BAYESIAN MECHANISMS IN SPATIAL COGNITION: TOWARDS REAL-WORLD CAPABLE COMPUTATIONAL COGNITIVE MODELS OF SPATIAL MEMORY

A THESIS SUBMITTED TO THE UNIVERSITY OF MANCHESTER
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
IN THE FACULTY OF ENGINEERING AND PHYSICAL SCIENCES

2015

By
Tamas Madl
School of Computer Science

Contents

Word Count: 79,848

List of Tables

Page numbers on the far right refer to the numbering used in the thesis. Page numbers in parentheses refer to the numbering used within the respective publication.

List of Figures

Page numbers on the far right refer to the numbering used in the thesis. Page numbers in parentheses refer to the numbering used within the respective publication.

Notation

$\boldsymbol{x}$	Location in 2D space
X	Path containing multiple locations
$\boldsymbol{l}$	Landmark location in 2D space
L	Set of all local (currently observable) landmark locations
$\boldsymbol{o}_i$	Observation (distance measurement in 2D space)
O_j	Set of all observations at time step j
d	Scalar (one-dimensional) distance measurement
S, Σ	Covariance matrix of a normal distribution
$\boldsymbol{\mu}$	Mean of a normal distribution
$\mathcal{N}(\boldsymbol{\mu}, \Sigma)$	Normal distribution with mean $\boldsymbol{\mu}$ and covariance Σ
$\boldsymbol{m}_t$	Motion vector in 2D space at time step t, based on motor command
$\boldsymbol{v}_t$	Speed vector in 2D space at time step t
$\boldsymbol{c}_i$	Constraint (measured displacement) between two locations, based e.g. on recognizing a previously visited place
A	Ratio of covariance matrices (ratio of the uncertainty associated with recognizing a previously visited place and the uncertainty of path integration)
C	All available constraints
$\boldsymbol{d}_i$	Discrepancy (difference between constraint $\boldsymbol{c}_i$ and the displacement between the two location representations)
$d_m(\boldsymbol{x}, \boldsymbol{y})$	Metric (distance function) specifying the distance between $\boldsymbol{x}$ and $\boldsymbol{y}$
D	Distance matrix
$[\Delta \boldsymbol{x}_{i,j}]_+$	Absolute pairwise difference vector between vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$
$p(c = 1 \mid \Delta \boldsymbol{x}_{i_j})$	Probability that $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ belong to the same representation (or cluster), given their absolute pairwise difference
$p(c = 0 \mid \Delta \boldsymbol{x}_{i_j})$	Probability that $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ do not belong to the same representation (or cluster), given their absolute pairwise difference
$\boldsymbol{\theta}$	Model parameters (e.g. means and variances of Gaussians)
α	Learning rate of gradient descent
γ	Normalization constant

Abbreviations

APD	Absolute Pairwise Difference
BICA	Biologically Inspired Cognitive Architecture
BVC	Boundary Vector Cell
CA1, CA3	Cornu Ammonis area 1 and 3 in the hippocampus
CD	Coincidence Detection in neurons
CNN	Convolutional Neural Network
DP-GMM	Dirichlet Process Gaussian Mixture Model
DIRECT	DIviding RECTangles algorithm
EC	Entorhinal Cortex
FLOPS	Floating-Point Operations Per Second
fMRI	functional Magnetic Resonance Imaging
GDA	Gaussian Discriminant Analsis
GP	Gaussian Process
GWT	Global Workspace Theory
LIDA	Learning Intelligent Distribution Agent
LIDAR	LIght raDAR
LTM	Long-Term Memory
MCMC	Markov Chain Monte Carlo
MDS	Multi-Dimensional Scaling
MTurk	Amazon Mechanical Turk
PAM	Perceptual Associative Memory
PHC	Parahippocampal cortex
PPC	Probabilistic Population Code
PRC	Perirhinal cortex
RBF	Radial Basis Function
RF	Random Forest
RI	Rand Index
ROS	Robot Operating System
SBC	Structure-Building Codelet
SLAM	Simultaneous Localization and Mapping
SSE	Sum of squared errors
STM	Short-Term Memory
SVM	Support Vector Machine
t-SNE	t-Distributed Stochastic Neighbor Embedding

Abstract

BAYESIAN MECHANISMS IN SPATIAL COGNITION:
TOWARDS REAL-WORLD CAPABLE COMPUTATIONAL COGNITIVE MODELS OF SPATIAL MEMORY

Tamas Madl

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy, 2015

Computational cognitive models of spatial memory often neglect difficulties posed by the real world, such as sensory noise, uncertainty, and high spatial complexity. However, since cognition and its neural bases have been shaped by the structure and challenges of the physical world, cognitive models should take them into account.

This work takes an interdisciplinary approach towards developing a cognitively plausible spatial memory model able to function in realistic environments, despite sensory noise and spatial complexity. We investigated how spatially relevant brain areas might maintain accurate location estimates, despite accumulating sensory noise, hypothesizing that hippocampal place cells might perform Bayesian localization, and that hippocampal reverse replay might play a role in correcting learned maps after revisiting known places. We developed computational models implementing these probabilistic mechanisms, which we argued to be psychologically plausible (producing human-like behaviour) as well as neurally plausible (implementable in brains). In support of the hippocampal Bayesian localization hypothesis, we reported modelling results of single-neuron recordings from rats (acquired outside this PhD), constituting the first evidence for Bayesian inference in place cells, as well as modelling behaviour data from humans. We also collected and modelled sketch map accuracy data in experiments performed online, substantiating the suggested map correction mechanism.

In addition to dealing with noise and uncertainty, in realistic environments, large-scale representations also have to be stored and used efficiently. Hierarchical representations help dealing with large amounts of spatial information by facilitating rapid and efficient retrieval search and route planning. It has been suggested that cognitive

maps in humans are hierarchical, but the computational principles underlying these hierarchies have remained unknown. We investigated features influencing cognitive map structure using spatial memory data concerning real-world and virtual reality environments collected in online experiments, and proposed a computational mechanism (clustering in psychological space) which might give rise to sub-map structures, showing that it can predict these structures in participants' spatial memory in advance.

We have extended a general cognitive architecture (the LIDA model of cognition) by these Bayesian mechanisms for localization and map learning, correction, and structuring; integrating them with the other cognitive phenomena accounted for by LIDA. We demonstrated the ability of the resulting model to deal with the challenges of realistic environments by running it in high-fidelity robotic simulations, modelled after participants' actual cities, showing that it can deal with noise, uncertainty and complexity, and that it can reproduce the spatial accuracies of human participants.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright

Acknowledgements

I would like to warmly thank Emer.Prof. Dr. Stan Franklin, for all the support, advice, inspiring discussions, and friendly encouragement over the last years; and for very productive and enjoyable periods of external research at the University of Memphis. I am grateful and indebted to my supervisor Dr. Ke Chen and co-supervisor Prof. Dr. Daniela Montaldi, who have supported and advised me in computational/mathematical and psychological/neural matters, respectively. Many thanks also to Emer.o.Univ.-Prof.Dr. Robert Trappl, for the opportunities of external research at the Austrian Institute for Artificial Intelligence. I am also indebted to Prof. Dr. Carol A. Barnes, and Dr. Sara N. Burke, for being kind enough to share their place cell recordings with me, which have both inspired and substantiated the core hypotheses of my work.

I am also eternally grateful to my family and loved ones for supporting me and being patient with me even if I was unreachable for days or weeks at a time, buried in work. Thank you to my dad, whose natural curiosity, rational world-view, and well-equipped workshop have inspired a scientific mindset in me even as a child; to my mum, whose creativity and perseverance have been as inspiring as her support has been comforting; to my sister, for showing me what it means to 'explore, dream, and discover'; and to my brother, for some of the best discussions I have had, and for his suggestions for improving this thesis. Thank you also to Maria for always being there for me. I also want to thank my friends, for the good times and for reminding me of what it means to be human; and for my colleagues, both at OFAI in Vienna, at the University of Manchester, and at the University of Memphis, for interesting discussions.

I take this opportunity to acknowledge my sponsors: EPSRC grant EP/I028099/1, FWF grant P25380-N23, OIST Computational Neuroscience Course 2013 accommodation and travel grant, TIMELY COST action TD0904 travel grant, and the accommodation grant for The First Örebro Winter School on Artificial Intelligence and Robotics.

Chapter 1

Introduction

Brains have evolved to move bodies through space in order to increase the chances of survival and reproduction, through numerous complex behaviours such as fleeing from threats or searching for nutrients or potential mates. The ability to remember spatial information, e.g. previously encountered food sources or shelters, has provided sufficient evolutionary advantage that all known organisms with brains (and even some without, such as the slime mold[1] - Reid et al. (2012)) have at least a rudimentary ability to utilize representations of space for more efficient navigation. Higher mammals have evolved a network of brain areas implementing spatial memory, a system for storing and recalling spatial information about the environment and about their location in it.

Representing spatial information accurately in the real world is hard, for several reasons. Sensors and actuators are limited, erroneous and noisy (in the sense of noise interfering with the signal). There are additional sources of uncertainty or unknown information, such as external events, actions of other organisms, unperceived or currently unperceivable objects or events. Furthermore, physical environments can be highly complex, and yet cognitive resources (amount of memory, processing power, time and energy available) are necessarily limited by biological and physical constraints.

In artificial intelligence (AI) and robotics research, probabilistic models have provided key tools for dealing with such challenges, facilitating the quantitative characterization of beliefs and uncertainty in the form of probability distributions, and the machinery of Bayesian inference for updating them with new data. They have also inspired the 'Bayesian brain' (Knill & Pouget, 2004) and 'Bayesian cognition' (Chater

[1] Slime molds are able to avoid previously explored areas using externalized spatial memories, and to solve mazes using nutrient gradients

et al., 2010) paradigms in the cognitive sciences. These paradigms have been successful in explaining human behaviour in tasks as diverse as the integration of sensory cues (Ernst, 2006) including spatial information (Cheng et al., 2007; Nardini et al., 2008), sensorimotor learning (Körding & Wolpert, 2004), visual perception (Yuille & Kersten, 2006) or reasoning (Oaksford & Chater, 2007). Their success suggests an answer to what biological cognition might be doing to cope with the above-mentioned challenges: approximate Bayesian inference.

1.1 Motivation

Despite of this success and of the suitability of probabilistic models to deal with uncertain and noisy spatial information, there have been few attempts to use them for modelling spatial memory within cognitive modelling, the branch of cognitive science concerned with computationally simulating mental processes. There is a gap in the literature between probabilistic spatial models in robotics and computational cognitive models of spatial memory. In robotics, Simultaneous Localization and Mapping (SLAM) models (Thrun & Leonard, 2008) are capable of dealing with real-world noise, uncertainty, and complexity to some extent, but are cognitively implausible[2]. On the other hand, current computational cognitive models of spatial memory, which are designed to model biological spatial cognition, cannot deal with all of these challenges, and are thus confined to simplistic simulations (see Chapter 3 for a review, and Figre 1.1 for an overview of the importance of spatial memory and the differences between information processing in robots and brains).

In addition, although spatial representations in humans have been argued early to be hierarchical (Hirtle & Jonides, 1985; McNamara et al., 1989; Greenauer & Waller, 2010), similarly to some robotic implementations having to deal with large, complex environments (Kuipers, 2000; Wurm et al., 2010), it is not known how (by which process) these hierarchical spatial maps might be structured. Although many computational models of spatial memory running in simplified environments exist, there is a lack of biologically and psychologically plausible 'algorithms' serving as models of human cognitive computations related to spatial information processing which can

[2]In our usage of the terms, a computational model is 'psychologically plausible' (or 'cognitively plausible') to the extent that it is consistent with psychological findings and can accurately reproduce psychology data, i.e. behaviours. Analogously, it is 'biologically plausible' (or 'neurally plausible') to the extent that it is consistent with neuroscience and can reproduce neural data, e.g. single-cell recordings or brain imaging results.

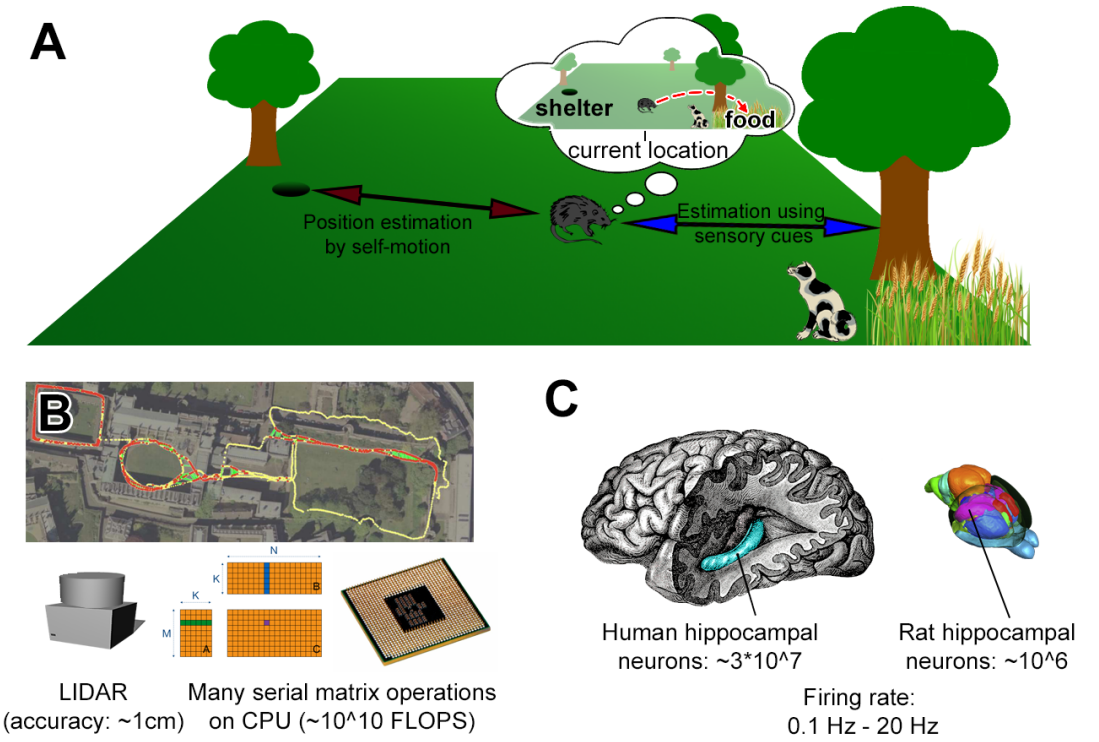


Figure 1.1: Motivation for proposing new computational cognitive models of spatial memory. A: Learning representations of the space around animals confers significant advantages, such as the ability to plan a detour out of sight (dashed red arrow) to reach a food source while avoiding danger in this example. In real environments, this task is made more difficult by the unreliability, errors and noise inherent in both the estimation of position by integrating self-motion and in estimated object distances (e.g. based on vision). Most existing cognitive models of spatial memory neglect these challenges. B: State of the art SLAM models in robotics are able to estimate locations and learn maps accurately, but rely on sensors and computations which are very different from biology - e.g. higher measurement accuracy using laser-based distance sensors (LIDAR), centralized control and coordination, and high number of serial operations per second - up to 10^{10} floating-point operations per second (FLOPS) needed for state of the art SLAM systems (Santos et al., 2013). C: In contrast, the hippocampus - the major brain area involved in world-centered spatial representations - contains only a few million neurons, of which only a subset is active at a time, each firing only a few times per second (Rapp & Gallagher, 1996; Šimić & Bogdanović, 1997); and relies on noisy, inaccurate sensory measurements. Although many models of spatial memory in brains exist, there is a lack of computational mechanisms which are both neurally and psychologically plausible, and can work in realistic environments and with noisy sensors. (Example SLAM data in Panel B from (Newman et al., 2011), and 3D rat brain in Panel C from (Calabrese et al., 2013), with permission.)

function in realistic, uncertain, complex environments.

The deprioritization of the problems of uncertainty and noise in favour of tractably modelling other human cognitive mechanisms is also pronounced in cognitive architectures, which try to account for a large number of mental processes in a unified, comprehensive, systems-level model (as opposed to computational cognitive models, which usually focus on a single phenomenon). In their overview of the field, Langley et al. (2009) argue that *" we should attempt to unify many findings into a single theoretical framework, then proceed to test and refine that theory"*, supporting the arguments of Newell (1973) that *"you can't play 20 questions with nature and win"*, highlighting the importance of systems-level research in the cognitive sciences. Although a few such cognitive architectures do model spatial mechanisms in navigation space (Harrison et al., 2003; Schultheis & Barkowsky, 2011; Sun & Zhang, 2004), they all run in simple, noise-free environments. According to a comparative table of cognitive architectures (Samsonovich, 2011) available in updated form online[3], there is currently no cognitive architecture implementing both Bayesian update and an empirically validated, psychologically plausible 'cognitive map' at the same time.

The present work was motivated by these gaps in the literature, and aims to take computational cognitive models of navigation-scale[4] spatial memory one step closer to modelling behaviour in realistic environments, such as high-fidelity robotic simulations or physical environments. It aims to do so by means of proposing probabilistic mechanisms of spatial cognition which are implementable in brains and can reproduce behaviour data, and by computationally implementing these mechanisms, in the form of cognitive models and within an existing cognitive architecture. Situated within the computational sub-fields of cognitive science (cognitive modelling and cognitive architectures), the goal of this work is to contribute to the understanding of information processing in human cognition. As such, although it is computational in nature, the extent of its success is determined by its ability to predict and explain the kinds of behaviour data it is intended to model, as well as its consistency with established findings in psychology and neuroscience. It is not aiming to maximize the accuracy of learned spatial representations, unlike robotics. Neither does it aim for neurobiological fidelity at the cellular level or below. Although building on neuroscientific evidence,

[3] http://bicasociety.org/cogarch/architectures.htm

[4] Human cognition needs to keep track of the space of navigation as well as the spaces immediately around the body (e.g. reachable objects) and of the body (e.g. body-part configurations). Although uncertainty and noise play are important in the latter two spaces as well, we will confine ourselves to navigation-scale spatial mechanisms in this work.

our concern is modelling spatial information processing on Marr's algorithmic level of analysis (Marr & Poggio, 1976; Poggio & Marr, 1977), as opposed to e.g. biological neural networks - see Table 1.1 -, with Chapter 4 being the single exception.

↓ Level of analysis	Description	In this work
1. Computational	What problem(s) does the system solve, and why?	Localization, Map error correction, Map structuring
2. Algorithmic/ Representational	How might it solve them? (Using what representations and processes?)	Cognitive models of spatial memory
3. Implementation	How is it implemented physically?	Place, grid, head-direction, border cells, ... (Hartley et al., 2014)

Table 1.1: Investigating spatial mechanisms on Marr's (1976) levels of analysis. The present work is mostly concerned with the second level.

We have investigated the plausibility of Bayesian spatial cue integration both on Marr's algorithmic (Chapter 6) and implementation level (Chapter 4), in order to maintain the desirable criteria of both psychological and neural plausibility for our other models. The possible neural implementation of this mechanism has been unknown, with current mechanistic models of Bayesian inference in brains making assumptions not fully consistent with the anatomy or activity of the hippocampal complex (the major brain areas representing world-centered spatial information) - see next Section. This doubt of biological implementability has motivated our investigation of single-cell electrophysiological data (acquired outside this PhD) to provide the first evidence for Bayesian updating in the hippocampus on a neuronal level, and our proposal of a plausible mechanism for implementing it. This evidence, presented in Chapter 4, affords a degree of biological plausibility to the models utilizing Bayesian mechanisms in the rest of our work (which is concerned with processes on the algorithmic/representational level).

1.2 Probabilistic models of space in brains and minds

Although the focus of most of this work is on the computational modelling of behaviour data, we would like the employed mechanisms to be plausibly implementable in the parts of the brain they functionally correspond to. Apart from the lack of

neuronal-level evidence that the hippocampal complex may perform Bayesian inference or even represent uncertainty, the possibility of the implementation of such a mechanism given the anatomical and electrophysiological constraints of this network of brain cells is also unclear.

Below, we briefly review probabilistic neural spatial models which have been proposed in the literature (Chapter 3 provides a more general review of computational cognitive models of spatial memory). We start with normative models of dealing with spatial uncertainty, which derive optimal solutions to the problem a system might be solving (Marr's computational level). We then continue describing mechanistic (implementation level) models which might facilitate these, and their consistency with what is known about the hippocampal complex. More extensive reviews of Bayesian models in brains can be found in (Pouget et al., 2013; Vilares & Kording, 2011). There is currently little experimental support for any of the proposed neural uncertainty representations.

Models of probabilistic estimation of spatial information have been pioneered by (Bousquet et al., 1997), who suggested to use a Kalman filter to model localization in the hippocampus. A Kalman filter is a dynamic Bayesian inference algorithm for estimating the values of unknown, not directly observable variables (such as location) from noisy observations, yielding statistically optimal estimates if the noise is normally distributed (Kalman, 1960). MacNeilage et al. (2008) also put forth arguments for dynamic Bayesian inference as a model of spatial orientation. They mention both Kalman filters and particle filtering (a related Bayesian filtering algorithm using samples instead of parameters to represent probability distributions), but leave the question of their neural implementation open. Particle filter-based models of localization on the algorithmic level have been suggested by (Fox & Prescott, 2010; Cheung et al., 2012). Osborn (2010) went beyond localization, suggesting a Kalman filtering approach to also account for localizing objects in the environment. Recently, Penny et al. (2013) argued that if one presupposes the existence of 'observation' and 'dynamic' models[5], required by Kalman filters, one might as well extend the inference to also use them for model selection ('which environment am I in?'), motor planning ('how do I get to place X?'), and to construct sensory imagery ('what does place X look like?') in addition to localization. They have combined these functions in a single probabilistic model, and argued that it is consistent with findings of pattern replay in the brain. An even

[5]Observation models and dynamic models are mathematical functions mapping from true states to observed states, and from pre-motion to post-motion states, respectively.

more general probabilistic formulation based on dynamic Bayesian inference is the Free-Energy Principle (Friston et al., 2006), which aspires to provide a unified theory of brain function, and has been argued to be consistent with aspects of hippocampal processing (Friston et al., 2011).

Despite their considerable theoretical elegance, the above-mentioned models do not provide a final and complete answer to the motivating question of this thesis (Section 1.1), which can be summarized as: 'how does biological cognition learn representations of navigation space from noisy sensors in an uncertain world?', for two reasons. First, none of them try to reproduce or show quantitative consistence with either behavioural or neural data concerning spatial cognition (although qualitative consistence with anatomical and neural findings is pointed out by the authors). Although these models provide explanations, their predictions regarding spatial processing have not been quantitatively evaluated.

Second, in addition to the lack of quantitative validation, their neural implementation is not known, and far from straightforward. For example, implementing the kinds of large matrix inversions and multiplications required by Kalman filters (Kalman, 1960) is easy on a computer, with centrally coordinated, serial, 'fast' computations, but difficult with the kind of distributed, parallel, 'slow' (on the level of single neurons, which only spike up to a few dozen times per second) computation performed by the brain. In the domain of world-centered, navigation-scale spatial mechanisms, any suggested neural implementation has to conform with not only the limitations imposed by biological neural networks, but also with the specific connectivity and activity observed in the hippocampal complex, in order to be considered biologically plausible.

In addition to such normative models, a number of mechanistic (implementation-level) models of how uncertainty and inference could be implemented in brains have also been proposed. They can be roughly grouped into three categories - see (Pouget et al., 2013; Vilares & Kording, 2011) for reviews. We briefly summarize these groups below, together with their consistency with what is known about the hippocampus.

- Probabilistic population codes (PPC) (Ma et al., 2006) encode probability distributions in the logarithmic domain by means of a set of coefficients of corresponding exponential basis functions, each coefficient encoded by the activity (spike count) of a neuron. They assume neural variability is independent and Poisson-distributed. However, hippocampal neurons exhibit more variability than a Poisson process (Fenton & Muller, 1998; Barbieri et al., 2001). Also,

if Bayesian inference were implemented in the hippocampus via a PPC, the encoded probability distributions would strongly depend on the firing rate of hippocampal neurons: increased firing rates should mean decreased levels of uncertainty. But empirically, this is not the case - for example, firing rates increase with movement speed (Maurer et al., 2005), which would mean the lowest uncertainties when running fastest (however, faster movements are harder to control and should thus lead to higher uncertainty).

- Instead of an encoding in the logarithmic domain, codes in which firing rates are proportional to probabilities have also been proposed, e.g. by Koechlin et al. (1999); Barber et al. (2003). The problem with their implementation in hippocampal neurons is that the firing rates of these neurons are also influenced by factors unrelated to probability, such as where the animal is headed (Ferbinteanu & Shapiro, 2003) or trial dependent features (Allen et al., 2012), and can change substantially if either the shape or colour of an environment is altered (Leutgeb et al., 2005). These influences would strongly interfere with the outcome of the Bayesian inference, if it were implemented in a code that directly utilizes firing rates.

