Evolution of Psychology Research: NLP Examination of changes in research focus in the past five decades

Tianyue Niu & Yinjiang Xiong

University of Chicago

Background & Research Question

The origin of psychology can be traced back to Greek philosophers in the 400s BC. Plato and Aristotle first debated about the role of nature vs nurture in psychological development (Lewkowicz, 2011). Since then, psychology has went through major transitions that transformed it from a theoretical subject to a well-established natural science with clearly defined sub-disciplines (Stangor & Walinga, 2018). In particular, cognitive psychology and neuroscience have seen major breakthroughs in the past few decades. For the first time, scientists uncovered details of our memory, perception, neural network, and brain structures (Westen & Gabbard, 2002; Pally 1997) that casted wide impacts in biological and medical sciences. The idea of applying artificial neural networks on machines also inspired brand new computing paradigms that led to the birth of artificial intelligence and deep learning.

The history of psychology does not only provide us with knowledge on how integrative science has developed in the past, but also gives us hints on how research might develop in the future. This project intends to explore the changes in past trends in psychological research. We intend to analyze psychology publications, particularly in neuroscience, from the past five decades in order to extract meaningful insights from them. We believe this is an interesting and meaningful topic to explore. With the insights of the current research, psychologists can gain a better understanding of the most 'trendy' topics in our time, so that they can leverage research resources to focus on certain domains where breakthroughs are most needed (e.g. finding a cure for Alzheimer). Furthermore, the change in focus on psychology can also be reflective of changes in cultural trends. For example, by examining the results, we would be able to answer questions such as 'how has our attitude toward mental health patients changed in the past 50 years?' These questions can have significant cultural and sociological implications.

Given the background, the research question that we aim to answer with this project is 'How has the focus of psychological research changed in the last 50 years?' To answer this question, we combined large scale computing techniques with LDA topic modeling and word embedding analysis to compare changes in research across time from a macro perspective. We then focused on research on Alzheimer and used text generation to explore advancements made in the field over the past 4 decades. We then performed network analysis to evaluate breakthroughs in Alzheimer in the past 5 years in greater details.

Method

Data

This project uses publication from MEDLINE as the basis of all analysis. The data is readily accessible from the PubMed baseline file[1] by any researcher. The entire PubMed baseline contains about 30 million citations. These publications are chronologically organized into 1015 different xml files. We downloaded and parsed every eighth file (127 files in total) and filtered articles in the file by three keywords — 'psyc', 'neur', and 'alzheimer' — to only keep psychology and neuroscience related articles that contained at least one of these keywords in their abstracts as our data. We included 'Alzheimer' because it is a neuroscience topic that has received wide attention from the public in recent years, so we wish to explore the advancement in Alzheimer-related research in our time in more details. Other than filtering abstracts by keywords, I also removed publications that had an empty abstract column. After processing the raw files as described, 299315 articles published in between 1927 and 2020 remained in our final data. Among these articles, more than 99% of them were published after 1970.

Out of all the articles included in our final data, 128812 articles (43.0%) were published in the United States, followed by 63609 articles (21.1%) published in England, 24198 articles(8.1%) in Netherlands, 17083 articles (5.7%) in Germany and 11024 articles (3.7%) in Switzerland. Each row in the data has eight columns: title of the article (*title*), article's abstract (*abstract*), authors of the article (*authors*), publication year (*pubdate*), keywords associated with the article (*keywords*), publication's digital object identifier (*doi*), country of publication (*country*), and publication journal (*journal*).

Large Scale Implementation

To examine past changes in psychology research on a macro level, large-scale computing techniques were applied to perform topic modeling and word-embedding analysis. This part of the analysis can be divided into three steps: 1) downloading & processing data, 2) LDA topic modeling analysis, and 3) word-embedding analysis.

---

[1] PubMed Annual Baseline: ftp://ftp.ncbi.nlm.nih.gov/pubmed/baseline

First, in order to parse and process the data without downloading all of the files to my local computer, I launched a t2.2xlarge EC2 instance on AWS to perform basic data wrangling as described in the data section. I then stored the cleaned data file in an Amazon S3 bucket for future access.

Given the magnitude of the data, it would take really long to train LDA and word embedding models on our local machines. PySpark is an ideal solution to this limitation because it uses Resilient Distributed Datasets (RDD) to perform parallel computing. This can help us speed up the modeling process significantly. Therefore, we chose to perform the second and third steps on an Amazon EMR notebook (8 m5xlarge) using PySpark. In both steps, data was accessed from our S3 bucket, and further divided into 6 different time periods: before 1970, 1970-80, 1980-90, 1990-2000, 2000-10, and after 2010. We then performed Latent Dirichlet Allocation (LDA, a statistical method that attempt to classify a document into k different topics composed of words) topic modeling on each period's data (set k = 8 based on preliminary analysis on my own computer) using the pyspark.ml package. Topic modeling results were then visualized by word clouds. The size of each term in a cloud was determined by its LDA weight. Word embedding analysis was also performed using pyspark.ml (Word2Vec). Each period's data was used to train a separate embedding model. Synonyms for 'mental', 'abnormal' and 'alzheimer' were extracted from each model. The synonyms extracted from different periods were then compared to evaluate differences in our interpretations of these terms across time.

