

AMERICAN UNIVERSITY OF BEIRUT

DATA-DRIVEN PRODUCT DEVELOPMENT: PATENT
DATA ANALYSIS USING NATURAL LANGUAGE
PROCESSING

by

RAGHED RABIH SAAB

A thesis
submitted in partial fulfillment of the requirements
for the degree of Master of Engineering Management
to the Department of Industrial Engineering and Management
of the Faculty of Engineering and Architecture
at the American University of Beirut

Beirut, Lebanon

July 2020

AMERICAN UNIVERSITY OF BEIRUT

DATA-DRIVEN PRODUCT DEVELOPMENT: PATENT DATA
ANALYSIS USING NATURAL LANGUAGE PROCESSING

by
RAGHED RABIH SAAB

Approved by:



29 August 2020

Dr. Ali Yassine, Professor

Advisor

Department of Industrial Engineering and Management



Dr. Hazem Hajj, Associate Professor

Member of Committee

Department of Electrical and Computer Engineering



29 August 2020

Dr. Jimmy Azar, Assistant Professor

Member of Committee

Department of Industrial Engineering and Management

Date of thesis defense: July 17,2020

AMERICAN UNIVERSITY OF BEIRUT

THESIS, DISSERTATION, PROJECT RELEASE FORM

Student Name:

_____ Saab _____ Raghed _____ Rabih
_____ Last _____ First _____ Middle

Master's Thesis
Dissertation

Master's Project

Doctoral

I authorize the American University of Beirut to: (a) reproduce hard or electronic copies of my thesis, dissertation, or project; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes.

I authorize the American University of Beirut, to: (a) reproduce hard or electronic copies of it; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes

after:

One ---- year from the date of submission of my thesis, dissertation, or project.

Two ---- years from the date of submission of my thesis, dissertation, or project.

Three ---- years from the date of submission of my thesis, dissertation, or project.

_____ Raghed Rabih Saab _____ 28/8/2020 _____

Signature

Date

ACKNOWLEDGMENTS

I would like to thank everyone who helped me along the way – Mom for her unconditional support, Dr. Yassine for keeping the work on the right track, and Dr. Azar and Dr. Hajj for dedicating their time to my thesis.

AN ABSTRACT OF THE THESIS OF

Raghd Rabih Saab for Master of Engineering
Major: Engineering Management

Title: Data-driven product development: patent data analysis using natural language processing

The product development lifecycle exhibits many big data flows of internal or external sources and destinations. Until recently, means of analyzing these data flows were severely limited due to performance and storage limits. With the advancement in technology, one can utilize these data flows to improve the product development process thus yielding better results. This thesis finds literature related to big data flows in the product development process and then classifies these flows and their position in the process. It also discusses the challenges and opportunities of utilizing big data analytics in the product development process.

This thesis also aims at developing a novelty measure for patents, a specific data flow inside the lifecycle, which is a basis to measure the level of patent innovation. Patents are a main proxy of invention, and each patent exhibits varying inventive value and novelty compared to the corpora. Prior studies of patent novelty have suggested a citation-based approach to measure the novelty across patents. Despite their progress in measuring patent novelty, several challenges remain: The inability to consider single class inventions, and the inclusion of patent-only citations. To address these challenges, we devise a novel approach using NLP techniques to find a text-based novelty measure. The proposed method is applied on patents that belong to a common category, which represents a subset of patents under a specific patent class. We then extract the novelty-value profile of those patents and discuss a use case for product development – extracting patent novelty and predicting inventive value. Product developers would benefit from our proposed approach by allowing them to predict value of patents being developed. These findings would contribute to having an alternative way to measure novelty, which complements previous citation-based methods. Future research would build upon text-based measures which will further improve data-driven approaches tackling novelty assessment.

CONTENTS

ACKNOWLEDGMENTS.....	V
ABSTRACT.....	VI
ILLUSTRATIONS.....	IX
TABLES.....	X
Chapter	
I. INTRODUCTION.....	1
II. BIG DATA OVERVIEW.....	6
A. Brief Summary on the Origin of Big Data.....	7
B. Properties of Big Data.....	8
C. Data Analytics Architecture.....	10
1. Data generation.....	10
2. Data aggregation.....	11
3. Analysis.....	11
4. Visualization.....	12
III. PRODUCT LIFECYCLE BIG DATA FLOWS.....	13
A. Paper Collection Methodology.....	14
B. Data Flow Framework.....	16
1. Data Flow into conception.....	18
2. Data Flow into design.....	19
3. Data Flow into manufacturing.....	24
C. Opportunities of having big data in product development.....	25

1.	Product quality improvement.....	26
2.	Better market needs prediction	26
3.	Better innovative value prediction.....	26
4.	Faster, more efficient development iterations.....	27
5.	More optimization to manufacturing, distribution processes	27
D.	Challenges of having big data in product development.....	28
1.	Challenges in the product lifecycle.....	28
2.	Challenges of technological factors.....	28
3.	Challenges based on regulations and legal issues.....	29
IV. FINDING PATENT INNOVATIVE VALUE COMPARED TO PRIOR ART		30
A.	Overview on Patents	30
B.	Literature Review	34
1.	Data-driven analytics on patents.....	34
2.	Data-driven measures of innovation value and novelty of patents.....	37
C.	Methodology	40
1.	Obtaining patents	41
2.	Measure of patent novelty.....	42
3.	Measure of patent value.....	50
4.	Validation of novelty approach.....	51
D.	Results of a case on a specific patent group according to the CPC	53
V. CONCLUSION		57
REFERENCES		60
APPENDIX		74

ILLUSTRATIONS

Figure		Page
1.	Data Information Knowledge flows (Aamodt and Nygard, 1995).....	3
2.	'Big data' term popularity according to Google queries, 2004 through 2019 (Google Trends, 2019).....	8
3.	Phases of data analytics.....	11
4.	The top 100 patents by citations (PatentsView, 2020)	13
5.	The data flow between various stages in the product life into the design phase, based on the ISO/IEC 15288 life cycle.....	15
6.	Papers used grouped by area of contribution across years.....	16
7.	Change in yearly patents in different offices in 2018, according to WIPO (WIPO IP Facts and Figures, 2019).....	34
8.	A sample patent document published in 2017.....	35
9.	The methodology flowchart.....	41
10.	From left to right, the general CBOW and skip-gram representations ..	48
11.	A general autoencoder (Mitchell, 1997).....	50
12.	The ROC curve.....	53
13.	From left to right, the scatter plot of the patent corpus value-novelty profile and the fitted curve.....	55

TABLES

Table		Page
1.	Properties of big data.....	10
2.	Internal and external data flow origins.....	19
3.	The selected fields to be saved for further analysis.....	43
4.	Some regex operations done non-destructively on the corpus	46
5.	An example string with stemming and lemmatization applied to	47
6.	The chosen CPC group code	54
7.	The 2 patents and their texts	57

CHAPTER I

INTRODUCTION

Since the early days of product design, data played an important role in the design process. With the ever-growing means of collecting data across the product lifecycle and the parallel scaling of storage and processing requirements, leveraging this data is becoming more and more relevant – if not critical – to the product design process and its continuous innovation and refinement requirements.

Until recently, data sets used in research and practice were small to medium in size due to collection, storage, and processing bottlenecks. Due to the advancements in information technologies and these being deployed across different industries and the ensuing sudden influx of data streams, data abundancy was suddenly increasing at an ever-increasing rate (The Economist, 2010). Having the means of capturing these data points as well as the processing power to analyze them shifted the way this data is utilized towards entirely new outcomes.

An ancient parable talks about six blind men who had never heard of an elephant before and were asked to identify its shape, with each coming up with his own description depending on the part he touched. An analogy can be drawn out here, where the elephant stands as the source of data; whatever it might be, the blind men as the people studying this data, and the sense of touch – albeit inferior here is the best they can use – is the tools used to capture, store, and analyze this data. It is clear that, depending on the frame of the observer, an image can be drawn out, and that unifying bigger numbers of different inspection points and extracting meaning increases the odds

of understanding the real subject and thus accurately decoding it to generate value and benefit.

The prevailing industrial revolution, Industry 4.0, holds the promise of intricate customizability, better productivity, and faster workflow from conceptualization to delivery (Zhong et al., 2017). The need for a shorter time to market, coupled with improved quality, led to the need for a revamped product design process. The constant push for this has led product designers to adopt various data sources in order to further optimize the design process. The volume of dynamically changing data generated throughout the lifecycle of products is growing (Kuo & Kusiak, 2018). Embedded within this data is vital information that underly the working of the product in case of data collected during the usage of the product, or spatial and temporal data generated by systems in the manufacturing phase. This data can then be analyzed, and meaningful results can be taken out, which feeds back into the design stage of the data-generating product itself so it can be further enhanced (Bertoni et al., 2017).

The big data analytics era is becoming prominent and more mainstream interest is being noticed, but the practice itself has been around for much more (e.g., mathematical and statistical research). The meshing between the industrial processes and technology, which is being implemented recently not only on the outside of the industrial process, but in the intricate details of the product such as sensors, process information, and other types of data being digitalized opens up to many possibilities. The wide-scale adoption of these technologies will lead to an industrial revolution, after the preceding internet revolution, which continues to transform our lives down to our daily activities. The change spans diverse industries across the globe, and as early adopters are continuously leveraging big data into their favor, other players will soon

follow suit. The resulting improvements in efficiency and optimization will affect economies positively as well as spark the next wave of product development and sustained refinement of these products.

The connection between data, information, and knowledge is often referred to by the Data-Information-Knowledge pyramid (Aamodt & Nygård, 1995) (Liew, 2007), as shown in figure 1. In this framework, data referred to as uninterpreted characters, signals, patterns, etc. is at the base of the pyramid. Once this data is interpreted and put into context, it is referred to as information. In turn, when this information is added to the reasoning structure and used in the decision-making process, it becomes known as knowledge, which ultimately is used back within every decision. This model could be further expanded into the Data-Information-Knowledge-Wisdom model (Ackoff, 1989), which encompasses a further stage, wisdom, that can be referred to when knowledge is tested and validated. This concept, when applied from the big data viewpoint, is of significant interest since big data, meaning larger amounts of data, would theoretically mean that more information is captured and thus more knowledge and wisdom in a shorter time frame. The data captured can create the product development knowledge necessary for more market-resonant products.

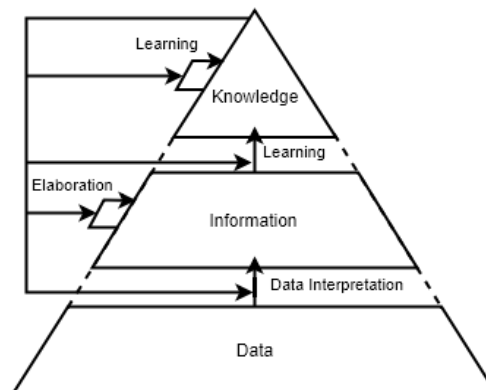


Figure 1 Data Information Knowledge flows (Aamodt and Nygard, 1995)

But is this really the case? How can big data in the product development cycle ensure bigger benefits that can move up the pyramid? As data flows get more complex over time, more data flows acquire qualities that set them under the ‘big data’ term (Sagiroglu & Sinanc, 2013) , and thus require for more advanced analysis in order to extract information (Gandomi & Haider, 2015) in order to leverage the big amounts of mostly heterogenous data. Many of these data flows exist within the product development cycle or interact with it. The information and knowledge extracted from these data sources is imperative to the success of the product development project. However, and while other fields already benefited from the surge of big data, product development particularly still falls behind. Most literature tackles a specific flow disregarding its relation to the overall product development process. While some previous research attempted to classify big data flows into the product lifecycle (J. Li et al., 2015), the framework used attempts to group data flows into three major categories, the Beginning-of-life (BOL), Middle-of-life (MOL), and End-of-life (EOL). In our research, we attempt to systematically classify big data flows in a more detailed framework in an attempt to build upon previous literature and highlight the interconnection of the sources and what step of the product development process they feed into.

We then select a relevant data flow, patent data, which plays a significant role in the conceptual and design phases of product development and apply a novel approach to extract information from it. Previous literature (He & Luo, 2017) links patents and innovation together, and attempts to study the variation of patent innovative value with novelty, and to measure patent novelty, it uses citations as a proxy. This method however has drawbacks, particularly the failure to account for non-patent citations

which a patent may cite. A sample of a non-patent citation is research published in a journal, which may have different qualities that are discarded when using this method. We develop a novel approach to measure patent novelty and its effect on value by proposing a text-based method. We rely on patent text to determine patent novelty with respect to the corpus. This allows us to have an alternative way to evaluate patent novelty while not relying on citations as a measure which mitigates its weaknesses. Then, we can leverage the text data to extract information that can be used by the product developer.

The motivation of this thesis is to take advantage of machine learning methods and apply them to product development where big data analysis was not possible due to collection, storage, and computing limitations and has much potential that is unutilized. The objective of this thesis is to identify and categorize big data flows in the product development life cycle, and to focus on one major data flow, patent data, and attempt to apply machine learning approaches in order to improve the knowledge gained from this data flow hence enhancing the product development cycle as a whole.

The remainder of this thesis is organized as follows. In section II, a literature review of big data in product development is made. In section III, we provide a methodology that was followed in order to conduct a search on literature of big data flows in the product development cycle, and we propose a product development lifecycle framework in which we classify incoming and outgoing big data flows according. In section IV, we focus on one big data flow, patent data, and use NLP techniques and other ML methods in order to apply a text-based novelty extraction method, then apply it to find the novelty-value characteristics of patents under a homogenous corpus. In section V, we summarize our findings and list future work.

CHAPTER II

BIG DATA OVERVIEW

‘Big Data’, meaning exponentially big amounts of data that are collected, aggregated, and analyzed, is seeping through into a multitude of industries and taking its foothold on international sectors ranging from healthcare and public sector to design and manufacturing (Manyika et al., 2011). A better interpretation of this not-so-well-defined term would not only include the size of the data being issued, but also other attributes equally contributing to making this data type a category of its own, such as the velocity and veracity of big data, among others. A quick search on this term on January 20, 2019 returns 7 billion results approximately, and a closely related term according to Google is ‘analytics’. The popularity of the term is on an increase as illustrated in Figure 2 (Google Trends, 2019). The figure shows that interest began to increase in 2012, followed a fairly sharp increase until 2015, and kept its position until 2019, which highlights the persistence of interest and signifies that this may not be a phase but rather a longer-term trend. This also accentuates the increasing interest globally in this concept, and the further implications it could have on a vast variety of fields, including product development, where the introduction of IoT and the gradual digitization of the whole product design and manufacturing chain is generating more and more big data points to utilize.

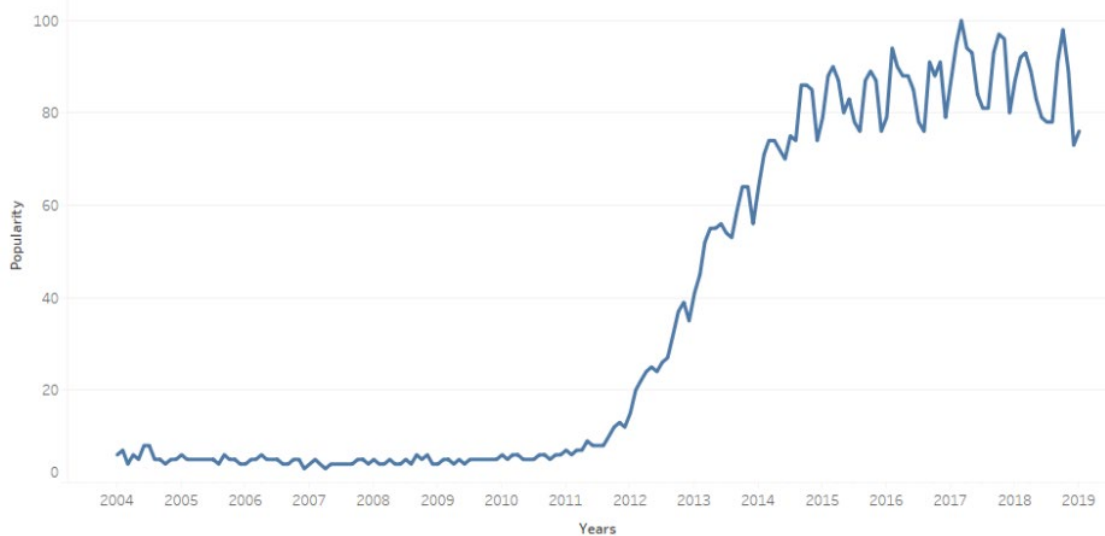


Figure 2 'Big data' term popularity according to Google queries, 2004 through 2019 (Google Trends, 2019)

This data, if understood and analyzed correctly, has the potential of making local exploitations all the way to redefining whole business scopes through business transformation as well as increase captured benefit in a parallel manner (Venkatraman, 1994). Clearly, the use and integration of this data are of surmountable importance for industries seeking a competitive edge in a vast pool of domains, namely manufacturing and product design.

