Decoding traces of memory during offline continuous electrical brain activity (EEG)

Dissertation

zur Erlangung des Grades eines
Doktors der Naturwissenschaften

der Mathematisch-Naturwissenschaftlichen Fakultät
und
der Medizinischen Fakultät
der Eberhard-Karls-Universität Tübingen

vorgelegt
von

Sarah Alizadeh

aus Mashhad, Iran

August – 2017

Tag der mündlichen Prüfung:	September 29, 2017
Dekan der Math.-Nat. Fakultät:	Prof. Dr. W. Rosenstiel
Dekan der Medizinischen Fakultät:	Prof. Dr. I. B. Autenrieth
1. Berichterstatter:	Prof. Dr. Steffen Gais
2. Berichterstatter:	Prof. Dr. Jan Born
Prüfungskommission:	Prof. Dr. Steffen Gais
	Prof. Dr. Jan Born
	Prof. Dr.-Ing. Moritz Grosse-Wentrup
	Prof. Dr. med. Martin Walter

Erklärung / Declaration:

Ich erkläre, dass ich die zur Promotion eingereichte Arbeit mit dem Titel:
„**Decoding traces of memory during offline continuous electrical brain activity (EEG)** “

selbständig verfasst, nur die angegebenen Quellen und Hilfsmittel benutzt und wörtlich oder inhaltlich übernommene Stellen als solche gekennzeichnet habe. Ich versichere an Eides statt, dass diese Angaben wahr sind und dass ich nichts verschwiegen habe. Mir ist bekannt, dass die falsche Abgabe einer Versicherung an Eides statt mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft wird.

I hereby declare that I have produced the work entitled “**Decoding traces of memory during offline continuous electrical brain activity (EEG)**”, submitted for the award of a doctorate, on my own (without external help), have used only the sources and aids indicated and have marked passages included from other works, whether verbatim or in content, as such. I swear upon oath that these statements are true and that I have not concealed anything. I am aware that making a false declaration under oath is punishable by a term of imprisonment of up to three years or by a fine.

Tübingen, den ....

Datum / Date Unterschrift /Signature

Abstract

Continuous electroencephalogram (EEG) provides an excellent possibility to track memory traces from brain rhythmic activity and to study the underlying neural signatures of memory processes. To do so, a promising approach is to employ multivariate pattern classification (MVPC). These methods lend themselves very well to decode the information that resides within the whole distributed spatiotemporal patterns of activity. Using these methods, it is possible to detect traces of memory during sleep or wakefulness, which will reveal valuable insights about the memory function in these brain states.

However, there are several methodological problems to decode memory traces from brain activity in paradigm-free (offline) periods. Continuous EEG is prone to elevated levels of noise and distortions and has much higher dimension than single-trial EEG, because of the longer recording time and lack of prior information about relevant time points that are informative for classification. In this case, detecting traces of memory involves searching the whole spatiotemporal feature space to find where memory representations reside. Such high-dimensional data, especially when signal-to-noise ratio and sample size are low, pose problems for classification and interpretation of MVPC result. To address these problems, in this thesis we aim: 1) to develop a proper classification algorithm that enables decoding of continuous EEG to detect memory traces in paradigm-free periods 2) to find EEG correlates of material-specific memory representations during offline periods of sleep and wakefulness, and 3) to provide a systematic method to interpret and validate the specificity of the MVPC results.

In chapter 2, we used our MVPC method to detect the 'when' and 'where' of sleep-dependent reprocessing of memory traces in humans. Although replay of neuronal activity during sleep has been shown in animal experiments, its

dynamics and underlying mechanism is still poorly understood in humans. We applied MVPC to human sleep EEG to see if the brain reprocesses previously learned information during sleep and looked for dynamics, neural signatures and relevance of different sleep stages to such process. Here, we developed a two-step classification algorithm that incorporates channel-based feature weighting as well as a tailored preprocessing scheme that is optimized to decode continuous EEG data for between-subject classification. With this method, we demonstrate that the specific content of previous learning episodes is reprocessed during post-learning sleep. We find that memory reprocessing peaks during two distinct periods in the night and both Rapid Eye Movement (REM) and non-REM sleep are involved in this process.

To detect traces of short-term memory representations, we employed MVPC in chapter 3 to test whether electrical brain activity during short-term memory maintenance satisfies the necessary conditions for mnemonic representations; i.e. coding for memory content as well as retrieval success. We found that it is possible to decode the content maintained in memory during delay period and if it is subsequently recalled mainly from temporal, parietal, and frontal areas. Importantly, the only overlap between electrodes coding for retrieval success and memory content was found in parietal electrodes, indicating that a dedicated short-term memory representation resides in parietal cortex.

Finally, chapter 4 aims at providing a systematic approach to validate the specificity of MVPC result. We investigate the consequences of the high sensitivity of MVPC for stimulus-related differences, which may confound estimation of class differences during decoding of cognitive concepts. We propose a method, which we call concept-response curve, to determine how much decoding performance is specific to the higher-order category processing and to lower-order stimulus processing. We show that this method can be used to quantify the relative contribution of concept- and stimulus-related components and to investigate the spatiotemporal dynamics of conceptual and perceptual processing.

TABLE OF CONTENTS

Chapter 1:

Synopsis

Sarah Alizadeh

Introduction

Memory processes

Memory formation and retrieval is one of the fundamental capabilities of humans, as well as other living organisms. It provides the ability to adapt behavior based on experience and is therefore essential for survival. It allows for a goal-directed behavior and has the capacity to integrate new experiences with the long-term knowledge network and make it accessible upon recall.

Memory functions comprise three essential processes: encoding, consolidation, and retrieval. Encoding is the process of getting information from perceived items into the memory, which results in formation of a new memory trace into the brain network. New memories are however labile and are susceptible to interference and forgetting. Later, a neuronal mechanism called consolidation stabilizes the new memory traces and integrates them into the pre-existing knowledge network. During consolidation, the newly encoded fresh memories are actively reprocessed and transformed into a stable state which is long-lasting and resistant to interference. During retrieval, the previously encoded memories are recalled and re-accessed by the brain. It is postulated that memory consolidation occurs most effectively during offline period of sleep while encoding and memory retrieval take place during wakefulness (Diekelmann and Born, 2010; Rasch and Born, 2013)

Today, it is evident that traces of memory, also known as engram, emerge from co-activation of one or multiple distributed brain networks which are manifested in the rhythmic electrical brain activities. Such neural oscillations are considered to promote communication of memory systems and are held to play a mechanistic role in all three aspects of memory processes (Duzel et al., 2010; Headley and Paré, 2017). Although the biological existence of engrams is accepted, the search for memory traces in the brain is still an ongoing research

and a consensus about the actual mechanism, locations of the process, and certain oscillations that code for specific aspects of memory is yet to emerge (Chadwick et al., 2010). In doing so, a major difficulty is that multiple brain areas are involved in encoding, maintaining and consolidation of memories and therefore, identifying the precise location and mechanisms involved in memory processes requires computational models that integrate information from multiple scales of temporal and spatial activities.

A promising approach to investigate memory traces, is when they are actively maintained after learning (e.g. working memory maintenance period) or when they are reactivated to be stabilized (e.g. during sleep). These offline periods of time appear to constitute critical windows during which memory traces are reprocessed, strengthened and transformed into the long-term memory representations. In this thesis, we investigate memory-related processes during post-learning offline periods of sleep and short-term memory maintenance to identify how brain reprocesses previously learned information.

Sleep and Memory reprocessing

Human sleep consists of two main stages; namely rapid-eye-movement (REM) and Non-REM sleep; which alternate and span the sleep period in a cyclic manner (see Figure 1). Both types of sleep are characterized by distinct and typical electroencephalogram (EEG) and physiologic patterns (see Figure 1B). Non-REM sleep is dominant in the first half of typical night sleep, whereas REM sleep becomes more prevalent and extensive towards the ends of sleep period.

Today, it has become clear that sleep is not a simple period of rest for the brain, but that it performs an important function for brain maintenance (Hobson, 2005). In particular, memory has been shown to benefit from sleep (Diekelmann and Born, 2010; Gais and Born, 2004; Rasch and Born, 2013). Memory recall is better after sleep than after wakefulness, and memories are more resistant to interference (Benson and Feinberg, 1975; Ekstrand et al., 1977; Plihal and Born, 1997). It has been proposed that during sleep, previously learned information is

reactivated, i.e. those neuronal circuits involved in learning/storing a certain memory become active again. Patterns of neuronal activity, similar to those during learning, occur repeatedly during the night, leading to a strengthening of the synaptic pathways involved and thus to a consolidation of the memory itself (Schwindel and McNaughton, 2011; Stickgold and Walker, 2007).

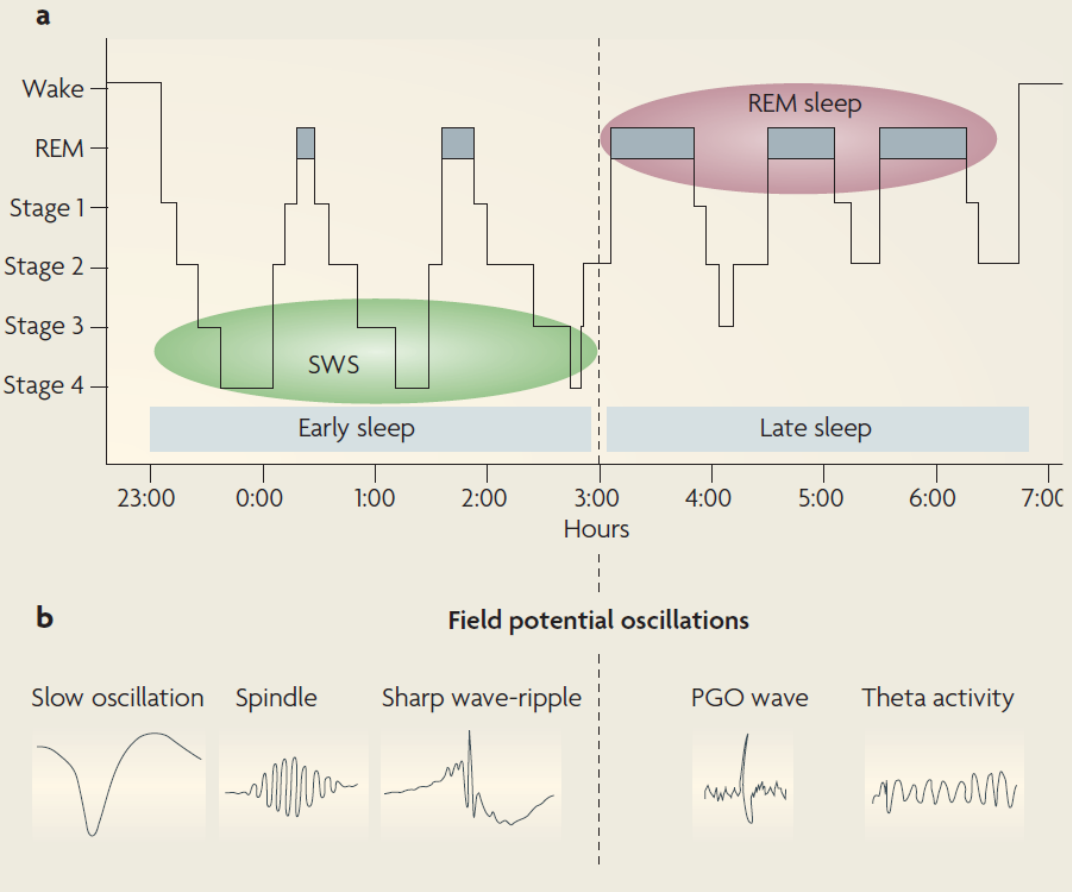


Figure 1: Typical human sleep structure (A) From electroencephalographic brain activity, sleep is characterized by the periodic patterns of rapid eye movement (REM) sleep and non-REM sleep. N-REM sleep includes slow-wave sleep (SWS, stages 3 and 4) and lighter sleep stages 1 and 2. SWS occurs predominantly in the first half of the night (early sleep), whereas REM sleep prevails in the second half of the night (late sleep) (B) Both types of sleep are determined by their specific patterns of electrical field potential. In particular, Non-REM sleep is characterized by the presence of slow oscillations, spindles and sharp wave ripples. On the other hand, REM sleep features ponto-geniculo-occipital (PGO) waves and theta activity. Figure adapted and reprinted with permission from Diekelmann and Born (2010).

Support for the idea of memory reactivation during sleep comes mostly from animal studies, which could show using single-cell recordings of neuronal activity in rats that individual neurons fire during sleep with the same correlational pattern and order as they did during previous learning (Ji and Wilson, 2007; Lee and Wilson, 2002; Wilson and Mcnaughton, 1994). This first evidence for reactivation of learning-related information in post-learning sleep was further investigated and extended by many more recent studies (Euston et al., 2007; Peyrache et al., 2009; Ribeiro et al., 2004). While these animal studies show highly specific reactivation patterns of neuronal activity, they did not show a relationship with behavioral memory performance that would indicate the functional significance of these reactivated patterns. On the other hand, human behavioral studies exist which show that an external reactivation of memories during sleep by presenting auditory or olfactory cues that were associated with learning task can lead to memory enhancement (Oudiette and Paller, 2013; Rasch et al., 2007; Rudoy et al., 2009). Moreover, some neuroimaging studies in humans using PET and fMRI have found learning-dependent off-line activity on the level of brain regions associated with learning during subsequent sleep (Peigneux et al., 2004), but also during wakefulness (Peigneux et al., 2006). These studies, however, can only determine if a certain brain area responsible for learning is active during post-learning sleep, and they neither can show if this activity is actually reflecting the content of previous learning experiences, nor whether it is actually replaying previous activity patterns (Duyn, 2012). In chapter 2, we investigated if human electrical brain activity during sleep contains information about previous learning episodes. We used electroencephalogram (EEG) to examine if material-specific memory reprocessing happens during sleep and when and how it preferentially occurs.

Short-term memory maintenance

Short-term memory is considered as a temporary buffer for holding a limited amount of information in an active and readily-available state in the absence of sensory input (Eriksson et al., 2015; Larocque et al., 2014). When we retain

information, multiple cognitive systems and brain areas are involved in active maintenance of information, in relating the information to the integrated knowledge coded in the brain, and in successfully retrieving the information from those activities. A large body of evidence from recent models propose that short-term memory maintenance results from an interaction between long-term memory representations, perceptual representations and basic processes, such as attention (D'Esposito and Postle, 2015; Eriksson et al., 2015; Jonides et al., 2008; Larocque et al., 2014). Based on this view, short-memory memory representations are linked to many distributed brain areas because those component processes that implement short-term memory involve distributed brain networks. This includes prefrontal cortex, parietal cortex, and the regions responsible for coding item-specific memory representations such as sensory areas which interact during maintenance period (Eriksson et al., 2015).

Recent studies show that persistent stimulus-related neural activation during offline periods underlie the capacity to maintain attended items (LaRocque et al., 2013; LaRocque et al., 2016; Lewis-Peacock et al., 2012), and may foster the encoding of new long-term memory representations (Olsson and Poom, 2005). However, it is still unclear if there are certain processes or brain structures unique and specific to short-term memory or whether its function emerges from combination of processes that can be explained by other terms than short-term memory. More importantly, if such a dedicated store for short-term memory existed, which brain region or processes would code such information? In chapter 3, we examined if electrical brain activity during short-term memory maintenance satisfies the mnemonic criteria, i.e. coding for memory content and retrieval success, and investigated where identified short-term memory representations reside.

Decoding memory traces from electrical brain activity using multivariate pattern classification

Oscillatory fluctuations of brain activity are held to play a mechanistic role in different aspects of memory, including encoding and maintenance of information, as well as consolidation and retrieval of stored memories (for a review see Duzel et al., 2010). Hence, electroencephalogram (EEG) which measures electrical brain activity with a high temporal resolution provides an excellent possibility to study the underlying mechanism of various memory processes. A promising approach to do so is to use multivariate pattern classification (MVPC). These methods lend themselves very well to decode the information represented within distributed activity patterns. They take into account the information that reside in the whole spatiotemporal pattern of activity, instead of looking for features that individually allow distinction between conditions. By taking the interdependencies between features into account, MVPC approaches provide increased sensitivity compared to their classical mass-univariate counterparts (for reviews see Haxby et al., 2014; Haynes, 2015; Norman et al., 2006). Moreover, for MVPC-based approaches, the problem of multiple comparison is bypassed and the generalizability of their findings does not depend on arbitrary significance thresholds and assumptions of statistical normality.

Aside from enhanced sensitivity and multivariate nature of MVPC which makes it a good fit for analyzing high-dimensional data, more arguments favor employing MVPC particularly for memory research. Importantly, MVPC assumes that neural activity is distributed over multiple brain areas, time, and frequency bands (Haynes, 2015; Pouget et al., 2000) and looks for multivariate patterns of activity that code a certain cognitive property. This complies well with the idea that neural representations of memory traces are distributed and it is the co-activation of all the sub-processes that code for a specific memory trace, in contrast to the mass-univariate approaches that rely on a local difference

between experimental conditions and try to find specific regions or time points that encode a memory.

Over the last decade, MVPC has been successfully applied to decode traces of memory from continuous brain activity in different states of consciousness. For example, it has been shown that using MVPC methods on brain activity measured by functional magnetic resonance imaging (fMRI), traces of individual episodic memories (Chadwick et al., 2010), as well as spatial memories (Hassabis et al., 2009) can be decoded from human hippocampus. Moreover, a recent study has shown that it is possible to decode the content of visual imagery occurring at sleep onset using fMRI data (Horikawa et al., 2013). Since MVPC has evolved into a quite well-established method in fMRI research, it has been employed by many studies in different ways to investigate memory function (for a comprehensive review seeRissman and Wagner, 2012). Nonetheless, these methods have recently begun to get momentum in the field of electrophysiological data as well (for example see Fuentemilla et al., 2010; Jafarpour et al., 2013; Newman and Norman, 2010).

Technically speaking, MVPC methods are a type of machine learning techniques where a classifier is trained to find a separation between neural activities belonging to different experimental conditions (see Figure 2). These methods can be understood as a four-step supervised pattern classification problem (Duda et al., 2000; Lemm et al., 2011). The first step is to extract features or attributes that quantify the neural activity with respect to experimental conditions (see Figure 2A-B). These features could be the activity of the selected voxels in fMRI data, or the amplitude of selected electrodes in certain time points or frequency power values in EEG data. After that, a classifier is selected to find a 'rule' that can correctly distinguish between conditions. In the simplest form, the decision boundary is a linear hyperplane which partitions the feature space into regions with different labels (see Figure 2E). To test the generalizability of the classifier to new unknown samples, only a part of data is used for training the classifier and the left-out part is used to test the accuracy (see Figure 2G).

Accuracy is defined by the proportion of correctly classified items and is typically estimated using cross-validation (Hastie et al., 2001). Finally, to test if the classifier can indeed extract information from the data, the resulting classification accuracy is compared to a distribution which is expected by chance. To estimate the chance distribution, the class labels are randomly relabeled and classification is repeated with the random labels, a procedure called permutation test (Nichols and Holmes, 2002).

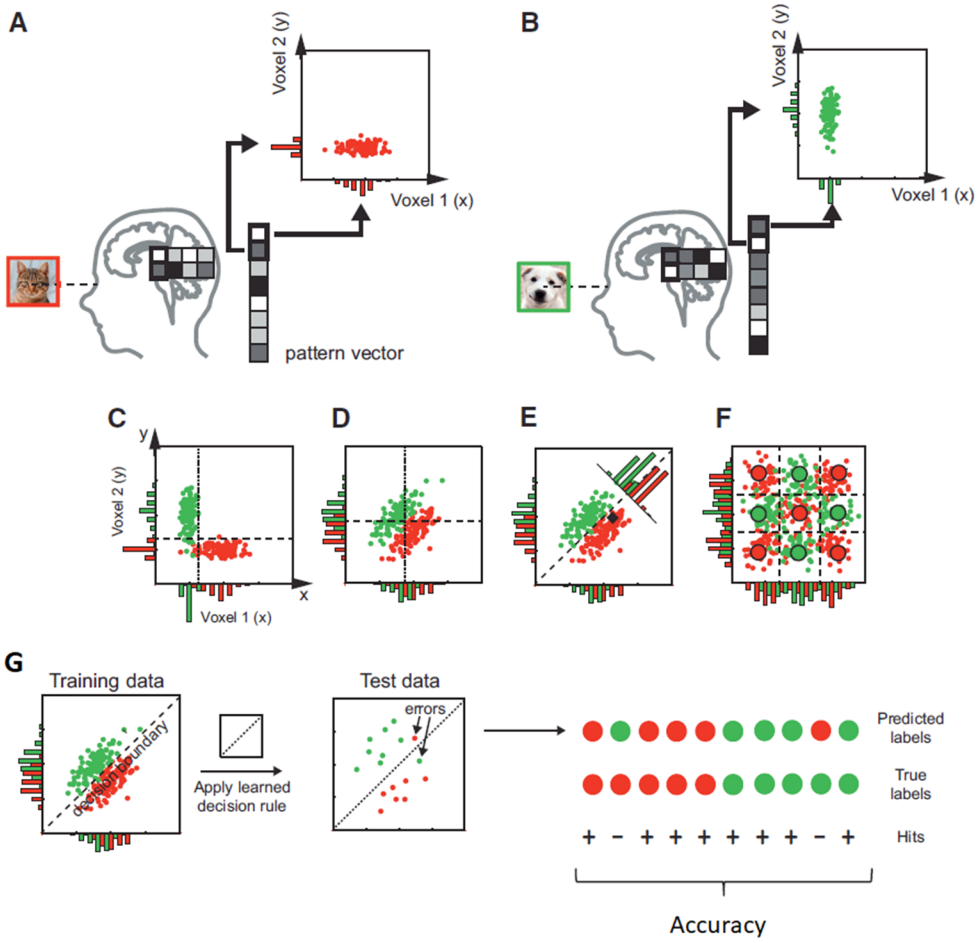


Figure 2: Basic principles of multivariate pattern classification approach. (A-B) Brain activity at different voxels when presenting different stimuli (e.g. cat or dog) can be used as features for classification. (C) When features are independent, the class effect can be detected with both univariate and multivariate methods. (D) When features are dependent, the class difference can only be detected with multivariate methods that assess both features simultaneously. (E) In this case, a linear classifier can find a linear decision boundary to distinguish two classes. Two common linear classifiers are linear

discriminant analysis (LDA; Fisher, 1936) and linear support vector machines (SVM; Duda et al., 2000; Pereira et al., 2009). (F) If linear decision boundaries cannot partition the data sufficiently well, nonlinear classifiers like k-nearest neighbor (KNN), Gaussian Naïve Base (GNB) or SVMs with nonlinear kernels can be used. (G) To test the generalizability of the classifier to new samples, cross-validation is typically done, where the data is randomly divided into multiple disjoint subsets of approximately equal size. The classifier is then trained repeatedly on the all except one partitions and tested on the remaining left-out subset. This results in predicted labels which can be compared with the true label to get the accuracy of classification. If there was enough information in the training set to make a proper class distinction, the resulting accuracy will be above chance level. Figure adapted and reprinted with permission from Haynes (2015).

Challenges of using MVPC for decoding memory traces

Technical challenges for decoding offline periods

Next to the many theoretical advantages of MVPC, there are a number of practical problems when using MVPC-based approaches for decoding memory traces from continuous electrical brain activity in offline periods.

First problem concerns the choice of the MVPC algorithm and the preprocessing steps necessary for decoding continuous EEG time series. There are currently only a few studies that used multivariate pattern classification on EEG data to analyze ongoing brain activity in offline periods. In EEG community, MVPC has been mainly applied to brain computer interfaces (BCI), motor imagery, or analysis of event related potentials (ERPs). However, because decoding continuous EEG in paradigm-free periods has different prerequisites compared to decoding trials with a clear onset of events, those approaches are often not helpful to decode spontaneous brain activity that is not induced by external stimuli. Importantly, continuous EEG has potentially much higher dimension than single-trial EEG, because of the longer recording time and lack of prior information about relevant time points that are informative for classification. Therefore, detecting traces of memory involves searching the whole

spatiotemporal feature space to find where memory representations reside. Such high-dimensional data, especially when sample size is small, pose specific problems for MVPC (Fan and Fan, 2008; Jamalabadi et al., 2016), which should be accommodated using proper preprocessing and classification algorithms.

In addition, EEG signals are prone to elevated levels of noise, missing data, and outliers that make the effective usage of EEG data difficult, especially in paradigm-free periods when the search space is huge and there is no clear onset of events in the signal. The analysis of EEG suffers from the abundance of irrelevant brain activities as well as multiple sources of noise and distortions which make generalization of signals over subjects a challenging task. Furthermore, since EEG signals recorded from different subjects show marked differences, the MVPC algorithm needs to deal with additional subject variabilities that are not related to the experimental conditions. Such variations, especially in low sample size and low effect size data, can explain most of the variance in the data and therefore, their effect should be carefully accommodated.

Challenges regarding interpretation of the MVPC results

Another critical issue is that when there are many features to be fitted and only a small number of samples are available (a case which is common in neuroimaging datasets), it is often possible to find a good fit for a certain sample of data, which, however, cannot be generalized to unknown new samples. Therefore, it is a widespread practice to validate the classifier using a test set which is statistically independent from training data and compare the accuracy to the distribution which is expected from chance. To estimate the chance distribution, permutation tests are often done, where the assignments between samples and class labels are randomly permuted and classification is repeated with random labels (Nichols and Holmes, 2002; Winkler et al., 2014). However, such permutation tests are designed only to test if decoding accuracy is indeed beyond chance level and cannot assess the source of information that classifier

has used to discriminate between experimental conditions. Importantly, when decoding traces of semantic memories or cognitive concepts, we are interested in identifying category-related memory representations that can generalize over items belonging to a category, rather than exemplar-specific representations. Therefore, it is important to determine which source of information, i.e. category-specific or stimulus-related differences, gives rise to decoding performance. Although MVPC-based approaches can resolve if there is information in the data about the experimental condition, they are inherently limited in their ability to specify the source of information that made the distinction possible. At the moment, it still remains an ongoing endeavor to provide guidelines and new methods that enable a more thorough interpretation concerning the specificity of MVPC results (Schreiber and Krekelberg, 2013).

Aims of this thesis

Following the challenges and motivated by the potentials of using MVPC for decoding memory traces, the main aim of this thesis is threefold: 1) to develop a proper classification algorithm that enables decoding of continuous EEG to detect memory traces in paradigm-free periods 2) to find EEG correlates of material-specific memory representations during sleep and wakefulness, and 3) to provide a systematic method to interpret and validate the specificity of the MVPC results.

In the last two decades, a vast amount of research has been conducted to investigate various aspects of memory reactivation during sleep and its role for consolidation of memories. Although sleep-dependent memory reactivation has been well studied in animals (Ji and Wilson, 2007; Louie and Wilson, 2001; Peyrache et al., 2009; Wilson and Mcnaughton, 1994), due to the restrictions of invasive imaging techniques, its mechanism is still poorly understood in humans. In particular, there is no systematic study yet that shows the dynamics

of learning-related memory reprocessing during sleep and if this re-expression of learning-related activity during sleep contains information about the specific content of a previous learning task. In chapter 2, we address these questions by applying multivariate pattern classification on human sleep EEG, to see whether electrical brain activity during sleep contains information about previously learned material. We hypothesized that if memory content is reactivated/reprocessed during sleep, it must be possible to determine, solely based on sleep EEG, which kind of material was learned before sleep. We used MVPC to test this hypothesis and to study the EEG correlates and dynamics of memory reprocessing during sleep and the relevance of different sleep stages to this process.

In chapter 3, we applied MVPC to EEG recordings during offline short-term memory maintenance period to see if a dedicated storage for short-term memory exists and where the corresponding memory representations reside. For that, we used MVPC to test whether electrical brain activity recorded during maintenance interval satisfies the necessary conditions of a mnemonic representation; namely coding for the specific memory content and the retrieval success upon recall. If activity in a brain region predicted subsequent memory performance and additionally carried information about the content kept in memory, it would be compelling evidence for a dedicated short-term memory storage. We employed MVPC to investigate brain areas and oscillations that separately code for retrieval success or memory content during maintenance period and identified those overlapping regions which would be potentially suited to harbor short-term memory representations.

Chapter 4 mainly aims at providing a systematic approach to validate the specificity of the MVPC results. Although MVPC is a statistically powerful and robust technique to study cognitive mental states (Kamitani and Tong, 2005; Tong and Pratte, 2012), its complexity can lead to important methodological and conceptual issues. Since these methods are designed to leverage all the information contained in the brain activity, any stimulus-related differences

between individual elements of the categories (e.g. orientation, shape, color, etc.) can drive the decoding performance to higher than chance, even if there is no overall difference between categories. In fact, MVPC is sensitive to both the effect of interest and to any other confounding factors that drive a difference between conditions. In chapter 4, we explore the consequences of the high sensitivity of MVPC for stimulus-related differences, which may confound estimation of class differences during decoding of cognitive concepts. We propose a systematic approach to determine the degree to which decoding performance is specific to the higher order category processing or lower order stimulus processing. We used this method to quantify the relative contribution of these two components and to investigate the spatiotemporal dynamics of conceptual and perceptual processing.

Conclusions and general discussion

A dedicated two-step classification algorithm to decode memory traces from continuous EEG

When decoding continuous EEG, we identified three main problems that restrain MVPC performance: 1) low signal-to-noise ratio, 2) large variability of EEG signal between subjects, 3) high dimensionality of the recorded data. Together, these problems lead to overfitting and instability of classification accuracies. To overcome these challenges, we developed a two-step procedure that uses channel-based feature weighting and independent sample validation as well as a tailored preprocessing scheme that is optimized to decode continuous EEG data for a between-subject classification (Schönauer et al., 2017; Schönauer et al., in prep).

We used power spectral density as the representational feature space to track memory traces. Spectral features provide a concise data representation which directly relates to brain rhythmic activity and is more comparable between subjects than EEG time series. However, even in frequency domain, EEG signals

from different subjects show marked differences which are often larger than the size of effects induced by experimental conditions. If neglected, this subject variability significantly reduces signal-to-noise ratio and the performance of between-subject classification. Because subject-specific baseline of EEG spectra remains fairly constant regardless of experimental conditions, we used spectral sharpening filter to remove this baseline, thus emphasizing the between-condition spectral differences. This significantly increases increased signal-to-noise ratio and makes data belonging to different subjects more comparable.