Network Analysis

For the network analysis, only data from 2015 to 2019 is examined to evaluate yearly evolution of research on Alzheimer disease. All abstracts containing the word "Alzheimer" are extracted. POS tags are then assigned to all words using natural language toolkit (nltk). The reason for POS tag assignment is to filter out all the non-noun words. The nouns are counted: only those with counts more than 270 and not included in the manually created stop words list (words like "control", "group", "function", "method", etc.) are created as nodes. The reason to set the threshold at 270 is to only graph the top 10-20 important nouns that co-occur with "Alzheimer". Edge weights are calculated by counting the total occurrence counts of both nodes. For example, if an abstract contains 2 "Alzheimer" and 1 "brain", then the edge weights between "Alzheimer" and "brain" will add 2 and this is repeated for all abstracts. Shorter edges between the nodes indicate larger edge weights (more co-occurrences). Network graphs are produced by the NetworkX package.

Text Generation

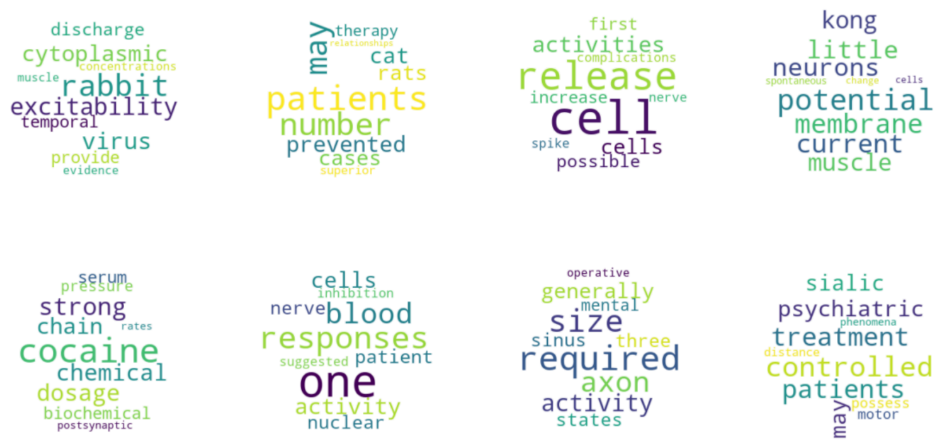
For the text generation part, GPT-2, a pre-trained deep learning model developed by OpenAI, is retuned with context-specific data from 1989, 1999, 2009 and 2019. The fine-tuned models are then used to generate texts.

Abstracts from each year are randomly split into 80% for training and 20% for validation. The sample size gets larger as the year becomes recent. 2005, 5378, 9287 and 18237 abstracts are trained for 1989, 1999, 2009 and 2019 respectively. Two prompt sequences are then fed to the models. The first sequence is “Alzheimer is a disease”, which is context specific and aims to summarize views on Alzheimer in different decades. The other sequence is “Rats are used”, which is less context specific and used to evaluate the effect of fine tuning.

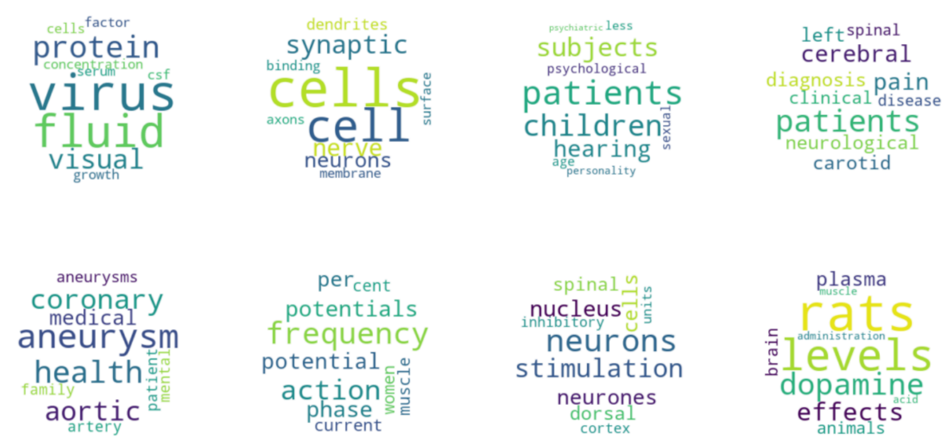
Results & Discussion

Topic Modeling

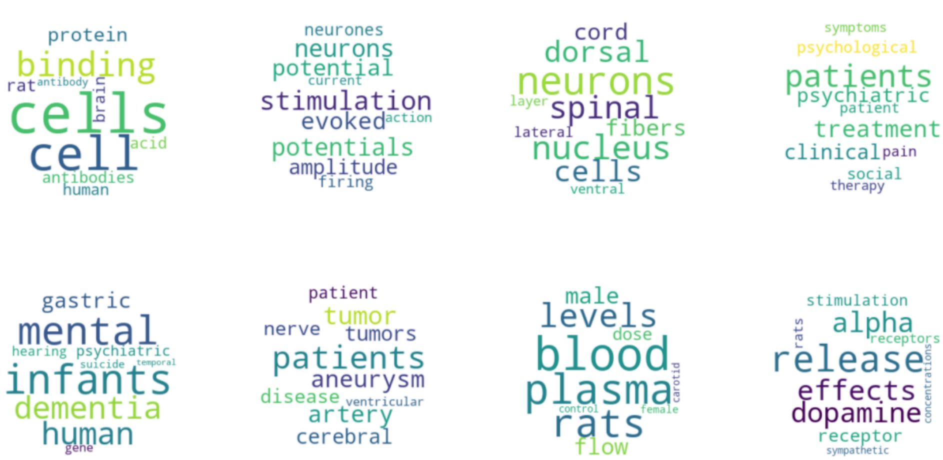
Word Clouds Based on LDA Weights Before 1970