A. Brief Summary on the Origin of Big Data

Before the current attention the term 'big data' is enjoying, many researchers attempted to address this concept long before it became known on a large scale. Though attempts to address similar concepts can be traced way back in history such as Fremont Ryder's speculation on the volumes stored in the Yale Library due to the fast growth of information; in 1997 Michael Cox and David Ellsworth were the first to use the term

across the ACM Digital Library, where they described the large amount of data and the inability to store it across media, calling it the problem of ‘big data’ (Cox & Ellsworth, 1997). In 1999, the article “Visually exploring gigabyte data sets in real-time” (Bryson et al., 1999), the opening paragraph states that “understanding the data resulting from high-end computations is a significant endeavor. As more than one scientist has put it, it is just plain difficult to look at all the numbers. And as Richard W. Hamming, mathematician and pioneer computer scientist, pointed out, the purpose of computing is insight, not numbers.” In 2001, an article published by Doug Laney, an analyst with Meta Group became the first to mention the volume, velocity, and variety of big data (Laney, 2001) although the term itself was not in the paper. The term ‘big data’ was then coined by Roger Mougallas in 2005 and used in its modern context, which is a large amount of data that is almost impossible to process using traditional tools. In the same year, Hadoop, a big data processing system, was built by Yahoo! to help make use of the big data sources and gain value out of them. Nowadays, businesses and researchers alike are using the capabilities of big data to drive progress and analyze and improve aspects that were not possible before.

B. Properties of Big Data

The first property that describes Big Data is the volume of the data; that is, it has a much larger size than conventional analytical tools can handle. But the volume of this data does not convey the whole story. Big data is mainly described by key attributes, known as the Vs of big data, which has various interpretations and are shown in table 1. The most basic interpretation gives big data its volume, velocity, and variety (Laney, 2001). Further models add new aspects such as veracity – coined by IBM - and

value. The 5Vs model classifies big data as data being too big, too fast, of many different types and origins, having a degree of unreliability, and of an inherent value that can be interpreted by analyzing this data.

<i>Aspect</i>	<i>Definition</i>
Volume	The large volume nature and quantity of data (Laney, 2001)
Velocity	The frequency of data generation and analysis (e.g. real-time) (Laney, 2001)
Variety	The structural heterogeneity of data across diverse sources and types of data (Laney, 2001)
Veracity	The unreliability of some data sources
Value*	The extracted benefit

**Describes the outcome of big data rather than qualities*

Table 1 Properties of big data

While these aspects attempt to provide a classification of the traits of Big Data, they also are inherently describing the problems faced when dealing with Big Data. It should be noted that the term Big Data serves as a catch-all phrase and holds no intrinsic accurate description but rather a fuzzy one. Depending on the environment the data analytics is done in, some of these aspects will differ and may not stay true. For example, sensors implemented across an area are considered big data, but the “big data” feed is relatively small in volume and lacks variety in this case. As such, this classification is misleading and may not apply in every case, and it is better to refer to qualities such as the exhaustivity (population-wide capture instead of samples), fine-grained resolution (the fineness of the data and it being sufficiently detailed) and

indexicality, as they are key descriptors of big data (Kitchin & McArdle, 2016) in addition to the attributes of velocity, veracity, and value. In other words, these properties should not be used as binary deciding criteria but rather as qualities that big data sources tend to have.

C. Data Analytics Architecture

It is substantial to know how the process of big data analytics works to further understand its impact on the design phase of the product development process. The big data collection, then analytics is a process consisting of different steps (Krishnan, 2013) that can be classified into data generation, data aggregation, analytics, and visualization, in this specific order, as shown in figure 3.

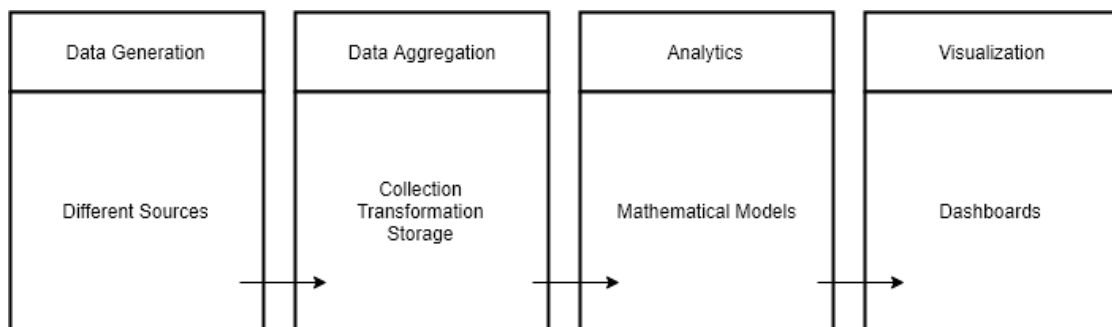


Figure 3 Phases of data analytics

1. Data generation

The data generation phase consists of the basic first step of realizing the data points. Big data can be of many different sources (Y. Zhang et al., 2017) e.g. sensor data, the internet, factory logs. It also comes as structured, semi-structured, and unstructured (Wu et al., 2014). These data points of different types (Wu et al., 2014)

(data feeds, data logs, manufacturing rig history data, sensor data, CAD/CAM versioning etc. have different source formats) are being generated throughout the product development process.

2. *Data aggregation*

This step deals with the acquisition and collection of the data, then correctly storing it in a suitable container for further use in later steps. It can be divided further into the collection step (Labrinidis & Jagadish, 2012) where the data is collected from the different sources through their respective method, the transformation step (Labrinidis & Jagadish, 2012) where the data is then cleaned, sorted, and made compatible with storage, then the data storage step (Labrinidis & Jagadish, 2012) where the treated data is then stored; the storage technology being NoSQL, MPP etc.

3. *Analysis*

Big data analysis is the means of transforming the collected and stored data into meaningful information. The analytics applied to the design phase of the product can be of aspirational, experienced, or transformed capabilities (LaValle et al., 2011), which affects and contribute to the design process in different manners, namely cutting costs and increasing efficiency in the case of the aspirational players, developing operational efficiency in case of the experienced subgroup, or looking into niche processes and structures beyond basic cost control in case of the transformed subgroup. Big data analytics in the product development realm means the analysis of data from different parts of the product lifecycle in order to discover insights previously

unattainable, in a prompt, convenient manner, decreasing possible errors and giving better value for feedback iterations inside the product development cycle. For example, it can be used to find market opinion that would affect the design of new products (Jin et al., 2016), and it can also be used to drive better manufacturing processes (Y. Zhang et al., 2017).

4. *Visualization*

The final stage in the process is the visualization stage, where the analyzed data stack is organized in a perceptible way so that insights can be grasped on the fly, depending on the targeted audience, in this case the scientists and engineers. As an example, big data from patents is analyzed and visualized (*PatentsView*, 2020) showing the top 100 patents in terms of citations, and how that relates to assignees and inventors as in figure 4.

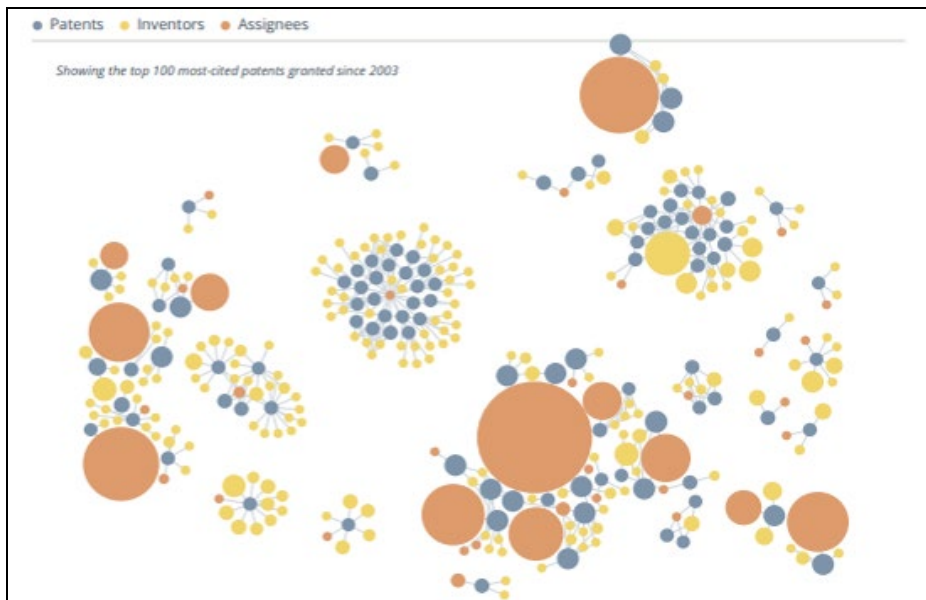


Figure 4 The top 100 patents by citations (*PatentsView*, 2020)

CHAPTER III

PRODUCT LIFECYCLE BIG DATA FLOWS

A product's lifecycle is made up of different steps that essentially explain the different states the product is going through, shown in figure 5. In the conception phase, an extensive market research is done to gauge the potential and applicability of the product ideas, followed by a design stage where the ideas are realized and developed into a fully detailed description for a production-ready product. The product is then produced in the production phase and later distributed and used by the customer, supported by the manufacturer, until its disposal or recycling at last.

During the lifecycle of a product, it generates huge amounts and diverse types of data along its way. The path from initial conception to the design, production, usage, and then lastly the decommissioning of the product has numerous data points that can be classified and investigated. Interestingly, every decision throughout these stages also depends on insights derived from previously collected data. Traditionally, this has been non-exhaustive and a fuzzy or a limited feedback within the process. By having the ability to potentially tap into data points previously inaccessible due to collection, storage, and analysis bottlenecks, this data can then be used to create knowledge to drive future decisions. The field of data-driven design is in its early stages of growth (Bertoni, 2018). The product life cycle and various data flow streams between internal stages and external sources where external sources mean from outside the PD stages are shown in figure 5 (Thimm et al., 2006).

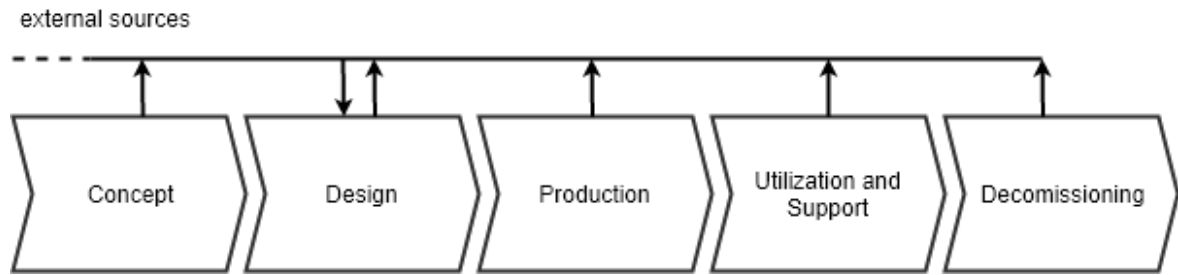


Figure 5 The data flow between various stages in the product life into the design phase, based on the ISO/IEC 15288 life cycle (ISO/IEC/IEEE 15288:2015, 2015)

A. Paper Collection Methodology

To search the literature for the usage and integration of big data in product design and development and its evolution in time, our approach was to systematically search for specific terms related to big data in product development. To do this, we followed the methodology of Biolchini et al. (2005). This consists of the three different phases of planning, executing, and analysis.

For the planning phase, we tried to pertain to a main question to guide our search for suitable terms. The main question was:

What are the different big data flows within the product development cycle?

To ensure unbiased, representative search results, the terms used were selected carefully to nullify any potential factor that can inhibit the appearance of some results. After these considerations, we settled upon the following terms: (“product development” OR “product design”) AND “product lifecycle” AND “big data”, to find results that relate big data to product lifecycle. These search strings are representative of our aim, and to further have papers that may not match the keyword search, we decided to look into each paper citations to add to our results. The source language was defined beforehand to only include papers published in English.

In the execution phase, we selected Google Scholar as the query engine, and no further conditions were introduced. The selected terms were inputted into Google Scholar and queries were recorded. The search query conducted in May 2019 returned nearly 1700 results. The paper titles were then scanned through and only relevant entries were kept, the papers that discuss an actual big data flow in PD. The distilled number of papers was then reviewed based on the abstract and relevancy was determined by filtering out entries that further do not mention big data in PD. The non-relevant papers were discarded. References in some review papers were also studied carefully to select papers that may explore data flows and big data.

In the analysis phase, the publications selected were then scanned carefully. Documents were tagged with areas in the framework they mention. The process was then to take similar documents and consolidate into groups of tags according to the framework. The remaining 31 papers that we chose to keep are classified into their respective publishing year in figure 6.

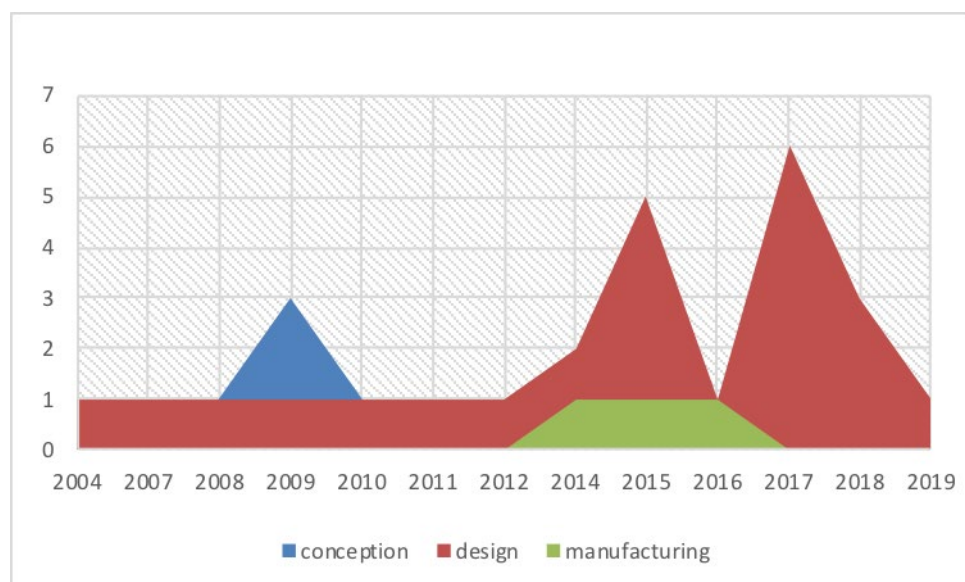


Figure 6 Papers used grouped by area of contribution across years

B. Data Flow Framework

The big data flows from different stages inside the product cycle that enter a specific stage are named as data flows with internal origins. Data flows from outside the product development lifecycle that happen to go into a specific stage inside it are named data flows with external origins. The different flows are collected and displayed in the table 2. The data flows, as they are in the table, can be from different internal stages of the product development lifecycle or from external origins, directly impacting one or more stages inside the product development cycle. By isolating each destination, the inflowing data streams are segmented as follows.

