Data Analytics and Law

Lesson 6

KEY CONCEPTS

■ Data ■ Analytics ■ Data Analytics ■ Legal Analytics ■ Machine Learning ■ Logistic Regression ■ Triple C Theory ■ Data Assessment

Learning Objectives

To understand:

- What is Data analytics?
- Different analytical methods and techniques used by data analysts
- What is Machine Learning?
- What is Artificial Intelligence and how it helps human being
- What is Triple C Theory?

Lesson Outline

- Introduction to Data Analytics
- Introduction to Legal Analytics
- Introduction to Machine Learning for Lawyers
- Quantitative Legal Prediction vis-à-vis Business of Law
- Bias/ Variance, Precision/Recall & Dimensionality
- Overfitting, Underfitting, & Cross-Validation
- Logistic Regression and Maximum Likelihood
- Triple C Theory and Data Assessment
- Network Analysis and Law
- Lesson Round-Up
- Test Yourself
- List of Further Readings
- List of Other References

INTRODUCTION

In modern times, industry whether legal or technology, academy, practitioners and scholars utilize Artificial Intelligence (AI) and machine learning to perform analysis which used to be labor-intensive endeavors decade back. Although technology can, in some instances, replicate human decision-making, yet professionals including lawyers, researchers and scholars play an essential role in compiling data sets, defining analytical queries, and, most importantly, interpreting findings and presenting them in an accessible way for a broad audience. Hence, this chapter aims to provide the basis for understanding data analytics, its role in various sectors and laws, if any, governing the same.

DATA ANALYTICS[1]

As data is becoming more prominent by the minute, organizations are becoming data-driven, which means adopting methods to collect more data. This data is then sorted, stored, and then analyzed to derive logical and valuable information. Data analytics makes the process possible.

Data analytics is the science of analyzing raw data to make conclusions about that information. Many of the techniques and processes of data analytics have been automated into mechanical processes and algorithms that work over raw data for human consumption.

- Data analytics is the science of analyzing raw data to make conclusions about that information.
- Data analytics help a business optimize its performance, perform more efficiently, maximize profit, or make more strategically-guided decisions.
- The techniques and processes of data analytics have been automated into mechanical processes and algorithms that work over raw data for human consumption.
- Various approaches to data analytics include looking at what happened (descriptive analytics), why something happened (diagnostic analytics), what is going to happen (predictive analytics), or what should be done next (prescriptive analytics).
- Data analytics relies on a variety of software tools ranging from spreadsheets, data visualization, and reporting tools, data mining programs, or open-source languages for the greatest data manipulation.

Data analytics is the process that refers to deriving valuable insights and information from data using quantitative and qualitative methods. It helps businesses and even in science - researchers use it to verify their theories. Data analytics is a broad term that encompasses many diverse types of data analysis. Any type of information can be subjected to data analytics techniques to get insight that can be used to improve things. Data analytics techniques can reveal trends and metrics that would otherwise be lost in the mass of information. This information can then be used to optimize processes to increase the overall efficiency of a business or system.

For example, manufacturing companies often record the runtime, downtime, and work queue for various machines and then analyze the data to better plan the workloads so the machines operate closer to peak capacity.

Data analytics is important because it helps businesses optimize their performances. Implementing it into the business model means companies can help reduce costs by identifying more efficient ways of doing business and by storing large amounts of data. A company can also use data analytics to make better business decisions and help analyze customer trends and satisfaction, which can lead to new—and better—products and services.

1. Source: Jake Frankefield at el (2023) Data Analytics: What it is, How It's Used and 4 Basic Techniques, Investopedia. Available at https://www.investopedia.com/terms/d/data-analytics.asp

Kinds of Data collected by Company

There are three primary kinds of data collected by the companies.

- *First-party data*: The data a company collects about its customers.
- *Second-party data*: The data a company gets from a known organization that collected it originally.
- *Aggregated data*: The data a company buys from a marketplace.

The Evolution of Data Analytics

Data analytics has become the next big thing in both large companies and small startups. The process of data analytics has evolved. Let's take a journey through the evolution of data analytics.

- **Data Analytics and Statistics:** Statistics has a pretty long history. Like, for example, in taxation, governments carried out planning activities for the creation of censuses. It was possible with the use of statistics. Data analytics stemmed from statistics, which analyzed the obtained data.

> Tabulating Machine* was counting machine which comprised of a punch for entering data onto a blank card, a tabulator for reading the cards and summing up information and a sorting box for sorting the cards for further analysis. The invention of the Tabulating Machine was a turning point as it transformed the census process and information processing in general.

- **Data Analysis and Computing:** Technology advancements were game-changers to how businesses adopted data analytics. In 1890, Herman Hollerith invented the "Tabulating Machine" to reduce the time taken to create the Census. This machine was highly useful in finishing the 1890 US Census in only 18 months.

- **Data mining:** Data mining got introduced in the 1990s, which is a process that discovers patterns in large data files. When data analytics moved from traditional methods to more modern means, you could obtain more positive results.

> Data mining refers to the process of sieving through large amounts of information for the purpose of extracting useful information.

- **Google Web Search:** When the Google search engine came into the picture, big data could be analyzed and processed quickly. It played an essential part in the evolution of Data Analytics because the search engine was more automated, scalable, and high-performing.

*(*Source: https://americanhistory.si.edu/collections/search/object/nmah_694410*)

- **Data Processing:** Today, Python & R are the leading technologies in data analytics. They are open-sourced and are capable of integrating with big data platforms and visualization tools. Businesses prefer R when the primary goal is exploratory analysis or modeling. At the same time, enterprises prefer Python to develop applications that have an embedded analytics engine.
- **Predictive Modeling:** Some advanced data analytics techniques that the data scientists and organizations are using are: Random Forests, Matrix Factorization, TensorFlow, Simulated Annealing, etc. For example, Random Forest is a machine learning algorithm which grows and combines multiple decision trees to create a forest. It is used for both classification and regression problems.
- **Visualization:** Many organizations are adopting more open-source technologies for their business. A few examples are D3 and Angular, which is popular development platform which strives to make web development easier through focus on developer's productivity, speed and testability. This decision relies on several factors like cost, customization options, visual appeal, and interactivity.

Data Analysis Steps

The process involved in data analysis involves several different steps:

- The first step is to determine the data requirements or how the data is grouped. Data may be separated by age, demographic, income, or gender. Data values may be numerical or be divided by category.
- The second step in data analytics is the process of collecting it. This can be done through a variety of sources such as computers, online sources, cameras, environmental sources, or through personnel.
- Once the data is collected, it must be organized so it can be analyzed. This may take place on a spreadsheet or other form of software that can take statistical data.
- The data is then cleaned up before analysis. This means it is scrubbed and checked to ensure there is no duplication or error, and that it is not incomplete. This step helps correct any errors before it goes on to a data analyst to be analyzed.

Types of Data Analytics

Following are the main types of data analytics: descriptive, diagnostic, predictive, and prescriptive. Each has its own set of goals and roles in the data analytics process.

1. **Descriptive Analytics:**

   This describes what has happened over a given period of time. Descriptive analytics answers the "what" questions in the data analytics process. It helps stakeholders understand large datasets by summarizing them. The descriptive analysis tracks the organization's past performance. It includes the following steps:

   - Data collection: It refers to the process of gathering data for a specific purpose.
   - Data processing: It refers to the process of converting raw data into useful information.
   - Data analysis: It refers to the process of inspecting, analyzing and transforming the data with the object of discovering useful information and conclusions to support decision making.
   - Data visualization: It refers to the process of representing the data into pictorial modes like graphs, charts, plots and animations.

2. **Diagnostic Analytics**

   This focuses more on why something happened. This involves more diverse data inputs and a bit of hypothesizing. Diagnostic analytics answers the "why" questions in the data analytics process. It analyzes the results from the descriptive analysis and then further evaluates it to find the cause. The diagnostic analysis process takes place in three steps:

   - Identifying any unexpected changes in the data.
   - Data related to the changes is collected.
   - Statistical techniques help find relationships and trends related to the changes.

3. **Predictive Analytics**

   This moves to what is likely going to happen in the near term. The purpose of predictive analytics is to answer questions about the future of the data analytics process. The past data identifies the trends. The techniques used in the process include statistical and machine learning techniques. A few of them are neural networks, decision trees, and regression.

4. **Prescriptive Analytics**

   This suggests a course of action. Prescriptive analysis helps businesses make well-informed decisions and predict the analytics. This type of data analytics uses machine learning strategies that are capable of finding patterns in large datasets.

> **Note:** Data Set refers to an organized collection of data related to a particular subject.

Data Analytics Techniques

There are several different analytical methods and techniques data analysts can use to process data and extract information. Some of the most popular methods are listed below.

- **Regression analysis** entails analyzing the relationship between dependent variables (i.e. the variable which is being tested and measured in a scientific experiment) to determine how a change in one may affect the change in another.
- **Factor analysis** entails taking a large data set and shrinking it to a smaller data set. The goal of this maneuver is to attempt to discover hidden trends that would otherwise have been more difficult to see.
- **Cohort analysis** is the process of breaking a data set into groups of similar data, often broken into a customer demographic. This allows data analysts and other users of data analytics to further dive into the numbers relating to a specific subset of data.
- **Monte Carlo simulations** is a data analysis model which assesses the possible outcomes of an uncertain event. This method conducts repeated random sampling to make estimations of unknown parameters. Often used for risk mitigation and loss prevention, these simulations incorporate multiple values and variables and often have greater forecasting capabilities than other data analytics approaches.
- **Time series analysis** tracks data over time and solidifies the relationship between the value of a data point and the occurrence of the data point. This data analysis technique is usually used to spot cyclical trends or to project financial forecasts.

Data Analytics Tools

- In addition to a broad range of mathematical and statistical approaches to crunching numbers, data analytics has rapidly evolved in technological capabilities. Today, data analysts have a broad range of software tools to help acquire data, store information, process data, and report findings.
- Data analytics has always had loose ties to spreadsheets and Microsoft Excel. Now, data analysts also often interact with raw programming languages to transform and manipulate databases. Open-source language, which is a programming language that can be freely used and modified by users, for example Python, are often utilized. More specific tools for data analytics like R can be used for statistical analysis or graphical modeling.
- Data analysts also have help when reporting or communicating findings. Both Tableau and Power BI are data visualization and analysis tools to compile information, perform data analytics, and distribute results via dashboards and reports.

> Tableau is a popular tool for business intelligence which is used for the purpose of visually analyzing the data.

- Other tools are also emerging to assist data analysts. SAS is an analytics platform that can assist with data mining, while Apache Spark is an open-source platform useful for processing large sets of data.

Data analysts now have a broad range of technological capabilities to further enhance the value they deliver to their company.

INTRODUCTION TO LEGAL ANALYTICS

Traditionally, lawyers have won cases with only two weapons in their arsenal- Research and Reasoning. However, with legal data analytics, lawyers are gaining a critical advantage in their practice.

Legal Data Analytics: Meaning

Data analytics is the process of analyzing data collected over a long period of time to discover useful information. It can help gain insights on how the variables behave under different circumstances. Legal analytics serves as a supplement that boosts legal research. It can provide lawyers with the margin necessary for victory.

Legal analytics, with the help of data processing technologies, helps lawyers clean up, collate, and structurally analyze data. This data is collected from various cases, instances, law books, or online legal directories.

For example, through legal analytics, a lawyer can analyze past data to find out how long will a case run for, what were the previous outcomes of similar cases based on similar facts, how the judge behaved in similar scenarios, and so on.

With data analytics, lawyers can find out answers within minutes, which would have taken hours or weeks of manual efforts otherwise. Additionally, analytics software can catch simple variations in data that the human mind might not be able to comprehend.

Advantage of Legal Data Analytics:

Majority of legal professionals agree that using legal analytics makes them a better, more efficient, and effective legal practitioner.

1. **Edge over Competitors:**

   With data analytics, you get insights on judges' behavior, arguments made in the past, plea deals that have worked and so on. This data can help you predict the shape of the present case you are dealing with.