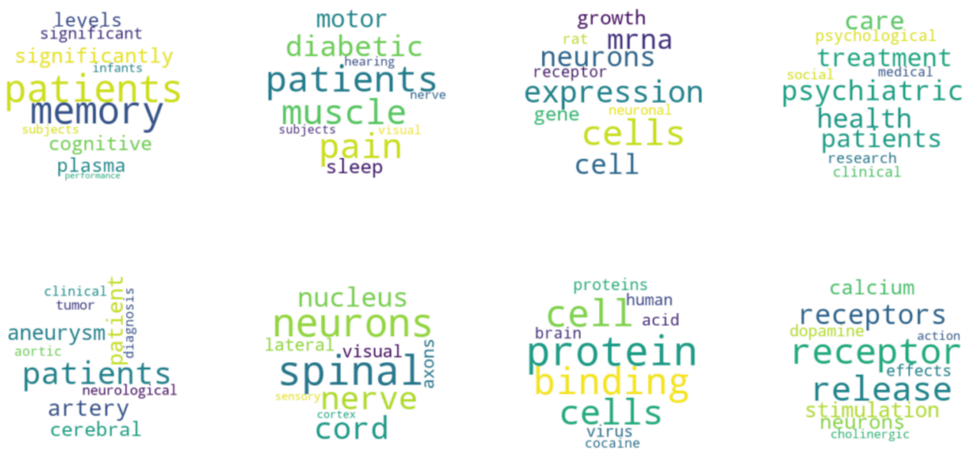
Word Clouds Based on LDA Weights 1970 - 1980



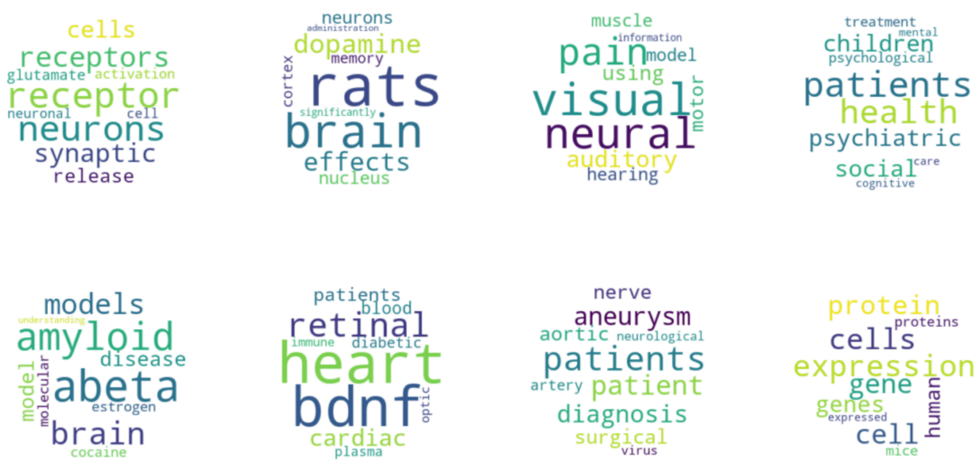
Word Clouds Based on LDA Weights 1980 - 1990



Word Clouds Based on LDA Weights 1990 - 2000



Word Clouds Based on LDA Weights 2000 - 2010



Word Clouds Based on LDA Weights After 2010

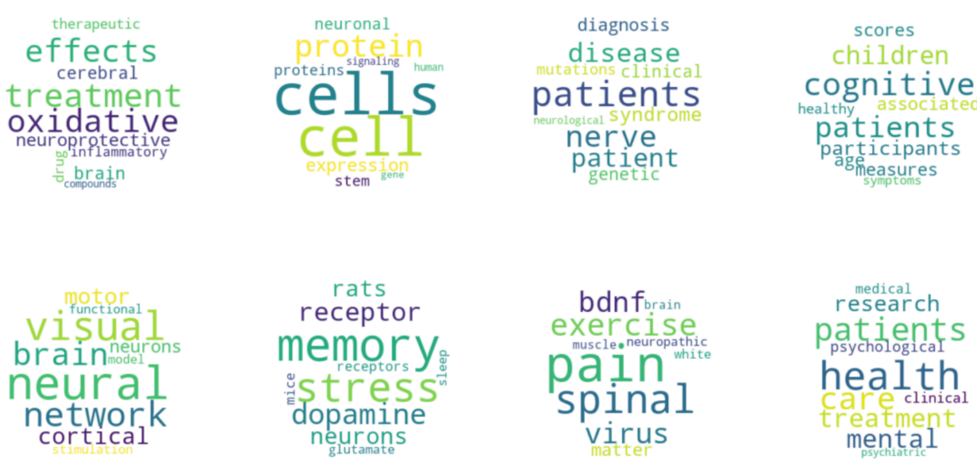


Figure 1. word clouds by LDA weights of terms from different topic models (by time period)

As described in the method section, results generated from each period's LDA topic modeling analysis are presented by word clouds shown in Fig 1. The size of each term is dependent on its LDA weight: higher LDA weight means greater influence over a topic. As we can see from Fig 1, progressive advances were made in the past few decades. Before 1970, scientists were interested in evaluating cell structure and rabbit-related experiments, whereas in 1970-1980, we see the research focus shifted to include more advanced topics such as neurotransmitters in rats (including words such as 'dopamine', 'effects', 'rats') and action potential ('action', 'potentials', 'frequency', 'phase', 'current'). Fast-forward to after 2010, scientists are now focused on topics such as brain disease ('brain', 'cerebral', 'oxidative', 'inflammatory'), visual neural network ('visual', 'neural', 'cortical', 'network', 'model'), memory ('memory', 'stress', 'glutamate', 'sleep') and child cognition ('scores', 'children', 'healthy', 'age', 'cognitive'). These results capture the macro-changes in research focus.