Aside from large subject variability, dimensionality is another factor that limits MVPC performance. When there are too many features compared to the number of subjects, like in continuous EEG, the classifier becomes instable (e.g. covariance estimation becomes systematically distorted in LDA Blankertz et al., 2011), resulting in prediction accuracies approaching chance (see also Hall et al., 2005). In our method, the input signal to the classifier was spatially down-sampled from 128 to 32 channels by averaging over neighboring electrodes which decreases the number of redundant features, increases signal-to-noise ratio, and further increases spatial similarity for the comparison between subjects.

Even after down sampling electrodes, the number of features remain too high (> 1000) compared to the number of subjects available (< 50). To circumvent this problem, we developed a stepwise classification procedure in which the spatial and spectral features are separated and used in two successive stages. Specifically, we first perform a channel-based classification and estimate the accuracy based on each channel. In the second step, we use the resulting accuracies of the first step to train another linear SVM based on a weighted average of data from all channels. Doing this, the parameters of the final hyperplane are essentially the product of two factors coding for the information either in time or in space. this procedure helps to avoid overfitting because the number of features in each step is effectively in the order of the number of subjects (samples).

With this method, only based on electrical brain activity, we could successfully decode 1) time course of memory reprocessing during sleep, 2) memory retrieval success, and 3) memory content in short-term maintenance interval. Importantly, in all three cases, the significant classification accuracy on validation data set was very close to the training accuracy which confirm lack of overfitting. In addition, the classification scores showed significant correlation with the behavioral performance in each case which further supports the relevance of the neural pattern found by our algorithm to the encoded memories.

Decoding reprocessing of memory traces during sleep

In chapter 2, we investigated human sleep EEG to see if the brain reprocesses previously learned information during sleep and looked for neural signatures of such process. To detect such reprocessing of material-specific memory traces, we employed an indirect approach using multivariate pattern classification. We hypothesized that if the content of memory is reprocessed during sleep, it should therefore be possible to distinguish between EEG recordings from nights that were preceded by different learning situations. If a classifier can detect such distinctive patterns in sleep recording to correctly predict the foregoing learned material, this can be taken as a sign of active reprocessing of learning-related information during sleep.

We employed MVPC on sleep EEG data that was recorded after participants learned pictures of either faces or houses. We found that electrical brain activity during sleep contains information about the types of visual stimuli that was learned before sleep, indicating that material-specific traces of memory are reprocessed during sleep. Using MVPC, we showed for the first time that our unconscious brain's activity directly reflects what we consciously learned before sleep. With the help of pattern classification algorithms, we traced the dynamics and neural correlates of memory reprocessing during sleep and its relation to subsequent memory performance. By linking sleeping brain activity with the

content of previous learning experience, our findings bridge studies from multicell recordings in animals, showing learning related reactivation, and human imaging studies, showing reactivation of brain regions during sleep (Schönauer et al., 2017).

Temporal dynamics of memory reprocessing during sleep

The benefits of MVPC combined with temporal precision of EEG enabled us to have a more fine-grained look at the timing underlying reprocessing in sleep. Using a time-resolved analysis, we found that the classifier detects generalizable learning-related information during two distinct periods of the night, three and six hours after learning, during which memory processing exhibits peaks at all sleep stages (both REM and NREM sleep). These are periods of the night, during which brain processing seems to be more strongly related to previous learning, whereas during others, no learning-related information can be detected. Importantly, these windows are congruous with periods of synaptic plasticity and "memory consolidation windows" that have been shown previously in animals (Davis, 2011; Igaz et al., 2002). In particular, this finding is also consistent with the concept of sleep windows, specific periods during which sleep has to occur after learning to strengthen memory. If sleep is prevented during these periods, memory performance deteriorates (Smith, 1995, 2001). Whether this consolidation window depends on learning or sleep onset cannot be determined by our data, but previous experiments indicate a dependency on the time after learning (Smith, 1995). Reprocessing, however, is cyclic in nature, initiates selectively at specific time points during sleep, and its occurrence depends more on timing than on sleep stages.

EEG correlates of memory reprocessing during sleep

In addition to the time course of memory reprocessing, sleep EEG can be explored regarding the frequency and spatial features that are most predictive for reprocessing of the previous learned content. By looking at the classification

weights, we found that different frequency bands have a different relative contribution to classification in different sleep stages and that the relevance of each frequency band for memory reprocessing varies depending on the spatial location in each sleep stage.

Slow frequencies (below 4 Hz) is relevant in both REM and NREM sleep. However, the topography of the related activities strongly differs in these sleep stages. In NREM sleep, frontal slow-wave activity is predictive for classification, whereas central slow frequencies have higher discriminative power in REM sleep, speaking for a different slow-wave-related process in REM than in NREM sleep. Sleep spindles (12-16 Hz) can distinguish previous learning conditions only in NREM sleep and is localized in parieto-temporal electrodes. This complies well with the previous findings that show sleep spindles increase after learning (Scholz et al., 2009) and correlate with subsequent memory performance (Schabus et al., 2004). On the other hand, frontal and temporal theta-band activity (4-8 Hz) shows relatively higher importance in REM sleep than in the other sleep stages. This supports older hypotheses about the role of REM sleep theta in memory processing that have only recently again received renewed attention (Grosmark et al., 2012; Walker and van der Helm, 2009).

Relation between classifier prediction and subsequent memory performance

We tested the relation between overnight change in memory performance after sleep, and the classifier performance which shows the strength of memory reprocessing during sleep. Interestingly, behavioral performance, i.e. overnight memory retention, was positively correlated with the strength of memory reprocessing in slow-wave sleep (SWS), which was inferred from the classification probability estimates provided by the classifier. We did not find this association for memory reprocessing during light sleep (sleep stage 2) and REM sleep. This finding is in line with the previous studies which show no behavioral benefit of external memory reactivation during REM sleep (Rasch et

al., 2007). On the other hand, the differential significance of memory reprocessing for behavioral performance between REM and NREM sleep stages indicates different functions of reprocessing during REM sleep and during SWS for memory consolidation.

Decoding memory traces during short-term memory maintenance interval

To detect traces of short-term memory representations, we employed multivariate pattern classification in chapter 3 to test whether electrical brain activity during short-term memory maintenance interval satisfies the necessary conditions for mnemonic representations; i.e. coding for memory content (stimulus specificity) as well as retrieval success (relation to performance). More specifically, we used MVPC to test whether we can predict solely based on EEG during maintenance interval (1) what kind of stimulus is maintained during the delay period, and (2) if the content of memory will be successfully recalled afterwards. For that, we used two types of Sternberg task (i.e. a short-term memory task), once with faces and houses stimuli which recruits maintenance of visual information and once with digits and letters stimuli which involves verbal rehearsal of information during maintenance period. We showed that the subsequent retrieval success can be reliably predicted across subjects for both short-term memory tasks. In addition, we can successfully decode if participants maintained pictures of faces or houses during the delay period. Interestingly, the ability to decode memory content positively correlated with the retrieval success of the participants, speaking for a causal relationship between strong and faithful memory reprocessing during retention and the success of memory maintenance (Schönauer et al., in prep).

EEG correlates of retrieval success

Using spatial as well as frequency band-based searchlight analyses, we found that retrieval success was mainly coded in the frontal and parietal areas,

regardless of the type of the content held in memory (Schönauer et al., in prep). In frontal areas, higher frequency activity in the beta and gamma band was informative about whether a trial was subsequently remembered. Similarly, beta and gamma as well as alpha activity in the medial parietal areas were predictive for successful memory maintenance.

Importantly, we found that frontal areas are involved in successful retention of both types of information, but are not predictive for memory content. This finding is consistent with previous studies that show frontal activity reflects memory-related control processes that are independent of the material content maintained in memory (deBettencourt et al., 2017; Sreenivasan et al., 2014). Based on this fact and our finding on the contribution of frontal beta for successful retrieval of both visuospatial and verbalizable materials, we therefore propose that frontal activity in beta band represents a domain-general mechanism which is functionally important for control of short-term memory processes.

EEG correlates of memory content

We could decode the type of visual stimuli (faces or houses) held in memory from temporal and medial parietal regions, with several informative channels also reaching into lateral occipital areas (Schönauer et al., in prep). In temporal cortex, only information in the beta band was predictive for short-term memory content, whereas oscillatory activity in theta as well as beta and gamma over medial parietal cortex held information about the material content kept in short-term memory.

Temporal and lateral occipital regions have been shown to be associated with the processing of category-specific visual information from images of faces and houses (Han et al., 2013; Jacques et al., 2016; Vuilleumier et al., 2001). Since activity in these regions was not informative about retrieval success, we suggest that these sensory processing areas harbor the relevant content-related information and their activity reflects a reinstatement of the sensory

information associated with the content that is maintained in short-term memory.

A dedicated storage for short-term memory representations

We found that frontal and parietal areas are predictive for subsequent memory retrieval, whereas temporal and medial parietal regions contain information about the short-term memory content. The only overlap between those regions that code simultaneously for retrieval success and memory content was found over the medial parietal areas. Therefore, we propose that a dedicated short-term memory representation resides in medial parietal cortex, where both mnemonic criteria are satisfied (Schönauer et al., in prep). This result is in-line with the recent literature that this region harbors item-specific memory representations (Brodt et al., 2016; Ester et al., 2015; Gilmore et al., 2015).

Controlling for nuisance variance when decoding cognitive concepts

Although MVPC is a sensitive and successful method to study cognitive mental states, its increased sensitivity makes it susceptible to any confounding factors that drives a difference between conditions (Todd et al., 2013; Woolgar et al., 2014). In contrast to classical statistical analyses where random effects average out in the group mean, the multivariate nature of MVPC allows differences to accumulate over dimensions (Fan and Fan, 2008; Jamalabadi et al., 2016). Therefore, any differences between individual items of categories (e.g. physical properties, familiarity, emotionality, etc.) can contribute to the discrimination power of the classifier, even if the categories themselves are not different. This susceptibility to nuisance effects is a major concern for MVPC, because it can lead to significant bias and higher than chance classification accuracy, even when the effect of interest is nonexistent (Alizadeh et al., 2017; Jamalabadi et al., in prep).

The high sensitivity of MVPC for nuisance effects has the important consequence that it is not clear which source of information, i.e. concept-related or stimulus-

specific feature, gives rise to decoding performance. Therefore, the specificity of the results to the concept under investigation remains unclear. To make sure that decoding results do not reflect nuisance effects, we proposed a method that can separate the actual concept-related effect form other nuisance factors, thus allowing for a correct interpretation of the source of MVPC results (Alizadeh et al., 2017). Inspired by dose-response curves, our method systematically manipulates the amount of concept-related information in the data using blocked permutation test while the stimulus-related concept-irrelevant factors are held constant. This results in a concept-response curve which shows how the performance of the classifier changes with varying levels of conceptual information. The shape of concept-response curve determines if significant nuisance effects are present in the data and if the primary effect of interest goes significantly beyond these effects.

Our results suggest that nuisance effects should be a general concern for all neuroimaging studies where there are differences between subgroups of trials that lead to existence of subclasses nested within each category. Nested subclasses can exist e.g. if several groups of trials are combined into one class, if stimuli or types of stimuli are presented repeatedly, or if multiple subjects or experimental sessions are included in one analysis. Importantly, it is usually difficult to account for confounds induced by nested subclasses because these nuisance effects are not systematic and cannot be avoided experimentally. Here, concept-response curves can help to quantify the contribution of nuisance variance induced by subclasses, by taking the structure of data into account. In addition, by introducing different experimental factors as subclasses, concept-response curves can be used to distinguish the effect of several factors of interest to classification.

In addition to the benefit of concept-response curves for correct interpretation of experimental results, our method makes it possible to separate the neural correlates of higher-level cognitive processing of concepts from lower-level stimulus processing. Providing such information is a challenging task because it

requires to fully disentangle, based on neural activity, which spatiotemporal aspects of data involve concept-related or stimulus-specific processes. Even if classification is possible with high accuracy, it will be questionable whether decoding of the concept was achieved on a purely conceptual or perceptual level (Murphy et al., 2011; Simanova et al., 2010; Wurm et al., 2015). Our method provides a solution for this question. By considering time-resolved windows of neural activity, concept-response curves can characterize the temporal dynamics of conceptual and perceptual information processing. This is particularly important because such effects often cannot be separated experimentally. In addition, our method can provide fine grained details about timing and spatial sites of information specific to each process.

Limitations and outlook

In this thesis, we developed a multivariate pattern classification algorithm to decode traces of memory from offline continuous EEG. We tracked temporal dynamics of material-specific memory reprocessing during sleep and found EEG correlates of retrieval success as well as content of short-term memory during memory maintenance interval. In addition, we investigated specificity of MVPC results, and provided a systematic approach to separate higher-level cognitive processing from lower-level stimulus processing and tracked the time course of the corresponding conceptual and perceptual processes. In relation to the current endeavor, we recognize two related open questions which still need to be investigated. However, they go beyond the scope of this thesis and can be considered interesting follow ups for this research.

Looking for reactivation: decoding from wakefulness to sleep EEG

The observation that spatiotemporal patterns of neural activity in hippocampus during exploration of a novel environment is re-activated during post-learning sleep is concretely shown in animals (Ji and Wilson, 2007; Lee and Wilson, 2002;

Wilson and Mcnaughton, 1994). However, given the lack of flexible intracranial recordings in humans, it is more difficult to demonstrate sleep-dependent memory reactivation directly in humans. In this thesis, we showed that using an indirect between-subject classification approach we can detect information pertaining to a previous learning experience in sleep data. Using an approach that trains and tests the classifier in the same state of consciousness enabled us to detect material-specific memory reprocessing (but not reactivation) during sleep and study its dynamics and relation to later behavioral performance.

Nonetheless, it remains an important problem to classify data from wakeful encoding to sleep EEG which would directly show memory reactivation in humans. For that, one would need to search for actual learning-related similarities between wake and sleep EEG to find replay of those patterns of activity that are specific to encoding. This is a tedious task because EEG activity differs fundamentally between wakefulness and sleep regarding amplitudes and frequencies. More importantly, replay of neuronal firing patterns may be compressed in time or otherwise transformed in time or space compared to the actual learning trials (Diekelmann and Born, 2010; Ji and Wilson, 2007). Thus, a direct search to find memory trace reactivation by comparing the changes in the power spectral density or amplitude of learning trials to other states of the mind becomes practically impossible. It would therefore be necessary to develop a set of invariant features that can be used in sleep as well as in wakefulness.

From decoding accuracy to accuracy maps

Next to the decoding accuracy which identifies whether two sets of data contain systematic differences, we are often more interested to know which subset of brain activity contains the most relevant information about experimental conditions. It is sensible to expect that machine learning algorithms can be used not only to decide whether a particular set of data contains information about a specific question, but also to provide insights about which part of the data was used to reach that decision. However, it is becoming clearer that multivariate

methods are not optimal to answer such univariate questions. That is, MVPC at the moment is not better than univariate methods in pinpointing which single feature of the two sets actually differs.

Unfortunately, it is difficult to relate classification accuracy to a subset of features (e.g. specific frequency at specific location), which would be more in line with the nature of typical univariate methods. While the pattern of MVPC result as a whole is significant, it is challenging to name individual data features that contribute to successful classification. At the moment, there are mainly three approaches in the MVPC literature to investigate the relevant aspects of data; namely searchlight analyses (Kriegeskorte et al., 2006), classifier weights (Haufe et al., 2014), and permutation based approaches (Ojala and Garriga, 2010). However, each of these methods has its own pros and cons and their result is often different from each other which causes confusion and difficulties regarding interpretation. The main problem is that when MVPC is used, it is not even necessary that a single feature contains class-related information to be an asset to increase classification accuracy. That is, a feature might not be even informative on its own but contributes to classification simply because it contains information about the structure of the noise and hence can de-noise other class-related features (Blankertz et al., 2011). Such problems make further complications in interpreting accuracy maps, especially in the presence of thousands of features as in our EEG data. Therefore, I think it is very important to conduct new studies to systematically investigate different approaches for estimating accuracy maps and to provide a robust algorithm to do so.

References

Alizadeh, S., Jamalabadi, H., Schonauer, M., Leibold, C., Gais, S., 2017. Decoding cognitive concepts from neuroimaging data using multivariate pattern analysis. NeuroImage, 159:449-458.

Benson, K., Feinberg, I., 1975. Sleep and memory: retention 8 and 24 hours after initial learning. Psychophysiology 12, 192-195.

Blankertz, B., Lemm, S., Treder, M., Haufe, S., Muller, K.R., 2011. Single-trial analysis and classification of ERP components--a tutorial. Neuroimage 56, 814-825.

Brodt, S., Pohlchen, D., Flanagin, V.L., Glasauer, S., Gais, S., Schonauer, M., 2016. Rapid and independent memory formation in the parietal cortex. Proc Natl Acad Sci U S A 113, 13251-13256.

Chadwick, M.J., Hassabis, D., Weiskopf, N., Maguire, E.A., 2010. Decoding Individual Episodic Memory Traces in the Human Hippocampus. Current Biology 20, 544-547.

D'Esposito, M., Postle, B.R., 2015. The cognitive neuroscience of working memory. Annual Review of Psychology 66, 115-142.

Davis, R.L., 2011. Traces of Drosophila Memory. Neuron 70, 8-19.

deBettencourt, M.T., Norman, K.A., Turk-Browne, N.B., 2017. Forgetting from lapses of sustained attention. Psychon Bull Rev.

Diekelmann, S., Born, J., 2010. The memory function of sleep. Nature Reviews: Neuroscience 11, 114-126.

Duda, R.O., Hart, P.E., Stork, D.G., 2000. Pattern Classification (2nd Edition). Wiley-Interscience.

Duyn, J.H., 2012. EEG-fMRI methods for the study of brain networks during sleep. Frontiers in Neurology 3.

Duzel, E., Penny, W.D., Burgess, N., 2010. Brain oscillations and memory. Current Opinion in Neurobiology 20, 143-149.

Ekstrand, B.R., Barrett, T.R., West, J.N., Maier, W.G., 1977. The effect of sleep on human long-term memory. Neurobiology of sleep and memory, 419-438.

Eriksson, J., Vogel, E.K., Lansner, A., Bergstrom, F., Nyberg, L., 2015. Neurocognitive Architecture of Working Memory. Neuron 88, 33-46.

Ester, E.F., Sprague, T.C., Serences, J.T., 2015. Parietal and Frontal Cortex Encode Stimulus-Specific Mnemonic Representations during Visual Working Memory. Neuron 87, 893-905.

Euston, D.R., Tatsuno, M., McNaughton, B.L., 2007. Fast-forward playback of recent memory sequences in prefrontal cortex during sleep. Science 318, 1147-1150.

Fan, J.Q., Fan, Y.Y., 2008. High-Dimensional Classification Using Features Annealed Independence Rules. Annals of Statistics 36, 2605-2637.

Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. Annals of Eugenics 7, 179-188.

Fuentemilla, L., Penny, W.D., Cashdollar, N., Bunzeck, N., Duzel, E., 2010. Theta-coupled periodic replay in working memory. Current Biology 20, 606-612.

Gais, S., Born, J., 2004. Declarative memory consolidation: mechanisms acting during human sleep. Learning and Memory 11, 679-685.

Gilmore, A.W., Nelson, S.M., McDermott, K.B., 2015. A parietal memory network revealed by multiple MRI methods. Trends Cogn Sci 19, 534-543.

Grosmark, A.D., Mizuseki, K., Pastalkova, E., Diba, K., Buzsaki, G., 2012. REM sleep reorganizes hippocampal excitability. Neuron 75, 1001-1007.

Hall, P., Marron, J.S., Neeman, A., 2005. Geometric representation of high dimension, low sample size data. Journal of the Royal Statistical Society Series B-Statistical Methodology 67, 427-444.

Han, X., Berg, A.C., Oh, H., Samaras, D., Leung, H.C., 2013. Multi-voxel pattern analysis of selective representation of visual working memory in ventral temporal and occipital regions. Neuroimage 73, 8-15.

Hassabis, D., Chu, C., Rees, G., Weiskopf, N., Molyneux, P.D., Maguire, E.A., 2009. Decoding neuronal ensembles in the human hippocampus. Current Biology 19, 546-554.

Hastie, T., Tibshirani, R., Friedman, J.H., 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

Haufe, S., Meinecke, F., Gorgen, K., Dahne, S., Haynes, J.D., Blankertz, B., Biessgmann, F., 2014. On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage 87, 96-110.

Haxby, J.V., Connolly, A.C., Guntupalli, J.S., 2014. Decoding neural representational spaces using multivariate pattern analysis. Annual Review of Neuroscience 37, 435-456.

Haynes, J.D., 2015. A Primer on Pattern-Based Approaches to fMRI: Principles, Pitfalls, and Perspectives. Neuron 87, 257-270.

Headley, D.B., Paré, D., 2017. Common oscillatory mechanisms across multiple memory systems. npj Science of Learning 2, 1.

Hobson, J.A., 2005. Sleep is of the brain, by the brain and for the brain. Nature 437, 1254-1256.

Horikawa, T., Tamaki, M., Miyawaki, Y., Kamitani, Y., 2013. Neural Decoding of Visual Imagery During Sleep. Science 340, 639-642.

Igaz, L.M., Vianna, M.R., Medina, J.H., Izquierdo, I., 2002. Two time periods of hippocampal mRNA synthesis are required for memory consolidation of fear-motivated learning. Journal of Neuroscience 22, 6781-6789.

Jacques, C., Retter, T.L., Rossion, B., 2016. A single glance at natural face images generate larger and qualitatively different category-selective spatio-temporal signatures than other ecologically-relevant categories in the human brain. Neuroimage 137, 21-33.

Jafarpour, A., Horner, A.J., Fuentemilla, L., Penny, W.D., Duzel, E., 2013. Decoding oscillatory representations and mechanisms in memory. Neuropsychologia 51, 772-780.

Jamalabadi, H., Alizadeh, S., Schonauer, M., Leibold, C., Gais, S., 2016. Classification based hypothesis testing in neuroscience: Below-chance level classification rates and overlooked statistical properties of linear parametric classifiers. Human Brain Mapping 37, 1842-1855.

Jamalabadi, H., Alizadeh, S., Schonauer, M., Leibold, C., Gais, S. (In prep.) Adjusting permutation tests in multivariate anaylsis of neuroimaging data with subclasses: introduction of biased null distributions.

Ji, D.Y., Wilson, M.A., 2007. Coordinated memory replay in the visual cortex and hippocampus during sleep. Nature Neuroscience 10, 100-107.

Jonides, J., Lewis, R.L., Nee, D.E., Lustig, C.A., Berman, M.G., Moore, K.S., 2008. The mind and brain of short-term memory. Annual Review of Psychology 59, 193-224.

Kamitani, Y., Tong, F., 2005. Decoding the visual and subjective contents of the human brain. Nature Neuroscience 8, 679-685.

Kriegeskorte, N., Goebel, R., Bandettini, P., 2006. Information-based functional brain mapping. Proceedings of the National Academy of Sciences of the United States of America 103, 3863-3868.

LaRocque, J.J., Lewis-Peacock, J.A., Drysdale, A.T., Oberauer, K., Postle, B.R., 2013. Decoding attended information in short-term memory: an EEG study. Journal of Cognitive Neuroscience 25, 127-142.

Larocque, J.J., Lewis-Peacock, J.A., Postle, B.R., 2014. Multiple neural states of representation in short-term memory? It's a matter of attention. Front Hum Neurosci 8, 5.

LaRocque, J.J., Riggall, A.C., Emrich, S.M., Postle, B.R., 2016. Within-Category Decoding of Information in Different Attentional States in Short-Term Memory. Cerebral Cortex.

Lee, A.K., Wilson, M.A., 2002. Memory of sequential experience in the hippocampus during slow wave sleep. Neuron 36, 1183-1194.

Lemm, S., Blankertz, B., Dickhaus, T., Muller, K.R., 2011. Introduction to machine learning for brain imaging. Neuroimage 56, 387-399.

Lewis-Peacock, J.A., Drysdale, A.T., Oberauer, K., Postle, B.R., 2012. Neural Evidence for a Distinction between Short-term Memory and the Focus of Attention. Journal of Cognitive Neuroscience 24, 61-79.

Louie, K., Wilson, M.A., 2001. Temporally structured replay of awake hippocampal ensemble activity during rapid eye movement sleep. Neuron 29, 145-156.

Murphy, B., Poesio, M., Bovolo, F., Bruzzone, L., Dalponte, M., Lakany, H., 2011. EEG decoding of semantic category reveals distributed representations for single concepts. Brain and Language 117, 12-22.

Newman, E.L., Norman, K.A., 2010. Moderate excitation leads to weakening of perceptual representations. Cerebral Cortex 20, 2760-2770.

Nichols, T.E., Holmes, A.P., 2002. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Human Brain Mapping 15, 1-25.

Norman, K.A., Polyn, S.M., Detre, G.J., Haxby, J.V., 2006. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends in Cognitive Sciences 10, 424-430.

Ojala, M., Garriga, G.C., 2010. Permutation Tests for Studying Classifier Performance. Journal of Machine Learning Research 11, 1833-1863.

Olsson, H., Poom, L., 2005. Visual memory needs categories. Proceedings of the National Academy of Sciences of the United States of America 102, 8776-8780.

Oudiette, D., Paller, K.A., 2013. Upgrading the sleeping brain with targeted memory reactivation. Trends Cogn Sci 17, 142-149.

Peigneux, P., Laureys, S., Fuchs, S., Collette, F., Perrin, F., Reggers, J., Phillips, C., Degueldre, C., Del Fiore, G., Aerts, J., Luxen, A., Maquet, P., 2004. Are spatial memories strengthened in the human hippocampus during slow wave sleep? Neuron 44, 535-545.

Peigneux, P., Orban, P., Balteau, E., Degueldre, C., Luxen, A., Laureys, S., Maquet, P., 2006. Offline persistence of memory-related cerebral activity during active wakefulness. Plos Biology 4, 647-658.

Pereira, F., Mitchell, T., Botvinick, M., 2009. Machine learning classifiers and fMRI: a tutorial overview. Neuroimage 45, S199-209.

Peyrache, A., Khamassi, M., Benchenane, K., Wiener, S.I., Battaglia, F.P., 2009. Replay of rule-learning related neural patterns in the prefrontal cortex during sleep. Nature Neuroscience 12, 919-926.

Plihal, W., Born, J., 1997. Effects of early and late nocturnal sleep on declarative and procedural memory. Journal of Cognitive Neuroscience 9, 534-547.

Pouget, A., Dayan, P., Zemel, R., 2000. Information processing with population codes. Nature Reviews Neuroscience 1, 125-132.

Rasch, B., Born, J., 2013. About Sleep's Role in Memory. Physiological Reviews 93, 681-766.

Rasch, B., Buechel, C., Gais, S., Born, J., 2007. Odor cues during slow-wave sleep prompt declarative memory consolidation. Science 315, 1426-1429.

Ribeiro, S., Gervasoni, D., Soares, E.S., Zhou, Y., Lin, S.C., Pantoja, J., Lavine, M., Nicolelis, M.A., 2004. Long-lasting novelty-induced neuronal reverberation during slow-wave sleep in multiple forebrain areas. PLoS Biology 2, E24.

Rissman, J., Wagner, A.D., 2012. Distributed representations in memory: insights from functional brain imaging. Annual Review of Psychology 63, 101-128.

Rudoy, J.D., Voss, J.L., Westerberg, C.E., Paller, K.A., 2009. Strengthening Individual Memories by Reactivating Them During Sleep. Science 326, 1079-1079.

Schabus, M., Gruber, G., Parapatics, S., Sauter, C., Klosch, G., Anderer, P., Klimesch, W., Saletu, B., Zeitlhofer, J., 2004. Sleep spindles and their significance for declarative memory consolidation. Sleep 27, 1479-1485.

Scholz, J., Klein, M.C., Behrens, T.E., Johansen-Berg, H., 2009. Training induces changes in white-matter architecture. Nature Neuroscience 12, 1370-1371.

Schönauer, M., Alizadeh, S., Jamalabadi, H., Abraham, A., Pawlizki, A., Gais, S., 2017. Decoding material-specific memory reprocessing during sleep in humans. Nature Communications 8, 15404.

Schönauer, M., Alizadeh, S., Jamalabadi, H., Emmersberger, M., Gais, S. (In prep.) Decoding retrieval success and memory content during short-term memory maintenance.

Schreiber, K., Krekelberg, B., 2013. The Statistical Analysis of Multi-Voxel Patterns in Functional Imaging. Plos One 8.

Schwindel, C.D., McNaughton, B.L., 2011. Hippocampal-cortical interactions and the dynamics of memory trace reactivation. Prog Brain Res 193, 163-177.

Simanova, I., van Gerven, M., Oostenveld, R., Hagoort, P., 2010. Identifying Object Categories from Event-Related EEG: Toward Decoding of Conceptual Representations. Plos One 5.

Smith, C., 1995. Sleep States and Memory Processes. Behavioural Brain Research 69, 137-145.

Smith, C., 2001. Sleep states and memory processes in humans: procedural versus declarative memory systems. Sleep Medicine Reviews 5, 491-506.

Sreenivasan, K.K., Vytlacil, J., D'Esposito, M., 2014. Distributed and Dynamic Storage of Working Memory Stimulus Information in Extrastriate Cortex. Journal of Cognitive Neuroscience 26, 1141-1153.

Stickgold, R., Walker, M.P., 2007. Sleep-dependent memory consolidation and reconsolidation. Sleep medicine 8, 331-343.

Todd, M.T., Nystrom, L.E., Cohen, J.D., 2013. Confounds in multivariate pattern analysis: Theory and rule representation case study. Neuroimage 77, 157-165.

Tong, F., Pratte, M.S., 2012. Decoding Patterns of Human Brain Activity. Annual Review of Psychology 63, 483-509.

Vuilleumier, P., Armony, J.L., Driver, J., Dolan, R.J., 2001. Effects of attention and emotion on face processing in the human brain: An event-related fMRI study. Neuron 30, 829-841.

Walker, M.P., van der Helm, E., 2009. Overnight therapy? The role of sleep in emotional brain processing. Psychol Bull 135, 731-748.