   In short, lawyers can build strong litigation strategies that succeed by deriving information from real-life scenarios. With data analytics, lawyers do not need to go over volumes of past case laws and law text books. Legal data analytics can strengthen the overall legal research methodology and enable the lawyers to arrive at a conclusion.

2. **Predict Results and Update Case Strategy:**

   Data analytics can aid lawyers in getting answers to client queries backed with illustrations and information. However, it is wrong to assume that legal analytics can only help lawyers in solving cases or doing research. It can also assist in making important business decisions related to the expansion or future planning of law firms. Other professionals in the legal domain can also use legal analytics. They can identify industry trends, run competitor analysis, study litigation history of opposition counsel, and so on.

3. **Exploring Strategies:**

   Lawyers can augment the ways of legal analytics to experiment with various legal strategies and scenarios. This way lawyers can predict case outcomes easily. Think of it as a science experiment where you play with the variables to come up with different results.

Implementing Data Analytics to Legal Practice:

Data in law industry can be broadly classified into individual data, law firm data, and industry data. Individual data is the data that one has in one's personal repository. The other two are broad-based organizational data available in law firm, companies, industry, on the internet, or in legal research libraries. With the advent of machine learning, artificial intelligence, and natural language processors, SaaS & service providers can make legal data analytics easy. They can do so by training computers to process a large amount of information at unmatchable speed and accuracy.

However, with the human element attached to each case, it cannot be said that legal output can be entirely depend on data analytics. We recommend to understand that legal analytics is tool of facilitation and improving efficiency of legal results rather than a replacement or a solution for all legal cases and client needs.

One has to note that Legal analytics support legal industry in decision-making in the following major areas:

- Compliance and regulatory support
- Internal investigations
- Fraud, waste, and abuse
- Incident response.

Business of Law Analytics vis-à-vis Practice Law Analytics[2]

Business of law analytics is used by legal operations and generally includes data related to legal spend, vendors, billing, and matters. These metrics enable functioning of legal operations as in a cost-controlled manner. Additionally, business of law analytics gives legal departments data-based evidence to prove how their work helps businesses.

1. **Spend analytics**

   Spend analytics are based on the total legal spend as well as details like spend to budget, spend by practice area, timekeeper rates, and more.

2. **Vendor analytics**

   Vendor analytics are based on assessment of the total list of vendors along with information like average matter cycle times, hourly rates, the use of alternative fee arrangements, and more. Vendor analytics also serve as evidence to back up departmental decisions to end vendor relationships. Vendor analytics enables assessment of the individual performance of each vendor, which is then compared with performance of other vendors, to understand which vendors are meeting expectations and thus deserve to be retained in the firm's portfolio.

   Without this type of detailed insight, it's difficult to figure out which vendors are (or aren't) worth working with.

3. **Billing analytics**

   Billing analytics are based on the total list of invoices along with information like status of invoice approval, status of accrual, violation of billing guidelines, and more. With insights from billing analytics, legal departments have a clearer idea of what's going on and can take the initiative to communicate with vendors and resolve issues before final invoices reach accounting.

2. Reproduced from Wen Kara (2022) What are the different types of legal data analytics? SimpleLegal. Available on https://www.simplelegal.com/blog/legal-data-analytics.

4. Matter analytics

Matter analytics are based on the total list of legal matters along with details like status of matters, vendors assigned to each matter, when the matter was worked, and more. Detailed matter analytics allow corporate legal departments to make educated decisions on using certain outside counsels vs. in-house counsels. After identifying the practice areas, they can use this information along with vendor costs to explore reducing the number of vendors and consolidating into law firms with multiple legal services. Legal departments can also assign standard, due diligence and compliance work to in-house attorneys to save money.

Practice of Law Analytics

Practice of law analytics gives in-house counsels insight that supports the legal practice, including data on legal contracts and cases. These metrics help attorneys with legal decision-making and risk management.

Contract Analytics

Contract analytics are based on the total list of legal contracts as well as detailed information on their status, type, renewal dates, content, and more.

As legal technology continues to evolve, contract analytics has also grown more advanced to include suggested revisions from artificial intelligence. While many basic legal platforms can sort and categorize contracts, AI legal software reviews them. With machine learning, these tools flag issues in contracts and offer changes based on their analysis of others contracts in the system. This saves lawyers "hundreds of hours" a year.

Case Analytics

Case analytics are based on the total list of legal cases along with information on prior and similar cases, case types, clients, judges, and more. Case analytics help in-house counsels to prepare their case strategy by learning from previous cases. With case analytics, attorneys can see trends on what worked and what didn't and apply those lessons to strengthen their legal approach towards complex legal issues.

INTRODUCTION TO MACHINE LEARNING FOR LAWYERS

Whenever any professional sector faces new technology, concern arises regarding how that technology will transform operations and the careers of those who choose that profession. And lawyers and the legal profession are no exception. Today, artificial intelligence (AI) is beginning to transform the legal profession in many ways and aims to take on higher-level tasks such as providing advice to clients, negotiating deals and drafting standard contracts.

Use of Artificial Intelligence and Machine Learning in Law[3]

Artificial intelligence mimics certain operations of the human mind and is the term used when machines are able to complete tasks that typically require human intelligence. The term machine learning is when computers use rules (algorithms) to analyze data and learn patterns and gain insights from the data. Artificial intelligence is aims to bring about a major shift in the way legal work is done.

Review Documents and Legal Research

AI-powered software improves the efficiency of document analysis for legal use and machines can review documents and flag them as relevant to a particular case. Once a certain type of document is denoted as

3. Bernard Marr and Company, How AI and Machine Learning Are Transforming Law Firms and The Legal Sector, Forbes. Available at https://bernardmarr.com/how-ai-and-machine-learning-are-transforming-law-firms-and-the-legal-sector/

relevant, machine learning algorithms can get to work to find other documents that are similarly relevant. Machines are much faster at sieving through documents as compared to humans and can produce output and results that can be statistically validated. They can help reduce the load on the human workforce by forwarding only those documents which are questionable rather than requiring humans to review all documents. Despite the monotony that legal research entails, it is important that legal research is done in a timely and comprehensive manner. AI systems such as the one offered by ROSS Intelligence leverages natural language processing to help analyze documents.

For example, when lawyers using AI-powered software flag certain documents as relevant, the AI learns what type of documents it's supposed to be looking for. Hence, it can more accurately identify other relevant documents. This is called "predictive coding." Predictive coding offers many advantages over old-school manual document review. Among other things, it:

- leverages small samples to find similar documents;
- reduces the volume of irrelevant documents that attorneys must wade through;
- produces results that can be validated statistically;
- is at least modestly more accurate than human review;
- is much faster than human review.

Help perform Due Diligence

In law offices around the world, paralegals and other legal professionals are kept busy conducting due diligence to uncover background information on behalf of their clients. This works includes confirming facts and figures and thoroughly evaluating the decisions on prior cases to effectively provide counsel to their clients. Artificial intelligence tools can help these paralegals to conduct their due diligence more efficiently and with more accuracy since this work is often tedious for humans.

Contract Review and Management

A big portion of work that legal professionals do on behalf of clients is to review contracts to identify risks and issues with how contracts are drafted that could have negative impacts for their clients. They redline items, edit contracts and counsel clients if they should sign or not or help them negotiate better terms.

For example, analysis of all contracts a company has signed can identify risks, anomalies, future financial obligations, renewal and expiration dates, etc. For companies with hundreds or thousands of contracts, this can be a slow, expensive, labour-intensive, and error-prone process (assuming the contracts aren't already entered into a robust contract management system). It's also boring for the lawyers (or others) tasked with doing it.

On a day-to-day basis, lawyers review contracts, make comments and redlines, and advise clients on whether to sign contracts as it is or try to negotiate better terms. These contracts can range from simple (e.g., NDAs) to complex (share subscription agreements). A backlog of contracts to review can create a bottleneck that delays deals (and the associated revenues). Lawyers (especially inexperienced ones) can miss important issues that can come back to bite their clients later.

AI can help analyze contracts in bulk as well as individual contracts.

Predict Legal Outcomes

AI has the capability of analyzing data to help it make predictions about the outcomes of legal proceedings better than humans. Clients are often asking their legal counsels to predict the future with questions such as "If we go to trial, how likely will it be that I win?" or "Should I settle?". With the use of AI that has access to years of trial data and record, lawyers are able to better predict the answers to such questions.

Legal Research

Any lawyer who has ever done research using legal research tools has used legal automation. Finding relevant cases in previous and recent years involves the laborious process of looking up headnote numbers and going through voluminous paper volumes. But AI takes legal research to the next level. For example, Ross Intelligence uses the power of IBM's Watson supercomputer to find similar cases. It can even respond to queries in plain English. The power of AI-enabled research is striking: using common research methods, a bankruptcy lawyer found a case nearly identical to the one he was working on in 10 hours; Whereas Ross's AI found it almost instantly.

Data Incorporated into Legal Analytics [4]

Legal analytics tools collect three broad categories of data and turn it into useful insights:

- Individual data,
- Internal data, and
- Legal industry data.

Individual data is information about a firm's current and future clients, including information that is collected whenever someone peruses the firm's website. This encompasses information collected from consultation request forms and data about how potential clients interact with the website, such as search terms used and time spent on particular pages. Individual data can also include client data stored within a firm or corporate law firm's secure file storage system. With that data, certain legal analytics tools can analyze documents and document metadata to inform case management and guide eDiscovery strategy.

> **Note:** eDiscovery is a resource that seeks to find evidence in emails, business communications and other electronic data and collects, preserves, reviews and exchanges information in electronic format for use in litigation and investigation.

Internal data is information about a firm or law department's business practices, such as billing rates, legal expenditure incurred in practice areas, and billable hours. This data might also include productivity data related to individual lawyers and support staff.

Legal industry data is data that has been externally collected from outside research groups, court dockets, and sources that cover the wider trends within the legal community. This might include data about case outcomes or trends about hot practice areas or changing client preferences.

Ways to use Legal Analytics[5]

Legal analytics has changed over the years, and providers are constantly updating their software and improving functionality to adapt. While early legal analytics software generally provided fewer features, the advanced tools available today serve a wide range of goals for legal professionals. Below are five ways to use legal analytics tools.

1. **Streamlining eDiscovery**

   While eDiscovery was once tedious and time-consuming, legal analytics tools have helped lawyers simplify its complex processes, making it both more efficient and more useful for determining likely

4. Reproduced from Wolff Jeffrey (2021) 5 Ways to Use Legal Analytics Tools to Work Smarter, Not Harder, JD SUPRA. Available at https://www.jdsupra.com/legalnews/5-ways-to-use-legal-analytics-tools-to-5039823/

5. Reproduced from Wolff Jeffrey (2021) 5 Ways to Use Legal Analytics Tools to Work Smarter, Not Harder, JD SUPRA. Available at https://www.jdsupra.com/legalnews/5-ways-to-use-legal-analytics-tools-to-5039823/

case outcomes and guiding litigation or settlement strategies. These tools detect trends and patterns in client data using artificial intelligence to help lawyers understand data better and make better strategic decisions.

- Structural analytics can give an overview of the scope of a data set. For example, analytics tools can study file metadata to organize files, while email threading and duplicate and near-duplicate detection can streamline data sets.
- Conceptual analytics can help lawyers understand the relationships between documents and provide a high-level overview of a corpus of documents. These tools can recognize concepts and use them to group related documents together and can even identify the sentiment or emotion behind documents, flagging concerning documents for review.
- Finally, predictive analytics, such as technology-assisted review, can prioritize the documents most likely to be relevant and make informed predictions about privileged documents or content; they can also detect and classify contract clauses.

With these insights, legal teams can conduct more accurate early case assessment, reduce the volume of eDiscovery data advancing to review, accelerate and improve the quality of document review, and substantially lower the total cost of eDiscovery.

2. Facilitating Law Firm Marketing and Increasing Client Satisfaction

A law firm is, quite literally, nothing without its clients. Fortunately, legal analytics tools have revolutionized lawyers' ability to recognize "ideal" clients and design effective marketing strategies to capture those clients. Many helpful tools can analyze the characteristics of the ideal client, identify the target market to maximize profitability, and channel business to the firm.