Data flow origin	Data flow end	Source
Conception	Conception and Design	<i>Concept generation tools</i> (English et al., 2010) (Wodehouse et al., 2004) (Arnold et al., 2008) (Kurtoglu et al., 2009)
Design	Design	<i>Digital-twin</i> (Tao et al., 2018) (Brossard et al., 2018) <i>Collaborative tools</i> (Son et al., 2014)
	Manufacturing	<i>CAD/CAM data</i> (Yang et al., 2014) (B. Li & Xie, 2015)
Manufacturing	Design and Manufacturing	<i>Product and manufacturing rig data</i> (Lei et al., 2016)
Utilization	Design	<i>Product usage data</i>

		(Bertoni et al., 2017) (Bae et al., 2015)
		<i>Environmental data</i> (Bertoni et al., 2017)
		<i>Maintenance and failure data</i> (Y. Zhang et al., 2017) (Baglee & Marttonen, 2015) (Yan et al., 2017)
Decommissioning	Design	<i>Product end-of-life data</i> (Parlikad & McFarlane, 2007) (Holler et al., 2017)
		<i>Recycling and disposal data</i> (J. Li et al., 2015)
Open data (Parraguez & Maier, 2017) and market (external)	Conception	<i>Company websites</i> (Unger & Eppinger, 2009)
		<i>Historical sales</i> (Song & Kusiak, 2009)
		<i>Social networks</i> (Tuarob & Tucker, 2015) (Jin et al., 2016)
		<i>Discussion boards and news aggregators</i> (Jin et al., 2016)
		<i>Competitor products</i> (Opresnik & Taisch, 2015)
		<i>IP Regulations and Patents</i> (Luo et al., 2014) (Xia et al., 2017)
		<i>Search engines</i> (Martí Bigorra & Isaksson, 2017)
		<i>Social and economic data</i> (Luo et al., 2014)

	(Jin et al., 2016)
Design	<i>Customer usage patterns</i> (Martí Bigorra & Isaksson, 2017) <i>Product reviews</i> (Ireland & Liu, 2018) <i>Open design repositories</i> (Qize Le & Panchal, 2012) <i>Public business registries and directories</i> (Michelino et al., 2015) <i>Research and publications</i> (Trappey et al., 2019) <i>Competitor products</i> (Opresnik & Taisch, 2015) <i>IP Regulations and Patents</i> (Luo et al., 2014) (Xia et al., 2017) <i>Knowledge databases</i> (Walthall et al., 2011)

Table 2 Internal and external data flow origins

1. Data Flow into conception

The company starts the ideation process by studying the market, looking into current news for business indicators, reaching out for insight from social media aggregated data (Jin et al., 2016; Tuarob & Tucker, 2015) and online discussion boards (Jin et al., 2016) as well as publicly available social and economic macro-data (Luo et al., 2014) to decide upon the feasibility and direction of the new product. Search engine

meta data is also a main source of business insight that can influence the product design (Martí Bigorra & Isaksson, 2017) in terms of feasibility, market traction, and expected product features. The firm checks relevant companies' websites for information, technologies, and a general sense of the company's business (Unger & Eppinger, 2009). It also looks into the competition public product lineup to further complete its profiling (Opresnik & Taisch, 2015). All these steps are done to estimate customer segmentation, demand, and preferences, so that a better estimate of what the customer needs can be reached and thus a better fitting product.

With the increasing digital shift of the former data sources, and the possibly recent existence of some, they can be captured in bulk and analyzed with big data frameworks to drive a viable conceptualization process as 96% of all innovations fail to return their cost of capital (Doblin Group, 2012). Delivering timely products that have a higher success rate should be an integral part of any product developer's plan, and this should be tackled early on in the product development cycle. Having these data sources to use can guide the product developer into better chances of success.

2. Data Flow into design

After selecting ideas that are relevant and are sufficiently feasible, the company has then to transition into the design phase to dive into the concept and expand it to a viable product design. Open data sources (Parraguez & Maier, 2017) are essential to utilize at this stage, and market data also provides a solid data source to tap into.

The product developer should be aware of patents and IP data (Luo et al., 2014) to gain a better understanding of the underlying technology and design data, have

a better knowledge of the intended area and the design limitations imposed, and to prevent any potential intellectual rights clash. Patents big data analysis can drive research and development efforts and point them in a specific way, for example helping in classifying biotech patents from others (Black & Ciccolo, 2004) which reduced the effort taken to find this specific type of patents. Another use case is using patent big data to automatically landscape the patent corpus (Abood & Feltenberger, 2018), thus allowing for better understanding of how different topics connect across different dimensions.

Open design repositories (Qize Le & Panchal, 2012) such as Github or GrabCAD can kickstart the design process by using mass-contribution, crowdsourced driven designs and community curated engineering design entities of various types, which can prevent reinventing the wheel especially for smaller companies. Using crowd shared hardware templates can have the same effect on physical products, and help cut down significantly on costs and time (Xu et al., 2016). The crowdsourced nature of this means that the crowd thinks as a whole resulting in a peer-rated ecosystem where the results are thoroughly tested (Qin et al., 2016). Using big data analysis on these, viable designs can be extracted from the data that otherwise cannot be manually scanned easily therefore contributing to the new product improvement (Xu et al., 2016).

Public registries and directories (Michelino et al., 2015) list diverse and rich metadata on aspects of the industry ranging from company and industrial data to product-related statistics. These registries can utilize big data tools to reduce the amount of information into guiding points for product development.

Publications and research are a main benchmark for engineering design, and they provide data on the current level of scientific exploration in the intended topic as

well as meta information on what surrounds it. Using big data analysis in terms of publications data allows for scientific impact evaluation which allows to evaluate authors, articles, and journals; academic recommendation to reduce the big amounts of data to actionable recommendations in terms of literature, collaboration, or venue; and expert finding (Xia et al., 2017).

Knowledge databases (Walthall et al., 2011) are a common ground for referral as they provide mostly factual and unbiased data intended to describe the current state, features, and timeline of any specific entry, and is sufficiently well-structured and cited, such as popular wiki sites or non-profit data collection organizations (archive.org). In order to store this data, a ratified version of information, fuzzy ontologies are proposed to attempt to find a specific order between the different pieces of data (Morente-Molinera et al., 2016).

Customer usage patterns (Martí Bigorra & Isaksson, 2017) are becoming more accessible especially in technologically advanced products, and usage patterns can be accessed in market studies. These can be used to study actual customer response to features and verify intended feature placement and usage with actual customer perception (Arora & Malik, 2015; Van Horn & Lewis, 2015). This will help relocate, remove, and redesign or segment complex features to maximize effectiveness and drive the customer-perceived value of the product. Customer preferences are as well collected and structured due to the increasing connectedness and subgroup preferences can be found from various sources and can be viewed changing with time as well as by geographical, gender-based, age-based, etc. categorization. Product reviews can also be a valuable source of data especially on platforms that can provide an aggregated view of all customer responses. These crowd generated reviews tend to accurately describe what

the masses perceive the product to be and have a hefty amount of details regarding pricing, features, quality, and other aspects of the product.

As for internal data flows, the product being developed can supply itself with many data streams from various parts of the process that can potentially be a game-changer if utilized correctly, especially knowing that this area is not tapped into until recently where technical means are - for the first time – allowing for big data collection, storage, and value capture.

Concept generation tools (Arnold et al., 2008; English et al., 2010) play an important role in making the conceptual research more effective as well as developing the first product concepts. By having an infinite number of possible outcomes, the product designer can tune the results to match his preferences while being exposed to much more simulated alternatives before choosing one.

The digital-twin (Brossard et al., 2018) is a data flow of design data into the design process itself. Constructing a virtual digital twin of the product being developed allows for many new testing schemes that are not possible to do on the physical product itself due to financial, physical, or other limitations. By upstreaming the data of the physical product experience in real-time to its digital alternative, the digital alternative can react expectedly to the same conditions the physical product is put in. However, the digital simulation can also be duplicated and tested for a multitude of factors, or combinations of, until a general consensus is reached in an order-of-magnitude time reduction compared to traditional means. This allows for testing scenarios that were unattainable due to cost and most importantly time constraints.

The collaborative tools product designers nowadays have access to in certain digitized product design environments are a major tool when it comes to idea sharing

and collaborative work. By sharing product design workflows with different teams of different specialties, groups would be communicating effectively that have a more productive routine and fewer feedback loops which makes the process faster and less error prone due to the big data streams being shared between different parties (Ekins et al., 2014). This also has an added advantage of version control, where designers can at any time return and reimplement previous designs which enhances the workflow by ensuring data retention along the path.

The manufacturing phase, which is being digitalized recently, is a viable source of data that takes the shape of temporal and spatial data of the product being manufactured itself (Lei et al., 2016) as well as the manufacturing rig. When this data is fed into the design phase, the product can be optimized and enhanced to better the manufacturing process as well as making it more efficient. For example, industrial robots working in manufacturing can be used to improve product design by using its various big data points to have product-specific rules (J. Zhang & Fang, 2015).

The released product can have the ability to call home, and supply valuable data of usage and handling (Bae et al., 2015; Bertoni et al., 2017) . This data stream can be utilized. An example would be the telematics implemented across the automotive industry to track car features and report back after being circulated in the market. With the abundance of IoT data, it can be utilized to mine customer needs through their usage (Bertoni, 2020). This data can scale to include every aspect of the product, which means product wear-and-tear, performance, and user habits can be internally acquired and stored for further analysis. Having a data stream from each deployment can pinpoint design weaknesses and areas of improvement that can be improved iteratively as soon as discovered. The product can also transmit its environment's data which serves the

same purpose. And in case of failure or unexpected events, the product designer can know the cause.

After the disposal of the product, the whole data generated over its lifetime can be assessed to determine its condition at end-of-life so to determine performance and certain design parameters to be used in future products. It also can influence optimal recycling or disposal plans (Bin et al., 2015) so that the company can make use of the parts as effectively as possible.

3. Data Flow into manufacturing

During the product manufacturing process, big data from different places can potentially intervene into the manufacturing and help improve it. CAD/CAM data (Yang et al., 2014) and metadata from the design phase, collected exhaustively can contribute to the efficiency of the manufacturing process by allowing to optimize between different runs, as well as interlinking the data together to better evaluate performance of the manufacturing processes, improve supply and material selection, and allow for better product inspection and maintenance (Roy et al., 2014).

Furthermore, the manufacturing rig can generate data (Lei et al., 2016), which can then be analyzed and used to improve the working of the design phase as in adopting better techniques that improves manufacturing and the phase itself regarding product handling and rig operation. Data from industrial machines can also be used to improve the manufacturing process by generating knowledge patterns (J. Zhang & Fang, 2015).

C. Opportunities of Having Big Data in Product Development

As big data flows become more available over time, the importance of their integration into the product development lifecycle would become more of an essential step rather than a luxury. Different big data streams can be harnessed to extract information and knowledge out of them, which would potentially increase the quality and responsiveness of the product and even the whole manufacturing process.

Many early literature discussed the importance of analyzing data on huge scales where human analysis would be limited (Fayyad & Stolorz, 1997). Due to the increasing amounts of data, the need of a decision support system that also is computer-based has been a major research area. Large-scale product development also tends to rely more on previous knowledge and experience of the team rather than data extracted from big data sources (Flanagan et al., 2007). However, the benefits of these ad hoc decisions are not necessarily better than data-supported decisions. Furthermore, data-driven design literature is present in diverse areas, such as using product usage information to enhance the definition of design parameters (Lützenberger et al., 2016) and finding customer demand variation by mining social media and online data (Pajo et al., 2015). To support the product development early design decisions, a two-stage scenario was suggested that allows for the use of data science in engineering design, then shows how that specific data can then be fed into the early design on the product to influence decision making and improve value-driven design (Bertoni et al., 2017).

The current big data flows in the product development lifecycle prove to be of much importance to data-driven design if utilized. As data flows are generated from various phases of the lifecycle or even outside it and hold potentially valuable information, the benefits of using this data to drive the product development process

would reap benefits across the lifecycle itself through various stages. The different opportunities captured by exploiting these data sources would span across different areas:

1. Product quality improvement

Product quality improvement can be made by using big data in various stages across the lifecycle. In the conception phase, various available data sources would allow for better product conceptualization. For example, building upon vast amounts of previous products and market indicators would help in avoiding common pitfalls when designing a new product (Zhan et al., 2018). In the design phase, validated engineering design data would significantly reduce faults due to utilizing previous knowledge solving them (Qin et al., 2016). Customer usage patterns would also play an important part in testing the product which would further contribute to the overall quality (Martí Bigorra & Isaksson, 2017).

2. Better market needs prediction

In the conception phase, various available data sources would allow for better market understanding which will then be internalized by the product development team and applied. This allows for better market understanding based on expressed sentiment but also on data collected from usage (Arora & Malik, 2015), which may reflect subconscious needs the market may align with.

3. Better innovative value prediction

By utilizing data sources such as patent filings, the product developer can help predict invention value and performance (Katila & others, 2000). Patents are often synonymous with innovation, yet value of each patent varies due to different metrics. By analyzing mass patent data, hidden relations of what constitutes successful innovation may arise, which in turn would be used and emulated in order to maximize patent value thus increasing the patented product value.

4. Faster, more efficient development iterations

By utilizing big data sources inside a product lifecycle specific step, much iterative efforts would become faster. For example, by utilizing design-to-design big data flows, such as the digital twin (Tao et al., 2018), much of the iterative process of building and testing would be offloaded to digital copies that would simulate different cases in a fraction of the time and cost.

5. More optimization to manufacturing, distribution processes

While products would obviously benefit, other processes that indirectly participate in the product lifecycle would also get affected by the use of big data. The manufacturing rig data generated can be used to optimize the operations and reduce faults by analyzing the various outputs (Roy et al., 2014). Distribution channels would also benefit from the availability of big data to better plan routes or to better track demand (Christopher & Ryals, 2014). These processes might not affect the product itself but would certainly reflect on the overall efficiency and the surplus gained which would be reinvested in the product.

D. Challenges of Having Big data in Product Development

At the same time, many limitations are found along the way that also need to be tackled. These challenges would hinder the usage of big data in many applications. The challenges can be categorized into 3 main groups:

1. Challenges in the product lifecycle

There is a not enough interest in applying big data inside proven methods that already work, which inhibits the adoption of it at a rate that allows for noticing the change (Raguseo, 2018). Furthermore, different lifecycle data flows have different characteristics, and that would mean large differences in measures used which would mean large variability. For example, market data has different volume or rate when contrasted with data supplied from the product. The data inside the product lifecycle would also exhibit some volatility due to the nature of some sources (Bughin, 2017), which drives down the interest in adopting big data techniques. The field is evolving at a fast pace that may not suit some product environments, especially traditional ones.

2. Challenges of technological factors

The big data analysis scene is still plagued by fragmentation, and there are no specifically tailored solutions that may help transitions. In addition, there are some clear technical bottlenecks when it comes to some data such as real-time use (Obitko et al., 2013), but that is getting better as time passes. There are still many obstacles when

trying to integrate unstructured data sources into a working and reliable stream (Adnan & Akbar, 2019).

3. Challenges based on regulations and legal issues

There are some concerns especially when it comes to management interaction with big data. The data analysis tools are mostly resembling a black box, which means that there is no way to explain the workings of the process and gain the trust that it truly gives good results (Sänger et al., 2014), over a long period of times. Also, there are some privacy conflicts when publishing or using mass data (Tene & Polonetsky, 2011). The data providers may have interest in keeping the data private, which would essentially kill the availability of usable data and that would affect the whole process. There are still many legal areas not developed enough to cover the usage of this data, which means that further exploration should be done in order for proper legal governance to work.

