The Progress and Prospective of Data Capital for Zero-Shot Deep Brain-Computer Interfaces

Wenbao Ma *, Teng Ma, Daniel Organisciak, Jude E.T. Waide, Xiangxi Meng, Yang Long

Posted Date: 5 December 2024

doi: 10.20944/preprints202412.0390.v1

Keywords: Brain-Computer Interfaces; Deep Learning; Zero-Shot Learning; Industrial Landscape; Conceptualisation

 doi:10.20944/preprints202412.0390.v1

Article

The Progress and Prospective of Data Capital for Zero-Shot Deep Brain-Computer Interfaces

Wenbao Ma [1,*,†], Teng Ma [1,†], Daniel Organisciak [2], Jude E.T. Waide [3], Xiangxi Meng [3] and Yang Long [3]

[1] School of Humanities and Social Science, Xi'an Jiaotong University, Xi'an 710049, China; matengsax@gmail.com
[2] Department of Computer and Information Sciences, Northumbria University, Newcastle upon Tyne, NE1 8ST, UK; d.organisciak@gmail.com
[3] Department of Computer Science, Durham University, Durham DH1 3LE, UK; jude.waide@durham.ac.uk, xiangxi.meng@durham.ac.uk and yang.long@durham.ac.uk
* Correspondence: smiling-ma@163.com
† These authors contributed equally to this work.

Abstract: The vigorous development of Deep Learning (DL) has been propelled by big data and high-performance computing. For Brain-Computer Interfaces (BCIs) to benefit from DL in a reliable and scalable manner, the scale and quality of data are crucial. Special emphasis is placed on the Zero-Shot Learning (ZSL) paradigm, which is essential for enhancing the flexibility and scalability of BCI systems. ZSL enables models to generalise from limited examples to new, unseen tasks, addressing data scarcity challenges and accelerating the development of robust, adaptable BCIs. Despite a growing number of BCI surveys in recent years, there is a notable gap in clearly presenting public data resources. This paper explores the fundamental data capital necessary for large-scale Deep Learning BCI models (DBCI) models. Our key contributions include: 1) A systematic review and comprehensive understanding of the current industrial landscape of DBCI datasets; 2) An in-depth analysis of research gaps and trends in DBCI devices, data and applications, offering insights into the progress and prospects for high-quality data foundation and developing large-scale DBCI models; 3) A focus on the paradigm shift brought by ZSL, which is pivotal for the technical potential and readiness of BCIs in the era of multi-modal large AI models.

Keywords: brain-computer interfaces; deep learning; zero-shot learning; industrial landscape; conceptualisation

1. Introduction

Neuralink's 'Telepathy' is a new intrusive Brain-Computer Interaction (BCI) device that places 1024 electrodes in the motor cortex. Unlike non-intrusive EEG technologies, intrusive methods are able to fully penetrate the brain for deep coverage. This, combined with Telepathy's high electrode count compared to the standard 32 or 64 electrodes for non-intrusive EEG, provides a quality of data unrivalled by non-intrusive means. Furthermore, a vast quantity of data can be gathered by the device due to being embedded in a user, unlike EEG devices that would typically be worn for no longer than a day, or to MRI machines that people would be in for only a couple of hours. The company has also developed a robot to implant their device, making it more scalable and accessible as it reduces the need for human expertise during the surgery. It is for these reasons that we believe Neuralink's technology could be the beginning of large-scale and high-quality data capital collection for BCI applications.

That being said, this technology is still in its infancy. At the time of writing, only 2 patients have had Telepathy implanted with varying success. Some of the 64 threads came loose with the first patient, although the user was still able to control a cursor with their thoughts. Questions still remain as to the safety of this technology, as well as its long-term durability. There is also the cost of the device to consider, as well as the fears people have about having a chip implanted in their brain, which are further barriers to the wide-spread adoption of this device. For this reason, we focus on non-intrusive EEG datasets in this paper, which contribute to the majority of existing data capital for BCI applications due to their price and ease-of-use compared to other brain-imaging techniques. It is also unclear if Neuralink, due to their nature as a private company, will release the data capital they collect to the broader academic community.

 doi:10.20944/preprints202412.0390.v1

The rapid development of Deep Learning (DL) embraces the prevalence of Internet technology. The increasing internet access and availability accumulate large-scale and diverse data which stimulates the demand for efficient computing and data storage. The past two decades have shown a significant trend from theoretical studies toward versatile applications. New commercial needs encounter technical challenges that in turn motivate the establishment of new theoretical foundations, such as multi-modal [1], multi-task [2], interpretable [3], causality [4] and AI-Generated Contents (AIGC) [5,6]. Putting DL technology in the context of the Industrial Landscape helps to understand the closed loop of theory-application-need and identify future trends, limitations and challenges.

On the path toward Artificial General Intelligence (AGI), models and data are the two fundamental pillars of AI development. The origins of AI models lie in logic formalisation [7]. Alan Turing, building on this foundation, introduced the Turing machine and the conceptualisation of AI from a deductive perspective. Deductive or rule-based symbolic systems [8] are characterised by their pursuit of rigour and precision but often sacrifice flexibility and generalisation capabilities. In contrast, inductive methods have evolved to follow a data-driven approach, deriving rules and models from patterns in observed data. Current deep learning (DL) models are primarily grounded in the inductive paradigm, processing empirical perception signals such as vision, natural language, and audio [9]. However, while these models excel at pattern recognition and representation, they are inherently limited to interpreting conscious levels of data, as conceptualised in cognitive theory [10]. According to this theory, an agent's cognition spans four levels: unconsciousness, consciousness, awareness, and meta-awareness. Traditional DL models predominantly operate at the level of behavioural data, reflecting perception and consciousness. Their goal is to align human attention signals (labels) with input data, such as semantic attributes. Despite their progress, supervised learning paradigms, which have dominated AI research for the past two decades, suffer from several limitations. These include issues such as subjective biases [11], vulnerability to adversarial attacks and data poisoning [12], the burden of data annotation, and significant ethical concerns [13]. These challenges highlight the constraints of existing methods and the need for more robust approaches to achieve AGI. The emergence of self-supervised learning and advances in parallel computing have prompted industrial efforts to pursue a top-down technological approach. This involves leveraging large-scale multi-modal interactive data to train powerful DL models that aim to achieve meta-awareness — a higher cognitive representation incorporating knowledge graphs and causal inference [14]. By addressing the divergence in individual awareness and improving moral generalisation, these models offer a pathway to mitigate bias and ensure more inclusive and fair outcomes. However, this approach relies on the assumption that data collection systems can comprehensively capture diverse users, thereby addressing the challenges of neurodiversity [15].

Emerging BCI technologies have brought new opportunities and challenges which push the AI and deep learning community to the next level. One aim of this technology is to explain fundamental brain mechanisms beyond perception and consciousness. For example, Rapid Serial Visual Presentation (RSVP) displays users with sequential images at high speed (e.g. 10 images per second). In face recognition tasks, users are given a well-known target face to find, e.g. Einstein, before being displayed a high-speed sequence of faces. A promising result is that the P300 signal, triggered when a person recognises a face, can be detected from the BCI signal when human participants are not aware that the face has been displayed. This shows that signals measured by BCI devices can indeed detect and analyse unconscious level information and suggests that BCI technology could lead AI to a new era by exploring the internal behaviour of brain activities beyond existing cognitive and conscious levels.

In this context, Brain-Computer Interface (BCI) systems represent a critical breakthrough. Unlike traditional methods that primarily rely on behavioural and conscious data, BCI systems have the potential to tap into unconscious levels of cognition, providing a fundamentally new layer of supervision for AI. By integrating neural signals directly from the brain, BCI can reduce the inherent subjectivity of supervised learning and provide richer, more diverse data inputs. This capability not only mitigates bias and enhances fairness but also paves the way for a more comprehensive and

 doi:10.20944/preprints202412.0390.v1

accurate alignment of AI systems with human cognition, thereby playing an indispensable role in the pursuit of AGI. However, the key barrier between BCI and contemporary deep learning research is data foundation. The data-hungry nature of deep models requires vast quantities of training data that can only be acquired through large-scale deployment. The polarised situation is that intrusive or fMRI-based data collection can provide high-resolution and reliable results, but are limited by cost and usability. However, the low-cost, lightweight, and commercialised devices, e.g. EEG, ECG, and EMG, are still limited in performance. In this paper, we investigate this problem through the lens of the Industrial Landscape [16], which provides a new perspective on data capital. This allows us to understand the progress and to predict the trend of deep neural network development in BCI domains. The contribution of this paper is threefold:

- First, we use the industrial landscape conceptualisation framework to conduct a systematic literature review. We summarise both established and emerging DBCI data capitals which help understand the progress of each identified core technical milestone of DBCI.
- Second, the motivation of this article aims to put the development of BCI models into the context of the industrial landscape framework. We identify key barriers preventing the development of large DBCI models in terms of devices, data, and applications.
- Third, we point those unaddressed technical challenges towards cutting-edge zero-shot learning techniques. Our findings establish a technical road-map through inter-sample, inter-person, inter-device, inter-domain and inter-task transfer paradigms, multi-modal visual-semantic neural signal models, and data synthesis and signal processing for higher SNR and scalable DBCI device adaption.

The organisation of this paper is as follows. In section 2, we systematically introduce the research background, the conceptualisation of the industrial landscape, and the current state of DBCI research. In section 3, we outline our survey methodology and the data we're going to collect. Section 4 discusses the survey results of existing BCI datasets and suggests how the emerging zero-shot neural decoding technique can overcome the barriers identified in the survey. We finalise our discussion and summarise the main findings in the last section.

2. Research Background

In this section, we introduce the research background of DBCI to put our review into the context of the Industrial Landscape (IL) [16]. The IL provides a framework to analyse the industrial trend of both existing digital technologies and AI. Our contribution focuses on making a mapping for DBCI development under the IL framework so as to understand and predict the progress in parallel with other AI and digital technologies.

The first return of our research on google scholar gives over 28,000 results based on the keywords of brain-computer/machine interface, EEG and review/survey. From the results, we select 677 papers ranging from 1986 to 2023 as our entry points. Through a narrative review approach, we extract key milestone papers to provide an overview of current BCI research as follows.