- Sampling-based codes represent probability distributions with a set of samples drawn from them (Fiser et al., 2010). They are asymptotically correct with infinitely many samples, and approximations otherwise. Apart from being able to represent complex, multi-modal distributions, not having to rely on any fixed-form parametrization such as Gaussians, this also allows reducing their accuracy and computational demands by restricting the number of samples used. This property has been used e.g. by (Shi et al., 2010) to explain the deviations from the statistical optimum in an exemplar model of a reproduction task. It is difficult to make a general statement as to the implementability of this class of models in the hippocampal complex, as there is a wide variety of suggested concrete neural implementations in non-spatial domains (Sanborn (2015) provides a review), and some applied to navigation space, e.g. (Fox & Prescott, 2010; Cheung et al., 2012). None of them have been quantitatively validated by neural (electrophysiological) measurements, although most of them are supported by behavioural observations.

How the brain might encode and utilize uncertainty is still an open question (Pouget et al., 2013), but based on the observations regarding the hippocampus outlined above,

we argue that a sampling-based code is most suitable in this brain area; in terms of violating as few empirical observations as possible. We will provide electrophysiological evidence of Bayesian inference from single neurons, and propose a possible sampling-based mechanism, in Chapter 4 (and in more detail in Appendix A).

1.3 Hypotheses

To achieve goals in a spatially extended, realistic environment, at a minimum, an agent (e.g. a biological agent such as an animal, or an artificial agent such as a robot) must be able to 1) move, and keep track of its movements, 2) sense, and interpret its sensations, 3) represent spatial locations in its environment, e.g. of itself and its goal, 4) update these representations when changes occur in the environment, and 5) utilize these representations to achieve its goals (e.g. navigate to its goal location, avoiding dangers). Extensive work on all levels of analysis has been carried out for 1)-3), with the most recent Nobel prize in physiology or medicine awarded on the topic of 3) to John O'Keefe, May-Britt Moser and Edvard I Moser for the discovery of *'cells that constitute a positioning system in the brain'* (Burgess, 2014). Specifically, it was awarded for the discovery of 'place cells' in the hippocampus (which show increased firing in a specific area in the environment, called its 'place field'), and of 'grid cells' which show a regular, grid-like firing pattern (see Chapter 3 below).

We have argued above that despite of the variety of existing models regarding 4)+5), new computational models are needed to move towards biological and psychological plausibility as well as real-world capability at the same time (since biological cognition has been shaped by the constraints and challenges of the real world, these should not be neglected in models of cognition). In particular, in accordance with the 'Bayesian brain' (Knill & Pouget, 2004) and 'Bayesian cognition' (Chater et al., 2010) paradigms, we have suggested approximate Bayesian inference to be a well-suited mechanism for tackling these challenges. Models on Marr's algorithmic (and implementation) level which utilize such a mechanism require a number of underlying assumptions, some of which can be stated and evaluated as hypotheses.

We summarize major hypotheses in one place in Table 1.2 below, and expand on them in the respective results chapters below. The first two concern the representation and manipulation of uncertainty in the hippocampus (required for maintaining approximately accurate location estimates despite noisy sensors and accumulating errors). Hypothesis 3 is needed since unless all remembered landmark locations are

corrected at every moment (which would likely be intractable), a discrepancy between remembered and actual locations might arise when revisiting a location encountered previously (when traversing a 'loop' in the environment). This discrepancy necessitates a backward correction of multiple recent self and landmark locations to maintain consistent representations. The last two are needed to formulate a computational mechanism of spatial representation structure. Structured, hierarchical representations provide clear computational advantages, such as increased speed and efficiency of retrieval search, and economical storage. However, although strong neural (Derdikman & Moser, 2010) and behavioural (Hirtle & Jonides, 1985; McNamara et al., 1989; Greenauer & Waller, 2010) evidence exists for such structure, underlying computational principles have remained largely unknown.

Hypothesis	**Prediction**	**Empirical support**
1 Hippocampal place cells can perform approximate Bayesian inference	Place field size depends on uncertainty (e.g. proximity of landmarks) in a Bayesian fashion	Place field sizes (recorded from hippocampal neurons of behaving rats) are correlated with uncertainties predicted by a Bayesian model (Chapter 4)
2 Spatial uncertainty is represented as the size of place cell firing fields		
3 When revisiting a place, estimates of recently traversed locations and encountered landmarks are updated in an approx. Bayes-optimal fashion	After revisiting parts of an environment, place fields should shift, and recently active place cells should re-activate. Errors should conform to Bayesian predictions	Neural: none in this work, but place fields seem to shift after revisits (Mehta et al., 2000), and recently active place cells do reactivate ('replay') (Carr et al., 2011). Behavioural: errors correlate with predictions (Chapter 6)
4 The structure of spatial representations arises from clustering	Landmarks which are co-represented (belong together) in participants' spatial memory should be closer in these features than those not belonging together	Neural: none in this work. Behavioural: the probability of two landmarks being co-represented is strongly correlated with distances along these specific features. These distances allow prediction of participant representation structure (Chapter 5)
5 This clustering mechanism operates on features including Euclidean distance, path distance, boundaries, visual and functional similarity		

Table 1.2: Hypotheses of the models presented in this work, and empirical support. Place cell electrophysiological recording data was acquired outside this PhD. All other data has been collected by the author, unless otherwise specified.

1.4 Outline and Contributions

This thesis is presented in the Alternative Format allowed by the University of Manchester presentation of theses policy [6], which allows incorporating sections in a format suitable for publication in peer-reviewed journals. We chose the alternative format to more easily accommodate already published work, to reduce risks of self-plagiarism, and because of the largely self-contained nature of our individual results chapters. Thus, in what follows, the literature review (Chapter 3) and the three chapters (4-6) reporting the results, are copies of papers either accepted by or submitted to peer-reviewed journals. The following list summarizes these papers and the contributions[7] therein:

- Chapter 3: Madl T., Chen K., Montaldi D. & Trappl R., 2015. Computational cognitive models of spatial memory in navigation space: A review. *Neural Networks, 65, 18-43.*
  Contributions: 1) a systematic review of representative cognitive models concerned with navigation-scale spatial memory, falling into symbolic, neural network, or cognitive architecture models, including a comparative table of the characteristics of these models.

- Chapter 4: Madl T., Franklin S., Chen K., Montaldi D. & Trappl R., 2014. Bayesian Integration of Information in Hippocampal Place Cells. *PLoS ONE 9(3), e89762*
  Contributions: 2) first quantitative electrophysiological validation of the representation of spatial uncertainty in the brain, and of Bayesian integration of spatial information in the brain, in three different environments (using data acquired outside this PhD). 3) Formulation and empirical support for an inference mechanism based on coincidence detection (falling into the camp of sampling-based models of neural inference)

- Chapter 5: Madl T., Franklin S., Chen K., Trappl R. & Montaldi D., submitted. Exploring the structure of spatial representations. *Cognitive Processing*

[6] http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=7420

[7] In all publications, Madl wrote the draft of the paper, developed the software, designed the experiments, recruited and tested the participants where applicable, and analysed the data. Corrections suggested by Chen, Montaldi, and Franklin were incorporated into the final drafts by Madl after discussions with these co-authors. All publications were supervised by Chen and Montaldi, with Chen mainly commenting on mathematical and computational issues, and Montaldi on psychological and neuroscientific issues.

Contributions: 4) behavioural evidence for clustering as the normative principle underlying spatial representation structure, and 5) the first computational model of navigation-scale spatial representation structure on the individual level (able to predict this structure in participants' long-term spatial memory from the geospatial properties of an environment)

- Chapter 6: Madl T, Franklin S, Chen K, Montaldi D & Trappl R, submitted. Towards real-world capable spatial memory in the LIDA[8] cognitive architecture. *Biologically Inspired Cognitive Architectures*
  Contributions: 6) integration of three spatial mechanisms capable of dealing with uncertainty and noise into a comprehensive cognitive architecture (localization, map structuring, map correction), and 7) embodying this architecture on a robot, allowing demonstration of the model functionality in a realistic robotic simulator. 8) Proposal of a biologically plausible mechanism for correcting errors in learned maps when revisiting an already known place (the 'loop closure' problem, well known in robotics, but neglected in cognitive science), and evaluation against behaviour data regarding cognitive map accuracy in human subjects.

The model best accounting for spatial memory structure presented in Chapter 5 also constitutes a novel kind of metric learning in machine learning, based on the idea of learning a similarity function in the space of absolute pairwise differences (as opposed to e.g. a Mahalanobis distance function). Although proposed before in a similar form for person re-identification in the computer vision community (Zheng et al., 2011), the insight that this space contains neglected information which can be utilized to improve performance in general (not just on image data), and the general formulation allowing arbitrary constituent models for learning a metric in this space, are a novel contribution (9). Since it is too far from the topic of this thesis, metric learning in absolute pairwise difference space is only described briefly (to the extent required to model cognitive map structure) in Chapter 5. Applications and results on other kinds of data, with other constituent models, and in a semi-supervised setting, and are briefly summarized in Appendix E.

Before presenting the mentioned papers constituting the literature review and results chapters, we briefly overview the computational methods employed during this research in Chapter 2 (they are also described in the respective results chapters). After

[8]LIDA stands for Learning Intelligent Distribution Agent, and is reviewed in a paper co-authored during this PhD but not included in this thesis: (Franklin et al., 2014)

the computational methods, we present the literature review (Chapter 3) and results (Chapter 4-6) in the form of published or submitted papers. Subsequently, we continue to discuss the implications of our results, the neural implementability of these mechanisms, and the shortcomings and limitations of our models in Chapter 7. We conclude in Chapter 8 with a conclusion and an outline of potential future work opened up by this research.

We note that the line of criticism mentioned regarding the neural implementability of the high-level probabilistic models of localization in the previous section also apply to our proposed mechanism of cognitive map structuring (Chapter 5). Although it is intended to be a cognitive and not a neural model, we have argued that consistency with the underlying neuroscience can and should play a role in constraining the space of possible models, and evaluating models, even on the algorithmic level. But the map structuring mechanism in Chapter 5 is, to our knowledge, the first formal model of the observed structure in cognitive maps, both on Marr's computational and algorithmic levels. We did not have the time and resources to extend it down to include a plausible neural implementation within this PhD.

Finally, work done during this PhD has contributed to two more publications which are not included in this thesis (the former because it is a conference paper, whereas University policy requires alternative format theses to contain journal papers instead; and the latter because it does not fit in well with the main topic):

Chapter 2

Computational Methods

As mentioned in the Introduction, the goal of this thesis is bringing computational cognitive models closer to being able to function in realistic environments under conditions of uncertainty, by proposing probabilistic models of spatial cognition which are implementable in brains. Probabilistic models have become successful and widespread in domains requiring the representation and manipulation of uncertainty, including artificial intelligence (Russell & Norvig, 2009), robotics (Thrun et al., 2005), and machine learning (Bishop, 2006). They have also been successfully employed in cognitive modelling (Chater et al., 2010) and in neuroscience (Knill & Pouget, 2004) - although there is little empirical evidence for particular neural implementations of probabilistic mechanisms as of yet (Griffiths et al., 2008; Vilares & Kording, 2011; Pouget et al., 2013).

This section briefly reviews the computational methods employed in this thesis. Figure 2.1 shows an overview over all employed methods, and the way they are utilized to support the mechanisms, algorithms, and cognitive models presented below. Figure 2.2 connects these computational mechanisms to their suggested implementation in brains (arguments and evidence for the neuroscientific plausibility of Bayesian localization are presented in Chapter 4).

To be able to plan novel routes in pursuit of its goals, an agent (whether biological or artificial), at a minimum, needs to be able to localize itself, its goal, and possible obstacles; and needs to do so in the face of a noisy and inaccurate sensory apparatus. From a probabilistic perspective, this localization problem can be described as a Bayesian network (see Figure 2.2B). In order to avoid having to perform calculations over every location ever visited, and every landmark ever observed, as done in many robotics solutions (Durrant-Whyte & Bailey, 2006; Bailey & Durrant-Whyte, 2006),

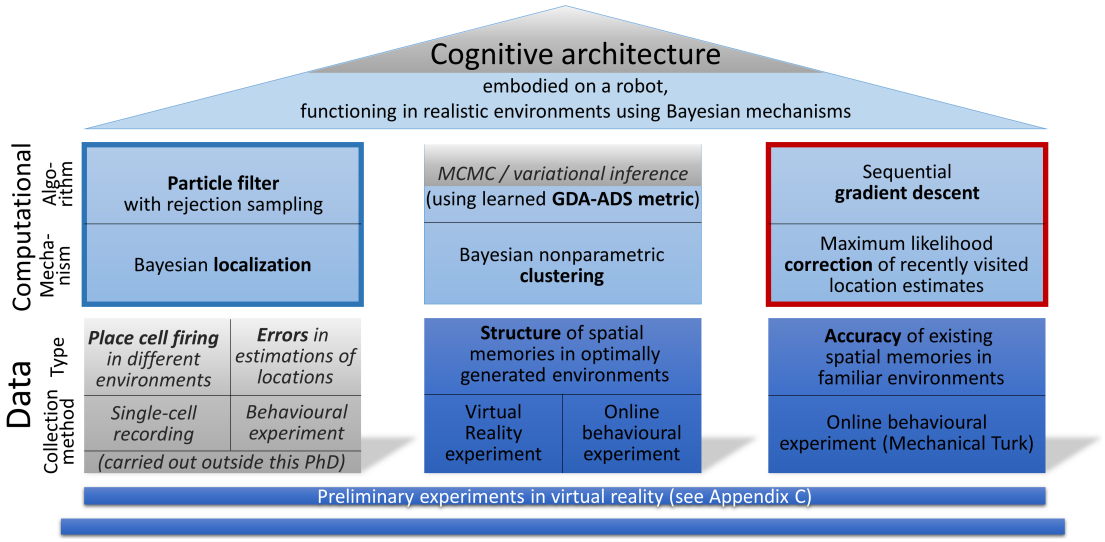


Figure 2.1: Overview of how the methods in this thesis help support real-world capable models of cognition, roughly divided into empirical methods (bottom half) and computational methods (top half). Gray boxes contain data/code used to substantiate or implement some models, but not gathered/implemented by us.

we split it into sub-problems.

Specifically, an approximate solution of this problem can be split into Bayesian cue integration for integrating noisy observations into a location estimate (Section 2.2), Bayesian localization for maintaining this location estimate through time (Section 2.3), and maximum likelihood-based correction for fixing the most recent location estimates when revisiting a location (Section 2.4). We suggest a rejection sampling-based algorithm for the former two, implementable through coincidence detection in hippocampal place cells (Chapter 4), and a gradient descent-based solution for the latter, implementable by reverse replay in the hippocampus (Chapter 6). We will present empirical evidence for these claims in those chapters, both from single-neuron recordings in live animals (collected outside this PhD) and from behavioural experiments performed online with participants recruited from Amazon's Mechanical Turk[1].

These mechanisms help inferring spatial locations in the environment from noisy observations, in a neurally and psychologically plausible fashion, as we will argue below. However, in a system operating under limited time and resources, these locations also need to be stored efficiently, such that they can be rapidly accessed. Hierarchical representations facilitate such desirable properties, and have been argued to be prevalent in human cognition (Cohen, 2000; Gobet et al., 2001). There is strong evidence

[1] https://www.mturk.com

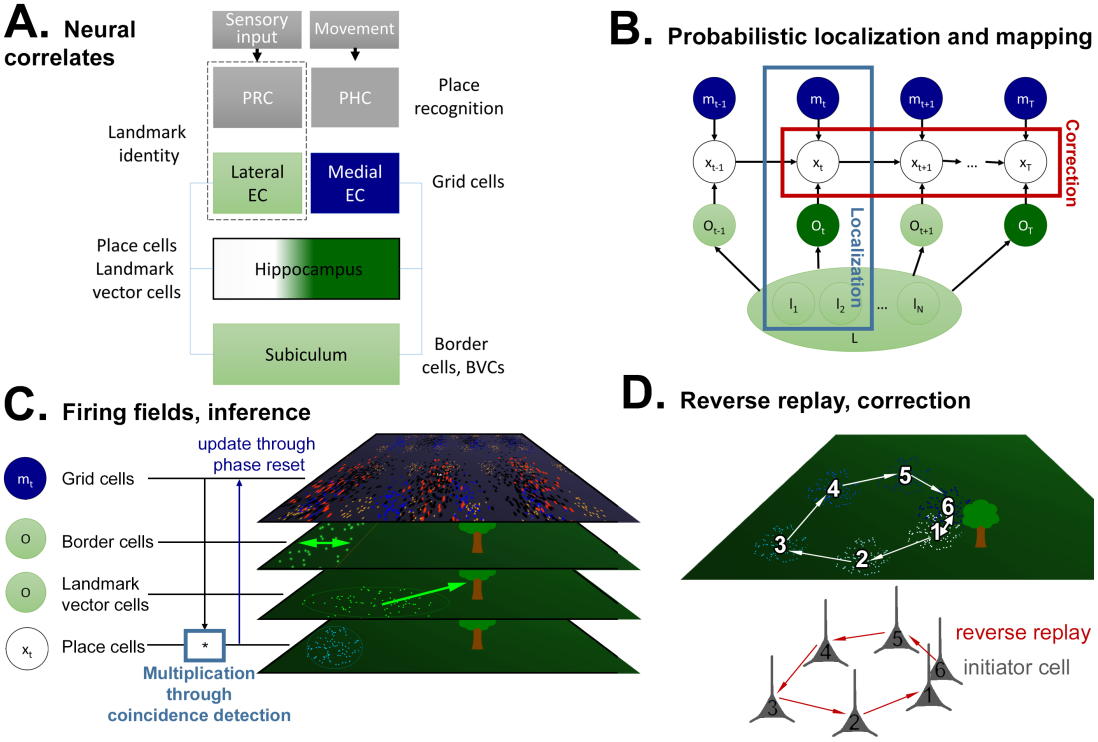


Figure 2.2: Probabilistic spatial localization and mapping implementable by brains. A: Neural correlates of localization. PRC: Perirhinal cortex, PHC: Parahippocampal cortex, EC: Entorhinal cortex (see Chapter 3 for details; and (Deshmukh et al., 2013) for evidence of landmark vector cells). B: Probabilistic graphical model of the simultaneous localization and mapping problem (Thrun & Leonard, 2008). Instead of capturing all correlations introduced through the landmarks, which requires vast computational resources, our model separately solves Bayesian localization with only local landmarks, and map correction ('pose optimization' in SLAM) with only loop closure constraints. See Chapter 2 for notation and details. C: Illustration of firing fields during localization. Coloured dots represent spikes of the respective cells at specific locations. Path integration (grid cells) and boundary and landmark information (border cells, landmark vector cells) is integrated in place cells, using coincidence detection (rejection sampling) to obtain a near-optimal location estimate. This new estimate is used to update grid cell representations via phase reset to combat accumulating path integration errors (see Chapter 4). D: Illustration of a small loop (firing fields 1-6) which can be corrected upon recognizing the same landmark at positions 1 and 6 via reverse replay, by reactivating place cells 6-1 and shifting their place fields proportionally (see Chapter 6).

that human spatial memories in particular are organized hierarchically (Hirtle & Jonides, 1985; McNamara et al., 1989; Greenauer & Waller, 2010), but the principles underlying these structures have not been known. We suggest a Bayesian nonparametric clustering model for structuring object representations under a subject-specific metric to account for human cognitive map structure (Section 2.5), and present empirical evidence for this claim gathered from virtual reality and real world environments in Chapter 5.

These probabilistic models for inferring self locations and object locations and structuring their representations constitute the pillars of a cognitive software agent able to function in a realistic robotic simulator, which provides the same interfaces as a real robot (and would allow this agent to run on a real robot without modifications to its code) (Rusu et al., 2007). We have implemented this agent within the LIDA (Learning Intelligent Distribution Agent) cognitive architecture, extending it with a spatial memory module and the described probabilistic models, integrating them with the other mechanisms already implemented in LIDA. Describing LIDA is outside the scope of this thesis, but see the review by Franklin et al. (2014), co-authored during this PhD.

Figure 2.2 above provides an overview over how the Bayesian mechanisms summarized above may be implemented in spatially relevant brain areas, and pointers to the parts of this thesis substantiating these connections; lending credence to our claim that our probabilistic models are neurally plausible (implementable in brains). Chapter 4 provides the first neural-level evidence for Bayesian inference in these brain areas.

2.1 Probabilistic modelling

Probabilistic models use probability distributions to represent quantities and the uncertainties associated with them, utilizing probability theory to manipulate these distributions (Ghahramani, 2015). Two basic rules provide the foundation, and together yield Bayes' theorem, which underlies Bayesian modelling. The *sum rule* takes the form

$$p(Y) = \sum_X p(Y,X), \tag{2.1}$$

where $p(X,Y)$ is the joint probability (i.e. the probability of random events X and Y) and the summation is over all values which Y could possibly take. $p(X)$ is also referred to as the marginal probability, and the summation in Equation 2.1 is also

called marginalization (which is especially useful to make inferences about variables of interest by summing out all other variables). The *product rule* states that

$$p(Y,X) = p(Y|X)p(X) = p(X|Y)p(Y), \tag{2.2}$$

where $p(Y|X)$ is the conditional probability (i.e. the probability of Y given X). Combined, they yield *Bayes' theorem*:

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} = \frac{p(X|Y)p(Y)}{\sum_Y p(X,Y)}. \tag{2.3}$$

In the context of a probabilistic model, defined by a number of parameters encoded in Y (such as the current coordinates of an agents location), and given some observed data encoded in X (such as the distances to landmarks), we can use Equation 2.3 to calculate a *posterior* probability distribution of model parameters, combining *prior* knowledge (or assumptions) $p(Y)$ with the *likelihood* $p(X|Y)$.

The sections below summarize computational-level solutions to the problems required for real-world spatial cognition outlined in Chapter 1 in this probabilistic framework. As mentioned there, the goal of this work is contributing to the understanding of spatial information processing in brains and minds, and not finding particularly accurate solutions to these problems. Numerous algorithms capable of much more accurate localization and mapping and making less restrictive assumptions have been proposed in probabilistic robotics (Thrun et al., 2005), more specifically simultaneous localization and mapping (SLAM) - see (Thrun & Leonard, 2008; Durrant-Whyte & Bailey, 2006; Bailey & Durrant-Whyte, 2006) for reviews and (Tuna et al., 2012) for a more recent evaluation.

Our particular computational-level solutions for estimating locations utilize stronger simplifications compared to the state of the art in SLAM. We are applying existing computational and mathematical tools to cognitive and neural mechanisms, following a long and successful history of this approach in the field of computational cognitive modelling (Sun, 2008), which can be seen as a branch of applied computer science. In this field, simplicity and approximations can be assets; since humans are unlikely to use computationally complex, optimal statistical models (see e.g. (Van Rooij, 2008; Simon, 1955)). A simpler, sub-optimal model which nevertheless explains empirical data better, and is more consistent with neural anatomy, is better suited to modelling cognition than an intractable or implausible optimal model. The implementation of these abstract methods in a way consistent with the neuroscience and psychology of

spatial memory is novel, as is their integration with a comprehensive cognitive architecture and their substantiation with empirical data (see Section 1.4 for the full list of novel contributions).