Wilson, M.A., Mcnaughton, B.L., 1994. Reactivation of Hippocampal Ensemble Memories during Sleep. Science 265, 676-679.

Winkler, A.M., Ridgway, G.R., Webster, M.A., Smith, S.M., Nichols, T.E., 2014. Permutation inference for the general linear model. Neuroimage 92, 381-397.

Woolgar, A., Golland, P., Bode, S., 2014. Coping with confounds in multivoxel pattern analysis: what should we do about reaction time differences? A comment on Todd, Nystrom & Cohen 2013. Neuroimage 98, 506-512.

Wurm, M.F., Ariani, G., Greenlee, M.W., Lingnau, A., 2015. Decoding Concrete and Abstract Action Representations During Explicit and Implicit Conceptual Processing. Cereb Cortex.

List of publications in this thesis

M. Schönauer*, **S. Alizadeh***, H. Jamalabadi, A. Abraham, A. Pawlizki, S. Gais (2017), "Decoding material-specific memory reprocessing during sleep in humans". *Nature Communications,* 8:15404 (*equal contribution)

M. Schönauer, **S. Alizadeh**, H. Jamalabadi, M. Emmersberger, S. Gais, "Decoding retrieval success and memory content during short-term memory maintenance", in submission process.

S. Alizadeh, H. Jamalabadi, M. Schönauer, C. Leibold, S. Gais (2017), "Decoding cognitive concepts from neuroimaging data using multivariate pattern analysis", *NeuroImage,* 159:449-458.

Statement of contributions

M. Schönauer*, **S. Alizadeh***, H. Jamalabadi, A. Abraham, A. Pawlizki, S. Gais (2017), "Decoding material-specific memory reprocessing during sleep in humans". *Nature Communications* **(*equal contribution)**

MS, AP, and SG planned and designed the experiment. MS, AA, and AP collected the data. SA and HJ developed and implemented the two-step classification framework for the between-subject classification. SA with the help of HJ developed and implemented the companion preprocessing steps dedicated for between-subject classification. SA and HJ classified the data and performed permutation test. MS and SG analyzed the behavioral data. SA and MS provided the analysis to relate behavioral data and machine learning output. HJ with the help of SA implemented the time domain analysis. SA provided the analysis to identify frequency contributions and spatial characteristics of memory reprocessing. SA wrote the first draft of the method part of the paper. MS and SG wrote the first draft of the content part of the manuscript. Revisions were done by HJ, SA, MS, and SG.

M. Schönauer, **S. Alizadeh**, H. Jamalabadi, M. Emmersberger, S. Gais, "Decoding retrieval success and memory content during short-term memory maintenance", in submission process.

MS, ME, and SG planned and designed the experiment. MS and ME collected the data. SA with the help of HJ preprocessed the data, implemented the classification framework for decoding memory content and retrieval success and performed permutation test for assessing the classification results. SA with the help of MS implemented the spatial searchlight analysis to identify the topography of predictive channels. SA came up with the idea of the

permutation-based frequency searchlight analysis to identify the contribution of different frequency bands. SA with the help of MS implemented this idea. SA and MS provided the analysis to relate classifier performance and behavioral data. MS wrote the first draft of the manuscript. SA, HJ, and SG revised the manuscripts.

S. Alizadeh, H. Jamalabadi, M. Schönauer, C. Leibold, S. Gais (2017), "Decoding cognitive concepts from neuroimaging data using multivariate pattern analysis", *NeuroImage*, 159:449-458.

SA, HJ and SG recognized that stimulus-related differences may confound estimation of class differences when multivariate pattern analysis is used for decoding cognitive concepts. SA developed concept-response curve to identify the relative contribution of stimulus-related and concept-related neural processing. SA showed with synthetic data that concept-response curve can have four theoretical shapes. CL with the help of SA, HJ, SG, and MS provided mathematical insight into the problem. MS and SG collected the EEG data. SA analyzed the data from EEG experiments. SA with the help of HJ implemented the time-resolved and searchlight analysis to identify the spatiotemporal signatures of conceptual and perceptual processing. SA wrote the first draft of the manuscript. SG, HJ, MS and CL revised the manuscript.

Chapter 2:

Decoding material-specific memory reprocessing during sleep in humans

Monika Schönauer*, Sarah Alizadeh*, Hamidreza Jamalabadi, Annette Abraham, Annedore Pawlizki, & Steffen Gais

* These authors contributed equally to this work

Published in Nature Communications (May 2017)

M. Schönauer*, S. Alizadeh*, H. Jamalabadi, A. Abraham, A. Pawlizki, S. Gais (2017), "Decoding material-specific memory reprocessing during sleep in humans". *Nature Communications,* 8:15404. (*equal contribution)

Abstract

Neuronal learning activity is reactivated during sleep but the dynamics of this reactivation in humans are still poorly understood. Here we use multivariate pattern classification to decode electrical brain activity during sleep, and determine what type of images participants had viewed in a preceding learning session. We find significant patterns of learning-related processing during rapid eye movement (REM) and non-REM (NREM) sleep, which are generalizable across subjects. This processing occurs in a cyclic fashion during time windows congruous to critical periods of synaptic plasticity. Its spatial distribution over the scalp and relevant frequencies differ between NREM and REM sleep. Moreover, only the strength of reprocessing in slow-wave sleep influenced later memory performance, speaking for at least two distinct underlying mechanisms between these states. We thus show that memory reprocessing occurs in both NREM and REM sleep in humans, and that it pertains to different aspects of the consolidation process.

Introduction

Sleep helps us retain new memories [1,2]. A reactivation of newly encoded memory traces in the sleeping brain is thought to underlie this effect. Replay of learning-related neuronal firing patterns has been observed in single cell recordings of the hippocampus and neocortex in animals [3-6]. Importantly, this sleep-dependent activation of neurons has recently been shown to promote synaptic plasticity [7]. Reactivation of neuronal ensembles involved in motor learning is associated with changes in the task-related spiking behavior of these neurons in the rodent brain [8]. Furthermore, oscillation related to memory replay during sleep have been linked to greater memory strength and precision in rats [9]. The dynamics of this memory trace reactivation in humans, however, are still poorly understood. When memory content was associated with auditory or olfactory cues during learning, a re-exposure to these cues during sleep can improve later recall performance [10,11]. Moreover, activity on the level of brain areas suggests reactivation during sleep [12,13]. It is unclear whether this re-expression of learning related activity reflects the specific content of a previous learning task. Recent advances in multivariate pattern classification (MVPC) methods have made it possible to investigate covert cognitive processes in continuous brain activity. Using such methods on brain activity measured with fMRI, Horikawa et al. [14] have recently shown that it is possible to decode the content of visual imagery occurring at sleep onset. In the present study, we used MVPC to test whether the human sleep electroencephalogram (EEG) contains information about what has previously been learned, and thus indicates reprocessing of memory content.

In our experiment, participants learned pictures of either faces or houses before sleeping in the laboratory for a whole night. During this time, brain activity was recorded using high-density EEG. We then employed MVPC methods to detect information about the previously learned material in electrical brain activity

during sleep (Fig. 1, also see Materials and Methods). We investigated continuous sleep EEG instead of evoked activity, because we were specifically interested in spontaneous information processing in sleep. Cued reactivation, which has already been demonstrated in humans with functional MRI, shows that stimulus processing in sleep can lead to memory improvement. Previous studies, however, have not shown that such activity actually occurs spontaneously in humans. After demonstrating the existence of such an activity, we were also interested in the time course of memory reprocessing across the night and in sleep-stage specific activity. It has been discussed previously whether such reactivation occurs during NREM or REM sleep, and both have been implicated in memory reactivation and consolidation [12,13,15,16]. Furthermore, activity that is present only at specific times during the night indicates that the underlying process is related to discrete periods of reprocessing rather than prolonged ongoing activity.

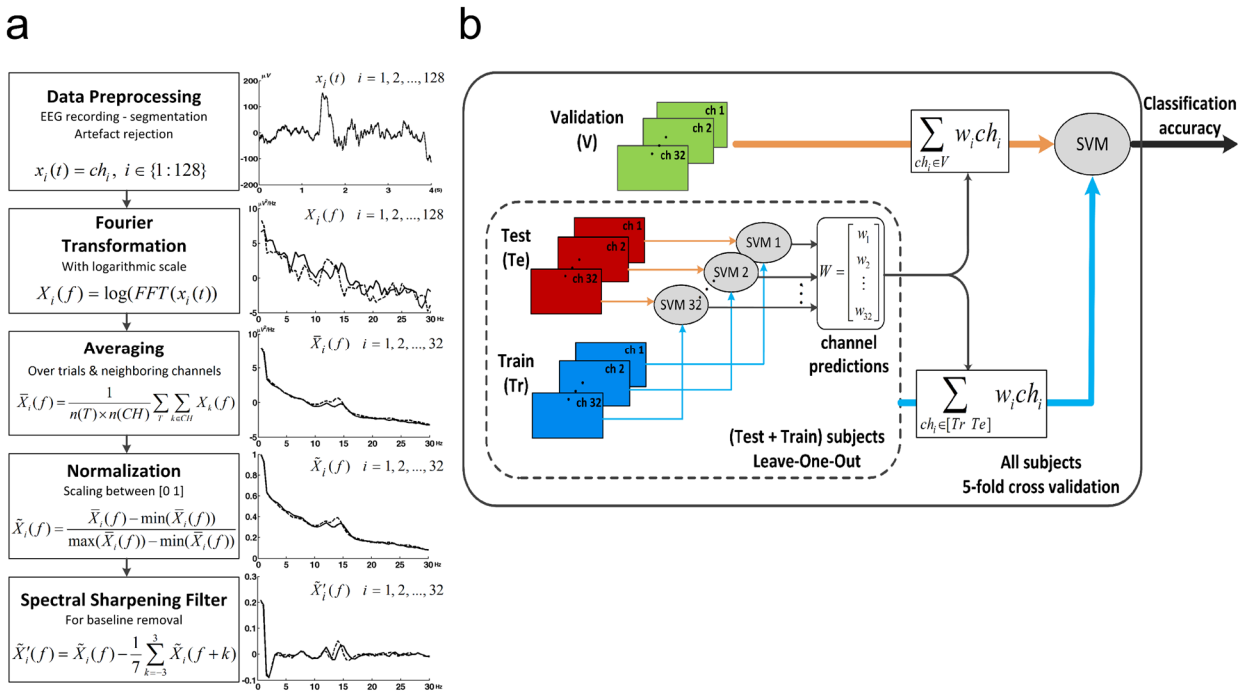


Figure 1: Data preprocessing and MVPC analysis. (a) After artefact rejection, data from the remaining 4-s trials of 128-channel sleep EEG data was frequency transformed. To reduce the dimensionality of the data and to increase the signal-to-noise ratio, spectra were averaged over trials and neighboring channels. Next, spectra of all channels were

normalized separately to make them comparable, and a spectral sharpening filter was applied to remove the baseline spectrum and enhance differences between neighboring frequency bins. (b) Training data was strictly separated from validation data in all MVPC analyses. Dimensionality of the data was further reduced in a two-step training procedure. Individual channel performance was determined using separate single-channel classifiers. An average of data from all channels weighted by their standalone performance was then used to train a classifier to distinguish between face and house stimulus conditions. Finally, classification was tested on independent validation data.

Results

Detecting memory reprocessing using MVPC

We tested whether MVPC can decode from the sleeping brain's activity what has been learned beforehand. Instead of looking for a single feature that can distinguish between conditions, MVPC methods take into account and compare the whole temporospatial pattern of activity. Given their multivariate nature, they are more suitable to deal with this kind of high-dimensional problem than is classical statistics, which usually relies on multiple univariate testing. Because EEG activity differs greatly between sleep stages and even more so between sleep and wakefulness, activity cannot be compared directly between these states. We therefore used between subject analyses to compare recordings from the same sleep state, i.e. the classifier was trained and tested on sleep data. If MVPC can determine from the sleep recording which type of visual stimulus a subject has learned before sleep, this implies that stimulus-specific reprocessing of the learned material occurs during sleep.

Our results show that human sleep EEG contains information about which kind of visual stimuli was learned before sleep (Fig. 2a). Classification accuracies for this distinction exceed classification rates expected from chance guessing of the classifier, as determined by randomization statistics, during two of the four 90-min segments (Fig. 2b). Thus, the sleep EEG reflects previous learning during

these intervals. Moreover, both NREM and REM sleep contain relevant information (Fig. 2a, b and c).

We used two different approaches to ensure that findings are significant and generalizable. First, we generated randomly labeled data, which, per se, cannot contain any information, and compared the performance of the classifier on these random data with its performance on the original observed data (see Supplementary Fig. 1). This test allows to determine the probability of an outcome by chance given that the data contain no actual information and thus provides exact significance values. Because this process, which repeats the whole analysis for each random iteration, is computationally intensive, we could complete only 1001 repetitions, which allows significance testing with a lower limit of precision of p=0.001. In the case of REM sleep of the 2nd 90-min sleep segment, none of these 1001 random iterations produced higher classification rates than the real data, thus allowing the conclusion of p<0.001.

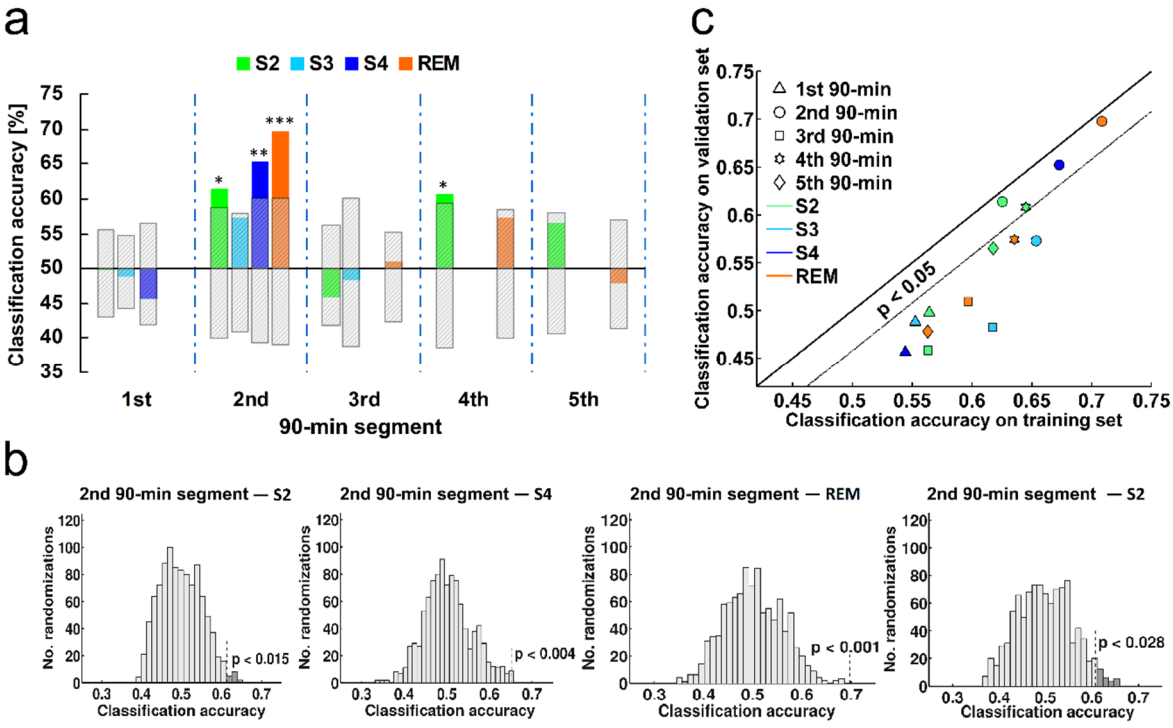


Figure 2: Classification results. (a) The content of a previous learning experience can be determined from sleep EEG during the second and fourth 90-min segment of the night. At these times, classification accuracy for all sleep stages is significant or approaches

significance. The hatched area shows the 95% confidence interval. Classification accuracies for S4 sleep as well as REM sleep in the second sleep segment remain significant after Bonferroni-Holm correction considering all tests (S4: p = 0.048, REM: p = 0.014). (b) Significance was assessed using permutation tests to ensure that classification rates are higher than can be expected from data sets with random labeling of the data, i.e. not containing any information. To estimate the displayed null-distribution from which exact significance levels of classification results can be determined, the MVPC analysis was repeated 1001 times on the actual data with randomly shuffled condition labels. Dark grey areas show those randomizations during which classification accuracy on randomly labeled data exceeded accuracy obtained on correctly labeled data. (c) If classification accuracies are similar between the training and validation sets, generalizable information could be extracted and the classifier was not overfitted on the training data set. This was the case for all analyses that were significant, i.e. for data from the second (circles) and fourth (stars) 90-min segments of the night. Here, patterns detected in one set of subjects during classifier training can be generalized to data from a new set of subjects. Data from the first (triangles) and third (squares) 90-min segments show low training accuracy low accuracy on validation data, indicating that the classifier could not extract information about previous learning content from these periods of the night.

The second approach to ensure generalizability was to compare classification accuracies of training and validation sets. If accuracy is higher during training than during validation testing, the classifier was overfitted to the training data set and uses random feature characteristics that allow separating classes only in the training data , which are not predictive for new data, and thus cannot be generalized. Ideally, classification rates for the validation data should resemble those for the training data. This shows that the classifier can extract meaningful information from the training set, and that the learned pattern can be generalized to new data. It can be seen in Fig. 2b that for data from the 1st (triangles) and 3rd (squares) 90-min sleep segment training accuracy was low (<0.625), but classification accuracy for the validation set was still worse. Thus, EEG from these periods does not seem to contain information pertaining to previous learning experience. On the other hand, EEG from the second (circles) and fourth (stars) 90-min sleep segment consistently shows higher training and

validation accuracies, and in some cases shows nearly perfect generalization between training and validation.

Relating reprocessing to behavioral memory performance

Participants showed good recognition performance in both the face and house learning conditions (see Supplementary Table 1). We did not observe forgetting across the night. This result is in line with other studies on declarative memory consolidation that have shown stable maintenance of memory performance over sleep but significant decline of memory performance after sleep-deprivation or daytime wakefulness [17,18]. Memory consolidation, i.e. the overnight change in performance, was positively correlated with time spent in sleep stage S4 (r_{64} = 0.254, p = 0.043; Supplementary Table 2), confirming that sleep was related to the consolidation of this task. We also tested the relation of memory consolidation with the strength of memory reprocessing, which was inferred from the classification probability estimates provided by the classifier. We find that memory reprocessing during SWS shows a positive relation with memory consolidation (r_{64} = 0.329, p = 0.008; Supplementary Table 3 and Fig. 3). This correlation remained significant after removing the three most influential values determined by leverage statistics (r_{61} = 0.28, p = 0.030). Memory reprocessing during sleep stage S2 and REM sleep were not related to memory performance (S2: r_{64} = 0.099, p = 0.436; REM: r_{56} = -0.199, p = 0.142). A regression model including strength of reprocessing in S2, SWS and REM sleep as predictors for memory consolidation found that only reprocessing during SWS correlated significantly with memory consolidation (β = 0.339, p = 0.020, explaining 9.7% of the variance), reprocessing in S2 and REM sleep was no significant predictor (S2: β = -0.064, p = 0.656, explaining 0.3% of the variance; REM: β = -0.112, p = 0.436, explaining 1% of the variance). Slopes differed significantly between SWS and REM sleep (strength of reprocessing × sleep stage interaction: p = 0.008), indicating that memory reprocessing in these sleep stages is differentially related to memory consolidation and could thus have different functions.

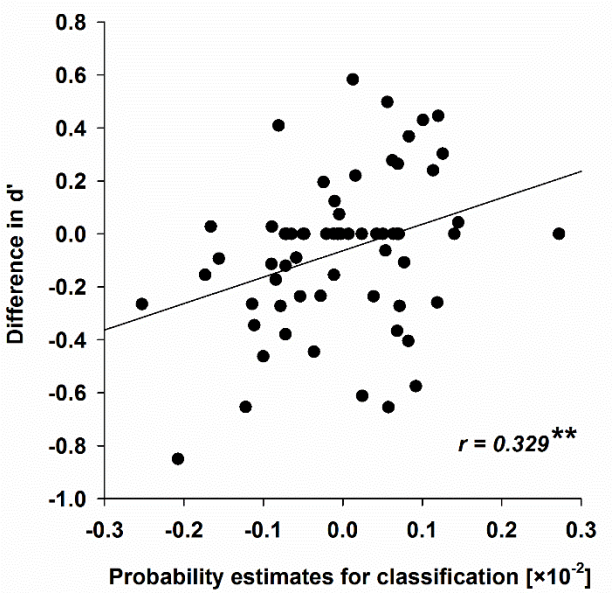


Figure 3: Correlation between classification probability estimates and overnight memory consolidation during SWS sleep. The more confident the classifier was in placing each subject in the correct condition, the more positive the change in memory performance during later recall. Spearman's rho is reported.

We then controlled whether general sleep features such as time spent in deep sleep could possibly account for an increase in both behavioral performance as well as classifiability of the data. Entering strength of reprocessing in SWS and time spent in this sleep stage in a regression model, we found that only strength of reprocessing in SWS was a significant predictor of memory consolidation and explained a larger part of the variance ($\beta = 0.335$, $p = 0.006$, explaining 11.2% of the variance), whereas duration of SWS was only marginally significant ($\beta = 0.214$, $p = 0.074$, explaining 5.2% of the variance). Strength of reprocessing in SWS was independent of time spent in that sleep stage ($r_{64} = -0.025$, $p = 0.423$) and the partial correlations support the view that strength of reprocessing in SWS and duration of SWS are independent predictors of overnight memory consolidation (partial correlation with strength of reprocessing during SWS controlling for the duration: $r_{64} = 0.342$, $p = 0.006$; partial correlation with duration of SWS controlling for strength of

reprocessing: $r_{64} = 0.226$, $p = 0.074$). Analogous regression analyses for strength of reprocessing and time spent in S2 and REM sleep yielded no significant results, as could be expected from the general lack of association with overnight memory consolidation (all $p > 0.143$).

While the proportion of variance in overnight memory consolidation that is explained by memory reprocessing during SWS is low in absolute terms, it should be noted that factors such as alertness or individual differences can introduce considerable variance in memory performance. Classifier performance similarly provides a measure of reprocessing strength that is affected by many sources of between-subject variance as it is estimated based on other participants' sleep EEG characteristics. Despite these difficulties, we demonstrate that memory reprocessing during SWS is significantly related to overnight memory retention, suggesting a robust underlying effect.

Temporal dynamics of reprocessing

We detected processing of learning material during sleep in the second and fourth 90-min segment of the night (Fig. 2). To investigate this pattern on a more fine-grained scale, we split the night into smaller intervals and analyzed the time course of classification accuracy across the night with a resolution of 4.5 min, using the same procedure as above. Again, we find two periods of the night during which brain processing seems to be more strongly related to previous learning, congruent with the time windows reported above. During other periods, no learning-related information was detected (Fig. 4).

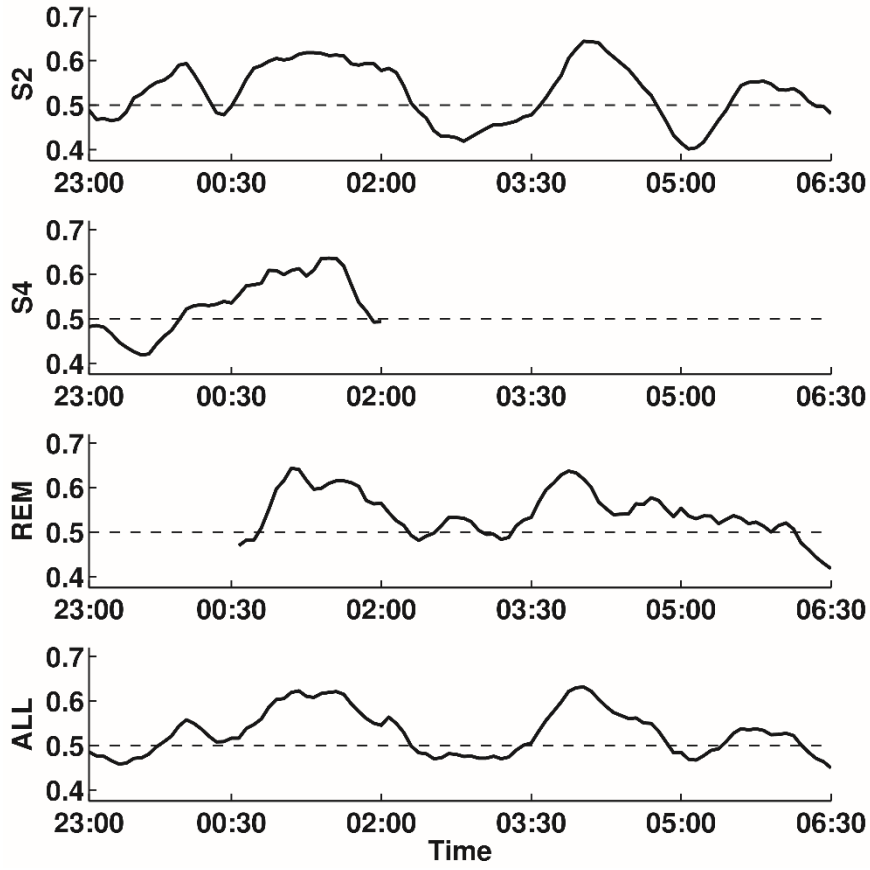


Figure 4: Time course of classification accuracy across the night. Separate analyses were performed for sleep stages S2, S4, and REM sleep. Classification performance follows an oscillatory pattern and peaks around three and six hours after learning in all stages. Timing therefore is more relevant to when memory reprocessing occurs than sleep stage

Spatial characteristics of reprocessing and frequency contributions

Brain activity in REM and NREM sleep is not alike. It is thus reasonable to assume that also information processing in these states will take different forms. To investigate this, the relative contribution of each frequency band to classification can be assessed in terms of classification weights and compared between sleep stages (Fig. 5). Our results show that the frequencies that are important for identifying previous learning content differ between sleep stages. Activity in the range of sleep spindles (11-16 Hz) can distinguish previous

learning conditions only in NREM sleep (Fig. 5a). Theta-band activity (4-8 Hz), on the other hand, has higher discriminative power in REM sleep. Slow frequencies below 4 Hz were informative in both NREM and REM sleep, but their topographies differ (Fig. 5b). Although there is some resemblance between the feature weight plots and power spectra of sleep, it has to be noted that the feature weights do not follow the typical 1/f logarithmic decrease of EEG power spectra, but remain essentially constant after a linear decrease in delta frequencies. Moreover, actual classifier input was not the power spectra but the preprocessed data seen in the lower panel of Fig. 1a.

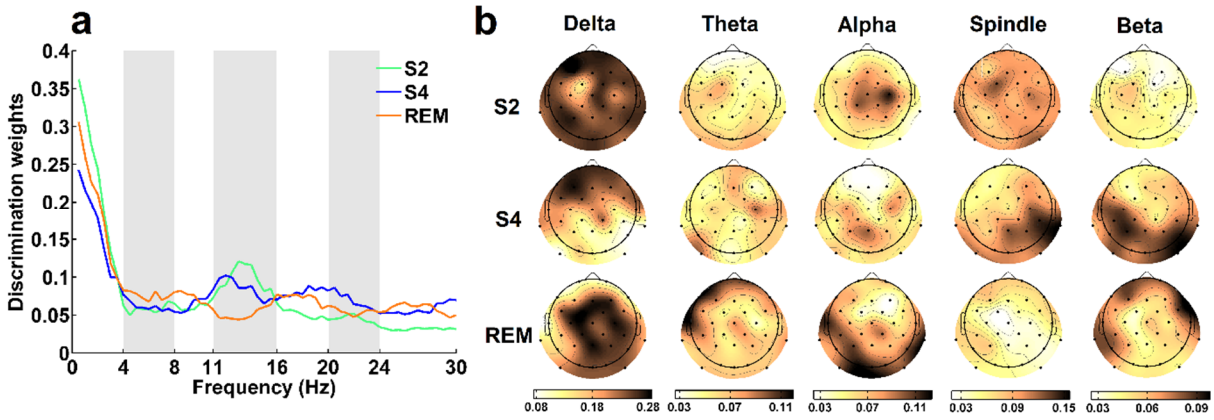


Figure 5: Frequency contributions to memory reprocessing in NREM and REM sleep. (a) Discrimination weights show that in NREM sleep stages S2 and S4 spindle activity in the frequency range between 11 and 16 Hz is predictive for learning content. In REM sleep, theta, alpha, and higher beta frequencies contributed more to correct classification. Slow frequencies below 4 Hz were informative in all sleep stages. (b) The topography of predictive channels clearly differs between NREM and REM sleep. In NREM sleep stage S2, mainly delta and spindle frequencies contributed to correct classification. Similarly, frontal delta power and right parieto-temporal spindle activity were most informative for classification during NREM sleep stage S4, together with posterior higher frequency activity. REM sleep shows a more complex pattern. Here, slow oscillations of central electrodes, frontal and temporal theta as well as occipital alpha contributed most to discrimination between learning conditions.

Discussion

We show that memory processing of a single memory task occurs during all stages of sleep. Reprocessing in REM and NREM sleep, however, has different effects on later memory performance. Although a large number of studies in rodents have observed the occurrence of spontaneous memory reactivation during NREM sleep [4-6,19,20], linking this reactivation with improvements in behavioral performance has remained a challenge. Contrary to rodents, task difficulty and training time can be easily adjusted in studies on humans, giving greater power to analyses on behavioral effects. It has early been suggested that memory reactivation during sleep has functional significance for strengthening new memories [21]. Indirect evidence for this assumption has accumulated over the last years [10,11,22-24]. A recent study in rats found that sleep-dependent reactivation of neurons involved in a simple motor learning task is associated with changes in the task-related spiking behavior of the same neurons [8]. In this way, reactivation may be related to later improvements in performance. We now show that content-related reprocessing of declarative learning material during NREM sleep influences later memory strength in humans. Conversely, memory reprocessing during REM sleep does not show this graded relation with overnight memory retention.

A number of animal studies detected reactivation of learning activity also in REM sleep [25,26], yet empirical evidence for this has remained ambiguous. We find that memory content is reprocessed during both NREM and REM sleep. The differential significance of memory reprocessing for behavioral performance between these states points towards at least two different mechanisms underlying memory reprocessing during sleep.