Closer examination of these results show that several topics were consistently present across decades, such as the study of neurotransmitters, neuron firing, psychiatry, and heart disease. Except for heart disease, the other topics are indeed the most important foundations of neuroscience in modern time. We will elaborate on each of these topic's development in details to examine the accuracy of our model.

In 1970, Julius Axelrod, Bernard Katz and Ulf von Euler were awarded the Nobel Prize for their work on neurotransmitters (Shafrir, 1994). We do see in our model that in 1970, one of topics extracted is related with neurotransmitters in rats ('dopamine', 'rat', 'level'). In 1980, a more clearly defined topic for neurotransmitter appeared that included words such as 'release', 'receptor', and 'simulation', as well as 'dopamine' and 'rats'. With the 1990-2000 model, a new word appeared in the the neurotransmitter topic cluster—'calcium'. It was exactly in 1991 that Erwin Neher and Bert Sakmann were awarded the Nobel Prize for their work on single ion channels (Neher & Sakmann, 1991), in which calcium ions were one of the main ions studied. While Nobel Prize studies do not necessarily represent the exact research focus of that decade, they are definitely accurate benchmark for general research trends. These evidences support the validity of our model in that the topics extracted accurately matched with real world research development across time.

Another consistently present topic was neuron firing. In 1973, Timothy and Bliss published their work on long-lasting potentiation of synaptic transmission (Bliss& Lømo, 1973), which proposed that regular activation can lead to persistent strengthening of synapses. This study serves as a foundation for memory and learning. In our model, the 7080 model indeed extracted action potential as one of the main topics. That cluster included words such as ‘potential’, ‘frequency’, ‘phase’. In the 1980s, this cluster evolved to include more targeted terms like ‘amplitude’ and ‘firing’. In the 1990s, action potential was no longer seen as a separate topic by our model, however, the term ‘memory’ appeared in one of the topic with high LDA weight. This probably suggest that researchers had since then shifted their attention from studying the underlying mechanism of long term potentiation (LTP) to evaluating the applications of LTP in fields such as memory and learning. In terms of capturing the development of research on neuron firing, our models were again able to capture major breakthroughs in the field with great accuracy.

We also see in Fig 1 that the term ‘psychiatric’ always appeared as one out of the eight topics extracted from each time period. ‘Treatment’ , ’patients’ and ‘psychological’ were also consistently present in the same topic with ‘psychiatric’. However, the other terms vary slightly across time. Before 1970, terms included in the psychiatric cluster are ‘sialic’, ‘controlled’, and ‘distance’. Between 1970 and 1980, ‘psychiatric’ had a much smaller weight in a seemingly mixed topic with terms such as ‘subjects’, ‘children’, ‘sexual’, ‘hearing’, ‘age’ and ‘personality’. Between 1980 and 1990, the weight of ‘psychiatric’ in the topic became significant again. Terms included in the same topic were ‘symptoms’,‘clinical’, ‘pain’, ‘social’ and ‘therapy’. After the 90s, however, terms like ‘care’, ‘medical’, ‘clinical’, ’social’, ‘research’ and ‘health’ became consistently present in the topic. After 2000, the term ‘mental’ also become included in the topic. These changes highlight different phases in the development of psychiatry over the past five decades. For example, in recent years, psychiatrists are increasingly aware of the variety of environmental reasons that can cause mental illness. The focus of psychiatry has therefore shifted from finding an innate cause for mental illness to providing proper care for mentally impaired patients (Kleinman, 2008). This probably explains why terms like ‘clinical’, ‘social’ and ‘care’ became consistently present in the ‘psychiatric’ topic.

Lastly, from 1970, we start to see a topic composed of heart and artery-related words (‘coronary’, ‘medical’, ‘aneurysm’, ‘aortic’, ‘artery’). This topic is consistently present as a distinct topic from 1970 to 2000. In 2000-2010, it became two separate but related topics, with one more focused on heart problems (‘heart’, ‘cardiac’, ‘diabetic’) and the other more focused on aortic problems (‘aneurysm’, ‘aortic', 'artery', 'surgical'). These changes accurately match the rise and fall of coronary heart disease (Jones and Greene, 2013). Until now, heart-related issue remains the leading cause of death in the United States. This explains why it persistently appears as a distinct topic in different models. In the early 2000s, obesity researchers started to link the increase in obesity and diabetes with heart disease (Ludwig & Ebbeling, 2001; Olshansky et al., 2005), which is also reflected in our topic models: in our 0010 model, the 'heart' topic included words like 'diabetic'. Although heart disease is not directly related with psychology (these articles probably got included in the data because their abstracts included words with the word stem 'neur'), the accurate match between these topics and real world heart-related research development support the effectiveness of our model.