CHAPTER IV

FINDING PATENT INNOVATIVE VALUE COMPARED TO PRIOR ART

In the previous chapters, segmented major data flows fell into their place in the product design process. One flow that stands out is IP and specifically patent data, which happens to be used in the conception and design phases according to our framework, that is essential in determining what is already patented and what is not which will greatly influence the actual design direction of the product and the patents filed later on. As patent data gets bigger and more detailed as of recent years, the need to have means of dissecting this data rather than manually paging through it is becoming more relevant, if not necessary. It is important for product designers to understand patent data and how it affects their outcome without investing too much time and effort.

A. Overview on Patents

A patent is a form of an intellectual property given to the patent filer that allows them to prevent others from using in any way their patented information until a specific time. The patents reveal important information on inventive activities undergone by the submitting entities. The patent is enforced in the location that the patent application covers. To file a patent, a person should make sure that his application is inventive and is able to be utilized in the industry.

Patents contain several different types of data: Dates (priority, application and publication dates), numbers (priority number, application number, citations, ...), names (Inventor name), classification codes (International or Cooperative Patent Classification), text fields (Title, Abstract, Claims, ...), images, additional information (Public registry info, ...)

This data represents the patent over many fields, where for example the patent application date may show when the patent filer applied to the patent office. What interests us in this paper is the abstract text of patents along with other fields such as patent number, code, publication date, among others. The patent abstract field provides a way to tap into the content of the patent. While the claims may have more information on the mechanisms, the abstract text upon inspection and previous literature appears to hold the gist of the patent, and would not be exposed to patent attorney abuse as much as the claims as they are the main criteria for accepting the patent.

Each country has its own patent office where applicants can file a patent application. The World Intellectual Property Organization (WIPO) is responsible for the global governance of the patent system as well as facilitating the sharing of patents and patent-related data.

Patents are mainly classified into different areas according to their topic. To classify a patent, a classification code is given to it according to its what area it belongs to. The classification topics are tree-like and subdivide all the way into specific topics in which patents are designated to. Many patent classification methods were designed as per each country's system at first, until the International Patent Classification (IPC) was made in 1968 and enforced in most jurisdictions. This however was a general system and did not remove the need for local patent offices which may have different standards.

The European Patent Office and the United States Patents and Trademark Office launched a new system, the Cooperative Patent Classification (CPC), to harmonize each proprietary classification system previously present. It was built on the IPC, but also contained other improvements to meet certain requirements. It should be known that a patent can belong to different classes.

The subareas of the CPC code are as follows:

- A letter at the start of the code, known as the section symbol, points to one of main areas under which the patent may be. There are 9 sections in total designated by letters from A to H and Y. For example, a patent under B is related to operations and transport
- A 2-digit number, which is the class symbol, which further divided the section into more specific topics
- A letter, which is a subclass, and expands the class into different areas
- A 1- to 3- digit number, which is a group
- A slash, followed by at least 2 digits, which is a main or sub group

For a quick look on the size of the patent data available, according to the WIPO the number of patents in force in 2018 in around 14.0 million, up 6.7% from the year before (*WIPO IP Facts and Figures*, 2019). The huge number of patents is on an increase, and main patent offices around the world are reporting bigger numbers year-over-year as shown in figure 7. For example, the Chinese patent office reported a 1.5 million new patents submitted in 2018 alone, a 11.6% increase over the past year as shown in the figure. Most of the offices worldwide are also exhibiting relative increases

in submitted patents. This trend is not showing a slow-down, at least in the short to mid-term, and patent numbers are expected to keep increasing.

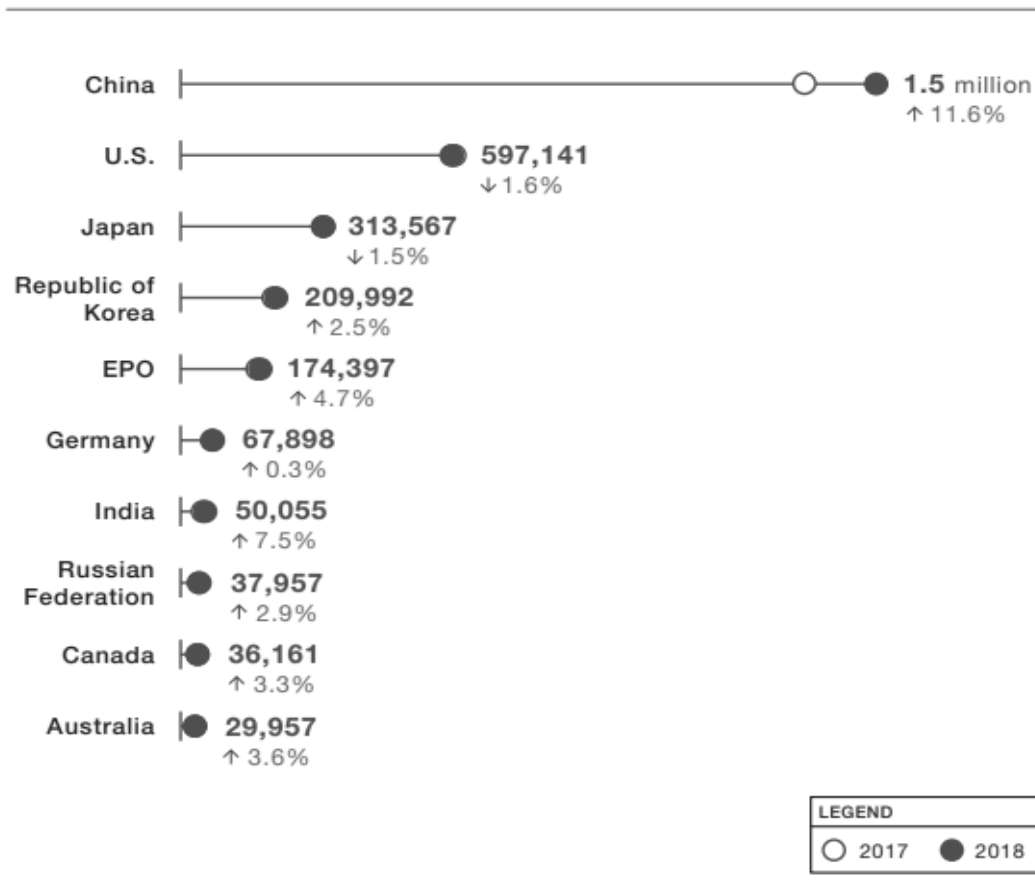



Figure 7 Change in yearly patents in different offices in 2018, according to WIPO (WIPO IP Facts and Figures, 2019)

These numbers aside, the patent documents themselves are of a certain degree of complexity. Most patents have details information in the claims on the details of the patent to prevent ambiguity. Patent attorneys also try to differentiate the patent claims to conform to a set of standards to cover all legal ends required for the patent to get approved. A sample patent from 2017 is shown in figure 8, showing the filing office at field (19), the patent title at field (54), the inventors at field (72), abstract at field (57),

and classes the patent belongs to according to different classification standards at (51) and (52), and other data points.



US 20170000565A1

(19) **United States**
 (12) **Patent Application Publication** (10) **Pub. No.: US 2017/0000565 A1**
 MURPHY et al. (43) **Pub. Date: Jan. 5, 2017**

(54) **ORTHOGNATHIC BIOMECHANICAL SIMULATION** filed on Feb. 14, 2014, provisional application No. 62/049,866, filed on Sep. 12, 2014.

(71) Applicant: **THE JOHNS HOPKINS UNIVERSITY**, Baltimore, MD (US) (51) **Int. Cl.**
A61B 34/10 (2006.01)
A61F 2/28 (2006.01)

(72) Inventors: **Ryan MURPHY**, Columbia, MD (US);
Ehsan BASAFA, Baltimore, MD (US);
Mehran ARMAND, Fulton, MD (US);
Chad GORDON, Lutherville, MD (US);
Gerald GRANT, Goshen, KY (US);
Peter Liacouras, North Potomac, MD (US) (52) **U.S. Cl.**
 CPC **A61B 34/10** (2016.02); **A61F 2/2803** (2013.01); **A61B 2034/105** (2016.02)

(21) Appl. No.: **15/100,241** (57) **ABSTRACT**
 Disclosed is a method of simulating mastication. The method includes obtaining computer-readable three-dimensional representations of a first skeletal fragment including a portion of at least one of a mandible and a maxilla and of a recipient skeletal fragment including a portion of at least one of a mandible and a maxilla. The method also includes obtaining placement data and obtaining muscle insertion data. The method also includes simulating a contraction of a muscle positioned according to the muscle insertion data in a representation of a surgical hybrid comprising at least a portion of the first skeletal fragment positioned according to the placement data relative to at least a portion of the recipient skeletal fragment. The method also includes outputting a representation of mastication represented by the simulating.

(22) PCT Filed: **Nov. 26, 2014**

(86) PCT No.: **PCT/US14/67581**
 § 371 (c)(1),
 (2) Date: **May 27, 2016**

Related U.S. Application Data

(60) Provisional application No. 61/910,204, filed on Nov. 29, 2013, provisional application No. 61/940,196,

Figure 8 A sample patent document published in 2017

Recently in 2010, patent data became available online through the efforts of the United States Patent and Trademark Office. This opened up the possibilities of getting this data and performing data mining on them. It is imperative to get familiar with the different data present in each patent in order to later apply relevant analysis.

B. Literature Review

1. Data-driven analytics on patents

There are many data-driven approaches applied on patent data in an attempt to extract information. With recent development across the storage and retrieval domains, access to bulk patent data was opened to the public, which in turn increased research interest in this specific type of data. A method to enhance keyword search was proposed (Sarica et al., 2019). This method trains an engineering knowledge graph to extract keywords relevant to the given query, which drastically reduces spent time as it substitutes a manual process of finding keywords that would return intended results. Previous research discuss finding similarities in patents used basic search queries and then tries to find synonyms for the keywords using a predefined dictionary file (Mahdabi & Crestani, 2014b; Stefanov & Tait, 2011). This method falls short as the relations between words need to be extracted by hand and thus rendering this method weak. Other means attempt to extract additional keywords from the bibliography of the initial results (Mahdabi & Crestani, 2014a) using the past results and then augment the search with newer values. Another study also used the word2vec scripting model to get the similar words that can be used to find matching patents (Singh & Sharan, 2016). Other measures attempt to model based on word frequency; (Kelly et al., 2018) used term-document frequency to find representations of the corpus, which in other words gets the proportional frequency of different terms inside a given document. They specifically measure for the innovativeness of each patent by overweighting rarely occurring terms up until the patent get released. These methods however do not account for the weaknesses of a frequency-based measures. They do not capture important aspects of the semantics of the text such as repetition of some words across different patents. They also do not distinguish words with similar meanings, completely removing the context out of the words extracted. Those methods also create sparse,

really high-dimensional data which would not be efficient to manipulate. Thus, these measures fail to capture different semantic aspects of the text at hand.

Patent data of leading telecom firms was collected and latent features identified by using Latent Dirichlet Allocation (Suominen et al., 2017), which allowed for the extraction of knowledge profiles per firm and for predicting future trends. To extract function knowledge automatically, an approach based on extracting subject-verb-object triplets from text then applying text mining techniques such as word2vec was made (Cheong et al., 2017), which when compared with manually constructed knowledge databases showed favorable results. To classify patents whether inside a specific group or as an outlier, a support vector machine approach was used (Black & Ciccolo, 2004) in order to have a supervised model that is able to classify topic-related patents. A framework to classify patents according to level of invention, as in the theory of inventive problem solving (Z. Li et al., 2012), allows for better design extraction for computer-aided design (CAD) applications. Creating ideas and conceptual design for engineering from vast knowledge volumes was distilled into knowledge that can be used for engineering design (Liu et al., 2020) by using big data and machine learning means to break down the idea space to clusters, and text mining to process different concept combinations.

Patent landscaping is a process to understand patents and their technological, scientific, and business trends. As patent databases are getting bigger and as the need for better landscaping techniques is getting more relevant due to the complex interactions between different sub-fields of patents, patent landscaping techniques evolved with data science and adopted much of its procedures. Automated patent landscaping processes allow for joint human and machine efforts to leverage human

domain knowledge and machine characteristics generating high quality patent landscape with reduced effort (Abood & Feltenberger, 2018).

2. *Data-driven measures of innovation value and novelty of patents*

When developing a patentable product, a product designer is faced with the question of how much novel or different the product or invention should be in order to have the most value. Previous research has indicated that there is exist a ‘sweet-spot’ when it comes to patent novelty (He & Luo, 2017) when compared to the existing corpus. Contrary to intuition, having a higher novelty would not translate into better value due to the additional risks introduced and thus more exposure to variability (Fleming, 2001). Should the product be of too low or too high novelty, the value decreases on average and thus would not be optimal to target. A product designer should target an in-between novelty level where historically, patents are most likely to have most value.

Patents are considered to be a representative of inventive activities which hold a certain value, and can be seen as a reliable measure of innovation (Di Guardo & Harrigan, 2016; Katila & others, 2000). Patents, however, vary in value which is the measure of innovative qualities of the patent. Forward citation are considered a proxy to measure patent value or impact (Albert et al., 1991; Fischer & Leidinger, 2014; Trajtenberg, 1990), and are considered a direct sign of invention size (Lee et al., 2007). It is also proven that the forward citations a patent gets is highly correlated with its value and reported economic value (Harhoff et al., 1999; Park & Park, 2006). As such, we would be using patent forward citations as a measure of value.

Previous studies attempt to measure novelty as the product of recombination of previous technologies (He & Luo, 2017; Youn et al., 2015) , especially when the combination is uncommon (Simonton, 1999). Citations are generally regarded as the tool for examining value (Harrigan, Di Guardo, & Cowgill, 2017; Harrigan, Di Guardo, Marku, et al., 2017), but also knowledge and technology structure (Castriotta & Di Guardo, 2016; Marku & Zaitsava, 2018). In this case, backward citations are used to extract technological areas the patent use as reference. Patent classification codes categories are considered as different technological areas, and combinations between those classification codes are given scores based on novelty (Uzzi et al., 2013). By counting occurrences of class pairs per each patent, a median as well as distribution of said z-scores can be made which indicates the central and extreme novelty of a patent (He & Luo, 2017). The citation-based novelty measure is limited due to the fact it does not consider single code inventions (Kim et al., 2016), neither does it include external references such as literature and academic sources which may contribute to the said measure and change the actual areas the patent is referring to, hence changing the novelty measure derived.

While citation-based novelty is a helpful measure to use, a text-based novelty measure would provide yet another way to spread out patents, tapping into text to identify unique patents rather than classification classes. Similar studies on engineering design attempted to automatically assess the novelty of engineering design by using similar text-based method (Siddharth et al., 2020). We attempt to use text embedding to measure novelty, as opposed to citation-based measures. Other novelty measure based on text utilize appearances of new words in patents and measure novelty based on it (Balsmeier et al., 2018). Unsupervised text corpora analysis is attempted on various data

sources in an attempt to detect novelty (Guthrie et al., 2008). Novelty detection in corpora of text is done using autoencoders by (Mei et al., 2018) to attempt to find which statements might be atypical to the overall corpus. Social media content is subjected to novelty detection (Amorim et al., 2019) by converting text and images into vectors that are then used to identify outlier cases. A closer area to patents, research papers, has an autoencoder neural network applied to in order to extract novelty (Amplayo et al., 2018) on both a macro-level, represented by metadata and a micro-level, represented by words. To find the novelty in a text corpus, topic homogeneity is must in order to accurately measure novelty (Kilgarrieff & Rose, 1998; Sahlgren & Karlgren, 2005). Patent classification allows for achieving this by labelling topics according to a defined set of codes. This means all patents falling under a specific code level are human labelled to be topically homogenous.

Text-based measures not only allow for novelty measure, but also provide a way to gauge patent performance by only using text excerpts – in this case abstract text – to attempt to have an expected patent innovative value before submitting the patent to the filing office or even allocating time to finish the patent document. This provides a certain benefit over citation-based measures.