Already early on, it has been suggested that memory is formed in a two-stage process. Labile memory traces are formed during exploratory behavior, when theta power is high. Later, during rest or sleep, long-lasting traces are formed [9,21]. Similarly, it has been proposed that during sleep, slow-wave-related NREM

activity and theta-related REM activity have complementary, mutually dependent functions [27]. We find that reprocessing occurs in both NREM and REM sleep. Interestingly, we can demonstrate a correlation between reprocessing and later memory performance only for NREM sleep. This supports the view that reprocessing during REM sleep and NREM sleep serves distinct functions. Our finding is in line with previous studies, which show no behavioral benefit of reactivating memories by cueing during REM sleep [10]. Interestingly, memory replay observed during REM sleep has also been shown to have different characteristics than that in NREM sleep, including a smaller time-compression factor, which is less suited for the induction of long-term potentiation [20,25].

A number of recent studies stress the importance of either light NREM sleep, SWS or REM sleep for memory consolidation, respectively [2,27,28]. Based on these findings, theoretical accounts have suggested that NREM and REM sleep may interact during memory consolidation, emphasizing different aspects of this process. The sequential hypothesis of sleep stresses that different sleep stages have to occur in succession to effectively influence memory function. It assigns specific and substantially different, but interdependent roles to NREM and REM sleep regarding the processing of memories [29]. Other accounts suggest the different processes contributing to memory processing during NREM and REM sleep are separate and independent. Thus, the function of NREM and REM sleep in consolidation is assumed to pertain to different aspects or forms of memory [30]. We find that relevant activity occurs in close temporal proximity over different stages, and that a single memory task triggers learning-related activity in both NREM and REM sleep EEG. It therefore seems possible that both sleep stages cooperate in the processing of memories. The differential function of NREM and REM sleep stages is still controversial [7,16,31]. One recent hypothesis is that cortical activity and long-range connectivity differs between sleep stages, allowing local memory reactivation and potentiation in SWS, and network-wide information integration in REM sleep [32,33]. This view fits with our findings.

Our data indicate that memory processing in sleep is cyclic in nature and its occurrence might depend more strongly on timing than on the stage of sleep. Instead of occurring in SWS throughout the whole night, reprocessing was detected in S2, S4 as well as REM sleep in the 2nd 90-min period, but not in the 1st or 3rd. Whether this consolidation window depends on time after learning, time after sleep onset, or circadian rhythm cannot be determined in the present study, because these were not varied independently.

Because reprocessing peaks during distinct times of the night, it is unlikely that the detected activity simply reflects ongoing reverberation of learning-related activity or selective fatigue in the involved brain areas. Instead, it must reveal a process that is selectively initiated at specific points during sleep. The finding that reprocessing is strongest around three and around six hours after learning fits well with experiments that found critical periods during memory consolidation, during which memory is particularly sensitive to disruption [34]. Thus, inhibiting protein synthesis 15 min and 3 h after learning, but not 1 h after learning impairs hippocampal one-trial avoidance learning [35]. Similarly, in drosophila, different behavioral memories and corresponding neuronal traces develop during different time windows over several hours after conditioning [36], a process that has been linked to systems memory consolidation in humans [37].

Moreover, our finding of discrete periods for memory reprocessing is reminiscent of previously reported 'sleep windows', i.e. times during which sleep has to occur after learning to strengthen memory [38,39]. Along the same lines, Stickgold et al. have found that, for consolidation of a visual discrimination task, mainly duration of SWS and REM sleep in the first and the last quartile of the night, respectively, are most critical parts of the night [40]. Although that task presumably does not rely on hippocampal memory reactivation and might therefore follow a different temporal trajectory, the similarities suggest the possibility of a common mechanism. Further behavioral, electrophysiological and molecular investigations are required to elucidate this underlying mechanism. Moreover, it has still to be ascertained whether the other periods of

the night have memory-related functions that cannot be detected by our method.

Because the amount of signal related to memory reprocessing across the whole night is very small compared to the unrelated noise, we used MVPA, which is a very sensitive method to detect systematic differences between large sets of data. However, multivariate approaches are not better suited to supply information about univariate hypotheses than classical tests. Using feature weights and individual channel accuracies (Fig. 5) can to some extent illustrate the features that are carrying relevant information. However, these features must be seen within the entire pattern. The following discussion of individual physiologic features should therefore be seen as a starting point for studies focusing on a smaller feature search space.

When looking at the frequencies contributing to correct classification, we find that spindle activity during NREM sleep contributes to the distinction of previous learning conditions. This is consistent with the fact that sleep spindles increases after learning [41] and correlate with performance [42]. Parietal sleep spindles accompany task specific reactivation seen in fMRI [43]. Moreover, frontal slow-waves, as they appear in our analysis for NREM sleep, have previously been shown to correlate with performance gains observed after memory reactivation induced by cueing during sleep [44].

Apart from confirming that learning-related information resides in frequency bands that have previously been implicated in memory consolidation, such as NREM spindles and slow oscillations, our results hint at promising objects for future study. We suggest that particular attention should be given to the function of REM sleep theta. Frontal theta power increases during successful memory encoding and retrieval, and theta is also involved in memory processing during wakefulness, such as in controlling, maintaining and storing memory content [45]. Theta has been linked to memory and sleep for a long time, but has only recently received renewed attention [16,46]. For instance, theta band activity during sleep has been shown to support formation of imprinting memory in chicks [47]. In

humans, another recent study found increased frontal theta power after presentation of cues related to a verbal learning task during sleep [44,48]. Moreover, frontal theta in REM sleep is predictive of successful dream recall [49]. These findings stress the active role of theta activity in memory reprocessing during sleep.

It is difficult to demonstrate reactivation directly in humans. Electroencephalographic activity during sleep differs greatly from that during wakefulness in both the time domain and the frequency domain. Thus, amplitude changes over time, as well as power spectral density cannot be compared between these states. This is owing to different modes of generation and transmission of electrical activity during sleep [50,51]. Previous data have shown that reactivation can be both time-compressed as well as changing in location (e.g. neocortical replay following hippocampal activity) [19,52]. Markers reflecting reactivation of neuronal firing patterns observed during learning can thus be altered by a large number of operations, which renders the search space virtually infinite. Because this makes wake-to-sleep classification problematic, and a within-subject design would have to rely on between-session classification that is confounded by various session differences (e.g. recording artefacts), we instead opted for a between-subject classification approach. This allowed us to detect information pertaining to a previous learning experience in data recorded in the same state of consciousness. Previous attempts to observe memory reactivation during off-line periods succeeded in showing memory reprocessing during wakefulness, but not during sleep [53-55]. Using an approach that trains and tests the classifier in the same state of consciousness made it possible for us to observe material-specific memory reprocessing during sleep and study its dynamics and relation to later behavioral performance.

We used multivariate pattern classification to decode the content of a previous learning experience from electrical brain activity during sleep. By linking brain activity during sleep with the content of previous learning, our findings bridge studies from multicell recordings in animals, which show learning-related

reactivation, to human imaging studies, which show reactivation of brain regions during sleep. Pattern classification methods are powerful tools for investigating the covert mechanisms that link electrical brain activity and behavior, and can thus contribute to our understanding of these complexities.

Materials and Methods

Subjects. In this study, we recorded EEG data from 32 healthy subjects with no history of neurological or psychiatric disorders. All participants were students, between 18 and 30 years old, native German speakers and non-smokers. They were right handed as measured by Edinburgh Handedness Inventory-test [56]. Chronotype was assessed via the Munich Chronotype Questionnaire [57] and experimental timing was adjusted to participants' usual sleep times (sleep midpoint 03:56h ± 01:33h [mean ± SD]). Subjects were regular sleepers with a habitual sleep duration of 6-9 h. They did not report any chronic or acute sleep-related problems in an initial interview. Moreover, they did no shift work and did not change time zones in the six weeks leading up to the experiment. Participants were told to refrain from drinking alcohol, coffee and tea on the days of the experiment and did not take any drugs that affect the central nervous system. All experimental procedures were approved by the local ethics committee (Department of Psychology, Ludwig-Maximilians-Universität München). Informed consent was obtained from all subjects.

Experimental Design. Participants slept in our laboratory on three different nights. The first of these served as an adaption night, to accustom subjects to the environment and to sleeping under the experimental conditions (e.g. wearing an EEG cap). In the subsequent two experimental nights, subjects completed an intensive image learning task, during which they studied pictures of either faces or houses. For an exemplary subject, learning took place from 8:30 p.m. to 10 p.m. after the EEG electrodes had been attached, and memory was tested immediately afterwards. The subject then went to bed at 11 p.m. for an 8-h sleep

period. Memory was tested once more in the morning. The times of the experiment were advanced or delayed such that time to bed corresponded to the individual habitual bedtime of the participants. All subjects participated in two experimental nights, each time learning only one type of images, in a counterbalanced fashion. The two nights were spaced at least 5 days apart. Sleepiness was tested with a visual analog scale in the evening and after sleep in the morning (Supplementary Table 4).

Learning Task. Subjects studied a set of 100 images of faces or houses in 30 repetitions. Pictures were shown in random order and individual images were always presented in one of the four quadrants of the screen. Participants had to remember the individual pictures and learn to associate the images with the quadrant in which it was presented. Participants were tested once immediately after learning and again in the next morning after a full night of sleep. During both immediate and delayed testing, 100 learned images were presented together with a set of 50 new images in random order. Participants first had to indicate via keypress whether they had seen the image before (with left hand on main keyboard: 1-sure, 2-probably, 3-probably not, 4-surely not. Responses 1 and 2 were counted as a "yes" response, responses 3 and 4 were counted as a "no" response). For "yes"-responses, also the quadrant in which the image had been presented was probed (with right hand on numerical pad: 1-lower left, 3-lower right, 7-upper left, 9-upper right). Image material was derived from two different sources: 300 pictures of houses were taken from German online real estate sites, 300 pictures of neutral faces were taken from Minear & Park [58].

This task was chosen because it is a declarative task that is supposed to involve the hippocampus, and sleep-related reactivation has mainly been shown in the hippocampus [10,19]. Face and house processing are clearly different in event-related EEG potentials and fMRI [59]. Face processing activates the mid-fusiform gyrus (fusiform face area) and the occipital face area in the occipito-temporal cortex as well as other temporal areas [60], whereas processing of houses activates the parahippocampal place area and the lateral occipital gyrus [61,62].

EEG Recording. Sleep EEG was recorded using an active 128 channel Ag/AgCl-electrode system (ActiCap, Brain products, Gilching, Germany) with 1 kHz sampling frequency and a high-pass filter of 0.1 Hz. Electrodes were positioned according to the extended international 10–20 electrode system. For sleep scoring, recordings were split into 30-s epochs and sleep stages were determined on electrodes C3/C4 according to standard rules by two independent raters [63]. Average sleep durations are reported in Supplementary Table 5.

Methodological Considerations. One of the challenges in sleep research is the difficulty of recording large sample sizes and the large amount of data that is recorded. The goal of classical analyses, which use multiple univariate comparisons (e.g. classical fMRI analysis), is to find single features that are strong enough independently to distinguish between conditions. Such features are unlikely to exist in high-density all-night EEG recordings, which thus present a problem better addressed by a multivariate approach. In multivariate analyses, it is of interest whether the overall pattern of data contains information that is relevant to distinguish conditions. A prominent method that can deal with large numbers of data dimensions is MVPC. However, high dimensional, low sample size data, like EEG recordings, pose specific problems for classical statistical testing as well as for MVPC [64,65]. For this kind of data, it is important to minimize the number of features. If the signal across features is highly correlated, as in EEG data, this can be achieved by averaging, which reduces dimensionality of the data and at the same time increases signal-to-noise ratio. We developed a two-step procedure that uses spatial averaging and a channel-based weighted average to improve classifiability of our data (Fig. 1). These steps are described in detail in the sections Data Preparation and Multivariate Pattern Classification (MVPC) below.

Data Preparation. For artefact rejection and further analysis, EEG data was split into 4-s trials. Artefact rejection was done in a semiautomatic process using custom MATLAB scripts. Based on the distributions of different parameters of

the raw data and power spectrum, rejection thresholds were chosen for each recording individually to make sure that only a minimal number of artefacts remained in the data. We tested for disconnected electrodes (outliers in overall spectral power), sudden jumps of the signal (outliers in amplitude changes) and muscle artefacts (outliers in spectral power between 110 and 140 Hz). Outlier thresholds were automatically suggested based on the variance of the data and manually confirmed upon visual inspection of parameter distributions and of the raw data. Trials containing artefacts were removed from the data set, channels that contained too many trials with artefacts were removed entirely and interpolated using routines provided by EEGLAB [66]. Whether individual epochs or channels were to be removed was determined automatically so that data loss was kept minimal. Artefact-free trials were then transformed into the frequency domain using Fourier transformation. To obtain smooth spectra, Welch's method was used for this, averaging over 10 Hamming windows of 2-s length with 95% overlap, resulting in a final data resolution of 0.5 Hz. Data was used up to a maximum frequency of 30 Hz.

The subsequent steps for data preparation were implemented to 1) increase signal-to-noise ratio, 2) reduce dimensionality of the data, and 3) adapt the signal for between-subject classification. First, we averaged power spectra across electrodes within a radius of approximately 3 cm around the 32 evenly spread locations of the extended 10-20-system to decrease the number of redundant features and increase signal-to-noise ratio as well as spatial similarity between subjects. We then separately averaged over all artefact-free trials available for each 90-min segment and sleep stage, to obtain a reliable estimate of spectral properties. This also ensures that an equal number of epochs per subject enters analysis, which is important for classification to remain unbiased. To remove amplitude differences between channels, which are caused by the distance of each channel to the reference electrode, spectra of all channels were separately normalized between zero and one. This also removed between-subject variability in general spectral power.

Because baseline EEG power spectra are highly similar and differences between conditions can be expected to be of smaller magnitude, these differences need to be enhanced within the spectra. We thus applied a spectral sharpening filter, which removes the baseline spectrum and emphasizes differences between neighboring frequencies in a final preparation step. To achieve this, we subtracted a moving average of six neighboring frequency bins (window size: 3 Hz) from the signal. This accentuates the smaller differences in power between frequencies within the spectrum. This is a valid procedure because neighboring data points in the power spectrum represent neighboring frequencies from the same signal and are therefore strongly correlated.

Subjects were only included in the analysis if they had at least 40 artefact-free trials within the respective sleep stage and segment (i.e. 160 s of data). Only segments and stages with at least 11 subjects were analyzed. The number of subjects and trials available for each 90-min segment and sleep stage can be found in Supplementary Table 6. As can be seen from that table, the amount of data available was unrelated to classifier performance.

Multivariate Pattern Classification (MVPC). In the present study, we tested whether electrical brain activity during sleep holds information about the content of previously learned visual stimuli. Instead of the typically used multiple univariate tests, we employed a multivariate classification approach, which can detect information contained in the overall pattern of brain activity, but is not distinguishable from single features.

Sleep EEG recordings from 64 nights (32 subjects, two conditions each) were analyzed using a classification algorithm developed on the basis of linear support vector machines (SVM). The aim was to detect material-specific information in the data. Please note that whereas the experiment followed a within-subject design, classification was done between subjects, with both nights of each participant (face and house conditions) assigned either to the training, test, or validation set. All analyses were done with the Matlab implementation of libsvm 3.1 (http://www.csie.ntu.edu.tw/~cjlin/libsvm). EEG

recordings pose problems typical of high dimensional, low sample size data (potential feature space of 128 channels times 60 frequency bins). We thus preprocessed the data to reduce the number of features and increase signal-to-noise ratio (see Fig. 1 and Data Preparation), averaging over neighboring channels to lower the number of channels to 32. To further enhance relevant features, we used a stepwise procedure for classification, which first regarded every channel as an independent classifier and then combined outcomes of this first step for the final analysis.

We split data into independent training and validation sets. In a first step, one linear SVM was trained for each of the 32 averaged EEG channels on all but one subject of the training set to see how much each channel contributes to distinguishing the content of learning conditions ('face' learning or 'house' learning). This channel-based classification was cross-validated in a leave-one-out procedure on each subject, and the obtained classification accuracies were averaged over all cross-validation runs. In the second step, this average classification accuracy from each channel was used as a weight to obtain a weighted average of the 32 channels. The main SVM was then trained on this weighted training set and classification accuracy tested on the independent validation set. The main reason for weighted averaging of channels was to reduce feature space dimensionality, because feature weights cannot be reliably determined if sample size is much smaller than the number of features [67]. Apart from this, weighted averaging can amplify relevant information in the data. This two-step classification process was cross-validated on independent data using 280 repetitions of a 5-fold procedure, which covers the whole data set with five independent validation sets.

We used permutation tests to assess significance. These tests sample the distribution of the null hypothesis by random shuffling of the original data, which is repeated a large number of times. To obtain the correct null-distribution for our data, we randomly shuffled condition labels, i.e. the two conditions of each subject were randomly labeled as 'face'/'house' or as

'house'/'face', effectively removing all relevant data pertaining to the effect of interest, while keeping other dependencies in the data constant. We then calculated classification accuracies for the randomly labeled data to estimate the random distribution. This was repeated 1001 times. Significance was calculated by determining the percentage of times that classification on randomly labeled data produced accuracies that were equal to or higher than the classification accuracy obtained from the actual data. If randomly labeled data did not result in a classification accuracy equal to or higher than the actual data, then the p value was determined by the number of random repetitions that were calculated (see Supplementary Fig. 1).

To assess whether reprocessing occurs uniformly across time, we split the night, starting from time to bed, into five 90-min segments, which are likely to include a whole sequence of sleep stages (S2, S3, S4, and REM sleep; see Supplementary Table 5 for details of sleep stage distribution). In this first analysis, we classified separately for all segments and sleep stages to assess the temporal dynamics of memory reprocessing. To determine a more fine-grained time course of classification accuracy, we moved a sliding window with a width of 22.5 min in steps of 4.5 min across the night. We then estimated classification accuracy within each window using the same two-step classification procedure as before. Analysis was done separately for each sleep stage and the same inclusion criteria were applied as in the main analysis.

To assess which features of the sleep EEG are particularly predictive, we analyzed classification weights. To assess which features of the sleep EEG are particularly predictive, we analyzed classification weights. The absolute value of the weights are informative about how much each frequency band and channel contributes to successful distinction. We averaged the classification weights over all repetitions of the training procedure, resulting in an averaged 32 (channels) × 60 (frequency bins) weight matrix. To examine frequency contributions to memory reprocessing, we further averaged the absolute values of these weights over all channels (see Fig. 5a). The topography of predictive

channels (see Fig. 5b) was obtained by averaging absolute values of classification weights for each channel over different frequency bands (delta: 0.5-3.5 Hz, theta: 4-7.5 Hz, alpha: 8-10.5 Hz, spindle: 11-15.5 Hz, beta: 16-30 Hz). We chose to analyze classification weights for frequencies obtained in the inner train-test loop (Fig. 1) because they can give additional information on the topography of predictive channels. These frequency weights are confirmed by weights from the outer validation loop (Fig. 1). Frequency contributions to classification assessed from both loops show the same pattern (see Supplementary Fig. 2).

Behavioral Performance. For assessment of memory performance, we calculated the memory sensitivity index d' as the difference of z-values between correctly recognized old items vs. falsely recognized new items (z[hits] – z[false alarms]). Performance pre and post sleep, as well as memory consolidation across the nights is reported in Supplementary Table 1. We correlated overnight memory consolidation with time spent in different sleep stages (see Supplementary Table 2). To examine whether memory reprocessing during sleep is associated with better memory performance, we correlated the probability estimates for classification given by the classifier with overnight memory consolidation measured as the difference between post sleep and pre sleep d' values. No such correlation was found for encoding or retrieval performance per se (see Supplementary Table 3). For each subject, results of all 280 repetitions of the 5-fold cross-validation procedure were averaged. We conducted this analysis separately for different sleep stages. All correlations report Spearman's rho.

Data availability

All data and codes are available from the corresponding authors upon request.

References

1 Walker, M. P. & Stickgold, R. Sleep, memory, and plasticity. *Annu. Rev. Psychol.* **57**, 139-166, doi:10.1146/annurev.psych.56.091103.070307 (2006).

2 Rasch, B. & Born, J. About sleep's role in memory. *Physiol. Rev.* **93**, 681-766, doi:10.1152/physrev.00032.2012 (2013).

3 Ribeiro, S. *et al.* Long-lasting novelty-induced neuronal reverberation during slow-save sleep in multiple forebrain areas. *PLoS Biol.* **2**, E24 (2004).

4 Lansink, C. S., Goltstein, P. M., Lankelma, J. V., McNaughton, B. L. & Pennartz, C. M. Hippocampus leads ventral striatum in replay of place-reward information. *PLoS Biol.* **7**, e1000173, doi:10.1371/journal.pbio.1000173 (2009).

5 Wilson, M. A. & McNaughton, B. L. Reactivation of hippocampal ensemble memories during sleep. *Science* **265**, 676-679 (1994).

6 Peyrache, A., Khamassi, M., Benchenane, K., Wiener, S. I. & Battaglia, F. P. Replay of rule-learning related neural patterns in the prefrontal cortex during sleep. *Nat. Neurosci.* **12**, 919-926, doi:10.1038/nn.2337 (2009).

7 Yang, G. *et al.* Sleep promotes branch-specific formation of dendritic spines after learning. *Science* **344**, 1173-1178, doi:10.1126/science.1249098 (2014).

8 Ramanathan, D. S., Gulati, T. & Ganguly, K. Sleep-Dependent Reactivation of Ensembles in Motor Cortex Promotes Skill Consolidation. *PLoS Biol.* **13**, e1002263, doi:10.1371/journal.pbio.1002263 (2015).

9 Girardeau, G., Benchenane, K., Wiener, S. I., Buzsaki, G. & Zugaro, M. B. Selective suppression of hippocampal ripples impairs spatial memory. *Nat. Neurosci.* **12**, 1222-1223, doi:10.1038/nn.2384 (2009).

10 Rasch, B., Büchel, C., Gais, S. & Born, J. Odor cues during slow-wave sleep prompt declarative memory consolidation. *Science* **315**, 1426-1429, doi:10.1126/science.1138581 (2007).

11 Rudoy, J. D., Voss, J. L., Westerberg, C. E. & Paller, K. A. Strengthening individual memories by reactivating them during sleep. *Science* **326**, 1079, doi:10.1126/science.1179013 (2009).

12 Peigneux, P. *et al.* Learned material content and acquisition level modulate cerebral reactivation during posttraining rapid-eye-movements sleep. *Neuroimage.* **20**, 125-134 (2003).

13 Maquet, P. *et al.* Experience-dependent changes in cerebral activation during human REM sleep. *Nat. Neurosci.* **3**, 831-836, doi:10.1038/77744 (2000).

14 Horikawa, T., Tamaki, M., Miyawaki, Y. & Kamitani, Y. Neural decoding of visual imagery during sleep. *Science* **340**, 639-642, doi:10.1126/science.1234330 (2013).

15 Marshall, L., Helgadottir, H., Molle, M. & Born, J. Boosting slow oscillations during sleep potentiates memory. *Nature* **444**, 610-613 (2006).

16 Grosmark, A. D., Mizuseki, K., Pastalkova, E., Diba, K. & Buzsaki, G. REM sleep reorganizes hippocampal excitability. *Neuron* **75**, 1001-1007, doi:10.1016/j.neuron.2012.08.015 (2012).

17 Himmer, L., Müller, E., Gais, S. & Schönauer, M. Sleep-mediated memory consolidation depends on the level of integration at encoding. *Neurobiol. Learn. Mem.* **137**, 101-106, doi:10.1016/j.nlm.2016.11.019 (2017).

18 Gais, S., Lucas, B. & Born, J. Sleep after learning aids memory recall. *Learn. Mem.* **13**, 259-262 (2006).

19 Ji, D. & Wilson, M. A. Coordinated memory replay in the visual cortex and hippocampus during sleep. *Nat. Neurosci.* **10**, 100-107, doi:10.1038/nn1825 (2007).

20 Euston, D. R., Tatsuno, M. & McNaughton, B. L. Fast-forward playback of recent memory sequences in prefrontal cortex during sleep. *Science* **318**, 1147-1150, doi:10.1126/science.1148979 (2007).

21 Buzsáki, G. Two-stage model of memory trace formation: a role for "noisy" brain states. *Neuroscience* **31**, 551-570 (1989).

22 Peigneux, P. *et al.* Are spatial memories strengthened in the human hippocampus during slow wave sleep? *Neuron* **44**, 535-545 (2004).

23 Dewald, J. F., Meijer, A. M., Oort, F. J., Kerkhof, G. A. & Bogels, S. M. The influence of sleep quality, sleep duration and sleepiness on school performance in children and adolescents: A meta-analytic review. *Sleep Med Rev* **14**, 179-189, doi:10.1016/j.smrv.2009.10.004 (2010).

24 Schönauer, M., Geisler, T. & Gais, S. Strengthening procedural memories by reactivation in sleep. *J. Cogn. Neurosci.* **26**, 143-153, doi:10.1162/jocn_a_00471 (2014).

25 Louie, K. & Wilson, M. A. Temporally structured replay of awake hippocampal ensemble activity during rapid eye movement sleep. *Neuron* **29**, 145-156 (2001).

26 Pavlides, C. & Winson, J. Influences of hippocampal place cell firing in the awake state on the activity of these cells during subsequent sleep episodes. *J. Neurosci.* **9**, 2907-2918 (1989).

27 Giuditta, A. *et al.* The sequential hypothesis of the function of sleep. *Behav. Brain Res.* **69**, 157-166 (1995).

28 Genzel, L., Kroes, M. C., Dresler, M. & Battaglia, F. P. Light sleep versus slow wave sleep in memory consolidation: a question of global versus local processes? *Trends Neurosci.* **37**, 10-19, doi:10.1016/j.tins.2013.10.002 (2014).

29 Ambrosini, M. V. & Giuditta, A. Learning and sleep: the sequential hypothesis. *Sleep Medicine Reviews* **5**, 477-490 (2001).

30 Ackermann, S. & Rasch, B. Differential effects of non-REM and REM sleep on memory consolidation? *Curr. Neurol. Neurosci. Rep.* **14**, 430, doi:10.1007/s11910-013-0430-8 (2014).

31 Abel, T., Havekes, R., Saletin, J. M. & Walker, M. P. Sleep, plasticity and memory from molecules to whole-brain networks. *Curr. Biol.* **23**, R774-788, doi:10.1016/j.cub.2013.07.025 (2013).

32 Boly, M. *et al.* Hierarchical clustering of brain activity during human nonrapid eye movement sleep. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 5856-5861, doi:10.1073/pnas.1111133109 (2012).

33 Sterpenich, V. *et al.* Memory reactivation during rapid eye movement sleep promotes its generalization and integration in cortical stores. *Sleep* **37**, 1061-1075, 1075A-1075B, doi:10.5665/sleep.3762 (2014).

34 Bourtchouladze, R. *et al.* Different training procedures recruit either one or two critical periods for contextual memory consolidation, each of which requires protein synthesis and PKA. *Learn. Mem.* **5**, 365-374 (1998).

35 Igaz, L. M., Vianna, M. R., Medina, J. H. & Izquierdo, I. Two time periods of hippocampal mRNA synthesis are required for memory consolidation of fear-motivated learning. *J. Neurosci.* **22**, 6781-6789, doi:20026642 (2002).

36 Davis, R. L. Traces of Drosophila memory. *Neuron* **70**, 8-19, doi:10.1016/j.neuron.2011.03.012 (2011).

37 Dubnau, J. & Chiang, A. S. Systems memory consolidation in Drosophila. *Curr. Opin. Neurobiol.* **23**, 84-91, doi:10.1016/j.conb.2012.09.006 (2013).

38 Smith, C. Sleep states and memory processes. *Behav. Brain Res.* **69**, 137-145 (1995).

39 Prince, T. M. *et al.* Sleep deprivation during a specific 3-hour time window post-training impairs hippocampal synaptic plasticity and memory. *Neurobiol. Learn. Mem.* **109**, 122-130, doi:10.1016/j.nlm.2013.11.021 (2014).

40 Stickgold, R., Whidbee, D., Schirmer, B., Patel, V. & Hobson, J. A. Visual discrimination task improvement: a multi-step process occurring during sleep. *J. Cogn. Neurosci.* **12**, 246-254 (2000).

41 Scholz, J., Klein, M. C., Behrens, T. E. & Johansen-Berg, H. Training induces changes in white-matter architecture. *Nat. Neurosci.* **12**, 1370-1371, doi:10.1038/nn.2412 (2009).

42 Schabus, M. *et al.* Sleep spindles and their significance for declarative memory consolidation. *Sleep* **27**, 1479-1485 (2004).

43 Bergmann, T. O., Molle, M., Diedrichs, J., Born, J. & Siebner, H. R. Sleep spindle-related reactivation of category-specific cortical regions after learning face-scene associations. *Neuroimage* **59**, 2733-2742, doi:10.1016/j.neuroimage.2011.10.036 (2012).

44 Schreiner, T. & Rasch, B. Boosting Vocabulary Learning by Verbal Cueing During Sleep. *Cereb. Cortex*, doi:10.1093/cercor/bhu139 (2014).

45 Lisman, J. E. & Jensen, O. The theta-gamma neural code. *Neuron* **77**, 1002-1016, doi:10.1016/j.neuron.2013.03.007 (2013).

46 Walker, M. P. & van der Helm, E. Overnight therapy? The role of sleep in emotional brain processing. *Psychol. Bull.* **135**, 731-748, doi:10.1037/a0016570 (2009).

47 Jackson, C. *et al.* Dynamics of a memory trace: effects of sleep on consolidation. *Curr. Biol.* **18**, 393-400, doi:10.1016/j.cub.2008.01.062 (2008).

48 Schreiner, T., Lehmann, M. & Rasch, B. Auditory feedback blocks memory benefits of cueing during sleep. *Nat. Commun.* **6**, 8729, doi:10.1038/ncomms9729 (2015).

49 Marzano, C. *et al.* Recalling and forgetting dreams: theta and alpha oscillations during sleep predict subsequent dream recall. *J. Neurosci.* **31**, 6674-6683, doi:10.1523/JNEUROSCI.0412-11.2011 (2011).