Early work of BCI can be traced up to 1924 [17] when the first-ever electroencephalogram signal was recorded by Hans Berger. A Bio-Neuro [18] feedback began in the late 1950s. Biofeedback refers to all physiological signals, e.g. blood pressure, heart rate etc. whereas Neuro feedback refers to brain signals only. The first seminal work that provided both a theoretical and technical review of BCI was given in 1973 [19]. Initial research focuses on Controlling Assistive Devices. Operant (Instrumental) conditioning refers to autonomous functions, e.g. blood pressure and heart rate which can be manipulated by operant conditions. In 1960 [20,21], Neil Miller demonstrated the first trial to disrupt the motor system of rats. The experiment extended to blood pressure, urine production, and gut control in [22]. Human learning, in contrast, takes the cognitive dimension into account. Controlling devices with BCI, end-users need to focus their attention throughout the tasks, which is cognitively demanding.

 doi:10.20944/preprints202412.0390.v1

One of the primary applications of BCI lies in the neuro-disorder domain. In particular, Locked-In Syndrome (LIS), which typically follows a stroke in the basilar artery of the brainstem, is characterised by the retention of vertical eye movements (e.g., looking up and down) [23,24]. LIS can also result from Amyotrophic Lateral Sclerosis (ALS), which leads to the loss of movement or complete motor paralysis. Both LIS and ALS are key target populations for restoring lost functionality through BCI. Compared to traditional voluntary assistive technologies, BCI offers four main advantages. First, Slow Cortical Potentials (SCP) provide the basis for long-term training, allowing individuals to communicate messages in the absence of peripheral muscular movement. Second, involuntary eye movements associated with LIS present a significant challenge for other assistive technologies, which BCI can bypass. Third, depression caused by LIS often makes it difficult for caregivers to interpret eye movements or spelling codes, which limits communication. Fourth, BCI eliminates the need for questionnaire-based assessments, providing a more direct and efficient interface. A more challenging scenario involves Complete Locked-In Syndrome (CLIS) [25], in which the loss of behavioural output [26] leads to "thought paralysis." This state, often resembling a vegetative state, poses limitations to operant learning approaches. Despite these challenges, contemporary BCI applications have embraced advancements brought by the AIGC era. For instance, Brain Painting replaces the traditional P300 matrix with icons representing painting tools, which are controlled by a cursor. This technology has enabled ALS patients to create art independently, without requiring researcher supervision [27,28]. Following painting sessions, satisfaction, joy, and frustration are evaluated by the BCI team, and favourable results have consistently been observed.

Neurophysiology has established key paradigms for BCI signal acquisition, such as Slow Cortical Potentials (SCP) and P300, which are widely applied in conditions like epilepsy and ADHD (Attention Deficit Hyperactivity Disorder). Techniques like voluntary control of Alpha Activity, Sensorimotor Rhythms (SMR), and μ-rhythm have been utilised in psychological therapy, behavioural studies, and medicine since the 1950s. Event-Related Potentials (ERP), SMR, SCP, and P300 (a positive potential occurring 300 ms after a stimulus) are frequently implemented with stimuli approaches like the oddball paradigm. For example, a 6x6 letter matrix [29] enables letter selection, while N400 (a negative potential 400 ms after stimulus) is used for face recognition tasks. Operant learning is commonly employed to increase SMR activity (8–15 Hz), which reflects Event-Related Desynchronisation (ERD). ERD was introduced for cursor control in 1991 and later expanded to motor imagery, though it requires users to learn to regulate their brain responses. The S1-S2 paradigm (S1: warning stimulus, S2: imperative stimulus requiring a motor response) is also used, where SCP measures slow EEG shifts, such as Contingent Negative Variation (CNV). For instance, a negative shift 800 ms before finger movement can be observed. SCP shifts are also associated with large negative DC shifts during epileptic seizures, and voluntary SCP modulation may help prevent them. These methods were first implemented for locked-in patients in 1999 [30] and remain foundational for smooth BCI control.

There are several traditional barriers preventing BCI to be widely applied. The first is the Signal-Noise Rate (SNR). SNR reflect the strength of the signal of interest in relation to artefacts like breathing and muscular movement [31]. These noist artifacts remain a fundamental challenge today. Second, BCI training is required for users, decreasing accessibility [32]. In 2010, Usability and User-Centred Design (UCD) [33] set the ISO 9241-210 as the usability standard. This norm requires BCI-controlled applications to be evaluated by user experience in terms of 1) *effectiveness* which considers the accuracy and completeness users can achieve; 2) *satisfaction* which measures comfort and acceptability while using the device; and 3) *usability* measurement. Information Transfer Rate (ITR) is also a key parameter to measure BCI accessibility. From the early work of 2 mins per letter [30], P300-based BCI progressed to 10 letters per min in [34]. However, it is still not suitable for independent home use. Device design is also an important factor. The trend in BCI technology development is moving towards lightweight, cost-effective solutions, ranging from compact RRG amplifiers integrated into caps [35], to artefact rejection techniques for smartphone applications during walking [36], and behind-the-ear designs

 doi:10.20944/preprints202412.0390.v1

[37]. The literature review highlights the essential need for advancements in Machine Learning, Communication, and Interaction technologies [38]. The key objectives are to:

- Reduce the training cost for both users and models
- Robust filters for improved SNR
- Transferable and generalised BCI without prior calibration

This presents a classic chicken-and-egg dilemma. On one hand, machine learning, particularly deep learning, requires large-scale data to achieve reliable transferability and generalisation. On the other hand, transferability and generalisation are essential features that must be established before a BCI device can be widely adopted. For instance, no long-term studies involving locked-in patients have been conducted using machine learning. Historically, neuro-feedback studies required significant time investments, such as 288 hours per user [39], or in 1977, 2.5 years of SMR data collected over 200 sessions. These efforts represent a foundational investment in data capital, which we consider critical for driving progress in models, devices, paradigms, and accessibility. We will further explore how data capital underpins these aspects in the context of the Industrial Landscape framework.

2.1. Industrial Landscape

Industrial Landscape (IL) generally refers to the physical and visual characteristics of areas where industrial activities take place, such as factories, mills, refineries, and other industrial facilities. It can also refer to the broader socio-economic and cultural impacts of industrialisation on the surrounding environment and communities, including changes to land use, infrastructure, and the built environment. Industrial landscapes can vary widely in appearance and character depending on the type of industry, the location, and the historical context, and may include features such as smokestacks, silos, pipelines, and rail yards.

The fast growth of internet companies and new technologies has resulted in a stark contrast to the traditional IL conceptualisation. The traditional labour theory of Karl Marx conceptualises economic development with key components of labour, value, property and production relationship. David Harvey [40] provides a modern interpretation with a significant influence on academic and political debates around the world. The work on urbanisation and the political economy of cities has been particularly influential and has been a vocal critic of the neoliberal policies that have shaped urban development in many parts of the world. The work has often addressed the intersections between political economy, social inequality, and environmental degradation. In this paper, we introduce the recent work which develops an industrial landscape conceptual framework in the context of AI and data capital [16]. We develop a consistent illustration of the IL framework in the work of David Harvey and that in the new contexts of data capital as illustrated in Figure 1.

The driving power in traditional IL conceptualisation is money capital, which meets producer-effective demands. The effect combines with free gifts of nature to facilitate labour and the means of production. Produced commodities stimulate the realisation of value in monetary form after deducting wage goods and the cost of the means of production. The production, reproduction and destruction of human nature and culture shape the free gift of human nature and fundamental wants, needs, and desires. The resulting consumer effective demands are matched to the realisation of value in money form through marketing activities and create distribution to the producer, consumer, and back-to-money capital. In this IL framework, the key gateways to control the flow of money capital are the means of production and distribution.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 December 2024 doi:10.20944/preprints202412.0390.v1

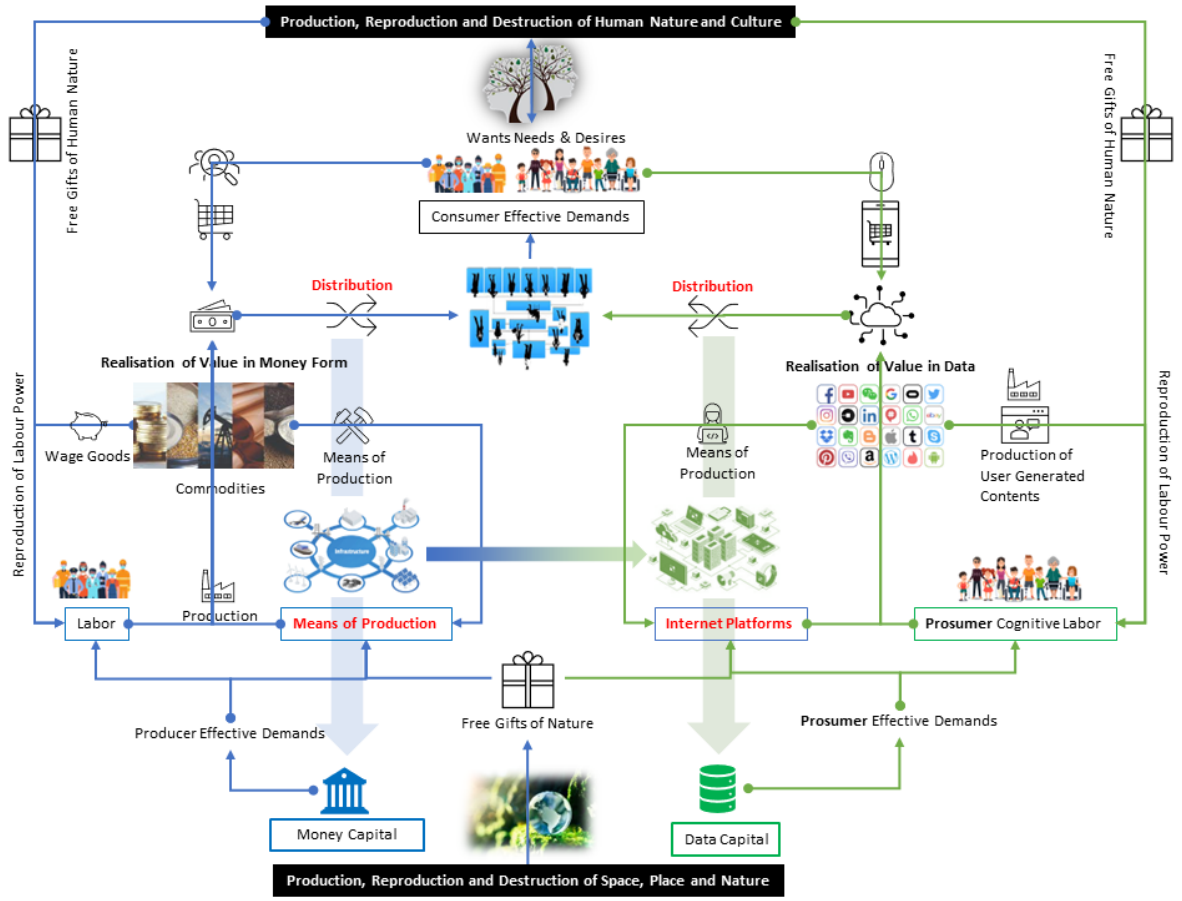

Figure 1. Comparison between traditional industrial landscape and that in the new contexts of digital technology. The new data capital model provides a roadmap to guide the DBCI development.