Previous studies have already integrated advanced statistical methods with patent analysis. A paper describing the state-of-the-art in IP analysis explains different ways to analysis patent data (Aristodemou & Tietze, 2018). It discusses the current approaches in many different domains such as artificial intelligence, machine learning, and other fields. Autoencoders in a patent context are not mentioned in the paper, which signifies that potentially our application of autoencoders is novel and would further contribute to the field.

We attempt to build on previous literature to introduce a method to assess patent value by using text-based novelty measures as opposed to citation-based methods. By assessing patent text, a product developer has yet another dimension at his disposal. Additionally, finding text-based novelty on different topic resolutions allows for local novelty assessment.

C. Methodology

We discuss an approach to harness patent data in order to measure patent innovative value relative to novelty under a specific area in the corpus. In this case, we are interested in having a homogeneous corpus of patents due to the fact the text-based novelty measures require it. These patents should fall under one area in order to have a common ground to measure similarity with. Figure 9 shows the various steps taken in the methodology.

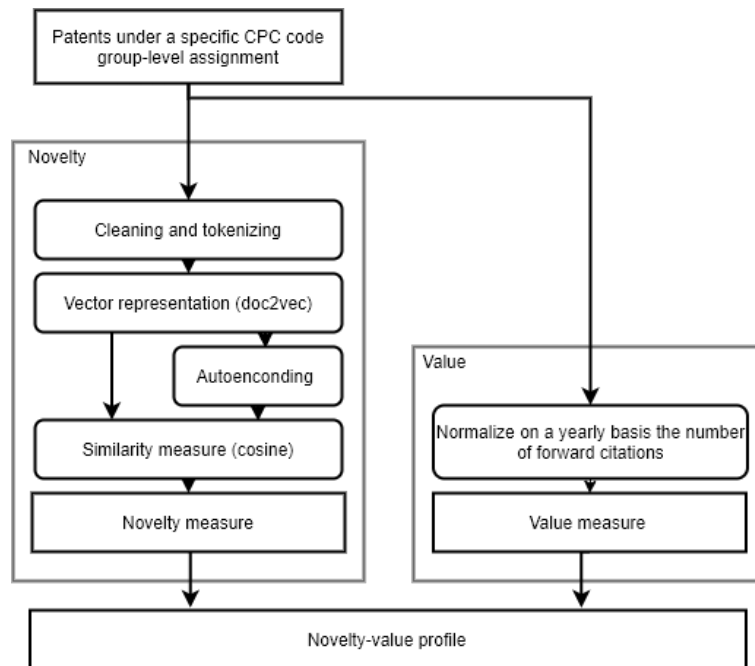


Figure 9 The methodology flowchart

To have a set of patents representative of one area, we select patents in a group-level class according to the CPC. Only patents older than a several number of years and thus known forward citations are selected because they have a known value measure. We then clean the data and convert it to its mathematical embedded form. We then try to detect the novelty of each patent relative to the group by comparing it to a special copy of itself thus finding a novelty measure. Finally, we find the innovative value profile of patents over novelty values. The complete code can be found in appendix A.

1. Obtaining patents

To have patent data in order to further study it, it is essential first to have a reliable source of patent data that encompasses the different filing regions as well as the different times at which patents were published. Many sources offer patent data with varying degrees of limitations and features. It is decided to work with a source that encompasses multiple patent registries in a readable format that is easily fetchable, and thus we chose Google Cloud Platform (hereafter GCP) patent database that is hosted by Google and allows for faster SQL queries due to the huge size of the data (1.8 TB) and the limited resources we have. This offers a more approachable way to interact with patent data, as well as the data being updated frequently. Other sources would have worked just as fine but GCP offers the better integration overall. Patent accessibility over the internet is less than a decade old allowing for getting this data for further analysis a relatively recent achievement.

In order to get the patent corpus, we run a query to get patents exhaustively under a specific classification group as dictated by the CPC. This is because we want to get area-specific patents where we know beforehand that they belong to one area, and thus we can assess similarity later.

The following fields were extracted to our offline storage as shown in table 3. We will be performing different operations on this data in the following sections.

<i>Value</i>	<i>Description</i>
Ab.text	The patent abstract text localized, with mixed text and HTML text formatting
Publication_date	The publication date of the patent, displayed as YYYYMMDD
Publication_number	Publication number of the patent, as in the patent signified 'number'
Ab.Language	The language of the patent to be used in filtering
country_code	The country code of the issuing party
c.code	The classification code (CPC) of the patent

Table 3 The selected fields to be saved for further analysis

2. *Measure of patent novelty*

a. Cleaning and tokenizing

While the data acquired from a data source would hold the information inside, fields themselves may not be clean enough in order to do any meaningful analysis later. Even though the data is present in a readable format, cleaning and tidying the data is important before attempting to manipulate it. Our data is segmented data that can be

parsed correctly, yet fields are not guaranteed to be error-free due to human mistakes and other factors. An example scenario is when an author's name is listed but spelled differently under 2 patents, which will in turn carry into and falsify the final results. Thus, cleaning the patents should be done either manually, or using other tools. Mainly, we would be having an iterative process of tidying the data until we reduce irregularities to a minimum.

Manual statistical measurements were done to verify the data integrity in terms of finding missing fields, errors, or unexpected mishaps after obtaining the data. Another method chosen to tidy the data is Open Refine, an open-source tool that allows for efficient tidying of fields to remove intruding characters, normalize the text case, among other cleaning functions. It is vastly better than manually cleaning in some places and this is because of the redundancy introduced when manually editing them which increases the time needed to perform it. Some approaches to clean on a document-basis were as follows:

Detecting duplicates – some patents has duplicate abstracts, being filed under different offices or for versioning and other reasons. By statistically sorting out the patents, and by constraining the saving algorithm, this was rectified. Iterative filtering based on topic – to ensure a stricter selection of patents rather than a blanket scan, measure are places to filter some keyword-similar yet not related patents. Filtering by language – to keep a homogenous language across all patents, which may or may not be written in English, a filter had to be made to only keep the localized copy of each patent. Then we could have international patents but only in one language. The patents acquired from the GCP are not standardized in a language and may vary according to the issuing office. For example, in our case some have non-Latin words such as the

“CN-103094218-B” patent under our search criteria. We avoid losing this information by selectively extracting the English language versions and validate that no duplicates exist. We also excluded patents older than 1980 due to their incorrect text syntax and since we are interested in newer patents, and also remove patent 9 years old or less due to levelling the time needed for patents to acquire forward citations thus affecting our value measure. This number can be studied further in order to find the optimal point patents are considered well-exposed and forward citations stalling, but we chose to go with a safe lower bound.

After getting the filtered output of the above cleaning, we have a preliminary copy to work on. Patent abstract text itself needed tidying, so we resorted to the following methods to make it conform before feeding it into the model:

Checking field length – to detect any irregularities in the data. For example, the abstract field in some patents was prepended with legal text that has many special characters and text that doesn’t fit with the abstract. Passing this to later stages will mess up the NLP, thus should be removed while keeping the abstract intact. Regex rules – Patents, being filed but many different parties all over the world, have a very unexpected way of formatting. The first issue would be that while some patent abstracts are presented as simple text, others were formatted as HTML. This introduces illegal characters that should be removed. Another issue would be references introduced between each claim in the abstract. Also, numbers and other special characters are present that may be detected as text and thus should be eliminated. The full regex rules are in table 4.

<i>Rule</i>	<i>Selects</i>
-------------	----------------

[0-9]	All numerical chars
^.*Summary - ^.*Abstract - ^.*Problem - ^.*Abstract of the Disclosure - ^.*Objective - A B S T R A C T	All prepending abstract introductory text, legal or otherwise
\[[^\]]*\] - \[[^]]*\)	Bibliography and misc. characters present
\&[^\;]*\;	All HTML characters
[^\w\s]	All punctuations

Table 4 Some regex operations done non-destructively on the corpus

After the text is cleaned, the next step is preparing it to be read correctly by the model. To do this, we had to:

Convert all text into lowercase – to prevent the model from detecting the same word with different case as separate. Tokenize the text – The corpus we had can be described as a list of documents. In order for the model to work the words in each document should be broken apart from their sentences and presented each alone as a token. Removing stop words – Many of the tokens extracted are words of common use such as “this, it, for” that are abundant in all documents and have no value, and thus are removed to limit the words used to meaningful words. We used the Natural Language toolkit (Klein et al., 2009) in order to determine the common words. A custom list of stop words is also appended after iteratively going through the model and refining the input. Lemmatization of words – For the remaining words, to prevent words of the same meaning taking different slots, we resorted to lemmatization, which standardizes the words used and transforms them back to the root word. Stemming of words – To limit the dictionary for the model further, we also applied stemming techniques that removes the suffixes of many words and keep the base form of each word. An example on both methods is shown in table 5.

<i>With lemmatization</i> 'element', 'device', 'slowing'	<i>With lemmatization and stemming</i> 'element', 'devic', 'slow' and base words where applicable
<i>Without lemmatization and stemming</i> 'elements', 'devices', 'slowing'	<i>With stemming</i> 'element', 'devic', 'slow'

Table 5 An example string with stemming and lemmatization applied to

The tokenized document is then processed and labelled with its index number and joined with others. This list of labelled documents would then be passed to the model.

b. Vectorization

In order to analyze the data, it has to be transformed into a mathematical form as opposed to text. To do this, we have to transform the text into a feature representation. To do this we used the doc2vec algorithm, a direct upgrade of word2vec, to learn constant-length representations from variable-length text as opposed to other frequency-based methods.

The word2vec method tries to solve one of the main problems of NLP, the loss of meaning of words after “one-hot” encoding (Mikolov et al., 2013). For example, if we encode “Beirut” “Lebanon” “Pizza” into labels they lose their meanings, and by this they also lose their relation to each other. Clearly, “Beirut” is closer to “Lebanon” as a term than to “Pizza”. Word2vec used 2 main methods to do this, one being using one word to predict the context, called Skip-gram, the other is to predict the target word knowing the context, Continuous bag-of-words, and both ways are illustrated in figure 10 where x is the input and y is the output. It can be noticed that in CBOW the representation of surrounding words (x_{ik}) beside the target word (y_j) are used to predict

it, while the input word (x_k) is used to determine the context words (y_{ij}). For Skip-gram, which is considered more slower, the objective function is:

$$J(\theta) = -\frac{1}{V} \sum_{t=1}^V \sum_{-m \leq j \leq m} \log p(w_{t+j} | w_t) \quad (1)$$

Here, θ contains both the input vector representation of words and the output representation vectors, and m is the length of the window around text selected. For $m = 1$, a center word would have the word before and after it fed as input. The objective function is then minimized by stochastic gradient descent which is our loss which we need to minimize.

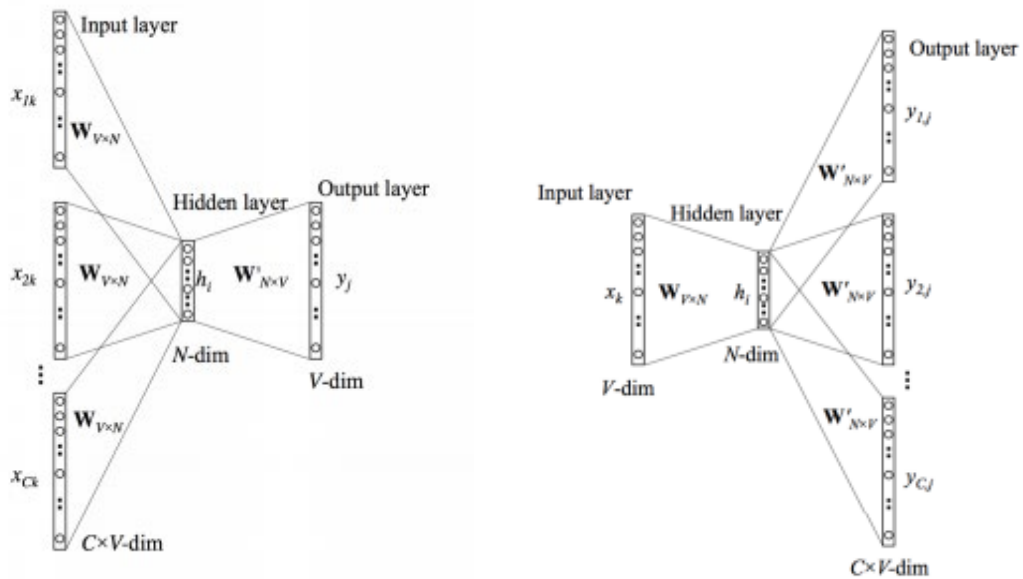


Figure 10 From left to right, the general CBOW and skip-gram representations (Mikolov et al., 2013)

Expanding on this, we will attempt to use the doc2vec method to evaluate the data at hand. The doc2vec algorithm (Quoc Le & Mikolov, 2014) is an upgrade on the word2vec model that allows to use the patent as a whole as one vector entity.

The Doc2vec method adds to the previous by adding a paragraph ID, which in our case would be each patent. This method has been proven to capture not only word but document similarity, which when applied to our context, patents, would mean capturing the details of each patent. It maps each patent to a vector that is then trained by a neural network to predict the context.

In our case, we would be using the distributed memory (DM) method, which is similar to the CBOW. This method introduces a new document-level token in addition to the words. However, the vectors are not summed but concatenated. The modified objective is to predict the target word knowing the concatenated word and document. The parameters of the classifier and the word vectors are not needed and backpropagation is used to tune the paragraph vectors.

It is important to say that although the model is fairly high in dimensions, a certain loss is expected to happen since some semantic value is lost in the conversion coupled with the limited size of the vector and other factors which may contribute to it.

c. Autoencoding

We then want to identify the novelty of each patent. An autoencoder learns a compressed representation of an input data of a fixed vector size, then tries to construct the initial data from that. It consists of 3 parts, the encoder which reduces the input into an encoded vector, the bottleneck which is the most compressed vector the data passes through, then the decoder which constructs the data from the previous step to try to emulate the input, with a varying degree of loss in the process. An autoencoder neural network tries to compress the input data through its bottleneck layer, then tries to rebuild it for the output. A general autoencoder that has same size input and output with a hidden layer is illustrated in figure 11. When passing an input that is expected, the

autoencoder can reconstruct it with considerably better results than when passing outliers. For example if we train the autoencoder on documents related to a certain topic C, then try to predict 2 documents one from the same topic and one an outlier, the outlier predicted form will look dissimilar to its initial form, while the common document will have a fairly closer predicted copy.

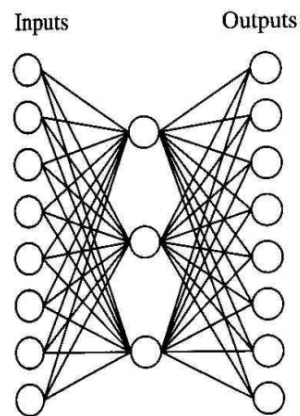


Figure 11 A general autoencoder (Mitchell, 1997)

An autoencoder was used in order to reduce and compress features in the input, then construct back a copy that when compared to the input, allows us to find a certain deviation. For each patent to have a correctly predicted vector, we only included chronologically older patents in the training data of the autoencoder for each patent we are trying to predict. This means the corpus training data includes patents up to the patent of interest, then the patent of interest is predicted using the autoencoder. This process is run on a random set of 100 patents to have enough data points to plot.