50 Steriade, M., McCormick, D. A. & Sejnowski, T. J. Thalamocortical oscillations in the sleeping and aroused brain. *Science* **262**, 679-685 (1993).

51 Massimini, M. *et al.* Breakdown of cortical effective connectivity during sleep. *Science* **309**, 2228-2232, doi:10.1126/science.1117256 (2005).

52 Nadasdy, Z., Hirase, H., Czurko, A., Csicsvari, J. & Buzsaki, G. Replay and time compression of recurring spike sequences in the hippocampus. *J. Neurosci.* **19**, 9497-9507 (1999).

53 Staresina, B. P., Alink, A., Kriegeskorte, N. & Henson, R. N. Awake reactivation predicts memory in humans. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 21159-21164, doi:10.1073/pnas.1311989110 (2013).

54 Deuker, L. *et al.* Memory consolidation by replay of stimulus-specific neural activity. *J. Neurosci.* **33**, 19373-19383, doi:10.1523/JNEUROSCI.0414-13.2013 (2013).

55 Tambini, A. & Davachi, L. Persistence of hippocampal multivoxel patterns into postencoding rest is related to memory. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 19591-19596, doi:10.1073/pnas.1308499110 (2013).

56 Oldfield, R. C. The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia* **9**, 97-113 (1971).

57 Roenneberg, T. *et al.* Epidemiology of the human circadian clock. *Sleep medicine reviews* **11**, 429-438, doi:10.1016/j.smrv.2007.07.005 (2007).

58 Minear, M. & Park, D. C. A lifespan database of adult facial stimuli. *Behav. Res. Methods* **36**, 630-633 (2004).

59 Iidaka, T., Matsumoto, A., Haneda, K., Okada, T. & Sadato, N. Hemodynamic and electrophysiological relationship involved in human face processing: evidence from a combined fMRI-ERP study. *Brain Cogn.* **60**, 176-186, doi:10.1016/j.bandc.2005.11.004 (2006).

60 Atkinson, A. P. & Adolphs, R. The neuropsychology of face perception: beyond simple dissociations and functional selectivity. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **366**, 1726-1738, doi:10.1098/rstb.2010.0349 (2011).

61 O'Craven, K. M., Downing, P. E. & Kanwisher, N. fMRI evidence for objects as the units of attentional selection. *Nature* **401**, 584-587, doi:10.1038/44134 (1999).

62 Pourtois, G., Schwartz, S., Spiridon, M., Martuzzi, R. & Vuilleumier, P. Object representations for multiple visual categories overlap in lateral occipital and medial fusiform cortex. *Cereb. Cortex* **19**, 1806-1819, doi:10.1093/cercor/bhn210 (2009).

63 Rechtschaffen, A. & Kales, A. *A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects*. (Brain Information Service, University of California, 1968).

64 Fan, J. & Fan, Y. High dimensional classification using features annealed independence rules. *Ann. Stat.* **36**, 2605-2637 (2008).

65 Norman, K. A., Polyn, S. M., Detre, G. J. & Haxby, J. V. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cog. Sci.* **10**, 424-430, doi:10.1016/j.tics.2006.07.005 (2006).

66 Delorme, A. & Makeig, S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* **134**, 9-21, doi:10.1016/j.jneumeth.2003.10.009 (2004).

67 Jamalabadi, H., Alizadeh, S., Schonauer, M., Leibold, C. & Gais, S. Classification based hypothesis testing in neuroscience: Below-chance level classification rates and overlooked statistical properties of linear parametric classifiers. *Hum. Brain Mapp.* **37**, 1842-1855, doi:10.1002/hbm.23140 (2016).

Acknowledgements

This work was supported by Deutsche Forschungsgemeinschaft grant GA730/3-1.

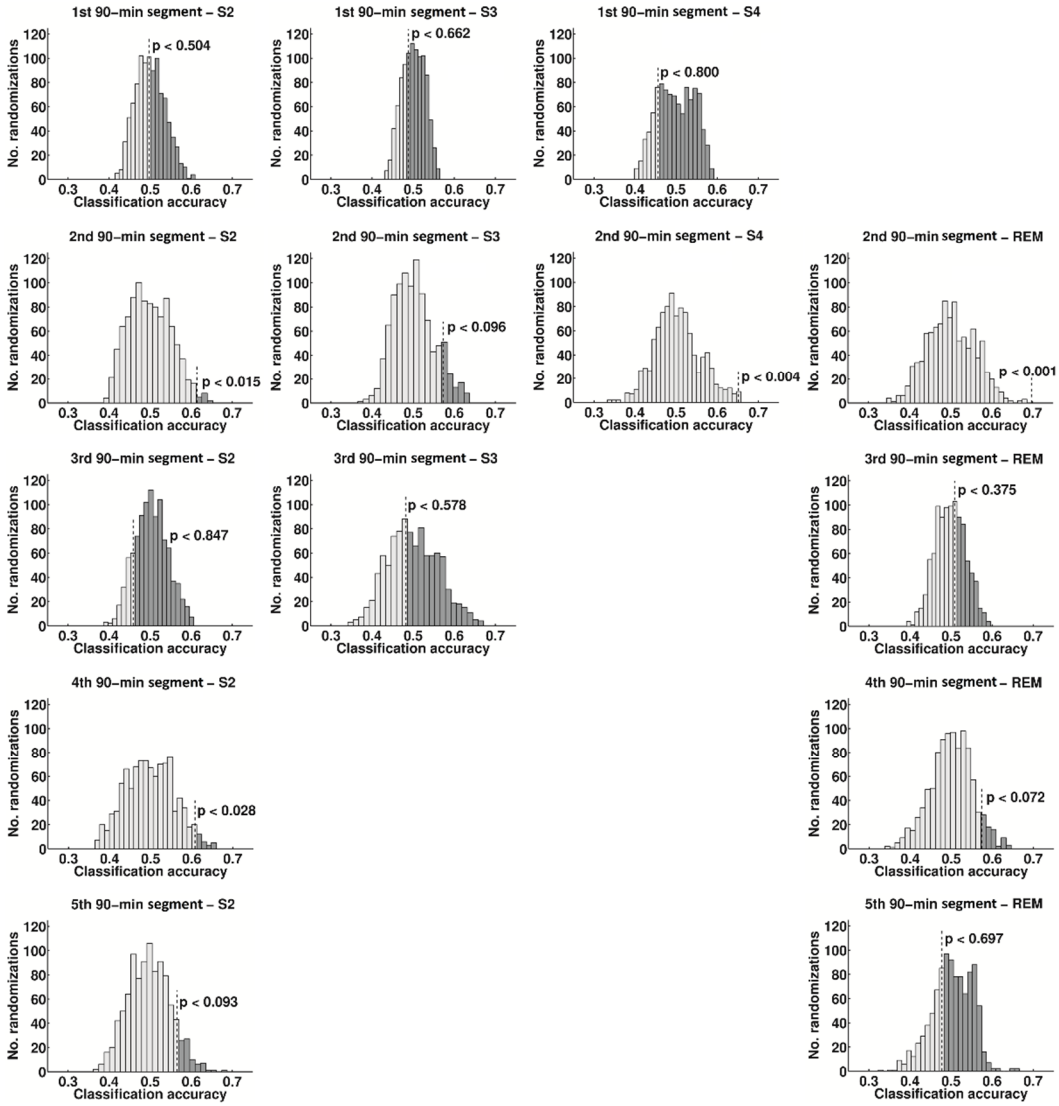
Author contributions

MS, AP, and SG designed the experiments. MS, AA, and AP collected the data. MS, SA, and HJ, analyzed the data. MS, SA, HJ, and SG wrote the manuscript.

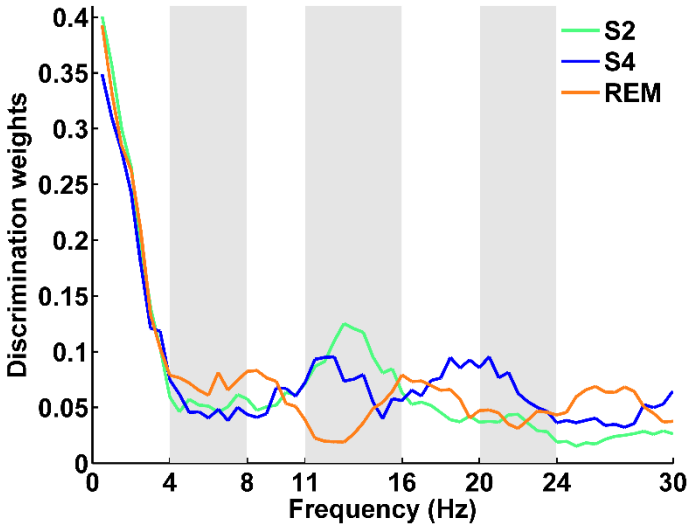
Competing financial interests

The authors declare no conflict of interest.

Supplementary Information



Supplementary Figure 1: Randomization statistics for classification in all segments (rows) and sleep stages (columns). Dark grey areas indicate those randomizations where classification accuracy for randomly labeled data exceeded the classification accuracy obtained with correctly labeled data.



Supplementary Figure 2: Absolute classification weights for the outer loop SVM. Note that weights estimated in the outer loop closely resemble those obtained in the inner loop of the two-step classification procedure (Figure 5).

Supplementary Table 1. Memory sensitivity d' in the face and house learning conditions

	pre	post	difference	p-value
Face pictures	3.72 ± 0.12	3.66 ± 0.12	-0.07 ± 0.04	0.116
House pictures	3.42 ± 0.13	3.34 ± 0.14	-0.08 ± 0.05	0.167

Values are given as mean ± SEM. Two sided t-test for dependent measures is reported. Note that no significant forgetting occurred across the night.

Supplementary Table 2. Correlations between total time spent in sleep stages and memory consolidation (difference in d' post-pre) over sleep for all available nights

	r	*p*	*n*
S2	-0.139	0.272	64
S3	0.106	0.405	64
S4	**0.254***	**0.043**	64
REM	-0.048	0.707	64

**Significant two-sided test at threshold of α < 0.05; Spearman's rho is reported.*

Supplementary Table 3. Correlations between classifier performance (probability estimates for classification) and memory consolidation (difference in d' post-pre) over sleep for all available nights

	difference		**pre**		**post**		
	r	*p*	*r*	*p*	*r*	*p*	*n*
S2 sleep	0.099	0.436	0.023	0.859	0.044	0.733	64
SWS sleep	**0.329****	**0.008**	-0.055	0.667	0.065	0.608	64
REM sleep	-0.199	0.142	0.069	0.611	-0.036	0.791	56

*** Significant two-sided test at threshold of α < 0.01; Spearman's rho is reported.*

Supplementary Table 4. Levels of fatigue in the face and house learning conditions

	Face night	House night	p-value
evening	5.3 ± 2.0	5.5 ± 1.8	0.772
morning	3.7 ± 1.9	3.6 ± 1.6	0.924

Values are given as mean ± SD. Participants were asked to rate their sleepiness on a visual analogue scale with the end points 0 (not tired at all) and 10 (very tired). Two sided t-test for dependent measures is reported.

Supplementary Table 5. Sleep data (mean ± SD)

	W	S1	S2	S3	S4	REM
1st 90-min segment	20.3 ± 11.8	4.8 ± 2.7	29.9 ± 11.8	14.2 ± 6.7	17.9 ± 13.8	2.4 ± 3.3
2nd 90-min segment	3.5 ± 7.8	2.1 ± 1.9	50.9 ± 12.8	11.1 ± 6.6	10.0 ± 8.7	11.0 ± 6.3
3rd 90-min segment	4.2 ± 10.9	2.2 ± 2.0	48.5 ± 10.9	8.0 ± 5.1	5.1 ± 5.7	20.3 ± 7.4
4th 90-min segment	6.9 ± 12.8	2.7 ± 2.2	49.0 ± 13.4	5.6 ± 5.1	1.8 ± 3.8	21.0 ± 8.2
5th 90-min segment	6.9 ± 11.4	4.9 ± 3.8	42.4 ± 12.0	3.3 ± 4.4	1.5 ± 4.1	26.4 ± 11.2
total	48.2 ± 41.5	18.7 ± 8.9	237.7 ± 40.4	42.5 ± 15.0	36.4 ± 23.7	96.0 ± 23.8

Average sleep latency was 20.1 ± 17.0 min (mean ± SD). Please note that total time does not correspond to the sum of 90-min segment values because participants slept slightly longer than five 90-min sleep segments.

Supplementary Table 6. Number of participants and trials that entered classification in different segments and sleep stages. Only data points with N ≥ 11 and number of trials ≥ 40 for both the face and house learning conditions were entered into analysis in each segment and stage.

	S2		S3		S4		REM	
	N	*trials*	*N*	*trials*	*N*	*trials*	*N*	*Trials*
1st 90-min segment	31	472 ± 47	30	355 ± 100	18	455 ± 84	3	279 ± 118
2nd 90-min segment	32	494 ± 33	20	321 ± 102	12	375 ± 93	18	360 ± 74
3rd 90-min segment	29	483 ± 46	16	300 ± 121	6	344 ± 111	24	417 ± 89
4th 90-min segment	24	478 ± 53	9	252 ± 110	2	257 ± 148	19	443 ± 59
5th 90-min segment	20	454 ± 94	0		0		18	415 ± 115

Values for total number of trials collapsed over the face and house conditions that entered classification, given as mean ± SD.

Chapter 3:

Decoding retrieval success and memory content during short-term memory maintenance

Monika Schönauer, Sarah Alizadeh, Hamidreza Jamalabadi, Mirjam Emmersberger, and Steffen Gais

Manuscript in submission process

Abstract

Apart from coding the particular content of a learning episode, a memory representation must permit successful memory retrieval. Using multivariate pattern classification, we tested whether electrical brain activity recorded during short-term memory maintenance satisfies these conditions, and where identified short-term memory representations reside. In our experiment, participants learned two short-term memory tasks, encoding either pictures of faces or houses, or sequences of digits or letters while brain activity was recorded with EEG. It was possible to decode retrieval success from electrical brain activity during the delay period of both short-term memory tasks. Moreover, we could distinguish whether participants kept pictures of faces or houses in memory, and classifier performance on this problem correlated with successful memory maintenance. Using spatial as well as frequency-based searchlight analyses, we found that distinct brain areas and frequency bands coded for the success versus the content of short-term memory. Frontal and parietal higher frequency bands and alpha activity predicted retrieval success, whereas memory content was represented in temporal and parietal higher frequency ranges, as well as theta activity. We propose that frontal cortex supports memory-related control processes, whereas temporal cortex shows a sensory reinstatement of material content and is part of the wider activated network during memory retention. Interestingly, the only overlap between electrodes coding for retrieval success and memory content was found over medial parietal regions, indicating that a dedicated short-term memory representation resides in medial posterior cortex.

Introduction

The term "short-term memory" describes the temporary maintenance of information in the absence of sensory input (Eriksson et al., 2015). Working memory, as a closely related term, additionally involves processing and manipulating information, next to holding it in a memory buffer (Roux and Uhlhaas, 2014). Most models consent that short-term memory maintenance involves an interaction between long-term memory representations, perceptual representations, and basic processes - such as attention - that are instantiated as a persistent reverberation in neural circuits (Eriksson et al., 2015; Jonides et al., 2008; Larocque et al., 2014).

In this way, short-term memory may be conceptualized as a state of temporarily enhanced accessibility of information that does not automatically entail the encoding of an independent memory trace (Cowan, 2008; D'Esposito and Postle, 2015; Eriksson et al., 2015; Fuster, 2009; Jonides et al., 2008; Lewis-Peacock and Postle, 2008; McElree, 2006; Oberauer, 2005). Consequently, it is still unclear whether there are mechanisms or brain structures unique to short-term memory, or whether these functions emerge from a combination of different processes that can be described in other terms than short-term memory. When conceptualizing short-term memory as such a combination of component processes (Cowan, 2001; D'Esposito and Postle, 2015; Eriksson et al., 2015; Fuster, 2009; Jonides et al., 2008), it is no longer necessary to assume a dedicated short-term memory storage. Yet if such a store existed, what would its properties be and how could it be identified?

In his search for the physical substrate of long-term memory in the brain, Semon proposed defining characteristics that such an engram must fulfill (Schacter, 2001; Semon, 1921). Apart from coding the particular content of a learning episode *(stimulus specificity)*, it should enable correct memory retrieval *(relation to performance)*. These criteria likewise apply to short-term memory

representations. Whereas long-term memory traces require persistent changes in the brain that can endure in a dormant state, short-term memory, contrary to this "passive trace" of long-term memory, may emerge from a temporary activation of neural representations. In line with this view, newer evidence shows that attended items are maintained in short-term memory by persistent neural activity during offline intervals (LaRocque et al., 2013; LaRocque et al., 2016; Lewis-Peacock et al., 2012). This persistent stimulus-related neural activity during short-term maintenance of such novel information may concurrently foster the encoding of new long-term memory representations (Olsson and Poom, 2005). Thus, it seems conceivable that regions related to long-term memory also harbor the specific trace currently kept in short-term memory.

New multivariate pattern classification approaches (MVPC) can test whether brain activity recorded during short-term memory maintenance satisfies necessary mnemonic criteria. For instance, MVPC analysis of functional magnetic resonance imaging (fMRI) data allows decoding memory content from brain activity during the offline short-term memory maintenance period (LaRocque et al., 2013; LaRocque et al., 2016; Lewis-Peacock et al.; Postle, 2015). Similarly, it would be possible to assess whether activity in the same or different areas is related to later retrieval success.

In our experiment, participants performed two types of short-term memory task, remembering different kinds of material. One task required encoding pictures of either faces or houses, recruiting visual short-term memory. During the other task, subjects encoded sequences of digits or letters, which involves verbal rehearsal during the maintenance period. Our main interest was to identify activity that reflects processes related to memory performance, as well as to detect item-specific persistent offline activity. We thus tested whether it is possible to predict retrieval success from electrical brain activity during the memory maintenance interval in both tasks, as well as whether we can decode

the content of the maintained information from brain electrical activity during this delay period.

Brain activity that reflects top-down control processes, such as attention or focus, is likely to support successful memory retrieval, but does not necessarily contain information about the specific material that is kept in short-term memory. Activity that codes for the content kept in short-term memory may either reflect continuous activity in the perceptual circuits that processed the learning material or else the activation of related long-term memory representations. It should be clearly noted that this does not automatically entail behavioral relevance for the task, because a sensory instatement together with functionally and regionally distinct control processes might suffice to give rise to short-term memory functions. If, however, activity in a brain region predicted subsequent memory performance and additionally carried information about the content kept in memory, it would be strong evidence for a dedicated short-term memory storage. Moreover, such a region would be optimally suited to harbor long-term memory representations arising from short-term memory processing.

Results

Retrieval Success (remembered vs. non-remembered)

In a first step, we determined whether it is possible to decode retrieval success from human electrical brain activity. For both the face/house (F/H) task as well as the digit/letter (D/L) task we could predict with an accuracy significantly exceeding chance level whether participants would answer the subsequent probe trial correctly (see Table 1).

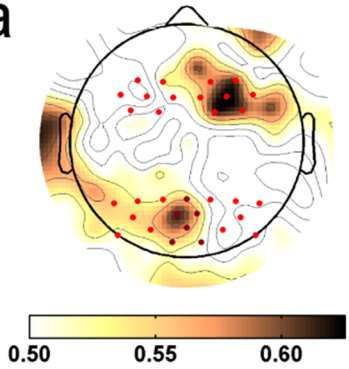
Table 1. Classifier performance when decoding retrieval success

Condition	N_{trial}	CCR	p-value
F/H	80	60.00	0.0464
D/L	76	61.84	0.0200

To explore which brain regions contributed most to successful short-term memory maintenance, we determined the topography of predictive channels in a spatial searchlight analysis, separately for the F/H as well as the D/L tasks (see Methods). During maintenance of face and house pictures, right frontal as well as parietal electrodes contributed most to successful retention (Fig. 1a). Based on these results we chose right frontal, left frontal and left, medial, and right parietal regions of interest (ROIs) and tested whether data from these electrodes alone carries sufficient information for successful classification (Fig 1b, also see Methods). Indeed, it was possible to decode only from activity over right frontal or medial parietal regions, whether short-term memory would be correctly retrieved.

We then proceeded to investigate which frequency bands contributed to short-term memory maintenance in the F/H task in the defined ROIs. For this, we removed class-related information from individual parts of the power spectrum by randomly shuffling the data between conditions. A frequency band contributes to memory maintenance if classification accuracy drops significantly after removing class-related information (see Methods). We found that frontal and parietal beta as well as parietal gamma, but also lower frequency activity and oscillations in the alpha band predicted successful short-term memory retrieval in medial parietal cortex (Fig. 1c).

a



b

CCR	ROI
0.425 *n.s.*	Left Frontal
0.625 *0.025*	Right Frontal
0.538 *n.s.*	Left Parietal
0.625 *0.024*	Medial Parietal
0.450 *n.s.*	Right Parietal

c

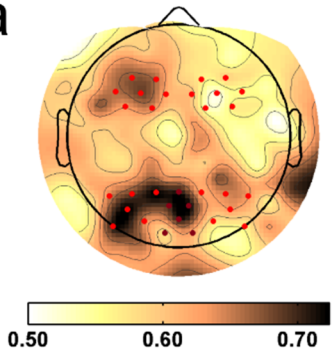
Delta	Theta	Alpha	Beta	Gamma
0.380 *0.101*	0.408 *0.385*	0.411 *0.397*	0.456 *0.761*	0.442 *0.709*
0.651 *0.963*	0.624 *0.606*	0.637 *0.898*	**0.516** *0.014*	0.583 *0.120*
0.551 *0.750*	0.574 *0.886*	0.574 *0.881*	0.512 *0.312*	0.512 *0.292*
0.557 *0.016*	0.587 *0.156*	**0.555** *0.023*	**0.515** *0.016*	**0.557** *0.036*
0.438 *0.351*	0.482 *0.842*	0.478 *0.904*	0.484 *0.783*	0.454 *0.590*

Figure 1. Decoding retrieval success in the F/H task (a) Topography of predictive channels based on a spatial searchlight. A searchlight with a window size of approximately 3.5 cm was moved across all 128 channels that covered the scalp. Topoplot shows smoothed average classification accuracy for the 128 spatial searchlights. Heat bar denotes classification accuracy. (b) When keeping pictures of faces or houses in working memory, right frontal and medial parietal areas contribute to successful maintenance. (c) To assess importance of individual frequency bands for classification, data was shuffled in the bands of interest, which removes class-related information. In frontal cortex, activity in the beta frequency range predicted retrieval success. Also in medial parietal cortex, beta as well as gamma activity informed about whether memory was correctly maintained. Next to these higher frequencies, lower frequency delta and alpha activity was predictive of retrieval success when keeping pictures in memory. Electrode positions in the ROIs are marked as red dots. Gray shading and bold font denote significance determined by permutation tests at a level of $p < 0.05$. Stars indicate significance after FDR correction at the levels of * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

During maintenance of digits and letters, left frontal as well as left and medial parietal electrodes were most informative about retrieval success (Fig. 2a). Thus, results showed a broadly similar topography to that of the face and house pictures. We again defined left and right frontal, as well as left, medial, and right parietal ROIs to assess which areas significantly contributed to successful memory trials. Our analysis shows that left frontal, left parietal and medial parietal regions are significantly involved in successful short-term maintenance in the D/L task (Fig. 2b). Note that these results remain significant after false discovery rate (FDR) correction. Again, we assessed which frequencies

promoted later correct memory retrieval. Higher oscillatory activity in the beta and gamma bands over left frontal cortex, but also over left and medial parietal cortex contributed significantly to successful maintenance of digits and letters. Moreover, left and medial parietal alpha activity as well as activity in the theta and lower delta frequency band over medial parietal cortex enhanced prediction of retrieval success. Thus, especially in the parietal cortex, but also in frontal cortex, activity in similar frequency ranges determined whether short-term memory content would later be correctly retrieved during maintenance in both the F/H and D/L tasks.

a



b

c

CCR ROI		Delta	Theta	Alpha	Beta	Gamma
0.658** *0.003*	**Left Frontal**	0.620 *0.112*	0.621 *0.157*	0.627 *0.165*	**0.584** *0.050*	**0.525*** *0.002*
0.500 *n.s.*	**Right Frontal**	0.521 *0.949*	0.529 *0.960*	0.512 *0.826*	0.545 *0.914*	0.486 *0.422*
0.750*** *< 0.001*	**Left Parietal**	0.677 *0.063*	0.686 *0.088*	**0.690*** *0.011*	**0.584*** *0.001*	**0.606*** *0.001*
0.697** *0.001*	**Medial Parietal**	**0.633** *0.044*	**0.630** *0.045*	**0.618*** *0.007*	**0.601*** *0.007*	**0.594** *0.022*
0.605 *0.051*	**Right Parietal**	0.591 *0.338*	0.578 *0.219*	0.623 *0.838*	0.561 *0.149*	0.585 *0.342*

Figure 2: Decoding retrieval success in the D/L task (a) Topography of predictive channels based on a spatial searchlight. A searchlight with a window size of approximately 3.5 cm was moved across all 128 channels that covered the scalp. Topoplot shows smoothed average classification accuracy for the 128 spatial searchlights. Heat bar denotes classification accuracy. (b) Left frontal and parietal areas were involved in successful retention of digits or letters. (c) To assess importance of individual frequency bands for classification, data was shuffled in the bands of interest. Frontal higher frequency activity in the beta and gamma bands predicted successful retrieval. Similarly, beta and gamma activity was informative about the success of memory maintenance in both left parietal as well as medial parietal cortex, where we additionally observe contributions of the alpha band. In medial parietal cortex, classification accuracy likewise dropped significantly if class-related information was removed from the theta band and the lower frequency delta band. Electrode positions in the ROIs are marked as red dots. Gray shading and bold font denote significance determined by permutation tests at a level of $p < 0.05$. Stars indicate significance after FDR correction at the levels of * $p < 0.05$, ** $p < 0.01$,

Short-term memory content (faces vs houses and digits vs letters)

Decoding the content held in short-term memory was only possible for the F/H Sternberg task, in the D/L Sternberg task, classification remained at chance level (see Table 2).

Table 2. Classifier performance when decoding memory content

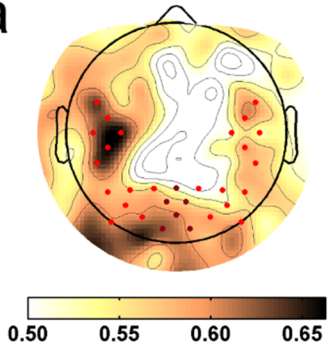
Condition	N_{trial}	CCR	p-value
F/H	80	61.25	0.022
D/L	76	48.68	n.s.

Stronger and more faithful memory processing during retention should result in improved classifiability of the content that is being maintained. Classifier performance may thus be a good indicator for both the strength and the fidelity of information maintenance. We found that classifier performance correlated positively with short-term memory performance measured as memory sensitivity index d' ($r = 0.313$; $p = 0.049$), indicating that continuous and faithful processing of the previously learned material is instrumental for successful short-term memory maintenance.

Analogous to the analysis of retrieval success reported above, we moved a spatial searchlight across all 128 channels to assess which electrodes carry the most information about the content kept in short-term memory. Mainly electrodes over temporal and lateral occipital, but also over parietal areas were informative about whether faces or houses were being maintained in the F/H task (Fig. 3a). We thus defined left and right temporal as well as left, right, and medial parietal ROIs to test which areas carry significant information about the content of short-term memory. We found that it is possible to decode from activity over both temporal and medial parietal cortex whether faces or houses are maintained (Fig. 3b). These results remain significant after FDR correction. In temporal cortex, only beta activity contributed significantly to this distinction.

Over medial parietal cortex, oscillatory activity in the theta, beta, and gamma bands was critical for decoding memory content.

a



b

c

CCR ROI		Delta	Theta	Alpha	Beta	Gamma
0.638** *0.002*	**Left Temporal**	0.599 *0.128*	0.610 *0.185*	0.624 *0.347*	**0.522** *0.006*	0.641 *0.595*
0.588 *0.067*	**Right Temporal**	0.598 *0.852*	0.573 *0.371*	0.553 *0.287*	0.524 *0.074*	0.569 *0.333*
0.550 *n.s.*	**Left Parietal**	0.508 *0.110*	0.544 *0.485*	0.573 *0.771*	0.532 *0.372*	0.513 *0.238*
0.625* *0.017*	**Medial Parietal**	0.600 *0.119*	**0.574** *0.034*	0.601 *0.194*	**0.547** *0.049*	**0.547** *0.043*
0.500 *n.s.*	**Right Parietal**	0.514 *0.806*	0.509 *0.675*	0.522 *0.864*	0.517 *0.679*	0.517 *0.687*

Figure 3: Decoding memory content in the F/H task. (a) Topography of predictive channels based on a spatial searchlight. A searchlight with a window size of approximately 3.5 cm was moved across all 128 channels that covered the scalp. Topoplot shows smoothed average classification accuracy for the 128 spatial searchlights. Heat bar denotes classification accuracy. (b) Memory content in the F/H task could be decoded from left temporal and medial parietal regions. (c) To assess importance of individual frequency bands for classification, data was shuffled in the bands of interest. In temporal cortex, only information in the beta band was crucial to predict memory content. In medial parietal cortex, classification accuracy dropped significantly if information from the theta band or from the higher frequency beta and gamma bands was removed. Electrode positions in the ROIs are marked as red dots. Gray shading and bold font denote significance determined by permutation tests at a level of $p < 0.05$. Stars indicate significance after FDR correction at the levels of * $p < 0.05$, ** $p < 0.01$.

As a proof of principle, we performed the same spatial searchlight and ROI analysis also for the D/L task, where it was not possible to decode memory content from electrical brain activity. As expected, classification accuracies over all electrodes are low and it was not possible to predict with an activity significantly exceeding the chance level whether digits or letters were maintained from activity over the individual ROIs (see Fig. 4).