In the new contexts of data capital [16], in particular the recent AIGC and large model era, internet platforms have become the fundamental infrastructure. The new IL framework is particularly useful in understanding the technical development of contemporary AI, such as computer vision and natural language processing. The data capital IL framework discusses the differences and commonalities between the traditional bourgeoisie and the new bourgeoisie, referred to as neo-bourgeoisie. The traditional bourgeoisie owns the means of production and has high fixed costs, while the neo-bourgeoisie owns the means of connection and has low fixed costs. For example, digital products, such as online videos and games are not limited by their physical forms and can serve the scalable need of customers. The factors involved in production for the traditional bourgeoisie are land, labour, and capital, while for the neo-bourgeoisie, they are data and information. For example, many online services and products are free to use as the owner of a digital gateway can gather valuable data and information. Data can be used to supply further development for business analysis and AI training while information is essential in controlling information distribution and matching market needs. Both data capital and money capital have monopoly power and high economic rents. The framework also discusses the differences and commonalities between the proletariat and the neo-proletariat. The former is paid for labour hours, while the latter receives free services in exchange for personal data and cognitive working load.

Our work focuses on analysing the progress and perspective of DBCI technologies in the context of the data capital IL framework. Different to previous surveys that are technique-driven, this paper provides a hybrid paradigm. Firstly, we derive the survey structure using scoping review approach using the IL framework. Based on the derived structure, we then match the development of DBCI

 doi:10.20944/preprints202412.0390.v1

models and data using the systematic literature review approach. Meta-analysis is also provided to compare key parameters, such as DBCI applications, data statistics, and BCI devices.

3. Methodology

3.1. Conceptualisation of DBCI Industrial Landscape

So far, there are more than 600 BCI survey papers published from 1986 to 2024. However, none of the surveys has put the technical development of BCI in the context of the industrial landscape which is crucial to understand how the factors of data, devices, commercialisation, etc. are shaping research. Therefore, we introduce the recent data capital IL framework [16] as an initial scoping review to narrow down and identify the following key topics. Our review methodology is summarised in Figure 2. As a result, there are a total of 53 datasets included in this paper which contribute to the following four key topics.

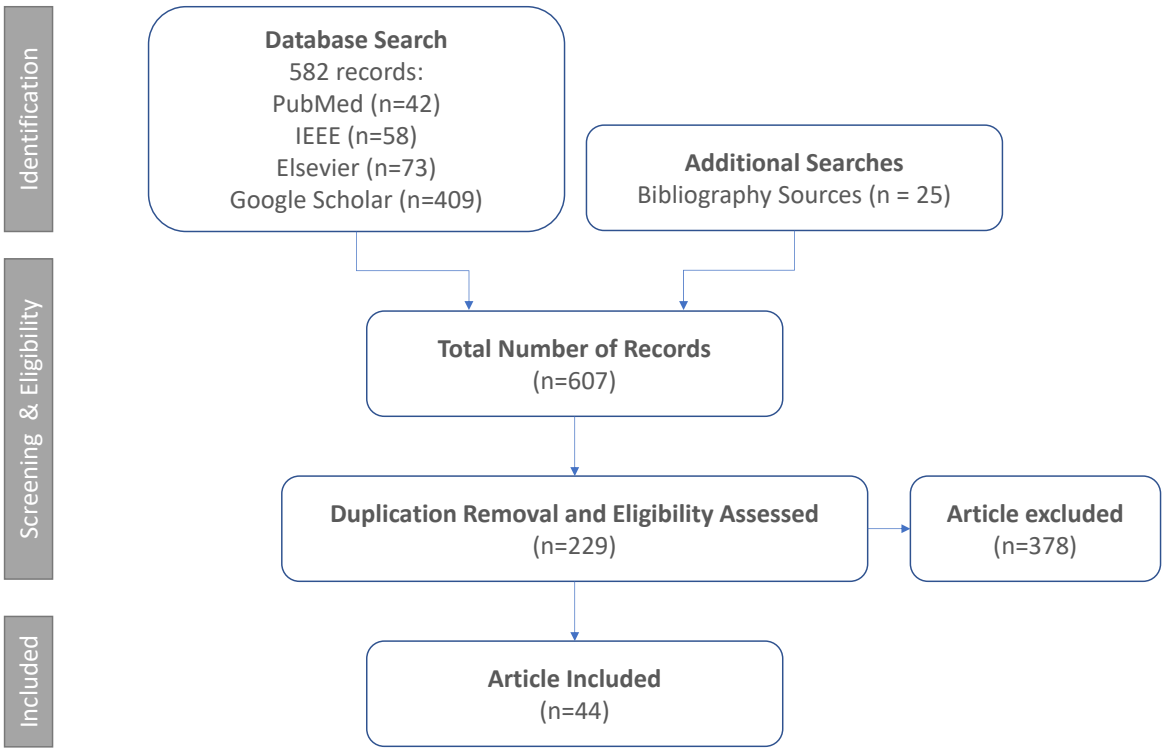

Figure 2. 607 survey and dataset papers are collected from our initial search. After screening and eligibility check, we found 229 papers to check whether they significantly contribute to our conceptual framework. As a result, 44 benchmarks from 38 papers are included in this study.

DBCI applications consider the impact of big data and artificial intelligence (AI) on the economic, social, and political systems of the world. AI has increased the ability to produce more for economic growth and development while also making human labour obsolete. This creates a trajectory where capitalism remains the ultimate system, controlling the lives of labour through big data. However, the growth of AI also promotes technological innovation and investment, leading to economic growth. The profit-driven technological singularity of AI creates social challenges and potentially fatal economic impacts under a neoliberal economic system. AI also creates a digital divide and potentially expands existing societal rifts and class conflicts. It is essential to develop policies to protect labour, privacy, trade, and liability and reduce the consequences of AI's impact on employment, inequality, and competition. The DBCI may create opportunities for individuals to monetise their personal data and potentially transfer control and ownership to actual data producers in a passive way, i.e. the mind

 doi:10.20944/preprints202412.0390.v1

activity and focused time consumption. Application is, therefore, a key parameter in evaluating the maturity and progress of the DBCI industrial landscape.

The Utility of DBCI The economic landscape has undergone major changes in the past few decades with the emergence of new Internet technologies and the creation of value through business model innovation using data and information. The factors of production have been redefined with data and information being recognised as new variables that have been made possible by technological breakthroughs in information and communications technology. The cost of computing power, data storage, and Internet bandwidth has decreased significantly, enabling the creation of increasingly rich digital information. This has given rise to new phenomena such as Big Data analytics and Internet platform companies. The democratisation of information and knowledge has also increased the bargaining power of workers and consumers whilst impacting Marxist philosophy in two areas related to value creation. The commodification of cognitive labour is the foundation of the new capitalist system in which modes of control over production, consumption, distribution, and exchanges are very different from earlier forms of capitalism in history. This new economy of capitalist transformation is referred to as 'cognitive capitalism.' A fundamental parameter of cognitive capital is access to a large-scale population. In the context of DBCI research, this is closely tied to utility, which is defined by the cost of devices and the flexibility of usage paradigms.

Value of Cognitive Workload The traditional idea that the value of products and services is measured in labour hours has been challenged by the process of datafication, which involves dematerialization, liquefaction, and density. Digitisation has made it possible for companies like Netflix to offer on-demand services and gather data on user behaviour. Digital products are also non-rivalrous and non-excludable, which means that they can be used by many individuals at the same time without reducing their availability to others. The availability of free digital services and products also challenges the use of labour hours to value a product or service, as many are provided through advertising or other business models. The concept of the "Prosumer" further undermines the traditional value creation process, as much of the online content is produced by the consumer for free. While existing AIGC technologies have provided the premises for creation, the *Cognitive Workload* in DBCI provides one step further. The research of cognitive workload can potentially encourage a healthy and fair ecosystem for DBCI and other large models for real-world applications.

Data and Model Ownership The scoping review discusses how the traditional Marxist dichotomy between bourgeoisie owners of the means of production and proletariat workers has been upended by the emergence of platform-based internet companies. These companies, such as Amazon, Google, and Facebook, do not own the means of production but rather the means of connection to the internet, and they leverage large amounts of customer data to create value. The article also discusses the democratization of information and the shift in power from traditional owners to individuals and entrepreneurs, as well as the emergence of the sharing economy and the de-linking of assets from value. In the AIGC era, the AI ecosystem is moving from the traditional data capital to the current model capital paradigm, such as ChatGPT. Largescale deep models, regardless open-source or not, are no longer accessible to common users for model fine-tuning. Deep model API or MLaaS have become the dominant practice. In the DBCI research, deep learning models are among the early stages in this model capital wave. Our review will discuss the influence of existing data and AI model capitals to the DBCI domain.

3.2. Process of DBCI Data Capital Liquidation

The process of data asset liquidation is intrinsically linked to the broader landscape of DBCI applications, encompassing their utility, the value of cognitive workload, and data and model ownership. By systematically evaluating and managing data assets, we can maximise their potential in driving forward DBCI applications, which rely heavily on high-quality and extensive datasets to develop and refine models that enable innovative solutions in healthcare, neurorehabilitation, and beyond. Understanding the utility of DBCI involves assessing the cost-effectiveness and accessibility of devices

 doi:10.20944/preprints202412.0390.v1

and paradigms, ensuring that the technology can be widely adopted and utilised. Moreover, the value of cognitive workload emphasises the importance of accurately measuring and leveraging cognitive data to enhance user experience and productivity, making it crucial to manage and assess data quality effectively. This comprehensive approach to data asset liquidation not only supports the advancement of DBCI technologies but also addresses the multifaceted challenges and opportunities within the industry.

We summarise our broken-down assessment factors in Table 1. Specifically, for BCI devices, frequency (Hz) indicates how often signals are sampled per second. Higher frequencies capture finer temporal resolution, which is crucial for tracking rapid brain activities. EEG channels represent the number of electrodes used in data collection. A higher number of channels offers better spatial resolution, capturing data from more regions of the brain. For DBCI Applications, high-frequency and multi-channel devices enable applications requiring precise brain activity mapping, such as neurorehabilitation and emotion recognition. For BCI utility, devices with higher frequency and channel count are more versatile but can be costlier and less portable. Optimising these metrics balances performance and usability in real-world settings. For the value of cognitive workload, accurate frequency and spatial resolution improve the fidelity of cognitive workload measurements, enabling deeper insights into attention, fatigue, and performance. For data and model ownership, high-resolution devices are often proprietary, with access to raw data or model training pipelines controlled by manufacturers. This raises questions about open standards and accessibility.

Table 1. Mapping Metrics of Devices, Data, and Applications to the DBCI Industrial Landscape. Note that trials refers to trials per user.

