on the U.S. antitrust jurisdiction, which includes two antitrust agencies: the Department of Justice (DOJ) Antitrust Division and the Federal Trade Commission (FTC). These two agencies have very different powers and structure. In a first analysis, no relevant patterns were detected when using the combined decisions of both agencies. For this reason, we focused only on the FTC enforcement action in the selection of cases in our dataset. Each case is described

[1] *See, e.g.*, Axel Gautier et al., *AI Algorithms, Price Discrimination and Collusion: A Technological, Economic and Legal Perspective*, 50 EUR. J. L. ECON. 405 (2020); Ulrich Schwalbe, *Algorithms, Machine Learning, and Collusion*, 14 J. COMPETITION L. & ECON. 568 (2019); Ariel Ezrachi & Maurice E. Stucke, *Sustainable and Unchallenged Algorithmic Tacit Collusion*, 17 NW. J. TECH. & INTELL. PROP. 217 (2020); Emilio Calvano, Giacomo Calzolari, Vincenzo Denicolò & Sergio Pastorello*, Algorithmic Pricing: What Implications for Competition Policy?*, 55 REV. INDUS. ORG. 155 (2019).

[2] *See, e.g.*, Eleanor M. Fox, *The End of Antitrust Isolationism: The Vision of One World*, 1992 U. CHI. LEGAL F. 221 (1992); Douglas H. Ginsburg, *International Antitrust: 2000 and Beyond*, 68 ANTITRUST L. J. 571 (2000).

[3] Assistant Attorney General Makan Delrahim, Remarks at the Thirteenth Annual Conference on Innovation Economics (Aug. 27, 2020), available at https://www.justice.gov/opa/speech/assistant-attorney-general-makan-delrahim-delivers-remarks-thirteenth-annual-conference.

along sixteen variables (or features), including, among others, the Industry, Type of Conduct, or Behavioral Remedies. The complete list is presented in Part III.

Then, we developed a machine learning pipeline to automatically analyze these cases, compute the similarity between them, and cluster similar cases together into well-formed, coherent groups. An important characteristic of our machine learning pipeline is that it relies on unsupervised learning (UL) algorithms, namely clustering methods, including K-Means, Bisecting K-Means and K-Modes (see Part II.B). Unlike their more popular supervised counterparts, unsupervised approaches operate with minimal human monitoring and intervention. In addition, we performed extensive analysis of each cluster of similar cases to determine which variables best characterized the various antitrust cases. This was achieved with two supervised learning algorithms, namely Random Forest and Support Vector Machines (see Part III.B).

We proceeded our investigation by interpreting the algorithms' results from an antitrust point of view. For example, we noted that cases from the data and computer industry were generally clustered with those in the healthcare industry, suggesting that these industries raise similar antitrust concerns. The algorithm also clustered cases in which conspiracy, the most commonly detected conduct, is strictly related to exchange of information, and as a consequence, "limitation in the exchange of information," one of the most common remedies, was recommended (see Part IV).

In summary, as our results suggest, AI and computational antitrust techniques in general can serve as a useful tool to assist competition agencies in enforcing antitrust law in the age of big data and AI. More specifically, our study investigates how ML techniques can be used to automatically discover insights from past antitrust decisions and extract underlying recurrent patterns from these decisions, as revealed in Part IV.

The idea for this paper stems from Giovanna Massarotto's book *Antitrust Settlements: How a Simple Agreement Can Drive the Economy*,[4] which analyzed a relatively large volume of antitrust cases in the primary antitrust jurisdictions (the US and EU). The book put forward the idea of an ML algorithm, trained on previous antitrust cases and used for assisting antitrust enforcers in regulating markets. In this paper, we go one step further by implementing and evaluating the algorithm trained on the FTC's cases and practices from 2005 to 2019.

This paper is structured in five Parts. Having introduced the project and aim of our paper, Part II serves as a background of 1) the FTC's role and powers within the U.S. antitrust law enforcement framework; 2) AI and AI methods; and 3) AI and antitrust.

[4] GIOVANNA MASSAROTTO, ANTITRUST SETTLEMENTS: HOW A SIMPLE AGREEMENT CAN DRIVE THE ECONOMY (2019).

Part III investigates the methodology used in the development of our AI model in the context of antitrust. Part III also describes how the AI model and data to build the model were selected.

Part IV analyzes the results of our AML model by assessing the adopted variables and their importance from an antitrust perspective. More specifically, we evaluate whether the variables detected by our algorithm as statistically relevant make sense. A similar analysis is conducted with respect to the clusters detected by our UL algorithm.

Part V ends with some final remarks. Our main argument is that AI cannot replace antitrust agencies such as the FTC, but it could be a valuable tool in making the work of antitrust agencies more efficient and effective in today's fast-moving technological environment.

II. Grasping the Meaning of Antitrust & AI

The 2020s have seen vast increases in investment and interest in AI and the data industry. Data is creating a variety of new opportunities for businesses, transforming markets with faster and more sophisticated technologies, including machine learning algorithms. These new data-driven markets, in turn, create new challenges for government agencies.[5]

Antitrust agencies are in the spotlight because they are charged with identifying and reducing monopolistic and collusive practices in cutting edge markets, where such practices might occur at scale through algorithms (e.g. algorithmic collusion).[6] A question the antitrust community is asking is whether antitrust agencies are equipped with the right tools and powers to tackle the present challenges in such a fast-moving technological environment.[7] Our study aims to respond to this question by building and testing an ML antitrust algorithm.

Before diving into the explanation of our ML antitrust algorithm developed in Part III, it is helpful to clarify why we focused on FTC enforcement actions, and why antitrust can be a good testing ground for future AI applications in the regulatory domain. This evaluation requires having a brief background on the role of antitrust economic regulation and the main AI techniques available.

[5] Bruce Schneier, *Click Here to Kill Everybody* Talks, YOUTUBE (Oct. 11, 2018), https://www.youtube.com/watch?v=GkJCI3_jbtg.

[6] Giovanna Massarotto, *From Standard Oil to Google: How the Role of Antitrust Law Has Changed*, 41 WORLD COMPETITION 395 (2018); MASSAROTTO, *supra* note 4 at 145.

[7] *See, e.g.*, Ariel Ezrachi & Maurice E. Stucke, *Artificial Intelligence & Collusion: When Computers Inhibit Competition*, 2017 U. ILL. L. REV. 1775 (2017).