By having a predicted output from the autoencoder, and as the autoencoder removes the noise when encoding to a fewer features layer, unique patents would lose their content when being encoded, then the decoded output would lack the content of the input. This means when comparing each unique patent with its predicted copy, the

two will not look alike. Alternatively, when comparing a patent that has no unique content compared to the corpus, its copy would keep most of the features intact if slightly different and would be relatively similar to its original form.

d. Novelty measure

Since the patent texts are represented as vectors extracted from textual data, the use of distance metric to measure similarity cannot be square distance or Euclidean as text data needs normalization to correctly get similar entries. The cosine similarity measure works with vector representations of text (Huang, 2008), and we use cosine similarity to find the similarity between the patent and the corpus. Equation (2) determines the similarity between the initial patent and its autoencoder version.

$$sim(P_i, P_j) = \frac{P_i P_j}{\|P_i\| \|P_j\|} \quad (2)$$

Our novelty measure is the complement of the similarity measure. Since unique patents would reconstruct poorly, the output will drastically change, which means it will tend to have a lower similarity or higher novelty when compared with the initial vectors. Likewise, common patents would have a lower novelty or better similarity value. Thus, a patent with a low similarity score is considered more novel compared to the corpus.

3. *Measure of patent value*

To get the patent value, we use the forward citations as a measure as suggested by (Albert et al., 1991; Fischer & Leidinger, 2014). The value of a patent or an invention is its significance in various social and economic aspects. The value of the patent is strongly correlated with the number of forward citations it gets, which permits

us to use it. We normalize the citations on a yearly basis to account for variation in average citations per patent changes.

After getting both the novelty and value measures, we can construct the plot that shows the change of novelty and how it affects value. Previous research tries to find the relation between patent novelty and value (He & Luo, 2017) using a citation-based novelty approach. Diverging from that, our novelty measure, which is based on text, attempts to rectify some of the limitations of the previous method and offering another point of view into novelty. We attempt to identify the novelty and value profile in order to identify any meaningful relations between these variables. Citation-based measures had a ‘sweet-spot’ of novelty where value was maximized. We try to verify if local patent corpus under a specific classification level feature the same characteristics of a certain ‘sweet spot’ where patents have the best value.

4. Validation of novelty approach

Since our novelty measure method is unsupervised, it is hard to validate due to the lack of any labelling. We treat the problem as a supervised one to have some insight on how well it performs.

To get labels on novel or non-novel data in order to validate our autoencoder, we trained our doc2vec model on our data then generated 5000 artificial data points outside the bounding hypercube of the patent vectors in the feature space with a maximum offset of 5 in either direction. The artificial data is around 7% of the normal data size. For example, for a specific feature with a minimum of -2 and a maximum of 1, the artificial data can only be in the intervals [-7,-2) or (1,6] respectively. This data is

then labelled as novel while the rest is labelled otherwise. This is the foundation behind our ground truth. We then split the data into training and testing data with an 80/20 split. We then train the autoencoder on the training data and predict the testing data output. The testing data predicted value is then compared with the input using our novelty measure. We assess the approach by using the ground truth we labeled earlier. We treat the problem as a classification problem, where one group of patents would be classified as novel and the rest would not be. This allows us to construct a receiver operating characteristic (ROC) curve to evaluate the method, as shown in figure 12. An ROC curve allows us to determine the separability measure, or in other terms the ability to distinguish between categories. The curve explains how well our approach was able to identify the novelty in patents compared to our ground truth, by measuring the True Positive Rate (TPR), the ratio of identified true novel patents to the total novel patents, and its variation with the False Positive Rate (FPR), the ratio of identified falsely novel patents to the total common patents. The method was able to identify our truths of novel patents fairly accurately with an Area Under Curve (AUC) value of 0.7.

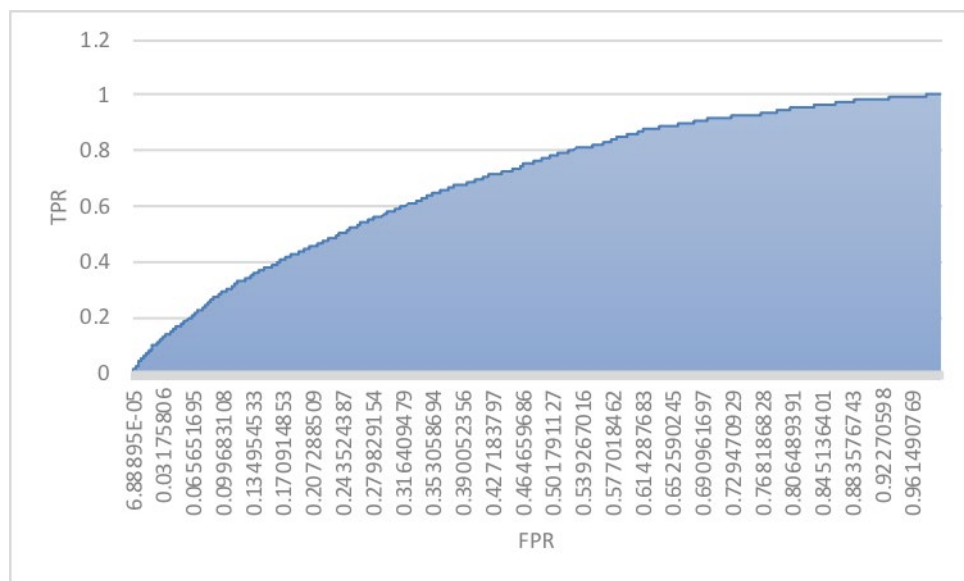


Figure 12 The ROC curve

D. Results of a Case on a Specific Patent Group According to the CPC

1. *Obtaining and converting patents to a vector representation*

We chose an arbitrary group of patents under the group-level classification. The reason would be that since we are aiming to identify similarity, our data has to take shape as reporting to a clearly defined description, which the group level in the CPC achieves. For example, the chosen patent set belongs mainly to prosthetics. Table 6 shows the chosen group level class code and description as shown in the CPC.

A61F2/	Filters implantable into blood vessels; Prostheses, i.e. artificial substitutes or replacements for parts of the body; Appliances for connecting them with the body; Devices providing patency to, or preventing collapsing of, tubular structures of the body
--------	--

Table 6 The chosen CPC group code

The patents are then prepared and the text cleaned from noise, filtered, then tokenized. The list of patents totaling 72313 is then fed into the doc2vec algorithm in order to get the document vector representation of the corpus.

2. *Applying the text-based novelty measure*

The vectors are then used to run several instances of the autoencoder where training data is used up to the patent being used to output a predicted version. The number of instances ran was 100. The cosine similarity of each patent's input vector and the predicted one is calculated, and the result was stored to be analyzed.

3. *The novelty-value profile*

After getting the novelty measure, we get the value measure which is the forward citations per each patent in the corpus. We then plot novelty and value, where novelty on the x-axis is compared with the value measure on the y-axis as shown in figure 13 (a) below. We then fit a regression curve to the data to determine the average value for different novelty levels as shown in figure 13 (b).

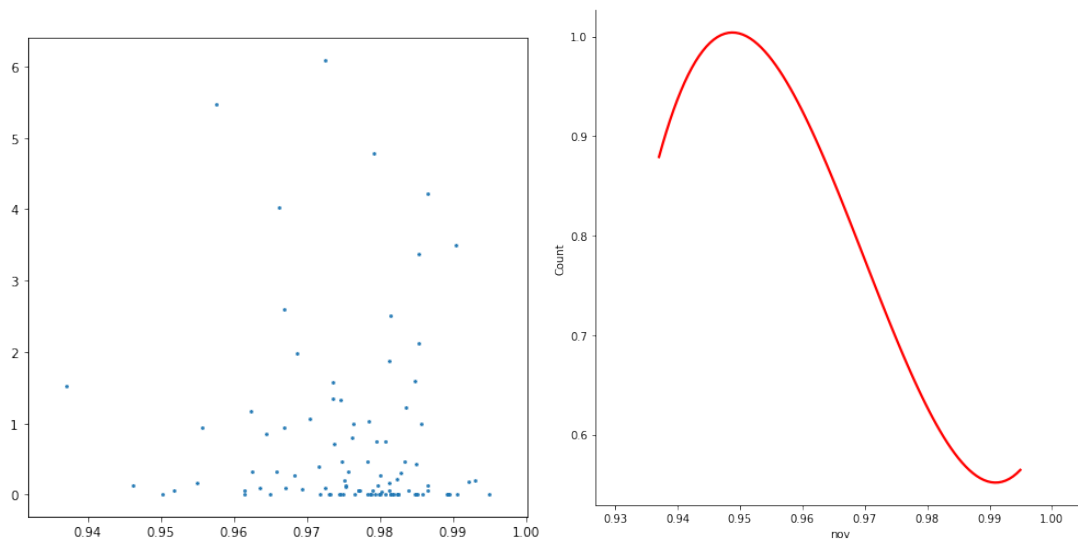


Figure 13 From left to right, (a) the scatter plot of the patent corpus value-novelty profile and (b) the fitted curve

Our results show that for a specific range of novelty for our local corpus (near 0.95), the average value measure reaches a crest which signifies better value for patents in this range. This shows a ‘sweet spot’ of novelty that should be achieved in order to have a better expected value measure, with the measure being text-based rather than citation-based (He & Luo, 2017).

4. *Forecasting a patent’s future value by using its abstract*

We took a highly relevant case in the product design field that is a possible application to our method. A product developer would want to assess a patent or invention’s actual value before submitting it or developing a full patent application. As our novelty measure is text-based, a patentee could simply predict the value of a draft abstract to a certain degree by using the trained model on existing corpora. In our case, we predict patent novelty using our model which will then be used to predict the innovative value of the patent in use, which is built using corpus data. This allows the patentee to estimate patent value relative to the corpus.

Patent number	US-2014148913-A1	US-2010211188-A1
Abstract text	<p>A joint prosthesis comprises a distal component for anchoring to a first bone a proximal component for anchoring to a second bone and a coupling piece that together with the first component forms a flexion bearing around a first axis and together with the second component forms a rotary bearing formed by the pin and the bearing bush around a second axis oriented transversely to the first axis The rotary bearing comprises a multilayer bearing insert having a sliding sleeve surrounding the pin and a support sleeve that encloses said sliding sleeve and is fastened to the coupling piece by means of a securing element wherein the securing element comprises an actuation unit within the support sleeve and can be connected to the</p>	<p>A temporary diagnostic prosthetic socket mounting system and kit including a generally circular test mounting block defined by an annular groove and four axial cutouts extending from the lower surface of the block to the upper surface The axial cutouts are at least as deep as the annular groove A band extending around the perimeter of the block is provided which spans the axial cutouts thereby forming a void in the cutouts beneath the tape The block is secured to a prosthetic diagnostic socket by adhesive with the band extending up over the upper edge of the block and onto the outer surface of the socket Casting tape is applied over the socketblock joint extending into the groove After diagnostic fitting and transfer of the alignment a cast saw can be passed around the joint through the casting tape and into the adhesive between the upper surface of the block and the rounded</p>

coupling piece such as to ensure tensile strength by means of two aligned bores in the support sleeve and the coupling piece	distal end of the socket The gap behind the band in the cutouts provides a void space for the saw that avoids damage to the block and the casting tape is severed and easily removed
--	--

Table 7 The 2 patents and their text

To test the model, we create a synthesized abstract text by concatenating both texts from 2 abstracts present in the corpus shown in table 7. This new abstract was then fed into the model which is trained on the corpus data. The model outputted a 0.96 novelty value. This in turn, when applied to the novelty and value relationship gave an expected innovative value of ratio 1, which is relatively high and falls in the sweet spot of the novelty measure.

Another aspect to consider is the sensitivity of the model to changes in the abstract text. We ran a sensitivity analysis on the text to measure the change in the output when the input changes. To do this, we randomly removed ten terms from the synthesized text and ran it through the model. The inferred vector from the summation of the two documents is shown below. The resulting value was 0.963, a small deviation from the initial value.

['joint', 'prothesi', 'compris', 'distal', 'compon', 'anchor', 'bone', 'proxim', 'compon', 'anchor', 'bone', 'coupl', 'piec', 'togeth', 'compon', 'form', 'flexion', 'bear', 'around', 'axi', 'togeth', 'compon', 'form', 'rotari', 'bear', 'form', 'pin', 'bear', 'bush', 'around', 'axi', 'orient', 'transvers', 'axi', 'rotari', 'bear', 'compris', 'multilay', 'bear', 'insert', 'slide', 'sleev', 'surround', 'pin', 'support', 'sleev', 'enclos', 'said', 'slide', 'sleev', 'fasten', 'coupl', 'piec', 'mean', 'secur', 'element', 'wherein', 'secur', 'element', 'compris', 'actuat', 'unit', 'within', 'support', 'sleev', 'connect', 'coupl', 'piec', 'ensur', 'tensil', 'strength', 'mean', 'align', 'bore', 'support', 'sleev', 'coupl', 'piec', 'temporari', 'diagnost', 'prosthet', 'socket', 'mount', 'system', 'kit', 'includ', 'general', 'circular', 'test', 'mount', 'block', 'defin', 'annular', 'groov', 'four', 'axial', 'cutout', 'extend', 'lower', 'surfac', 'block', 'upper', 'surfac', 'axial', 'cutout', 'least', 'deep', 'annular', 'groov', 'band', 'extend', 'around', 'perimet', 'block', 'provid', 'span', 'axial', 'cutout', 'therebi', 'form', 'void', 'cutout', 'beneath', 'tape', 'block', 'secur', 'prosthet', 'diagnost', 'socket', 'adhes', 'band', 'extend', 'upper', 'edg', 'block', 'onto', 'outer', 'surfac', 'socket', 'cast', 'tape', 'appli', 'socketblock', 'joint', 'extend', 'groov', 'diagnost', 'fit', 'transfer', 'align', 'cast', 'saw', 'pass', 'around', 'joint', 'cast', 'tape', 'adhes', 'upper', 'surfac', 'block', 'round', 'distal', 'end', 'socket', 'gap', 'behind', 'band', 'cutout', 'provid', 'void', 'space', 'saw', 'avoid', 'damag', 'block', 'cast', 'tape', 'sever', 'easili', 'remov']

CHAPTER V

CONCLUSION

As the data collected and used within the product development lifecycle increases with the transition into a digital world, it becomes increasingly conforming with the definition of big data, thus making it big data. This thesis attempts to define a framework to classify incoming and outgoing data sources in the product development process. It also showcases a description of data sources with their origins and destinations that a product developer has to map out and convert into applicable knowledge in order to streamline the product development process while discussing the challenges and benefits of using them. A specific data flow, patent data, proves to be a potential major element to be considered when approaching product development. We proposed a model to extract text representations from a specific patent group under a defined classification code, then extract a text-based patent novelty measure with respect to other patents in a local corpus along with a value measure. We then study the variation of patent value with novelty and identify a target range of novelty that the value of patents is the best at. The text-based novelty measure would complement other citation-based measures and would help product developers acquire a better idea on the novelty-value relation.

There are several decisions made when constructing our model, which could introduce limitations and drawbacks in the intended result in retrospect. These limitations are discussed with the reasoning behind each and its respective effects.

We chose to run text mining on the abstracts rather than the claims due to the following reasons. First, while naturally claims text is the more descriptive part of the patent stating each claim with detail, abstract text is sufficiently inclusive (Adams, 2010) of what

the patent is about. Additionally, claims often contain complex elements and characters, which are a challenge to clean in an automated way, and may introduce further noise in the result which diminishes any perceived improvement. Finally, although claim text is more important when determining the patent grant and is the most edited and revised text, it also means that patent lawyers are more exposed to it and that would mean more specifically reworded text to match certain criteria.

For the level of detail chosen as the group level classification, we can argue that we can encompass more general classes or even dive deeper into the hierarchy. While this argument is valid, we chose the current level of detail since it pertains to a fairly specific topic that has a sufficient number of documents for our model to work. More exploration can be done in this regard in future work.

For the novelty method validation, we could validate the results of the autoencoder itself, which means that the novelty measure at least correctly behaves when it comes to textual uniqueness. However, we did not use expert patent novelty assessment and that is an area to be tackled later.

In addition to that, having more blackbox-y ways of getting novelty measures of patents would impose a challenge in clearly understanding the approach and reproducing it. This would potentially affect acceptance of the methods used and would require careful validation of the approach to prevent getting false output.

Since the novelty method is composed of fairly different components, certain degree of loss should be expected after each step. For example, to convert text into its doc2vec representation entails partial loss since context would not be preserved 1:1 when transformed into mathematical numbers.