Aspect	**Metric**	**Why It's Used**	**Connection to Industrial Landscape**
Devices	Frequency (Hz)	Captures temporal resolution of brain activity.	Enables high-precision DBCI applications, improves cognitive workload modelling, but is often tied to proprietary devices.
	EEG Channels	Indicates spatial resolution of brain activity.	Supports diverse applications, increases utility and workload fidelity, but raises ownership challenges.
Data	Length (s)	Determines duration of captured data for each trial.	Supports long-term applications, increases utility, and enhances workload assessment across varied contexts.
	Trials	Reflects dataset robustness and reliability.	Ensures applicability in diverse scenarios, increases model reliability, and requires careful ownership considerations.
	Users	Represents diversity and generalisability of the dataset.	Enables cross-population applications, improves utility, and raises ethical issues about ownership and privacy.
Applications	Stimuli	Defines the context of recorded brain activity.	Links directly to DBCI use cases, increases task-specific utility, and impacts workload relevance and accessibility.
	Task	Defines the dataset's relevance to specific DBCI applications.	Drives model training for targeted use cases, improves cognitive workload insights, and ties to ownership of annotations.
	Response	Determines modalities available for analysis (e.g., EEG, behavioural responses).	Increases flexibility across applications, improves model utility, but raises accessibility challenges due to ownership.

 doi:10.20944/preprints202412.0390.v1

The second metric we consider is the data. For DBCI applications, longer trial lengths and diverse participant pools make the datasets applicable to a wider range of use cases, such as personalised neurofeedback or cross-cultural studies. For BCI utility, more trials and participants increase the dataset's statistical power but also its complexity and storage requirements. This impacts its usability for researchers and practitioners. For the value of cognitive workload, repeated trials and diverse user data improve the accuracy and generalisability of cognitive workload models, ensuring they work effectively across different scenarios. And, for data and model ownership, datasets with longer trials and diverse users often require substantial investment. Ownership can dictate access, limiting opportunities for public or collaborative research. Specifically, we use length (s) for the duration of each recorded trial, which impacts the total data volume and its utility in capturing prolonged cognitive states. Trials reflect the number of repetitions per user, affecting dataset reliability and robustness. Users indicate the number of participants in a dataset, determining its diversity and generalisability across populations.

Finally, we consider applications in terms of stimuli, task, and response as metrics. The type of stimuli (e.g. visual, auditory) presented to participants defines the context of the dataset and its relevance to specific applications. Task describes what participants were asked to do (e.g., motor imagery, attention tasks), directly linking the dataset to specific DBCI use cases. Response refers to the recorded data types (e.g., EEG signals, behavioural responses), which determine the modalities available for model training and application. For DBCI applications, stimuli, tasks, and responses define the real-world scenarios where the dataset can be applied. For example, datasets with motor imagery tasks are crucial for prosthetics, while emotional stimuli datasets are vital for affective computing. For utility, datasets with diverse stimuli and task types are more flexible but require sophisticated annotation and preprocessing, impacting ease of use. The value of cognitive workload is often associated with the stimuli and task types that influence cognitive demands, making these metrics critical for accurately modelling workload and designing adaptive systems. Data and model ownership can be reflected by datasets with complex stimuli and multi-modal responses which are often proprietary due to the cost and effort involved in collection, limiting broader accessibility and collaboration.

4. Survey Results

Our survey results are summarised in Table 2. The table summarises a comprehensive survey of BCI datasets, focusing on metrics across devices, data, and applications. These metrics provide a valuable foundation for understanding the current landscape of BCI data capital and its alignment with key technical and industrial challenges. Below are the general descriptions of the dataset characteristics based on the metrics presented.

Table 2. Summary of BCI datasets survey results.

	Devices		Data			Application		
Dataset Name	Freq	Chan	Len	Tri	Use	Stimuli	Task	Response
WAY-EEG-GAL[41]	500	32	10	328	12	Visual Cue	Motor Imagery	EEG, EMG, Event Timings, Object Positions, Object Forces
GigaDB-EEG-MI[42]	512	64	3	260	52	Visual Cue	Motor Imagery	EEG, EMG, EOG, Hand Movement Data, Questionnaire
PhysioNet-EEG-MI[43]	160	64	120	12	109	Visual Cue	Motor Imagery	EEG, Annotations
Largescale-EEG[44]	200	19	3	900	13	Visual Cue	Motor Imagery	EEG
BCI Comp II dataset 1a [45]	256	6	3.5	293	1	Visual Feedback	Motor Imagery	EEG

 doi:10.20944/preprints202412.0390.v1

BCI Comp II dataset 1b [45]	256	6	4.5	200	1	Visual Feedback, Audio	Motor Imagery	EEG
BCI Comp II dataset 2a [45]	160	64	30	60	3	Visual Feedback	Motor Imagery	EEG
BCI Comp II dataset 3 [45]	128	3	9	280	1	Visual Feedback	Motor Imagery	EEG
BCI Comp II dataset 4 [45]	1000	28	0.5	416	1	None	Motor Imagery	EEG, Typing
BCI Comp III dataset 1 [45]	1000	64	3	378	1	N/A	Motor Imagery	ECoG
BCI Comp III dataset 2 [45]	240	64	2.5	92	2	Character Matrix	P300	EEG
BCI Comp III dataset 3a [45]	240	64	7	80	3	Visual Cue, Audio Cue	Motor Imagery	EEG
BCI Comp III dataset 3b [45]	125	2	8	40	3	Visual Cue	Motor Imagery	EEG
BCI Comp III dataset 4 [45]	1000	118	3.5	280	2	Visual Cue	Motor Imagery	EEG
BCI Comp III dataset 5 [45]	512	32	240	4	3	Audio Cue	Motor Imagery	EEG
BCI Comp IV dataset 1 [45]	1000	64	3.5	42	7	None	Motor Imagery	EEG, Artificial EEG
BCI Comp IV dataset 2 [45]	250	22	6	576	9	Audio Cue	Motor Imagery	EEG, EOG
High-Gamma[46]	500	128	4	880	14	Visual Cue	Motor Imagery	EEG
Planning-Relax[47]	256	8	5	10	1	Audio Cue	Motor Imagery	EEG, EOG
DAEP[48]	512	32	60	40	32	Music, Video	Emotion Recognition	Face Recordings, Questionnaire, EOG, EMG, Blood Pressure, GSR, Respiration
HeadIT[49]	256	256	218	15	32	Audio	Emotion Recognition	EEG, ECG, Infra-ocular
Enterface06[50]	1024	54	2.5	450	5	Image	Emotion Recognition	EEG, fNIRS, GSR, Respiration, Video
NeuroMarketing[51]	128	14	4	42	25	Image	Neuromarketing	EEG, Questionnaire
SEED[52]	1000	62	240	45	15	Video	Emotion Recognition	EEG, Eye Movement, Self Assessment Questionnaire
HCI Tagging[53]	512	32	135	20	30	Image, Video	Emotion Recognition	EEG, GSR, ECGG, Eye Tracking, Audio, Video, Questionnaire
Regulation of Arousal[54]	500	64	45	24	18	Audio, Simulation	Neurofeedback	EEG, ECG, EDA, Respiration, Pupil Diameter, Eye Tracking
BCI-NER Challenge[55]	600	56	10.51	340	26	Character Matrix	P300	EEG, MEG
Face-House[56]	1000	N/A	0.8	300	7	Image	Neural Decoding	ECoG, ERPS
Synchronised Brainwave[57]	512	1	319	1	30	Video	Neural Decoding	EEG

 doi:10.20944/preprints202412.0390.v1

Target vs Non-target[58]	512	16	300	3	64	Character Matrix	P300	EEG
Impedance[59]	1024	10	1.5	1280	12	Text	Neural Decoding	EEG, EOG
Sustained Attention[60]	500	30	5400	2.5	27	Simulation	Driving	EEG, Questionnaire
Dryad-Speech[61]	512	128	105	20	92	Audio	N400	EEG
SPIS Resting State[62]	256	64	300	1	10	None	Resting State	EEG, EOG
Alpha-waves[63]	512	16	10	10	20	None	Resting State	EEG, Questionnaire
Music Imagery Retrieval[64]	400	14	11.5	12	10	Music	Music Imagery	EEG
EEG-eye State[65]	128	14	117	1	1	None	Eyestate	EEG
EEG-IO[66]	250	19	3.5	25	20	N/A	Eyestate	EEG, Annotations
Eye State Prediction[67]	N/A	14	117	1	1	None	Eyestate	EEG, Video, Annotations
Kara-One[68]	1024	64	2100	1	8	Text, Audio	Speech Imagery	EEG, Video, Audio
MNIST Brain Digits[69]	161	11	2	12066111		Image	Neural Decoding	EEG
ImageNet Brain[69]	128	5	3	14012	1	Image	Neural Decoding	EEG
EEGLearn[70]	500	64	3.5	240	13	Text	Neural Decoding	EEG
Deep Sleep Slow Osciallation[71]	125	N/A	10	1261	N/A	None	Slow Oscillation Prediction	EEG, Sleep Stage, Time Sleeping
Genetic Predisposition to Alcholism[72]	256	64	1	120	122	Image	Neural Decoding	EEG
Confusion During MOOC[73]	2	1	60	10	10	Video	Education Feedback	EGG, Questionnaire
TUH EEG Corpus[74]	250	31	167	1.56	10874	None	Seizure Detection	EEG, Clinitian Report
Predict-UNM[75]	500	64	3.6	200	25	Medication, Audio	Neural Decoding	EEG
ERP CORE[76]	1024	30	600	6	40	Image, Video, Audio	Face Perception	EEG, ERP
Statistical Parametric Mapping[77]	2048	128	1.8	172	1	Image, Audio	Face Perception	EEG, fMRI, MEG, sMRI, EOG
GOD-Wiki[78]	N/A	N/A	3	590	5	Image	Neural Decoding	fMRI, Image, Text
DIR-Wiki[78]	N/A	N/A	2	2400	3	Image	Neural Decoding	fMRI, Image, Text
ThingsEGG-Text[78]	1000	64	0.235	8216	10	Image	Neural Decoding	EEG, Image, Text

For devices, the datasets span a wide range of sampling frequencies, from low frequencies such as 128 Hz in the NeuroMarketing dataset to very high frequencies like 2048 Hz in the Statistical Parametric Mapping dataset. High-frequency datasets, such as ThingsEEG-Text (1000 Hz), are ideal for capturing rapid neural dynamics, essential for decoding precise temporal brain activity. Lower frequencies are generally sufficient for static tasks or simple signal processing, such as motor imagery. The number of EEG channels varies significantly, ranging from 1 channel (e.g., Synchronised Brainwave) to 256 channels (e.g., HeadIT dataset). Multi-channel setups are crucial for high spatial resolution, supporting applications like emotion recognition (SEED dataset) or complex neural decoding (GOD-Wiki).

 doi:10.20944/preprints202412.0390.v1

In the data category, trial durations vary widely, with some datasets focusing on short, event-related trials (e.g., DIR-Wiki (2 seconds)) and others providing longer continuous recordings (e.g., Sustained Attention (5400 seconds)). Shorter trials are suitable for tasks like P300 spellers, whereas longer recordings are necessary for sustained attention or neurofeedback studies. Trials and Users: The number of trials and participants reflects the dataset's diversity and robustness. For instance: DIR-Wiki includes 2400 participants, making it highly suitable for inter-person generalisability. ThingsEEG-Text provides 8216 trials per user, supporting inter-sample learning for robust model training. Smaller datasets, like BCI Competition IV dataset 1 (7 participants), are ideal for exploring targeted applications or algorithms.