A – Antitrust

1. Antitrust Economic Regulation

U.S. antitrust enforcement action is primarily economic in nature because it occurs mostly outside of courts and it is explicitly grounded in economics.[8] In the U.S., "over the last three decades the Agencies [DOJ and FTC] have resolved nearly their entire civil enforcement docket by consent decrees."[9] Specifically, more than ninety percent of civil antitrust lawsuits filed by the U.S. government (excluding mergers) are settled by means of an agreement.[10] This consent solution puts in place remedies agreed on by the company under investigation and the agency before or during a trial. The wide adoption of consent decrees results in a regime of minimal case law,[11] leaving the same antitrust agencies and companies to regulate markets through agreed remedies enshrined in a consent decision.[12] Unlike the DOJ, the FTC can settle proceedings without the need for adjudication by a court.[13] Consent decrees identify certain behavioral or structural remedies based on economic analysis, which *de facto* imposes an economic regulatory regime.[14]

Because antitrust agencies are empowered to enforce competition principles in any market, antitrust is often the first type of regulation to reach a new market.[15] For example, while the Federal Communication Commission (FCC) has the authority and duty to regulate the telecommunications industry specifically,[16] the FTC, through Section 5 of the FTC Act, may exercise wide discretion in regulating markets generally.[17] As a result, antitrust may be considered the "first arm" of government regulation because it reaches new markets for which Congress has yet to draft a more specific regulatory scheme.

[8] *See* Barry E. Hawk, *System Failure: Vertical Restraints and EC Competition Law*, 32 COMMON MARKET L. REV. 973, 986 (1995). According to Barry E. Hawk, "economics must play a predominant (if not exclusive) role in the examination of particular agreements." *Id.*; *see also* MASSAROTTO, *supra* note 4 at 3.

[9] Joshua D. Wright & Douglas H. Ginsburg, *The Economic Analysis of Antitrust Consents*, 46 EUR. J. L. & ECON. 245 (2018).

[10] *See* MASSAROTTO, *supra* note 4 at 8; Giovanna Massarotto, *The Deterrent and Enunciating Effects of Consent Decrees*, 11 J. COMPETITION. L. & ECON. 493 (2015).

[11] *See, e.g.*, Harry First, *Is Antitrust "Law"?*, 10 ANTITRUST 9 (1995); Douglas A. Melamed, *Antitrust: The New Regulation*, 10 ANTITRUST 13 (1995); MASSAROTTO, *supra* note 4 at 209-213.

[12] Consent decrees enable antitrust agencies to save time and reduce the costs of an expensive trial. Similarly, companies avoid the risk of uncertain outcomes, as well as reputational harm even when no finding of liability is made. MASSAROTTO, *supra* note 4 at 7.

[13] *See A Brief Overview of the Federal Trade Commission's Investigative, Law Enforcement, and Rulemaking Authority*, FEDERAL TRADE COMMISSION (revised 2019), https://www.ftc.gov/about-ftc/what-we-do/enforcement-authority ("When the Commission has 'reason to believe' that a law violation has occurred, the Commission may issue a complaint setting forth its charges. If the respondent elects to settle the charges, it may sign a consent agreement (without admitting liability), consent to entry of a final order, and waive all right to judicial review.").

[14] Giovanna Massarotto, *The Deterrent and Enunciating Effects of Consent Decrees*, 11 J. COMPETITION L. & ECON. 493, 497 (2015).

[15] Herbert Hovenkamp, Progressive Antitrust (Jan. 25, 2018) (unpublished manuscript) (on file with the University of Pennsylvania Legal Scholarship Repository), available at https://scholarship.law.upenn.edu/cgi/viewcontent.cgi?article=2766&context=faculty_scholarship, ("[A]ntitrust policy is an extended arm of regulation.").

[16] Ronald Coase, *The Federal Communications Commission*, 2 J. L. & ECON. 1, 5, 6 (1959) ("The Commission was … provided with massive powers to regulate the radio industry. … In 1934 the powers exercised by the Federal Radio Commission were transferred to the Federal Communications Commission, which was also made responsible for the regulation of the telephone and telegraph industries.").

[17] Under Section 5 of the FTC Act, the FTC regulates "unfair methods of competition … and unfair or deceptive acts or practices." 15 U.S.C. § 45(a).

In summary, FTC antitrust enforcement mechanisms resemble economic regulation because, as outlined above, the FTC enforcement action occurs mostly outside of courts and its decisions are grounded in economics. Economic concepts drive decisions on what the FTC considers anticompetitive conduct as well as the types of antitrust remedies to adopt.

2. The FTC and Section 5 of the FTC Act

In contrast to the DOJ Antitrust Division, which represents the U.S. in criminal as well as civil antitrust cases and traditionally plays a prosecuting role, the FTC is an administrative agency with regulatory powers in addition to its prosecutorial powers.[18] Our study specifically focuses on FTC antitrust enforcement action under Section 5 of the FTC Act, because the FTC is the only agency with authority to enforce the FTC Act and it is not technically in charge of enforcing the Sherman Act.[19]

According to Section 5, the FTC has exclusive authority to regulate "unfair methods of competition . . . and unfair or deceptive acts or practices,"[20] preventing individuals, partnerships, or corporations from unfairly disrupting competitive markets.[21] In other words, Section 5 grants the FTC a wide range of discretion in controlling and regulating markets generally.[22] Since there are almost no litigated Section 5 cases, [23] our ML algorithm was trained mainly on regulatory settlements, known as consent decrees.

The structure and powers of the FTC resemble those of many antitrust agencies all over the world. EU National Antitrust agencies are mostly administrative agencies with similar powers, although they generally enjoy less open-ended delegations of power than does the FTC. Therefore, the same or similar AI techniques that have been applied in building the ML algorithm at hand are likely to be helpful for many other agencies across the world.

[18] The FTC "was meant to practice preventive law through administrative and regulatory activities as well as by the initiation and conduct of adversary proceedings . . . Both agencies were to work in the same field, but with different tools." Hon. Edward F. Howrey, Address before the Section of Antitrust Law of the New York State Bar Association (Jan. 28, 1954).

[19] *The Antitrust Laws*, FEDERAL TRADE COMMISSION, https://www.ftc.gov/tips-advice/competition-guidance/guide-antitrust-laws/antitrust-laws (last visited Feb. 28, 2021) ("The Supreme Court has said that all violations of the Sherman Act also violate the FTC Act. Thus, although the FTC does not technically enforce the Sherman Act, it can bring cases under the FTC Act against the same kinds of activities that violate the Sherman Act.").