2.2 Bayesian cue integration

One concrete application of Equation 2.3 is the inference of the most likely current location of an animal, given some observations regarding the distance of a number of landmarks. For simplicity, we assume 1) a uniform prior over these observations, and 2) conditional independence of the observations given the location. The posterior probability of the current location $p(\boldsymbol{x}|O)$, given a location prior $p(\boldsymbol{x})$ and some observations $\boldsymbol{o}_1, ..., \boldsymbol{o}_N \in O$ (and a normalization constant γ), is

$$p(\boldsymbol{x}|O) = \frac{p(\boldsymbol{x})p(O|\boldsymbol{x})}{p(O)} = \gamma p(\boldsymbol{x})p(O|\boldsymbol{x}) \tag{2.4}$$

The prior can be obtained by adding up self-motion signals (a process called 'path integration' or dead reckoning - see Chapter 3). Individual observation distributions can express distance measurements to landmarks, and can be multiplied due to their conditional independence:

$$p(\boldsymbol{x}|O) = \gamma p(\boldsymbol{x}) \prod_{i=1}^{N} p(O_i|\boldsymbol{x}). \tag{2.5}$$

For now, we further assume that each of these variables is normally distributed. We will use this simplified formulation to predict the sizes of place cell firing field in Chapter 4; but will implement our localization model without this restrictive assumption (see next section). The Gaussian assumption makes it straightforward to derive the variance S_L of the normal/Gaussian posterior location distribution $p(\boldsymbol{x}|O) = \mathcal{N}(\boldsymbol{x}; \mu_L, S_L)$ from the variances of the prior and of the likelihood distributions S_x and $S_{o,i}$ (see e.g. Wu (2004) for the derivation of the parameters of products of Gaussian distributions):

$$S_P = (S_x^{-1} + \sum_{i=1}^{N} S_{o,i}^{-1})^{-1}. \tag{2.6}$$

In the one-dimensional case, the variance is the square of the standard deviation σ. We can say that the standard deviation of a Gaussian distribution is a measure of the 'uncertainty' associated with it (as it measures the spread among possible values - the more certainly a value is known, the lower the associated σ of the distribution

describing it). Assuming that the observation uncertainties $\sigma_{o,i}$ depend linearly on the respective distances d_i, such that $\sigma_{o,i} = s \cdot d_i$ (Chapter 4 provides justifications and evidence for this linear relationship), we obtain the standard deviation of the location posterior for a given set of measurement distances:

$$\sigma_P(d_1,...,d_N) = \sqrt{(\sigma_x^{-2} + s\sum_{i=1}^{N} d_i^{-2})^{-1}}. \tag{2.7}$$

Chapter 4 uses Equation 2.7 to test the hypotheses that place cells may represent uncertainty and perform Bayesian cue integration. Although place cells constitute a two-dimensional representation, this one-dimensional treatment of observation likelihoods is an acceptable approximation in the kinds of environments from which the data was collected (rectangular boxes without landmarks, where the axes can be assumed to be independent as they are orthogonal, and a very narrow, circular track with landmarks, where the width can be neglected as it is less than 3% of the length).

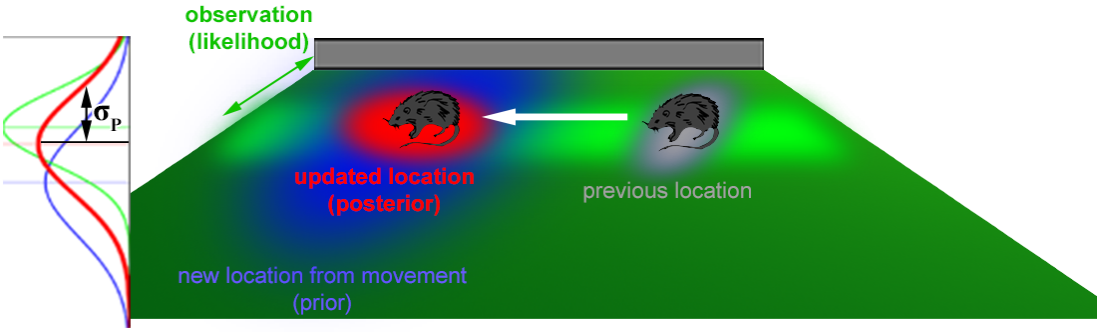


Figure 2.3: Bayesian cue integration for localization. Illustration of how an animal might use its prior location belief (blue) estimated from its movement, and distance distributions e.g. to a boundary (green) to obtain a corrected location estimate (red) using Bayesian inference.

2.3 Bayesian localization

To maintain a location estimate through time, the kind of cue integration described above has to be performed regularly (after every time step). One source of location information is adding up each movement vector, a process called odometry in robotics and 'path integration' in cognitive science and biology. However, movements are not accurate and noise free in real-world environments - each movement vector contains a

slight error, and these errors add up over time. Eventually, these accumulating errors render the location estimate useless, if sensory information is not used to correct it.

Bayesian localization is concerned with correcting the location estimate in time using noisy observations (Thrun et al., 2005). Conceptually, it entails performing the Bayesian cue integration to correct location estimates *recursively*, after every movement / time step. Its operation can be summarized in three stages, which are performed iteratively at every time step: 1) movement (adding the current movement), 2) correction of the location estimate via Bayesian cue integration, 3) updating of the path integration estimate for use in the next iteration.

Unlike the simplified treatment above, which has considered only one snapshot in time, Bayesian localization considers the posterior at any time step t. This posterior distribution has to depend on all movements until now: $\boldsymbol{m}_{1:t}$, on all observations until now: $O_{1:t}$, as well as the locations of known landmarks $\boldsymbol{l}_{1:N}$. Extended by these dependencies, the posterior location distribution from Equation 2.4 becomes

$$p(\boldsymbol{x}_t|\boldsymbol{m}_{1:t}, O_{1:t}, \boldsymbol{l}_{1:N}) = \gamma p(O_t|\boldsymbol{x}_t, \boldsymbol{l}_{1:N}) p(\boldsymbol{x}_t|\boldsymbol{m}_{1:t}), \tag{2.8}$$

through simple application of Bayes' theorem. We can use the sum rule (with the sum replaced by an integral for dealing with continuous distributions) to model the 'path integration' (odometry) mechanism which provides the prior in Equation 2.8:

$$p(\boldsymbol{x}_t|\boldsymbol{m}_{1:t}) = \int p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{m}_{t-1}) p(\boldsymbol{x}_{t-1}|\boldsymbol{m}_{1:t-1}) \mathrm{d}\boldsymbol{x}_{t-1}. \tag{2.9}$$

This equation allows inferring the current location prior based on the most recent movement $\boldsymbol{m}_{t-1}$ and on the previous location estimate $\boldsymbol{x}_{t-1}$ by marginalizing (integrating out) the previous location. This is a recursive formulation which yields a path integration estimate based on a starting location and a number of movements. This estimate is subject to accumulating errors. However, crucially, the corrected previous location estimate (previous posterior) can be used instead of the uncorrected previous path integration estimate. Using this insight, replacing $p(\boldsymbol{x}_{t-1}|\boldsymbol{m}_{1:t-1})$ in Equation 2.9 by the previous location posterior $p(\boldsymbol{x}_{t-1}|\boldsymbol{m}_{1:t-1}, O_{1:t-1}, \boldsymbol{l}_{1:N})$ and plugging the resulting prior into Equation 2.8 yields

$$\begin{aligned} p(\boldsymbol{x}_t|\boldsymbol{m}_{1:t}, O_{1:t}, \boldsymbol{l}_{1:N}) = \gamma p(O_t|\boldsymbol{x}_t, \boldsymbol{l}_{1:N}) \int p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{m}_{t-1}) \cdot \\ p(\boldsymbol{x}_{t-1}|\boldsymbol{m}_{1:t-1}, O_{1:t-1}, \boldsymbol{l}_{1:N}) \mathrm{d}\boldsymbol{x}_{t-1} \end{aligned} \tag{2.10}$$

This recursive equation for updating location estimates is a Bayes-optimal solution to the localization problem and allows inferring the current location based on two conditional densities: a model specifying the effect of movements on the location (a 'motion model'):

$$p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{m}_{t-1}) \tag{2.11}$$

and a model specifying the probability distribution of the current measurements O_t at a position $\boldsymbol{x}_t$ given the landmarks $\boldsymbol{l}_{1:N}$ (a 'sensor model'):

$$p(O_t|\boldsymbol{x}_t, \boldsymbol{l}_{1:N}). \tag{2.12}$$

Equation 2.10 is the mathematical formulation of Bayesian localization, which, conceptually, iterates over the three stages mentioned above: movement (application of the motion model), correction (via Bayes' theorem), and update.

As argued in Chapter 4 and Appendix A, the activity of hippocampal place cells can be viewed as samples from probability distributions, and the size of their firing fields can be partially predicted by a Bayesian model. We will also argue based on existing evidence that the 'motion model' is implemented by a neural path integrator in the entorhinal cortex, and that neurons with boundary-related firing might implement the 'sensor model'.

Such a sampling-based representation of uncertainty in these spatially relevant brain areas naturally suggests employing a sequential Monte Carlo method (Doucet et al., 2000) to computationally evaluate the integral in equation 2.10 (the same model using samples for representation might as well use them for inference). Although the usual method of choice in robotics is importance sampling (Montemerlo & Thrun, 2007; Thrun et al., 2005), we approximate the integral using rejection sampling (Doucet et al., 2000), and will argue in Chapter 4 and Appendix A that coincidence detection (CD) in hippocampal place cells can implement this mechanism (since CD can filter out samples at locations where different measurements and path integration disagree, and keeps the ones where they agree - see illustration in Figure 2.2C, and Appendix A for mathematical details).

From a computational point of view, instead of inferring the parameters of the location posterior distribution (e.g. the mean and variance in case of a Gaussian), we represent it by sampling multiple location hypotheses. The mean of these hypotheses corresponds to the 'best guess' estimate, and their standard deviation to the associated uncertainty. Apart from the empirical evidence for sampling based mechanisms in the

brain (see Chapter 4, as well as (Fiser et al., 2010) for a more general review), the main advantage of this approach is the ability to represent free-form distributions (irregular, non-Gaussian, multimodal distributions etc.).

Particles (samples, hypotheses) $\boldsymbol{x}^i$ are generated regularly based on self-motion information (linear and angular movement speed v) according to the motion model (Equation 2.11), performing path integration - in the simplest case: $\boldsymbol{x}_t^i = \bar{\boldsymbol{x}}_{t-1} + \boldsymbol{v}'\Delta t$ - at simulated timesteps Δt. Gaussian noise is multiplied to the estimated speed to obtain a distribution of hypotheses reflecting the path integration / odometry uncertainty (neither animals nor robots can estimate their movement speed with perfect accuracy):

$$\boldsymbol{v}' = \boldsymbol{v}_{true} \cdot \mathcal{N}(\mathbf{1}, \begin{bmatrix} \sigma_v^2 & 0 \\ 0 & \sigma_\omega^2 \end{bmatrix}),$$

where σ_v^2 and σ_ω^2 are model parameters representing the variance in the linear and angular speeds, respectively. Since the estimate of $\boldsymbol{v}$ is noisy, accumulating errors would lead to an increase of uncertainty and the corruption of the distribution represented by the set of particles, which is why correction with the sensor model is required.

Under Gaussian assumptions, this correction can be implemented simply by multiplying a path integration prior and a number of sensory likelihoods and solving for the means and variances (Equation 2.5). The ensuing algorithm for Bayesian localization is trivial. When using samples instead of a Gaussian to represent the posterior, the correction can be implemented by rejection sampling (Doucet et al., 2000), i.e. by deleting hypotheses inconsistent with sensory measurements (see Figure 2.4). The derivation of why this rejection sampling scheme approximates the true Bayesian posterior can be found in Appendix A. Details regarding how brains could implement this algorithm are discussed in Chapter 4.

2.4 Maximum likelihood map error correction

Landmark location estimates can be updated in the same way as the agents' location estimates $\boldsymbol{x}$, by integrating new observations into the posterior distribution representing these locations (either in the form of Gaussians or of samples from this distribution). With infinitely many particles, the algorithm presented in Figure 2.4 would suffice to maintain correct location estimates.

However, there are practical limits on the particle budget (due to limited computational resources in computers, and due to limited firing rates in neurons). This neces-

Algorithm 2.3.1: MOVEMENT($samples, \mathbf{v}, N$)

$1: prevmean \leftarrow mean(samples)$
$2: newsamples \leftarrow \{\}$
$3:$ **for each** $particle \in samples$
$4:$ $\quad newsamples \leftarrow newsamples \cup motionModel(particle, \mathbf{v})$
$5:$ **while** $count(newsamples) < N$
$6:$ $\quad newsamples \leftarrow newsamples \cup motionModel(prevmean, \mathbf{v})$
$7: return(newsamples)$

Algorithm 2.3.2: CORRECTION($samples, \mathbf{O}, \mathbf{L}$)

$1: newsamples \leftarrow \{\}$
$2:$ **for each** $particle \in samples$
$3:$ $\quad likelihood \leftarrow sensorModel(particle, \mathbf{O}, \mathbf{L})$
$4:$ $\quad$ **if** $random() < likelihood$
$5:$ $\quad\quad newsamples \leftarrow newsamples \cup particle$
$6: return(newsamples)$

Algorithm 2.3.3: LOCALIZATIONSTEP($posteriorsamples, \mathbf{v}, \mathbf{O}, \mathbf{L}, N$)

$1: timestep++$
$2: movedsamples \leftarrow movement(posteriorsamples, \mathbf{v}, N)$
$3: correctedsamples \leftarrow correction(movedsamples, \mathbf{O}, \mathbf{L})$
$4: return(correctedsamples)$

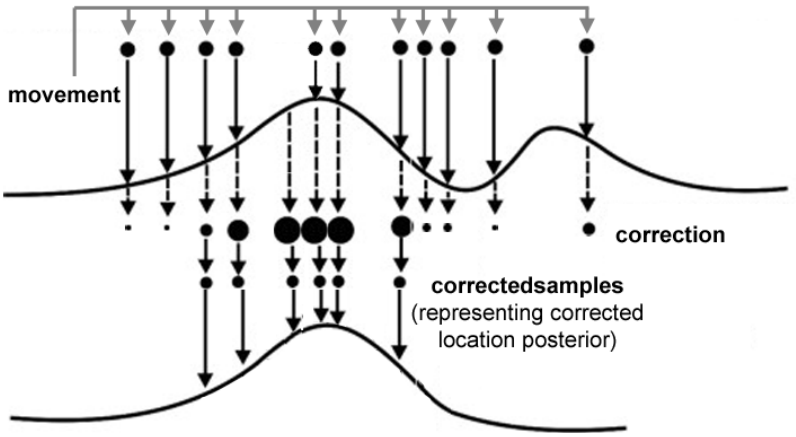


Figure 2.4: Bayesian localization algorithm with rejection sampling, producing updated posterior samples given the samples from the previous posterior, speed vector $\boldsymbol{v}$ and observations O at the current time step, landmarks L, and a particle budget N

sarily leads to errors whenever there is no particle at the unobservable true location. Unfortunately, these errors add up as well. They become most pronounced when revisiting an already known part of the environment, i.e. when traversing a loop - although the agent has returned to its starting location, it will think that it is at a new location, and form new representations of the same place. Multiple such loops can lead to multiple redundant, erroneous representations.

The problem of how to correct spatial representations when revisiting a known place (not only the location estimate but also the estimated recent path and landmark locations) is the 'loop closing' problem in robotics (see e.g. (Williams et al., 2009; Thrun & Leonard, 2008)). Brains need to solve this problem as well - although human spatial representations are not perfectly accurate, humans are able to correct mistaken estimates when they recognize a revisited place. Interestingly, despite the abundant robotics literature on the topic of closing loops, this problem has been largely neglected in cognitive science literature.

Our cognitive model of loop closing is described in more detail Chapter 6. Here, we will briefly summarize its purely computational and mathematical aspects. We will assume that it is sufficient to correct the route taken during the loop, i.e. the most recent locations of the agent; and that the landmarks are corrected by the same amount as the location closest to them. That is, when performing large-scale loop closing, the model in Chapter 6 applies the same correction to a position and the local landmarks around it (a simplification justified based on neuroscientific evidence in that Chapter). We also make the assumption that correction only concerns position representations and not angular representations, once again based on neural evidence. Hippocampal 'reverse replay' (Carr et al., 2011) (the re-activation of recently active place cells) is a plausible mechanism for correcting the recent route when revisiting a location, as argued in Chapter 6, but such a mechanism has not been found for neurons with direction-specific firing.

When revisiting a known place, the recently traversed path has to be corrected using the discrepancy between the previously and recently estimated location of the revisited place. Naturally, when an agent recognizes that it is in the same place it has visited before, the current estimate has to be reset to be equivalent to the previous estimate of the same location. However, it is not obvious how to correct the other recently visited locations $\boldsymbol{x}_0, ..., \boldsymbol{x}_m \in X$ along the recent path X. Let $\boldsymbol{c}_1, ...\boldsymbol{c}_m \in C$ denote a set of vectors we will call constraints, each expressing how far apart two locations should be according to some measurement. That is, each constraint specifies the difference

between two locations $\boldsymbol{c} = \boldsymbol{x}_a - \boldsymbol{x}_b$, and each is associated with a measurement uncertainty S_c in the form of the covariance matrix of a normal distribution. For locations traversed in sequence, $\boldsymbol{c}$ and S_c is given by the motion model (by path integration). For revisited locations, $\boldsymbol{c}$ is zero.

According to Bayes' theorem, and assuming that constraints are independent given the location, the recent path depends on the product of the constraint distributions; and the best path estimate is the one that maximizes:

$$P(X|C) \propto \prod_{i=1}^{m} P(\boldsymbol{c}_i|X) \tag{2.13}$$

Each $P(\boldsymbol{c}_i|X)$ expresses the likelihood that this constraint is satisfied by the path X, as a Gaussian distribution: $P(\boldsymbol{c}_i|X) \propto \mathcal{N}(\boldsymbol{x}_a - \boldsymbol{x}_b; \boldsymbol{c}_i, S_i)$ (where $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$ are the location estimates which should have the distance $\boldsymbol{c}_i$ according to this constraints). We are interested in the maximum of Equation 2.13, which is equivalent to the minimum of its negative logarithm. Let $\boldsymbol{d}_i = \boldsymbol{x}_a - \boldsymbol{x}_b - \boldsymbol{c}_i$ be the discrepancy between the constraint and the locations it concerns within the path. With noise-free measurements, all d_i would be zero; but since sensory errors may add up, there will be discrepancies (e.g. after traversing a loop, the estimate of the first visit $\boldsymbol{x}_a$ and second visit $\boldsymbol{x}_b$ may differ, but $\boldsymbol{c}_i = 0$ for the revisited place). Then, the most likely path is given by:

$$X_{ml} = \arg\max_X P(X|C) = \arg\min_X -logP(X|C) = \arg\min_X \sum_{i=1}^{m} ||\boldsymbol{d}_i||_{S_i^{-1}}. \tag{2.14}$$

Equation 2.14 mathematically describes the maximum likelihood error correction problem for loop closing. It tries to minimize the discrepancies between the constraints and the estimated locations, taking into account the constraint uncertainties S_i by utilizing the Mahalanobis distance[2] to measure the discrepancy.

There are several ways to solve Equation 2.14. For our cognitive model (Chapter 6), we chose sequential gradient descent, because it can be implemented in biological neurons (Bengio et al., 2015). Olson et al. (2006) derive the starting point for this solution. They suggest the following gradient with respect to constraint i, depending on a learning rate α, a full Jacobian J of the constraints with respect to the path, and the Jacobian J_i of constraint i:

[2]The Mahalanobis distance is defined as $||\boldsymbol{x}_1 - \boldsymbol{x}_2||_S = \sqrt{(\boldsymbol{x}_1 - \boldsymbol{x}_2)^T S(\boldsymbol{x}_1 - \boldsymbol{x}_2)}$

$$\Delta X \approx \alpha (JS^{-1}J)^{-1} J_i^T S_i^{-1} \boldsymbol{d}_i. \tag{2.15}$$

Because of the incremental structure of the Jacobian, it is possible to simplify this expression (see Chapter 6). Making use of this structure, and defining a loop precision parameter $A_i = S_i/S_P$ specifying the ratio of the uncertainties of loop closure constraints (added when revisiting a place) and path integration constraints, the gradient for each individual location within the loop becomes:

$$\Delta \boldsymbol{x}_j \approx \alpha d_i \frac{\sum_{k=a+1}^{j} S_i^{-1}}{\sum_{k=a+1}^{min(j,b)} S_P^{-1}} = \alpha A_i \boldsymbol{d}_i p_j, \tag{2.16}$$

where $p_j = (min(j,b_i) - a_i - 1)/(b_i - a_i - 1)$ denotes how far x_j lies along the loop, with $0 \leq p_j \leq 1$. Unlike usual gradient descent procedures, in this particular case we know that $\Delta \boldsymbol{x} \leq \boldsymbol{d}_i$ must hold, and can prevent the algorithm from overshooting, accelerating its convergence. Figure 2.5 contains the algorithm using this gradient to correct location estimates when revisiting a place. We will use this algorithm in Chapter 6 to account for human cognitive map accuracy, as a part of a cognitive architecture embodied on a robot and learning maps in realistic simulated environments.

Algorithm 2.4.1: CORRECTPATH($X, loopConstraints, \alpha, A, N$)

```
1 : while i < N and not converged
2 :   i++
3 :   for each a,b ∈ loopConstraints
4 :     discrepancy ← X_a − X_b
5 :     for each j ∈ (a,b]
6 :       p ← (min(j,b) − a − 1)/(b − a − 1)
7 :       β ← min(αA · discrepancy, discrepancy)
8 :       X_j ← X_j + βp
9 : return(X)
```

Figure 2.5: Algorithm for correcting location estimates when revisiting places ('loop closing'), producing a corrected path given the estimates of locations X along that path (from Bayesian localization), a list of loop constraints indicating the same (revisited) places (from landmark recognition or place recognition), a learning rate α, a loop precision parameter A and an iteration budget N

2.5 Bayesian nonparametrics for map structuring

It has been suggested that map-like spatial representations are structured hierarchically (Hirtle & Jonides, 1985; McNamara et al., 1989; Greenauer & Waller, 2010), but no formal model has been put forth for a process that might account for this structure. We hypothesize in Chapter 5 that this process might be clustering. Computationally, we chose a Dirichlet Process Gaussian Mixture Model (DP-GMM) to account for the behaviour data we collected (see Chapter 5), for two reasons. First, DP-GMMs (unlike most clustering algorithms) are able to infer the number of clusters, not just cluster memberships; and are infinitely extensible (Rasmussen, 1999). Second, Bayesian nonparametric models with Dirichlet priors have a successful history in psychological modelling, e.g. of category learning and causal learning (Tenenbaum et al., 2011), transfer learning (Canini et al., 2010), and human semi-supervised learning (Gibson et al., 2013).

By 'map structure', here and in Chapter 5, we mean sub-map memberships. There is evidence that human spatial maps are hierarchical (Hirtle & Jonides, 1985; McNamara et al., 1989; Greenauer & Waller, 2010), just as geographical maps are - e.g. there is a map of the country and a map of the cities therein; and any given building may be represented not only on the country map but also on one of the city maps. Similarly, any object (e.g. building) memorized by a participant belongs to her map-like spatial representation ('cognitive map'), as well as to one of its sub-maps. We only consider a two-level hierarchy (map and sub-maps); thus, sub-map memberships fully describe our modelled map structure.

A number of features can influence spatial representation structure, including spatial distance and visual and functional similarity of landmarks. The importance of these features varies across participants, and these subject-specific importances have to be accounted for before the clustering process. We chose to implement a new metric learning method to do so (see below). Our model of spatial representation structure consists of these two components: a subject-specific metric, expressing the 'similarity function' between two buildings, and the DP-GMM model for clustering buildings under this metric.

As noted in the Introduction, unlike the rest of our work, we have not shown what the neural implementation of such a structuring process might look like. Some prior work exists showing the possibility of inference in hierarchical Bayesian models such as the DP-GMM, e.g. (Shi & Griffiths, 2009) - see (Sanborn, 2015) for a review. We have substantiated the psychological plausibility of this model by showing that it can

explain and predict human behavior data (Chapter 5), and leave the investigation of the biological plausibility of this specific mechanism for future work.

2.5.1 Dirichlet Process Gaussian Mixture Models for clustering

We will only describe the DP-GMM model very briefly, since it is a well-established model and since we did not implement it ourselves in this work (we used the *bnpy* Python library instead). See e.g. (Rasmussen, 1999) for its introduction, or (Gershman & Blei, 2012) for a tutorial. The DP-GMM partitions a number of data points x into K clusters by fitting a mixture of K Gaussian distributions to the data. It infers the number of clusters, as well as the means $\boldsymbol{\mu}_k$ and covariances Σ_k of each Gaussian, by inverting the generative process defined as follows:

$$\begin{aligned} \phi_k &\sim Beta(1, \alpha_1) \\ \boldsymbol{\mu}_k &\sim Normal(0, \mathbf{I}) \\ \Sigma_k &\sim Wishart(D, \mathbf{I}) \\ \pi_k &\sim SBP(\phi) \\ \boldsymbol{x}_t &\sim Normal(\boldsymbol{\mu}_{z_i}, \Sigma_{z,i}^{-1}), \end{aligned} \tag{2.17}$$

where SBP stands for the stick-breaking process for generating mixture weights: $\pi_k = v_k \prod_{j=1}^{k-1}(1 - v_j)$. Data can be generated from this model by first choosing a cluster with probabilities specified by mixture weights: $z \sim Cat(\pi)$, and then drawing an observation from the parameters of that cluster $\boldsymbol{x} \sim Normal(\boldsymbol{\mu}_z, \Sigma_z)$.