For applications, the datasets incorporate a variety of stimuli types, such as visual cues, audio cues, and videos, to simulate diverse cognitive and sensory tasks. For example: HCI Tagging utilises both images and videos for emotion recognition. GOD-Wiki integrates images and text, making it a prime example for visual-semantic decoding applications. Most datasets focus on motor imagery, a staple task in BCI research. However, emerging tasks like neural decoding (e.g., GOD-Wiki) and emotion recognition (e.g., SEED, DEAP) indicate growing interest in expanding the scope of BCI applications. Multi-modal datasets that include EEG and additional modalities (e.g., EEG, fMRI, Image, and Text in GOD-Wiki) are increasingly prevalent. These datasets support advanced tasks like zero-shot neural decoding and multi-modal integration, critical for expanding BCI applications. Summary of Dataset Contributions Support for BCI Applications:

Overall, motor imagery remains the most common task, providing a benchmark for BCI algorithm development. Novel tasks like neural decoding and emotion recognition reflect the evolution of BCI datasets toward more complex and versatile applications. Datasets with high temporal (e.g., ThingsEEG-Text) and spatial resolution (e.g., HeadIT) are critical for improving utility in advanced modelling techniques. Large participant pools (e.g., DIR-Wiki) ensure generalisability across diverse populations. Multi-modal datasets like SEED and HCI Tagging are invaluable for studying cognitive workload in realistic scenarios, enabling adaptive DBCI systems. Datasets with diverse trial designs and stimuli improve the fidelity of cognitive workload modelling. Open datasets like BCI Competitions and BraVL promote accessibility and collaborative research. Proprietary datasets with restricted access, particularly those involving high-cost modalities like fMRI, highlight the ongoing need for equitable data-sharing practices.

The diversity and richness of datasets summarised in the table provide a strong foundation for advancing DBCI research. The wide range of device specifications, data configurations, and application contexts ensures that these datasets are well-suited to address the challenges of generalisability, scalability, and adaptability in BCI systems. Next, we provide in-depth analysis with enhanced insights for each category

4.1. Device

The scatter plot in Figure 3, illustrating the relationship between frequency (Hz) and the number of EEG channels, was generated using the dataset information provided in the table. Both frequency and channel data were transformed into a log-2 scale to allow for a more interpretable comparison across datasets with varying magnitudes. The dataset "Confusion During MOOC," which had an unusually low frequency of 2 Hz, was excluded to avoid distortion of the plot. The x-axis represents the log-2 of the number of channels, while the y-axis represents the log-2 of the frequency. Each point on the scatter plot corresponds to a dataset, allowing us to visualise the distribution and clustering of datasets based on their device configurations. This approach highlights key patterns and outliers in the data, such as datasets with exceptionally high temporal or spatial resolution, facilitating deeper analysis of trends in DBCI devices.

 doi:10.20944/preprints202412.0390.v1

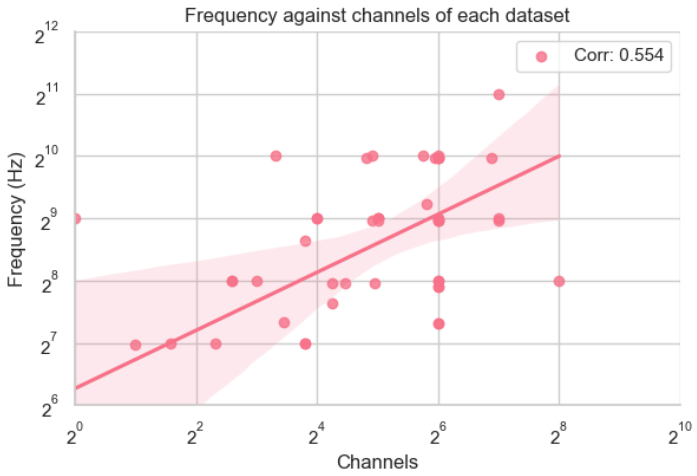

Figure 3. The frequency of each dataset plotted again the number of channels on a log-2 scale. The Confusion During MOOC[73] dataset was an outlier with a frequency of only 2Hz, and so was cut off from this graph.

Analysing the scatter plot reveals several notable trends and insights into the current state of device metrics in DBCI datasets. The majority of datasets cluster around 32–64 channels and 128–512 Hz frequencies, reflecting the most common experimental setups in EEG research. This range balances temporal and spatial resolution, making it suitable for general-purpose applications such as motor imagery, emotion recognition, and cognitive workload studies. A few datasets stand out as outliers. For example, HeadIT features an exceptionally high number of channels (256), which enhances spatial resolution and is particularly valuable for advanced applications like high-resolution neural decoding or emotion recognition. On the other hand, datasets like Enterface06 (1024 Hz) and Statistical Parametric Mapping (2048 Hz) offer exceptionally high temporal resolution, enabling precise tracking of rapid neural dynamics. These high-frequency datasets are critical for applications such as speech imagery, real-time neurofeedback, or fine motor control.

Interestingly, there is a moderately positive correlation of 0.554 between the frequency and number of channels. This may be a result of technology improving over time with more modern devices having a greater bandwidth, resulting in both a higher frequency and a large number of channels. Alternatively, this may not be due to fundamental bandwidth limitations and instead due to financial limitations where devices with high bandwidth may be too expensive. This highlights a limitation of our study and we leave it to future work to investigate the price BCI devices.

From an industrial landscape perspective, the clustering of datasets around 32–64 channels and 128–512 Hz frequencies reflects the standardisation of EEG devices. This standardisation ensures compatibility and widespread usability across research and clinical settings, contributing to the utility of these devices. However, datasets that rely on higher-channel and higher-frequency devices often involve proprietary equipment, raising challenges related to data and model ownership. Furthermore, datasets with extreme configurations, such as high-channel or high-frequency setups, cater to niche applications but may face scalability and cost-effectiveness challenges in real-world DBCI deployment. Overall, the diversity in device configurations highlights the ongoing need to balance spatial and temporal resolution to meet the varying demands of DBCI applications. While standard configurations dominate due to their general usability, high-resolution setups offer unique opportunities for advanced research, albeit with limitations in accessibility and scalability.

 doi:10.20944/preprints202412.0390.v1

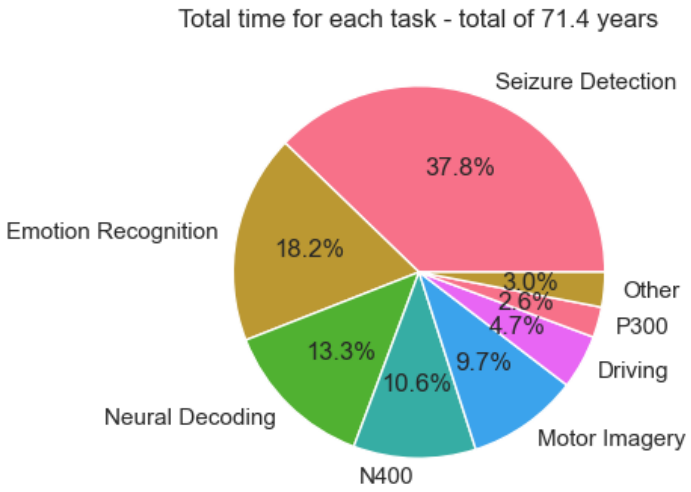

Figure 4. The fraction of data belonging to each task, measured in years. In total, the datasets we looked at had 71.4 years of data. Tasks with less than 1.5 years of data were merged into 'Other'.

4.2. Data

The pie chart was created to represent the proportion of accumulated data for each task, such as seizure detection, emotion recognition, and neural decoding, based on the provided formula:

$$time = channels \times trial\ length \times trials\ per\ subject \times subjects$$

This formula calculates the total recording time for each dataset in years by multiplying the number of EEG channels, the length of each trial, the number of trials per subject, and the number of subjects. The datasets were then grouped by task, and the total recording time for each task was summed. Tasks with less than 1.5 years of data were combined into the "Other" category to simplify the visualisation. The total accumulated data across all tasks was 71.4 years, and the pie chart shows the fraction of this total for each task. We summarise our findings according to the industrial landscape framework:

Applications The pie chart analysis highlights the dominance of seizure detection, accounting for 37.8% of the total data. This reflects the clinical priority of seizure detection in healthcare, where its applications in epilepsy diagnosis and monitoring are highly established. It's worth noting that the data for seizure detection comes from a single large data set, the TUH EEG Corpus[74]. The impressive size of this dataset shows that a large volume of data can be gathered when a device is widely deployed. Furthermore, this is a very diverse dataset with data coming from over 10,000 patients, meaning that a model trained on this data will be robust due to the high inter-subject variability. These factors combined make the dataset well-suited for real-world deployment, showing that seizure detection is a mature task in the DBCI application landscape. On the other hand, tasks like emotion recognition (18.2%) and neural decoding (13.3%) represent expanding frontiers in BCI research. These emerging applications cater to the rising demand for adaptive systems in mental health, emotion-aware technologies, and cognitive analysis, showcasing their growing relevance in the industrial framework. However, tasks like driving (4.7%) and P300 paradigms (2.6%) remain underrepresented despite their direct applicability to safety-critical applications and assistive devices, indicating the need for further investment to enhance their practical deployment.

Utility The dataset distribution underscores the significant utility of core tasks like motor imagery (9.7%) and N400 (10.6%) in the DBCI landscape. Motor imagery serves as a cornerstone for neurorehabilitation and prosthetic control, while N400 supports applications in linguistic processing and cognitive workload analysis. Their substantial data representation highlights their importance for developing reliable and scalable BCI systems. In contrast, the other category (3%) and specialised tasks

 doi:10.20944/preprints202412.0390.v1

like driving-related paradigms reflect limited utility due to insufficient data accumulation. Expanding data collection efforts for these underrepresented areas could significantly enhance their scalability and integration into diverse real-world applications, fostering a more balanced utility across the DBCI domain.

Value of Cognitive Workload The significant proportion of datasets dedicated to emotion recognition and neural decoding reflects a growing emphasis on modelling cognitive workload within the DBCI landscape. These tasks enable the development of systems that adapt to users' cognitive and emotional states, supporting advanced applications such as emotion-aware interfaces, cognitive workload management, and mental health monitoring. However, the limited data availability for tasks in the other category suggests missed opportunities for expanding cognitive workload research into less-explored domains. A more diversified dataset ecosystem could provide deeper insights into user cognition and behaviour, enhancing the adaptability and personalisation of DBCI systems.