Word Embedding Analysis

Synonyms for 'mental', 'abnormal' and 'alzheimer' were fetched from each period's word embedding model. Top 10 synonyms for each term from different periods are shown in Table 1. From the results, we see that each decade's model returned different terms, and these synonyms become increasingly close to our current interpretation of the original terms as we get closer to 2020.

	mental	alzheimer	abnormal
Before 1970	psychotic 0.989485 geniculate 0.970594 three 0.969684 since 0.967741 treated 0.963939 psychiatric 0.945798 members 0.943704 glial 0.943315 free 0.939168 services 0.938498	NOT IN VOCABULARY	NOT IN VOCABULARY
1970 - 1980	applicable 0.992386 care 0.990944 resolve 0.990482 states 0.990316 asked 0.990215 fluoroscopy 0.988398 needed 0.987659 setting 0.987610 relapse 0.986823 today 0.986600	alkaline 0.997977 desirable 0.991022 added 0.988666 glyoxylic 0.986688 elimination 0.984513 efficiently 0.983935 oxidation 0.983633 suspension 0.982340 periodate 0.981462 equilibrium 0.981461	last 0.994673 preferred 0.989865 myopia 0.986158 divided 0.984945 fit 0.983737 regressed 0.983211 unknown 0.982618 orthopaedic 0.982572 hypothyroidism 0.982553 words 0.982152

1980 - 1990	cost 0.997516	annelid 0.995656	asystole 0.987548
	situational 0.997437	cmv 0.992732	irradiation 0.985809
	qualified 0.997254	organize 0.989907	later 0.980812
	maladjustment 0.996797	neuropathologically 0.988815	laser 0.979930
	alcoholism 0.996228	hepatitis 0.983152	meningomyelocele 0.977162
	people 0.996172	reversion 0.975064	another 0.973331
	disorder 0.995956	semliki 0.974556	fifth 0.971242
	word 0.995616	abandonment 0.963284	noted 0.971085
	manage 0.995438	leukaemia 0.962314	celiac 0.971066
	recruited 0.995340	maturity 0.962280	stage 0.970795
1990 - 2000	advisors 0.999361	newcastle 0.989006	stainless 0.994486
	financial 0.999260	scl 0.978717	cytology 0.991041
	nurses 0.998518	screened 0.977543	fruste 0.990458
	offering 0.997657	shore 0.972582	vater 0.990200
	informed 0.997073	promyelocytic 0.970696	video 0.985222
	vignettes 0.996838	leukodystrophy 0.968767	late 0.982006
	consent 0.996311	endorse 0.967614	subclassified 0.981908
	duties 0.996205	rtt 0.965389	hemiballismus 0.979936
	calls 0.996186	multigenerational 0.961682	guarantees 0.979609
	applicants 0.996122	knows 0.958546	extraforaminal 0.975690
2000 - 2010	welfare 0.998368	autoimmune 0.994168	ultrafine 0.987578
	alliance 0.998236	provisionally 0.984787	precocious 0.987300
	supportive 0.998204	vsp 0.982683	innocua 0.985736
	announcement 0.996551	infectious 0.982446	extrapolating 0.983501
	dispositions 0.996072	progression 0.981692	allegedly 0.982127
	objectify 0.995649	millennium 0.980344	der 0.979466
	unmet 0.995394	recognized 0.976911	centromeric 0.979247
	safety 0.994998	disease 0.976163	embryologic 0.978305
	caregiving 0.994920	pejorative 0.975901	digestive 0.978137
	community 0.994829	diseases 0.975734	mucopolysaccharidosis 0.976334
After 2010	stave 0.994489	neurometabolic 0.993835	quinolones 0.998634
	health 0.994171	microbiologically 0.993593	acylated 0.998501
	psychosocial 0.993910	fragile 0.993417	pentapeptide 0.998122
	lifestyles 0.992894	foodborne 0.987663	congophilic 0.997481
	lebanon 0.992530	rett 0.984513	ascs 0.996644
	narrated 0.991415	disease 0.982809	myoinhibitory 0.992948
	psrfs 0.991413	wolfram 0.981952	glycosylated 0.992174
	rita 0.991092	affecting 0.980769	heptapeptide 0.992013
	bereavement 0.989977	fibrinolytic 0.980083	dodecyl 0.990582
	affordability 0.988072	encephalopathies 0.978776	vasoconstrictive 0.990035

Table 1. Top 10 synonyms from word embedding analysis

For example, before 1970, top synonym for 'mental' was 'psychotic', which carries a negative connotation that almost resembles 'madness'. This matches with the general public's understanding of mental illness at that time. In early 20th century, Kraepelin, a pioneer German psychiatrist, connected degeneration theory with his psychiatry practice. The main message conveyed was that mental illness is innate, and that it can degenerates from generation to generation (Hoff, 2008). This exerted wide and negative impact on how people view mental illness (Zubin, Oppenheimer & Neugebauer, 1985), which might help to explain why our before70 model identified 'psychotic' as the top synonym for 'mental'. From 1980 to 2000, we see terms such as 'financial', 'cost', and 'qualified' appear as synonyms for 'mental', which probably reflect the high expenses associated with mental illness. In fact, in 1980, mental illness was indeed the third most expensive class of disorders in the United States (Trevino & Moss, 1983). This detail associated with the cost of mental illness at this time period was again accurately captured by our word embedding models. After 2010, top synonyms for 'mental' became 'psychosocial', 'lifestyles', and 'bereavement'. These words match closer with our current understanding of mental health related issues. In the past two decades, psychiatrists have indeed started to recognize the social cultural environment as one of the main causes for mental diseases like depressive disorder (Mirowsky & Ross, 2003). Furthermore, these synonyms extracted also reflect development in research. For example, in 2013, the 'bereavement' exclusion in the diagnosis of depression was removed in DSM-5 (Pies, 2014). This was an extremely controversial change. The word 'bereavement' was probably mentioned so frequently in research with mental health related issues that it was picked up as a synonym for 'mental' by the after10 word embedding model. These evidences show that our word embedding models are reflective of cultural trends, which can be further explored in more granularity to reveal deeper implications in future work.