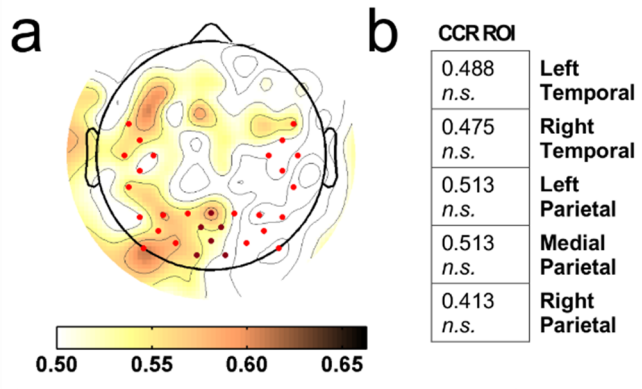


Figure 4: Decoding memory content in the D/L task. (a) Topography of predictive channels based on a spatial searchlight. A searchlight with a window size of 4 channels was moved across all 128 channels that covered the scalp. Topoplot shows smoothed average classification accuracy for the 128 spatial searchlights. Heat bar denotes classification accuracy. In line with the finding that whole-brain classification of memory content did not reach significance in the D/L task, overall accuracies in the spatial searchlight are low and classification did not reach significance in any of the defined regions of interest (b).

Discussion

It is possible to predict, based on brain activity recorded during the delay period of a short-term memory task, whether the memory content will later be correctly recalled. We find that activity in frontal as well as parietal areas critically contributes to successful maintenance, regardless of whether pictures of faces and houses or digits and letters are kept in memory. Similarly, we could decode whether participants were keeping pictures of faces or pictures of houses in memory, with activity over temporal and parietal areas most informative. Interestingly, classification accuracy on this problem correlated with behavioral performance in the short-term memory task, indicating that processing of the learning material critically contributes to successful memory retrieval. Frontal brain activity selectively coded for retrieval success whereas temporal brain activity selectively informed about memory content. The only overlap of electrodes predictive for both successful retrieval as well as the

particular content kept in memory was found over medial parietal areas. We thus suggest that a short-term memory representation is formed and rehearsed in medial posterior cortex.

Frontal higher frequency activity codes for retrieval success

We find that frontal higher frequency activity in the beta band predictive of retrieval success regardless of whether faces and houses or digits and letters are kept in short-term memory. Some studies on short-term memory retention have also reported beta activity (Palva et al., 2011; Roberts et al., 2013), yet its specific function has so far remained unclear (Roux and Uhlhaas, 2014). Since we find an involvement of beta activity in the retention of both visuospatial as well as verbalizable material, we propose that it represents a domain-general mechanism of memory maintenance. Moreover, frontal gamma activity contributed to successful working memory maintenance in the D/L task, which is in line with previous observations (Gotts et al., 2013). Interestingly, frontal areas held no information about memory content, which supports the idea that frontal activity reflects memory-related control processes that are independent of the material content that is being retained (deBettencourt et al., 2017; Sreenivasan et al., 2014). The relevant frontal activity was right lateralized for pictures of faces and houses and left lateralized for digits and letters. It has been shown that participants internally rehearse letter or digit stimuli during the maintenance interval using speech related processing, which may explain the left-sided lateralization of predictive signals (Baddeley, 2012). Maintenance of visual information as face and house pictures, on the other hand, is not facilitated by verbalization, and may rather reflect the scenic and spatial visual aspects of the learned material, for which right-sided lateralization has been observed previously (Roux et al., 2012; Tallon-Baudry et al., 1998). Higher frequency beta in frontal cortex thus represents a domain-general mechanism of short-term memory control that displays domain-specific lateralization, but is independent of the specific content that is being processed.

Sensory processing areas harbor information about short-term memory content

We could decode whether pictures of faces or houses were held in short-term memory from temporal regions and lateral occipital areas. These are associated with the processing of material-specific visual information of the two image categories. A recent study detected category-specific signatures of face and house processing using MVPC methods on EEG data while the stimuli were online (Jacques et al., 2016). Temporal areas were involved in processing of faces while medial and dorsal occipital cortex were activated for houses, which is in line with findings from functional magnetic resonance imaging (fMRI) data (Epstein and Kanwisher, 1998; Haxby et al., 1999; Kanwisher et al., 1997; Vuilleumier et al., 2001). Moreover, fMRI studies using MVPC approaches have found that short-term retention of familiar object, faces and scene and body stimuli can be decoded from the ventral occipito-temporal cortex (Han et al., 2013; Nelissen et al., 2013; Sreenivasan et al., 2014). Since activity in both temporal and lateral occipital regions was not related to retrieval success, we suggest that it reflects a reinstatement of the sensory information or activated long-term memory associated with the content retained in short-term memory which accompanies the activation of a wider network of brain regions during memory processing. In temporal cortex, short-term memory content could be decoded from activity in the beta frequency range. We thus suggest that the relevant content-related activity resides in beta frequencies, again underlining the importance of beta activity from short-term mnemonic functions.

Medial parietal cortex related to both retrieval success as well as memory content

The only overlap of channels that were informative about both retrieval success as well as memory content was found over medial parietal regions. The posterior parietal cortex has often been implicated in the context of short-term memory (D'Esposito and Postle, 2015; Harvey et al., 2012). Recent studies found

item-specific memory representations during short-term memory tasks in parietal cortex (Ester et al., 2015; Sarma et al., 2016). This finding further strengthens the notion that short-term memory representations reside in posterior cortical regions that are both material-specific and relate to later memory performance. Interestingly, it has recently proposed that posterior parietal regions form a long-term memory network (Gilmore et al., 2015). Activity over medial parietal regions could thus reflect such activated long-term memory, which has been proposed to play a major part in short-term memory retention (Eriksson et al., 2015). We have shown that material-specific memory presentations are rapidly established in parietal cortex during the course of visuo-spatial learning (Brodt et al., 2016). It is therefore enticing to speculate that transient memory representations observed in parietal cortex during short-term memory tasks may be stabilized and become long-lasting over rehearsal or time.

When considering results from both the F/H as well as the D/L task, the whole frequency range contributed to successful memory maintenance in parietal cortex. It should however be noted that a large number of independent tests were conducted and theta activity was only predictive of retrieval success in the D/L task. Activity in the alpha, beta and gamma frequency ranges, however, remained significant predictors of successful memory retention after correction for multiple comparisons. Decoding of memory content in parietal cortex depended on activity in the theta, beta and gamma frequency bands. Given previous literature, we would suggest that theta might coordinate processing of content-related higher frequency activity like beta and gamma oscillations during the delay period in short-term memory tasks (Jensen and Lisman, 2005; Roux and Uhlhaas, 2014).

A role of beta in short-term memory maintenance going beyond control processes

If information in the beta band was removed from the power spectrum, classification accuracy dropped significantly in all analyses. This strongly indicates that beta activity holds a functionally important role both in control of short-term memory processes as well as in coding more specifically for memory content. In support of this idea, beta predictive of retrieval success was observed over both frontal and parietal areas, whereas beta that coded for memory content was observed over both lateral temporal regions, which have been shown to be involved in face and house stimulus processing, as well as over parietal cortex. Roberts (2013) has found enhanced beta band activity over posterior regions during the delay period in correct short-term memory trials. Similarly, Palva (2011) reported load-dependent strengthening in frontoparietal beta activity during a visual short-term memory delay period.

Conclusions

We found that electrical brain activity in the frontal, temporal and parietal cortices is related to either successful working memory maintenance or coding the content of what needs to be remembered. Our data suggest that frontal cortex supports memory-related control processes that are domain general, whereas activity in the temporal lobe reflects a sensory reinstatement of memory-related content. Since the only overlap between electrodes coding for retrieval success as well as memory content was found over medial parietal electrodes we would argue that a dedicated short-term memory representation is formed in medial posterior cortex, a region recently found to also harbor item-specific memory representations (Brodt et al., 2016; Chen et al., 2017; Gilmore et al., 2015).

Materials and Methods

Subjects. 20 healthy subjects with no history of neurological or psychiatric disorders participated in this experiment. All were students, between 18 and 30 years old, native German speakers and non-smokers. They were right handed as measured by Edinburgh Handedness Inventory (82 ± 18 [mean score ± SD]) (Oldfield, 1971). Each subject visited our laboratory for two separate experimental sessions, each time performing the same two short-term memory tasks. Daytime of testing was kept constant across participants. Participants were told to refrain from drinking alcohol, coffee and tea and from taking any drugs that can affect the central nervous system on the days of the experiment.

Learning Task. During each of the two experimental sessions, subjects learned two Sternberg short-term memory tasks that assessed maintenance of different kinds of material. The two tasks were performed consecutively. In the first Sternberg task, participants memorized 8-item image sequences of either faces or houses (F/H task). In the second Sternberg task, they memorized 7-item sequences of either digits or consonant letters (D/L task; see Fig. 5). Thus, for each kind of stimulus material, short-term memory content was derived from two distinct categories. Sequence length was pretested to achieve intermediate levels of maintenance performance for the different kinds of material. Participants completed 80 maintenance trials in both tasks. Individual trials contained only items of one content category. Stimulus categories were evenly distributed and trial order was randomized. One participant did not participate in the D/L task.

During each trial, individual stimuli from one content category were presented consecutively for 100 ms in random order with an interstimulus-interval of 1 s showing a black screen. The sequence of memory items was followed by a 4-s maintenance interval during which a black screen with fixation dot was shown. Then, subjects were presented with a probe item for 100 ms followed by a 2-s black screen. Then, they had to indicate via key-press whether this stimulus was

part of the previous sequence, yes or no. They had maximally 5 s to give an answer. After an inter-trial interval of 1 s, the next trial was initiated.

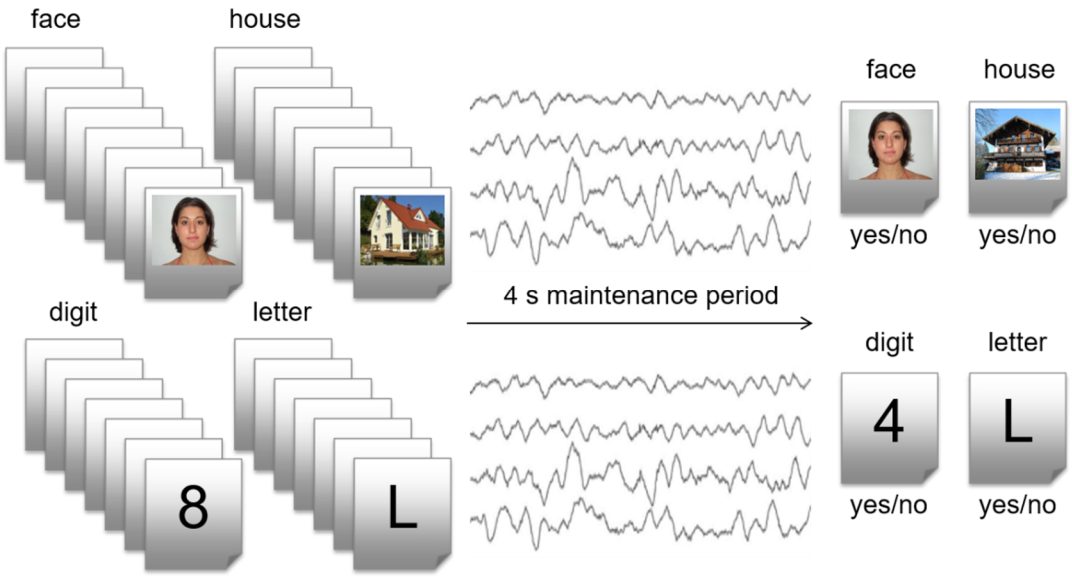


Figure 5: Sternberg task. During the F/H task, participants were instructed to memorize 8 pictures of faces or houses that were presented to them sequentially on a screen. During the following 4-s maintenance period, they had to fixate a dot in the middle of the screen and keep the previously presented information in mind. When the probe item appeared, they fixated it for 2 s until prompted to answer whether it appeared in the studied row of pictures. The D/L task followed the exact same procedure, yet only 7 stimuli were presented in one learning trial.

EEG was recorded throughout the experiment. Participants were instructed to fixate the middle of the screen with eyes open, blinking and moving their eyes as little as possible. To allow relaxation, brief breaks were introduced every 5 min that were terminated by the participants when they felt ready to continue the experiment. Both the F/H task and the D/L task lasted approximately 25 min.

For stimuli in the first Sternberg task, 300 images were taken from German online real estate sites, as well as 300 pctures of neutral faces from Minear & Park (Minear and Park, 2004). For the second Sternberg task, individual stimuli were chosen from the digits 0 to 9 and the consonant letters of the alphabet.

Digits and letters were presented in equally sized light gray sprites using dark gray font. Overall luminance was adjusted by slight modifications to the background color for all stimuli. Pictures of faces and houses had the same format and size. All stimuli were presented centered on the middle of the screen.

EEG Recording. EEG was recorded using an active 128 channel Ag/AgCl-electrode system (ActiCap, Brain products, Gilching, Germany) with 1 kHz sampling frequency and a high-pass filter of 0.1 Hz. Electrodes were placed according to the extended international 10–20 electrode system.

EEG Data Preparation. EEG data was analyzed using support vector machine (SVM) multivariate pattern classification (MVPC). Before classification, the EEG data was preprocessed to minimize problems associated with high dimensional, low sample size data (Jamalabadi et al., 2016). First, EEG data from the 4-s maintenance period was artefact corrected and transformed into the frequency domain using Fourier transformation. Artefact rejection was done in a semiautomatic process using custom MATLAB scripts, ensuring that only a minimal number of artefacts remained in the data. We assessed open channels (outliers in overall power), jumps (outliers in amplitude changes) and muscle artefacts (strong amplitudes in power > 25 Hz). Thresholds were automatically detected based on the variance of the data and manually confirmed upon visual inspection of parameter distributions and concurrent inspection of the raw data. Trials containing artefacts were removed from the data set, channels that contained too many trials with artefacts were removed and interpolated using EEGLAB (Delorme and Makeig, 2004). Trials or channels to be removed were determined by an optimization algorithm so that data loss was kept minimal. To get smooth spectra, Welch's method was used for Fourier transformation, averaging over 10 Hamming windows of 4-s length with 95% overlap, resulting in a final data resolution of 0.25 Hz. Data was used up to a maximum frequency of 45 Hz.

Next, we reduced the number of features entering classification in a two-step procedure using both spatial averaging and a channel-based weighted average

(Schönauer et al., 2017)(Fig. 6a). First, electrode power spectra were averaged within a radius of approximately 3 cm around the 32 evenly spread locations of an extended 10-20-system to increase signal-to-noise ratio and reduce dimensionality. We then averaged over all available artefact-free maintenance trials to obtain a reliable estimate of spectral properties. To remove amplitude differences between channels, which are caused by the distance of each channel to the reference electrode, spectra of all channels were separately normalized between zero and one, removing between subject variability in general spectral power. Baseline EEG power spectra are very similar and differences between conditions are of comparably smaller magnitude. In a final data preparation step, we thus emphasized the relevant differences between neighboring frequencies, by applying a spectral sharpening filter. For this, the moving average of 23 neighboring frequency bins (window size: 5.5 Hz) was subtracted from the signal to remove the baseline spectrum.

EEG Multivariate Pattern Classification (MVPC) Analysis. The aim of the present study was to test whether EEG activity during a short-term memory maintenance interval can predict whether retrieval from short-term memory will be successful and whether the EEG contains information about the kind of stimuli that are retained. We thus conducted two separate MVPC analyses on both the face/house and the digit/letter Sternberg data. The first analysis assessed which features of the EEG data predict if the trial can be solved correctly. The second analysis considered whether the EEG data reflects which content category (faces vs. houses, digits vs. letters) is kept in short-term memory, and whether the strength of such off-line content processing is related to memory performance.

For the F/H task, EEG recordings from 40 experimental sessions were analyzed using a classification algorithm developed on the basis of linear support vector machines (SVM). Please note that one participant did not participate in the D/L task, thus only 38 experimental sessions were available for analysis. During data preparation, we reduced the number of channels from 128 to 32 and averaged

over all available trials of each condition in each subject and session, leaving 32 × (45 × 4) data features (# channels, # frequency bins [high-cutoff = 45 Hz, fs = 0.25 Hz]). Because generalizability of the data decreases with number of features that enter classification, we used a stepwise procedure during classification (Fig. 6b). First, data was split into independent training and validation sets. Please note that all 4 data points of a subject (2 conditions × 2 sessions) were allocated to either the training set or the test set following a between-subject classification approach. We then trained one linear SVM per EEG channel to determine how well the different categories can be distinguished based on individual channel data. This channel-based classification was cross-validated on each subject of the training set in a leave-one-out (LOO) procedure, and the resulting accuracies were averaged over all possible cross-validation runs (n-1, with n denoting the number of subjects; again note that all 4 data points of a subject for each task were treated as an individual fold in this procedure). In the second step, the resulting average classification accuracy for each channel was used to calculate a weighted average of data. The main SVM was trained on this weighted training set and classification accuracy was tested on the independent validation set. This complete two-step process was cross-validated in a LOO procedure.

Significance of the classification accuracies in the whole-brain analysis was tested using randomization statistics. The distribution of the null hypothesis was generated by randomly shuffling condition labels of the original data and repeating the complete classification procedure 1001 times. Significance was calculated by determining the percentage of times that a randomly labeled data produced a classification accuracy that was equal or higher to the one found in the correctly labeled data (lower limit of estimate $p < 0.001$).

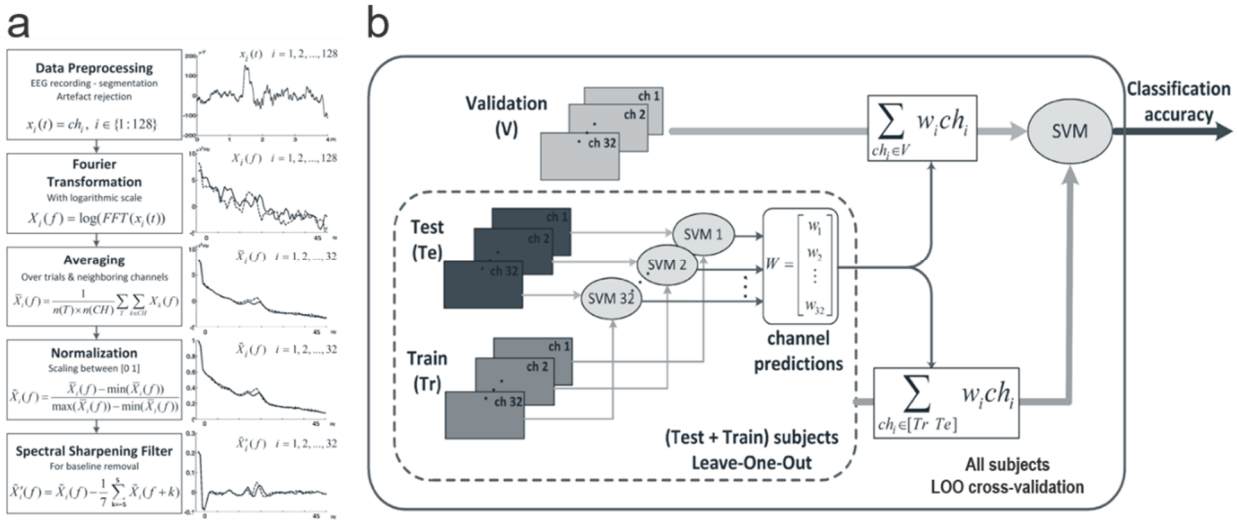


Figure 6: Algorithms used for data preparation and MVPC analysis. (a) After artefact rejection, power spectra of the 4-s memory maintenance trials of the 128-channel EEG recordings were calculated. To reduce the dimensionality of the data and to increase the signal-to-noise ratio, all trails for each content category and session, as well as neighboring channels were averaged. Next, spectra of all channels were normalized separately to make them comparable, and a spectral sharpening filter was applied to remove the baseline spectrum and enhance differences between neighboring frequency bins. (b) In MVPC analysis, training data was strictly separated from validation data. During training, it was again an important goal to reduce dimensionality of the data. Therefore, channels were weighted according to their performance in separate single-channel classifiers. A weighted average of data from all channels was then used to train a classifier to distinguish between two conditions. Finally, classification was tested on independent validation data.

If classification from this whole-brain MVPC analysis yielded significant results, we ran a spatial searchlight to assess the relative importance of different brain regions to classification. Thus, for retrieval success, searchlight analysis was done for both the F/H as well as for the D/L data. In the analysis of short-term memory content, searchlight analysis was similarly conducted for the F/H data, and additionally for the D/L data, where whole-brain MVPC analysis was not possible, to assess the specificity of our method. Searchlight analysis used the same data preparation procedure as before, except that it omitted spatial down-sampling during data preparation, thus considering data from all 128 channels.

The searchlight moved a spatial window with a 3.5 cm radius across the scalp, estimating average classification accuracy in the same two-step classification procedure as the main MVPC analysis. That is, we trained separate linear SVMs for each individual channel in the searchlight in the first step, then used the resulting channel-based averaged data to train and cross-validate the main SVM. For this part of the analysis, power spectral values for all 45 × 4 frequency bins entered classification.

We defined separate regions of interest (ROI) based on the results of the foregoing spatial searchlight analysis to test which brain regions carry information about retrieval success and memory content, respectively. For the analysis of retrieval success, we defined ROIs over the left and right frontal cortex (left frontal electrodes: F1, F3, FFC1h, FFC3h, FFC5h, FC1, FC3; right frontal electrodes: F2, F4, FFC2h, FFC4h, FFC6h, FC2, FC4) as well as the parietal cortex (left parietal electrodes: P1, P3, P5, PPO5h, PO3, P07; medial parietal electrodes: Pz, POz, POO1, POO2, PPO1h, PPO2h; right parietal electrodes: P2, P4, P6, PPO6h, PO4, P08). For the analysis of short-term memory content, we defined ROIs over the left and right temporal cortex (left temporal electrodes: FC5, FCC5h, C3, C5, CP5h, CP5; right temporal electrodes: FC6, FCC6h, C4, C6, CP6h, CP6) as well as the parietal cortex (left parietal electrodes: P1, P3, P5, PPO5h, PO3, P07; medial parietal electrodes: Pz, POz, POO1, POO2, PPO1h, PPO2h; right parietal electrodes: P2, P4, P6, PPO6h, PO4, PO8). To assess which specific frequency features contributed most to classification, we then ran additional searchlight analyses on individual frequency bands of the EEG power spectrum (delta: 1 – 2.75 Hz, theta: 3 – 7.75 Hz, alpha: 8 –11.75 Hz, beta: 12 – 29.75 Hz, gamma: 30 – 45 Hz). That is, keeping data in other frequency bands unchanged, we shuffled data in the target frequency band of the power spectrum 1001 times to remove class-related information, and tested whether this leads to a significant drop in classification accuracy, to examine whether this frequency band critically contributed to retrieval success or to coding memory content. Frequency searchlights were done separately in the pre-defined ROIs

reported above. For the ROI and frequency searchlight analysis, false discovery rate (FDR) corrected significance estimates are reported in addition to the values obtained by permutation testing.

Behavioral Performance. For assessment of memory performance, we calculated the memory sensitivity index d' as the difference of z-values between correctly recognized old items vs. falsely recognized new items (z [hits] – z [false alarms]). To examine whether content reprocessing during short-term memory maintenance is associated with better memory performance, we correlated classifier performance with memory sensitivity d' for the testing probes. Classification accuracy on each of the four data points from a participant's LOO fold in the outer loop was used as an estimate of classifier performance and related to the respective average retrieval success for this content category and session.

References

Baddeley, A. Working memory: theories, models, and controversies. *Annu. Rev. Psychol.* 63: 1-29.

Brodt, S., Pöhlchen, D., Flanagin, V.L., Glasauer, S., Gais, S., Schönauer, M. Rapid and independent memory formation in the parietal cortex. *Proc Natl Acad Sci U S A* 113 (46): 13251-13256.

Chen, J., Leong, Y.C., Honey, C.J., Yong, C.H., Norman, K.A., Hasson, U. Shared memories reveal shared structure in neural activity across individuals. *Nat. Neurosci.* 20 (1): 115-125.

Cowan, N. The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *Behav. Brain Sci.* 24 (1): 87-114; discussion 114-185.

Cowan, N. What are the differences between long-term, short-term, and working memory? *Prog. Brain Res.* 169: 323-338.

D'Esposito, M., Postle, B.R. The cognitive neuroscience of working memory. *Annu. Rev. Psychol.* 66: 115-142.

deBettencourt, M.T., Norman, K.A., Turk-Browne, N.B. Forgetting from lapses of sustained attention. *Psychon. Bull. Rev.*

Delorme, A., Makeig, S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* 134 (1): 9-21.

Epstein, R., Kanwisher, N. A cortical representation of the local visual environment. *Nature* 392 (6676): 598-601.

Eriksson, J., Vogel, E.K., Lansner, A., Bergstrom, F., Nyberg, L. Neurocognitive Architecture of Working Memory. *Neuron* 88 (1): 33-46.

Ester, E.F., Sprague, T.C., Serences, J.T. Parietal and Frontal Cortex Encode Stimulus-Specific Mnemonic Representations during Visual Working Memory. *Neuron* 87 (4): 893-905.

Fuster, J.M. Cortex and memory: emergence of a new paradigm. *J. Cogn. Neurosci.* 21 (11): 2047-2072.

Gilmore, A.W., Nelson, S.M., McDermott, K.B. A parietal memory network revealed by multiple MRI methods. *Trends Cogn Sci* 19 (9): 534-543.

Gotts, S.J., Jo, H.J., Wallace, G.L., Saad, Z.S., Cox, R.W., Martin, A. Two distinct forms of functional lateralization in the human brain. *Proc Natl Acad Sci U S A* 110 (36): E3435-3444.

Han, X., Berg, A.C., Oh, H., Samaras, D., Leung, H.C. Multi-voxel pattern analysis of selective representation of visual working memory in ventral temporal and occipital regions. *Neuroimage* 73: 8-15.

Harvey, C.D., Coen, P., Tank, D.W. Choice-specific sequences in parietal cortex during a virtual-navigation decision task. *Nature* 484 (7392): 62-68.

Haxby, J.V., Ungerleider, L.G., Clark, V.P., Schouten, J.L., Hoffman, E.A., Martin, A. The effect of face inversion on activity in human neural systems for face and object perception. *Neuron* 22 (1): 189-199.

Jacques, C., Retter, T.L., Rossion, B. A single glance at natural face images generate larger and qualitatively different category-selective spatio-temporal signatures than other ecologically-relevant categories in the human brain. *Neuroimage* 137: 21-33.

Jamalabadi, H., Alizadeh, S., Schonauer, M., Leibold, C., Gais, S. Classification based hypothesis testing in neuroscience: Below-chance level classification rates and overlooked statistical properties of linear parametric classifiers. *Hum. Brain Mapp.* 37 (5): 1842-1855.

Jensen, O., Lisman, J.E. Hippocampal sequence-encoding driven by a cortical multi-item working memory buffer. *Trends Neurosci.* 28 (2): 67-72.

Jonides, J., Lewis, R.L., Nee, D.E., Lustig, C.A., Berman, M.G., Moore, K.S. The mind and brain of short-term memory. *Annu. Rev. Psychol.* 59: 193-224.

Kanwisher, N., McDermott, J., Chun, M.M. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J. Neurosci.* 17 (11): 4302-4311.

LaRocque, J.J., Lewis-Peacock, J.A., Drysdale, A.T., Oberauer, K., Postle, B.R. Decoding attended information in short-term memory: an EEG study. *J. Cogn. Neurosci.* 25 (1): 127-142.

Larocque, J.J., Lewis-Peacock, J.A., Postle, B.R. Multiple neural states of representation in short-term memory? It's a matter of attention. *Front Hum Neurosci* 8: 5.

LaRocque, J.J., Riggall, A.C., Emrich, S.M., Postle, B.R. Within-Category Decoding of Information in Different Attentional States in Short-Term Memory. *Cereb. Cortex*.

Lewis-Peacock, J.A., Drysdale, A.T., Oberauer, K., Postle, B.R. Neural evidence for a distinction between short-term memory and the focus of attention. *J. Cogn. Neurosci.* 24 (1): 61-79.

Lewis-Peacock, J.A., Postle, B.R. Temporary activation of long-term memory supports working memory. *J. Neurosci.* 28 (35): 8765-8771.

McElree, B. Accessing recent events. *Psychol Learn Motiv* 46: 155-200.

Minear, M., Park, D.C. A lifespan database of adult facial stimuli. *Behav Res Methods Instrum Comput* 36 (4): 630-633.

Nelissen, N., Stokes, M., Nobre, A.C., Rushworth, M.F. Frontal and parietal cortical interactions with distributed visual representations during selective attention and action selection. *J. Neurosci.* 33 (42): 16443-16458.

Oberauer, K. Control of the contents of working memory--a comparison of two paradigms and two age groups. *J. Exp. Psychol. Learn. Mem. Cogn.* 31 (4): 714-728.

Oldfield, R.C. The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia* 9 (1): 97-113.

Olsson, H., Poom, L. Visual memory needs categories. *Proc Natl Acad Sci U S A* 102 (24): 8776-8780.

Palva, S., Kulashekhar, S., Hamalainen, M., Palva, J.M. Localization of cortical phase and amplitude dynamics during visual working memory encoding and retention. *J. Neurosci.* 31 (13): 5013-5025.

Postle, B.R. The cognitive neuroscience of visual short-term memory. *Curr Opin Behav Sci* 1: 40-46.

Roberts, B.M., Hsieh, L.T., Ranganath, C. Oscillatory activity during maintenance of spatial and temporal information in working memory. *Neuropsychologia* 51 (2): 349-357.

Roux, F., Uhlhaas, P.J. Working memory and neural oscillations: alpha-gamma versus theta-gamma codes for distinct WM information? *Trends Cogn Sci* 18 (1): 16-25.

Roux, F., Wibral, M., Mohr, H.M., Singer, W., Uhlhaas, P.J. Gamma-band activity in human prefrontal cortex codes for the number of relevant items maintained in working memory. *J. Neurosci.* 32 (36): 12411-12420.

Sarma, A., Masse, N.Y., Wang, X.J., Freedman, D.J. Task-specific versus generalized mnemonic representations in parietal and prefrontal cortices. *Nat. Neurosci.* 19 (1): 143-149.

Schacter, D.L. *Forgotten Ideas, Neglected Pioneers: Richard Semon and the Story of Memory*. Psychology Press, 2001.

Schönauer, M., Alizadeh, S., Jamalabadi, H., Abraham, A., Pawlizki, A., Gais, S. Decoding material-specific memory reprocessing during sleep in humans. *Nat. Commun.*

Semon, R. *The Mneme*. G. Allen & Unwin, 1921.