Data and Model Ownership The dominance of seizure detection datasets highlights a relatively mature ecosystem for data collection, sharing, and model development in this domain. This maturity offers opportunities to refine data-sharing frameworks, ensuring equitable access and fostering collaborative research. However, the limited representation of lesser-explored tasks, grouped under the other category, presents challenges related to data ownership and accessibility. Addressing these challenges requires the establishment of robust frameworks for data sharing and ownership, particularly for underrepresented tasks. This would support a more equitable and innovative landscape for developing open-access datasets and models across the DBCI spectrum.

4.3. Application

To analyse the distribution of stimuli, tasks, and responses across datasets, three bar charts were created. For stimuli, a bar chart was generated to show the number of times each type of stimulus (e.g., visual cues, audio cues, video) was featured in a dataset. For tasks, another bar chart represented the frequency of each task (e.g., motor imagery, emotion recognition, seizure detection) in the datasets. For responses, the chart depicted the number of datasets that recorded various responses (e.g., EMG, EOG, fMRI). EEG, being the dominant response type, was excluded from the responses chart to avoid overshadowing other data modalities. In total, the analysis considered 47 datasets that recorded EEG, allowing a detailed exploration of how stimuli, tasks, and responses are distributed in the DBCI landscape.

Figure 5 shows that the distribution of stimuli reveals a strong focus on visual stimuli, which dominate the datasets. Visual cues feature heavily in tasks like motor imagery, whilst images appear more in tasks that require more complicated stimuli, like neural decoding and emotion recognition. However, the inclusion of audio cues and video stimuli in several datasets reflects the expanding diversity of applications, such as emotion recognition and cognitive workload assessment, which demand multi-modal data to mimic real-world environments. The growing use of diverse stimuli suggests a shift toward broader applicability of DBCI systems, including multimedia interactions and adaptive user interfaces.

The analysis of tasks in Figure 6 underscores the dominance of foundational paradigms, like motor imagery and seizure detection, which are critical for clinical and rehabilitative applications. However, the emergence of tasks like emotion recognition and neural decoding signals the diversification of DBCI utility into consumer-oriented applications, such as mental health monitoring and cognitive enhancement tools. These trends indicate that DBCI research is moving beyond traditional clinical use cases toward more general-purpose systems that align with evolving user needs and technological capabilities.

The response data in Figure 7 highlights the inclusion of multi-modal recordings, such as EOG, EMG, and fMRI, alongside EEG. The use of these additional modalities supports the modelling of complex cognitive and emotional states, which are critical for understanding cognitive workload in diverse scenarios. For example, datasets incorporating fMRI and EOG responses provide high-

 doi:10.20944/preprints202412.0390.v1

resolution insights into brain activity and eye movements, respectively, enriching the development of adaptive and context-aware DBCI systems. This multi-modal approach aligns with the growing emphasis on cognitive workload evaluation, ensuring that systems can dynamically respond to users' mental states.

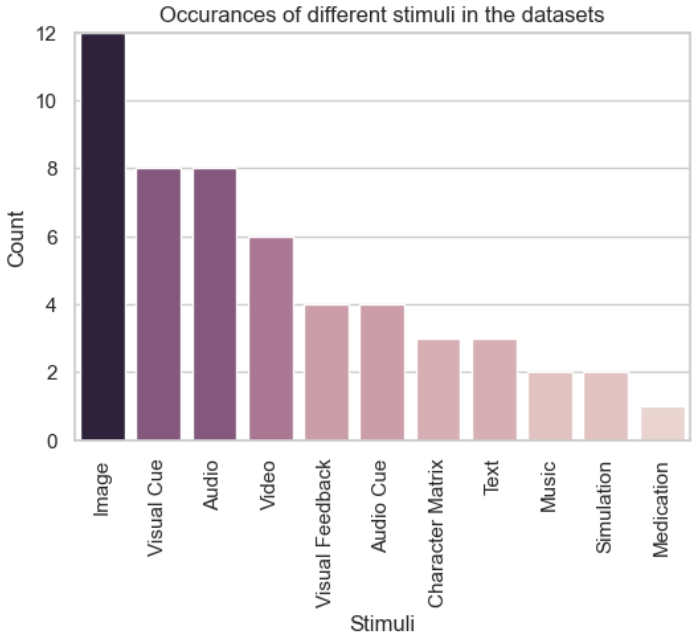

Figure 5. Bar chart showing the number of times each stimuli was featured in a dataset.

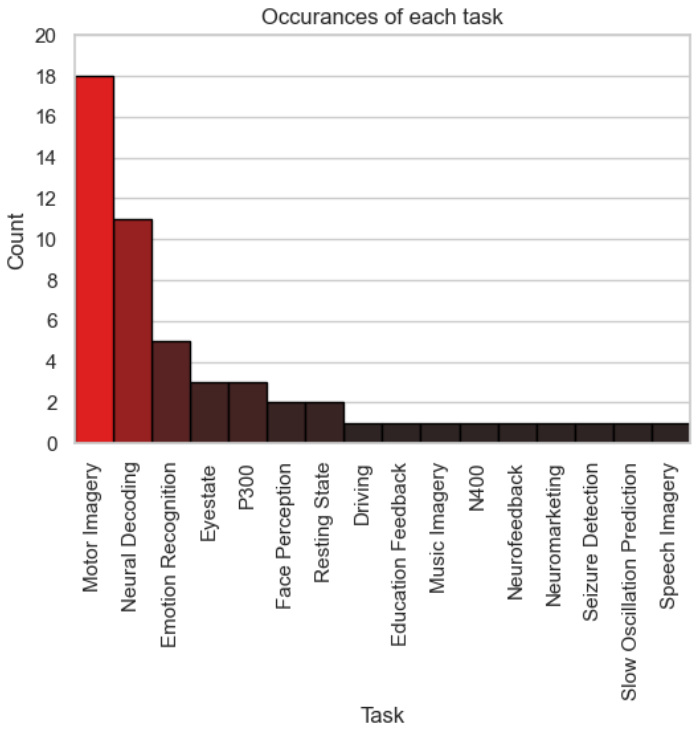

Figure 6. Bar chart showing the number of times each task featured in a dataset.

 doi:10.20944/preprints202412.0390.v1

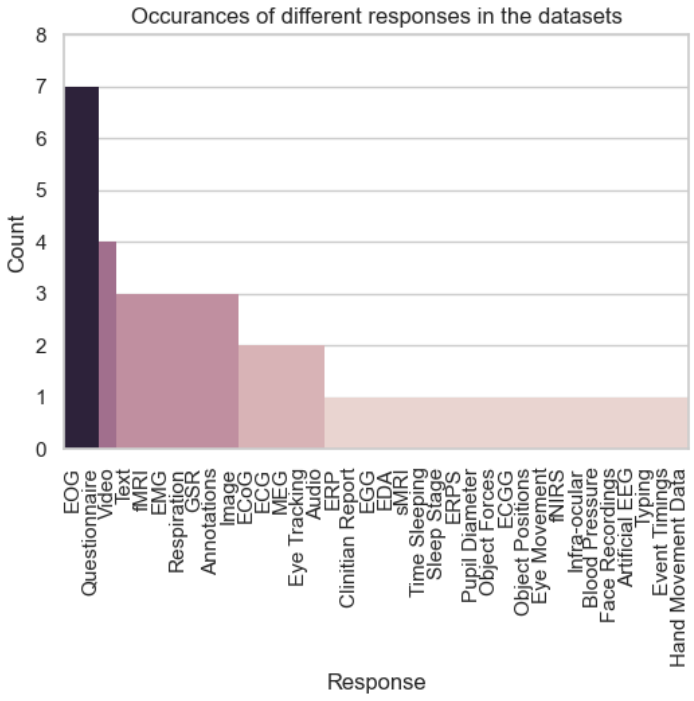

Figure 7. Bar chart showing the number of times each response was featured in a dataset. EEG was removed as the other data became dwarfed. In total, there were 47 datasets we looked at that recorded EEG.

Finally, the growing inclusion of alternative responses such as EMG and fMRI indicates a diversification of data modalities, which has implications for data and model ownership. Proprietary restrictions associated with high-cost modalities like fMRI may limit accessibility and collaboration. On the other hand, the widespread use of EEG reflects a more open ecosystem, promoting data sharing and model development. Addressing ownership challenges for multi-modal datasets is crucial for fostering equitable innovation in the DBCI domain.

4.4. Zero-Shot Neural Decoding Techniques

The above analysis of DBCI data capital has shown a critical juncture. While traditional deep learning or machine learning approaches may still struggle with the limited devices, data, and applications, emerging techniques such as Zero-Shot Neural Decoding (ZSND) and the availability of high-quality, multimodal datasets from other domains are enabling solutions to longstanding challenges. In this section, we synthesise the contributions of ZSND techniques and dataset metrics to address our claim of establishing a technical roadmap for overcoming barriers in DBCI applications, utility, cognitive workload, and data and model ownership.

ZSND techniques [78] enable DBCI systems to generalise across unseen samples, individuals, devices, domains, and tasks without extensive retraining. These capabilities are made possible by cutting-edge frameworks such as BraVL, which integrates brain activity with visual and linguistic information through trimodal learning approaches. The use of multimodal data ensures that models can transfer knowledge effectively, mitigating the following challenges:

Inter-Sample and Inter-Person Transfer ZSND datasets, such as DIR-Wiki with 2400 participants and ThingsEEG-Text with 8216 trials per participant (10 participants), provide the diversity necessary for robust inter-person generalisation. These datasets allow models to adapt to neural variability across individuals, a critical requirement for DBCI applications like personalised neurorehabilitation. Inter-sample transfer is enhanced by the trial-level richness of datasets, as seen in ThingsEEG-Text, which captures high temporal resolution (1000 Hz) data across multiple conditions.

 doi:10.20944/preprints202412.0390.v1

Inter-Device and Inter-Domain Transfer By incorporating multiple modalities such as EEG, fMRI, image, and text, ZSND datasets bridge the gap between invasive and non-invasive techniques, facilitating inter-device adaptability. For example, BraVL supports the alignment of brain signals recorded via EEG or fMRI with visual and semantic stimuli, ensuring models remain functional across diverse hardware environments. Inter-domain transfer is critical for applying DBCI systems in new contexts, such as transitioning from laboratory settings to real-world applications. The multimodal design of GOD-Wiki and DIR-Wiki exemplifies how datasets can support cross-domain learning.
Inter-Task Transfer Neural decoding tasks in datasets like GOD-Wiki and ThingsEEG-Text demonstrate the capability of ZSND techniques to generalise across tasks. Models trained on image decoding tasks can seamlessly adapt to semantic decoding tasks due to shared latent representations. This inter-task flexibility is crucial for multi-purpose DBCI systems, enabling applications ranging from motor imagery control to emotion recognition.
Unility Enhancement Frameworks like BraVL leverage multimodal data integration to create robust visual-semantic neural signal models. These models align brain activity with both visual and linguistic information, expanding the scope of DBCI applications to include cognitive workload assessment, attention monitoring, and adaptive feedback systems. The inclusion of high-resolution data (e.g., 64-channel EEG in all datasets and 1000 Hz sampling in ThingsEEG-Text) enables advancements in signal processing techniques to improve signal-to-noise ratio (SNR). Enhanced SNR is essential for the scalable adaptation of DBCI devices in real-world environments.