Those different aspects are main weaknesses of the model and should be actively improved upon. Future research may allow for bypassing and minimizing several of those, yet as of today, the methods used in this study exhibit the limitations listed above.

This research was implemented as a proof-of-concept of the pipeline used. For further integration into the patentee's environment, an intuitive interface should be developed and compiled to automate the process for it to be viable to use in a performant manner. This can be built upon the methods discussed in the paper. Other future work would attempt improving the output of the model in terms of reduction of noise.

REFERENCES

- Aamodt, A., & Nygård, M. (1995). Different Roles and Mutual Dependencies of Data, Information, and Knowledge-An AI Perspective on their Integration. *Data Knowl. Eng.*, *16*(3), 191–222.
- Abood, A., & Feltenberger, D. (2018). Automated patent landscaping. *Artificial Intelligence and Law*, *26*(2), 103–125.
- Ackoff, R. L. (1989). From data to wisdom. *Journal of Applied Systems Analysis*, *16*(1), 3–9.
- Adams, S. (2010). The text, the full text and nothing but the text: Part 1—Standards for creating textual information in patent documents and general search implications. *World Patent Information*, *32*(1), 22–29.
- Adnan, K., & Akbar, R. (2019). Limitations of information extraction methods and techniques for heterogeneous unstructured big data. *International Journal of Engineering Business Management*, *11*, 1847979019890771.
- Albert, M. B., Avery, D., Narin, F., & McAllister, P. (1991). Direct validation of citation counts as indicators of industrially important patents. *Research Policy*, *20*(3), 251–259.
- Amorim, M., Bortoloti, F. D., Ciarelli, P. M., Salles, E. O., & Cavalieri, D. C. (2019). Novelty Detection in Social Media by Fusing Text and Image Into a Single Structure. *IEEE Access*, *7*, 132786–132802.
- Amplayo, R. K., Hong, S., & Song, M. (2018). Network-based approach to detect novelty of scholarly literature. *Information Sciences*, *422*, 542–557.

- Aristodemou, L., & Tietze, F. (2018). The state-of-the-art on Intellectual Property Analytics (IPA): A literature review on artificial intelligence, machine learning and deep learning methods for analysing intellectual property (IP) data. *World Patent Information*, 55, 37–51.
- Arnold, C. R. B., Stone, R. B., & McAdams, D. A. (2008). *MEMIC: an interactive morphological matrix tool for automated concept generation*. 1196.
- Arora, D., & Malik, P. (2015). Analytics: Key to Go from Generating Big Data to Deriving Business Value. *2015 IEEE First International Conference on Big Data Computing Service and Applications*, 446–452.
- Bae, H.-R., Ando, H., Nam, S., Kim, S., & Ha, C. (2015). Fatigue design load identification using engineering data analytics. *Journal of Mechanical Design*, 137(1), 011001.
- Baglee, D., & Marttonen, S. (2015). The need for Big Data collection and analyses to support the development of an advanced maintenance strategy. *Proceedings of the International Conference on Data Mining (DMIN)*, 3.
- Balsmeier, B., Assaf, M., Chesebro, T., Fierro, G., Johnson, K., Johnson, S., Li, G.-C., Lück, S., O'Reagan, D., Yeh, B., & others. (2018). Machine learning and natural language processing on the patent corpus: Data, tools, and new measures. *Journal of Economics & Management Strategy*, 27(3), 535–553.
- Bertoni, A. (2018). Role and challenges of data-driven design in the product innovation process. *IFAC-PapersOnLine*, 51(11), 1107–1112.
- Bertoni, A. (2020). Data-driven design in concept development: Systematic review and missed opportunities. *Proceedings of the Design Society: DESIGN Conference*, 1, 101–110.

- Bertoni, A., Larsson, T., Larsson, J., & Elfsberg, J. (2017). *Mining data to design value: A demonstrator in early design*. ICED17 21st International Conference on Engineering Design.
- Bin, S., Zhiquan, Y., Jonathan, L. S. C., Jiewei, D. K., Kurle, D., Cerdas, F., & Herrmann, C. (2015). A big data analytics approach to develop industrial symbioses in large cities. *Procedia CIRP*, 29, 450–455.
- Biolchini, J., Mian, P. G., Natali, A. C. C., & Travassos, G. H. (2005). Systematic review in software engineering. *System Engineering and Computer Science Department COPPE/UFRJ, Technical Report ES*, 679(05), 45.
- Black, D., & Ciccolo, P. (2004). *Machine learning for patent classification*. Camus. Available from <http://www.stanford.edu/class/cs229/proj2005>
- Brossard, M., Hepp, D., & Erntell, H. (2018, June). *Accelerating product development: The tools you need now*.
- Bryson, S., Kenwright, D., Cox, M., Ellsworth, D., & Haimes, R. (1999). Visually exploring gigabyte data sets in real time. *Communications of the ACM*, 42(8), 82–90.
- Bughin, J. (2017). Ten big lessons learned from big data analytics. *Applied Marketing Analytics*, 2(4), 286–295.
- Castriotta, M., & Di Guardo, M. C. (2016). Disentangling the automotive technology structure: A patent co-citation analysis. *Scientometrics*, 107(2), 819–837.
- Cheong, H., Li, W., Cheung, A., Nogueira, A., & Iorio, F. (2017). Automated extraction of function knowledge from text. *Journal of Mechanical Design*, 139(11).
- Christopher, M., & Ryals, L. J. (2014). The supply chain becomes the demand chain. *Journal of Business Logistics*, 35(1), 29–35.

- Cox, M., & Ellsworth, D. (1997). *Application-controlled demand paging for out-of-core visualization*. 235–244.
- Di Guardo, M. C., & Harrigan, K. (2016). Shaping the path to inventive activity: The role of past experience in R&D alliances. *The Journal of Technology Transfer*, 41(2), 250–269.
- Doblin Group. (2012). *96% of all innovations fail to return their cost of capital*.
- Ekins, S., Clark, A. M., Swamidass, S. J., Litterman, N., & Williams, A. J. (2014). Bigger data, collaborative tools and the future of predictive drug discovery. *Journal of Computer-Aided Molecular Design*, 28(10), 997–1008.
<https://doi.org/10.1007/s10822-014-9762-y>
- English, K., Naim, A., Lewis, K., Schmidt, S., Viswanathan, V., Linsey, J., McAdams, D. A., Bishop, B., Campbell, M. I., & Poppa, K. (2010). Impacting designer creativity through IT-enabled concept generation. *Journal of Computing and Information Science in Engineering*, 10(3), 031007.
- Fayyad, U., & Stolorz, P. (1997). Data mining and KDD: Promise and challenges. *Future Generation Computer Systems*, 13(2–3), 99–115.
- Fischer, T., & Leidinger, J. (2014). Testing patent value indicators on directly observed patent value—An empirical analysis of Ocean Tomo patent auctions. *Research Policy*, 43(3), 519–529.
- Flanagan, T., Eckert, C., & Clarkson, P. J. (2007). Externalizing tacit overview knowledge: A model-based approach to supporting design teams. *AI EDAM*, 21(3), 227–242.
- Fleming, L. (2001). Recombinant uncertainty in technological search. *Management Science*, 47(1), 117–132.

- Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2), 137–144.
<https://doi.org/10.1016/j.ijinfomgt.2014.10.007>
- Google Trends. (2019, January 20). *Google Trends—'Big Data'*. Google Trends.
trends.google.com
- Guthrie, D., Guthrie, L., & Wilks, Y. (2008). An unsupervised approach for the detection of outliers in corpora. *Statistics*, 3409–3413.
- Harhoff, D., Narin, F., Scherer, F. M., & Vopel, K. (1999). Citation frequency and the value of patented inventions. *Review of Economics and Statistics*, 81(3), 511–515.
- Harrigan, K. R., Di Guardo, M. C., & Cowgill, B. (2017). Multiplicative-innovation synergies: Tests in technological acquisitions. *The Journal of Technology Transfer*, 42(5), 1212–1233.
- Harrigan, K. R., Di Guardo, M. C., Marku, E., & Velez, B. N. (2017). Using a distance measure to operationalise patent originality. *Technology Analysis & Strategic Management*, 29(9), 988–1001.
- He, Y., & Luo, J. (2017). The novelty 'sweet spot' of invention. *Design Science*, 3.
- Holler, M., Neiditsch, G., Uebernickel, F., & Brenner, W. (2017). *Digital Product Innovation in Manufacturing Industries-Towards a Taxonomy for Feedback-driven Product Development Scenarios*.
- Huang, A. (2008). Similarity measures for text document clustering. *Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008)*, Christchurch, New Zealand, 4, 9–56.

- Ireland, R., & Liu, A. (2018). Application of data analytics for product design: Sentiment analysis of online product reviews. *CIRP Journal of Manufacturing Science and Technology*, 23, 128–144.
<https://doi.org/10.1016/j.cirpj.2018.06.003>
- ISO/IEC/IEEE 15288:2015. (2015). *Systems and software engineering—System life cycle processes*. ISO.
- Jin, J., Liu, Y., Ji, P., & Liu, H. (2016). Understanding big consumer opinion data for market-driven product design. *International Journal of Production Research*, 54(10), 3019–3041.
- Katila, R., & others. (2000). Using patent data to measure innovation performance. *International Journal of Business Performance Management*, 2(1/2/3), 180–193.
- Kelly, B., Papanikolaou, D., Seru, A., & Taddy, M. (2018). *Measuring technological innovation over the long run*. National Bureau of Economic Research.
- Kilgarriff, A., & Rose, T. (1998). Measures for corpus similarity and homogeneity. *Proceedings of the Third Conference on Empirical Methods for Natural Language Processing*, 46–52.
- Kim, D., Cerigo, D. B., Jeong, H., & Youn, H. (2016). Technological novelty profile and invention's future impact. *EPJ Data Science*, 5(1), 1–15.
- Kitchin, R., & McArdle, G. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society*, 3(1), 205395171663113. <https://doi.org/10.1177/2053951716631130>
- Klein, E., Loper, E., & Steven, B. (2009). *Natural Language Processing with Python*. O'Reilly Media.
- Krishnan, K. (2013). *Data warehousing in the age of big data*. Newnes.

- Kuo, Y.-H., & Kusiak, A. (2018). From data to big data in production research: The past and future trends. *International Journal of Production Research*, 1–26.
- Kurtoglu, T., Campbell, M. I., & Linsey, J. S. (2009). An experimental study on the effects of a computational design tool on concept generation. *Design Studies*, 30(6), 676–703.
- Labrinidis, A., & Jagadish, H. V. (2012). Challenges and opportunities with big data. *Proceedings of the VLDB Endowment*, 5(12), 2032–2033.
- Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META Group Research Note*, 6(70), 1.
- LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big data, analytics and the path from insights to value. *MIT Sloan Management Review*, 52(2), 21.
- Le, Qize, & Panchal, J. H. (2012). Analysis of the interdependent co-evolution of product structures and community structures using dependency modelling techniques. *Journal of Engineering Design*, 23(10–11), 807–828.
- Le, Quoc, & Mikolov, T. (2014). Distributed representations of sentences and documents. *International Conference on Machine Learning*, 1188–1196.
- Lee, Y.-G., Lee, J.-D., Song, Y.-I., & Lee, S.-J. (2007). An in-depth empirical analysis of patent citation counts using zero-inflated count data model: The case of KIST. *Scientometrics*, 70(1), 27–39.
- Lei, Y., Jia, F., Lin, J., Xing, S., & Ding, S. X. (2016). An intelligent fault diagnosis method using unsupervised feature learning towards mechanical big data. *IEEE Transactions on Industrial Electronics*, 63(5), 3137–3147.

- Li, B., & Xie, S. (2015). Module partition for 3D CAD assembly models: A hierarchical clustering method based on component dependencies. *International Journal of Production Research*, 53(17), 5224–5240.
- Li, J., Tao, F., Cheng, Y., & Zhao, L. (2015). Big data in product lifecycle management. *The International Journal of Advanced Manufacturing Technology*, 81(1–4), 667–684.
- Li, Z., Tate, D., Lane, C., & Adams, C. (2012). A framework for automatic TRIZ level of invention estimation of patents using natural language processing, knowledge-transfer and patent citation metrics. *Computer-Aided Design*, 44(10), 987–1010.
- Liew, A. (2007). Understanding data, information, knowledge and their inter-relationships. *Journal of Knowledge Management Practice*, 8(2), 1–16.
- Liu, Q., Wang, K., Li, Y., & Liu, Y. (2020). Data-Driven Concept Network for Inspiring Designers' Idea Generation. *Journal of Computing and Information Science in Engineering*, 20(3).
- Luo, J., Olechowski, A. L., & Magee, C. L. (2014). Technology-based design and sustainable economic growth. *Technovation*, 34(11), 663–677.
- Lützenberger, J., Klein, P., Hribernik, K., & Thoben, K.-D. (2016). Improving product-service systems by exploiting information from the usage phase. A case study. *Procedia CIRP*, 47, 376–381.
- Mahdabi, P., & Crestani, F. (2014a). The effect of citation analysis on query expansion for patent retrieval. *Information Retrieval*, 17(5–6), 412–429.
- Mahdabi, P., & Crestani, F. (2014b). Query-driven mining of citation networks for patent citation retrieval and recommendation. *Proceedings of the 23rd ACM*

International Conference on Conference on Information and Knowledge Management, 1659–1668.

Manyika, J., Chui, M., Brown, B., Roxburgh, C., Bughin, J., & Dobbs, R. (2011). *Big data: The next frontier for innovation, competition, productivity*.

Marku, E., & Zaitsava, M. (2018). *Smart Grid Domain: Technology structure and innovation trends*.

Martí Bigorra, A., & Isaksson, O. (2017). Combining customer needs and the customer's way of using the product to set customer-focused targets in the House of Quality. *International Journal of Production Research*, 55(8), 2320–2335.

Mei, M., Guo, X., Williams, B. C., Doholi, S., Kenworthy, J. B., Paulus, P. B., & Minai, A. A. (2018). Using semantic clustering and autoencoders for detecting novelty in corpora of short texts. *2018 International Joint Conference on Neural Networks (IJCNN)*, 1–8.

Michelino, F., Lamberti, E., Cammarano, A., & Caputo, M. (2015). Measuring open innovation in the Bio-Pharmaceutical industry. *Creativity and Innovation Management*, 24(1), 4–28.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *ArXiv Preprint ArXiv:1301.3781*.

Mitchell, T. (1997). *Machine learning*. WCB. McGraw-Hill.

Morente-Molinera, J. A., Pérez, I. J., Ureña, M. R., & Herrera-Viedma, E. (2016). Creating knowledge databases for storing and sharing people knowledge automatically using group decision making and fuzzy ontologies. *Information Sciences*, 328, 418–434.

- Obitko, M., Jirkovský, V., & Bezdiček, J. (2013). Big data challenges in industrial automation. In *Industrial Applications of Holonic and Multi-Agent Systems* (pp. 305–316). Springer.
- Opresnik, D., & Taisch, M. (2015). The value of big data in servitization. *International Journal of Production Economics*, *165*, 174–184.
- Pajo, S., Verhaegen, P.-A., Vandevenne, D., & Duflou, J. R. (2015). Fast lead user identification framework. *Procedia Engineering*, *131*, 1140–1145.
- Park, G., & Park, Y. (2006). On the measurement of patent stock as knowledge indicators. *Technological Forecasting and Social Change*, *73*(7), 793–812.
- Parlikad, A. K., & McFarlane, D. (2007). RFID-based product information in end-of-life decision making. *Control Engineering Practice*, *15*(11), 1348–1363.
- Parraguez, P., & Maier, A. (2017). *Data-driven engineering design research: Opportunities using open data*. *7*, 21–25.
- PatentsView. (2020). <https://www.patentsview.org/web/#viz/relationships>
- Qin, S., Van der Velde, D., Chatzakis, E., McStea, T., & Smith, N. (2016). Exploring barriers and opportunities in adopting crowdsourcing based new product development in manufacturing SMEs. *Chinese Journal of Mechanical Engineering*, *29*(6), 1052–1066.
- Raguseo, E. (2018). Big data technologies: An empirical investigation on their adoption, benefits and risks for companies. *International Journal of Information Management*, *38*(1), 187–195.
- Roy, U., Zhu, B., Li, Y., Zhang, H., & Yaman, O. (2014). Mining big data in manufacturing: Requirement analysis, tools and techniques. *ASME 2014 International Mechanical Engineering Congress and Exposition*.