[20] 15 U.S.C. § 45(a).

[21] William E. Kovacic & Marc Winerman, *Competition Policy and the Application of Section 5 of the Federal Trade Commission Act*, 76 ANTITRUST L. J. 929, 930 (2010).

[22] Richard Posner, *The Federal Trade Commission: A Retrospective*, 72 ANTITRUST L.J. 761, 765-66 (2005).

[23] Jan M. Rybnicek and Joshua Wright observed that "Section 5 enforcement has resulted in no litigated cases," focusing instead on regulatory settlements defined by the Commission. Jan Rybnicek & Joshua D. Wright, *Defining Section 5 of the FTC Act: The Failure of the Common Law Method and the Case for Formal Agency Guidelines*, 21 GEO. MASON L. REV. 1287, 1305 (2014); *see also* William E. Kovacic & Winerman Marc, *Competition Policy and the Application of Section 5 of the Federal Trade Commission Act*, 76 ANTITRUST L.J. 929, 936-937 (2010).

B – AI - Primary Concepts

We now provide a brief overview of the primary AI concepts relevant for our AML algorithm. In particular, we provide an overview of ML and the main approaches and techniques available in the learning process. The following discussion is limited to the main concepts and terminology relevant to the construction of our AML model.

1. Machine Learning

Today, ML is the main AI paradigm for a wide variety of applications, such as speech recognition (Alexa or Siri) or machine translation (Google Translate). ML algorithms enable machines to learn how to perform a task, such as playing chess or translating text, through experience. Experience here manifests itself from large volumes of data, annotated or labeled with some information of interest (e.g., credit risk score). Models are trained on the data and subsequently used to make predictions concerning the information of interest.[24] Deep Learning (DL) is a specific form of machine learning, relying solely on complex neural network architectures.[25]

There are three main ML approaches: 1) supervised learning (SL); 2) unsupervised learning (UL); and 3) reinforcement learning (RL). Below, we describe only the SL and UL paradigms as they adopt completely opposite learning procedures, and our proposed method is based on UL. The comparison between SL and UL better explains our motivation in opting for the method described in Part III.B.[26]

2. Supervised Leaning (SL)

In SL, an algorithm is presented with huge volumes of example data collected from the past. These data will consist of a number of variables (e.g., age, education level, salary) as well as a label, which corresponds to the information of interest (e.g., credit risk level).

The algorithm will then identify relationships between the independent variables and the information of interest (known as the dependent variable or target). For example, the algorithm may learn that customers of a certain age and salary tend to have high credit risk. These relationships are usually established using methods originating from mathematics or statistics. This phase of learning from past annotated data is known as training. Once the algorithm has been trained, it can be applied to make predictions on new unseen cases.

[24] *See* TOM MITCHELL, MACHINE LEARNING 2 (1997) ("A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience.").

[25] *See* Yann LeCun et al., *Deep Learning*, 521 NATURE 436 (2015).

[26] Francois-Lavet et al., *An Introduction to Deep Reinforcement Learning*, 11 FOUNDATIONS AND TRENDS IN MACHINE LEARNING 219 (2018).

The peculiarity of SL is its requirement for large volumes of training data, annotated with the label of interest. There are several different types of SL algorithms, such as random forests, neural networks, and support vector machines. These algorithms differ in the way in which they learn the relationships between variables.

3. Unsupervised Leaning (UL)

Unlike SL and RL methods, Unsupervised Learning (UL) algorithms are generally not concerned with learning how to perform specific tasks. As such, there is neither a training phase in which the algorithms learn from past annotated data nor an exploration phase in which they explore their environments to determine an optimal action sequence to maximize specific rewards. Indeed, UL algorithms operate without any type of supervision or external reward signal. The main aims in UL are to discover latent structures and extract rules or associations from data, without any prior training or exploration phases. That is, the algorithms operate completely on their own (in an "unsupervised" manner). Whereas SL algorithms typically try to *predict* outcomes, UL algorithms are typically used to *describe* data.

The approach we have adopted in our algorithm is known as *clustering*, which identifies latent structures within a dataset and involves estimating the similarity between various data points and grouping similar data points into clusters. Clustering is widely adopted in marketing in order to identify groups of consumers with similar price or product preferences. The similarity between data points is often estimated by first projecting these data points as vectors in a multi-dimensional space and then computing the distance between them using measures like Euclidean distance. Subsequently, data points that are found to be close to each other (based on the computed distance) are grouped into clusters.

Literature distinguishes between two types of clustering approaches, partitional and hierarchical. Partitional clustering methods partition the data points into clusters based on the estimated similarity. In hierarchical clustering, the data points are organized in a hierarchical fashion based on their similarity. The resulting hierarchies are a nested sequence of clusters, known as dendrograms. As can be expected, partitional and hierarchical methods rely on different similarity measures.

Several clustering algorithms exist, including K-means and bisecting K-Means (both partitional) and divisive and agglomerative methods (both hierarchical), all of which we have applied to our AML model. Clustering algorithms will be described in more detail in Part III.B.

C – Why Antitrust and AI?

Having clarified the role and powers of the FTC within the U.S. antitrust law enforcement framework and the main AI techniques relevant for our project, we now explain why AI techniques may be relevant to antitrust enforcement.

As already mentioned, AI algorithms have been applied to other areas of law enforcement unsuccessfully, causing some to question the usefulness of AI in the law. For example, Compas is an algorithm employed in the U.S. legal system with the aim to make judicial decisionmaking more efficient. Compas was trained to assist judges in Florida in deciding whether a defendant was likely to re-offend[27] and should remain in jail or be released while the trial was pending.[28] However, the algorithm showed a clear bias. According to a study conducted by Propublica, "defendants predicted to re-offend who actually did not were disproportionately black."[29] This algorithm exhibits the risks related to the adoption of AI techniques, which can lead to bias at scale if the algorithm is not correctly built and trained.

Despite initial skepticism of the utility of AI applied to law enforcement, as observed in Part II.A.1, FTC antitrust enforcement resembles economic regulation, which means it is neither a law nor best practice, but "the pattern of government intervention in the market"[30] designed to achieve market efficiency.[31] Therefore, antitrust enforcement might serve as a safer testing ground for the exploitation of AI techniques in future regulatory interventions based on economic reasoning and goals rather than on protecting human rights. In this way, government agencies can potentially gain in efficiency and companies can have a better understanding of what constitutes an anticompetitive practice.