Other than development in mental health research, research on 'alzheimer' has also made huge progress over the past few decades. Before 1970, the word 'alzheimer' was not even found in the vocabulary. In between 1900 and 2000, however, researchers had begun to understand Alzheimer as 'multigenerational'. After 2010, we see terms such as 'fragile', 'rett', 'wolfram' , 'microbiologically' and 'neurometabolic' at the top of the synonym list, which showcases that researchers' now see Alzheimer as a neurodegenerative disease that is similar to the Rett Syndrome and the Wolfram Syndrome. The Rett syndrome is a genetic neurological disorder (almost exclusively found in girls) that can cause severe impairments in patients' ability to walk, eat and speak in everyday life (Hagberg et al, 1985). The Wolfram Sydnrome on the other hand is also a rare genetic disorder that can lead to severe consequences such as deafness and blindness (Barrett, Bundey & Macloed, 1995). As Rett, Wolfram and Alzheimer are all genetic disorders, the fact that our after10 embedding model can identify Rett and Wolfram as Alzheimer's synonyms validates the effectiveness of our model. One reason that might explain the high level of accuracy exhibited by the after10 embedding model may be advancement in research. However, another explanation is that, when compared with the other periods' data, after10 had the largest number of publications. If this is is true, it might imply that we should attempt to sample more data in the future in order to obtain more accurate results from large scale content analysis.

We also extracted synonyms for 'abnormal' in our word embedding analysis. When we first decided to include the term 'abnormal', we were hoping to find synonyms associated with abnormal psychology. However, as shown by the result in Table 1, 'abnormal' is clearly used more often in contexts other than abnormal psychology. In fact, 'abnormal' was probably used in so many different contexts to the extent that our word embedding models were unable to accurately extract its synonyms. For example, in the 7080 model, top synonyms of 'abnormal' were 'last', 'preferred', 'myopia', whereas in the 8090 model, top synonyms found were 'asystole', 'irradiation', 'later', and 'laser'. The after10 model, which is also the model trained with the largest data, returned 'quinolones', 'acylated', 'pentapeptide', 'congophilic' and 'ascs' as 'abnormal's top synonyms. With some research, quinolones are chemotherapeutic bactericidal drugs, acylated is a chemical process, pentapeptide is a type of polypeptide, congophilic is an adjective used to describe tissue whereas 'ascs' is probably an abbreviation for abnormal squamous cells. Except for ASCs, the other terms do not seem to be directly associated with 'abnormal'. These results demonstrate the limit of our word embedding model: for terms that are not so frequently used, but are at the same times used in very different linguistic contexts, word embedding model would not be able to accurately capture its synonyms.

Network Analysis

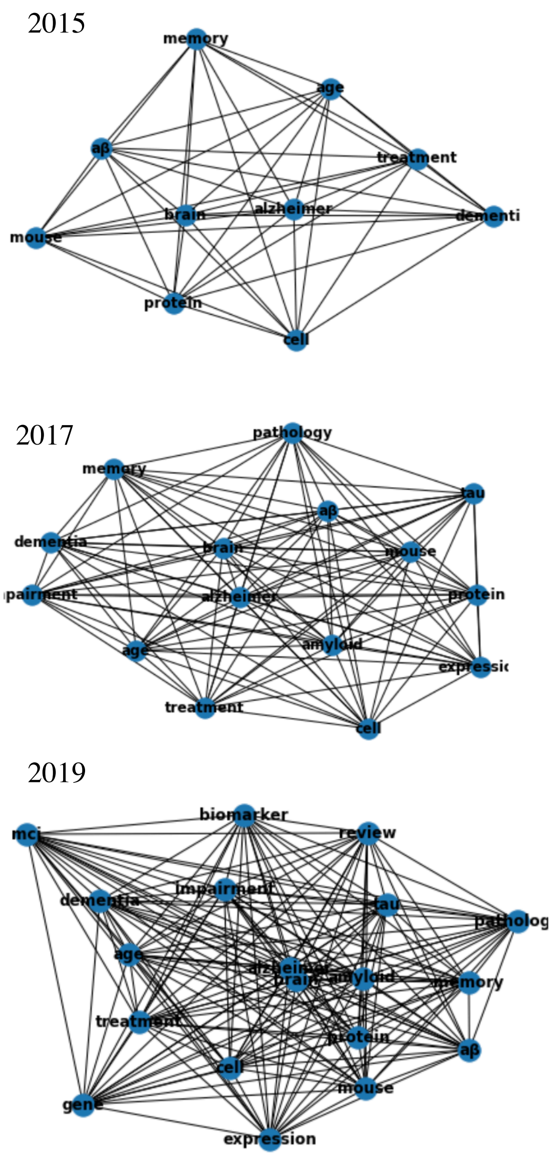


Figure 2. Network analysis for 'Alzheimer'

Network analysis has a deeply rooted history in sociology. As we maneuver around the world, an invisible mesh of connections is formed and knitted together for each of us. As early as 1930s, Moreno developed the so-called "sociometry" in order to capture friendship patterns and informal interactions. Bavelas and Festinger built upon the concept of "sociometry" and significantly extended the study of "group dynamics" (Scott, 1988). The structural networks, which can be most intuitively represented by human interactions, is not confined to human social relationships. Gravitational attraction interacts between planets in astrophysics; atoms interact to form molecules in molecular chemistry; and capacitors and resistors interact to influence circuit in electrical engineering (Freeman, 2004). We rarely find any subject in our life that is not composed of numerous visible or invisible webs of interactions. Language is one of the subjects that contains rich interactions. In English, for example, multiple words at different positions belonging to different parts of speech (POS) construct the intrinsic logic of a sentence. In the current research, we selected a specific topic, Alzheimer's disease, to explore its evolution of network from 2015 to 2019.