These tools can also help with benchmarking. For example, law firm can compare their performance against peer firms. Similarly, instead of wasting money on unsuccessful marketing campaigns, legal analytics tools can help a firm craft a plan—based on real data—to target and draw in lucrative clients and increase the firm's profitability.

3. Making Informed, Risk-based Litigation Decisions

Clients are in the pesky habit of asking their lawyers to predict the future, but few lawyers are naïve enough to make the attempt. Predictive legal analytics tools, however, give lawyers the next best thing: a way to accurately predict the specific characteristics of a case based on real case data.

For example, modern analytics tools can predict the likelihood of winning a case, the probable length of litigation, and the number of hours that the firm should allocate to working on it, all based on past data. This insight is invaluable for saving time and money because it lets the counsel know which battles are worth fighting and which should be promptly settled (and for how much). The best part is that these predictions are personalized based on the specific jurisdiction, judge, and even opposing counsel in question.

Note: Jurisdiction refers to the authority of a court to hear, decide and rule on cases within a particular geographical limit. In other words, it is the territory over which the court can exercise its authority.

4. Improving Legal Research

Lawyers are intimately familiar with legal research and are accustomed to digging through reams of case law to identify the nuggets of wisdom they need. Luckily, legal research has been revolutionized in the past few decades so that lawyers no longer need to spend hours flipping through dense books in the law library. Instead, relevant legal precedents from around the country, and even around the world, are available with a few keystrokes and clicks on a keypad through popular legal research websites like SCC and Manupatra.

Analytics tools for legal research save lawyers time and effort—and helps them win cases—by rapidly identifying relevant and significant cases that can inform their arguments and strategy. Searches can be narrowed by practice area, date, location, or even the number of times that a search term appears in the result. All of these capabilities arise from data analytics.

Therefore, in this manner, not only do legal research tools create an easy way to sort through legal cases, but they also provide access to treatises, pleadings, guidelines laid down by the court, and other helpful materials. In fact, many state bar associations provide free access to legal research tools so that all lawyers can benefit from this powerful technology.

5. Promoting Lawyer Productivity

Legal analytics don't just help clients; legal analytics software can also tell partners and heads of law firm which lawyers are the most productive, efficient, and cost-effective and which lawyers need to improve. This data is critical for making decisions about assignments and promotions and deciding how to manage attorneys and their workload.

For instance, with legal analytics tools that track billable hours, you can see how long a lawyer spends working on a specific case and dedicate more expensive resources (partners and senior associates) for only the most important work. Law firms and corporate law offices can also explore hiring remote lawyers and see how the costs and benefits stack up against full-time employees.

Not only can managers get unparalleled access to data about associates, but legal analytics can also provide tools to help underperforming lawyers get back on track. Many products have attorney dashboards, automated document management, and client collaboration tools to boost productivity.

QUANTITATIVE LEGAL PREDICTION VIS-À-VIS BUSINESS OF LAW[6]

Overview of Quantitative Legal Prediction (QLP)

This current wave of legal artificial intelligence has abandoned the idea of algorithms mimicking the thought process of lawyers. Rather, the predictive justice algorithms turn to quantitative approaches, i.e., utilizing brute-force processing of data.[7] The development in this field is propelled by the increasing computing power, declining data storage costs, better access to data and improvements in machine learning and other artificial intelligence technologies.[8]

Specifically, quantitative legal prediction is based on supervised machine learning techniques which employ data of previous cases as input to predict the result of a future case.[9] This method applies statistical means to "induce a prediction model (or function) from a dataset that can be used to predict an outcome for a new case."[10]

6. Reproduced from Trasberg Henrik (2019) Quantitative Legal Prediction and the Rule of law, Master Thesis, Law and Technology LLM, TILBURG University Law School.

7. Greenleaf (n 17) 312; For a further description of utilizing big data for building legal prediction models see Richard Susskind and Daniel Susskind, The Future of the Professions: How Technology Will Transform the Work of Human Experts (Oxford University Press 2015) 226–228 and 276–278.

8. An overview of advances in computer hardware and artificial intelligence powering the development of quantified legal prediction is provided by Katz (n 5) 913–923.

9. It is regarded as "supervised" since it involves inferring a classification model from labeled training data. See Ashley (n 6) 109. Ashley further explains the functioning of supervised machine learning: "The training data comprise a set of examples that have been assigned outcomes. Each example is a pair consisting of an input object (often a vector of feature values) and a desired output value. The learning algorithm needs to generalize from the training data to unseen situations."

10. ibid.

Quantitative Legal Prediction thus creates an "inverse" model in which correlations are created between certain elements available in the case (such as specific words used) and the outcome of a case, as a result of which certain value is assigned to each element.

As described by Katz, "simply put, one uses the observables to build the model rather than using the model to assign causal weight to those observables."[11] In context of legal prediction, this means that certain parameters or features of a case (e.g. who is the judge, what is the subject-matter of the case, which words or phrases are present in the case documentation, etc.) are assigned a value that would indicate which way a case would be decided based on past patterns.[12] It is worth noting that the QLP tools will not discover the features that influence outcomes, but instead learn the weights of such features.[13]

Quantified legal prediction is what Daniel Katz categorizes as "soft artificial intelligence" in the sense that the algorithms aim to achieve outcome that would mimic human intelligence while the process to achieve the outcome does not.[14] There is thus no legal reasoning employed in the predictions of the quantified prediction models. Importantly, data-based prediction models require elements that are relatively simple to extract – there is no input needed from the larger body of legal system. There is also no knowledge representation bottleneck that arises with the legal expert systems as the model does not require semantic understanding or the context of the information extracted.

Applications of Quantitative Legal Prediction

QLP can serve many purposes, one of the more intriguing of which is predicting outcomes of court cases – seemingly with a better precision than legal experts. An example of such functionality is the algorithm developed by Katz et al. which enables prediction of US Supreme Court case outcomes with 70.2 % accuracy.[15] While such prediction models largely dismiss legal causality, Katz considers that "it is not always necessary to have a deep theory in order to generate a well-functioning prediction engine."[16] The algorithm by Katz et al. uses input information about the case (e.g., who are the parties, type of law, the source circuit court, the judgement at lower court, issue area), background information (e.g. name of judge(s), age and gender of judge, party of the appointing president) and trends (e.g., historic trends of Supreme Court, current trends of Supreme Court, trends of lower courts, trends specific to individual Supreme Court justice) in creating the outcome prediction.[17] In a study conducted by Blackman et al. in 2012, three distinguished legal experts were tasked with predicting outcomes of 171 SCOTUS cases, predicting correctly only 59 % of the cases.[18] While this sample is too small to make exhaustive conclusions, it does give promise to the idea that QLP can predict case outcomes, on average, with a better accuracy than expert lawyers and could thus emerge as a genuine mechanism for making litigation decisions.

Using meta-data from previous cases to draw conclusions about emerging legal matters is also employed by many commercialized predictive justice products (for example, Premonition, Blue J Legal, Courtquant and Predictice). In some domains of law this may be particularly effective: a prediction model developed by Dunn et al. was able to predict judge's ruling in asylum applications with 80 % accuracy while merely having information about the judge and the applicant's nationality.[19] Blue J Legal, which creates its predictions by combining data-

11. Katz (n 5) 952

12. See ibid 952–953.

13. Ashley (n 6) 125.

14. See Katz (n 5) 913 and 918.

15. Katz, Bommarito and Blackman (n 7).

16. Katz (n 5) 950.

17. Katz, Bommarito and Blackman (n 7) 7.

18. Blackman and Carpenter (n 8).

19. Dunn and others (n 25).

based prediction based on patterns from prior case laws with input provided by the user, is able to consistently predict correctly over 90 % of outcomes of tax law cases in Canada.[20]

Note: Meta-data refers to that data which provides information about one or more aspects of data.

The algorithm developed by Aletras et al., which predicts cases of European Court of Human Rights with a 79 % accuracy, has adopted a different direction – instead of assigning value to certain meta-data, it creates correlations between sequences of words (n-grams)[21] available in certain parts of a case (e.g., the legal circumstances and facts) and case outcome. For example, the term 'second applicant' has a significant correlation with ECtHR identifying an infringement. Another similar model was able to predict ECtHR i.e. European Court of Human Rights, case outcomes with 75 % accuracy.[22] Potentially, such algorithm could be provided with the pleadings/complaint concerning the case and/or other case documentation as input and, based on the language used in the input, it predicts (the likelihood of) positive/negative outcome.

There are many other models developed to predict, for example, security fraud class actions and settlement amounts,[23] likelihood of patent being litigated,[24] and appeal decisions in tax law in Germany.[25] The knowledge that these tools provide can enable valuable input for significant litigation decisions, for example whether to litigate or not and whether to seek compromise or discard specific claims. One could envision that if the prediction accuracy of these models continues to significantly rise, they may be adopted by courts and other public institutions for making administrative decisions and solving legal disputes.

In addition to case outcome prediction, QLP also provides other benefits. To bring a few examples, QLP-based algorithms can mine and aggregate information from past case precedents that enables, for example, to predict how long the case might take, what are the likely legal costs, what is the best way to remedy a particular legal dispute, etc.[26] But most significantly, predictive justice tools can identify certain patterns about how the courts and the judges operate, such as which prior case laws a particular judge likes to refer to, how the judge tends to decide in non-obvious cases in respect of a particular subject-matter, whether there are certain lawyers that tend to win with a particular judge, what are the types of arguments that the judge tends to embrace, etc. These patterns, which are largely based on the data about how the judge has ruled in the past, render the adjudication process more transparent as they provide insight into what a particular judge might find relevant in respect of a certain legal debate. This enables legal counsels to make better strategic decisions and tailor the legal arguments within the adjudication process for the particular judge. In light of the notion that the judge's decision is influenced by its attitudes, biases and other cognitive processes, the input into tendencies of a judge can have transformational impact on the transparency of adjudication and the predictability of a judge's actions.

As another use-case, there are some QLP-based software being adopted by the stake-holders before the case even proceeds for trial, i.e. at the pre-trial stage. For example, a few US courts are using data-based tools to assess the risk of recidivism of convicted criminals – a major factor in contemplating the choice and extent of

20. Benjamin Alarie, Anthony Niblett and Albert Yoon, 'Using Machine Learning to Predict Outcomes in Tax Law' (2017) SSRN <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2855977>

21. N-grams are contiguous word sequences within the text. See Aletras and others (n 9) 8–9.

22. Masha Medvedeva, Michel Vols and Martijn Wieling, 'Judicial Decisions of the European Court of Human Rights: Looking into the Crystall Ball' (2018) University of Groningen, Department of Legal Methods 1 <http://martijnwieling.nl/files/Medvedeva-submitted.pdf>.

23. Blakeley B Mcshane and others, 'Predicting Securities Fraud Settlements and Amounts: A Hierarchical Bayesian Model of Federal Securities Class Action Lawsuits' (2012) 9 Journal of Empirical Legal Studies 482.

24. Colleen V Chien, 'Predicting Patent Litigation' (2011) 90 Texas Law Review 283.

25. Bernhard Waltl and others, 'Predicting the Outcome of Appeal Decisions in Germany's Tax Law' (2017) 10429 (eds) Electronic Participation. ePart 89 <http://link.springer.com/10.1007/978-3-319-64322-9_8>.

26. Katz (n 5); Hildebrandt, 'Law as Computation in the Era of Artificial Legal Intelligence. Speaking Law to the Power of Statistics' (n 12) 11.

punishment.[27] In law enforcement, predictive justice tools are used for assessing whether a suspect should be held in pre-trial detention or not.[28]

Note: Recidivism refers to the tendency of a convicted criminal to relapse into criminal behavior.