III. Antitrust & AI Techniques

Part III describes the steps we took to build the dataset for our AML algorithm. Subpart A explains in detail how we collected the data included in our dataset. Subpart B describes the ML methods applied to analyze the dataset, including those for computing similarities between antitrust cases, identifying pertinent characteristics of these cases, and assessing the results' quality.

A – AML Dataset

1. Data Collection

Data collection was the first step for building our AML system. Specifically, we had to identify what were the appropriate data to collect, as well as the significant variables. A readily available source of data for the proposed ML algorithm was the collection of cases analyzed in Giovanna Massarotto's book *Antitrust Settlements: How a Simple Agreement Can Drive the Economy*.[32] These cases spanned both the U.S. (FTC and DOJ) and Europe, with a concentration from 2013 to the end of 2018.

However, as outlined in Part II, a closer inspection revealed significant heterogeneity among the various jurisdictions and antitrust agencies because each agency enforces its own laws. Thus, as explained in Subpart B, our final data set was narrowed to the FTC cases opened under Section 5 of the FTC Act over a wider

[27] *See* MARK COECKELBERGH, AI ETHICS 127 (2020).
[28] *See* Karen Hao & Jonathan Stray, *Can You Make AI Fairer Than a Judge? Play Our Courtroom Algorithm Game*, MIT TECHNOLOGY REVIEW (Oct. 17, 2019), https://www.technologyreview.com/2019/10/17/75285/ai-fairer-than-judge-criminal-risk-assessment-algorithm.
[29] COECKELBERGH, *supra* note 27, at 127.
[30] Richard A. Posner, *Theories of Economic Regulation*, 5 BELL J. ECON. & MGMT. SCI. 335 (1974).
[31] *See* MASSAROTTO, *supra* note 4 at 2.
[32] MASSAROTTO, *supra* note 4.

range of time: from September 2005 to November 2019. We included a total of seventy-two proceedings. We collected data directly from the FTC official website, from the Cases and Proceedings dataset.[33]

2. Variables

From each FTC proceeding, we extracted data regarding the year, the name of the proceeding (which usually identifies the parties involved), the affected industry, the investigated anticompetitive conduct, and the adopted remedies.

Initially, we also considered data concerning the affected markets, the market shares of the entities subject to investigation, and if the market at hand was either a natural monopoly or a two-sided market. We dispensed with some of these variables because there were too few examples of each category for meaningful statistical analysis. The fact that the market was "two-sided" resulted in no significant statistical conclusions as there were very few cases related to such markets. Similarly, it did not make sense to consider the variable related to the "affected markets" as markets were too various and narrowly defined, while "market shares" and the fact that there was a "natural monopoly" were rarely identified in the dataset, and thus were also useless for our AML model.

We initially also considered whether the FTC imposed a monetary sanction or disgorgement and whether the case ended with a consent decree or the identification of no antitrust violation. However, due to the fact that a monetary sanction or disgorgement and outcomes other than consent decrees were present in very few cases, we consider these variables statistically unhelpful for training our ML algorithm. Next, we describe how we coded the data.

3. Data Coding

With respect to the variable "Year," we defined the following five years or ranges of time: 2005 to 2013, 2014 to 2016, 2017, 2018, and 2019, ensuring an equal number of cases per range, giving preference to the most recent cases. The Industry variable had the following categories: Data industry; Computer industry; Healthcare/Pharmaceutical; Professional/Trade associations; Gas&Oil; Barcode; Sport industry; Telecommunications; Transportation industry; Real estate; and Funeral service industry.

Because no cases included more than four investigated anticompetitive conduct types, we considered from a minimum of one to a maximum of four conducts per case. We identified the following categories of anticompetitive conduct: exclusionary conduct (A); predatory conduct (B); refusal to deal (C); tying conduct (D); price fixing (E); rebates (F); discriminatory practice (G); customer allocation agreement (H); pay for delay (I); disruption in the bidding process (J); agreement orchestration (K); invitation to collude (L); agreement not to compete (M); unlawful exchange of information (N); concerted practices (O); conspiracy (P); no poach (Q);

[33] *Cases and Proceedings*, FED. TRADE COMM'N, https://www.ftc.gov/enforcement/cases-proceedings (last visited Feb. 26, 2021).

and no anti-competition (NOCOND). We also specified if the anticompetitive conduct fell into the broader Monopolization Conduct or Agreements in Restraint of Trade categories. We coded Monopolization Conduct and Agreement in Restraint of Trade as binary variables, i.e., taking values of either 0 or 1.

With respect to the antitrust remedies employed in each FTC order, initially we distinguished between structural and behavioral remedies, and we specified for each case the types of remedies adopted in the FTC order. We identified a maximum of five remedies imposed per proceeding, and we classified the "type of remedy" imposed in the following categories: amendments to contract provisions (R1); amendments to the code of ethics (R2); obligation to disclose information (R3); limitation to enter into specific markets (R4); refraining from the investigated conduct (R5); compliance obligations (R6); implementation of an antitrust compliance program (R7); contract limitations (R8); divestiture (R9); impose specific contract requirements (R10); limitation in the exchange of information (R11); permanent injunction (R12); other performance obligations (e.g. equipment interoperability in the Intel case) (R13); and no remedy (NOREM).

4. Final Dataset

To sum up, the final dataset we used to construct our algorithm includes seventy-two proceedings under Section 5 of the FTC Act from September 2005 to November 2019. Each case was described though 16 variables: 1) case name, 2) year, 3) industry, 4) Agreement in Restraint of Trade, 5) Monopolization Conduct, 6) Type of Conduct **1**, 7) Type of Conduct **2**, 8) Type of Conduct **3**, 9), Type of Conduct **4**, 10) Structural Remedies, 11) Behavioral Remedies, 12) Remedy **1**, 13) Remedy **2**, 14) Remedy **3**, 15) Remedy **4**, and 16) Remedy **5**.

As will be discussed later in our experiments, the variables: Remedy **1**, Behavioral Remedies, and Structural Remedies were discarded, as they were not considered to be relevant by our algorithm. Also, the case name variable was not included in our analysis, as it lacks informational content—thus it was dropped in our analyses.