Given the data, the parameters of this model can be inferred using either a Monte Carlo chain sampling method (Neal, 2000) or variational inference (Blei et al., 2006). We did not implement an inference algorithm in this work; instead, we have used the *bnpy* Python library for this purpose. See (Hughes & Sudderth, 2013) for implementation details.

2.5.2 Metric learning in absolute pairwise difference space

In order to learn a suitable metric for our data, we had to develop a novel metric learning method, since the assumptions made by existing methods do not hold in our case. Neither the linear separability assumption (made by linear metric learning), nor the prerequisite of roughly isotropic variances along the features (made by RBF-based methods (Ong et al., 2005)) is the case for all subjects in our dataset (see Appendix E for further motivation and evaluation from a machine learning perspective).

Furthermore, our metric can naturally incorporate the hypothesis that building pairs belonging to the same representation should be located close to the origin in pairwise difference space (i.e. they should not be very different), and should be separable from building pairs belonging to different representations. These two distributions of pair differences can be naturally modelled using Gaussian distributions - see Chapter 5.

Our proposed method can be seen as a novel approach to perform non-linear metric learning using weak supervision in the form of pairwise constraints, in order to improve clustering performance, as pioneered by Xing et al. (2002). The problem to be solved can be defined as follows. Let $\mathcal{X} = (\boldsymbol{x}_i, ..., \boldsymbol{x}_n)$ be the feature vector representation of n objects (buildings on a cognitive map) which are to be clustered (assigned to representations we will call 'sub-maps'), where $\boldsymbol{x}_i \in \mathbb{R}^D$ are vectors with D dimensions. Let the set of m given labelled pairwise co-representation constraints be denoted by $\mathcal{C}$, where $|\mathcal{C}| = m$, and $c_{i,j} \in \mathcal{C}$ is

$$c_{i,j} = \begin{cases} 1, & \text{if } i \text{ and } j \text{ belong to the same sub-map (co-represented)} \\ 0, & \text{if } i \text{ and } j \text{ belong to different sub-maps (not co-represented)} \end{cases} \tag{2.18}$$

Our ultimate goal is to group the n objects into K clusters ('sub-maps'), such that objects of the same cluster are more similar to each other than to those of different clusters; taking into account the provided pairwise constraints to learn a good similarity metric for the given data.

Conventional approaches leveraging non-linear metric learning for this problem try to find a kernel Φ such that the clustering resulting from using the distance metric defined by that kernel, $d_m^2(\boldsymbol{x}_1, \boldsymbol{x}_2) = (\Phi(\boldsymbol{x}_1) - \Phi(\boldsymbol{x}_2))^T(\Phi(\boldsymbol{x}_1) - \Phi(\boldsymbol{x}_2))$, does not violate the provided constraints (ensures co-represented pairs are closer than other pairs, if possible), and often employ RBF kernels for this purpose, e.g. (Baghshah & Shouraki, 2010; Chitta et al., 2011).

In contrast, the proposed framework aims to learn the distribution of co-representation probabilities (whether or not two object should be linked) from the provided set of constraints, and constructs a pseudo-metric based on a generative model of co-representation probabilities. Crucially, this probabilistic model is defined on the vector space of absolute pairwise differences (APD), which allows learning the importance of each feature (a challenge for RBF kernels for data with non-isotropic variance). Learning in APD space has been proposed before by Zheng et al. (2011) (specifically for person re-identification in computer vision), but not as a general metric learning method. The

metric based on this generative model is a pseudo-metric, because it does not satisfy the conditions of subadditivity, $d_m(\boldsymbol{x},\boldsymbol{z}) \leq d_m(\boldsymbol{x},\boldsymbol{y}) + d_m(\boldsymbol{y},\boldsymbol{z})$ and the identity of discernibles, $d_m(\boldsymbol{x},\boldsymbol{y}) = 0$ if and only if $\boldsymbol{x} = \boldsymbol{y}$.

Let $[\Delta \boldsymbol{x}_{i,j}]_+ = \left(|\boldsymbol{x}_{i,k} - \boldsymbol{x}_{j,k}|\right)_{k=1}^{m}$ be the representation of each pair of objects (i,j) in APD vector space. The co-representation probability distribution, i.e. the posterior probability of any pair of objects belonging to the same cluster, given a pair of objects and some model parameters $\boldsymbol{\theta}$ is then

$$p(c = 1|\Delta \boldsymbol{x}, \boldsymbol{\theta}) \propto p(\Delta \boldsymbol{x}|c = 1, \boldsymbol{\theta}) p(c = 1|\boldsymbol{\theta}) \tag{2.19}$$

The likelihood $p(c = 1|\Delta \boldsymbol{x}, \boldsymbol{\theta})$, the model parameters $\boldsymbol{\theta}$ (as well as the prior) can be estimated from $\mathcal{X}$ and $\mathcal{C}$, even in closed form, using Gaussian Discriminant Analysis (GDA). This yields a suitable non-linear pseudo-metric based on this probability distribution - see Equation 2.20 -, such that objects likely to belong to the same cluster will be close, and those likely to belong to different clusters will be far apart; with these distances directly depending on co-representation probabilities.

$$d_m(\boldsymbol{x}_1, \boldsymbol{x}_2; \boldsymbol{\theta}) = 1 - p(c = 1|\Delta \boldsymbol{x}, \boldsymbol{\theta}) = p(c = 0|\Delta \boldsymbol{x}, \boldsymbol{\theta}) \tag{2.20}$$

A metric is well-suited for clustering if within-cluster instances are closer than across-cluster instances according to it. That is, if for any co-represented $\Delta \boldsymbol{x}_r$ and not co-represented $\Delta \boldsymbol{x}_n$ it holds that $d_r(\boldsymbol{x}_{r,1}, \boldsymbol{x}_{r,2}; \boldsymbol{\theta}) < d_n(\boldsymbol{x}_{n,1}, \boldsymbol{x}_{n,2}; \boldsymbol{\theta})$. It follows from Equation 2.20 that this is the case if the generative model learns to separate the absolute differences of within-cluster instance pairs from across-cluster pairs.

In the generative **GDA** model (Bensmail & Celeux, 1996), the likelihoods of a pair of instances either being co-represented (i.e. belonging to the same sub-map), or not being co-represented (i.e. belonging to different sub-maps) are each modelled using a multivariate Gaussian:

$$p(\Delta \boldsymbol{x}|c = i; \boldsymbol{\mu}_i, \Sigma_i) = (2\pi)^{-\frac{D}{2}} |\Sigma_i|^{-\frac{1}{2}} e^{-\frac{1}{2}(\Delta \mathbf{x} - \boldsymbol{\mu}_i)^\intercal \Sigma_i^{-1} (\Delta \mathbf{x} - \boldsymbol{\mu}_i)}, \tag{2.21}$$

where $i \in \{0, 1\}$. $(\boldsymbol{\mu}_1, \Sigma_1)$ are the means and covariances of the APD distances of co-represented pairs, and $(\boldsymbol{\mu}_0, \Sigma_0)$ those of not co-represented pairs. These parameters can be easily estimated from the two given sets of co-represented and not co-represented object pairs, respectively, by calculating their means and covariances.

From Equation 2.21 and Bayes' theorem, we obtain the generative probability required for the metric in 2.20, which then becomes:

$$d_m(\boldsymbol{x}_1, \boldsymbol{x}_2; \boldsymbol{\theta}) = 1 - \frac{p(\Delta\boldsymbol{x}|c=1; \boldsymbol{\mu}_1, \Sigma_1)}{\sum_{i\in\{0,1\}} p(\Delta\boldsymbol{x}|c=i; \boldsymbol{\mu}_i, \Sigma_i)} \tag{2.22}$$

Thus, the trained GDA-model can be used to calculate distances (Equation 2.22) between all pairs of objects in any testing data set. The data is projected under the metric in Equation 2.22 using distance-preserving embedding. We have used multi-dimensional scaling (MDS) for this purpose (Borg & Groenen, 2005). The result of this projection is a data set embedded such that Euclidean pairwise distances therein reflect the distances 2.20 in the original dataset.

We subsequently perform clustering of this resulting data, using a Dirichlet Process Gaussian Mixture Model (DP-GMM) (Rasmussen, 1999), since the number of clusters is unknown (see previous section). The resulting algorithm for structuring map representations is shown in Figure 2.6. It requires training data in the form of buildings for which it is known which representation they belong to (this can be inferred from recall lists of participants). We use this algorithm to predict the representation structure of participants' cognitive maps in advance in Chapter 5.

We point out that Equation 2.20 constitutes a general framework for metric learning using any model capable of producing probability estimates that two instances belong together. This includes the entire family of generative models in machine learning (see e.g. (Bishop, 2006)), as well as any discriminative model when combined with Platt scaling (Platt et al., 1999) for transforming discrete outputs into probabilities. Two example applications of this general metric learning framework are semi-supervised clustering (extending the algorithm in Figure 2.6 by using semi-supervised GDA), or semi-supervised classification. See Appendix E for these examples and the evaluation of their performance with different constituent models.

Algorithm 2.5.1: PREDICTMAPSTRUCTURE($X, knownX, knownStructure$)

$corepresented \leftarrow \{\}$
$notcorepresented \leftarrow \{\}$
for $i \in (1, |knownX|)$
  for $j \in (i+1, |knownX|)$
    if $knownStructure_i = knownStructure_j$
      $corepresented \leftarrow corepresented \cup (knownX_i - knownX_j)$
    else
      $notcorepresented \leftarrow notcorepresented \cup (knownX_i - knownX_j)$
$\mu_{co} \leftarrow mean(corepresented)$
$\Sigma_{co} \leftarrow cov(corepresented)$
$coprior \leftarrow \frac{|corepresented|}{|knownX|}$
$\mu_{not} \leftarrow mean(notcorepresented)$
$\Sigma_{not} \leftarrow cov(notcorepresented)$
$notprior \leftarrow \frac{|notcorepresented|}{|knownX|}$
$D \in \mathbb{R}^{|X|x|X|}$
for $i \in (1, |X|)$
  for $j \in (i+1, |X|)$
    $D_{i,j} \leftarrow 1 - \frac{coprior \cdot \mathcal{N}((X_i - X_j); \mu_{co}, \Sigma_{co})}{coprior \cdot \mathcal{N}((X_i - X_j); \mu_{co}, \Sigma_{co}) + notprior \cdot \mathcal{N}((X_i - X_j); \mu_{not}, \Sigma_{not})}$
$embedding \leftarrow MDS(D)$
$structure \leftarrow DPGMM(embedding)$
$return(structure)$

Figure 2.6: Algorithm for predicting participants' spatial representation structure, given the features of the new buildings to be structured, and given buildings with known structure (from a previous experiment) specifying which of these buildings were co-represented.

Chapter 3

Review of computational cognitive models of spatial memory

Publication 1 / 4. Madl T., Chen K., Montaldi D. & Trappl R., 2015. Computational cognitive models of spatial memory in navigation space: A review. *Neural Networks, 65, 18-43.*

Neural Networks 65 (2015) 18–43

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Review

Computational cognitive models of spatial memory in navigation space: A review

Tamas Madl [a,c,*], Ke Chen [a], Daniela Montaldi [b], Robert Trappl [c]

[a] School of Computer Science, University of Manchester, Manchester M13 9PL, UK
[b] School of Psychological Sciences, University of Manchester, Manchester M13 9PL, UK
[c] Austrian Research Institute for Artificial Intelligence, Vienna A-1010, Austria

ARTICLE INFO

Article history:
Received 30 May 2014
Received in revised form 15 December 2014
Accepted 12 January 2015
Available online 20 January 2015

Keywords:
Spatial memory models
Computational cognitive modeling

ABSTRACT

Spatial memory refers to the part of the memory system that encodes, stores, recognizes and recalls spatial information about the environment and the agent's orientation within it. Such information is required to be able to navigate to goal locations, and is vitally important for any embodied agent, or model thereof, for reaching goals in a spatially extended environment.

In this paper, a number of computationally implemented cognitive models of spatial memory are reviewed and compared. Three categories of models are considered: symbolic models, neural network models, and models that are part of a systems-level cognitive architecture. Representative models from each category are described and compared in a number of dimensions along which simulation models can differ (level of modeling, types of representation, structural accuracy, generality and abstraction, environment complexity), including their possible mapping to the underlying neural substrate.

Neural mappings are rarely explicated in the context of behaviorally validated models, but they could be useful to cognitive modeling research by providing a new approach for investigating a model's plausibility. Finally, suggested experimental neuroscience methods are described for verifying the biological plausibility of computational cognitive models of spatial memory, and open questions for the field of spatial memory modeling are outlined.

Contents

* Correspondence to: School of Computer Science, University of Manchester, Oxford Road, M13 9PL Manchester, UK.
E-mail address: tamas.madl@gmail.com (T. Madl).

http://dx.doi.org/10.1016/j.neunet.2015.01.002

1. Introduction

A wealth of neurophysiological results from human and animal experiments have, in recent years, helped shed light on the mechanisms and brain structures underlying spatial memory. Although it is possible to investigate spatial cognition purely from the point of view of one of the cognitive sciences, interdisciplinary analyses at the level of behavior as well as underlying neural mechanisms provide a more solid foundation and more evidence. Within the broader scope of cognitive sciences involved in investigating memory systems, such as psychology and neuroscience, computational models play a unique and important role in helping to integrate findings from different disciplines, as well as generating, defining, formalizing, and testing, and generating hypotheses, and thus helping to guide research in cognitive science.

There are multiple relevant reviews concerning the psychology of spatial cognition (Allen, 2003; Tommasi & Laeng, 2012) as well as its underlying neuroscience (Avraamides & Kelly, 2008; Burgess, 2008; Moser, Kropff, & Moser, 2008; Tommasi, Chiandetti, Pecchia, Sovrano, & Vallortigara, 2012). Although some of these reviews also mention the occasional computational model, no systematic review of computational models of spatial memory has been published in the last decade (note that Trullier, Wiener, Berthoz, & Meyer, 1997 have reviewed biologically based artificial navigation systems, and Mark, Freksa, Hirtle, Lloyd, & Tversky, 1999 published a review of models of geographical space). The main contributions of the current paper lie in providing a review of computational cognitive models of spatial memory (taking into account implemented models of cognition across disciplines, including psychology, neuroscience, and AI); providing a comparison of these models; reporting possible underlying neural correlates corresponding to parts of these models to aid comparison and verification; and finally outlining open questions relevant to this field which have not been fully addressed yet.

1.1. Spatial memory and representations

Biological agents such as mammals, as well as embodied autonomous agents, exist within spatially extended environments. Given that these environments contain objects relevant to the agent's survival, such as nutrients or other agents, they need to take the positions of these objects into account. The purpose of spatial memory is to encode, store, recognize and recall spatial information about the environment, and the objects and agents within it.

Spatial representations can be categorized based on the reference frame used. Egocentric representations represent spatial information relative to the agent's body or body parts. In contrast, allocentric representations represent spatial information relative to environmental landmarks or boundaries, independent of their relation to the agent. We will return to these types of representations, and the way they are encoded in mammalian brains, in Section 2.

In addition to navigation space – the space of potential travel – other forms of spatial representation have also been considered in the literature (e.g. representations of the positions of body parts or external representations such as maps or diagrams—Tversky, 2005).

In this review, we will focus on representations of navigation space and the space around the body, because the largest number of computational cognitive models account for them, and also because they are the most ubiquitous and generalizable representations. Whereas information concerning the space of the body strongly depends on the specific form of embodiment (such as body size and shape), and the use of external spatial representations is exclusive to humans, the types of representations and strategies required for navigation space are similar for different kinds of bodies and agents.

1.2. Relevance of computational cognitive models to spatial memory research

Computational models attempt to formally describe a part (or parts) of cognition in a simplified fashion, allowing their simulation on computers (McClelland, 2009; Sun, 2008b), and providing more detail, precision, and possibly more clarity than qualitative descriptions. In addition, computational models might facilitate the understanding and clarification of the implications of a theory or idea, in ways that would be difficult for humans without simulation on computers (McClelland, 2009). Since spatial memory is an interdisciplinary research area (drawing on at least psychology, neuroscience, and artificial intelligence), involving multiple representations and processes, it is especially important to formulate theories precisely, using a common language. Computational models can provide such a common ground.

The development of computational cognitive models also requires making a large number of design decisions, possibly leading to novel hypotheses, which can then be evaluated. This process usually constitutes an ongoing cycle of development, testing, and revision. Critically, most of this is performed on a computer and thus can be quick and efficient.

This efficiency is especially important for modeling mechanisms with representations that are not easily explicated or measured directly, such as in the case of spatial cognition. Humans cannot easily report the structure of their spatial representations and the mechanisms operating on them. There are a large number of structures and mechanisms that could partially account for spatial skills (e.g. navigation), and a time-efficient way of defining them, and investigating their implications in an automated fashion is important to facilitate the evaluation of their plausibility.

Once a theory or hypothesis has been encoded computationally, generating predictions from it is a straightforward matter of providing model parameters and input data, and running the model on a computer. This is usually more efficient than obtaining experimentally verifiable predictions from a verbal/conceptual theory. The predictions can subsequently be tested or verified using data obtained from empirical experiments with humans or animals, and comparing this data with the model predictions

(usually employing some statistical measure of model fit; Pitt, Myung, & Zhang, 2002).

Once in possession of the empirical data, both the prediction and the testing can be performed by running computer programs. Since this process is automated, it takes little human effort. This is a general advantage of computationally formulated models, but is especially useful for spatial memory models, since experiments investigating spatial cognition using the classical, iterative cycle of hypothesis formulation, prediction derivation and testing (Godfrey-Smith, 2003) usually require multiple, sometimes large environments (especially for navigation-scale spatial memory), and are thus impractical and time consuming to perform in the real world. In contrast, computational cognitive models of spatial memory can be run in a large number of different simulated environments, with different parameterizations, over a short period of time and with little effort.

1.3. Motivation for the proposed neural mappings

Since cognitive modeling is concerned with describing and explaining cognitive phenomena, they should behave the same way as humans (or animals) do. Comparison of model predictions with behavioral evidence, 'goodness of fit', is the most widespread quantitative method of evaluating, judging and comparing cognitive models (Pitt et al., 2002). In addition to fit, model complexity and generalizability can (and should) also be analyzed qualitatively. Frequently employed qualitative criteria include explanatory adequacy, interpretability, and biological plausibility or realism (Cassimatis, Bello, & Langley, 2008; Myung, Pitt, & Kim, 2005).

Despite these criteria, the space of models possibly accounting for experimental data is under-constrained. There can be multiple models of comparable complexity achieving comparable goodness of fit, and there might not be enough empirical data available for full evaluation. Furthermore, it is often difficult to compare cognitive models along qualitative dimensions. For example, there is no consensus on which models are biologically plausible (there are large differences between different approaches, ranging from spiking neural network models with parameters derived directly from electrophysiological measurements to AI-based methods described as 'biologically inspired' based on vague functional similarity). Many authors of cognitive models describe their work without establishing how parts of their model relate to the functionally similar biological implementation, making it difficult to judge the degree of correspondence to the brain.

Since cognition is implemented by the brain, cognitive modelers would do well to take into account the known neuronal mechanisms underlying the cognitive phenomena they are trying to model, even if not aiming to be highly biologically accurate. We will propose tentative neural mappings of the models reviewed in this paper for the following reasons. First, such mappings might help assess the biological realism of models claiming to be biologically plausible, based on the degree of structural and functional correspondence between models and the neural areas implementing the cognitive mechanisms they account for. Since cognition is implemented by the brain, close similarity between cognitive models and their neural counterparts is desirable (whether structural, functional, paradigmatic, or otherwise). Clarifying neuronal correspondence might also help provide an additional quantitative evaluation criterion, by facilitating possible future verification using neuronal data—such as imaging data from humans or electrophysiological data from animals.

Interestingly, such neuronal data can help in substantiating a model even if there is very little similarity between the elementary units of a model and the brain (as is the case with symbolic models, which usually employ local and amodal symbols for representations, as opposed to the distributed and grounded representations of the brain). A good example is the ACT-R cognitive architecture, which is primarily symbolic but nevertheless has been shown to be capable of not only fitting brain imaging data, but roughly predicting activation levels of brain areas (Anderson, Fincham, Qin, & Stocco, 2008; Qin, Bothell, & Anderson, 2007). This shows that it is possible even for high-level cognitive models which have little to do with biological neurons to contribute to and guide research in neuroscience; and that results in neuroscience can guide the development and parameter adjustment of such models despite their structural differences. Thus, the mapping between model components and brain areas might be interesting even for neuroscientists uninterested in pure cognitive modeling, or cognitive modelers uninterested in pure neuroscience.

Finally, relating models and their components to brain areas with known functions can facilitate their explanation, especially for readers with a background in cognitive neuroscience or psychology. Such mappings also help clarify and explicate structural differences and similarities between individual cognitive models.

2. Neural correlates of spatial representations

Since this review is targeted mainly at researchers in cognitive modeling, who might not be deeply familiar with the details of the neurophysiology of spatial memory and spatial cognition, we briefly summarize the neuroscientific literature concerning how mammalian brains represent navigation space.[1]

This section is intended to provide a basis for the neural mappings of model parts (to provide further plausibility constraints, an additional basis for comparisons, and a functional guide for model parts). Our descriptions of the neural correlates of spatial representations are biased toward describing areas known to be important and with (more or less) known functions, and are not meant to be a complete review of all brain areas related to spatial cognition. See Burgess (2008) and Moser et al. (2008) for more comprehensive reviews of spatial cognition in the brain, and Kravitz, Saleem, Baker, and Mishkin (2011) for an overview of areas associated with visuospatial processing.

2.1. Allocentric spatial memory

Four types of cells play an important role in processing allocentric spatial representations in the mammalian brain, established mostly through single-cell electrophysiological recording studies from mammals (the following list is based on Madl, Franklin, Chen, Montaldi, & Trappl, 2014)—see also Fig. 1:

1. **Grid cells** in the medial entorhinal cortex (MEC) show increased firing at multiple locations, regularly positioned in a grid across the environment consisting of equilateral triangles (Hafting, Fyhn, Molden, Moser, & Moser, 2005). Grids from neighboring cells share the same orientation, but have different and randomly distributed offsets, meaning that a small number of them can cover an entire environment. It has been suggested that grid cells play a major role in path integration (PI),[2] since their activation is updated depending on the animal's movement speed and direction (Burgess, 2008; Hafting et al., 2005; McNaughton, Battaglia, Jensen, Moser, & Moser, 2006). There is evidence to

[1] We apologize to readers who are already familiar the information in this section.

[2] Path integration refers to the integration of self-motion signals to maintain a location estimate; also called dead reckoning. A disadvantage of exclusively using path integration to estimate current location is that errors or noise accumulate upon each movement, increasing until it eventually renders the location estimate useless, unless corrected by allothetic sensory information Etienne, Maurer, and Séguinot (1996).

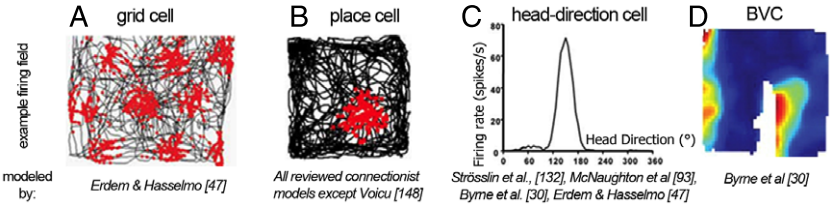


Fig. 1. Grid cells, place cells, boundary-related cells, head-direction cells, and the neuronal basis of self-motion information. A.–D.: Four cell type firing fields associated with allocentric spatial representation; as well as reviewed models accounting for them. A. Regular grid cell firing pattern from rat intracranial recording (black lines: rat trajectory, red dots: places where grid cell showed increased firing). B. Hippocampal place cell firing pattern (A and B from Burgess, 2008). C. Firing pattern of a head-direction cell tuned to about 150 allocentric direction (relative to distal landmarks or boundaries). D. Firing fields of 'boundary vector cells' identified in the rat entorhinal cortex. In specific areas of the environment (highlighted with hot colors) these cells exhibit increased firing rates (from Solstad et al., 2008). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

suggest that grid cells exist not only in mammals, but also in the human entorhinal cortex (EC) (Doeller, Barry, & Burgess, 2011). In contrast to MEC, neurons in the lateral EC exhibit little spatial modulation, and are instead highly selective to sensory stimuli.