What QLP thus promises is an augmented insight into what might be the outcome of a particular case, enabling the counsel to make smarter whether to commence litigation or not. This value should not be understated – lawyers are surprisingly bad at predicting outcomes of court cases that exceed certain basic level of difficulty.[29]

Furthermore, while a lawyer's ability to reason far exceeds any algorithm, humans' capacity to sift through thousands of court cases and identify correlations from historic patterns that may play a significant role in the outcome of a current case or provide insight into behavior of the judge is dwarfed compared to the ability of AI. QLP thus enables lawyers to overcome some of the cognitive limitations of the human brain.

Limitations of Quantitative Legal Prediction

Despite its potential, QLP simultaneously contains some significant limitations which hinders its adoption within the adjudication process.

- Firstly, the usefulness of a prediction algorithm depends on identifying variables that reliably correlate with case outcome while also being readily available prior to or in an early stage of the adjudication. For example, the model developed by Aletras et al. was trained by seeking correlations between case outcome and the procedure, facts, law and circumstances of the case that were written in the actual judgment. It is evident that the choices made in drafting the document on facts and circumstances of the case will be significantly influenced by the outcome that is established in the same judgment. Thus, it is unknown (not to say: extremely unlikely) that the model of Aletras et al. would have any reliability when it is ordered to predict outcome based on the input that is extracted from the case pleadings/document or any other documentation available before the judgment.
- Secondly, many of the correlations between input data and output that the QLP algorithms identify may be misleading. As framed by Cashwell and Pasquale, "Any student of statistics knows, if one tests enough data sets against one another, spurious correlations will emerge."[30] There is thus a considerable likelihood that the correlations between majority of the n-grams or other factors taken into account by the QLP algorithms, in reality, do not provide any useful knowledge about the outcome but are, in fact, spurious.
- In addition to the (current) technical limitations which restrict the usefulness of the QLP algorithms, their approach to solving legal disputes fundamentally clashes with the notion of law and the rule of law, causing a great number of legal challenges. The issue emerges from the fact that the output of QLP is based on the seemingly random correlations that the algorithm has identified, rather than "inferences from any causal model."[31]As such, the insights it provides ignore the substantial merits of the case. In fact, the current quantitative legal prediction models give no regard to semantic understanding of legal

27. For example, Compas software by Equivant is used by several courts in US. See Adam Liptak, 'Sent to Prison by a Software Program's Secret Algorithm', The New York Times, 01.05.2017. <https://www.nytimes.com/2017/05/01/us/politics/sent-to-prison-by-a-software-programs-secretalgorithms.html>. However, it must be noted that using the Compas software has been met with significant scepticism – see for example Jason Tashea, 'Courts Are Using AI to Sentence Criminals. That Must Stop Now', Wired Magazine, 17.04.2017 <https://www.wired.com/2017/04/courts-using-ai-sentence-criminals-must-stopnow/>. See also Ed Yong, 'A Popular Algorithm Is No Better at Predicting Crimes Than Random People', The Atlantic, 17.01.2018. <https://www.theatlantic.com/technology/archive/2018/01/equivant-compasalgorithm/550646/>

28. Nick Statt, 'UK police will start using AI to decide whether suspects should be kept in custody', The Verge, 10.05.2017. <https://www.theverge.com/2017/5/10/15614980/uk-durham-police-ai-risk-assessment-policing>

29. See analysis of the study by Jane Goodman-Delahunty, Maria Hartwig, and Elizabeth F Loftus in Section 3.3.

30. Pasquale and Cashwell (n 19) 9.

31. See Greenleaf (n 15) 312.

arguments or legal deduction. Consequently, these algorithms are not able to provide an explanation, not to mention legal reasoning, for the conclusions reached. On this note, Susskind argues that in many areas "we can develop high performing, non-thinking machines that can outperform the best human experts, even though they go about their business in quite unhuman ways" and thus "we will not need to understand and then replicate the way human experts work, nor will we need to develop thinking machines to replace much of the work currently undertaken by human professionals."[32] Yet, does our idea of law permit a situation where important (legal) decisions have no causative basis in law?

- Evidently, it raises some significant concerns. For example, as the output of QLP is not cast into form of legal argumentation and the reasoning (logic) based on which the algorithms reach conclusions remains opaque and unintelligible, the decision or a suggestion of a QLP-based algorithm effectively cannot be contested.
- Furthermore, these models are never unbiased, objective nor neutral in their prediction; but are affected by a series of design decisions by its developers."[33] Notably, developing the model requires a trade-off, between volume, relevance, completeness, accuracy and correctness of the training dataset, the dimensionality and aptitude of the hypotheses space, the time taken for iterant testing, and the availability of the relevant domain expertise."[34] Since these choices will significantly influence the conclusions reached by the model, it essentially means that the developers of the algorithms will have the power to direct the output of a model and thereby influence legal decision-making.

The above-described limitations hint towards some of the rule of law implications that arise with the emergence of data-based legal decision-making.

Using QLP to Predict Court Case Outcomes

Adoption of QLP algorithms to predict case outcome essentially provides individuals a risk management tool when deciding whether to adjudicate or not. This risk management tool can predict a case outcome, on average, with a better accuracy than a lawyer, promising a new era in how litigation decisions are made. Yet, I claim that widespread adoption of QLP for predicting case outcome and using these predictions for making litigation decisions undermines the value of adjudication and thereby the realization of rule of law, as it dismisses the role of adversarial legal argumentation.

> **Note:** Adjudication refers to the legal process by which a judge gives a decision which determines the rights an liabilities of the parties involved in a case.

The legal disputes for which insight from QLP is sought are presumably cases where no axiomatic solutions exist. Questions of law in such cases cannot be solved purely by having knowledge of the legal system or the existing case law, which is why relying on a data-based prediction tool that has a good track record seems tempting. Yet, such approach misrepresents the adjudication process: It is argued that it is the adversarial legal argumentation that takes place within the adjudication process, which forges the result of non-obvious cases. Therefore, the outcome of a dispute (i.e., how legal rule applies in the particular situation) is created within the adjudication process. Consequently, what a lawyer does in the adjudication process is not just a prediction – the legal arguments it creates and submits are proposals for regulation of behavior which, if convincing to the court, are materialized into a legal norm. It is only through this adversarial argumentation that the resolution on how the disputed behavior is to be regulated comes into existence. As such, QLP misrepresents adjudication by purporting predetermination where there is none.

32. Susskind and Susskind (n 35) 276; See also Greenleaf (n 17) 312.

33. Mireille Hildebrandt, 'Algorithmic Regulation and the Rule of Law' (2018) <http://dx.doi.org/10.1098/rsta.2017.0355>.

34. Hildebrandt, 'Law as Computation in the Era of Artificial Legal Intelligence. Speaking Law to the Power of Statistics' (n 12) 10.

The principal issue with the above emerges from the fact that the adjudication process carries a fundamental role in safeguarding the realization of rule of law. The principle of rule of law postulates that no one is above the law and law applies equally to all citizens irrespective of their ranks, other distinctions, etc. The rule of law is evidently an elusive concept with a wide array of objectives, but one of its key aspects is treating individuals as moral agents entitled to dignity and respect. The adjudication process ensures such treatment by accepting that law has an arguable character and by providing a procedure where a question of how a certain behavior should be governed is argued. Thus, the adjudication process is not only a place for reaching a conclusion, but functions as a forum where the understanding of the law is debated and formed, and the individual is entitled to participate and influence that procedure. The preemptive character of QLP impedes the adjudication process in acting as such an institution.

Using QLP for Profiling Judges and its Rule of Law Implications

The second significant rule of law interference relates to applying QLP technology for identifying and exploiting specific patterns of judges, with the purpose of exploiting these tendencies for one's advantage. The commercialized products that have adopted QLP technology increasingly promise the possibility to outline certain patterns of a judge to better understand what are the arguments, cases, types of evidence, etc., that a particular judge likes, enabling a more insightful prediction on what might be the most fruitful strategy or a set of argumentations for a particular case. The value proposal of such use of QLP is thus a more transparent and predictable adjudication through identification of adjudication patterns otherwise inaccessible to human cognition.

Simultaneously, the adjudication process is riddled with biases and inconsistencies. The more significant the patterns these predictive justice tools are able to identify, the more it enables to exploit the disparities that exist within the judiciary. If patterns of judges become increasingly visible, the extent to which the biases are exploited will inevitably increase, tilting case outcomes by engaging factors that are not rooted in law or the legal system at large. This apparent lack of concern for substantive justice would undermine fair adjudication as well as erode public confidence in the adjudication system.

Joseph Raz states that insufficient regard for the rule of law can lead to uncertainty or it may lead to frustrated and disappointed expectations.[35] Adoption of QLP-based algorithms for identifying patterns of adjudication affects both of these aspects – on one hand it enhances transparency and foreseeability; on the other, it enables unfair exploitation of the system, eroding public confidence and giving rise to frustration about the adjudication process by the courts.

Reconciliation comes in the form of adequate reaction to the developments enabled by QLP – in order to mitigate the risks on fairness of adjudication and the perception thereof. The emergence of QLP requires attention and willingness from the judiciary to evolve adjudication in line with the capabilities of the QLP technology. This includes the need to further understand, embrace and develop a theoretical framework about the role of biases within adjudication and provide judicial education on understanding significance of the adjudication data extracted by QLP.

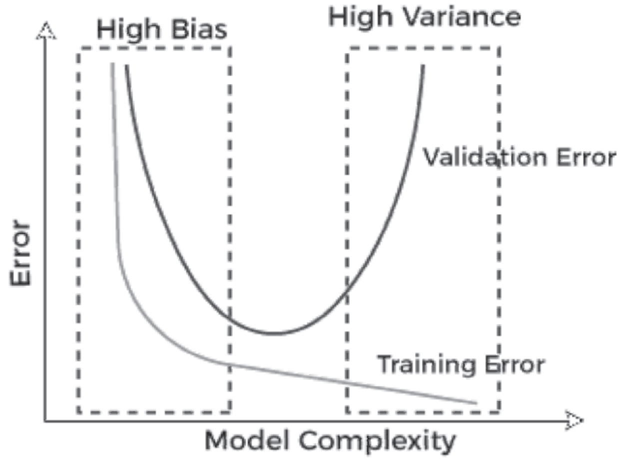
BIAS/ VARIANCE, PRECISION/RECALL & DIMENSIONALITY

Bias and Variance in Machine Learning[36]

Machine learning is a branch of Artificial Intelligence, which allows machines to perform data analysis and make predictions. However, if the machine learning model is not accurate, it can make predictions errors, and these prediction errors are usually known as Bias and Variance. In machine learning, these

35. See Raz (n 22) 222.

36. Bias and Variance in Machine Learning, Java T Point. Available at https://www.javatpoint.com/bias-and-variance-in-machine-learning errors will always be present as there is always a slight difference between the model predictions and actual predictions. The main aim of ML/data science analysts is to reduce these errors in order to get more accurate results. In this portion, we are going to discuss bias and variance, Bias-variance trade-off, Underfitting and Overfitting.

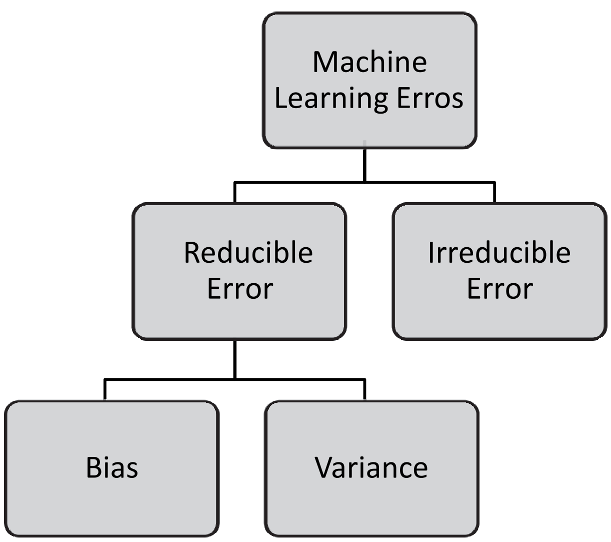


Source: https://www.javatpoint.com/bias-and-variance-in-machine-learning

Errors in Machine Learning

In machine learning, an error is a measure of how accurately an algorithm can make predictions for the previously unknown dataset. On the basis of these errors, the machine learning model is selected that can perform best on the particular dataset. There are mainly two types of errors in machine learning, which are:

- ***Reducible Errors:*** These errors can be reduced to improve the accuracy of the mode. Such errors can further be classified into bias and Variance.