Beyond the ZSND techniques, the high-quality data published also established a foundation for DBCI Progression. Our proposed metrics highlight the contributions of the datasets to the industrial landscape conceptualisation framework:

- Devices: High-frequency datasets such as ThingsEEG-Text ensure precise temporal resolution for decoding dynamic neural activity. The consistent use of 64-channel setups across datasets provides the spatial granularity necessary for diverse applications.
- Data: Datasets like DIR-Wiki, with its 2400 participants, address the need for diversity in neural data, improving inter-person generalisability.
- Applications: Multimodal stimuli in GOD-Wiki and DIR-Wiki datasets, including image and text, expand the applicability of DBCI systems to multi-modal tasks. Neural decoding tasks recorded in these datasets align directly with the practical needs of applications such as neurorehabilitation, cognitive monitoring, and emotion recognition.

Overall, our work establishes a roadmap for DBCI research by identifying key barriers and demonstrating how ZSND techniques and dataset metrics address them. ZSND datasets and techniques enable generalisation across diverse stimuli and tasks, expanding the applicability of DBCI systems. The inclusion of diverse participants and trials increases dataset reliability and usability, supporting scalable and robust model training. Multimodal data and advanced signal processing improve the fidelity of workload modelling, ensuring adaptive and context-aware systems. While proprietary aspects of devices and datasets remain a challenge, open frameworks like BraVL and publicly available datasets mitigate access barriers, fostering collaboration and innovation.

While ZSND techniques and datasets provide significant advancements, further work is needed to fully realise the potential of DBCI systems. Future efforts could focus on: 1) Expanding Modalities: Incorporating additional data modalities such as MEG or wearable EEG devices to enhance data diversity and usability; 2) Self-Supervised Learning: Leveraging unsupervised techniques to reduce dependency on large-scale annotated data, improving efficiency and scalability. 3) Standardisation: Establishing universal standards for dataset annotation and evaluation to enable seamless integration and benchmarking across research groups. By leveraging ZSND techniques and metrics, this roadmap provides a clear pathway for overcoming barriers and advancing the industrial framework of DBCI systems.

 doi:10.20944/preprints202412.0390.v1

5. Conclusions

In this work, we introduced the industrial landscape conceptualisation framework for data capital to understand and evaluate the development of DBCI domain. This framework highlights how data capital is established within the DBCI domain, identifying four key barriers: *Applications, Utility, Data and Model Ownership, and Cognitive Workload*. We further demonstrated how publicly available datasets can be assessed through metrics categorised into *Devices, Data, and Applications*. Using this measurement approach, we identified and analysed 53 top DBCI datasets to reflect the progression of the current DBCI data capital industrial framework. Moreover, we emphasised the role of emerging techniques such as Zero-Shot Neural Decoding, which has shown significant potential in mitigating the barriers by enabling more generalisable, scalable, and efficient utilisation of DBCI data.

While our study provides a comprehensive assessment of the current state of DBCI data capital, further work is needed to address the limitations and expand the scope of this research. Future directions include exploring the integration of additional data modalities, such as fMRI or MEG, into the industrial landscape framework to ensure a more holistic understanding of data capital. Additionally, the framework can be extended to evaluate existing DBCI models and accessibility challenges associated with data and model ownership. Advancements in self-supervised learning [79] and federated learning [80] techniques also present opportunities for enhancing data capital by improving data efficiency and privacy. Lastly, establishing an open standard for dataset annotation and evaluation could foster collaboration across academia and industry, accelerating innovation in DBCI applications.

Author Contributions: Conceptualization, Y.Long. and T.Ma; methodology, T.Ma; validation, Y.Long and J.Waide.; investigation, X.X.; resources, W.Ma; data curation, X.X.; writing—original draft preparation, D.Organisciak; writing—review and editing, T.Ma, J.Waide and Y.Long; visualization, Y.Long and J.Waide; supervision, W.Ma. All authors have read and agreed to the published version of the manuscript.

References

1. Hosseini, M.P.; Tran, T.X.; Pompili, D.; Elisevich, K.; Soltanian-Zadeh, H. Multimodal data analysis of epileptic EEG and rs-fMRI via deep learning and edge computing. *Artificial Intelligence in Medicine* **2020**, *104*, 101813.
2. Li, C.; Wang, B.; Zhang, S.; Liu, Y.; Song, R.; Cheng, J.; Chen, X. Emotion recognition from EEG based on multi-task learning with capsule network and attention mechanism. *Computers in Biology and Medicine* **2022**, *143*, 105303.
3. Cui, J.; Lan, Z.; Sourina, O.; Müller-Wittig, W. EEG-based cross-subject driver drowsiness recognition with an interpretable convolutional neural network. *IEEE Transactions on Neural Networks and Learning Systems* **2022**.
4. Fang, S.X.; Chiu, T.F.; Huang, C.S.; Chuang, C.H. Leveraging Temporal Causal Discovery Framework to Explore Event-Related EEG Connectivity. HCI International 2022–Late Breaking Posters: 24th International Conference on Human-Computer Interaction, HCII 2022, Virtual Event, June 26–July 1, 2022, Proceedings, Part I. Springer, 2022, pp. 25–29.
5. Takagi, Y.; Nishimoto, S. High-resolution image reconstruction with latent diffusion models from human brain activity. *bioRxiv* **2022**, pp. 2022–11.
6. Floridi, L.; Chiriatti, M. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines* **2020**, *30*, 681–694.
7. Vazquez Hernandez, A. Wittgenstein and the Concept of Learning in Artificial Intelligence. Master's thesis, The University of Bergen, 2020.
8. Payani, A.; Fekri, F. Inductive logic programming via differentiable deep neural logic networks. *arXiv preprint arXiv:1906.03523* **2019**.
9. Cao, L. A new age of AI: Features and futures. *IEEE Intelligent Systems* **2022**, *37*, 25–37.
10. Morin, A. Levels of consciousness and self-awareness: A comparison and integration of various neurocognitive views. *Consciousness and cognition* **2006**, *15*, 358–371.

11. Roselli, D.; Matthews, J.; Talagala, N. Managing bias in AI. Companion Proceedings of The 2019 World Wide Web Conference, 2019, pp. 539–544.
12. Terziyan, V.; Golovianko, M.; Gryshko, S. Industry 4.0 intelligence under attack: From cognitive hack to data poisoning. *Cyber defence in Industry* **2018**, *4*, 110–125.
13. Morley, J.; Machado, C.C.; Burr, C.; Cowls, J.; Joshi, I.; Taddeo, M.; Floridi, L. The ethics of AI in health care: a mapping review. *Social Science & Medicine* **2020**, *260*, 113172.
14. Savulescu, J.; Maslen, H. Moral enhancement and artificial intelligence: moral AI? *Beyond artificial intelligence: The disappearing human-machine divide* **2015**, pp. 79–95.
15. Walkowiak, E. Digitalization and inclusiveness of HRM practices: The example of neurodiversity initiatives. *Human Resource Management Journal* **2023**.
16. Walton, N.; Nayak, B.S. Rethinking of Marxist perspectives on big data, artificial intelligence (AI) and capitalist economic development. *Technological Forecasting and Social Change* **2021**, *166*, 120576.
17. Berger, H. Über das elektrenkephalogramm des menschen. *DMW-Deutsche Medizinische Wochenschrift* **1934**, *60*, 1947–1949.
18. Gruzelier, J. A theory of alpha/theta neurofeedback, creative performance enhancement, long distance functional connectivity and psychological integration. *Cognitive processing* **2009**, *10*, 101–109.
19. Vidal, J.J. Toward direct brain-computer communication. *Annual review of Biophysics and Bioengineering* **1973**, *2*, 157–180.
20. Kübler, A.; Kotchoubey, B.; Hinterberger, T.; Ghanayim, N.; Perelmouter, J.; Schauer, M.; Fritsch, C.; Taub, E.; Birbaumer, N. The thought translation device: a neurophysiological approach to communication in total motor paralysis. *Experimental brain research* **1999**, *124*, 223–232.
21. Miller, Leo V., N.E.D. Instrumental learning of heart rate changes in curarized rats: shaping, and specificity to discriminative stimulus. *Journal of comparative and physiological psychology* **1967**, *63*, 12–19. doi:10.1037/h0024160.
22. Taub, E. What Psychology as a Science Owes Neal Miller: The Example of His Biofeedback Research. *Biofeedback* **2010**, *38*, 108–117. doi:10.5298/1081-5937-38.3.108.
23. Bruno, Steven; Demertzi, A.M.A.L. *Coma and Disorders of Consciousness*; Vol. NA, 2017. Issue: NA Pages: NA Publication Title: NA.
24. Smith, Mark, D. Locked-in syndrome. *BMJ (Clinical research ed.)* **2005**, *330*, 406–409. doi:10.1136/bmj.330.7488.406.
25. Hill, TN; Schröder, M.H.T.W.B.N.F.M.U.W.G.E.C.E.S.B.K.A.B.N.N.L. Classifying EEG and ECoG signals without subject training for fast BCI implementation: comparison of nonparalyzed and completely paralyzed subjects. *IEEE transactions on neural systems and rehabilitation engineering : a publication of the IEEE Engineering in Medicine and Biology Society* **2006**, *14*, 183–186. doi:10.1109/tnsre.2006.875548.
26. Kübler, Niels, A.B. Brain-computer interfaces and communication in paralysis: extinction of goal directed thinking in completely paralysed patients? *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology* **2008**, *119*, 2658–2666. doi:10.1016/j.clinph.2008.06.019.
27. Hoesle, A. Between Neuro-potentials and Aesthetic Perception. Pingo Ergo Sum. In *The International Library of Ethics, Law and Technology*; 2014; Vol. NA, pp. 99–108. doi:10.1007/978-94-017-8996-7_8.
28. Zickler, Sebastian; Kleih, S.C.H.C.K.A.C.H. Brain Painting. *Artificial intelligence in medicine* **2013**, *59*, 99–110. doi:10.1016/j.artmed.2013.08.003.
29. Farwell, Emanuel, L.A.D. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and clinical neurophysiology* **1988**, *70*, 510–523. doi:10.1016/0013-4694(88)90149-6.
30. Birbaumer, N.; Hinterberger, T.I.I.H.K.B.K.A.P.J.T.E.F.H.N.G. A spelling device for the paralysed. *Nature* **1999**, *398*, 297–298. doi:10.1038/18581.
31. Sterman, M.B. Basic Concepts and Clinical Findings in the Treatment of Seizure Disorders with EEG Operant Conditioning. *Clinical EEG (electroencephalography)* **2000**, *31*, 45–55. doi:10.1177/155005940003100111.
32. Irimia, Rupert; Poboroniuc, M.I.B.G.C.D.C.O. High Classification Accuracy of a Motor Imagery Based Brain-Computer Interface for Stroke Rehabilitation Training. *Frontiers in robotics and AI* **2018**, *5*, 130–NA. doi:10.3389/frobt.2018.00130.
33. Kübler, Elisa Mira; Riccio, A.Z.C.K.T.K.S.C.S.S.P.D.L.H.E.J.M.D.A.H. The user-centered design as novel perspective for evaluating the usability of BCI-controlled applications. *PloS one* **2014**, *9*, e112392–NA. doi:10.1371/journal.pone.0112392.