2. **Head-direction cells** (HD cells) fire whenever the animal's head is pointing in a certain direction. The primary circuit responsible for head direction signals projects from the dorsal tegmental nucleus to the lateral mammillary nucleus, anterior thalamus and postsubiculum, terminating in the entorhinal cortex (Taube, 2007). There is evidence that head direction cells exist in the human brain within the medial parietal cortex (Baumann & Mattingley, 2010).
3. **Border cells** and **boundary vector cells** (BVCs) are cells with boundary related firing properties. The former (Lever, Burton, Jeewajee, O'Keefe, & Burgess, 2009; Solstad, Boccara, Kropff, Moser, & Moser, 2008) seem to fire in proximity to environment boundaries, whereas the firing of the latter (Barry et al., 2006; Burgess, 2008) depends on boundary proximity as well as direction relative to the mammal's head. Cells with these properties have been found in the mammalian subiculum and entorhinal cortex (Lever et al., 2009; Solstad et al., 2008), and there is also some behavioral evidence substantiating their existence in humans (Barry et al., 2006).
4. **Place cells** are pyramidal cells in the hippocampus which exhibit strongly increased firing when the animal is in specific spatial locations, largely independent from orientation in open environments (Burgess, 2008; O'Keefe & Dostrovsky, 1971), thus providing a representation of an animal's (or human's Ekstrom et al., 2003) location in the environment. A possible explanation for the formation of place cell firing fields is that they emerge from a combination of grid cell inputs on different scales (Moser et al., 2008; Solstad, Moser, & Einevoll, 2006). It has also been proposed that place fields might be mainly driven by environmental geometry, arising from a sum of boundary vector cell inputs (Barry et al., 2006; Hartley, Burgess, Lever, Cacucci, & O'Keefe, 2000); or by a combination of grid cell and boundary vector cell inputs (Madl et al., 2014). Apart from information about the current spatial location, hippocampal place cells also participate in place–object associations (Kim, Delcasso, & Lee, 2011; Manns & Eichenbaum, 2009), associating place cell representations of specific locations with the representations of specific objects in recognition memory (the perirhinal cortex, among others, is heavily involved in recognition memory for objects—Brown & Aggleton, 2001; Yonelinas, Otten, Shaw, & Rugg, 2005). In addition, in the primate hippocampus, view-dependent instead of place-dependent cells have also been identified (dubbed spatial-view cells Rolls & Xiang, 2006). Finally, an interesting cell type with spatially localized firing activity has been found in the medial prefrontal cortex (mPFC), representing **goal** or **reward** locations (Hok, Save, Lenck-Santini, & Poucet, 2005).

Hippocampal place cells seem to encode long-term allocentric spatial representations of environments (this is suggested by the spatially localized firing of place cells, the observation that this firing did not depend on heading direction and remains stable in an environment for several weeks, and finally the associations between place cells and specific objects). It has been argued that multiple such representations are learned for different environments, with different frames of reference and on different scales. Evidence for this includes the observation that place cells 're-map' when rats enter a new environment (the firing fields of the same cells reflect a completely different map in different environments), and the observation that their firing field sizes can significantly differ (Derdikman & Moser, 2010).

Allocentric representations allow not only the storage and subsequent recall of remembered routes, they also allow the calculating of novel routes, shortcuts or detours (important especially after changes in the environment, e.g. when a known route is blocked). Furthermore, it is possible to keep track of more allocentrically encoded object positions than egocentric positions—since the latter are encoded relative to the agent and thus require updates as the agent moves through the environment, making accurate egocentric representations of large numbers of objects intractable.

Such allocentric representations of physical locations in the environment have been called 'cognitive maps' – a term coined by Tolman (1948) – and there is substantial evidence that the hippocampal–entorhinal complex is the main neural correlate involved in their storage and recall (Moser et al., 2008).

Another proposed form of allocentric representation is a topological map. Topological maps lack metric information (such as distances or directions), but provide adjacency and containment information and thus allow route planning as well (although planning optimal routes can be difficult) (Booij, Terwijn, Zivkovic, & Krose, 2007). There is no well-established neural correlate of possible topological representations in the brain; although computational models with topological assumptions have successfully accounted for some hippocampal experimental data (Chen, Kloosterman, Brown, & Wilson, 2012; Dabaghian, Cohn, & Frank, 2011) (and there is some neural evidence for the involvement of posterior parietal cortex (PPC) Calton & Taube, 2009 and retrosplenial cortex (RSC) Epstein, 2008).

2.2. Egocentric spatial memory

For humans and primates, vision is the primary perceptual modality, having the largest cortical area associated with its processing. There are multiple pathways originating from the visual cortices. Apart from a pathway supporting object vision along ventral areas (the 'what' pathway), two others have been proposed which are relevant for spatial memory.

The primary visual cortex (V1) located in the occipital lobe projects visual information through higher visual cortices to the Posterior Parietal Cortex (PPC). The parieto-medial temporal

pathway connects this occipito-parietal circuit with areas in the medial temporal lobe including hippocampal, entorhinal and subicular areas involved in processing long-term allocentric spatial representations supporting spatial navigation (see above) (Kravitz et al., 2011).

On the other hand, many brain areas involved in the representation of egocentric space reside in the posterior parietal cortex. Posterior parietal areas can be said to extract object positions relative to the agent from sensory information. Patients with parietal lesions might have intact primary sensory and motor representations, but often suffer from spatial neglect—they are unable to perceive one side of space (Husain, 2008).

Evidence suggests that the **precuneus** is the main brain area concerned with multiple types of egocentric representations, as well as transformations between them (Kravitz et al., 2011; Vogeley et al., 2004; Zaehle et al., 2007). The precuneus seems to coordinate spatial processing in the reference frames of the eyes and the head with controlling body and limb-centered actions (in addition to the intraparietal and postcentral sulci and the parieto-occipital region; Plank, 2009; Vogeley et al., 2004)—for example, area 5d within this parietal area seems to represent reach vectors (hand position relative to reach target).

Neuropsychological studies have also implicated the lateral intraparietal area (LIP) in representing visual stimuli in the reference frame of the body (Snyder, Grieve, Brotchie, & Andersen, 1998), the ventral intraparietal area (VIP) containing receptive fields with head-centered reference frames Duhamel, Colby, and Goldberg (1998), the medial intraparietal area (MIP) in the encoding of object locations in eye-centric coordinates (Pesaran, Nelson, & Andersen, 2006), and area 6a (Marzocchi, Breveglieri, Galletti, & Fattori, 2008). The latter two areas have also been called the 'parietal reach region' and seem to encode the location of reach targets in an eye-centered reference frame (Bhattacharyya, Musallam, & Andersen, 2009). (See Kravitz et al., 2011 for a detailed review of visuospatial processing in the brain.)

Finally, the retrosplenial cortex (RSC) and the parahippocampal place area (PPA) in the parahippocampal cortex both seem to be involved in the visual representations of places, since they respond strongly to scenes such as landscapes or cityscapes but weakly to non-scene objects (such as animals or small objects).

Apart from visuospatial representations, the basal ganglia also play an important role in egocentric navigation, and are thought to associate a cue with a reward (Packard & McGaugh, 1996), triggering guidance behavior along a known route. The basal ganglia can thus encode the body turns/directions to take when landmarks are recognized, depending on the spatial relationship between the landmark and the body (e.g. turn left at the big tree). This encoding allows navigation based on simple associations between actions and egocentric spatial relations (also called 'taxon navigation', as opposed to 'locale navigation' which requires allocentric spatial representations). This taxon strategy seems to be in use mainly when a route is well-known (Hartley, Maguire, Spiers, & Burgess, 2003). In contrast, novel route planning requires additional allocentric representations (see previous section).

2.3. Structures involved in transformation

Since sensory information is perceived from the reference frame of the observing agent, allocentric spatial representations must be built via transformation of the sensory input. Furthermore, allocentric information has to be transferred back into an egocentric reference frame in order to allow spatial actions.

Because of its interconnections with brain areas associated with both egocentric and allocentric spatial representations, it has been suggested that the **RSC** is involved with translations of frames of reference. The RSC receives direct inputs from visual areas V2 and V4, and egocentric sensory information from parietal areas 7a and LIP, among others; as well as inputs from the hippocampal formation and the anterior thalamus usually associated with allocentric position and heading information (Vann, Aggleton, & Maguire, 2009).

Area 7a in the posterior parietal cortex is another area strongly connected to both the medial temporal areas associated with allocentric representations, and the parietal areas associated with egocentric representations. Thus, area 7a could also play a role in transforming between reference frames. For example, neurons in area 7a can transform viewer- to object-centered spatial information (Byrne, Becker, & Burgess, 2007; Crowe, Averbeck, & Chafee, 2008).

2.4. Structures involved in associative and reward-based learning

Hebb's rule is a prevalent and frequently modeled associative learning rule, which is based on the idea of activity-dependent synaptic modification, and proposing that a change in the strength of a connection is a function of the neural activities of the connected neurons. Hebbian learning is often summarized as 'neurons that fire together, wire together'. There is strong empirical evidence for such a learning mechanism ubiquitously occurring in brains (Song, Miller, & Abbott, 2000). This learning rule is critical for associative learning in spatial memory paradigms—for example, for learning associations between the representation of a rat's current location, and sensory stimuli at that location. In a variant of Hebbian learning, called competitive learning, neurons of one population compete with each other to respond to the pattern appearing in another population from which they receive input (the more strongly a neuron responds to the input, the more it inhibits other neurons, and the more its connection strengths to highly active input neurons increase) (Grossberg, 1987; Kaski & Kohonen, 1994; Rumelhart & Zipser, 1985).

As opposed to the unsupervised, associative Hebbian learning rule, reward-based learning is also frequently observed in spatial memory experiments. As mentioned above, the mPFC seems to be involved in representing goal or reward locations (Hok et al., 2005) (and has also been suggested to be involved in responding to rewards). Animals including humans have a propensity to seek out rewards, and are able to learn the spatial locations of such rewards. The primary neural correlates of reward-learning include the orbitofrontal cortex (OFC, which seems to encode stimulus reward value), the amygdala, and the ventral striatum; all three show increased activity during the expectation of a reward. The dopamine system also plays an important role, being involved in the signaling of error in the prediction of reward (presumably aiding learning and facilitating the improvement of reward predictions). To select an action based on an expected reward, stimulus–response or response–reward associations have to be learned; empirical evidence implicates the dorsal striatum in this process (which exhibits increased activity when a contingency is established between responses and reward) (Maia, 2009; O'Doherty, 2004).

Reinforcement learning theory has been used in attempts to mathematically formalize the process of reward-based learning through interacting with an environment. Reinforcement learning (RL) agents represent the world as a set of states S, a set of actions A possible in each state and leading to a new state, and possible rewards r. They learn from the consequences of their actions, and try to select actions based on past experiences (exploitation) as well as novel choices (exploration). The name comes from the reinforcement signal – a numerical reward – used in such models; RL agents aim to choose actions that maximize the reward they obtain over time (Woergoetter & Porr, 2008). It has been argued that mathematically derived solutions to RL can plausibly be implemented in brains, based on the reward-relevant brain areas

listed in the previous section (explaining RL and its correspondence to brains would exceed the scope of this paper; see Maia, 2009 for an explanation and review of evidence). Reinforcement learning can be used to learn which action to take in each location, e.g. to learn how to navigate to a food source (see also Fig. 4 and some of the models reviewed below).

3. Computational cognitive models of spatial memory

3.1. Introduction

Computational models attempt to formally describe an aspect (or aspects) of cognition in a simplified fashion, allowing their simulation on computers (McClelland, 2009; Sun, 2008b). Computational cognitive modeling is concerned with achieving a better understanding of various cognitive functionalities through computational models of representations, mechanisms, and processes.

Cognitive models should be functional—they should perform well at the task they were designed for (which can be difficult, especially for challenging tasks such as trying to robustly map real-world environments).

Psychological or cognitive plausibility are also important—these models aim to model cognitive phenomena (spatial memory and associated processes), and should correspond to them as closely as possible in terms of the mechanisms, processes, and representations employed, and behavioral measures produced. They should account for empirical data as well as possible (high 'goodness of fit'), and should do so in the simplest possible way (low complexity), making as few unsubstantiated assumptions as possible. They should also have the ability to generalize to new data, not only account for the data provided to the model during development and training (Myung et al., 2005).

Cognitive models should also be as biologically plausible as possible within their paradigm. Although many cognitive models are not concerned with the physiological details of neural functioning (with the exception of biological/spiking neural networks—see Section 3.4), the underlying neuroscience of the modeled cognitive phenomena is nevertheless arguably relevant. Functions of the mind are implemented in brains; thus, neuroscience can provide valuable input regarding the structure and function of plausible models, even for those not intending to model the neuron level. Further advantages of taking neural implementation into consideration include constraining the model space (reducing the large number of algorithms possibly accounting for given behavior data), providing additional evaluation criteria, and facilitating model comparison by establishing analogies between representations in models and in brains (see also Section 1.3).

Clarification of the elemental units used by models, and how they relate to neural substrate, is critical in evaluating biological plausibility. The correspondence does not need to be on the neuron level—symbolic cognitive models can also structurally correspond to brains on a higher level (e.g. on the level of brain areas and their connectivity). Explicating the correspondence between model components and brain areas, as done by the researchers of ACT-R (Anderson et al., 2008) (who have also performed brain imaging experiments for validation), helps to verify structural similarity between the model and the corresponding neural substrate, and thus also to evaluate claims of biological plausibility. Describing such neural mappings is one of the aims of this section, as well as establishing tentative mappings based on functional correspondence in cases where the authors did not explicitly describe them in their work, as is the case for the majority of models outside of computational neuroscience.

Clarification of the following properties is also important in characterizing computational models of spatial memory (partially based on O'Reilly, 1998 and Webb, 2001[3]).

- The level of modeling (characterizing the elemental units),
- The types of representation accounted for (e.g. egocentric, allocentric, metric, topological)
- The learning mechanism, if any (e.g. Hebbian learning, reinforcement learning)
- The generality and abstraction of the models (the range of phenomena accounted for, and complexity relative to the modeled phenomena)
- Structural similarity (how well models represent the underlying neural mechanisms)
- Performance match or 'goodness of fit' to behavior data (to what extent the model can match target behavior; useful for comparing different models of the same phenomena).

It is important to note that this review is limited to computational models of cognition concerned with navigation space, that were published in the last two decades,[4] and as such excludes models of diagrammatic spatial reasoning, models of low-level sensory representations, robotic models unconcerned with biological cognition, and other models which might include spatial information on a different scale or for a different purpose. Furthermore, we exclude reactive navigation models without representations, which might allow agents to solve problems in space, but cannot be said to model spatial memory.

Finally, we do not claim to review every single model involving spatial memory (such an endeavor could fill a book); the aim of this review is to summarize representative models for major modeling directions (of any set of models which are highly similar in terms of paradigm, structure and functionality, only the most recent one is reviewed; similarly, if the same first author publishes multiple times on a model, only the most recent version of the model is included).

3.2. Overview

The spatial memory models reviewed in this section are divided into three categories, inspired by major modeling paradigms in the field of computational psychology (Sun, 2008a). The section 'symbolic spatial memory models' describes models emphasizing explicit rules and localist representations based on symbolic logic (Bringsjord, 2008). In contrast, 'neural network-based spatial memory models' are based on a number of simple processing units affecting each other via weighted connections, operate in parallel, usually employ distributed representations, and commonly learn rules from training data instead of encoding explicit rules (Thomas & McClelland, 2008). Finally, we also review a number of spatial memory models that are a part of cognitive architectures (which are concerned with modeling a wide range of cognitive phenomena in addition to spatial memory, and are often employing a combination of the mentioned paradigms).

We have confined our survey to these relevant categories and model types to keep it within the limited space available.

Each of these types of models have different strengths and roles in modeling and understanding spatial memory. Symbolic models

[3] Criteria specific to neuroscience and unimportant for characterizing purely cognitive models have been excluded.

[4] We used the academic search engines Scopus, JSTOR, Google Scholar, Microsoft Academic, and arXiv; searching for keywords (and their combinations) relevant to this review, including *computational, cognitive, spatial, models, spatial memory, cognitive map, hippocampus, place cells, egocentric representations, allocentric representations, navigation, orientation, localization, mapping, SLAM, symbolic, connectionist, cognitive architectures*. Furthermore, we manually searched the Comparative Repository of Cognitive Architectures (by the BICA society) for relevant models.

operate on a high-level of abstraction (they are usually not concerned with neuron level phenomena), and are often functionally more powerful than neural network models (they can often perform more complex tasks). They usually have less structural similarity to brains, and are thus less constrained (even if validated against behavior data, it is difficult to evaluate multiple symbolic models performing a similar task with comparable goodness-of-fit). In contrast, neural network models are often more similar to the neurophysiological implementation (both in terms of representation and mechanism) and are thus easier to constrain by established neuroscientific knowledge and by additional types of data (such as neural recordings or brain imaging). However, this paradigm often makes it difficult to implement complex cognitive processes, especially those requiring serial processing steps (for example, none of the neural network models are able to perform spatial reasoning or loop closure, in contrast to some symbolic approaches). Finally, cognitive architectures can follow either or both of these paradigms, and have the additional advantage of incorporating multiple cognitive mechanisms—thus, they can perform, and be evaluated against, different tasks and datasets.

Apart from categorizing the models based on their underlying modeling paradigm, we will also group them into models evaluated in simplified, simulated environments, and into models which are capable of dealing with – and being evaluated in – real world environments (such as robotic implementations). In general, robotics emphasizes high-performance solutions to low-level 'sensor problems' (e.g. dealing with sensory uncertainty/noise or processing or recognizing complex sensory data), and aims for high performance (accuracy, efficiency, etc.) instead of cognitive plausibility (Jefferies & Yeap, 2008). However, as Gallistel (2008) points out, the nature of the computational problems of navigation and map making based on limited information does not depend on whether one is studying biological or artificial systems. Thus, the latter could help in understanding the former.

Robots and animals must perform similar computations when trying to make sense of space. Computational models of cognition operating in similar environments to the modeled biological agent, and dealing with similar difficulties posed by the real world (such as complexity, limited knowledge, uncertainty, or noise), can be regarded as being more plausible than models not accounting for such difficulties (Webb, 2000). This is the main motivation for dedicating subsections to cognitive models evaluated in the real-world (but excluding systems concerned with practical robot performance rather than investigating cognition).

The following list presents an overview of the models reviewed below. Models embodied on robots capable of running in the real world are printed in bold, and, for clarity, the first mention of a model in each subsection below is underlined. A comparative table of all reviewed models, with additional properties for comparison, can be found at the end of this section (Table 1).

- Symbolic models (Section 3.3)
  - Allocentric models
    - (Yeap, Wong, & Schmidt, 2008)
    - (Jefferies, Baker, & Weng, 2008)
    - perceptual wayfinding model (Raubal, 2001)
  - Egocentric models
    - NAVIGATOR (Gopal & Smith, 1990)
  - Allocentric + egocentric
    - **HSSH** (Beeson, Modayil, & Kuipers, 2010)
    - (Franz, Stürzl, Hübner, & Mallot, 2008)
    - DP-model (Brom, Vyhnánek, Lukavský, Waller, & Kadlec, 2012)
- Neural network-based models (Section 3.4)
  - Allocentric models
    - (Burgess, Jackson, Hartley, & O'Keefe, 2000)
      - (later extended in simulation as the BVC model by Barry et al., 2006)
    - (Strösslin, Sheynikhovich, Chavarriaga, & Gerstner, 2005)
    - (Barrera, Cáceres, Weitzenfeld, & Ramirez-Amaya, 2011)
    - (Schölkopf & Mallot, 1995)
    - (Voicu, 2003)
    - (McNaughton et al., 1996)
    - (Erdem & Hasselmo, 2012)
  - Allocentric + egocentric
    - (Byrne et al., 2007)
- Cognitive Architectures (Section 3.5)
  - Allocentric models
    - LIDA (Madl, Franklin, Chen, & Trappl, 2013)
  - Egocentric models
    - ACT-R/S (Harrison et al., 2003)
    - CLARION (Sun & Zhang, 2004)
  - Allocentric + egocentric
    - Casimir (Schultheis & Barkowsky, 2011)

3.3. Symbolic spatial memory models

Symbolic models of spatial memory are concerned with explicitly representing spatial knowledge in a declarative form as facts and rules. They are based on the assumption that cognition consists of discrete mental states (representations), which can be modeled as localist symbols (in contrast, in neural network-based models the representations are not discrete, but constitute distributed and potentially overlapping patterns of activation—see next section). A number of processes operate on these representations, creating, modifying, or deleting them (Smolensky, 1987). One of the earliest definitions of such symbolic models has been put forth by Newell and Simon (1976), coining the term of a 'physical symbol system', a class of systems having symbols, being capable of manipulating them, and being realizable within our physical universe.

Symbolic models are often based on cognitive science theories (most frequently information processing models), and thus are able to claim a degree of cognitive plausibility. There is usually very little similarity between the elementary representations of symbolic models and biological neurons (mainly because of the choice of localist and amodal representations, in contrast to the distributed and more modal representations of the brain; Barsalou, 2008; Martin & Chao, 2001). However, they can still correspond to the brain on a higher level (e.g. functional correspondence to brain areas, as established for ACT-R). Despite the structural and paradigmatic difference, and for reasons mentioned in the Introduction, brain areas corresponding to model parts will be pointed out based on such functional correspondences where applicable.

3.3.1. Models evaluated in real-world environments

A few cognitive models of spatial memory have been implemented in robotic systems capable of navigating in the real world. Jefferies and Yeap (2008) provides a survey of such cognitive mapping approaches that have been designed to work on robots. Usually, robotic implementations following the symbolic[5] approach build metric representations of the local environment using an approach called SLAM (Simultaneous Localization and Mapping). Recent SLAM approaches are capable of recognizing a place the robot has seen before (this is called 'loop closing'), and correcting errors in the map representation by exploiting and correcting for the difference between expected and observed location on the map.

[5] A notable exception is RatSLAM (Milford & Wyeth, 2010), a model based on attractor neural networks (which however is not a cognitive model, and is not intended to model behavior or biology).

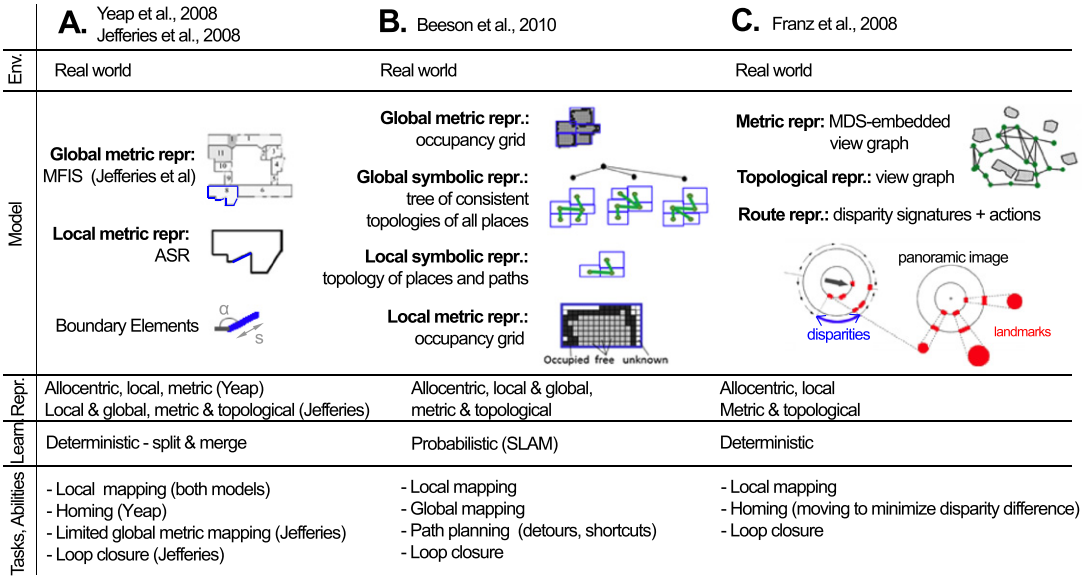


Fig. 2. Overview of symbolic models evaluated in real-world environments. A: (Jefferies et al., 2008; Yeap et al., 2008); both models create local metric maps (absolute space representations – ASRs – consisting of boundary elements); the latter model also builds a global metric map (Memory for Immediate Surroundings—MFIS) with which it can perform loop closing. B: HSSH (Beeson et al., 2010). C: (Franz et al., 2008). *Deterministic learning* is a collective term for all mechanisms that learn by adding new symbolic representations to memory upon perceiving a new object (as opposed to probabilistic or neural network learning mechanisms). *Local maps* represent spaces appearing to enclose the agent (such as a room). *Global maps* can represent and align multiple local maps in the same reference frame.