As figure 2 suggests, abstracts from 2015 to 2019 have an increasing number of words that meet the threshold. This indicates a growing interest in the research related to Alzheimer disease. Specifically, the word "tau" does not appear in the 2015 graph and locates at the far ends of the graphs from 2016 to 2018, but moves considerably closer to "Alzheimer" in 2019. According to the research development of Alzheimer disease, researchers in 2016 published an article named "A Novel Alzheimer Disease Locus Located near the Gene Encoding Tau Protein" on Nature. This paper revealed association between microtubule-associated protein tau (MAPT) and early onset of Alzheimer disease, which sparked a large volume of following research endeavors (Jun et al., 2016). In December 2018, another paper on Nature titled "Tau Impairs Neural Circuits, Dominating Amyloid-β effects, in Alzheimer Models in Vivo" enlightened the underlying mechanisms between Amyloid-β and tau and suggested a larger role of tau than previously expected (Busche et al., 2019). The greater influence of tau is reflected on its relative position to "Alzheimer" and "aβ" on the graph of 2019. Tau is closer to both of the words in 2019 than the previous years. "Biomarker" and "mci" made their first appearance on 2019's graph. As the research on the causes and indicators of Alzheimer disease develops, there have been an increasing volume of discussions on detecting biomarkers for Alzheimer disease. This marks a transition from grasping the fundamental causes of Alzheimer disease to utilizing the risk factors we know (amyloid-β, tau) to predict, control and eventually cure the disease. Moreover, mild cognitive impairment (mci) has been frequently mentioned with biomarkers and Alzheimer disease, suggesting that the research focus of Alzheimer disease is expanding and generalizing.

Text Generation

```
Original Model
Alzheimer is a disease that affects about 1 in 10 people. It is caused by a
genetic mutation that causes the brain to become more active.
1989
Alzheimer is a disease of the brain that affects the brain's ability to
process information. It is a neurodegenerative disease that affects the
brain's ability to process information.
1999
Alzheimer is a disease of the brain that affects the central nervous system.
It is a neurodegenerative disease of the brain that affects the central
nervous system.
2009
Alzheimer is a disease of the brain that affects the brain's ability to
process information. The disease is associated with a wide range of
neurological and psychiatric disorders, including Alzheimer's disease,
Parkinson's disease, Huntington's disease.
2019
Alzheimer is a disease of the central nervous system that affects the brain
and spinal cord.

Original Model
Rats are used to kill the most common types of birds, including the black-
footed, the black-tailed, and the black-tailed eagle.
1989
Rats are used in the study of the effects of the various drugs on the nervous
system.
1999
Rats are used to study the effects of a variety of drugs on the immune
system. The purpose of this study was to determine the effects of a variety
of drugs on the immune system of rats.
2009
Rats are used to study the effects of a novel antipsychotic on the
development of schizophrenia. The aim of this study was to determine the
effects of a novel antipsychotic on the development of schizophrenia in rats.
2019
Rats are used to study the effects of a variety of drugs on the brain. The
aim of this study was to investigate the effects of a novel drug, the
selective serotonin reuptake inhibitor (SSRI), on the brain of rats.
```

Machine-generated text has been confined by its resemblance to human writing styles in

Machine-generated text has been confined by its lack of resemblance to human writing styles in, so most of its application have been in email auto-completion, chatbot and other short, well-structured contexts. With the development of computing techniques and neural models, however, natural language generation is making a breakthrough to greatly mimic different human writing styles and produce plausible content.

GPT-2, launched by OpenAI in 2019, is capable of generating human-like text in any context without task-specific training. According to OpenAI, the model is “chameleon-like”, as it can adapt to the writing style and content of the input. Its outstanding performance and flexibility have raised concern for potential misuse of the model. OpenAI team believes that the full model can be used to generate false information, impersonate people, produce phishing content and fulfill other malicious purposes (Radford et al., 2019).

In the current research, we aim to use the pre-trained GPT-2 model and fine-tune it with abstracts in 1989, 1999, 2009 and 2019. The goal is to extract developmental insights of specific topics by comparing the generated texts. We used two sequences, “Alzheimer is a disease” and “Rats are used”, as input. First, we compare the output between the original pre-trained models and the fine-tuned models. For the input “Alzheimer is a disease”, the original output gives 3 pieces of information: some statistics for the commonality of Alzheimer disease, its cause and its effect. This output seems to greatly resemble common knowledge. On the other hand, for the context-trained output, the use of words is more professional and specific. As for the more general input “Rats are used”, the original model has outputted content that is completely unrelated to neuroscience and psychology, whereas the pre-trained models have all identified rats as experiment subjects for neuroscience studies. Second, we analyze whether there exists any insightful trend for the fine-tuned models trained in 4 different years. We found that the text moves from general to specific information. For example, the model using 1989 data describes Alzheimer disease as “affects the brain’s ability to process information” and the later models add domain-specific nouns such as central nervous system, associated disorders and so on. Similar pattern can be found for the “Rats are used” sequence. From the 1989 model, the output only generally describes the use of rats in drug studies. The 1999 model specifies that it is for the immune system, the 2009 model identifies the drug as a new antipsychotic for schizophrenia and the 2019 model made clear the name of the drug. This trend suggests that neuroscience research has been developing to uncover more granular details on general concepts.