Our dataset size might appear to be relatively small compared to those traditionally employed in ML, particularly in Deep Learning. Arguably, the availability of more data, and of better quality data, significantly enhances the performance of these algorithms and makes the results statistically more reliable. However, in our study, the small size of our dataset is determined by the number of antitrust cases, dealt with by a single agency and confined to a recent (and thus restricted) time frame. The availability of more data could impact our findings in two ways. First, we might observe that the clusters generated and the distribution of cases within clusters do not change, thereby further confirming our current results. Second, we might observe the formation of more clusters, with a different distribution of cases across clusters. However, it is important to mention that even if our small dataset size could be perceived as compromising our findings' reliability, we performed several steps (EDA, silhouette score, and feature selection) to mitigate this effect.

B – Methodology

After gathering and coding the required data, we studied a variety of UL techniques (clustering methods: K-Means, Bisecting K-Means, K-Modes), which attempted to automatically identify similar cases in our dataset. This Part explains in detail that study and what clusters our algorithm generated, as well as the most significant dimensions (variables) detected through the adoption of random forests and support vector machines.

1. EDA & Pre-Processing

Before actually processing data, we performed an exploratory data analysis (EDA), which examines a dataset and detects any peculiarities or incoherencies exhibited by the data points. Note that each data point here corresponds to an antitrust case in our dataset.

EDA permits correction of any errors that might have been introduced during data collection. EDA is fundamental in machine learning endeavors because it determines how to make data more amenable to the application of specific machine learning algorithms. Essentially, EDA cleans data in a dataset to improve the performance of a model’s learning process by ensuring that the data fed to the algorithms are correct, of good quality, and will not compromise the algorithms’ accuracy and results.

The EDA here determined that two variables, Structural Remedies and Behavioral Remedies, appeared to be irrelevant and would not provide any meaningful information for the subsequent steps of our analysis. With respect to the Structural Remedies variable, all the cases recorded in our dataset had a value of 0, implying that there were no structural remedies adopted in the analyzed cases. Thus, it was dropped from our dataset.

On the other hand, the second variable, Behavioral Remedies, contained values that were distributed in an imbalanced manner. Specifically, we noted that only two cases in the dataset had a value of 0 for Behavioral Remedies, while all the remaining cases had a value of 1. Thus, this variable was also dropped from our dataset because such an uneven distribution of values causes either the performance of machine learning algorithms to deteriorate or the algorithms to learn incorrect relationships. For instance, in our given situation (2 cases with a value of 0 and the rest with a value of 1), algorithms might be biased into giving more weight to the cases with a value 1 because they constitute the majority. For a similar reason, we dropped the first remedy variable (namely Remedy 1), as in all cases except for one, the FTC required the entity subject to investigation to refrain from the investigated conduct.

Techniques for dealing with data imbalance exist. For example, with undersampling, we could remove cases with the majority value (e.g., Behavioral Remedies=1) from our dataset so as to balance the distribution. However, this removal would have resulted in a dataset so small that any further analyses would not yield any meaningful information. Another solution to address data imbalance is SMOTE (Synthetic Minority Oversampling Technique).[34] However, the basic premise of SMOTE is the creation of synthetic data points based on their similarity with other data points. Such an approach would be unsuitable for our study as we are concerned with *real* antitrust cases. The data points generated by SMOTE to balance our dataset would therefore be of very limited value, if any.

In addition, as part of the EDA, we determined whether certain variables were correlated with each other. In statistics and machine learning, correlated variables tend to have a significant impact on the final clustering. Thus, they could bias the algorithm's learning procedure. To address this issue, we visualized the variables' correlations according to a heat map, depicted in Appendix 1. The horizontal and vertical axes correspond to variables. The greener the cell, the higher the correlation. As can be expected, each variable is highly correlated to itself, which is natural. The heat map revealed no apparent significant correlation among variables, which do not warrant further multi-collinearity tests.

2. Methods for Computing Similarity Between Antitrust Cases

After pre-processing, we focused on analyzing the data using machine learning in order to address our research questions and achieve our objectives. In particular, we worked on discovering underlying patterns characterizing antitrust cases and using these patterns to generate clusters of similar cases. With this information, we identified the most pertinent variables characterizing antitrust cases in our dataset. These variables tend to have the highest impact or importance when computing the similarity between cases. Also, these variables, if properly identified, could enable antitrust enforcers and legislators to better comprehend antitrust cases and even predict their outcomes.

3.Unsupervised Learning - Clustering

We adopted clustering algorithms, which were briefly discussed in Part II. Among the two families of clustering algorithms, partitional and hierarchical, the former lent itself more naturally to our task. Furthermore, hierarchical approaches are generally more computationally expensive and have higher run-time complexity than their partitional counterparts.

Several partitional clustering algorithms exist. Because it is extremely difficult to determine which algorithms will give the best performance, it is common to experiment with several algorithms before choosing the best one. Consequently, we decided to investigate three partitional clustering algorithms: K-Means, Bisecting K-Means, and K-Modes. The algorithms project the data, here antitrust cases, as vectors in a multi-dimensional space. Then, the distance between cases is estimated based on the variables characterizing each case. Finally, similar (less distant) cases are grouped into clusters while minimizing a criterion, such as the error. For a more elaborate description, please see Appendix 2. Different approaches and metrics exist for estimating the quality of the generated clusters. In

[34] *See generally* Alberto Fernández et al., *SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary*, 61 J. A.I. RSCH. 863 (2018).

our experiments, we rely on the standard elbow method with silhouette scores, as described in Appendix 3.

4. Clustering Results

Appendix 4 depicts the silhouette scores achieved by the K-Means, Bisecting K-Means, and K-Modes algorithms with different numbers of clusters. As outlined in Appendix 4,[35] the best-performing algorithm (in terms of silhouette scores) is the K-Means with four clusters (red and green plots—each indicating different ways by which the K-Means algorithm was initialized). It can also be observed that the best silhouette scores for both K-Modes and Bisecting K-Means (cyan and black plots, respectively) were achieved with six clusters. K-Modes and Bisecting K-Means significantly underperformed both K-Means algorithms.