SLAM is usually implemented by a probabilistic state estimation method, integrating self-motion information and landmark observations in a statistically optimal fashion (Jefferies & Yeap, 2008; Thrun & Leonard, 2008).

The core ideas of SLAM – using probabilistic inference to deal with uncertainty and noise and to infer near-optimal estimates of the locations of the agent, and objects in its environment – do not contradict the cognitive sciences. They fit in well with the recent 'Bayesian brain' hypothesis (Knill & Pouget, 2004); the idea that the brain integrates information in a statistically optimal fashion. There is evidence that spatial cues might be integrated statistically optimally in humans (Nardini, Jones, Bedford, & Braddick, 2008) and animals (Cheng, Shettleworth, Huttenlocher, & Rieser, 2007) on the behavioral level. Computational models resembling SLAM – using probabilistic state estimation – have been proposed to explain spatial orientation and cognitive mapping (Cheung, Ball, Milford, Wyeth, & Wiles, 2012; Fox & Prescott, 2010). It has also been suggested that hippocampal place cells might be able to perform approximate Bayesian inference on the neuronal level, based on electrophysiological recording evidence (Madl et al., 2014).

However, the representation implementation is highly important for judging the plausibility of such probabilistic models (e.g. in terms of their structural accuracy, and levels of abstraction and modeling). In SLAM approaches in robotics, maps are stored in different ways, most commonly as covariance matrices, or as occupancy grids (two-dimensional matrices with entries storing the probability of occupancy), or tree-based representations (Thrun & Leonard, 2008). In the absence of psychological or neuroscientific data substantiating the existence of explicit covariance representations in human or animal cognition, and of biologically realistic implementations, it would be difficult to argue for the plausibility of covariance matrices as cognitive models of spatial memory. Here we only include models where authors explicitly address the relationship of their models to cognitive science or neurobiology (unfortunately, although citing empirical evidence, few of these authors evaluate their models against empirical data from humans or animals). For reviews of robotic SLAM, see e.g. Bailey and Durrant-Whyte (2006), Durrant-Whyte and Bailey (2006) and Thrun and Leonard (2008).

Building on work by Yeap (1988) – one of the first symbolic computational models of cognitive maps – a number of robotic systems have been built (many of which have departed from the original claim of being computational theories of cognitive maps and will therefore be omitted). Yeap suggests the computation of abstract allocentric maps of a region (from the shape and disposition of surfaces/boundaries relative to the agent) which the author calls 'absolute space representation' or ASR (see Fig. 2(A)). Multiple ASRs can be interconnected as a traversable network to form a cognitive map of the entire environment, and afford the notion of 'places'; a network of ASRs can model a network of places, with exits leading from one to the other, such as rooms in a building. The elemental representation in ASRs is a list of triplets, each representing a boundary element (BE), and containing its size, angle to the next adjacent BE, and whether it is empty space, not empty, or occluded.

- Based on this model, Yeap et al. (2008) developed a robotic system capable of building allocentric maps. The robot uses a simple exploration strategy (move forward in a straight line, stop when encountering an obstacle, turn away from the obstacle but maintain forward direction), after which it has to find its way 'home' (back to its starting location).

It used 8 simple sonar sensors to measure distances to obstacles and boundaries, and built a metric map based on both the robot path, and linear surfaces around it approximated from sonar data. This map was subsequently split, or merged, into distinct regions (e.g. corridors and rooms) using features such as average width (e.g. corridors are long and narrow), and employing the split and merge algorithm (Pavlidis & Horowitz, 1974) to find continuous regions. Each continuous motion segment of the robot (without stopping or turning) was represented as an ASR (consisting of multiple boundary elements from sonar data). The robot was able to use the final network of ASRs it has built using the split and merge algorithm to find its way back 'home' by backtracking the distances traveled.

The robot could localize itself using 'confidence maps' computed from the similarity between the currently perceived region or ASR (current sonar readings), and all stored ASRs. The authors reported that the localization was accurate with respect to the occupied region (i.e. the error was smaller than the size the regions).

The robot could also robustly estimate a homing vector and return to its starting position even when the ASRs computed during the outward and inward journey were inconsistent—not requiring correct and consistent metric representations for homing is the main strength of the model. However, it could not match re-observed boundaries with those in its memory, and thus it is unable to 'close the loop'.

The model does not make any claims of structural accuracy with regard to its neurobiological equivalent. Based on functional similarity, ASR regions contain some of the information represented by place cells ('confidence maps' on ASR regions, similarly to place cells, carry information regarding the currently occupied space), goal cells (ASRs regions, like goal cells, can constitute goal representations) and by boundary vector cells (boundary elements carry boundary size and angle information).

• Also drawing on the ideas of Jefferies et al. (2008) and Yeap (1988) proposed that a cognitive map might consist of a topological global map containing metric local space representations, aiming to benefit from the advantages of both—simple localization and metric consistency of the local maps, and easier 'loop closing' with the help of global maps (as well as the confinement of location errors to the local maps). The idea of separate local and global representations is consistent with most empirical cognitive science research (Hirtle & Jonides, 1985; Poucet, 1993; although there is some debate regarding whether and which mechanisms/areas are metric or topological). In contrast to the previously described model, the robot by Jefferies et al. (2008) used laser rangefinders as sensors (which provide more accuracy and resolution than simple sonar sensors).

Their approach turns the raw laser data (distance measurements) into lines representing boundaries, finds the exits (gaps in the boundary), and then computes ASRs based on this information. Different ASRs representing local regions can subsequently be connected topologically via the identified exits to form a global map.

Finally, with the help of this topological map, they also build a global map of limited extent containing the last few local spaces visited (called 'Memory for the Immediate Surroundings', MFIS), providing easier recognition that the robot has re-entered a previously observed part of the environment (loop closing). This model is one of only two reviewed models capable of building a global map and of loop closing (the other being Beeson et al., 2010—see below).

The authors argue for the psychological plausibility of their model using the empirical evidence for local and global spatial maps (Poucet, 1993) and multiple reference frames for different parts of an environment (Derdikman & Moser, 2010; McNaughton et al., 2006).

The model does not aim for structural resemblance to the brain. As it is also based on Yeap (1988), tentative arguments of functional similarity can be made between ASR regions and place cells, and boundary elements and boundary vector cells. No equivalent of a consistent, metric, global map has been found in the brain (the same place cells participate in representing very different locations in different environments; there is no one-to-one mapping as in the MFIS).

• Beeson et al. (2010) also propose a spatial memory model combining the strengths of topological and metrical approaches, calling it *HSSH* (Hybrid Spatial Semantic Hierarchy), an extension of the SSH model proposed by Kuipers (2000). The HSSH has four major levels of representation: a local metrical level (in which the agent builds a metric Local Perceptual Map—LPM), a local topological level (in which the agent identifies discrete places in a large-scale environment and describes paths in it), a global topological level (for resolving structural ambiguities and determining how the environment is best described as a graph of places, paths and regions), and a global metrical level (describing the environment in a single metric global map using the same reference frame).

On the first level, the LPM is built using probabilistic SLAM (Thrun & Leonard, 2008) based on laser rangefinders, and represented as an occupancy grid (a discretized grid in which each cell contains the probability of being occupied by an obstacle). On the local topological level, a discrete set of 'places' and 'path segments' connecting them are identified (using an approach based on Voronoi graphs and recognized gateways/doors). The global topological map is built by creating a tree of all possible topological maps (map hypotheses) consistent with current experience. After each travel action, every map hypothesis is extended; if it leads to a predicted transition to a known state, the hypothesis can be updated or refuted based on the subsequent observation. This allows 'closing the loop' and pruning the tree of topological maps when places are revisited.

Finally, on the global metrical level, a metric map of the entire environment in a global reference frame can be assembled on the structural skeleton provided by the global topological map (and based on the known robot trajectory and the displacements between places to appropriately translate local frames of reference). HSSH is the only model except for Jefferies et al. (2008) capable of building a global map and closing the loop.

Although not aiming for structural similarity to the brain, the HSSH and its predecessors claim to be 'theories of robot and human commonsense knowledge of large-scale space: the cognitive map' (Kuipers, 2008). Unfortunately, no comparisons of the model's performance with human data have been performed. In terms of functional similarity, the occupancy grid employed as the low-level metric representation bears some resemblance to hippocampal place cells, as both can be used to infer the most likely location of the agent, as well as the most likely locations of boundaries (Barry et al., 2006). However, there are also significant differences, including the resolution (1 cm in some SLAM approaches, as opposed to the sizes of place cell firing fields,[6] which range from 20 cm or less to multiple meters, Kjelstrup et al., 2008; O'Keefe & Burgess, 1996), constancy (place fields can be destroyed or changed by adding barriers or making other changes in the environment), shape properties (occupancy grid cells are square, place cells can have multiple firing fields of different round shapes), representation (occupancy grid cells contain probabilities, place cell firing rates almost certainly do not, since they strongly depend on factors such as running speed), among others (Moser et al., 2008). Independently of the differences in the representation employed, probabilistic inference (the mechanism which SLAM is based on) has been argued to be plausible based on empirical data (see above).

• Franz et al. (2008) have developed a robotic system on a Khepera miniature robot that accounts for egocentric route navigation, as well as allocentric topological navigation and global metric navigation (with the first two working on the robot and the latter implemented in a simulation), building on their earlier work (Franz & Mallot, 2000).

Route navigation (or taxon navigation) works by storing simple associations of actions to egocentric spatial relations. Several such associations can be concatenated to routes that might lead from the current location to a goal location (see Section 2). Franz et al. (2008) use a panoramic stereo camera to calculate the disparities of $N = 72$ image sectors, after identifying each sector in both images (disparities are defined as how much an image sector

[6] Despite these sizes of individual place fields, it is possible to decode the animal's position more accurately using the cumulative activity of multiple overlapping cells and statistical methods (up to an error of about 8 cm based on the spike train alone, Brown, Frank, Tang, Quirk, & Wilson, 1998, or about 3 cm based on theta phase coding, Jensen & Lisman, 2000).

appears shifted in the second image relative to the first; from these disparities, distances can be computed using elementary trigonometry). They represent a place using a 'disparity signature', a list of disparities and their corresponding reliabilities. Storing such place representations allows a simple homing by using a strategy of calculating the disparity signatures for several possible movements, and then choosing the movement that minimizes the difference between the current and the goal disparity signature. Sequences of distinguishable disparity signatures can constitute a route and allow taxon navigation.

Topological navigation integrates routes leading through the same place to a representation that can be used for navigating to multiple goals. In this model, topological navigation is afforded by a 'view graph', which is built by measuring similarities between views (using maximal pixel-wise cross-correlation), and connecting two routes whenever two views are sufficiently similar and whenever the robot succeeds in homing to this similar view. This system could successfully explore an environment, and perform homing and shortcut planning in the real world. However, it requires views to be unique (since it connect routes when views match)—thus, it cannot close the loop in environments with non-unique views.

The model was also extended by an approach to survey navigation in a simulation. This requires a representation in a common frame of reference. The model attempts to construct a global metric map by metrically embedding the view graph using an approach based on multi-dimensional scaling (MDS).

Franz et al. (2008) argue that their navigation strategies are 'biomimetic'; citing behavioral evidence from studies with insects, which lend strong support to the claim that insects seem to use mainly view-based homing for navigation (Graham & Collett, 2002)—a strategy resembling the 'disparity minimization' approach of the authors. However, they do not claim any structural similarity to brains, whether mammalian or insectile. The taxon navigation strategy can be implemented in principle by the basal ganglia storing routes as stimulus–reaction mappings, in combination with neurons encoding views, such as spatial-view cells or PPA neurons (see Section 2). However, even on the functional level, this similarity is highly tentative, since mammals can robustly navigate to goals even in dynamic environments, or after changes in the environment (which would interfere with the simple correlation-based similarity measure of this model), and also because it is highly unlikely that mammals recognize views based purely on disparities (for example, scene recognition works almost as well on computer screens as in real scenes, suggesting that stereo vision does not play a major role).

3.3.2. Models evaluated in simulations

Computational experiments in simulations have to deal with fewer issues such as complexity, sensory inaccuracy, or noise. Thus, their developers often have the resources to endow them with a larger range of abilities and to account for more tasks and paradigms (at the expense of less similarity to the actual environment of the modeled biological cognition).

• One of the first symbolic models of spatial memory in urban environments, growing out of the symbolic AI paradigm of the last century, was the *NAVIGATOR* model by Gopal, Klatzky, and Smith (1989), implemented in LISP. It runs in a simple environment consisting of horizontal and vertical streets, as well as 'plots'—locations and associated sets of objects (such as houses)—and associated decision points—points at which navigational decisions can be made. The environment is represented in a predicate calculus-based language. The agent (called NS, navigating system) can perceive information from the plot associated with its location, as well as other plots visible in each feasible direction of view; and can either turn in the four modeled directions (to perceive in that direction) or move in one of those directions.

Upon receiving an input, the NS selects the most salient objects, and stores them in a working memory (WM). WM has the function of processing perceived information, transferring it to long-term memory (LTM), monitoring instructions, and planning paths to a goal through pattern matching. LTM in turn permanently stores conceptual, spatial, and goal knowledge; in the form of semantic network representations (e.g. decision-point 2 *associated-with* house, house *color-of* red) which can decay ('forgetting'). These representations are connected by 'links' which represent spatial relations, and can be learned either from perceptual input (when two locations are present in WM at the same time), or from explicit instructions connecting two locations (e.g. 'go from A to B').

Based on its semantic network representation, NAVIGATOR is able to find goal locations and plan novel paths. The agent runs in a very simple (simulated, discrete and static) environment. The model is claimed to qualitatively replicate several aspects of human spatial behavior, such as way-finding errors (three types of errors made by NAVIGATOR appear similar to humans—errors made at locations with more information, at locations requiring complex navigational actions, and errors due to misidentification of the goal) (Gopal & Smith, 1990). No quantitative comparison against human data is performed.

The NAVIGATOR model is based on information processing theories of cognitive psychology and thus can claim a degree of cognitive plausibility. No structural similarity to neural architecture is claimed. The spatial parts of the semantic networks constituting the representations in NAVIGATOR bear some resemblance to hippocampal representations (plots and decision points to place cells), but are too simplistic for an actual functional correspondence (e.g. not every point of the environment is represented, and the distance metric is city-block, not Euclidean).

• Raubal (2001) describe the *perceptual wayfinding model*, a cognitively based computational model for wayfinding which, unlike NAVIGATOR, considers the information needs of navigators at each decision point. The model is based on the 'Sense-Plan-Act' framework, as well as affordance theory (affordances are possibilities for action)—the idea that *animals perceive the environment in terms of what they can do with it and in it* (Gibson, 1986). It is a goal-based agent—given a current state description, goal information, and the results of possible actions, it chooses actions to achieve a goal.

Its main components are its observation schema (containing spatial and temporal location, goal, and measuring limitations, in fixed frame-like structures), a wayfinding strategy (decision rules for wayfinding), and 'commonsense knowledge' (including procedural knowledge—how to move in a direction, what to do upon reading specific symbols such as arrows on signs). The implemented agent runs in a very simple simulated environment which is static and discrete, having a limited number of possible percepts and actions at each point. The agent can observe the entire environment at any given time (unlike NAVIGATOR, which also used static and discrete environments but accounted for partial observability). The environment is represented as a graph of decision points, where each node has a position and a state, and each edge represents a transition between positions and states. Since the evaluation scenario is set in an airport, each position has information regarding how to reach goals (signs containing arrows to gates). The agent first perceives the environment (senses), then decides which action leads it toward its goal (a trivial decision given the signs at each node), and then carries out the action (acts).

The perceptual wayfinding model is evaluated in an airport wayfinding task (successfully finding gates), but not compared against any human or animal data. Because of the amount of information pre-programmed into the implemented agent, and because of the fully observable static environment, the agent needs

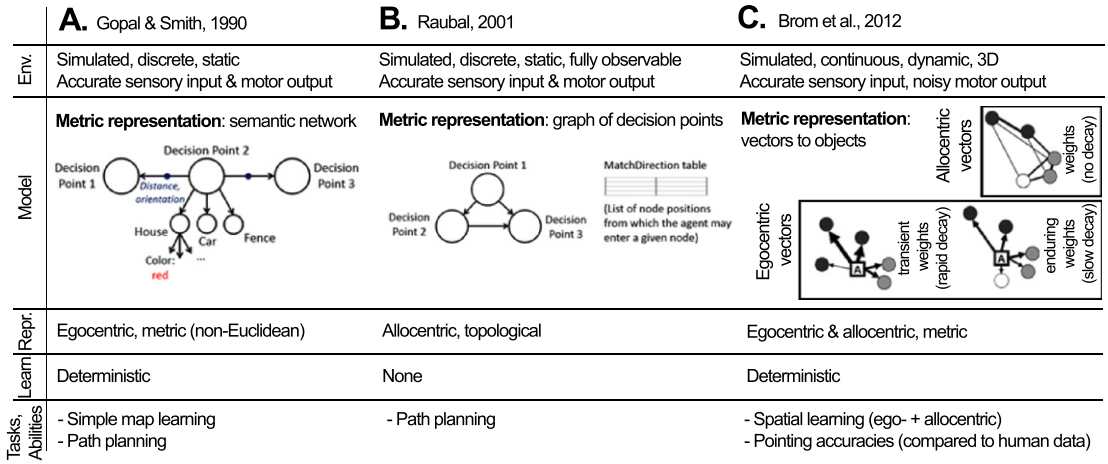


Fig. 3. Overview of symbolic models evaluated in simulated environments.

and has no learning mechanism. There is little functional similarity of the components of this model to the brain.

• Due to more recent improvements in computer graphics, it has become possible to simulate virtual agents in more complex, three-dimensional environments. In a recent model, Brom et al. (2012) have proposed a computational model of both egocentric and allocentric spatial memory for intelligent virtual agents (IVAs), calling their spatial model the *DP-model* since it was evaluated in a disorientation paradigm (see below). IVAs can be considered to be embodied, although in a much simpler and more predictable environment than the real world.

The information flow in the DP-model is as follows: sensory systems assemble information in the 'perception field', based on which egocentric representations (spatial vectors to objects in the agent's own reference frame) are built in the 'egocentric subsystem', which has both an STM and LTM component. Egocentric representations can consolidate into the LTM component of the egocentric subsystem, as well as allocentric representations in the long-term 'allocentric subsystem'. Both egocentric and allocentric representations are weighted, and weights serve as a representation of accuracy—how well a representation was learned (they are required to model errors, since the vectors are represented precisely, without modeling sensory inaccuracies or noise). The agent's perception field contains all objects in the agent's visual field (which is 120 wide). Eye movements, foveation, attention, and visual recognition are not modeled; objects are represented as state-less and static symbols. The egocentric component contains the agent's current heading (with respect to the south–north axis), a set of weighted egocentric vectors from the agent to objects, and the egocentric updating configuration (containing the rates of increasing or decreasing the weights of egocentric vectors). The allocentric component contains a set of weighted allocentric vectors between all objects, and an allocentric updating configuration (specifying the speed of increasing weights of the allocentric vectors). Egocentric vector weights are increased at every time step if the associated object is still part of the perceptual field, and decreased if it is not. The vectors themselves are updated whenever the agent moves to point correctly from the agent's position to the associated object. Allocentric vectors are learned from egocentric vectors.

The agent is also endowed with an action selection mechanism, enabling it to follow a specified trajectory to learn a representation of space during the learning phase, as well as to perform pointing tasks. In these tasks, the agent first observes and learns a number of object locations, and subsequently has to point to these locations after the objects have been removed. The pointing error in this task is a function of the vector weights (themselves depending on how often and how long the associated object has been seen during the learning phase). An advantage of this model compared to the previous two models is that it runs in a more complex (continuous, dynamic, three-dimensional) simulated environment.

Brom et al. (2012) successfully replicate human data from two pointing paradigm experiments previously performed using their model, experiment 7 of Holmes and Sholl (2005), in which subjects learned the locations of objects in a room and then had to point to the remembered locations of the objects with their eyes closed after a 45 rotation left or right (both in an oriented and in a disoriented condition induced by slow rotation on a swiveling chair), and experiment 1 of Waller and Hodgson (2006), a similar pointing paradigm.

This model builds on theories from cognitive psychology and produces error patterns consistent with humans in pointing paradigms, but does not claim structural similarity to brains. Some tentative functional correspondence between egocentric vectors and representations in the parietal reach region (and other correlates of egocentric spatial memory) might be identified, since they encode the positions of targets in an egocentric reference frame.

3.4. Neural network-based spatial memory models

Unlike symbolic systems, neural network models usually employ non-local and distributed representations (also called subsymbolic representations), within interconnected networks of simple units. NNs are simplified models of the brain composed of a number of units (analogs of neurons) with weighted connections between them. Mental states are represented as numeric activation values of the units (or subsets of the units), and learning is usually implemented by modifying connection strengths between the units (Thomas & McClelland, 2008).

There is a variety of flavors and implementations of neural networks, ranging from the simplest perceptrons (which sum up a number of inputs multiplied by incoming weights and threshold the result to yield a binary output) over the commonly used feed-forward artificial neural networks, networks of perceptrons without cycles such as feed-forward ANNs and self-organizing maps,[7]

[7] A self-organizing map (SOM) is a typically two-dimensional neural network learning a discretized representation ('map') of its N-dimensional inputs. Unlike other ANNs, they preserve the topology of the input space. Each unit stores an N-dimensional weight vector. During a set number of training iterations, for each input, the nodes with weight vectors closest to the input (smallest Euclidean distance) are 'pulled closer ' to the input (weight vectors are updated to be more similar to the input)—see Kohonen (1990), or Willshaw and Von Der Malsburg (1976) for a similar, more biologically plausible model.

and recurrent neural networks allowing feed-back connections and cycles (such as attractor networks[8]), over neural networks aiming to make only biologically plausible assumptions (BNNs, 'biological NNs'), to spiking neural networks (SNNs, which are the most biologically realistic, and are the most computationally expensive to run; Jain, Mao, & Mohiuddin, 1996).

Of these, only the latter two (BNNs and SNNs) explicitly aim to be biologically realistic, with this claim being extensively verified only for SNNs (they are able to account for electrophysiological recording data from biological brains). In addition to modeling neuronal and synaptic state, they also model temporal dynamics, and use short and sudden increases in voltage ('spikes') to transmit information (Ghosh-Dastidar & Adeli, 2009). BNNs, although not directly modeling electrophysiology, also aim to be biologically realistic in terms of brain connectivity and their learning mechanisms. We shall collectively refer to all other types of neural networks (the ones not aiming to closely model biological neurons) as artificial neural networks (ANNs). ANNs, unlike BNNs and SNNs, are usually driven by mathematical reasoning instead of biological accuracy.

Because of the biological inspiration and the clear analogy between units of neural networks and neurons in brains, neural networks have been claimed to be more biologically plausible than symbolic models. This is verifiably true for many SNNs (spike trains, firing rates, membrane potentials etc. can be compared with biological neurons). For ANNs, the claim of biological realism can be cast in doubt, since they make undefended design decisions (e.g. elements not having clear biological counterparts such as fixed biases, nonmonotonic activation functions, or the commonly used back-propagation learning algorithms) (Dawson & Shamanski, 1994).

Still, even if their degree of realism is debatable, ANNs are structurally more similar to brains than symbolic cognitive models—the representations employed by both are mostly distributed, grounded and modal Barsalou (2008). Furthermore, on a higher level, neural network-based models incorporate properties characteristic of biological cognition, such as content-addressable memory, context-sensitive processing, and graceful degradation under damage or noise Thomas and McClelland (2008). Finally, such models can accommodate the anatomical connections and functional distinctions known from neuroscience in a more straightforward fashion than symbolic models. Fig. 1 depicts anatomical connections between the spatially relevant regions described in Section 2, and shows some example recorded firing fields of cells with spatially localized firing. Most neural network models reviewed below attempt to be consistent with at least a subset of these results. For example, all of them model place cells, except for the SOM model by Voicu (2003). The model by Byrne et al. (2007) accounts for all of these cell types (with a simplified anatomy).

3.4.1. Models evaluated in real-world environments

• A large number of biological ANN-based models have been proposed based on the hippocampus and other neuroanatomical bases of spatial memory. Burgess et al. (2000) proposed one such model that was implemented on a Khepera robot, based on the influential idea that place cell firing is driven by inputs with

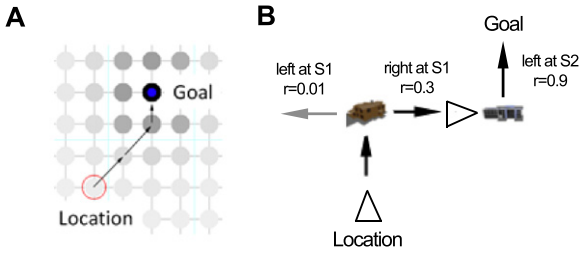


Fig. 4. Two navigation strategies. A: Allocentric navigation using a gradient ascent strategy on a heavily interconnected network of place representations, as used by the biological ANN model by Burgess et al. (2000), the ANN model by Schölkopf and Mallot (1995), as well as the LIDA hybrid cognitive architecture (on a hierarchical network). B: Egocentric navigation by always executing the action associated with the highest reward r at each state S, learned by reinforcement learning (used in the neural network models by Barrera et al., 2011, and Strösslin et al., 2005, as well as in the CLARION cognitive architecture).