Conclusion

To conclude, our results suggest that psychology research trends had changed dramatically in the past 50 years. Scientists have made huge progresses in multiple areas. For example, In terms of research related to neuron, we see from our topic modeling analysis that scientists have went from exploring basic structure of neurons to examining how neurons interact with each other via neurotransmitters, as well as how neurons fire together to form long-term potentiation. In the past 20 years, scientists have also delved deeper into the field of neural network and memory. These progresses were accurately captured by our LDA models, which proved the effectiveness of our method.

In terms of specific topics such as mental health and Alzheimer, we have also seen great advancements made. According to our word embedding models, before 1970, researchers tended to view 'mental' as similar to 'psychotic', which degrades mental illness as similar to madness. In between 1980 and 2000, however, the expensive nature of mental illness treatment became captured by our topic modeling model. After 2010, 'mental's synonyms became 'bereavement', 'care', and 'research', which are closer to our current understanding of mental health. These changes match with real development in how our understanding of mental illness has changed over the past 50 years. These results support the validity of our model, and showed that word embedding analysis was effective in capturing changes in our understanding of specific topics across time.

In general, large-scale content analysis techniques presented in this project show great effectiveness in exploring past research trends as well as term-related cultural trends. The same methods can be applied to other large data sets to examine topic changes across time on a macro-level and evaluate changes in our interpretation for specific terms in greater details. To extend on this project, future work can sample the entire PubMed baseline and divide the past few decades into shorter time periods to examine more nuanced changes in past research trends.

Though only on a 5-year span, the network analysis centering around the topic of Alzheimer disease reflects significant development in research. A growing body of literature is indicated by the increasing number of nodes included in the graph. The evolving scientific attention on tau is also embodied by the appearance as a node since 2016 and increasingly close proximity to the center word, "Alzheimer". We also observed a potential transition starting in 2019 to utilizing biomarkers such as amyloid-β and tau for prediction and treatment of Alzheimer and other mild cognitive impairment. The effectiveness of our network analysis can be attributed to our micro-level focus on Alzheimer disease over a relatively short amount of time. To expand on the current research, future endeavors can attempt at generating graphs for a longer period on various topics.

Text generation produces insightful results as an uncommon technique in NLP analysis. We are able to observe that fine-tuned models can reflect the data on which they are trained. We conclude that neuroscience research has gradually deepened its understanding in Alzheimer disease. This technique can be deployed for qualitative presentations for future research.

References

Barrett, T. G., Bundey, S. E., & Macleod, A. F. (1995). Neurodegeneration and diabetes: UK nationwide study of Wolfram (DIDMOAD) syndrome. *The Lancet*, *346*(8988), 1458-1463.

Busche, Marc Aurel, et al. "Tau impairs neural circuits, dominating amyloid-β effects, in Alzheimer models in vivo." Nature neuroscience 22.1 (2019): 57-64.

Freeman, Linton. "The development of social network analysis." A Study in the Sociology of Science 1 (2004): 687.

Hagberg, B., Goutières, F., Hanefeld, F., Rett, A., & Wilson, J. (1985). Rett syndrome: criteria for inclusion and exclusion. *Brain & development*, *7*(3), 372-373.

Hoff, P. (2008). Kraepelin and degeneration theory. European archives of psychiatry and clinical neuroscience, 258(2), 12.

Jones, D. S., & Greene, J. A. (2013). The decline and rise of coronary heart disease: understanding public health catastrophism. *American journal of public health*, *103*(7), 1207-1218.

Jun, Gyungah, et al. "A novel Alzheimer disease locus located near the gene encoding tau protein." Molecular psychiatry 21.1 (2016): 108-11

Kleinman, A. (2008). *Rethinking psychiatry*. Simon and Schuster.

Ludwig, D. S., & Ebbeling, C. B. (2001). Type 2 diabetes mellitus in children: primary care and public health considerations. *Jama*, *286*(12), 1427-1430.

Mirowsky, J., & Ross, C. E. (2003). Social causes of psychological distress. Transaction Publishers.

Neher, E., & Sakmann, B. (1991). The Nobel Prize in Physiology or Medicine 1991.

Olshansky, S. J., Passaro, D. J., Hershow, R. C., Layden, J., Carnes, B. A., Brody, J., ... & Ludwig, D. S. (2005). A potential decline in life expectancy in the United States in the 21st century. *New England Journal of Medicine*, *352*(11), 1138-1145.

Radford, Alec, et al. "Better language models and their implications." OpenAI Blog https://openai. com/ blog/better-language-models (2019).

Scott, John. "Social network analysis." Sociology 22.1 (1988): 109-127.

Shafrir, E. (1994). Julius Axelrod, Bernard Katz and Ulf von Euler--Nobel Prize winners for the discovery of mechanisms of nerve signal transmission. *Israel journal of medical sciences*, *30*(11), 869.

Stangor, C., & Walinga, J. (2018). Introduction to Psychology-1st Canadian Edition.’

Trevino, F. M., & Moss, A. J. (1983). Health, United States, 1983. *Department of Health and Human*.

Pally, R. (1997). Developments in neuroscience. II: how the brain actively constructs perceptions. *The International journal of psycho-analysis*, *78*(5), 1021.

Pies, R. (2014). The bereavement exclusion and DSM-5: an update and commentary. *Innovations in clinical neuroscience*, *11*(7-8), 19.

Westen, D., & Gabbard, G. O. (2002). Developments in cognitive neuroscience: I. Conflict, compromise, and connectionism. *Journal of the American Psychoanalytic Association*, *50*(1), 53-98.

Zubin, J., Oppenheimer, G., & Neugebauer, R. (1985). Degeneration theory and the stigma of schizophrenia. *Biological Psychiatry*, *20*(11), 1145-1148.