In our next set of experiments, our aim was to confirm that K-Means was actually the best-performing algorithm. We adopted an intuitive approach to confirm our empirical findings and relied on the distribution of data points achieved by each clustering algorithm. Specifically, for each of these algorithms, we considered only their optimal number of clusters, i.e., that number of clusters that yielded the best silhouette scores. Appendix 4 shows that for K-Means, the optimal score was achieved with K=4 clusters, while the optimal K-Modes and Bisecting K-Means were achieved with K=6. Then, for each of these configurations, we plotted: 1) the silhouette scores in Appendix 5; 2) a 2-D plot showing how the various data points and clusters were distributed in Appendix 6; and 3) a corresponding 3-D plot in Appendix 7.

These plots reveal that the best performing algorithm is the K-Means with four clusters. Besides the higher silhouette scores, it can also be seen from the 2-D and 3-D plots that the clusters generated by K-Means are more coherent. Specifically, the data points are more uniformly distributed within each cluster. This can be visually observed by the arrangement of the colored dots in the plots. It can also be observed that there is much less overlap between clusters.

This can be conceptualized as the algorithm having more certainty that a certain data point belongs in a given cluster. Therefore, when presented with a new case, the algorithm is more confident in predicting the cluster to which it could potentially belong. Conversely, we note that the clusters produced by the K-Mode and Bisecting K-Means are very near to each other, with lots of data points overlapping between clusters. This reveals that the K-Mode and Bisecting K-Means both have difficulties in precisely determining the best cluster in which a given data point could be classified.

5. Identifying Pertinent Characteristics of Antitrust Cases (Feature Selection)

As discussed earlier, in our dataset, we defined each of the seventy-two antitrust cases according to a number of variables. The final dataset used in our experiments

[35] Note that such plots are referred to as "elbow plots" (or "elbow methods").

was comprised of the variables as shown in the figure below. Note that variables marked with "*" were not considered in the analysis.

Industry	Type of Conduct 4	Remedy 3
Year	Structural Remedies*	Remedy 4
Type of Conduct 1	Behavioral Remedies*	Remedy 5
Type of Conduct 2	Remedy 1*	Case Name*
Type of Conduct 3	Remedy 2	Monopolization Conduct
Agreement in Restraint of Trade		

Figure 1: Variables/Features of Antitrust Dataset

Identifying which variables have high information content from a machine learning perspective also resulted in interesting and useful insights from a legal perspective. The ML algorithm performs better with highly informative variables than a number of less significant variables because the latter can introduce noise to the algorithm. The aim is therefore to find a reduced subset of variables that best characterize the cases. Armed with this particular subset of variables, antitrust enforcers and legislators should be able to draw preliminary conclusions or make predictions about an antitrust situation before them. Furthermore, this inquiry is relevant as it could reveal certain underlying patterns that antitrust enforcers and legislators implicitly, or even unconsciously, consider when looking at antitrust cases. We develop this point further in Part IV.

Here, we describe our feature selection experiments. After applying our clustering algorithms to our dataset, each data point is assigned to a unique cluster (see Appendices 6 and 7). We treat each cluster as a category, resulting in four categories (labeled 1-4) for K-Means and in six categories (labeled 1-6) for both K-Modes and Bisecting K-Means.

Then, we determine which variables of the data points in a given category were more predictive of that cluster. In other words, we measured the degree to which a given variable (e.g., Industry) influenced the cluster to which a data point would be assigned. For example, data points with {Industry=Telecom, Type_of_Conduct="exclusionary conduct"} have a higher probability of being assigned a given cluster (e.g., cluster 3). To accomplish this, we relied on a random forest algorithm, which attempts to transform the data points (and variables) into a collection of tree-like structures, and in doing so, estimates how much each variable is predictive of a given cluster. Specifically, it relies on a measure of entropy derived from Shannon's Information Theory, which, broadly speaking, estimates the order or disorder in the data if a given variable is removed.[36]

We used the same procedure to determine the most influential variables, which we repeated for the clusters generated by K-Means, K-Mode and Bisecting K-Means. However, we will focus our discussion on the K-Means (with four clusters)

[36] *See generally* Gérard Biau & Erwan Scornet, *A Random Forest Guided Tour*, 25 *TEST* 197 (2016) (providing an overview of the Random Forest algorithm).

configuration, which was the best-performing according to the silhouette scores (see Appendix 5).

The top five variables thus identified were namely Type of Conduct 2, Type of Conduct 3, Type of Conduct 4, Remedy 3, and Type of Conduct 1. We could use these top five features as independent features to predict, for instance, the outcomes of cases using a supervised learning algorithm. Working with a subset of highly relevant features makes the algorithm less susceptible to noise that might be introduced by less relevant ones. However, a feature not within the top five (and thus discarded) might have been relevant for a handful of cases. In such a situation, the algorithm might be unable to correctly predict the cases' outcomes. Nevertheless, feature selection generally improves performance. Variables and their relative weights (importance), as computed by the random forest algorithm, are presented in Appendix 8. For the sake of illustration, we also included the Behavioral Remedies variable in the plot. This plot confirms that that variable is among the least important, which reaffirms the decision to remove it. We went a step further and investigated variable importance per cluster. Results are depicted in Appendix 9.

To validate these results, we employed another method, namely a support vector machine (SVM), to estimate the degree to which a variable influences the cluster to which a data point is assigned. Broadly speaking, an SVM is a machine learning approach that attempts to find the best hyperplanes separating points that belong to different clusters. The SVM was trained using Recursive Feature Elimination. Given the limited size of our dataset, training was achieved using 5-fold cross validation. Specifically, given the clusters produced earlier, the SVM was trained to predict the cluster based on data points that had been assigned to each category. The same experiment was repeated by varying the number of features and the accuracy of the cluster predictions was assessed. Results are shown in Appendix 10. They indicate that the best accuracy is achieved with nine features. A closer inspection revealed that these nine features are: Year, Industry, Type of Conduct 1, Type of Conduct 2, Type of Conduct 3, Type of Conduct 4, Remedy 2, Remedy 3, and Remedy 4.

Finally, we compared the set of pertinent variables identified by both methods (random forest and SVM). We kept only those features identified by both. They were: Remedy 3, Type of Conduct 1, Type of Conduct 2, Type of Conduct 3, and Type of Conduct 4, as described in Part III.A. It is important to note that the order or magnitude of the importance of these variables differed. For example, according to the random forest algorithm, the most pertinent variable was Type of Conduct 2. However, according to the SVM algorithm, this variable was the third most important. Such differences are common in machine learning studies due to noise, some degree of randomness inherent in the learning procedure, and the inherent differences between the algorithms.