Gaussian responses tuned to the presence of walls at particular distances (O'Keefe & Burgess, 1996) (later expanded and called Boundary Vector Cell model Barry et al., 2006, which successfully accounted for rat neural and human behavior data, but was not implemented in a real-world robot). The model is mainly designed to account for the place specificity of hippocampal cells and their contribution to behavior.

It consists of a population of 'sensory cells', projecting to 'entorhinal cells', which map to 'place cells' via competitive learning, which in turn map to 'goal cells' by one-shot Hebbian learning. 'Goal cells' also receive inputs from a reward signal and from four 'head-direction cells' (north, south, east, west). Sensory cells are a rectangular array of cells, each representing a different possible distance and allocentric direction to a wall, just like BVCs (Barry et al., 2006) (however, unlike BVCs, only the four orthogonal compass directions are represented). Each entorhinal cell receives hard-wired connections from two sensory cells related to two orthogonal walls. Entorhinal cells are connected to place cells, with the connection weights being adjusted by competitive learning in order to increase the spatial specificity of place cells. Finally, connections between place cells and goal cells are learned by one-shot Hebbian learning—when the agent encounters a location with a reward, a goal cell is excited, and the connection between it and the corresponding place cells increased. When the rat moves away from the reward location, the activity of these place cells will decrease; thus, the activation of goal cells will encode the proximity to the reward, allowing a gradient ascent based navigation strategy.

The robot running the model is able to navigate to local goals. It is running in a single small environment without objects, and cannot plan novel paths. However, the modeled place cell firing fields resemble empirically observed firing fields (including changes in their amplitude and shape when the environment is changed in size or shape—these firing field changes are reported to be consistent with experimental data).

The model is largely based on the neural basis of allocentric spatial memory. Although the goal learning model is speculative, both the anatomical connections and arising firing fields of the 'place cells' in the model are plausible, and qualitatively resemble empirically recorded firing fields. Later extensions of the model – which however have not been implemented on a real-world system – include comparison to empirical data, electrophysiological data recorded from rats as well as human behavior data (Barry et al., 2006) (the model could successfully account for the effect of changed environment size on both the firing fields of rat place cells and on object locations remembered by humans).

• Another biological ANN model that is also capable of controlling a real-world Khepera robot was proposed by Strösslin et al.

[8] A recurrently connected network of units whose time dynamics settle to a stable pattern (e.g. a stationary point or a time-varying pattern; Eliasmith, 2007). A type useful for spatial representations is called continuous attractor neural network (CANN), which is able to represent a point in space by means of an activity packet in the network centered on a specific spatial location. The activity packet stays stationary with no inputs, but if a unit near it receives activation it moves toward that unit—see e.g. the path integration model of Samsonovich and McNaughton (1997).

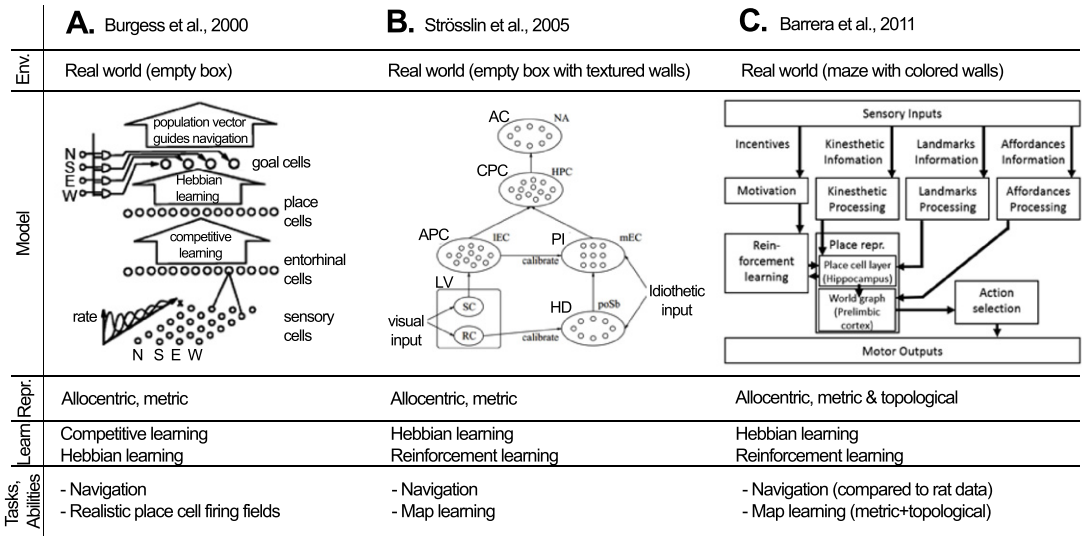


Fig. 5. Overview of neural network models evaluated in real-world environments.

(2005), building on earlier modeling work (Arleo & Gerstner, 2000). Unlike (Burgess et al., 2000), this model includes full visual processing, not just distance measurements to boundaries. The model consists of multiple interconnected populations of neurons (subnetworks).

The 'local view' (LV) processes and stores visual stimuli, and contains rotation cells and step cells. The 'head direction system' (HD), corresponding to the postsubiculum, contains head-direction cells (driven by rotation cells in LV). The 'allothetic place cells' (APC) represent the agent's position in the environment (driven by step cells in LV). The 'position integrator' (PI) is a path integration system (driven by step cells in LV). Both the APC and PI project to the 'combined place code' (CPC), corresponding to the hippocampus and subiculum. Finally, 'action cells' in nucleus accumbens perform navigation learning based on place cells in CPC. The model uses V1-inspired ensembles of units with Gabor wavelet-like receptive fields (filters) to represent visual input in LV. Rotation cells (RCs) in LV discriminate headings regardless of position, based on average relative distance between stored and current filter activity; whereas step cells (RCs) discriminate positions – regardless of headings – based on perceived angular differences between landmarks (firing rates of SCs depend on the most similar column difference of the associated filters, similarly to the 'disparities' in the model by Franz et al. (2008) described above). The HD system updates head directions based on both idiothetic cues (dead reckoning) and allothetic cues (from the rotation cells). APC place cells are driven by multiple step cells (connections are set by one-shot Hebbian learning), and thus their firing is based on the current view. APC place cells help calibrate PI cells using allothetic information to correct accumulating errors Etienne et al. (1996). Finally, information from APC (allothetic) and PI (idiothetic) converge in the CPC place cells.

Connections between APC and CPC are modified using Hebbian learning. Goal-driven actions are learned in AC using Q-learning, a variant of reinforcement learning (the ACs would correspond to neurons in the nucleus accumbens). Each action cell encodes a motor command, determining the allocentric direction of the next movement.

The model is capable of learning a map in the form of a consistent place cell code, and is able to solve navigation tasks and learning tasks such as the Morris water-maze task.[9] It cannot plan novel routes.

Although not using spiking dynamics, the model incorporates insights from the neuroscience of spatial cognition known at the time of its development, and, unlike many ANNs, does not include neuroscientifically questionable design decisions. Furthermore, it is consistent with the neuroanatomy of the hippocampal–entorhinal complex. Thus, it can claim a high amount of neural plausibility. In addition to the neurally plausible models reviewed in the next section, it also functions in the real world, with realistic input. However, it is not evaluated against neural or behavior data.

- Barrera et al. (2011) proposed another biological ANN model based on brain neurophysiology, which they evaluated against rat behavior data, unlike the previously reviewed models (extending their earlier work Barrera & Weitzenfeld, 2008). Similarly to the model by Strösslin et al. (2005) above, they use modeled 'place cells' to represent spatial locations, and use reinforcement learning to learn appropriate reward-oriented actions at spatial locations. Their model receives four kinds of sensory inputs: incentives (providing the motivation/reinforcement signal), kinesthetic self-motion information, visual landmark information (driving the place cell representation), and affordances information (providing possible actions to the action selection module).

These kinds of input are processed by four corresponding modules, a 'motivation' module (calculating a reward signal from the incentives), a 'path integration' module (updating position based on self-motion), a 'landmarks processing' module (representing the current view of the animal, based on all perceived landmarks; suggested to correspond to the EC), and an 'affordance processing' module (encoding possible turns the rat can perform at a given location and orientation). The reward signal from 'motivation' drives the 'learning' module (learning by reinforcement; corresponding to the VTA, NA and striatum in brains), and the outputs of the path integration and landmarks processing modules drive the 'place representation module', which in turn project to the 'action selection' module. The 'place representation' module includes 'place cells' (PCs, the activity of which arises from a weighted linear combination of the path integration and landmark inputs; corresponding to the hippocampus) as well as a 'world graph layer' (WGL, suggested to correspond to the prelimbic cortex). The WGL learns a topological map by learning associations between overlapping

[9] In the Morris water-maze task, rats are placed into a pool of water in which they have to swim. The pool contains a hidden platform. The rats search for and eventually find the platform, and remember its location in their spatial memory. Subsequently, they immediately head for the remembered location of the platform when placed into the pool.

place fields, as well as learning actor units representing actions with high expected rewards associated with place cells (actor unit weights are learned by reinforcement learning). The WGL also performs place recognition, by classifying the currently active PCs.

Finally, the action selection module computes a motor output (the next moving displacement and direction), given the current possible affordances, current location (place cells), and the expectations of maximum reward from the actor units in WGL. The model is able to learn metric (PCs) as well as topological maps (WGL) in the place representation layer, and is able to navigate to reward locations.

The authors evaluated their model against rat behavior data in a simple maze navigation paradigm, in which water-deprived rats were looking for a water dispenser, learning its location during a number of training sessions. They used AIBO robots in the same paradigm, in similar mazes. The robots could learn near-accurate metric and topological maps of the mazes, and exhibited learning curves (during learning the reward location) and numbers of incorrect trials and optimal trials (during test trials) similar to those of the rats.

The model is based on rat neurophysiology, and thus is neuronally plausible. It is also able to function in the real world, and has also been evaluated against rat behavior data (the learning curves in a simple maze were comparable), lending credence to the authors' claim that their model can be used by experimentalists to predict rodents' spatial behavior, and test neuroscientific hypotheses. Additionally, although not replicating neural data, the authors present results verifying the engagement of the proposed neural correlates of their models (reporting gene expression data) in the rats they used in their experiments (Barrera et al., 2011).

A further neural network based model of mapping very successful in robotics which was also inspired by rat neurophysiology is RatSLAM (Milford & Wyeth, 2010). It will not be reviewed here, since RatSLAM is not a cognitive model, and is not compared to or intended to model either behavior or biology (the authors aim for practical robot performance instead of plausibility).

3.4.2. Models evaluated in simulations

• Schölkopf and Mallot (1995) proposed a neural network model of cognitive map learning in a maze, a model aiming for cognitive rather than biological plausibility (but nevertheless pointing out similarities to neural substrate). Their agent employs a central perception–action cycle (Fuster, 2002) (similarly to the sense–plan–act cycle of the symbolic perceptual wayfinding model; Raubal, 2001). The model assumes it is dealing with a maze environment consisting of at least two places, with corridors connecting the places; and also assumes a direct correspondence between these corridors and 'views' (a view is thought of as being attached to the wall opposite to the entry of the respective corridor); and that views are uniquely distinguishable.

The model is based on the idea of a 'place graph' (an allocentric graph of places, connected by corridors) and a 'view graph' (a graph of local views connected by edges with labels representing egocentric movements; and connected only if they can be experienced in immediate temporal sequence). The view graph is learned using a SOM-type (self-organizing sequence map) neural network (Kohonen, 1990), which has three layers: an input layer (with units representing views), a movement layer (representing the movements left, right or back; with only one of these three units active at each time), and a 'map layer'. The map layer receives sequences as inputs, from both the movement layer (a sequence of movements), and the view layer (a sequence of views represented by the activity of the view layer units). A map of the current maze is learned by 'random exploration', i.e. a large number of random movements and views are passed to the network, which uses learning by self-organization (Kohonen, 1990) to assign map units in a way that they closely resemble the view graph (i.e. near views are represented by near units, and distant views by distant units). After learning, path planning to arbitrary views can be performed by a gradient ascent strategy (spreading activation from the goal, and then at each map unit, progressing to the adjacent map unit with the highest activation), a planning strategy that the authors implemented algorithmically (not in a neural network).

Unlike the previously reviewed neural network models, this model is able to plan novel routes algorithmically. It is also one of only three neural network models implementing topological maps (the other two being Barrera et al., 2011 and Erdem & Hasselmo, 2012).

Since there is little direct correspondence between this model and neuroanatomy, and since planning is implemented algorithmically, this model cannot be called biologically plausible. However, it is argued by the authors to functionally resemble some aspects of biological spatial memory (such as free/passive exploration and expectations of future views).

• A model also based on self-organized learning was proposed by Voicu (2003), extending their earlier work (Voicu & Schmajuk, 2000). Unlike the model above, it is capable of running in a full two-dimensional metric simulation instead of a restricted maze-like environment. A further difference is that it learns hierarchical instead of flat spatial representations—which is frequently argued to be the structure of cognitive maps (see Derdikman & Moser, 2010, Hirtle & Jonides, 1985 and McNamara, 1986, for behavioral and Derdikman & Moser, 2010, for neural evidence).

The model architecture consists of a hierarchical allocentric cognitive map and four additional modules (a localization system providing landmark representations, a working memory for planning paths, a motor system translating them into actions and a control system supervising information flow between these modules). The cognitive map itself uses types of SOM (recurrently connected hetero-associative networks; Kohonen, 1990) to build associations. There are three different networks representing associations between all landmarks, associations between landmarks having the largest number of associations at the first level, and associations between landmarks having the largest number of associations at the second level, respectively. The map is learned in two stages: an exploration stage for building the first level at the highest resolution (moving randomly at the beginning, avoiding previous places, and then, over ten acquired landmarks, moving toward those having the fewest associations), and a second stage, building the hierarchical cognitive map (selecting the landmarks with the largest number of associations and associating them). Weights are adjusted depending on distance (lower distances yielding lower weights), so that activation gradients can serve to plan a path toward a goal.

The model can learn hierarchical metric maps, and can plan novel paths. It succeeded in reproducing the empirically observed hierarchical cognitive maps by Hirtle and Jonides (1985), and also produced similar distance judgment errors as humans (distances spanning multiple clusters or submaps are overestimated both by humans and by the model).

This model uses SOMs, types of ANNs, and does not aim to be neurobiologically plausible. The spatial specificity of its SOM units is also a property of hippocampal place cells, but its units correspond to much larger areas than the observed PFs of place cells. However, it is able to reproduce human behavior data, and thus can make empirically validated claims for cognitive (if not biological) plausibility.

• The *map-based path integrator (MPI)* model by McNaughton et al. (1996) was an influential and still highly relevant model of spatial representation and path integration in brains, implemented as a SNN. It was tested and evaluated by Samsonovich and McNaughton (1997) and later reviewed and argued to be plausible

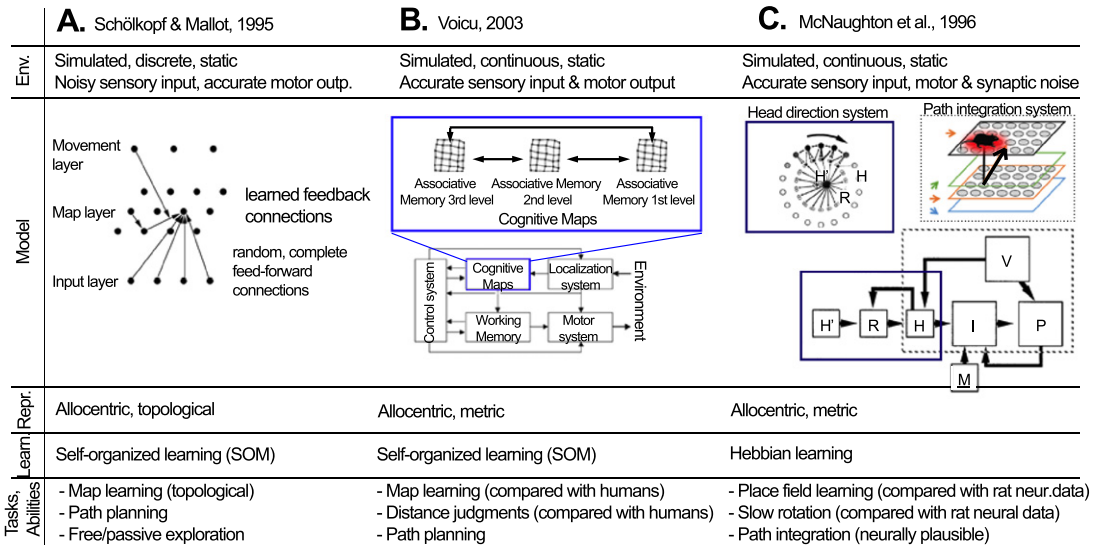


Fig. 6. Overview of neural network models evaluated in simulated environments.

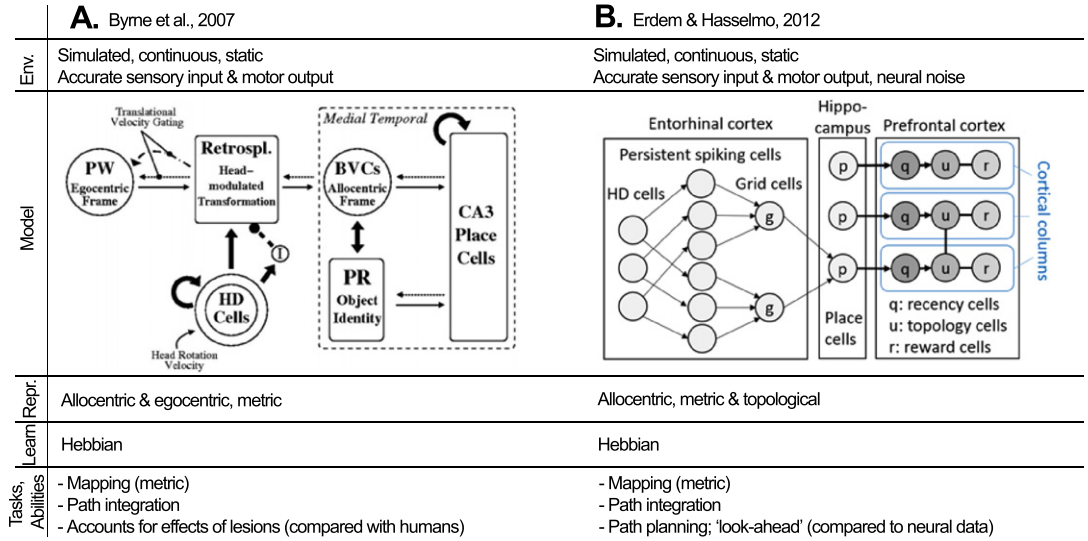


Fig. 7. Overview of neural network models evaluated in simulated environments.

based on neural evidence by McNaughton et al. (2006). The model is based on 'attractor maps', continuous attractor networks in which the mobility threshold for transitions between neighboring attractors is negligibly small, as opposed to the large thresholds for jumps between distant points (and with global feedback inhibition limiting total activity in the network)—this leads to activity focused on one maximum unit and declining with distance from that unit (i.e. an activity packet), tending to move toward the maximal input into the network or staying stationary in the absence of input.

The two most important modules are H, a one-dimensional cyclic attractor map (CANN) encoding the head direction of an agent (containing 'HD cells' arranged on a circle in the order of their head-direction preference), and a P, a two-dimensional attractor map used to encode the agent's current position, as well as for path integration (containing 'place cells' arranged in a plane, with weights that decrease with distance). The head direction estimate and position estimate correspond to the maxima of the activity packets on the circular CANN and two-dimensional CANN, respectively. To implement path integration for the HD cells, two additional layers are required, one with units representing angular velocity (H′), and a conjunctive layer representing both current head direction and velocity (R—receiving connections from H and H′), and projecting back to the appropriate HD units. The R layer drives the HD activity packet in the right direction whenever the agent is turning, since R units project to the right of the currently most active HD cell for positive angular velocity, and to the left for negative velocity (and with below-threshold activity if the velocity is zero).

Similarly, path integration for 'place cells' in P works by employing a number of intermediate 2D CANN layers in the I module, each layer corresponding to a different possible head direction (and receiving activation from that HD cell in H), with connections that project to units in the P layer, but displaced in the respective head direction (e.g. if the 'north' HD cell activates the corresponding I layer, units of this layer would project to a place cell that is associated with more northern locations in the P layer, instead of equivalent units in P corresponding to the same locations as units in I). Thus, the projection to P from the currently active I layer (depending on the most active HD cell) can move the 'place cell' activity packet in the correct direction.

Finally, HD cells and place cell firing is not only driven by path integration, but also by associated sensory representations,

encoded in an additional module called V. Associating spatially localized place cells with sensory representations can correct accumulating path integration errors, as well as represent stimuli encountered in a specific location. Such sensory associations can be learned by Hebbian learning, whereas the weights driving the path integration mechanism (such as from H to I) which are preconfigured and fixed.

The model is implemented as an integrate and fire SNN, and is able to numerically reproduce several single-cell experimental findings, such as place field stretching upon changing environment size, dependence of place field location on the entry site, slow rotation of place fields in disoriented rats, and learning in novel environments; and also makes a novel prediction which was verified experimentally after publication of the model (activity jumps in P upon significant unexpected changes in sensory input). However, navigation or path planning or the representation of objects on the map is not explicitly modeled by the authors. The model's main strength lies in proposing the first plausible neural network model of path integration.

The model and its elements are neuroanatomically plausible; MEC might perform path integration (passing activation to hippocampal place cells), and the analogy between modeled and biological HD cells is clear (see McNaughton et al., 2006 for evidence). Despite the anatomical plausibility of the elements, and the common use of attractor networks to model head direction, it should be noted that no empirically validated mechanism has been proposed yet that could result in the very specific connectivity required by continuous attractors in brains.[10] The model is implemented as a SNN, and is thus biologically more realistic than the reviewed ANN models. Finally, it also succeeds in reproducing and even predicting empirical data, further substantiating its plausibility.

- Another influential model was the *'BBB' model* proposed by Byrne et al. (2007) and based on the BVC model (Barry et al., 2006) (the predecessor of which was implemented on a robot, and reviewed in the previous subsection; Burgess et al., 2000). The model is based on the brain areas involved in allocentric spatial representations in the medial temporal lobe, as well as the egocentric areas in the parietal lobe (see Section 2), and thus accounts for both kinds of reference frames.

In the model, egocentric maps are represented by a set of neurons in a grid, each tuned to respond most strongly to an object at a particular distance and direction from the agent's head. Allocentric maps are represented similarly, using neurons with specific preferred distances and directions, with the difference that the neurons reference direction is fixed to features of the environment, instead of the agent's current head direction (these are equivalent to BVCs). The model consists of an 'egocentric frame' module (representing egocentric maps, corresponding to the precuneus), a 'HD cells' module representing head direction, a 'transformation' module (translating between egocentric and allocentric maps, corresponding to RSC), an 'allocentric frame' module (representing allocentric maps, suggested to correspond to BVCs), a 'place cell modules' (representing current location and associating sensory representations with locations), and an 'object identity module' (for sensory representations, with each unit representing an object or landmark; corresponding to the perirhinal cortex).

The network has a 'top down' (temporal to parietal) and a 'bottom up' (parietal to temporal) phase, during which the allocentric map updates the egocentric one and vice versa (the information flow in the opposite direction is blocked in each phase). Similarly to the BVC model and its predecessors (Burgess et al., 2000), place cell firing is driven by BVCs (the firing of which in turn depends on the distances and directions of boundaries). The 'transformation' module contains N identical subpopulations, each tuned to a specific head-direction, and connected to the egocentric map so as to rotate it by the angle of that head direction (to translate it to a north-oriented allocentric reference frame). At each time step, only the subpopulation corresponding to the currently active HD cell is active. Just like in the previous model, HD cell activities are updated using CANN dynamics and angular velocity input; however, unlike the MPI, linear path integration is not performed by the allocentric representation. Instead, the 'transformation module' performs this function as well, by having an alternative set of pre-trained weights that result not only in the rotation but also in the translation of a map by a constant amount (the model only accounts for constant velocities).