Source: https://www.javatpoint.com/bias-and-variance-in-machine-learning

- ***Irreducible Errors:*** These errors will always be present in the model regardless of which algorithm has been used. The cause of these errors is unknown variables whose value cannot be reduced.

Let's understand the entire concept of bias, in detail.

What is Bias?

In general, a machine learning model analyses the data, find patterns in it and make predictions. While training, the model learns these patterns in the dataset and applies them to test data for prediction. While making predictions, a difference occurs between prediction values made by the model and actual values/expected values, and this difference is known as bias errors or Errors due to bias. It can be defined as an inability of machine learning algorithms such as Linear Regression to capture the true relationship between the data points. Each algorithm begins with some amount of bias because bias occurs from assumptions in the model, which makes the target function simple to learn. A model has either:

- **Low Bias:** A low bias model will make fewer assumptions about the form of the target function.
- **High Bias:** A model with a high bias makes more assumptions, and the model becomes unable to capture the important features of our dataset. A high bias model also cannot perform well on new data.

Generally, a linear algorithm has a high bias, as it makes them learn fast. The simpler the algorithm, the higher the bias it has likely to be introduced. Whereas a nonlinear algorithm often has low bias.

Some examples of machine learning algorithms with low bias are Decision Trees, k-Nearest Neighbors and Support Vector Machines. At the same time, an algorithm with high bias is *Linear Regression, Linear Discriminant Analysis and Logistic Regression.*

Ways to reduce High Bias:

High bias mainly occurs due to a much simple model. Below are some ways to reduce the high bias:

- Increase the input features as the model is underfitted.
- Decrease the regularization term.
- Use more complex models, such as including some polynomial features.

What is a Variance Error?

The variance would specify the amount of variation in the prediction if the different training data was used. In simple words, *variance tells that how much a random variable is different from its expected value.* Ideally, a model should not vary too much from one training dataset to another, which means the algorithm should be good in understanding the hidden mapping between inputs and output variables. Variance errors are either of *low variance or high variance.*

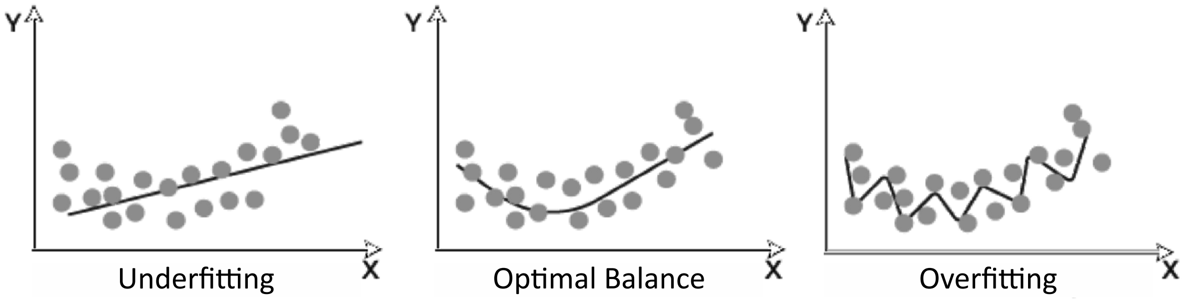
Low variance means there is a small variation in the prediction of the target function with changes in the training data set. At the same time, High variance shows a large variation in the prediction of the target function with changes in the training dataset.

A model that shows high variance learns a lot and performs well with the training dataset, and does not generalize well with the unseen dataset. As a result, such a model gives good results with the training dataset but shows high error rates on the test dataset.

Since, with high variance, the model learns too much from the dataset, it leads to overfitting of the model. A model with high variance has the below mentioned problems:

- A high variance model leads to overfitting.
- Increased model complexities.

Usually, nonlinear algorithms have a lot of flexibility to fit the model, have high variance.



Source: https://www.javatpoint.com/bias-and-variance-in-machine-learning

Some examples of machine learning algorithms with low variance are, Linear Regression, Logistic Regression, and Linear discriminant analysis. At the same time, algorithms with high variance are decision tree, Support Vector Machine, and K-nearest neighbours.

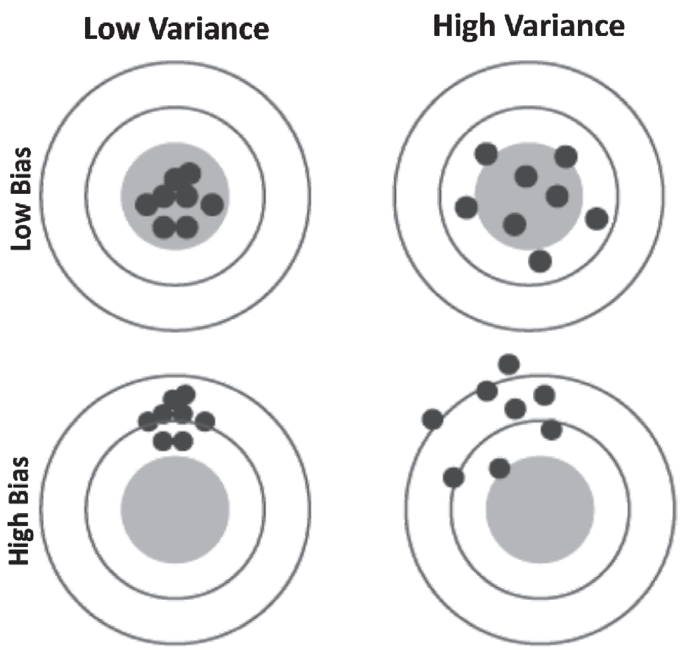
Ways to Reduce High Variance:

Reduce the input features or number of parameters as a model is overfitted.

- Do not use a much complex model.
- Increase the training data.
- Increase the Regularization term.

Different Combinations of Bias-Variance

There are four possible combinations of bias and variances, which are represented by the diagram below:



Source: https://www.javatpoint.com/bias-and-variance-in-machine-learning

1. **Low-Bias, Low-Variance:**

   The combination of low bias and low variance shows an ideal machine learning model. However, it is not possible practically.

2. **Low-Bias, High-Variance:**

   With low bias and high variance, model predictions are inconsistent and accurate on average. This case occurs when the model learns with a large number of parameters and hence leads to an overfitting

3. **High-Bias, Low-Variance:**

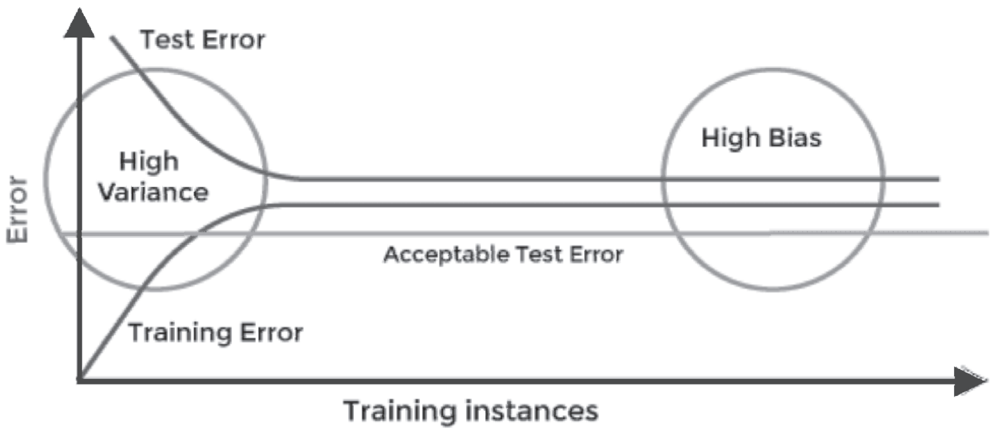
   With high bias and low variance, predictions are consistent but inaccurate on average. This case occurs when a model does not learn well with the training dataset or uses few numbers of the parameter. It leads to underfitting problems in the model.

4. **High-Bias, High-Variance:**

   With high bias and high variance, predictions are inconsistent and also inaccurate on average.

How to identify High variance or High Bias?

High variance can be identified if the model has:



Source: https://www.javatpoint.com/bias-and-variance-in-machine-learning

- Low training error and high-test error.

High Bias can be identified if the model has:

- High training error and the test error is almost similar to training error.

Trade Offs and Bias

Trade-offs are the one thing that is between sane but boring life and complicated, risky but adventurous life. At every point in life, even every second, we make some kind of 'Trade-Off'. Risky or not, Trade-Offs always help us to find the sweet spot or middle ground. As we are turning the machine to think like humans, they too, are haunted by 'Trade-Offs'.

Machine Learning mostly have to deal with two Trade-offs:

- Bias-Variance Trade-offs
- Precision-Recall Trade-offs

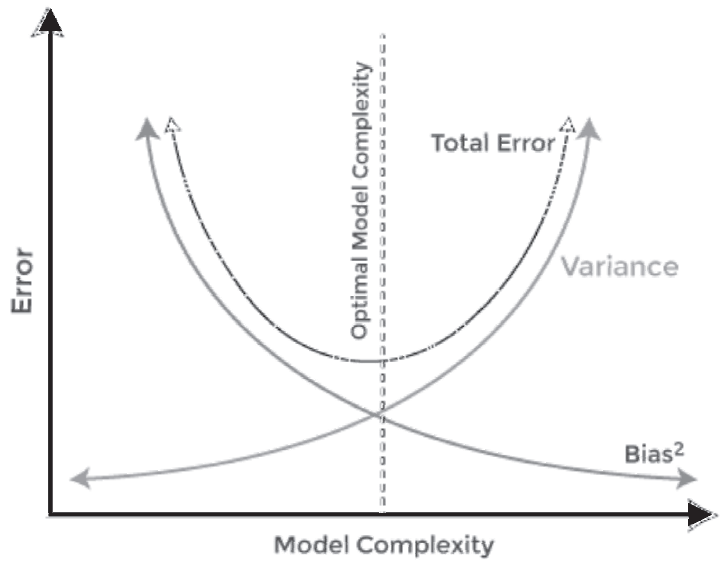
Bias-Variance Trade-Off

While building the machine learning model, it is really important to take care of bias and variance in order to avoid overfitting and underfitting in the model. If the model is very simple with fewer parameters, it may have low variance and high bias. Whereas, if the model has a large number of parameters, it will have high variance and low bias. So, it is required to make a balance between bias and variance errors, and this balance between the bias error and variance error is known as the Bias-Variance Trade-Off.

For an accurate prediction of the model, algorithms need a low variance and low bias. But this is not possible because bias and variance are related to each other:

- If we decrease the variance, it will increase the bias.
- If we decrease the bias, it will increase the variance.

Bias-Variance trade-off is a central issue in supervised learning. Ideally, we need a model that accurately captures the regularities in training data and simultaneously generalizes well with the unseen dataset. Unfortunately, doing this is not possible simultaneously. Because a high variance algorithm may perform well with training data, but it may lead to overfitting to noisy data. Whereas, high bias algorithm generates a much simple model that may not even capture important regularities in the data. So, we need to find a sweet spot between bias and variance to make an optimal model.



Source: https://www.javatpoint.com/bias-and-variance-in-machine-learning

Hence, ***the Bias-Variance trade-off is about finding the sweet spot to make a balance between bias and variance errors.***

Overfitting, Underfitting, & Cross-Validation[37]

Overfitting and Underfitting are the two main problems that occur in machine learning and degrade the performance of the machine learning models. It is pertinent to note that 'Machine Learning' is a type of Artificial Intelligence that enables systems and software applications to become more accurate at making prediction of outcomes, without being explicitly programmed to do so.