IV. Antitrust Interpretation of Results

In Part IV, we analyze the outcomes of our ML algorithms and the different techniques applied from an antitrust point of view. In particular, Part IV is divided into three Subparts, which include an evaluation of: A. Clusters; B. Variables; C. UL & Limitations.

In Subpart A, we assess whether the four clusters resulting from our ML algorithm make sense from an antitrust perspective. In Subpart B, we make the same evaluation with respect to the detected variables to determine if their importance from a technical point of view accurately reflects their importance from an antitrust perspective. Finally, we critically discuss the limitations of our approach and consider possible avenues for improvement.

A – Clusters

As outlined in Part III.B, the best-performing algorithm was the K-Means algorithm with four clusters. Here, we analyze the four clusters from an antitrust perspective to assess their validity.

1. Cluster 1

The most interesting result in cluster 1 is related to the industry variable. All cases concerning the computer industry and the data industry were placed in cluster 1. There were only three cases in those industries, but it is curious to notice that all of them appear in the same cluster. Fourteen cases out of the twenty in cluster 1 concern the healthcare/pharmaceutical industry, and the remaining three cases involve professional/trade associations.

The conduct that is most frequent is Conduct E, price fixing, which appears in eight cases. Conduct A, related to exclusionary conduct, is present in seven cases. The remedies that appear most frequently in cluster 1 are R8, contract limitation, which occurs eight times, and R6, compliance obligations, identified in seven cases. In summary, cluster 1 seems to suggest that the algorithm looks at price fixing and exclusionary conduct in the healthcare/pharmaceutical industry by adopting contract limitations and compliance obligations as remedy. Cases in the computer/data industry seem to raise antitrust concerns similar to those present in the healthcare industry.

2. Cluster 2

All twenty cases identified in cluster 2 occurred in the same time frame, from 2005 to 2013. Nine out of the ten cases concerning the real estate industry appear in cluster 2. This cluster is characterized by the number of different conducts investigated as, in the majority of cases (thirteen out of twenty cases), the FTC looked at three or more anticompetitive practices. Conduct P, conspiracy, appears in seventeen out of twenty cases. As we will see, this is relevant for identifying the employed remedies.

In nineteen out of twenty cases, the FTC adopted R6, compliance obligations. Seven cases include the remedy R11, limitation in the exchange of information. Although the detected practices are too different from each other to be able to determine a specific rule linking the investigated antitrust practices and the remedies imposed, this result is still interesting from an antitrust point of view. Conspiracy, the most detected conduct, is strictly related to exchange of information, and as a consequence, the "limitation in the exchange of information" remedy is one of the most common remedies revealed in cluster 2.

To sum up, the ML algorithm suggests that the FTC, in cases involving more than two anticompetitive practices, adopts "compliance obligations" as a remedy very commonly. Because seventy percent of cases concern "conspiracy" as one of the conducts, the algorithm recommends considering "limitation in the exchange of information" as a remedy. The results revealed in this cluster appear to be particularly valuable for cases concerning the real estate industry, given that they were almost entirely assigned to this cluster.

3. Cluster 3

The third cluster included eighteen cases. In seventeen, the FTC investigated merely one or two anticompetitive practices. More precisely, fifteen out of eighteen cases concern proceedings involving only one type of anticompetitive conduct. Seven cases featured Conduct M, agreement not to compete, and three cases had Conduct L, invitation to collude.

With respect to remedies, we notice that in eight cases, the FTC did not require any remedy; in seven cases, it imposed one remedy; in one case, two remedies; and in two cases, the FTC required compliance with three different remedies. The most common remedies detected are: R6, compliance obligations; R7, implementation of an antitrust compliance program; and R11, limitation in the exchange of information. All of these remedies appear in three cases of cluster 3. In addition, we noticed that this cluster mainly concerns trade or professional associations, which appear in nine cases out of eighteen. Finally, the year of the adopted decision looks interesting, as eleven out of eighteen cases referred to proceedings decided between 2014 and 2016, four cases in 2017, and only three between 2005 and 2013.

Looking at this data, we can argue that the algorithm learned to detect cases in which the FTC investigated merely one or two anticompetitive practices. In this situation, the algorithm seems to suggest that the FTC does not impose remedies or adopts only one of the following: 1) compliance obligation; 2) the implementation of a compliance program; 3) or limitation in the exchange of information. This recommendation stems from and seems to be justified by the fact that some cases in this cluster concerned "agreement not to compete" or "invitation to collude." In sum, what the algorithm suggests seems to be logical from an antitrust point of view.

4. Cluster 4

This last cluster detected by our algorithm is likely the most interesting from an antitrust perspective. Cluster 4 identifies twelve cases, all of them occurring between 2005 and 2013. Eight cases out of twelve concern the healthcare/pharmaceutical industry (which includes professional associations in this industry). Nine cases detected in this cluster concern Conduct E, price fixing; four cases involve both Conduct A, exclusionary conduct, and Conduct P, conspiracy. With respect to remedies, the most frequent remedies in this cluster are: R6, compliance obligations, detected in nine cases out of twelve, and R11, limitation in the exchange of information, revealed in seven cases.

From an antitrust perspective, this cluster appears useful because it detects price fixing as the most common anticompetitive practice in the healthcare/pharmaceutical industry. In addition, the cluster reveals as common remedies in these cases compliance obligations and limitation in the exchange of information, which make perfect sense in price fixing cases.

B – Variables

Having explained the validity of clusters generated by our UL algorithm, we now evaluate the validity of variables that resulted from the technical testing process. The adoption of different algorithms enabled us to identify the most valuable variables from a technical point of view. Here, we analyze if the results of EDA and pre-processing performed in Part III.B and the identified important variables make sense from an antitrust perspective.

1. EDA & Pre-Processing

During the EDA process, we found that two columns of our dataset, “Structural Remedies” and “Behavioral Remedies,” were statistically irrelevant. In all cases recorded in our dataset, “Structural Remedies” had a value of 0. Except for two cases, “Behavioral Remedies” always had a value of 1. These two variables were dropped.

From an antitrust perspective, this decision makes perfect sense. Generally, antitrust remedies are classified into two macro-categories: structural remedies and behavioral remedies. Because we were able to identify in each case the specific types of remedies (e.g., compliance obligations, amendments to contract provisions, divestiture) adopted by identifying a maximum of five remedies per case, a broader distinction between “Structural Remedies” and “Behavioral Remedies” is meaningless.