Sreenivasan, K.K., Vytlacil, J., D'Esposito, M. Distributed and dynamic storage of working memory stimulus information in extrastriate cortex. *J. Cogn. Neurosci.* 26 (5): 1141-1153.

Tallon-Baudry, C., Bertrand, O., Peronnet, F., Pernier, J. Induced gamma-band activity during the delay of a visual short-term memory task in humans. *J. Neurosci.* 18 (11): 4244-4254.

Vuilleumier, P., Armony, J.L., Driver, J., Dolan, R.J. Effects of attention and emotion on face processing in the human brain: an event-related fMRI study. *Neuron* 30 (3): 829-841.

Chapter 4:

Decoding cognitive concepts from neuroimaging data using multivariate pattern analysis

Sarah Alizadeh, Hamidreza Jamalabadi, Monika Schönauer, Christian Leibold, and Steffen Gais

Published in NeuroImage (July 2017)

S. Alizadeh, H. Jamalabadi, M. Schönauer, C. Leibold, S. Gais (2017), "Decoding cognitive concepts from neuroimaging data using multivariate pattern analysis". *NeuroImage,* 159:449-458.

Abstract

Multivariate pattern analysis (MVPA) methods are now widely used in life-science research. They have great potential but their complexity also bears unexpected pitfalls. In this paper, we explore the possibilities that arise from the high sensitivity of MVPA for stimulus-related differences, which may confound estimations of class differences during decoding of cognitive concepts. We propose a method that takes advantage of concept-unrelated grouping factors, uses blocked permutation tests, and gradually manipulates the proportion of concept-related information in data while the stimulus-related, concept-irrelevant factors are held constant. This results in a concept-response curve, which shows the relative contribution of these two components, i.e. how much of the decoding performance is specific to higher-order category processing and to lower order stimulus processing. It also allows separating stimulus-related from concept-related neuronal processing, which cannot be achieved experimentally. We applied our method to three different EEG data sets with different levels of stimulus-related confound to decode concepts of digits vs. letters, faces vs. houses, and animals vs. fruits based on event-related potentials at the single trial level. We show that exemplar-specific differences between stimuli can drive classification accuracy to above chance levels even in the absence of conceptual information. By looking into time-resolved windows of brain activity, concept-response curves can help characterize the time-course of lower-level and higher-level neural information processing and detect the corresponding temporal and spatial signatures of the corresponding cognitive processes. In particular, our results show that perceptual information is decoded earlier in time than conceptual information specific to processing digits and letters. In addition, compared to the stimulus-level predictive sites, concept-related topographies are spread more widely and, at later time points, reach the frontal cortex. Thus, our proposed method yields insights into cognitive processing as well as corresponding brain responses.

Introduction

Advances in electrophysiological, genetic, and neuroimaging methods generate ever growing volumes of data. These massively multivariate data sets require methods of analysis which go beyond traditional statistical ANOVA-based approaches (Haynes and Rees 2006; O'Toole et al. 2007; Tong and Pratte 2012). Particularly machine learning methods have seen growing adoption in the life sciences because they can be used to analyze high-dimensional data with great sensitivity (Norman et al. 2006; Haxby et al. 2014). In neuroimaging, multivariate pattern analysis (MVPA) has made it possible not only to investigate differences in brain regional activity during the performance of a task, but also to decode perceptual and mental representations as well as conceptual and semantic information (Kamitani and Tong 2005; Kay et al. 2008; Mitchell et al. 2008; Schwarzlose et al. 2008; Rissman et al. 2010; Simanova et al. 2014).

The complexity of multivariate analysis, however, leads to unexpected problems (Todd et al. 2013; Woolgar et al. 2014; Haynes 2015; Jamalabadi et al. 2016). Here, we will explore the consequences of the high sensitivity of MVPA for differences found between subgroups of trials in cognitive experiments. In classical analyses, two conditions with identical means are considered identical. Differences between trials (caused by different stimuli, subjects, etc.) usually average out on the dependent variable and therefore do not influence the group average. The multivariate nature of MVPA, however, allows differences to accumulate over dimensions (Fan and Fan 2008; Jamalabadi *et al.* 2016). Any differences between individual elements of the categories will be used by MVPA to distinguish between categories, even if the categories themselves have identical centroids. For example, if concept-related features are the intended focus of study, different combinations of low-level, stimulus-specific features like orientation, shape, color, etc. can drive decoding although there is no overall

average difference in these features between both concepts (Haynes and Rees 2006). In fact, MVPA is sensitive to both the effect of interest and to any other confounding factors that drive a difference between conditions (Todd *et al.* 2013; Woolgar *et al.* 2014). Thus, if a data set consists of groups of trials that differ in some stimulus-specific features, MVPA can detect differences that might then be mistakenly attributed to the concept under investigation. In other words, the classifier can use stimulus-specific rather than category-specific features to decode data, effectively predicting stimuli instead of conceptual categories. Therefore, the present paper explores a method to determine the degree to which classification performance is specific to higher order category processing and to lower order stimulus processing.

Consider the following neuro-cognitive experiment, in which the concepts of animate and inanimate objects are to be distinguished based on electrical brain activity. 40 pictures each of six different types of animals (e.g. cow, bear, dog, frog, …) and tools (e.g. knife, scissors, hammer, saw, …) are presented to subjects, with the aim to decode the two conceptual categories from event-related EEG. Since different types of stimuli have features that distinguish them from the other types, the classifier will detect brain responses to individual stimuli based on combinations of their physical features alone (e.g. cows and frogs differ in size, shape and color). As we will show below, these differences between stimulus types will contribute to classification even in the absence of an actual effect of the superordinate concept. We will investigate the relative contribution of these two components, i.e. how much of the decoding performance originates from concept-related information and how much is caused by stimulus differences.

In the following, we will consider the concept-related information as the factor of interest (primary effect) and all the other contributing, concept-irrelevant factors as the nuisance effects. By relabeling the data, we can manipulate the relative contribution of concept (animate, inanimate) and stimulus (cow, frog, knife, scissors, …) to determine the presence of the effect of interest when

nuisance effects are controlled for. The basic idea resembles that of a dose-response curve, in that we systematically vary the amount of concept-related information in the training data set of the classifier to assess how classification performance changes with varying levels of conceptual information. When the effect of concept-related information is completely counterbalanced, decoding performance originates solely from concept-irrelevant nuisance effects, which constitutes our null hypothesis for statistical testing. We will apply this method here in several examples, showing how to separate high-level cognitive concepts from low-level stimulus processing. In particular, we will show how this method can be used to describe the detailed time-course of cognitive concept processing. However, we believe that the basic method can find application in many similar problems.

Method & Results

Suppose that an experiment has the aim to decode conceptual information (e.g. the semantic category) from brain activity. Different exemplars of each category are presented to the subjects and the brain response is recorded. For the sake of simplicity, and without loss of generality, we assume that there are two semantic categories $\mathfrak{A}$ and $\mathfrak{B}$. Each category consists of stimuli coming from $j = 1, 2, \dots, k$ subclasses (see Fig. 1A). For instance, in our example of animals and tools, there are six subclasses per category (cow, bear, dog, frog, … for animals and knife, scissors, hammer, saw, … for tools). We assume that each stimulus is presented n times, resulting in $k \times n$ trials per category. We consider all of the n trials that belong to the jth subclass as one block of data and denote it with $\mathfrak{A}_j$ or $\mathfrak{B}_j$. Therefore, each category consists of k blocks and can be defined as a set.

$$\mathfrak{A} = \{\mathfrak{A}_1, \mathfrak{A}_2, \dots, \mathfrak{A}_k\} \quad , \quad \mathfrak{B} = \{\mathfrak{B}_1, \mathfrak{B}_2, \dots, \mathfrak{B}_k\}$$

Here we are interested in decoding neural activity that is specific to concept processing. To do this, we adopt a systematic approach that gradually eliminates the amount of concept-related information in the data while preserving concept-

irrelevant information. To do this, we generate a number of new sets S_1 and S_2 that comprise varying proportions of elements of $\mathfrak{A}$ and $\mathfrak{B}$. More precisely, S_1 is built by randomly selecting $m = 0, 1, \ldots, k/2$ blocks from category $\mathfrak{A}$ and $(k - m)$ blocks from category $\mathfrak{B}$. Therefore,

$$S_1 = \mathfrak{a}_m \cup \mathfrak{b}_{k-m}\,, \quad \mathfrak{a}_m \subseteq \mathfrak{A}\,, \quad \mathfrak{b}_{k-m} \subseteq \mathfrak{B}$$

where $\mathfrak{a}_m$ and $\mathfrak{b}_{k-m}$ are random subsets of $\mathfrak{A}$ and $\mathfrak{B}$ of size m and $k - m$, respectively. Accordingly,

$$S_2 = (\mathfrak{A} \backslash \mathfrak{a}_m) \cup (\mathfrak{B} \backslash \mathfrak{b}_{k-m})$$

where $\mathfrak{A} \backslash \mathfrak{a}_m$ denotes the set of elements in $\mathfrak{A}$ but not in $\mathfrak{a}_m$, and $\mathfrak{B} \backslash \mathfrak{b}_{k-m}$ represents the elements in $\mathfrak{B}$ but not in $\mathfrak{b}_{k-m}$. Thus, each set S_1 contains m out of k blocks belonging to category $\mathfrak{A}$, while the corresponding sets S_2 contain m out of k blocks belonging to category $\mathfrak{B}$. The ratio of data from categories $\mathfrak{A}/\mathfrak{B}$ in S_1 therefore varies between 0 and $1/2$, and is complemented by S_2. We apply a linear support vector machine (SVM) with cross-validation to distinguish the two sets S_1 and S_2. This process can be repeated up to $\binom{k}{m}\binom{k}{k-m} = \binom{k}{m}^2$ [or $\frac{1}{2}\binom{k}{m}^2$ for $m = k/2$] times to account for random subset selection of $\mathfrak{a}_m$ and $\mathfrak{b}_{k-m}$. Resulting classification accuracies are averaged. The whole procedure is repeated for m ranging from 0 (sets containing only elements of either category $\mathfrak{A}$ or $\mathfrak{B}$) to $k/2$ (two sets with an equal number of elements belonging to categories $\mathfrak{A}$ and $\mathfrak{B}$). It is worth noting that the ratio of m/k which represents the proportion of relabeled subclasses has always a range of 0 to 0.5, regardless of the number of subclasses. Thus, we gradually manipulate the amount of concept-related information differentiating between sets S_1 and S_2.

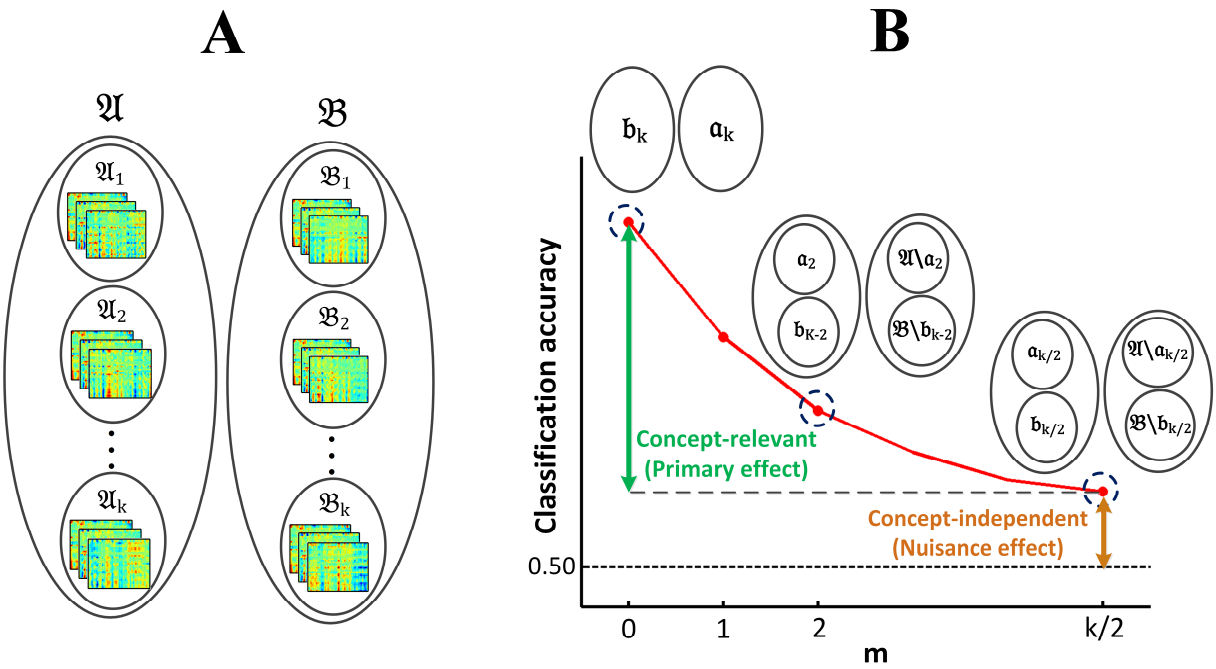


Figure 2: Example of a concept-response curve. (A) Structure of data with two experimental conditions ($\mathfrak{A}$ and $\mathfrak{B}$, e.g. animate and inanimate objects) and k blocks of data per condition. Each block consists of all trials that belong to one subclass (e.g. frogs, cows, hammers, scissors, ...). (B) By changing the number of blocks m in set S_1 belonging to category $\mathfrak{A}$ from 0 to $k/2$, we can change the amount of concept-relevant information distinguishing between sets S_1 and S_2. Each point of the curve is derived from the classification of S_1 versus S_2. $\mathfrak{a}_m$ and $\mathfrak{b}_m$ represent m-block subsets of $\mathfrak{A}$ and $\mathfrak{B}$, respectively. $\mathfrak{A}\backslash\mathfrak{a}_m$ denotes the set of blocks in $\mathfrak{A}$ but not in $\mathfrak{a}_m$ (similar for $\mathfrak{B}\backslash\mathfrak{b}_m$).

We can plot classification accuracy depending on values of m to get a graph that indicates how the response of a classifier changes with varying levels of conceptual information (Fig. 1B). The first point of this concept-response curve ($m = 0$), which corresponds to the classification of category $\mathfrak{A}$ versus $\mathfrak{B}$, represents the total discrimination power driven by both the effect of interest and nuisance effects. In the last point ($m = k/2$), discrimination originates solely from nuisance effects, because the effect of categories $\mathfrak{A}$ and $\mathfrak{B}$ cancel out. This is also the classification performance that we would expect if the null hypothesis that there is no primary effect in the data is true, i.e. the concept in question does not affect brain activity. A concept-response curve as shown in Figure 1B can have several theoretical shapes. Figure 2 shows the four possible, idealized curves that can be obtained. The shape of the curve reveals which

sources of information (primary effect or nuisance effects) drive decoding performance. Depending on the shape of the curve, only a primary effect (Fig. 2A), only a nuisance effect (Fig. 2B), a combination of both (Fig. 2C), or no effect can be detected (Fig. 2D).

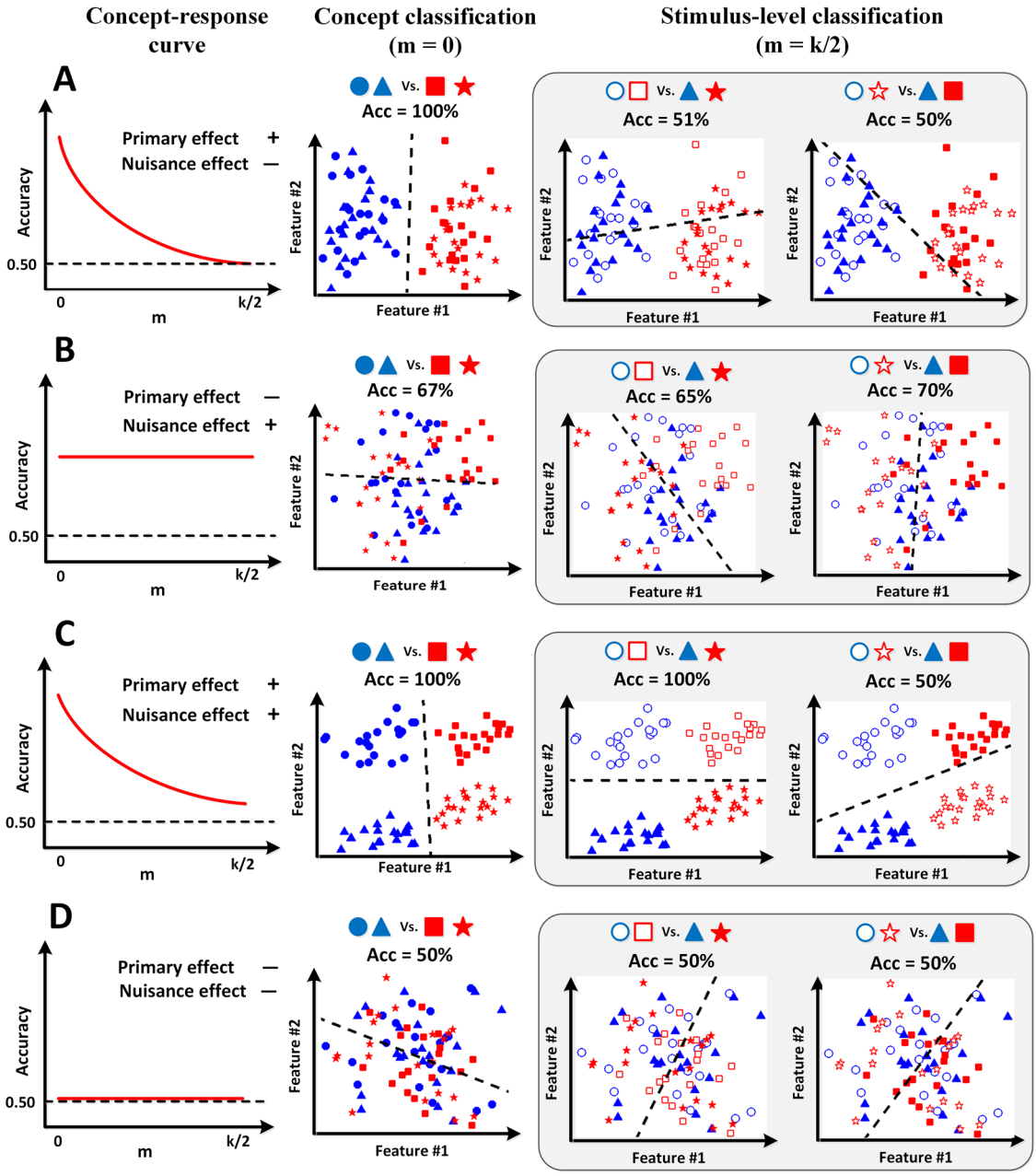


Figure 3: Four possible, idealized concept-response curves that can be obtained from our proposed method. The '+' sign represents a contribution of the primary/nuisance effect to the decoding accuracy. The '-' sign signifies the absence of the corresponding effect. *m*

represents the number of relabeled subclasses with a discrete value from 0 to $k/2$. Scatter plots in the second column illustrate two-dimensional sample data sets with two subclasses per category ($k = 2$). Color and markers represent categories and subclasses respectively. Scatter plots in the gray boxes show the two possible ways of random relabeling of subclasses for $m = k/2$. Filled and open markers represent random sets S_1 and S_2. Classification accuracies (Acc) and decision boundaries were calculated using linear discriminant analysis (LDA). Average classification accuracy for $m = k/2$ is higher than 50% when nuisance effects contribute to decoding performance (B, C). A value of 50% signifies that no relevant nuisance effects exist (A, D).

Although one might assume that classification performance should be at chance level (50%) if the effect of concept cancels out between S_1 and S_2, this is not necessarily the case. As we will show below, classification accuracy in case of missing concept information is determined by the subclasses of $\mathfrak{A}$ and $\mathfrak{B}$. If subclasses have some distinguishing features, the chance level for classification of any sets S_1 and S_2 of subclasses will be above 50%. Therefore, the difference between the first ($m = 0$) and the last points ($m = k/2$) of the concept-response curve in Figure 1B indicates the contribution of concept-relevant information above the influence of the subclass-related nuisance effects. The point $m = k/2$ also represents the correct null hypothesis against which the effect of interest has to be compared. Because the effect of interest cancels out only for $m = k/2$, all the other points of the curve are partially biased by this effect. Therefore, it is strictly the point $m = k/2$ which should be used to test the null hypothesis that there is no relationship between classes in the data if one wants to avoid overly conservative statistical testing. The true null distribution for $m = k/2$ is produced by balanced permutation on blocks of trials belonging to different subclasses and contrasts to the typical trial-wise permutation test, where single trials are relabeled and different proportions of data from two classes can potentially exist in the randomized data sets. In other words, the proper exchangeability unit in the permutation scheme for data sets with subclasses is the subclass and not the overarching category/concept. Using the 95% confidence interval (CI) of the classification accuracy for $m = k/2$, which can be determined from the distribution of random permutations of subclass labels, we

can assess whether decoding of neural activity is specific to processing of the cognitive concepts of interest. It must be noted that if k (number of blocks in each category) is an odd number, m ranges from 0 to $\lfloor k/2 \rfloor$ (the largest integer no greater than $k/2$). For the case of $m = \lfloor k/2 \rfloor$, the primary effect due to the categories $\mathfrak{A}$ and $\mathfrak{B}$ is not completely balanced between sets S_1 and S_2, which results in a slightly more conservative test.

Experiment 1A: Decoding digits and letters from visually evoked potentials

In the first experiment, we aim to decode the semantic categories of 'digits' and 'letters' from event-related EEG-potentials (ERP) elicited by presentation of visual stimuli. 19 healthy subjects with no history of neurological or psychiatric disorders underwent EEG recording in two sessions while individual digits and letters were repeatedly presented to the subjects in randomly ordered sequences of 6 characters in the context of a Sternberg task, i.e. with the instruction to remember all elements of the sequence. Each stimulus appeared for 100 ms, to avoid eye movement during presentation, and was followed by a black screen for 1 s. The stimuli were the digits from 0 to 9 and 10 consonant letters, which were selected randomly but remained the same for all of the subjects (see Fig. 3A). Each stimulus was presented 18 times, resulting in a total of 180 trials per category. EEG was recorded during the whole task using an active 128-channel Ag/AgCl-electrode system (ActiCap, Brain products, Gilching, Germany) with 1 kHz sampling frequency and a high-pass filter of 0.1 Hz. Electrodes were placed according to the extended international 10-20 electrode system. Because the most relevant components of the visual ERP have a duration of 40 – 70 ms, which corresponds to maximum frequency of 25 Hz, we have applied a 40 Hz low-pass filter to reduce the number of features entered into the classification. Data was then divided into epochs of one second starting 50 ms before stimulus onset. Artefact rejection was done in a semiautomatic process using custom MATLAB scripts. Epochs containing artefacts were removed from the data set. Channels that contained too many epochs with

artefacts were removed and interpolated using routines provided by EEGLAB (Delorme and Makeig 2004).

In order to decode brain activity, we employed a linear SVM with 2-fold cross-validation to identify on a single trial level which of the two stimulus categories (digits or letters) was presented to the subject. 2-fold cross-validation was chosen because resulting classification accuracies have a lower variance than those obtained with a higher number of folds. It therefore has a higher sensitivity for the purpose of hypothesis testing (Jamalabadi *et al.* 2016). As input to the classifier, we used the 1-s ERP response in all 128 channels. The classifier was trained and tested within each subject. Performance was evaluated using the average percentage of the correctly classified trials in the test set (classification accuracy). No outliers have been removed from analysis, because classification accuracies can have a strongly asymmetric null distribution with a mean of 50% and a median above 50%. Removal of individual data points with low classification accuracies would lead to false positive results in this case (for details, see Jamalabadi *et al.* 2016).

As Figure 3B shows, single trial classification of digits and letters in individual subjects resulted in classification accuracies ranging from 47.0% to 60.2%, with a mean value of 54.2% across all subjects and sessions. Classification accuracy is positively correlated with the performance of the subjects in the Sternberg task (r_{38} = 0.372, p = 0.02), confirming the behavioral relevance of the classification results.

To determine if category-related information specific to processing of digits and letters is present on the group level, we varied the amount of concept-relevant information by changing the ratio of digit and letter stimuli in the classification sets S_1 and S_2 according to the method proposed above. Since there are 10 subclasses (digits, letters) per category ($k = 10$), we varied m (number of different letters in S_1) from 0 to 5, decreasing the primary effect of stimulus category gradually to zero. For each value of m, we repeated the random sampling of m letter and $(10 - m)$ digit stimuli for all possible permutations and

averaged classification accuracies over all repetitions in case the number of possible permutations was lower than 100. When more permutations were possible, we limited random sampling to 100 times, because the group null distribution which is needed for statistical inference on group level, converges already with 100 random permutations on the single subject level (Stelzer et al. 2013). This resulted in a concept-response curve for each subject and session. By averaging all curves, we obtained the group mean concept-response curve, which is shown in Figure 3C. The first point of the curve ($m = 0$), which corresponds to the classification of digits versus letters, shows an average classification accuracy of 54.2%. With increasing m, which is equivalent to decreasing the amount of conceptual information, the average classification accuracy monotonically decreases. For $m = 5$, although the primary effect is completely balanced between sets S_1 and S_2, the average classification performance is still 50.9%, and not 50.0% as might be expected. To obtain the confidence interval for $m = 5$, we generated the group null distribution by combining the subject-wise distributions of classification accuracies over 100 random combinations of 5 letters and 5 digits (Stelzer *et al.* 2013). This was done by randomly drawing (with replacement) from each subject one of the 100 classification accuracies. These subject-level accuracies were then averaged to obtain the group-level accuracy. This procedure was repeated 10^5 times, resulting in a distribution of 10^5 group-level accuracies. The resulting distribution shows that for two sets, each consisting of 5 random digits and 5 random letters, classification accuracy was still significantly above chance level (95% CI: [50.2%, 51.6%], $p < 0.018$). This means that besides the concept of digits and letters, the stimuli themselves (individual digits/letters) represent subclasses that also influence classification performance. It also signifies that the correct null distribution for the digit/letter concept classification cannot be derived from trial-wise permutation, which results in exactly 50% mean classification accuracy, but must be derived from subclass-wise permutation, which retains the bias produced by the similarity of subclass stimuli. Comparing the classification accuracy for $m = 0$ with the distribution for $m = 5$ shows that

the former is significantly above the latter ($p < 10^{-5}$). Thus, classification accuracy is significantly higher when stimuli are sorted according to the concept of digits/letters than when different subclasses of digits and letters are randomly combined. We therefore showed that the ERP contains information specific to processing the concepts of digits and letters.

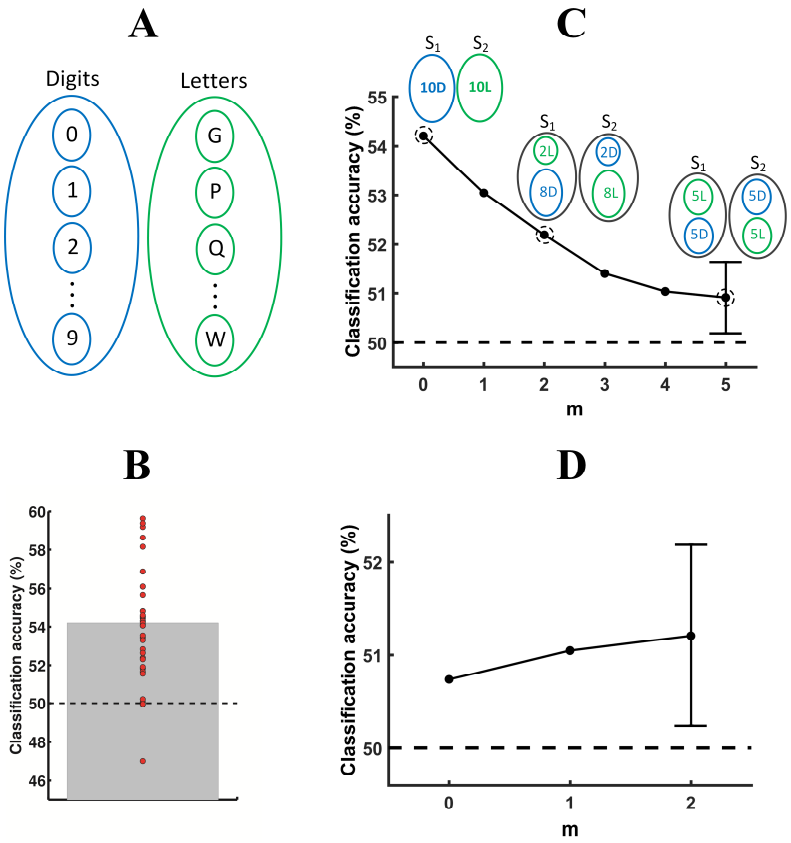


Figure 4: Decoding concepts of digits and letters (A-C) and even and odd digits (D). (A) Structure of the data for digit versus letter classification. Each category consists of 10 different stimuli. Stimuli were pictures of 10 digits (0-9) and 10 consonant letters (G, J, L, P, Q, R, S, W, X, Z) which were presented to the subjects. All trials that presented the same stimulus are considered as one block of data. (B) Digit/letter classification accuracy. The bar represents group average classification accuracy, each dot indicates results from one subject and session. The dashed line represents chance performance (50%). (C) Concept-response curve for different proportions of digits and letters per set. The procedure is shown for three points of the curve (m=0,2,5), representing combinations of 0/10, 2/8, and 5/5 digits/letters, respectively. 'D': digit, 'L': letter. The error bar shows the 95% CI of the stimulus-level classification on the group level. (D) Concept-response curve for decoding even and odd digits.

Experiment 1B: Decoding the concepts of 'even' and 'odd' digits

In continuation of the digit/letter analysis above, we used the same procedure as above to classify digits into 'even' and 'odd' numbers. Each category consists of 5 subclasses ({0, 2, 4, 6, 8} and {1, 3, 5, 7, 9}). We manipulated the amount of concept in sets S_1 and S_2 by varying m (number of even digit exemplars in S_1) from 0 to 2. Figure 3D illustrates the resulting concept-response curve. The average performance of odd/even classification ($m = 0$) is 50.7%. For $m = 2$, the average classification accuracy is 51.2% (95% CI: [50.2%, 52.2%]). This shows that in contrast to the concept of digits/letters, no information specific to whether a stimulus is odd or even can be detected in the ERP. The shape of the concept-response curve, which resembles Fig. 2B, indicates that the discrimination can be explained solely by the nuisance effects and no contribution of the primary effect to decoding performance can be concluded.