37. Reproduced from Overfitting and Underfitting in Machine Learning, Java T Point. Available at https://www.javatpoint.com/overfitting-and-underfitting-in-machine-learning

The main goal of each machine learning model is to generalize well. Here generalization defines the ability of an ML model to provide a suitable output by adapting the given set of unknown input. It means after providing training on the dataset, it can produce reliable and accurate output. Hence, the underfitting and overfitting are the two terms that need to be checked for the performance of the model and whether the model is generalizing well or not.

Before understanding the overfitting and underfitting, let's understand some basic term that will help to understand this topic well:

- **Signal:** It refers to the true underlying pattern of the data that helps the machine learning model to learn from the data.
- **Noise:** Noise is unnecessary and irrelevant data that reduces the performance of the model.
- **Bias:** Bias is a prediction error that is introduced in the model due to oversimplifying the machine learning algorithms. Or it is the difference between the predicted values and the actual values.
- **Variance:** If the machine learning model performs well with the training dataset, but does not perform well with the test dataset, then variance occurs.

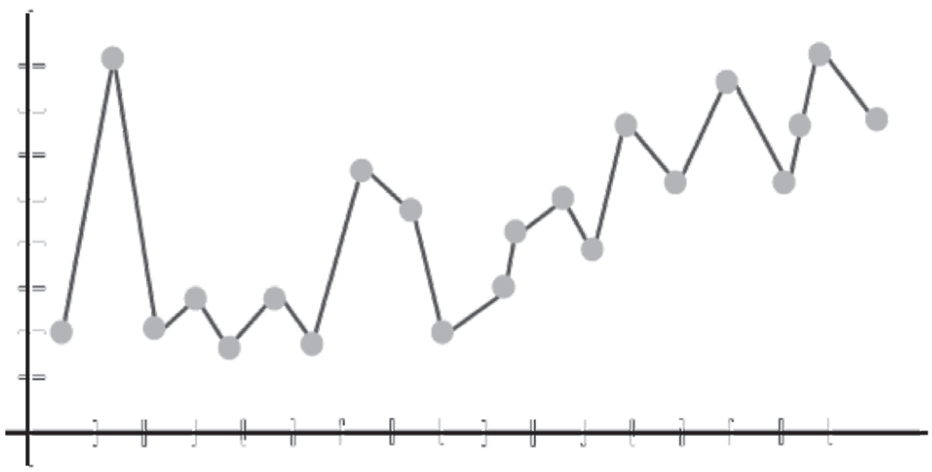
Overfitting

Overfitting occurs when our machine learning model tries to cover all the data points or more than the required data points present in the given dataset. Because of this, the model starts caching noise and inaccurate values present in the dataset, and all these factors reduce the efficiency and accuracy of the model. The overfitted model has **low bias** and **high variance.**

The chances of occurrence of overfitting increase as much we provide training to our model. It means the more we train our model, the more chances of occurring the overfitted model.

Overfitting is the main problem that occurs in supervised learning.

Example: The concept of the overfitting can be understood by the below graph of the linear regression output:



Source: https://www.javatpoint.com/overfitting-and-underfitting-in-machine-learning

As we can see from the above graph, the model tries to cover all the data points present in the scatter plot. It may look efficient, but in reality, it is not so. Because the goal of the regression model to find the best fit line, but here we have not got any best fit, so, it will generate the prediction errors.

How to avoid the Overfitting in Model

Both overfitting and underfitting cause the degraded performance of the machine learning model. But the main cause is overfitting, so there are some ways by which we can reduce the occurrence of overfitting in our model.

- Cross-Validation
- Training with more data
- Removing features
- Early stopping the training
- Regularization
- Ensembling

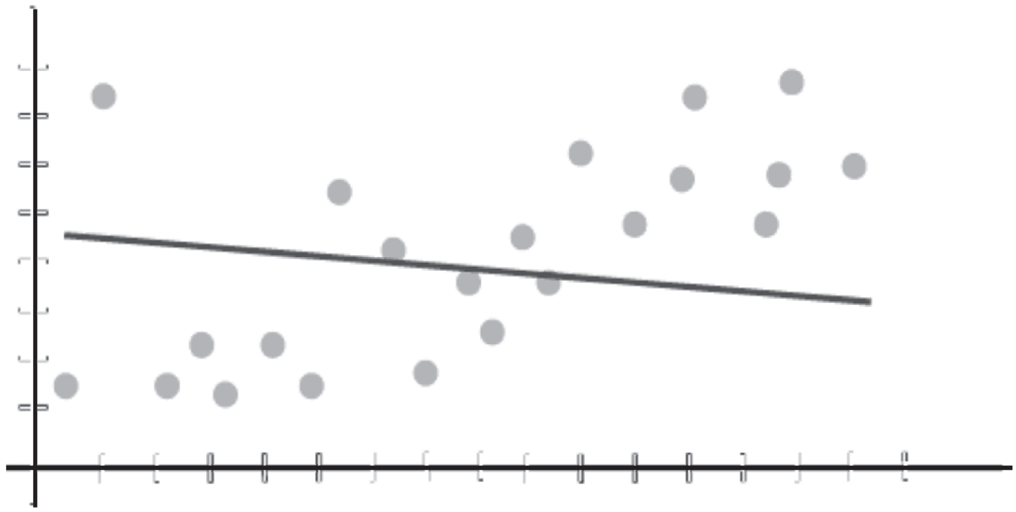
Underfitting

Underfitting occurs when our machine learning model is not able to capture the underlying trend of the data. To avoid the overfitting in the model, the feed of training data can be stopped at an early stage, due to which the model may not learn enough from the training data. As a result, it may fail to find the best fit of the dominant trend in the data.

In the case of underfitting, the model is not able to learn enough from the training data, and hence it reduces the accuracy and produces unreliable predictions.

An underfitted model has high bias and low variance.

Example: We can understand the underfitting using below output of the linear regression model:



As we can see from the above diagram, the model is unable to capture the data points present in the plot.

How to avoid underfitting:

- By increasing the training time of the model.
- By increasing the number of features.

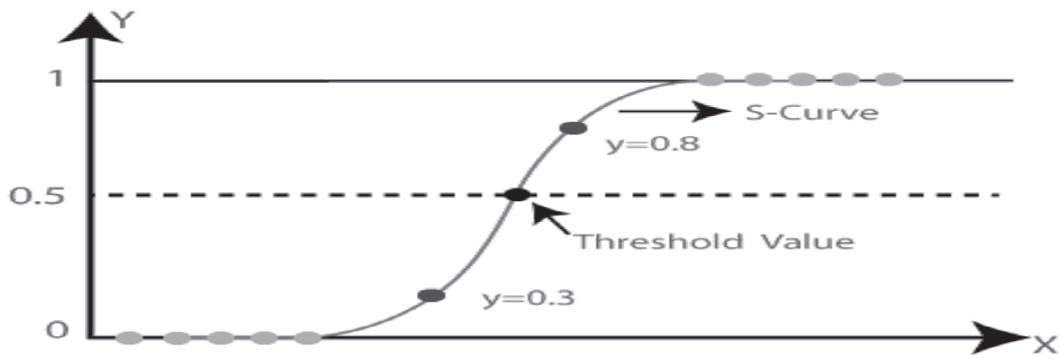
LOGISTIC REGRESSION AND MAXIMUM LIKELIHOOD

Logistic Regression in Machine Learning

- Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

> **Note:** Logistic Regression is a supervised machine learning algorithm which is mainly used for classification tasks where the goal is to predict a probability that an instance is belonging to a particular class or not.

- Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.
- Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.
- In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).
- The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.
- Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.
- Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The below image is showing the logistic function:



Maximum Likelihood Estimation (MLE)[38]

Maximum Likelihood Estimation (MLE) is a technique used for estimating the parameters of a given distribution, using some observed data. MLE is a probabilistic based approach to determine values for the parameters of the model. MLE is a widely used technique in machine learning, time series, panel data and discrete data. Maximum Likelihood Estimation (MLE) is a probabilistic based approach to determine values for the parameters of the model. Parameters could be defined as blueprints for the model because based on that, the algorithm works. The motive of MLE is to maximize the likelihood of values for the parameter to get the desired outcomes.

38. Mehta Sourabh (2022) How is Maximum Likelihood Estimation used in Machine Learning, The Analytics India MAG.

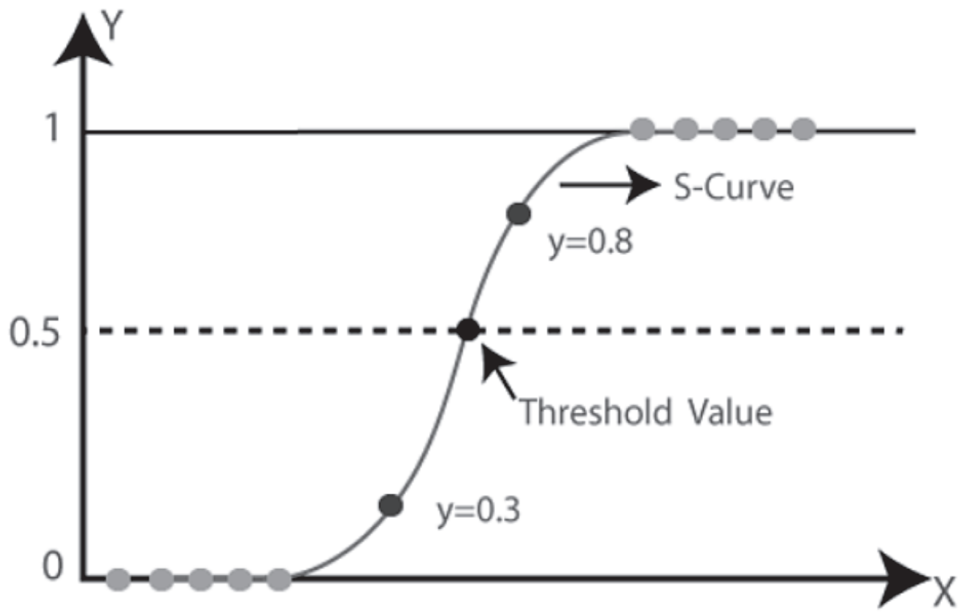
What is the likelihood?

The likelihood function measures the extent to which the data provides support for different values of the parameter. It indicates how likely it is that a particular population will produce a sample. For example, if we compare the likelihood function at two-parameter points and find that for the first parameter the likelihood is greater than the other it could be interpreted as the first parameter being a more plausible value for the learner than the second parameter. More likely it could be said that it uses a hypothesis for concluding the result. Both frequentist and Bayesian analyses consider the likelihood function. The likelihood function is different from the probability density function.

Working of Maximum Likelihood Estimation

The maximization of the likelihood estimation is the main objective of the MLE. Let's understand this with an example. Consider there is a binary classification problem in which we need to classify the data into two categories either 0 or 1 based on a feature called "salary".

So MLE will calculate the possibility for each data point in salary and then by using that possibility, it will calculate the likelihood of those data points to classify them as either 0 or 1. It will repeat this process of likelihood until the learner line is best fitted. This process is known as the maximization of likelihood.



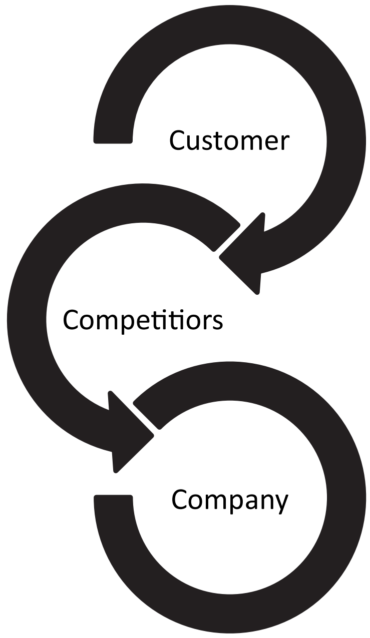
Source: https://analyticsindiamag.com/how-is-maximum-likelihood-estimation-used-in-machine-learning/

The above explains the scenario, as we can see there is a threshold of 0.5 so if the possibility comes out to be greater than that it is labelled as 1 otherwise 0.