- Sagiroglu, S., & Sinanc, D. (2013). Big data: A review. *2013 International Conference on Collaboration Technologies and Systems (CTS)*, 42–47.
<https://doi.org/10.1109/CTS.2013.6567202>
- Sahlgren, M., & Karlgren, J. (2005). Counting lumps in word space: Density as a measure of corpus homogeneity. *International Symposium on String Processing and Information Retrieval*, 151–154.
- Sänger, J., Richthammer, C., Hassan, S., & Pernul, G. (2014). Trust and Big Data: A Roadmap for Research. *2014 25th International Workshop on Database and Expert Systems Applications*, 278–282.
- Sarica, S., Song, B., Low, E., & Luo, J. (2019). Engineering Knowledge Graph for Keyword Discovery in Patent Search. *Proceedings of the Design Society: International Conference on Engineering Design*, 1(1), 2249–2258.
- Siddharth, L., Madhusudanan, N., & Chakrabarti, A. (2020). Toward Automatically Assessing the Novelty of Engineering Design Solutions. *Journal of Computing and Information Science in Engineering*, 20(1).
- Simonton, D. K. (1999). Creativity as blind variation and selective retention: Is the creative process Darwinian? *Psychological Inquiry*, 309–328.
- Singh, J., & Sharan, A. (2016). Relevance feedback-based query expansion model using ranks combining and Word2Vec approach. *IETE Journal of Research*, 62(5), 591–604.
- Son, S., Na, S., Kim, K., & Lee, S. (2014). Collaborative design environment between ECAD and MCAD engineers in high-tech products development. *International Journal of Production Research*, 52(20), 6161–6174.

- Song, Z., & Kusiak, A. (2009). Optimising product configurations with a data-mining approach. *International Journal of Production Research*, 47(7), 1733–1751.
<https://doi.org/10.1080/00207540701644235>
- Stefanov, V., & Tait, J. I. (2011). An introduction to contemporary search technology. In *Current challenges in patent information retrieval* (pp. 45–65). Springer.
- Suominen, A., Toivanen, H., & Seppänen, M. (2017). Firms' knowledge profiles: Mapping patent data with unsupervised learning. *Technological Forecasting and Social Change*, 115, 131–142.
- Tene, O., & Polonetsky, J. (2011). Privacy in the age of big data: A time for big decisions. *Stan. L. Rev. Online*, 64, 63.
- The Economist. (2010). *A special report on managing information: Data, data everywhere*.
- Thimm, G., Lee, S. G., & Ma, Y.-S. (2006). Towards unified modelling of product life-cycles. *Computers in Industry*, 57(4), 331–341.
<https://doi.org/10.1016/j.compind.2005.09.003>
- Trajtenberg, M. (1990). A penny for your quotes: Patent citations and the value of innovations. *The Rand Journal of Economics*, 172–187.
- Trappey, A. J., Chen, P. P., Trappey, C. V., & Ma, L. (2019). A machine learning approach for solar power technology review and patent evolution analysis. *Applied Sciences*, 9(7), 1478.
- Tuarob, S., & Tucker, C. S. (2015). Automated discovery of lead users and latent product features by mining large scale social media networks. *Journal of Mechanical Design*, 137(7), 071402.

- Unger, D. W., & Eppinger, S. D. (2009). *Comparing product development processes and managing risk*.
- Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science*, 342(6157), 468–472.
- Van Horn, D., & Lewis, K. (2015). The use of analytics in the design of sociotechnical products. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing: AI EDAM*, 29(1), 65.
- Venkatraman, N. (1994). IT-enabled business transformation: From automation to business scope redefinition. *Sloan Management Review*, 35, 73–73.
- Walthall, C. J., Devanathan, S., Kisselburgh, L. G., Ramani, K., Hirleman, E. D., & Yang, M. C. (2011). Evaluating wikis as a communicative medium for collaboration within colocated and distributed engineering design teams. *Journal of Mechanical Design*, 133(7), 071001.
- WIPO IP Facts and Figures*. (2019). WIPO.
https://www.wipo.int/edocs/pubdocs/en/wipo_pub_943_2019.pdf
- Wodehouse, A., Grierson, H., Ion, W., Juster, N., Lynn, A., & Stone, A. (2004). *TikiWiki: A tool to support engineering design students in concept generation*. Conference Proceedings of IEPDE04.
- Wu, X., Zhu, X., Wu, G.-Q., & Ding, W. (2014). Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, 26(1), 97–107.
- Xia, F., Wang, W., Bekele, T. M., & Liu, H. (2017). Big Scholarly Data: A Survey. *IEEE Transactions on Big Data*, 3(1), 18–35.

- Xu, Z., Frankwick, G. L., & Ramirez, E. (2016). Effects of big data analytics and traditional marketing analytics on new product success: A knowledge fusion perspective. *Journal of Business Research*, *69*(5), 1562–1566.
- Yan, J., Meng, Y., Lu, L., & Li, L. (2017). Industrial big data in an industry 4.0 environment: Challenges, schemes, and applications for predictive maintenance. *IEEE Access*, *5*, 23484–23491.
- Yang, H., Park, M., Cho, M., Song, M., & Kim, S. (2014). *A system architecture for manufacturing process analysis based on big data and process mining techniques*. 1024–1029.
- Youn, H., Strumsky, D., Bettencourt, L. M., & Lobo, J. (2015). Invention as a combinatorial process: Evidence from US patents. *Journal of The Royal Society Interface*, *12*(106), 20150272.
- Zhan, Y., Tan, K. H., Li, Y., & Tse, Y. K. (2018). Unlocking the power of big data in new product development. *Annals of Operations Research*, *270*(1–2), 577–595.
- Zhang, J., & Fang, X. (2015). On-site usage-data drives industrial robot design improvement. *2015 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 704–709. <https://doi.org/10.1109/ROBIO.2015.7418851>
- Zhang, Y., Ren, S., Liu, Y., & Si, S. (2017). A big data analytics architecture for cleaner manufacturing and maintenance processes of complex products. *Journal of Cleaner Production*, *142*, 626–641.
- Zhong, R. Y., Xu, X., Klotz, E., & Newman, S. T. (2017). Intelligent manufacturing in the context of industry 4.0: A review. *Engineering*, *3*(5), 616–630.

APPENDIX

Appendix A The code

To build our model we used several libraries, which are explained followed by the code.

Sci-kit learn (scikit-learn.org) was used for the autoencoder, using the MLPRegressor class. Gensim (<https://pypi.org/project/gensim/>) was used for the the vector space model, the doc2vec method. Nltk (<https://www.nltk.org/>) was used for language operations, such as the stemmer, lemmatizer, and stopwords. Other python packages such as pandas (for dataframes), numpy (for mathematical operations), matplotlib (for visualization) were also used.

To get the patents, we using the GCP BigQuery platform (<https://console.cloud.google.com>) and run the query shown

```
#SELECT DISTINCT MIN(publication_number), ab.text, MIN(publication_date),
#MIN(ab.language) , MIN(country_code) ,MIN(c.code)
        #FROM `patents-public-data.patents.publications`, UNNEST(cpc) as c,
        #UNNEST(abstract_localized) as ab WHERE (c.code LIKE 'A61F2/%') AND
        ab.language LIKE 'en'
#GROUP BY ab.text
```

We also get the value measure citation count by running this query

```
# SELECT
        # c.publication_number AS publication_number,
        # COUNT(DISTINCT REGEXP_EXTRACT(p.publication_number, r'(.+-
        .+)-')) AS Count
```

```

        # FROM `patents-public-data.patents.publications` AS p, UNNEST(p.citation)
AS c

        # WHERE c.publication_number IN (

        # SELECT publication_number

        # FROM `dataset-aub.list_pa.list_of_pa`

        # )

# GROUP BY c.publication_number

```

Then, we run this code using python

```

import pandas as pd
import numpy
import re
import os
import gensim
import numpy as np
from gensim.models.doc2vec import Doc2Vec, TaggedDocument
from nltk.tokenize import word_tokenize
import nltk
from nltk.corpus import stopwords
from gensim import utils
from nltk.stem.snowball import SnowballStemmer
import matplotlib.pyplot as plt
from nltk.stem.wordnet import WordNetLemmatizer
import plotly.graph_objects as go
import seaborn as sns

def rem_stopwords(inp):
    stop_words = set(stopwords.words('english'))

```



```

tagged_data=[TaggedDocument(words=stem(rem_stopwords(word_tokenize(_d.lower()
))), tags=[str(i)]) for i, _d in enumerate(data['text'])]
model = Doc2Vec(vector_size=100,
                alpha=0.025,
                min_alpha=0.001,
                min_count=2,
                dm =1)
model.build_vocab(tagged_data)

for epoch in tqdm(range(2)):
    model.train(tagged_data,
                total_examples=model.corpus_count,
                epochs=model.epochs)
    # decrease the learning rate
    model.alpha -= 0.0002
    model.min_alpha = model.alpha

df = data.copy()

dt = pd.read_csv('list_cit.csv', skipinitialspace=True)
df.drop_duplicates('publication_number',inplace=True)

print (len(df))
df = df.merge(dt, how='left')

df = df.fillna(0)
df['text'] = df['text'].map(lambda x: stem(rem_stopwords(word_tokenize(x.lower()))))

d2v_model= model

doc2vec_vectors = d2v_model.docvecs.vectors_docs

```

```

from sklearn.neural_network import MLPRegressor

dl = int(len(data)/2)

dfb = pd.DataFrame(columns=['input','predict','index'])
columns = list(dfb)
store = []

for k in range(100):

    auto_encoder = MLPRegressor(hidden_layer_sizes=(
                                600,
                                100,
                                600,
                                ), random_state=53)
    auto_encoder.fit(doc2vec_vectors[:dl + 360*k],doc2vec_vectors[:dl +
360*k])
    predicted_vector = auto_encoder.predict([doc2vec_vectors[dl + 360*k]])
    zp = zip(columns,[doc2vec_vectors[dl + 360*k],predicted_vector,dl +
360*k])
    d = dict(zp)
    store.append(d)
dfb = pd.DataFrame(store)
from scipy.spatial.distance import cosine

cos_list = []
for k in range(len(dfb)):
    cosine_nov = (1 - cosine(dfb.loc[k]['predict'][0], dfb.loc[k]['inp']))
    cos_list.append(cosine_nov)

```



```
dfb['nov'] = cos_list
```

```
df['Count'] /= df.groupby(df.pub_date.astype(str).str[:4]).Count.transform('mean')
```

```
df = df.fillna(0)
```

```
dfb = dfb.merge(df['Count'],how='left', left_on='index', right_index=True, )
```

```
from scipy.interpolate import CubicSpline
```

```
plt.figure(figsize=(7,7))
```

```
plt.scatter(dfb['nov'],dfb['Count'], s=3)
```

```
sns.lmplot(x='nov',y='Count',data=dfb[dfb.columns[-2:]], fit_reg=True, order=3,ci=None,scatter=False,line_kws={'color':'red'}, height=7)
```

```
plt.show()
```

For the validity test

```
size_dv = doc2vec_vectors.shape[0]
```

```
ymax = numpy.amax(doc2vec_vectors,axis=0)
```

```
ymin = numpy.amin(doc2vec_vectors,axis=0)
```

```
outlier_mat = []
```

```
for i in range(5000):
```

```
    tem = []
```

```
    for k in range(ymax.size):
```

```
        gen = ymin[k]
```

```
        while(gen >= ymin[k] and gen <= ymax[k]):
```

```
            gen =numpy.random.uniform(ymin[k]-5,ymax[k]+5)
```

```
        tem.append(gen)
```

```

        outlier_mat.append(tem)

from sklearn.model_selection import train_test_split

data_set = pd.DataFrame(outlier_mat)

data_set['isdiff'] = 1

tmp = pd.DataFrame(doc2vec_vectors)

tmp['isdiff']= 0

data_set = data_set.append(tmp)

data_set = data_set.reset_index(drop=True)

x_only = data_set.drop(columns=['isdiff'])

y = data_set['isdiff']

xtrain, xtest, ytrain, ytest = train_test_split(x_only, y, test_size=0.2)

xtrain = xtrain.to_numpy()

xtest = xtest.to_numpy()

auto_encoder = MLPRegressor(hidden_layer_sizes=(
                                600,
                                100,
                                600,
                                ), random_state=53)

auto_encoder.fit(xtrain, xtrain)

predicted_vector = auto_encoder.predict(xtest)

cos_list = []

from scipy.spatial.distance import cosine

for j in range(len(xtest)):

```

```

cosine_si = (1 - cosine(predicted_vector[j], xtest[j]))

cos_list.append(cosine_si)

dfco = pd.DataFrame({'nmeasure':cos_list,'isdiff':ytest.to_numpy()})

dfco.to_csv('count_wd.csv') # which is then used to do the roc

```

For the forecasting case

```

patenta = ['joint', 'prothesi', 'compris', 'distal', 'compon', 'anchor', 'bone',
'proxim', 'compon', 'anchor', 'bone', 'coupl', 'piec', 'togeth', 'compon', 'form', 'flexion',
'bear', 'around', 'axi', 'togeth', 'compon', 'form', 'rotari', 'bear', 'form', 'pin', 'bear', 'bush',
'around', 'axi', 'orient', 'transvers', 'axi', 'rotari', 'bear', 'compris', 'multilay', 'bear', 'insert',
'slide', 'sleev', 'surround', 'pin', 'support', 'sleev', 'enclos', 'said', 'slide', 'sleev', 'fasten',
'coupl', 'piec', 'mean', 'secur', 'element', 'wherein', 'secur', 'element', 'compris', 'actuat',
'unit', 'within', 'support', 'sleev', 'connect', 'coupl', 'piec', 'ensur', 'tensil', 'strength', 'mean',
'align', 'bore', 'support', 'sleev', 'coupl', 'piec']

```

```

patentb =['temporari', 'diagnost', 'prosthet', 'socket', 'mount', 'system', 'kit',
'includ', 'general', 'circular', 'test', 'mount', 'block', 'defin', 'annular', 'groov', 'four', 'axial',
'cutout', 'extend', 'lower', 'surfac', 'block', 'upper', 'surfac', 'axial', 'cutout', 'least', 'deep',
'annular', 'groov', 'band', 'extend', 'around', 'perimet', 'block', 'provid', 'span', 'axial',
'cutout', 'therebi', 'form', 'void', 'cutout', 'beneath', 'tape', 'block', 'secur', 'prosthet',
'diagnost', 'socket', 'adhes', 'band', 'extend', 'upper', 'edg', 'block', 'onto', 'outer', 'surfac',
'socket', 'cast', 'tape', 'appli', 'socketblock', 'joint', 'extend', 'groov', 'diagnost', 'fit',
'transfer', 'align', 'cast', 'saw', 'pass', 'around', 'joint', 'cast', 'tape', 'adhes', 'upper', 'surfac',
'block', 'round', 'distal', 'end', 'socket', 'gap', 'behind', 'band', 'cutout', 'provid', 'void',
'space', 'saw', 'avoid', 'damag', 'block', 'cast', 'tape', 'sever', 'easili', 'remov']

```

```

vec_new = d2v_model.infer_vector(patenta + patentb)

```

```

print(vec_new)

```

```

patents_ae =auto_encoder.predict([vec_new])

```

`cosine_sim_v1 = (1 - cosine(vec_new, patents_ae))`