Experiment 1C: The spatiotemporal dynamics of conceptual and perceptual processing

Tracking the time course of brain activity to separate between different components of information processing is an interesting possibility which is put forward by time-resolved analysis of decoding accuracy (Bode and Haynes 2009; Simanova et al. 2010; Sudre et al. 2012). Here, we show that the method that we propose here can not only dissociate primary and nuisance effects, it can also characterize their spatiotemporal dynamics. Using the same digit/letter ERP data as above, we performed a time resolved decoding using classification accuracies from a sliding 70-ms window, which was gradually shifted in 5-ms steps over the whole 1-s duration of the ERP. For each point, we repeated the classification procedure for $m = 0$ (digit/letter concept present) and $m = 5$ (no concept present). For $m = 5$, we calculated the 95% CI from the group null distribution by combining the distribution of classification accuracies obtained from 100 randomly selected sets S_1 on the subject level with a bootstrapping procedure on the group level as above (Stelzer *et al.* 2013). The bootstrapping

process was repeated 10^5 times, resulting in a group null distribution with 10^5 group accuracies.

Figure 4A shows the time course of decoding accuracy averaged over nineteen subjects and two sessions. The blue line represents the time course of digit/letter decoding ($m = 0$), driven by both conceptual and perceptual differences between stimuli. The red line represents the time course of subclass-level decoding ($m = 5$), driven by perceptual, stimulus-related differences. This curve characterizes the portion of the ERP signal that is unrelated to the concept of digits vs. letters. Our data show that perceptual information can be reliably decoded between 150 and 350 ms after stimulus onset, when the lower bound of the 95% CI exceeds chance level (50%). Where digit/letter classification exceeds the upper bound of the 95% CI, i.e. from 90 to 635 ms after stimulus onset, concept-related information can be reliably decoded. Stimulus-level decoding (red line) shows peak performance around 220 ms after stimulus onset while digit/letter decoding (blue line) reaches its peak 35 ms later at 255 ms after stimulus onset. We can assume that this time lag occurs because lower-level, stimulus-specific information processing is faster and terminates earlier than higher-level concept processing. To further look into stimulus-level information processing, we repeated the same time-resolved analysis by one-versus-one classification of digits and letters, separately, and averaged over all 45 possible binary classifications of 10 stimuli in each category. Figures 4B-C show the time course of average classification accuracy for stimulus-level classification within each stimulus category. The results show that single digits and letters can be decoded reliably from 150 to 300 ms after stimulus onset, which overlaps substantially with the interval for successful subclass-level classification above. This indicates that it is stimulus-specific differences that make subclasses distinguishable. Moreover, one-versus-one stimulus classification peaks earlier than concept-level digit/letter classification, reflecting the slower nature of higher-level concept processing.

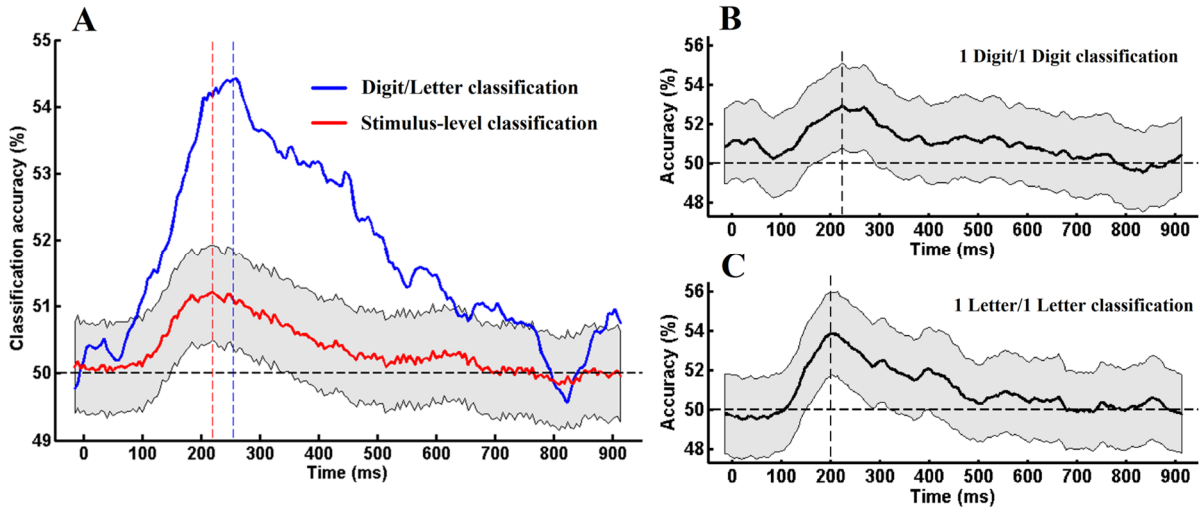


Figure 5: Time-resolved decoding of digits and letters (A). The blue line shows the time course of classification accuracy for digit/letter classification ($m = 0$). The red line represents the results for stimulus-level classification ($m = 5$). The shaded area around the red line indicates the 95% CI for stimulus-level classification on the group level. Where the lower margin of the CI exceeds 50%, significant stimulus-level information is present in the data. Where the blue line exceeds the upper margin of the CI significant concept-level classification is possible. The vertical dashed lines indicate the latencies at which the blue and red curves peak. (B, C) Time course of average classification accuracy for one-versus-one classification of digits and letters, respectively. The shaded area indicates the 95% CI on the group level.

Going beyond temporal localization, we can use a so-called searchlight approach to perform spatiotemporal localization (Kriegeskorte et al. 2006). We used a temporal window size of 70 ms with a 20 ms resolution and a spatial window size of 4 cm on-scalp radius around the 32 evenly spaced locations of the extended 10-20 system. For each spatiotemporal searchlight, we calculate a linear SVM with 2-fold cross-validation as proposed above, once for $m = 0$ and once for $m = 5$ (Fig. 5). To get the significance maps for digit/letter classification ($m = 0$), we compare the classification accuracy of each spatiotemporal searchlight with the group null distribution obtained by the permutations of $m = 5$. On the other hand, significances for stimulus-level classification are calculated based on the probability that the group distribution of permutations of $m = 5$ exceed 50%. Resulting topographies show areas of the cortex surface

that hold information relevant to the distinction between digits and letters and to the distinction of individual digits and letters, respectively. The results show that predictive sites for the digit/letter classification overlap with those sites responsible for the stimulus-level distinction, speaking for a contribution of these sites to both lower and higher-level processing. In particular, both include the occipital and temporal cortices. Concept-related topographies, however, are spread more widely and, at later time points, reach the frontal cortex, which is completely spared by stimulus-level processing. These results are in line with a previous study by Sudre et al. (2012) that used machine learning to track neural coding of perceptual and semantic features of concrete nouns in MEG data. In particular, they showed that perceptual features related to visual stimuli are decodable earlier in time than higher-level semantic features (e.g. animacy, manipulability and size), which were best decoded only after 250 ms post stimulus onset. Similarly, the lateral occipital cortex was shown to be preferentially related to encoding of perceptual features whereas activity in parietal and temporal regions were mainly associated with encoding semantic information.

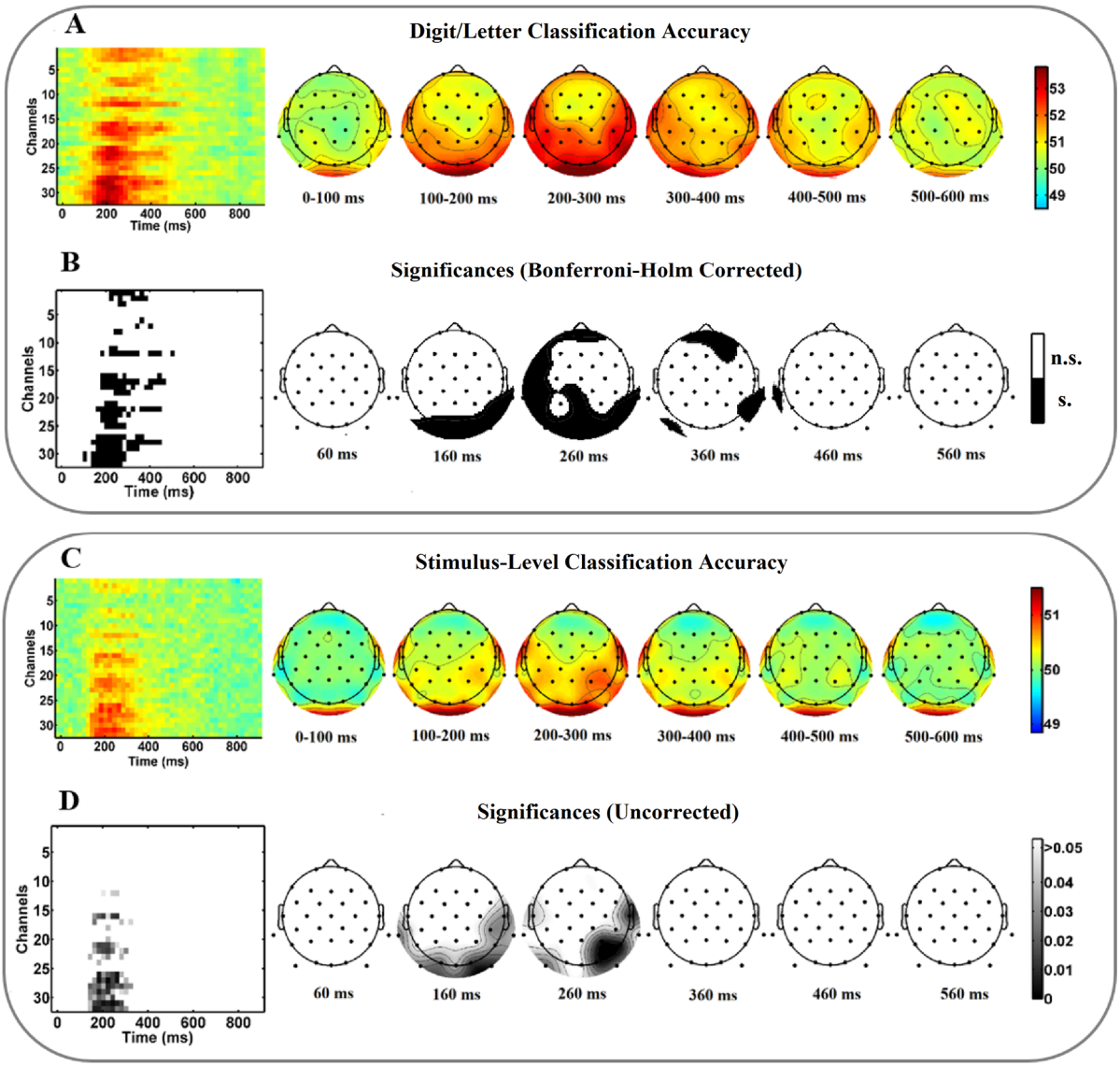


Figure 6: Searchlight classification of digits and letters. Performance of the digit/letter (A) and stimulus-level (C) classifier for all the spatiotemporal searchlights. (B) Significance maps for digit/letter classification after Bonferroni-Holm correction for 32×47 tests. Significant searchlights after correction are shown in black. Significances are calculated by comparing the classification accuracy of each spatiotemporal searchlight with the corresponding group null distribution obtained by permutations of $m = 5$ (stimulus-level classification). (D) Significance maps for stimulus-level classification. P-values are based on the distribution of permutations of $m = 5$ compared to chance performance of 50%. Significant searchlights did not survive the correction for multiple comparison.

Experiment 2: Decoding faces and houses from EEG with repeated stimulus presentation

We can use concept-response curves in any experiment where MVPA is used to decode category information from brain activity when different exemplars of each object category are presented multiple times. Here, we used EEG recordings from 19 healthy subjects in two sessions during presentation of visual stimuli belonging to the two categories of faces and houses. Similar to Experiment 1, subjects were presented with randomly ordered sequences of 8 pictures of either faces or houses in the context of a Stenberg task, i.e. they had to remember all pictures of the sequence and later report whether a target stimulus was present or not. We used totally 10 exemplars per category (10 different pictures of faces/houses). Each exemplar was presented 4 times throughout the experiment, resulting in a total of 40 trials per category. Each trial consisted of 100 ms visual presentation followed by a 1-s black screen. Recording, preprocessing and artefact rejection procedures were done as in Experiment 1.

We employed linear SVM with 2-fold cross-validation to identify on a single trial level whether an image of a face or house was presented to the subject. To analyze whether the concepts can be decoded from our data, we generated a concept-response curve as above. Since stimuli were selected from a set of 20 different pictures (2 categories with 10 exemplars each), each category consists of 10 blocks of 4 trials each ($k = 10$). We manipulated the amount of category-specific information by changing the ratio of face/house exemplars in sets S_1 and S_2. By changing m (number of face exemplars in S_1) from 0 to 5 we obtained the concept-response curve shown in Figure 6A. The curve shows a clear dependence of classification rate on the amount of concept present in the data. The average classification accuracy for $m = 0$ (maximum separation of concepts) over all subjects and sessions is 62.0%. For $m = 5$, the average classification accuracy due to category-irrelevant information is slightly but not significantly above chance (50.8%; 95% CI: [49.6%, 51.9%], p = 0.132). Based

on the shape of the concept-response curve we can therefore conclude that the primary effect of 'face' and 'house' category is present in the data and no significant nuisance effect due to the presentation of multiple category exemplars can be detected. However, the correct null-distribution to test for significance is still the one defined by the permutations of $m = 5$.

Experiment 3: Decoding the concepts of 'animal' and 'fruit'

In this last experiment, we aim to decode the two semantic categories 'animal' and 'fruit' from event-related EEG potentials. EEG was recorded with the same setup as in Experiment 1 above. 19 healthy subjects participated in two sessions during which visual stimuli belonging to the two categories were presented in a learning and recognition task. For the present analysis, ERP responses to 120 pictures (60 different pictures per category) were analyzed. Each picture was presented to the subjects once for 300 ms, followed by a black screen for 1.5 s. ERPs were calculated for epochs of 1 s starting at stimulus onset. Recording, preprocessing, and artefact rejection procedures were done as above. We used a linear SVM with 2-fold cross validation on the whole 128-channel ERP in order to decode for each trial whether a fruit or an animal had been presented. To investigate whether concept-irrelevant variance induced by different stimuli can affect classification, we generate a concept-response curve, which presents the relationship between the amount of concept in the data and classification accuracy. Notably, since there were no obvious subclasses in the data, the number of blocks in each category is equal to the number of trials ($k = 60$). We varied the amount of category-related information by changing the ratio of animal and fruit trials in the classification sets S_1 and S_2. We repeated the procedure for 6 points of the curve ($m = 0, 6, 12, 18, 24, 30$), equivalent to a 0%, 10%, 20%, 30%, 40% and 50% combination ratio, respectively. Figure 6B shows the resulting concept-response curve. The average classification accuracy over all subjects and sessions is 57.23% for $m = 0$. It decreases monotonically and converges to 50.0% (95% CI: [49.0%, 50.9%]) for $m = 30$. The shape of the

curve confirms that there is category-specific information related to the concepts of 'animal' and 'fruit' in our EEG data.

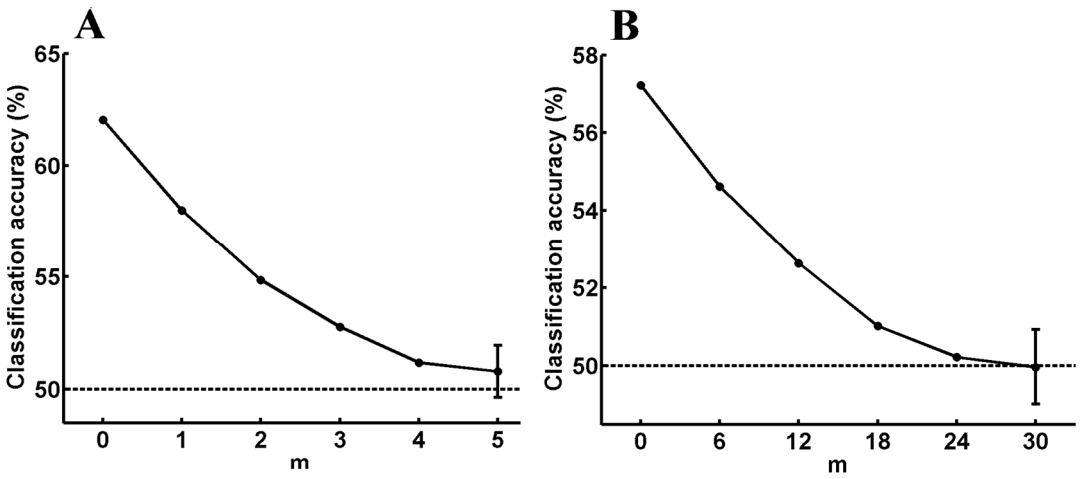


Figure 7: Concept-response curve for (A) decoding face and house stimuli and (B) decoding animal and fruit stimuli from event-related EEG potentials. The error bars show the 95% CI of the stimulus-level classification on the group level.

Discussion

The high sensitivity of MVPA for any kind of structure in a data set allows to detect subtle differences between conditions of interest, e.g. the distinct patterns of brain activity during processing of separate concepts (Haxby et al. 2001; Cox and Savoy 2003; Shinkareva et al. 2008; Simanova *et al.* 2010; Wang et al. 2013). In principle, although MVPA is a statistically powerful and robust method, its complexity can lead to important methodological and conceptual issues. Classification rates should not be tested for significance with classical parametric tests because their distribution can be strongly skewed for small effect sizes and it does not fulfill prerequisites for these tests (Noirhomme et al. 2014; Jamalabadi *et al.* 2016). Moreover, the sensitivity of MVPA makes it susceptible to effects of nuisance variables, which cannot be completely counterbalanced in some circumstances. This is usually the case if there are distinct subclasses in the data set. Subclasses can exist e.g. if several groups of

trials are combined into one class, if stimuli or types of stimuli are presented repeatedly, or if multiple subjects or experimental sessions are included in one analysis. In principle, these nuisance effects differ from systematic confounds (e.g. decoding black and white inanimate objects versus colorful animals) because they cannot be avoided experimentally. Therefore, confounds induced by subclasses are a general concern for MVPA, because they can lead to significant bias and higher than chance classification accuracy, even when the primary effect is nonexistent. To test against the correct null hypothesis, the influence of these nuisance effects has to be accounted for. Previous literature noted the challenges posed by nuisance variance and proposed to identify proper exchangeability blocks when constructing the null distribution (Nichols and Holmes 2002; Schreiber and Krekelberg 2013; Winkler et al. 2014). Here, we propose a method to present and test MVPA results which can quantify the contribution of nuisance variance by taking the data set structure into account. Concept-response curves enable us to show whether significant nuisance variables are present in the data and test whether the actual effect in question goes significantly beyond these effects. In the context of hypothesis testing, our method provides a permutation inference framework for the case when exchangeable units in the relabeling scheme are defined by the subclasses in the data. Importantly, our method is meant to be useful for cases when the confounds are not systematic and therefore cannot be avoided experimentally.

Similar to dose-response curves, concept-response curves also provide a convincing way to show that classification accuracy is increasing with the amount of conceptual information in a data set and increase confidence in the validity of a finding, especially if effects are small and classification rates are close to chance levels. Using concept-response curves provides an additional measure of validity because multiple classification steps are involved. An accurate decoding of concept-related information can only be confirmed if accuracy lies significantly above the rightmost point of the concept-response curve and if the concept-response curve shows a monotonic decay. If the curve

has an irregular structure, this indicates that classification accuracies are not stable enough and therefore cannot be attributed to the concept under study.

Using a grouping factor unrelated to the actual classification can not only be used to derive the correct null hypothesis when decoding cognitive concepts from brain activity, but is also helpful when separating the effects of different experimental factors. Because such factors (e.g. concept-related and perceptual influences as in Experiment 1C) often cannot be separated experimentally (Simanova *et al.* 2010; Murphy et al. 2011; Wurm et al. 2015), we believe that it is a worthwhile approach to manipulate the amount of concept-related information in the data during analysis and thus separate the actual concept from other (nuisance) factors. This method can also be used, e.g., to identify the temporal and spatial aspects of the signal related to each process, by determining where decoding accuracies related to the concept exceed those from concept-irrelevant classification, or to characterize the spatiotemporal dynamics of mental representations. Finally, by deliberately introducing other experimental factors as subclasses, it is possible to distinguish the independent contributions of several factors to classification.

It has been recently proposed by Höhne et al. (2016) that additional label information (i.e. subclass labels) should be incorporated into the classifier to improve the accuracy of pattern classification in neuroimaging studies. It is important to note that this is only true for designs with crossed factors, i.e., when every subclass coexists in both categories. While exploiting the information that is shared between crossed subclasses can improve classification performance (Hohne *et al.* 2016), the contribution of such information in nested designs, i.e. when each subclass pertains only to one of the categories (see Experiments 1 and 2), represents a confound and can lead to false positive results. Nested data are characterized by a hierarchal, multi-level structure (e.g. recordings using repeated stimuli, multiple sessions per subject, or multiple cells per animal). It has been reported that more than 50% of neuroscience papers included nested data, although this is largely ignored (Aarts et al. 2014). The nested structure

introduces dependency in the data that must be statistically accommodated. Although such considerations are not new for classical statistics (Galbraith et al. 2010; Lazic 2010; Aarts *et al.* 2014; Aarts et al. 2015; Moen et al. 2016), the implications for the use of MVPA must be further explored.

When planning to use MVPA for decoding cognitive concepts, and if confounding subclasses cannot be avoided, we recommend increasing the number of subcategories per condition. This makes subclass-specific information less prominent (see Experiment 2). Particularly, when more than five distinct subclasses are available, the correct null distribution and the corresponding 95% CI can easily be determined by random permutation. If only a smaller number of groupings is available, e.g. because the nuisance feature is dichotomic by nature, statistical inference on the group level must be applied to correct for the subclass-related bias instead. Together, we suggest that including concept-unrelated grouping factors into analyses, using blocked permutation tests, and gradually manipulating the proportion of concept-related information in MVPA to achieve concept-response curves is a viable, sensible and often necessary approach to data analysis when investigating brain responses to cognitive concept processing.

Acknowledgements

This research was supported by DFG grant (GA730/3-1) and BMBF Bernstein Center grant (01 GQ 1004A - B4).

References

Aarts E, Dolan CV, Verhage M, van der Sluis S. 2015. Multilevel analysis quantifies variation in the experimental effect while optimizing power and preventing false positives. BMC Neurosci 16:94.

Aarts E, Verhage M, Veenvliet JV, Dolan CV, van der Sluis S. 2014. A solution to dependency: using multilevel analysis to accommodate nested data. Nat Neurosci 17:491-496.

Bode S, Haynes JD. 2009. Decoding sequential stages of task preparation in the human brain. Neuroimage 45:606-613.

Cox DD, Savoy RL. 2003. Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. Neuroimage 19:261-270.

Delorme A, Makeig S. 2004. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J Neurosci Methods 134:9-21.

Fan JQ, Fan YY. 2008. High-Dimensional Classification Using Features Annealed Independence Rules. Annals of Statistics 36:2605-2637.

Galbraith S, Daniel JA, Vissel B. 2010. A Study of Clustered Data and Approaches to Its Analysis. J Neurosci 30:10601-10608.

Haxby JV, Connolly AC, Guntupalli JS. 2014. Decoding neural representational spaces using multivariate pattern analysis. Annu Rev Neurosci 37:435-456.

Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P. 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293:2425-2430.

Haynes JD. 2015. A Primer on Pattern-Based Approaches to fMRI: Principles, Pitfalls, and Perspectives. Neuron 87:257-270.

Haynes JD, Rees G. 2006. Decoding mental states from brain activity in humans. Nat Rev Neurosci 7:523-534.

Hohne J, Bartz D, Hebart MN, Muller KR, Blankertz B. 2016. Analyzing neuroimaging data with subclasses: A shrinkage approach. Neuroimage 124:740-751.

Jamalabadi H, Alizadeh S, Schonauer M, Leibold C, Gais S. 2016. Classification based hypothesis testing in neuroscience: Below-chance level classification rates and overlooked statistical properties of linear parametric classifiers. Hum Brain Mapp 37:1842-1855.

Kamitani Y, Tong F. 2005. Decoding the visual and subjective contents of the human brain. Nat Neurosci 8:679-685.

Kay KN, Naselaris T, Prenger RJ, Gallant JL. 2008. Identifying natural images from human brain activity. Nature 452:352-355.

Kriegeskorte N, Goebel R, Bandettini P. 2006. Information-based functional brain mapping. Proc Natl Acad Sci U S A 103:3863-3868.

Lazic SE. 2010. The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis? BMC Neurosci 11:5.

Mitchell TM, Shinkareva SV, Carlson A, Chang KM, Malave VL, Mason RA, Just MA. 2008. Predicting human brain activity associated with the meanings of nouns. Science 320:1191-1195.

Moen EL, Fricano-Kugler CJ, Luikart BW, O'Malley AJ. 2016. Analyzing Clustered Data: Why and How to Account for Multiple Observations Nested within a Study Participant? PLoS One 11:e0146721.

Murphy B, Poesio M, Bovolo F, Bruzzone L, Dalponte M, Lakany H. 2011. EEG decoding of semantic category reveals distributed representations for single concepts. Brain Lang 117:12-22.

Nichols TE, Holmes AP. 2002. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum Brain Mapp 15:1-25.

Noirhomme Q, Lesenfants D, Gomez F, Soddu A, Schrouff J, Garraux G, Luxen A, Phillips C, Laureys S. 2014. Biased binomial assessment of cross-validated estimation of classification accuracies illustrated in diagnosis predictions. Neuroimage-Clinical 4:687-694.

Norman KA, Polyn SM, Detre GJ, Haxby JV. 2006. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn Sci 10:424-430.

O'Toole AJ, Jiang F, Abdi H, Penard N, Dunlop JP, Parent MA. 2007. Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data. J Cogn Neurosci 19:1735-1752.

Rissman J, Greely HT, Wagner AD. 2010. Detecting individual memories through the neural decoding of memory states and past experience. Proc Natl Acad Sci U S A 107:9849-9854.

Schreiber K, Krekelberg B. 2013. The statistical analysis of multi-voxel patterns in functional imaging. PLoS One 8:e69328.

Schwarzlose RF, Swisher JD, Dang S, Kanwisher N. 2008. The distribution of category and location information across object-selective regions in human visual cortex. Proc Natl Acad Sci U S A 105:4447-4452.

Shinkareva SV, Mason RA, Malave VL, Wang W, Mitchell TM, Just MA. 2008. Using FMRI brain activation to identify cognitive states associated with perception of tools and dwellings. PLoS One 3:e1394.

Simanova I, Hagoort P, Oostenveld R, van Gerven MA. 2014. Modality-independent decoding of semantic information from the human brain. Cereb Cortex 24:426-434.

Simanova I, van Gerven M, Oostenveld R, Hagoort P. 2010. Identifying object categories from event-related EEG: toward decoding of conceptual representations. PLoS One 5:e14465.

Stelzer J, Chen Y, Turner R. 2013. Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): random permutations and cluster size control. Neuroimage 65:69-82.

Sudre G, Pomerleau D, Palatucci M, Wehbe L, Fyshe A, Salmelin R, Mitchell T. 2012. Tracking neural coding of perceptual and semantic features of concrete nouns. Neuroimage 62:451-463.

Todd MT, Nystrom LE, Cohen JD. 2013. Confounds in multivariate pattern analysis: Theory and rule representation case study. Neuroimage 77:157-165.

Tong F, Pratte MS. 2012. Decoding Patterns of Human Brain Activity. Annu Rev Psychol 63:483-509.

Wang J, Baucom LB, Shinkareva SV. 2013. Decoding abstract and concrete concept representations based on single-trial fMRI data. Hum Brain Mapp 34:1133-1147.

Winkler AM, Ridgway GR, Webster MA, Smith SM, Nichols TE. 2014. Permutation inference for the general linear model. Neuroimage 92:381-397.

Woolgar A, Golland P, Bode S. 2014. Coping with confounds in multivoxel pattern analysis: what should we do about reaction time differences? A comment on Todd, Nystrom & Cohen 2013. Neuroimage 98:506-512.

Wurm MF, Ariani G, Greenlee MW, Lingnau A. 2015. Decoding Concrete and Abstract Action Representations During Explicit and Implicit Conceptual Processing. Cereb Cortex 26:3390-3401.

Acknowledgements

First and foremost, I would like to express my special thanks to my supervisor, Steffen Gais, for all his time, expertise and continuous support in guiding and mentoring me step by step through the PhD process. I am truly grateful for his patience, encouragement and the many things that I learned from him. His advices on both research as well as my career have been invaluable to me.

Special thanks also go to my colleague and co-author, Monika Schönauer who had a great contribution in the completion of this research work. Her support, enthusiasm and friendship was always heartwarming and her scientific contributions significantly improved this work.

I would like to thank Christian Leibold for stimulating and fruitful collaborations on topics related and beyond this thesis, which I enjoyed a lot.

I am truly grateful to my friends, fellow colleagues and lab mates Farid Shiman, Andreas Ray, Ander Ramos, Thiago Figueiredo, Svenja Brodt, Lea Himmer, Monika Schönauer, Jingyi Wang, Frederik Weber and Paulo Rogerio with whom I shared so many memorable experiences and had a lot of fun. Thanks to them for being always around when I needed someone to talk to.

I am thankful to BCCN and LMU Munich, UKT and GTC Tübingen for the funding and supports. Furthermore, I would like also to thank Jan Born and Moritz Grosse-Wentrup for their helpful feedbacks and comments as members of my advisory board committee.

My sincere thanks go to my parents, for their moral and emotional support which encouraged me to efficiently overcome the difficulties along the way.

Finally, and most importantly, I would like to thank a very special person, my husband and colleague, Hamidreza Jamalabadi for his unconditional love, understanding and constant support during my Ph.D. that made the completion of this thesis possible. I would have not been able to accomplish this feat without having him by my side.