TRIPLE C THEORY AND DATA ASSESSMENT

In a perfect world, data would always be complete, accurate, current, pertinent, and unambiguous. In the real world, data is generally flawed on some or all of these dimensions. Data assessment in practice has tended to focus on completeness and accuracy, and that is the focus of these notes. Currency, pertinence and clarity deserve more attention than they receive, perhaps, but their assessment requires very different methods.

Assessment is sometimes thought of as a preliminary to analysis proper. This is a useful distinction in some circumstances, but in general the assessment of error and the drawing of substantive conclusions are two sides of the same coin. This is suggested by the symbolic equation “Data = Reality + Error”, in which “Reality” represents conclusions drawn from the data that are valid despite the error and “Error” represents spurious conclusions suggested by the data as a result of error. Since all conclusions fall into one or the other of these two categories, conclusions about error are at the same time conclusions about reality, and conversely.



Direct Assessment

There are two general approaches to the assessment of data, direct and indirect. Direct assessment consists of evaluating the coverage and content of a data set. Coverage refers to the faithfulness of the correspondence between the records that constitute the data set and the statistical aggregate the data set represents. Data sets may omit records for some entities that should be represented and include records that should not be included. Improper inclusions occur when a data set includes more than one record for the same entity, includes records for entities not in the statistical aggregate, or includes fictitious records. Content refers to the completeness and accuracy of the information contained on the records in the data set.

Indirect Assessment

Direct assessment of data sets is expensive, both because a second data set is required for comparison and because matching is often a complex and difficult process. The results of direct assessment are, moreover, limited by response correlation bias and by the tendency of data sets collected at the same or nearly the same time to have similar content error. The indirect approach, by which data sets are assessed by analyzing the accuracy of statistics derived from them, is generally far less expensive and will often give results as good as or better than direct assessment.

Triple C Theory and Data Analysis

- **The First C – Customer Analysis**
  - Doing in-depth consumer research is the best way for you to figure out how to appeal to your target market. Being able to create catchy catchphrases and creative ads is going to be your bread and butter.

- Demographic data, that is, information on groups of people or population as per certain attributes like age, sex and place of residence, plays a huge part in this analysis. Figuring out your business's target market and their desires will drastically improve the success rate of your marketing strategies after they are put into circulation.

- **The Second C – Competitor Analysis**
  - Competitor analysis is mainly done by visiting their websites, subscribing to their newsletters, visiting their stores and/or receiving the service (heuristic analysis) they offer.

- **The Third C – Corporation Analysis**
  - The last step you'll want to take with this method requires you to analyze your own client's corporation. You'll want to know what marketing strategies have worked for them in the past and what ideas have failed. The best way for you to do this is, again, from the customer's viewpoint.

NETWORK ANALYSIS AND LAW[39]

While legal scholars have not adopted network analytic methods to the same extent as those in other fields of study, there have been a variety of legal network studies, and the trend appears to be increasing. This section will briefly review the legal scholarship applying network analyses before the following part explores more fully the future potential for legal network studies and the challenges faced by legal scholars working in the field.

Researchers have used network analytic techniques in a variety of contexts relevant to legal scholars. These include the analysis of legal social networks[40] examining statutes and regulatory codes as networks,[41] studying the networks of criminals and terrorists,[42] and studying the structure created by case law citations.[43] As case law citation network analyses are the most common and perhaps the most accessible of these, we will begin the review by looking at the history of this body of study.

> **Note:** Case law citation refers to a referencing system used by legal professional to identify judicial decisions, either in series of books called Reporters, or on Legal research websites, or a neutral style which identifies the case regardless of where it is reported.

Examples of Network Analysis and Law

a. Judicial Citation Networks

One very fruitful application for the tools of network science is to study the 'evolution' of the common law through the prism of judicial citations. Indeed, a distinguishing feature of a common law system is the precedential weight that judicial actors attach to prior decisions. Judges presented with questions in a given case consider how to apply doctrines from prior cases. Taken in the aggregate, common law systems produce vast amounts of citation data and although there is a rich literature studying these citations, relatively little scholarship has applied the tools of network science.

b. Visualisation of Relations

Visualising connected nodes leverages humans' perceptual abilities to discover patterns from data

39. Whalen Ryan (2016) Legal Networks: The Promises and Challenges of Legal Network Analysis, 2016 Mich. ST. L. Rev. 539

40. See, e.g., John P. Heinz & Edward O. Laumann, Chicago Lawyers: The Social Structure of the Bar (1982).

41. See, e.g., Romain Boulet, Pierre Mazzega & Danièle Bourcier, A Network Approach to the French System of Legal Codes—Part I: Analysis of a Dense Network, 19 Artificial Intelligence & L. 333 (2011).

42. See, e.g., Jialun Qin et al., Analyzing Terrorist Networks: A Case Study of the Global Salafi Jihad Network, in 3495 Lecture Notes in Computer Science: Intelligence & Security Informatics 287 (Paul Kantor et al. eds., 2005).

43. See, e.g., Fowler & Jeon.

associated with nodes and edges. The network data in this article were acquired from external data sources. The network is represented by nodes and edges: the nodes are the relational data such as sources (cases on the one hand, and Roman law, customary law and case law on the other hand), while the edges represent the citations between these nodes. On all occasions, the cases records are the sources because only these records contain references.

Research Case of Network Analysis in Law[44]

Court proceedings and its records are interesting for the legal historian, specifically from a perspective of law and its development over the centuries. Such cases are commonly analyzed by intensive reading and note-taking and by identifying commonalities, differences and relationships between documents or elements of documents.

We use these cases to explore whether human analysis can be improved, or at least complemented, by applying network analysis. The latter is a computer science method that allows for the mapping, measuring and visualising of relationships between individuals, groups and other types of information.[45] In this form of analysis, nodes are connected through edges, with the nodes being individuals, groups or information, and the edges being used to link the nodes. By treating court decisions as nodes, and linking the allegations in those cases to legislation and to other cases or scholarly work, network analysis can create a citation network that signals how information flows or has flown, and the extent to which certain nodes in the network are authoritative.

Network analysis has been used for several purposes in the legal field:

- for analyzing criminal behaviour and terrorist networks;
- for finding authoritative cases at courts; and
- for examining legal social networks, networks of statutes and regulatory codes and patent citations.

Volkaert[46] argues that network analysis can be both conceivable and useful in legal history. His recent article provides an overview of research in digital legal history using network analysis and focusing on case citation networks and what he refers to as 'digital-dogmatic legal history'. According to Volkaert, network analysis can complement dogmatic and contextual legal history, although qualitative juridical interpretations for understanding law remain necessary. By applying network analysis to a selected number of cases of the Court of Friesland, we further explore its potential for legal history research purposes.

Recovery from a third party

For the purposes of the examination, three civil court records with similar cases were selected dealing with the same legal question. Two of these cases could be joined by close reading the court records, while the third was found in a footnote in one of the other case studies. The facts in these three cases were broadly similar in that all three involved an object being sold at an auction. In the first two cases, the object was a black mare, whereas in the third case it was a red spotted cow. All three objects sold were encumbered with a mortgage and all three buyers failed to pay the purchase price, even after being reminded. The auctioneer who had sold the object subsequently had to recover it from a third party. In all three cases, the latter refused to surrender the object, claiming that it had been sold to him in a legally valid way and that he was therefore an (unassailable) possessor in good faith. Each time, the corresponding legal question was whether the security right (i.e. the

44. Reproduced from Hylkje De Jong and Gijs Van Dijck (2022) Network analysis in legal history: an example from the Court of Friesland, The Brill. Available at https://brill.com/view/journals/lega/90/1-2/article-p250_9.xml?language=en

45. J.H. Fowler et al., Network analysis and the law: measuring the legal importance of precedents at the U.S. Supreme Court, Political Analysis, 15 (2007), p. 324-325. See also M.G.H. Schaper, A computational legal analysis of acte clair rules of EU law in the field of direct taxes, World Tax Journal, 6 (2014), p. 77 and 80, which provides references.

46. F. Volkaert, OK computer? The digital turn in legal history: a methodological retrospective', Tijdschrift voor Rechtsgeschiedenis, 89 (2021), p. 1-46, especially p. 37 et seq mortgage) was valid only if the object was still with the debtor. Or could it also be valid and invoked if the object had been sold to a third party? In other words: did the security right include a right of pursuit, and could the auctioneer recover the object from the possessor in good faith?

Although the facts in the three cases were largely the same, the conditions of the sales were not. In the first case, the auctioneer had added the auction condition clausula constituti (a clause from respective constitution) and two sureties, which the plaintiff initially failed to call upon. As a result, and as far as we can reconstruct, the claim was rejected by the court. In the second case (, the auction condition clausula constituti was not added, but a specific mortgage was taken out, and there was no surety. Here, it would appear, the security right included the right of pursuit. In the final judgment, the defendant was ordered to hand over the black mare or pay the purchase price to the plaintiff (the auctioneer). The conditions in the third case were the same as in the second case, with the exception of the mortgage. In this case, the mortgage was not a specific mortgage, but a general one, which implied the privilege of the right of eviction. The plaintiff should, therefore, have first sued the non-paying buyer, which he had neglected to do. The Court of thus rejected the claim.

Simple interrelationships

Both the second and third case contains references to one of the preceding cases. In his reply, the plaintiff in the second case, copied 26 articles – the claim, the reply etc. were enumerated – verbatim from the first case. Indeed, he explicitly mentioned this first case at the end of his reply. In the third case, the plaintiff, copied 11 articles verbatim from the second case. These articles also happened to correspond with articles in the first case, although the judgement mentioned only the second case. The three lawyers would appear to have considered their cases to be similar and may have thought they could benefit from the earlier case(s). This interrelationship between the cases was found coincidentally, through close reading. New questions arise from this finding, including how did the lawyers acquire procedural documents from the other cases? Did they have access to them in some way? Was it common to refer verbatim to articles from other court cases? Interestingly, and even without going into the cases in depth, we can also identify another simple interrelationship: the auctioneers from the first two cases and the red spotted cow from the third case all came from a different locational background. Did the lawyers then work together? Did they exchange procedural documents? Linking more cases of a similar nature could provide new answers to questions about legal practice in the early modern period.

Problematic in the study of court records is that it takes a long time to identify allegations in the text and to find relationships between them. One has to be both conscientious and fortunate to recognize interrelationship between litigations. Moreover, based on the sample used for the present study it is hard to draw conclusions. And although the interrelationships mentioned in the three cases studies here were easy to recognize, recognition will be more difficult where the numbers of entities are higher.

Relations between references: Roman law, customary law and case law

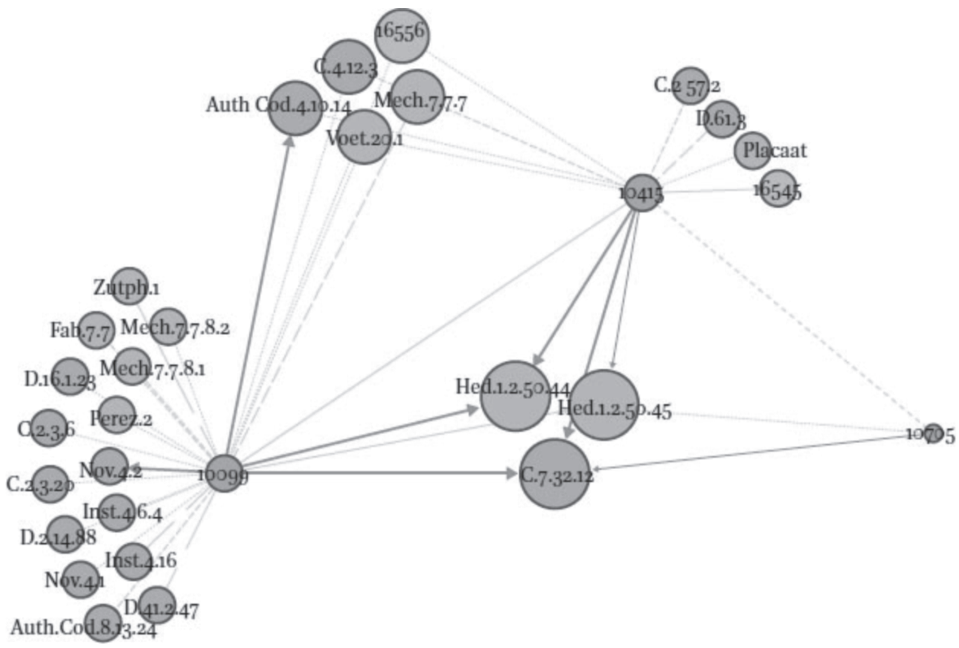
Since the facts and the legal questions in the three cases under consideration are similar, it would not be surprising if the references to the sources also prove to be similar. And, indeed, this turned out to be true. The three cases contain a total of 55 references; of these, 30 were unique references, not including the two references to the other two cases studied here. There is thus an overlap: more than once, the lawyers used the same reference. It is difficult to identify patterns of references through close reading. Instead, therefore, we labelled the nodes (i.e. information entities) and analysed the relationships between them by using network analysis. Before applying this analysis, however, we first had to categorise the various sources to which the allegations referred, given that conclusions can only be drawn on well-defined entities.