34. Kaufmann, Stefan M.; Köblitz, A.R.G.W.C.K.A.T.S. Face stimuli effectively prevent brain-computer interface inefficiency in patients with neurodegenerative disease. *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology* **2012**, *124*, 893–900. doi:10.1016/j.clinph.2012.11.006.
35. De Vos, Katharina; Debener, S.M.G. Towards a truly mobile auditory brain–computer interface: Exploring the P300 to take away. *International journal of psychophysiology : official journal of the International Organization of Psychophysiology* **2013**, *91*, 46–53. doi:10.1016/j.ijpsycho.2013.08.010.
36. Blum, Stefan; Emkes, R.V.N.F.S.B.M.G.S.D. EEG Recording and Online Signal Processing on Android: A Multiapp Framework for Brain-Computer Interfaces on Smartphone. *BioMed research international* **2017**, *2017*, 3072870–3072870. doi:10.1155/2017/3072870.
37. Bleichner, Stefan, M.G.D. Concealed, Unobtrusive Ear-Centered EEG Acquisition: cEEGrids for Transparent EEG. *Frontiers in human neuroscience* **2017**, *11*, 163–163. doi:10.3389/fnhum.2017.00163.
38. Blankertz, Guido; Lemm, S.K.M.C.G.M.K.R.B.D. The Berlin brain-computer interface: Machine learning based detection of user specific brain states. *Journal of Universal Computer Science* **2007**, *12*, 581–607.
39. Lantz, M. B., D.S. Neuropsychological assessment of subjects with uncontrolled epilepsy: effects of EEG feedback training. *Epilepsia* **1988**, *29*, 163–171. doi:10.1111/j.1528-1157.1988.tb04414.x.
40. Harvey, D. *Marx, capital, and the madness of economic reason*; Oxford University Press, 2017.
41. Luciw, M.D.; Jarocka, E.; Edin, B.B. Multi-channel EEG recordings during 3,936 grasp and lift trials with varying weight and friction. *Scientific data* **2014**, *1*, 1–11.
42. Cho, H.; Ahn, M.; Ahn, S.; Kwon, M.; Jun, S.C. EEG datasets for motor imagery brain–computer interface. *GigaScience* **2017**, *6*, gix034.
43. Goldberger, A.L.; Amaral, L.A.; Glass, L.; Hausdorff, J.M.; Ivanov, P.C.; Mark, R.G.; Mietus, J.E.; Moody, G.B.; Peng, C.K.; Stanley, H.E. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *circulation* **2000**, *101*, e215–e220.
44. Kaya, M.; Binli, M.K.; Ozbay, E.; Yanar, H.; Mishchenko, Y. A large electroencephalographic motor imagery dataset for electroencephalographic brain computer interfaces. *Scientific data* **2018**, *5*, 1–16.
45. Sajda, P.; Gerson, A.; Muller, K.R.; Blankertz, B.; Parra, L. A data analysis competition to evaluate machine learning algorithms for use in brain-computer interfaces. *IEEE Transactions on neural systems and rehabilitation engineering* **2003**, *11*, 184–185.
46. Schirrmeister, R.T.; Springenberg, J.T.; Fiederer, L.D.J.; Glasstetter, M.; Eggensperger, K.; Tangermann, M.; Hutter, F.; Burgard, W.; Ball, T. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human brain mapping* **2017**, *38*, 5391–5420.
47. Bhatt, R. Planning-relax dataset for automatic classification of eeg signals. *UCI Machine Learning Repository* **2012**.
48. Koelstra, S.; Muhl, C.; Soleymani, M.; Lee, J.S.; Yazdani, A.; Ebrahimi, T.; Pun, T.; Nijholt, A.; Patras, I. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing* **2011**, *3*, 18–31.
49. Onton, J.A.; Makeig, S. High-frequency broadband modulation of electroencephalographic spectra. *Frontiers in human neuroscience* **2009**, p. 61.
50. Savran, A.; Ciftci, K.; Chanel, G.; Mota, J.; Hong Viet, L.; Sankur, B.; Akarun, L.; Caplier, A.; Rombaut, M. Emotion detection in the loop from brain signals and facial images. Proceedings of the eNTERFACE 2006 Workshop, 2006.
51. Yadava, M.; Kumar, P.; Saini, R.; Roy, P.P.; Prosad Dogra, D. Analysis of EEG signals and its application to neuromarketing. *Multimedia Tools and Applications* **2017**, *76*, 19087–19111.
52. Duan, R.N.; Zhu, J.Y.; Lu, B.L. Differential entropy feature for EEG-based emotion classification. 6th International IEEE/EMBS Conference on Neural Engineering (NER). IEEE, 2013, pp. 81–84.
53. Soleymani, M.; Lichtenauer, J.; Pun, T.; Pantic, M. A multimodal database for affect recognition and implicit tagging. *IEEE transactions on affective computing* **2011**, *3*, 42–55.
54. Faller, J.; Cummings, J.; Saproo, S.; Sajda, P. Regulation of arousal via online neurofeedback improves human performance in a demanding sensory-motor task. *Proceedings of the National Academy of Sciences* **2019**, *116*, 6482–6490.
55. Margaux, P.; Emmanuel, M.; Sébastien, D.; Olivier, B.; Jérémie, M. Objective and subjective evaluation of online error correction during P300-based spelling. *Advances in Human-Computer Interaction* **2012**, *2012*, 4–4.

56. Miller, K.J.; Schalk, G.; Hermes, D.; Ojemann, J.G.; Rao, R.P. Spontaneous decoding of the timing and content of human object perception from cortical surface recordings reveals complementary information in the event-related potential and broadband spectral change. *PLoS computational biology* **2016**, *12*, e1004660.
57. of Information, B..U.B.S. Synchronized brainwave dataset, 2019.
58. Korczowski, L.; Ostaschenko, E.; Andreev, A.; Cattan, G.; Rodrigues, P.L.C.; Gautheret, V.; Congedo, M. Brain Invaders calibration-less P300-based BCI using dry EEG electrodes Dataset (bi2014a). PhD thesis, GIPSA-lab, 2019.
59. Kappenman, E.S.; Luck, S.J. The effects of electrode impedance on data quality and statistical significance in ERP recordings. *Psychophysiology* **2010**, *47*, 888–904.
60. Cao, Z.; Chuang, C.H.; King, J.K.; Lin, C.T. Multi-channel EEG recordings during a sustained-attention driving task. *Scientific data* **2019**, *6*, 19.
61. Broderick, M.P.; Anderson, A.J.; Di Liberto, G.M.; Crosse, M.J.; Lalor, E.C. Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology* **2018**, *28*, 803–809.
62. Torkamani-Azar, M.; Kanik, S.D.; Aydin, S.; Cetin, M. Prediction of reaction time and vigilance variability from spatio-spectral features of resting-state EEG in a long sustained attention task. *IEEE journal of biomedical and health informatics* **2020**, *24*, 2550–2558.
63. Cattan, G.; Rodrigues, P.L.C.; Congedo, M. Eeg alpha waves dataset. PhD thesis, GIPSA-LAB, 2018.
64. Stober, S.; Sternin, A.; Owen, A.M.; Grahn, J.A. Towards Music Imagery Information Retrieval: Introducing the OpenMIIR Dataset of EEG Recordings from Music Perception and Imagination. ISMIR, 2015, pp. 763–769.
65. Roesler, O. UCI Machine Learning Repository: EEG Eye State Data Set. *2013* **2013**.
66. Agarwal, M.; Sivakumar, R. Blink: A fully automated unsupervised algorithm for eye-blink detection in eeg signals. 2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton). IEEE, 2019, pp. 1113–1121.
67. Rösler, O.; Suendermann, D. A first step towards eye state prediction using eeg. *Proc. of the AIHLS* **2013**, *1*, 1–4.
68. Zhao, S.; Rudzicz, F. Classifying phonological categories in imagined and articulated speech. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015, pp. 992–996.
69. Vivancos, D.; Cuesta, F. MindBigData 2022 A Large Dataset of Brain Signals. *arXiv preprint arXiv:2212.14746* **2022**.
70. Bashivan, P.; Rish, I.; Yeasin, M.; Codella, N. Learning representations from EEG with deep recurrent-convolutional neural networks. arXiv 2015. *arXiv preprint arXiv:1511.06448* **2015**.
71. Challenge Data.
72. Begleiter, H.; Ingber, L. UCI Machine Learning Repository: EEG Database Data Set, 1999.
73. Wang, H.; Li, Y.; Hu, X.; Yang, Y.; Meng, Z.; Chang, K.m. Using EEG to Improve Massive Open Online Courses Feedback Interaction. AIED Workshops, 2013.
74. Picone, J. Electroencephalography (EEG) resources.
75. Cavanagh, J.F.; Napolitano, A.; Wu, C.; Mueen, A. The patient repository for EEG data+ computational tools (PRED+ CT). *Frontiers in neuroinformatics* **2017**, *11*, 67.
76. Kappenman, E.S.; Farrens, J.L.; Zhang, W.; Stewart, A.X.; Luck, S.J. ERP CORE: An open resource for human event-related potential research. *NeuroImage* **2021**, *225*, 117465.
77. Penny, W.D.; Friston, K.J.; Ashburner, J.T.; Kiebel, S.J.; Nichols, T.E. *Statistical parametric mapping: the analysis of functional brain images*; Elsevier, 2011.
78. Du, C.; Fu, K.; Li, J.; He, H. Decoding visual neural representations by multimodal learning of brain-visual-linguistic features. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2023**, *45*, 10760–10777.
79. Doersch, C.; Gupta, A.; Efros, A.A. Unsupervised visual representation learning by context prediction. *Proceedings of the IEEE international conference on computer vision* **2015**, pp. 1422–1430.
80. Konečný, J.; McMahan, H.B.; Ramage, D.; Richtárik, P. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527* **2016**.