2. Variable importance

As outlined in Part III.B, paragraph 5, an ML approach is more precise, and hence performs better, with a small subset of highly informative variables. In Part III.B, we adopted the Random Forest algorithm to detect the most important variables of our dataset, which are displayed in order of importance in Appendix 8. Specifically, the most significant five variables, in order of importance, are: Type of Conduct 2, Type of Conduct 3, Type of Conduct 4, Remedy 3, and Type of Conduct 1. Then follow Remedy 2, Industry, Year, Remedy 4, Monopolization Conduct, Agreements in Restraint of Trade, and Remedy 5.

Appendix 9 shows variable importance per cluster. In addition to the random forest, we also relied on an SVM to assess each feature's importance. Both the random forest and the SVM identified a common subset of pertinent variables. They were: Remedy 3, Type of Conduct 1, Type of Conduct 2, Type of Conduct 3, and Type of Conduct 4. This means that the algorithms would exclude the following variables: Year, Industry, Agreements in Restraint of Trade, Monopolization Conduct, Remedy 2, Remedy 4, and Remedy 5.

The exclusion of Agreements in Restraint of Trade and Monopolization Conduct makes sense from an antitrust perspective because we have already identified the type of conduct in more detail. For example, we have defined price fixing, invitation to collude, and conspiracy, which are subcategories of Agreements in Restraint of Trade. Similarly, we defined exclusionary conduct, refusal to deal, and tying conduct, which are subcategories of Monopolization Conduct.

With respect to years, it is true that we identified more proceedings in some years than in others with similar characteristics, but it is also true that, to our knowledge, Section 5 enforcement in such proceedings did not change over the years we examined.

Remedy 4 and Remedy 5, which appeared last in their importance ranking in both the random forest and SVM algorithms, might also be less relevant from an antitrust perspective. In most cases the FTC adopted one, two, or zero remedies, in addition to restricting the company from engaging in the investigated conduct, which as we have seen is a remedy adopted by default. The variables that look more problematic to evaluate are Industry and Remedy 2. The industry usually is an important feature to evaluate in an antitrust proceeding, as antitrust enforcement is centered on the concept of market definition. However, we admit that we were unable to identify either the market definition or the related market share in the analyzed FTC decisions, and industry is a more generic term that describes a variety of different markets. Thus, we leave open the possibility to exclude the industry from the variables considered, although this variable at first glance seems meaningful from an antitrust perspective.

We note that the decision in each case to classify a given remedy as Remedy 2 or Remedy 3 was random. For example, two cases included both R6 and R11, and the decision of which was assigned to Remedy 2 and which to Remedy 3 was random and could just as well have been reversed. Therefore, we suggested that Remedy 2 remain as variable in the dataset.

C – Unsupervised Learning & Limitations

In Part II.C, we described the primary AI concepts relevant for this paper. In particular, we highlighted the distinction between supervised (SL) and unsupervised (UL) learning algorithms. In our application we opted for an unsupervised learning for the following reasons. Unlike SL methods, UL methods look for structures within data without any prior training or exploration phase (see Part II.B). Second, we were more interested, from a scientific perspective, in whether an algorithm could learn the salient features of antitrust cases and suggest relevant groupings of cases than in asking an algorithm to predict features of cases from incomplete data. In other words, we were interested in seeing what the algorithm learned on its own, instead of asking the algorithm to predict something in particular (e.g., a specific variable).

As outlined in Part III.B, there are a variety of approaches available within UL. We opted for clustering to see if the algorithm identified within our dataset clusters that made sense from an antitrust perspective. In other words, we wanted to test the validity of an unsupervised learning process in the context of antitrust by using clustering techniques. These techniques are similar to those used in advertising. For example, clustering algorithms are often at the crux of recommendation systems to discover and suggest items similar to what customers have purchased or viewed in the past.[37] Thus, we asked the question: *Could a similar algorithm be built to suggest to the FTC possible anticompetitive conduct and remedies to enforce?*

This is why we developed the ML approach at hand by evaluating the clusters and variables detected from an antitrust point of view. The clusters revealed in Part III.B and analyzed in Part IV.A seem to be valid not only technically, but also in the context of antitrust, although we note some limitations. For example, we admit that we used a limited amount of data as we reduced our data collection to the FTC proceedings under Section 5 of the FTC Act from 2005 to 2013. It might be useful to continue feeding the algorithm by considering older cases and to verify whether the algorithm continues to detect similar clusters. Similar AI techniques might also be adopted and tested in the context of mergers.

We also leave room to explore the adoption of different types of algorithms and techniques. As observed in Part II.B, there are a number of possible algorithms and techniques available that can be used and tested in a similar fashion. In summary, the fact that our study produced valid results from both a technical and legal point of view might justify the investigation of more sophisticated and different AI applications in the antitrust domain.

[37] *See* Chih-Fong Tsai & Chihli Hung, *Cluster Ensembles in Collaborative Filtering Recommendation*, 12 APPLIED SOFT COMPUTING 1417 (2012).

V. Final Remarks

Our study shows that a relatively basic ML approach can automatically discover the most important underlying characteristics of antitrust cases and identify similarities between those cases. The clusters that our algorithm generated seem to make sense. Although it is clear that extant AI cannot replace the FTC or other antitrust agencies, AI remains a valuable tool to assist regulators and enforcers in decision making. Companies can also benefit from AI techniques as they help in predicting whether their conduct will attract antitrust scrutiny, which behavior could be considered anticompetitive, and which remedies could be imposed.

The final conclusions we draw are that this achievement has been obtained through the combination of our expertise on antitrust and AI techniques. We could not attain this result without knowing exactly what antitrust elements could be valuable to train an ML algorithm and the different ML techniques available to train and test our results. Overall, it is our hope that our contribution will pave the way for future AI applications in market regulation to increase consumer welfare by fostering innovation in the age of data. The government's adoption of emerging technologies, such as AI, and, more generally, of computational antitrust techniques, seems to us the best way to achieve higher consumer welfare in today's data-driven economy by providing the "adequate resources"[38] that antitrust enforcers need.[39]

[38] Richard A. Posner, *Antitrust in the New Economy*, 68 ANTITRUST L.J. 925 (2001).

[39] *See* Giovanna Massarotto, *Can Antitrust Trust Blockchain?*, *in* ALGORITHMIC ANTITRUST (forthcoming 2021), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3622979 ("Governments must anticipate today's fast moving technologies to be effective.").