For this case study, we distinguished three categories of references, defined more or less by content, Roman law (15 times)25, customary law (10 times) and case law (5 times). The three cases are separated from these categories. We are aware that this categorisation is not exhaustive and can be more accurately expanded in future studies, focusing more on the defining of the information entities. Some shortcomings arising from the use of inadequate definitions are that the legal commentaries and works on Roman-Frisian law were classified – for the sake of convenience – under customary law. We also did not take the classification of the use of allegations into consideration, for example, are they copied or are they used by analogy? In addition, no distinction was made between data pertaining to the plaintiff and data pertaining to the defendant. The extent to which references were used at different stages in the legal proceedings was also neglected. However, the three categories chosen were considered sufficient for the purposes of this article, which sets out only to demonstrate the benefit of network analysis for this dataset.

Visualisation of relations

Visualising connected nodes leverages humans' perceptual abilities to discover patterns from data associated with nodes and edges. The network data in this article were acquired from external data sources. The network is represented by nodes and edges: the nodes are the relational data such as sources (cases on the one hand, and Roman law, customary law and case law on the other hand), while the edges represent the citations between these nodes. On all occasions, the cases records are the sources because only these records contain references.

The visualisation of the references in the three cases is presented in figure 1. The layout of the network is based on the Force Atlas 2 algorithm28, with the three cases illustrated in orange as set out below. As mentioned above, the first case has inventory number 10099 (1716), the second 10415 (1718) and the third 10705 (1720). The figure should be read chronologically, from left to right, with references to Roman law being shown in purple, references to customary law in green and those to case law in blue. The node size depends on the incoming references: the more incoming references, the larger the node size (or, in network analysis terms, the larger the in-degree value). The same applies in respect of the thickness of the arrows: the thicker the arrow, the more often the lawyer referred to the particular source.



Source: https://brill.com/view/journals/lega/90/1-2/article-p250_9.xml?language=en

What can we generally deduce – despite the limited categories – from the network? The three nodes in the middle are the largest and therefore have the highest number of incoming references. All three cases refer to these fragments. The high (in-degree) centrality of these nodes suggests that these fragments are authoritative in this kind of legal issue. The first case in time (10099) contains the most references, while the second case contains substantially fewer references. This second case refers to the first case, as the third case does to the second one. In its references, the third case is limited to the most important references, while the first two cases have some references in common, as shown at the top left. Both these cases refer to other case law in blue (16556). This case is from 20 December 1687. The second case also refers to other case law, specifically the judgment issued on 1 February 1676 (shown as 16545).

References to Roman law account for 50.0% of the references, those to customary law for 33.3% and those to case law for 6.7%, with the remainder comprising references to either or both of the other two cases. It is not surprising that most of the references are to Roman law, given that the Frisians chose Roman law to be applicable. If customary law could be further specified within the category, for example, of Roman-Frisian law, the reception of Roman law could be calculated more precisely. The more accurate the definitions of information entities are, the more specific the information generated will be. However, qualitative juridical research will continue to be necessary.

Future of network analysis in legal history

For what other legal questions may network analysis be useful? As well as identifying meaningful relations between entities within a dataset, it could be helpful for combining similar sets of entities. If other court archives in the other countries were to be unlocked, relations between these courts could also be revealed.

Today, entire archives of centuries-old material are available digitally and have been made searchable. Although such initiatives have commonly been undertaken in (digital) history departments, initiatives in the legal domain are under development. Fortunately, important initiatives are taken to change this situation. In, for instance, a pilot study, focusing on legal history, at the Department of Legal Theory and Legal History at VU Amsterdam, 48 metres of archives from the Court of Friesland (1499-1811) are being made accessible. The resultant data will be from 2273 civil cases from the period 1716-1730 (all documents in manuscript). This will be the first time that such a large amount of historical material from legal practice will be available.

A challenge resulting from the increased availability of data is that identifying relationships between documents or parts of documents become unmanageable if only human analysis is applied. Computational methods can assist in analysing historical data by, for instance, automatically recognising references in the text. This can then result in more data becoming available metadata; that is, data about the data (e.g. citations). The availability of more digitalised documents in the field of legal history will also allow leveraging of computational methods and techniques for analysing the documents, with network analysis then being used, for instance, to explore relationships between large numbers of cases and between references to and from those cases. However, this type of research requires new (digital) skills and implies a transformation, at least in part, of academic research in the sense that the very nature and scope of historical research will change.

Concluding Remark

This article discusses three similar cases, in which the legal question at stake was whether security rights (i.e. mortgages) included a right of pursuit. Could the auctioneer recover the object if the buyer failed to pay? The answer to this question depended on the contractual conditions. The lawyers in the three cases from the Court of Friesland appear to have used some of the same references. Network analysis of these references was used to visualize the relationships between these three cases, with the most important references also being made visible. A network can also be customized to suit specific purposes: the more entities (big data) that are defined, the more information that can be generated.

By comparing legal practice in civil court records in various provinces, or even internationally, network analysis will also make it possible to explore the mode of operation in these courts. This type of research in legal history will then generate new information that will add knowledge to historical legal practice, as well as uncovering information about daily life in early modern history and possibly also leading to a new understanding of the identities of the various provinces.

LESSON ROUND-UP

- In the modern times, industry whether legal or technology, academy, practitioners and scholars utilize artificial intelligence (AI) and machine learning to perform analyses which used to be labor-intensive endeavors a decade back.
- Data analytics is the science of analyzing raw data to make conclusions about the information.
- Many of the techniques and processes of data analytics have been automated into mechanical processes and algorithms that work over raw data for human consumption.
- Data analytics is the process that refers to deriving valuable insights and information from data using quantitative and qualitative methods.
- Legal analytics, with the help of data processing technologies, helps lawyers clean up, collate, and structurally analyze data.
- Data in the law industry can be broadly classified into individual data, law firm data, and industry data.
- Individual data is the data that one has in one's personal repository.
- Legal analytics is tool of facilitation and improving efficiency of legal results rather than a replacement or a solution for all legal case and client needs.
- Practice of law analytics gives in-house counsel insight that supports the legal practice, including data on legal contracts and cases.
- Artificial Intelligence (AI) is beginning to transform the legal profession in many ways and aims to take on higher-level tasks such as advising to clients, negotiating deals and drafting of standard contracts and agreements.
- Adoption of QLP algorithms to predict case outcome essentially provides individuals a risk management tool when deciding whether to adjudicate or not. This risk management tool can predict a case outcome, on average, with a better accuracy than a lawyer, promising a new era in how litigation decisions are made.
- Machine learning is a branch of Artificial Intelligence, which allows machines to perform data analysis and make predictions.
- Maximum Likelihood Estimation (MLE) is a probabilistic based approach to determine values for the parameters of the model.
- The likelihood function measures the extent to which the data provide support for different values of the parameter.
- While legal scholars have not adopted network analytic methods to the same extent as those in other fields of study, there have been a variety of legal network studies, and the trend appears to be increasing.

TEST YOURSELF

(These are meant for recapitulation only. Answer to these questions are not to be submitted for evaluation.)

1. Write a short note on Data Analytics.
2. What is Artificial Intelligence (AI) and Machine Learning (ML)? Discuss the role of AI and ML in Law and Legal Industry.

3. Write Short Note on any two of the following:
   a. Bias/Variance
   b. Overfitting and Underfitting
   c. Legal Analysis.
4. Discuss 3 Ways to use Legal Analytics. Substantiate the same with suitable examples.

LIST OF FURTHER READINGS

- Machine Learning for Lawyers, LWN. Available at https://lwn.net/Articles/721540/
- Novotna Tereza, Network Analysis in Law: A Literature Overview and Research Agenda
- James H. Flower et al (2007) Network Analysis and the Law: Measuring the Legal Importance of Precedents at the U.S. Supreme Court, Advance Access Publication, 2007. Available at https://home.gwu.edu/~wahlbeck/articles/Fowler-Johnson-Spriggs-Jeon-Wahlbeck%202007%20PA.pdf
- Park So-Hui et al (2021) A Survey of Research on Data Analytics-Based Legal Tech, MDPI, Volume 13, Issue 14
- Trasberg Henrik (2019) Quantitative Legal Prediction and the Rule of law, Master Thesis, Law and Technology LLM, TILBURG University Law School. Available at http://arno.uvt.nl/show.cgi?fid=149307

LIST OF OTHER REFERENCES

- An Introduction to Artificial Intelligence for Law Firms (2021) Lateral
- Artificial Intelligence for Lawyers Explained, Bloomberg Law
- Biswal Avijeet (2023) What is Data Analytics and its future scope in 2023, Simplilearn. Available at https://www.simplilearn.com/tutorials/data-analytics-tutorial/what-is-data-analytics#:~:text=Data%20analytics%20is%20the%20process,and%20efficiency%20of%20your%20business
- Columbia Law School on Data Analytics. Available on https://www.law.columbia.edu/areas-of-study/data-analytics
- Embroker (2022) What is the Legal Analytics and How you can use it to benefit your law firm. Available on https://www.embroker.com/blog/what-is-legal-analytics/
- Emily Stevens (2023) What is data Analytics? A Complete Guide for Beginners, CF Blog. Available at https://careerfoundry.com/en/blog/data-analytics/what-is-data-analytics/
- Jake Frankefield at el (2023) Data Analytics: What it is, How It's Used and 4 Basic Techniques, Investopedia. Available at https://www.investopedia.com/terms/d/data-analytics.asp
- James H. Fowler et al (2017) Network Analysis and the Law: Measuring the Legal Importance of Precedents at the U.S. Supreme Court, Cambridge University Press.
- Jong and Dijck (2022) Network Analysis in Legal History: An Example from the Court of Friesland, Brill. Available at https://brill.com/view/journals/lega/90/1-2/article-p250_9.xml

- Marios Koniaris et al (2018) Network analysis in the legal domain: a complex model for European Union legal sources. Get access Arrow, Journal of Complex Networks, Volume 6, Issue 2, April 2018, Pages 243–268.
- Network Analysis and Law: Introductory Tutorial, Computer Legal Studies, Jurix 2011
- Lauri Donahue (2018) A Primer on Using Artificial Intelligence in the Legal Profession, JOLT Digest, Harvard University. Available on https://jolt.law.harvard.edu/digest/a-primer-on-using-artificial-intelligence-in-the-legal-profession
- Legal Analytics: Definition and Advantage, Legal Practice, Lego Desk. Available at https://legodesk.com/legopedia/legal-analytics-definition-advantage/#:~:text=This%20data%20is%20collected%20from,similar%20scenarios%2C%20and%20so%20on.

  Lexis Insights (2023) What is Legal Analytics, Lexis Nexis. Available on https://www.lexisnexis.com/community/insights/legal/b/thought-leadership/posts/what-is-legal-analytics

PART II

TECHNOLOGICAL PERSPECTIVE