The model is able to learn allocentric as well as egocentric representations of the local space surrounding the agent in a simulation, and is the only reviewed neural network-based model with the ability to translate between the two. It is also able to mentally explore representations, and to plan routes, by mentally generating velocity signals ('mock motor efference') which are decoupled from the motors. However, it cannot plan novel routes (e.g. shortcuts/detours).

Because of the clear correspondence of model parts and brain areas, the authors are able to simulate 'lesions' (by selectively deactivating model parts or connections) and to account for lesion studies (failure to identify landmarks in half of the egocentric space hemispheric neglect patients; and place cell firing with HD cell lesions). They could also model mapping, path integration, and a paradigm in which visual and path-integrative inputs were conflicting.

The model was implemented as a biological neural network (with rate-coded instead of spiking neurons). Its modules and connections are based on neuroscientific, and psychological evidence, and are highly plausible. The model was further strengthened by evaluating it in lesion study paradigms and qualitatively comparing the results with human and rat data.

Most reviewed neural network models accounting for navigation make use of either place cell-like units associated with units representing motor actions, or a gradient ascent strategy, propagating activation from a goal location in a heavily interconnected place cell-like network, and always selecting directions that increase the current activation, until eventually reaching the goal. There is no direct evidence for either of these strategies actually being used by brains (no action representations monosynaptically connected to place cells have been found; and except for area CA3, place cells do not seem to be heavily interconnected—and in any case, such activity diffusion is inherently limited in range due to signal decay in biologically realistic networks).

- In contrast to these navigation strategies, Erdem and Hasselmo (2012) have proposed a SNN model of navigation based on probing linear look-ahead trajectories in several candidate directions to find a trajectory leading to the goal location.[11] This model is also based on the neural correlates of allocentric spatial memory in the medial temporal lobe, and incorporates hierarchical spatial representations. It incorporates four modeled medial temporal cell types, and an additional three cell types in a 'PFC' module.

[10] However, there is some empirical evidence substantiating the existence of continuous attractors in brains (Yoon et al., 2013).

McNaughton's continuous attractor networks are also prone to accumulating errors, requiring external sensory input to correct them, and have distorted firing fields at the edges of the network. Later work has improved these issues (e.g. Burak & Fiete, 2009).

[11] Earlier, less neurally plausible models of the same group have also used omnidirectional probing for navigation (Gorchetchnikov & Hasselmo, 2005).

Suggested to correspond to the entorhinal cortex, it models 'head-direction cells', 'persistent spiking cells', and 'grid cells', and corresponding to the hippocampus, it models 'place cells'. The prefrontal module in turn contains 'recency cells', 'topology cells', and 'reward cells' (presumably corresponding to mPFC). HD cells are modeled to have a receptive field at a specific preferred angle from an anchor cue (they are only driven by sensory input, not by self-motion, unlike CANN models of HD cells). Modeled grid cell firing is based on the persistent spiking cell model (briefly, grid fields arise from an interference oscillation in persistent spiking cells) (Hasselmo, 2008). Place cells are driven by grid cells in the model, as suggested before by theoretical models (Moser et al., 2008; Solstad et al., 2006); place fields arise from a thresholded product of the grid fields (the multiplication is implemented using coincidence detection in the model).

In contrast to the metric place cell map, a topological map is created in the PFC module. Each place cell is associated with a corresponding recency, topology, and reward cell; and topology cells are laterally interconnected. The activity of recency cells decays exponentially in time; their firing depends on the time elapsed since the last visit of the associated place cell. Each time the agent visits a place cell, the topology layer's lateral connections are reinforced by Hebbian learning, depending on thresholded current activities of recency cells, with the threshold controlling what time window is considered 'recently visited' and which topological weights should be reinforced. Finally, reward cells are also associated with place cells, and fire persistently if their corresponding place cell marks the location of a goal or reward.

During goal-directed navigation, the agent decides on what direction to choose by probing several linear look-ahead trajectory probes with different directions starting from its current location. Each probe engages the HD cell–persistent spiking cell–grid cell–place cell circuit as if the agent was physically moving along the probe trajectory. If the probe leads to the activation of a reward cell at the goal location, associated with a place cell, the rat proceeds to move in the direction of the probe. In order to avoid the probes missing the goal location, and to allow reaching intermediate goals, the reward signal is diffused in the PFC module. Thus, secondary goals associated with place cells close to the reward cell (and thus receiving diffused activation from it) can be navigated to first, until the agent gets close enough to find the actual, highest-activated reward cell with a probe. Finally, since only directions not obstructed by an obstacle can be probed, the agent can navigate around obstacles (but also find a novel shortcut once an obstacle is removed and a novel probe direction to the reward becomes possible). The model was able to produce grid cell ensemble activity resembling recorded rat medial entorhinal neurons demonstrating 'look-ahead' activity in a T-maze navigation task (Gupta, Erdem, & Hasselmo, 2013).

The model is able to learn both metric and topological maps, and can perform path planning on the learned maps, including planning novel routes such as shortcuts or detours.

This model is implemented as a SNN, using biologically realistic modules and connectivity; furthermore, its look-ahead mechanism results in activity patterns resembling data from biological neurons.

3.5. Spatial memory models in cognitive architectures

In contrast to computational cognitive models focused on accounting for one or few specific processes, systems-level cognitive architectures aim to comprehensively model a wide range of cognitive phenomena, attempting to account for behavior and structural properties of minds (Sun, 2007). Cognitive models of specific processes can be implemented within the framework of a systems-level cognitive architecture. Such models also play an important role in cognitive science, providing detailed, formal explanations, providing hypotheses, and guiding research (see Introduction). However, the goal of being integrated with a broadly scoped, domain generic model – and desirability of being able to function using the same mechanisms and internal parameters as an agent running the same cognitive architecture in a completely different task – sets the task of modeling with cognitive architectures apart from the task of developing cognitive models.

A large number of cognitive architectures have been proposed (many of which deal with modeling spatial representations in some way), too many to review here; we will aim to outline a representative sample contributing to spatial memory modeling instead of exhaustiveness, and only include architectures explicitly claiming to model human or animal cognition (we omit the large number of robotic or AI architectures uninterested in biological cognition). More comprehensive reviews can be found in Duch, Oentaryo, and Pasquier (2008), Goertzel, Lian, Arel, de Garis, and Chen (2010) and Samsonovich (2010).

There is some intersection here with the previous two categories, since there are cognitive architectures that are exclusively symbolic, exclusively neural network-based, or hybrid (combinations of symbolic and neural network parts); we shall point out the corresponding paradigm in the text, as well as in the comparison in Table 1. To the authors knowledge, there exists no cognitive architecture explicitly aiming to be cognitively plausible (i.e. model humans or animals) which would account for navigation-space spatial memory as well as being implemented on a real-world robot in current literature. Thus we omit the 'real-world' category from this section—all reviewed models run in simulations.

The popular ACT-R (Adaptive Control of Thought Rational) cognitive architecture by Anderson, Matessa, and Lebiere (1997) follows a production-rule based approach (productions consist of sensory preconditions or 'IF' statements, and associated actions or 'THEN' statements executed when the precondition matches the state of the world). It utilizes two types of memory: declarative memory, encoding factual knowledge about the world (as symbolic entities called 'chunks'), and procedural memory, containing procedural knowledge in the form of productions (IF-THEN rules). The general usefulness of these chunks and production rules is stored in a neural network reflecting previous usage (which has led some researchers to categorize ACT-R as a hybrid cognitive architecture, despite it being primarily symbolic Duch et al., 2008).

Apart from memory, the central components of ACT-R are perceptual-motor modules interfacing with the environment, buffers, and a central pattern matcher for productions (matching, selecting and executing production rules). This central module is hypothesized to correspond to the basal ganglia in the brain. ACT-R has been used to replicate a large number of psychological experiments (Anderson et al., 2004). Although the original version did not explicitly account for spatial cognition, it has later been extended to include spatial memory models.

- One such extension, called *ACT-R/S* was proposed by Harrison et al. (2003), adding two additional systems to ACT-R: a 'manipulative system' (representing spatial characteristics of objects facilitating manipulation), and a 'configural system' (representing the relative, approximate configuration of objects in space). The latter consists of a 'path integrator' and a buffer containing a number of spatial chunks called 'configurals', each storing an egocentric vector to an object along with its identity (ACT-R/S only includes egocentric representations). Objects attended to enter this configural buffer, which holds the two or three most recent objects—when this capacity is exceeded, the least recent chunk will be discarded from this buffer (but will still exist in 'declarative memory' for later retrieval).

The 'path integrator' – instead of updating an allocentric location representation – updates all egocentric representations in the

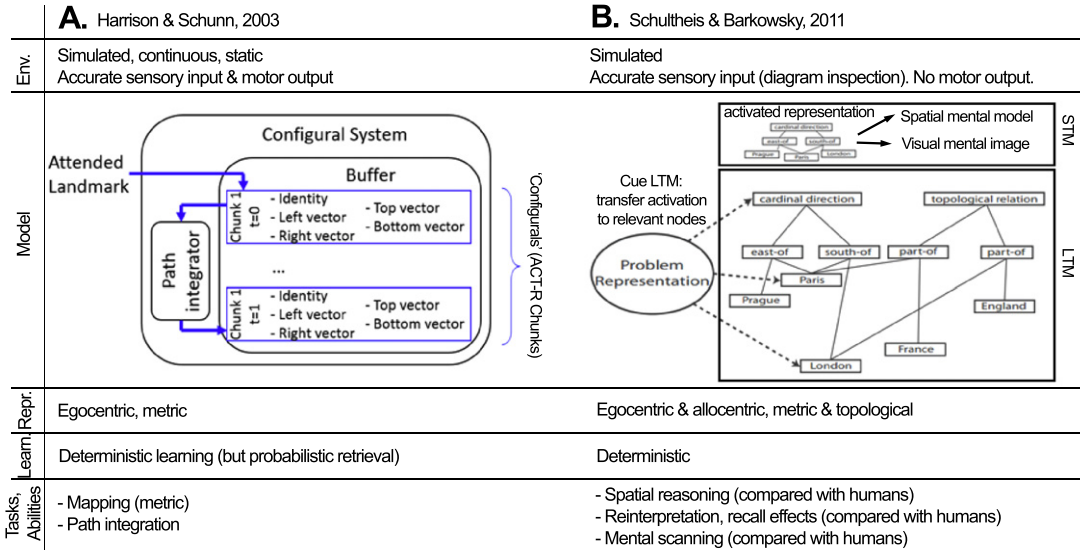


Fig. 8. Overview of cognitive architectures evaluated in simulations.

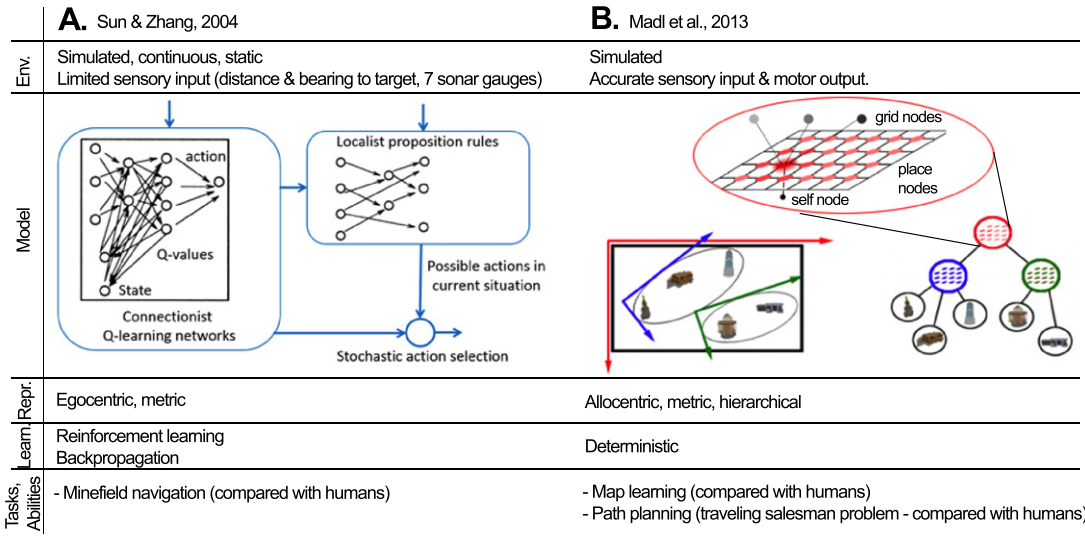


Fig. 9. Overview of cognitive architectures evaluated in simulations.

configural buffer after each movement by the motor system (this is feasible due to the small number of configurals actively maintained in the buffer). Apart from object identity, configurals store multiple vectors, to all edges of an object—in the implemented model, which was two-dimensional, objects were approximated by their bounding box, and four vectors were stored to the edges of that bounding box (to the left, top, right, and bottom sides of an object). Multiple configurals referring to the same object from different points of view can be present in the model, which would have different edge vectors but the same identity tag.

The authors implemented a food search model, which can randomly explore an environment, try to recall a food location, or visually search for food. The search is performed by requesting unattended objects from the configural system, identifying it using the visual system, and continuing the search if it is not food, or setting it as a goal if it is. In the latter case, the agent orients itself toward the food location, and begins another search (this time for obstacles—any object that intersects its path to the food location). If obstacles are found, the agent adds a subgoal to move to the left or right of it, depending on which brings it closer to the goal. If no obstacles are left, it moves to the goal location. During navigation, the agent repeatedly checks if it has arrived at its destination, and also repeatedly corrects path integration errors using its visual system (that is, if the egocentric representations updated by path integration do not match their perceived correct location, they are corrected).

Furthermore, it encodes 'episodic traces' (current contents of the configural buffer) at each step. If visual search fails to find food, these episodic traces can be recalled to find previously identified food locations as well as nearby objects (after which it can perform another visual search for those nearby objects and navigate to them to get closer to the food location). The authors functionally evaluate this model of path integration and navigation, and point out functional similarities between configural chunks and primate spatial-view cells.

The psychological plausibility of the ACT-R model and its parameters (buffer capacities, timings etc.) have been extensively strengthened in a large number of different paradigms. There is also functional similarity between the egocentric representations in this model, and egocentric representations in the brain

(e.g. spatial-view cells). However, there is no clear structural similarity between this model and neurobiology.

• *Casimir* by Schultheis and Barkowsky (2011) is a cognitive architecture explicitly devised as a framework for computationally modeling human spatial knowledge processing. Its main parts are a long-term memory (LTM), working memory (WM), and a diagram interaction component (externalizing WM representations on diagrams, or visually inspecting diagrams to build WM representations).

The LTM stores hierarchical, semantic network-like representations (nodes and connections between them; categories and objects as well as spatial relations are represented as nodes, whereas connections signify associations; e.g. three nodes and two connections could represent the relation 'Paris'-'south of'-'London'). The WM can be split into three parts, one concerned with retrieving representations from LTM based on a 'problem representation', one performing memory updates of WM and LTM, and a 'visuo-spatial WM' part storing and manipulating short-term representations relevant to the current problem. The problem representation also takes the form of a semantic network, and allows the specification of a query (such as the cardinal direction to a location, or a distance between locations).

Retrieval from LTM works by spreading activation over the nodes in LTM from the problem representation; the subnet ('fragment') with the highest sum of activation is retrieved to the visuo-spatial WM (retrieved subnets also have to be directly or indirectly interlinked). This LTM structure and retrieval process can account for some human memory phenomena. Knowledge from different sources can enter visuo-spatial WM, including knowledge retrieved from LTM, built by visual inspection, or constructed from previous representations; and is represented not symbolically but in a spatio-analogical form (i.e. there is a structural correspondence between the representations and what they represent in the world).

Casimir assumes that there is no strict division between spatial and visual representations, but, rather, a continuum between the extremes of simple nonmodal spatial mental models (spatial) and mental images (visual). Representations are deemed more visual with increasing numbers of relations, involved knowledge types (such as distance, direction, topological knowledge), specificity, and exemplarity (concrete exemplars or prototypes). A 'conversion' process in working memory can construct and extend representations, adding retrieved fragments if necessary, or converting fragments to spatial mental models. An 'exploration' process in turn can extract spatial information from existing representations, or infer knowledge using spatial reasoning.

Because of its emphasis on structural modeling (spatio-analogical instead of symbolic representations), Casimir is argued to exceed the modeling capabilities of other cognitive architectures in the spatial domain (Schultheis & Barkowsky, 2011). The architecture was tested on paradigms involving eye movements in a spatial reasoning task (Sima, Lindner, Schultheis, & Barkowsky, 2010), mental scanning (the effect of the time to scan between entities in a mental image increasing linearly with the distance between them), mental reinterpretation of spatial relations (Sima, 2011), and recall effects (Schultheis, Lile, & Barkowsky, 2007). The model has a simple visual perception implementation facilitating the replication of such experiments. However, navigation has not been implemented.

The model is heavily based on prevalent cognitive science theories of mental representations (e.g. analogical representations Barsalou, 2008, mental models Mani & Johnson-Laird, 1982, mental images Shepard & Metzler, 1971), and replicates human behavior data in a number of paradigms. However, it does not aim to be biologically plausible, and its parts do not clearly correspond to brain areas or neurons.

• *CLARION* by Sun and Zhang (2004) is a hybrid cognitive architecture accounting for spatial representations. It incorporates explicit (symbolic) as well as implicit (subsymbolic) knowledge through its four memory modules: the action-centered subsystem (regulating procedural knowledge and actions), non-action-centered subsystem (maintaining general declarative knowledge), motivational subsystem (providing motivation for action), and metacognitive subsystem (monitoring and directing the operations of the other subsystems).

Each module has a localist-distributed representation (explicit knowledge) and a distributed section stored in a neural network (implicit knowledge). Spatial representations can be acquired by associating explicit knowledge in the form of 'chunks' (similarly to ACT-R chunks—e.g. a chunk representing a reward) with the corresponding implicit representation of sensory input.

CLARION's ability to represent and navigate in space is shown in the complex minefield navigation (MN) task implemented by Sun, Merrill, and Peterson (2001). In this task, an agent has to navigate through a two-dimensional minefield to reach a target. The agent only has access to limited sensory information (short-range sonar readings to mines, range and bearing gauges showing distance and direction to the target, and the remaining time), and has to reach the target in a limited amount of time. Only egocentric spatial relations were used (distances and directions to nearby mines). The agent used a type of reinforcement learning called Q-learning (with a gradient reward depending on target distance, and a second reward at the end depending on the agents success—depending on how close it got to the target) to learn an optimal action policy.

The model was evaluated against human behavior data, and produced trajectories and learning curves similar to humans in this paradigm. It does not learn an allocentric map; rather, it uses reinforcement learning to learn the optimal actions to reach its goal given the obstacles in the environment. Information about the current obstacles is represented as implicit knowledge in the 'state' layer of CLARION's neural network (see Fig. 9).

Since the model uses very general modules (there is no specialized spatial memory module), and since it consists of both symbolic and neural network parts, it is difficult to identify structural correspondences to neurobiology. CLARION has succeeded in modeling human behavior data from a large number of paradigms – including the above mentioned minefield navigation task – and thus can be called cognitively plausible (Sun & Zhang, 2004).

• Another hybrid cognitive architecture is *LIDA*[12] by Franklin, Madl, D'Mello, and Snaider (2014), with recently developed spatial capabilities Madl et al. (2013). Although not modeling neurons, LIDA is biologically inspired, with each major part of the model functionally mapped to brain areas (Franklin et al., 2014; Goertzel et al., 2010), and is largely based on the Global Workspace Theory of functional consciousness (Baars & Franklin, 2009; Baars, Franklin, & Ramsoy, 2013), as well as a number of psychological and neuropsychological theories including grounded cognition (Barsalou, 2008), working memory (Baddeley, 1992), and Slomans H-CogAff cognitive framework (Sloman, 1998) among others. It is a recent architecture and only partially implemented, but has replicated a number of psychological experiments (Franklin et al., 2014).

LIDA's cognitive cycles, corresponding to the action–perception cycles in neuroscience Fuster (2002), consist of three phases. The 'understanding' phase includes sensing the environment, detecting features, recognizing objects and categories, and building internal representations. The 'attending' phase is responsible for deciding

[12] Learning Intelligent Distribution Agent (Learning IDA), where IDA is a software personnel agent hand-crafted for the US Navy that automates the process of finding new billets (jobs) for sailors at the end of a tour of duty (Franklin, 2003). LIDA adds learning to IDA and extend its architecture in many other ways.

what portion of this representation should be attended to and broadcast to the rest of the system, making it the current contents of consciousness. This portion allows the agent to choose an appropriate action to execute in the 'action' phase. During the understanding phase, percepts are recognized based on LIDA's perceptual knowledge base, the Perceptual Associative Memory (PAM), which is a connectionist structure containing nodes with activation connected by links. Recognized objects, categories, etc. are stored in LIDA's preconscious 'Working Memory', and are represented by structures of PAM nodes and links between them.

These PAM node structures – parts of the PAM network – are hierarchical, modal representations similar to Barsalou's perceptual symbols Barsalou (2008). Since they are hierarchical and associative, they are well-suited to represent 'hierarchical cognitive maps', by associating PAM nodes representing objects or landmarks with 'place nodes'. Place nodes are special kinds of PAM nodes representing a spatial location; they are arranged in layers of two-dimensional rectangular grids with different resolutions (distances between the place nodes). The layers are interconnected, multiple high-resolution place nodes project to a single low-resolution place node (with overlap); which implements spatial clustering. This can account for systematic position errors in humans due to hierarchical representation (Madl et al., 2013).

LIDA agents use a gradient ascent based navigation strategy (passing activation from a goal location through the place node network), similarly to some of the neural network models above. However, a significant difference is that hierarchical map representation is used during navigation (first a rough route is planned using the lowest resolution layer, and then successively refined on the higher resolution layers). It can be shown that in multi-goal navigation tasks, gradient ascent on a single map leads to a sub-optimal nearest-neighbor strategy (as does the 'look-ahead' approach (Erdem & Hasselmo, 2012) and RL with simple goal-distances as rewards (Barrera et al., 2011; Strösslin et al., 2005), although RL with different reward functions can improve this). Humans significantly outperform the nearest-neighbor strategy in multi-goal paradigms such as the traveling salesperson problem[13] (TSP), planning near-optimal routes. The gradient ascent strategy on a hierarchical cognitive map in LIDA significantly improves route optimality, without sacrificing the biological plausibility of a connectionist map for a symbolic planning mechanism.

LIDA-based agents have been shown to be able to perform mapping and navigation, and model human behavior in different tasks, including modeling map recall errors, capacity limits of spatial working memory, and errors in the TSP paradigm (Madl et al., 2013) (work is underway to embody LIDA on a robot (Franklin et al., 2014) and to extend it with both egocentric and allocentric real-world spatial memory). Although not a biological neural network, spatial memory in LIDA is connectionist; and there is similarity between 'place nodes' and hippocampal place cells (also accounting for hierarchies in an empirically substantiated fashion, unlike most other models).

3.6. Comparative table

Table 1 shows a comparison of the reviewed models, characterizing them according to the criteria outlined in Section 3. It compares the level of modeling by stating the elemental position representation for each model, as well as the reference frames or types of representations accounted for, the learning mechanism, the structural similarity between models and underlying neural mechanisms, and the complexity of the environments and types of tasks in which the models have been evaluated (to help assess their generality and complexity). Quantitative 'goodness of fit' was not included because most models did not perform quantitative statistical evaluations against data; and the exceptions that did used different tasks.

[13] The traveling salesperson problem requires planning the shortest route visiting each location among a fixed number of locations exactly once, and then returning to the starting location.

4. Discussion

Direct comparison of the reviewed models is made difficult by their very different goals and paradigms. Although computational cognitive models should be evaluated quantitatively as well as qualitatively, the majority of the reviewed models were not quantitatively evaluated against actual behavior data. Exceptions include:

- (Symbolic—Brom et al., 2012): replication of human accuracies in pointing tasks (subjects/agent had to remember locations of several objects in a room, and subsequently asked to point to the locations after the objects have been removed)
- (Neural network-based—Barry et al., 2006; Burgess et al., 2000): This model is the only reviewed model which was compared to both electrophysiological data from rat place cells, and behavioral data human subjects. It could successfully account for the effects of changed environment size on both place fields and on remembered locations of objects.
- (Neural network-based—Barrera et al., 2011): The model's learning curve when learning to reach a goal in a maze was comparable to that of rats in an experiment
- (Neural network-based—Voicu, 2003): The model imposed hierarchies comparable to human hierarchical cognitive maps, and resulted in comparable distance estimation biases
- (Cognitive architecture-based—Schultheis & Barkowsky, 2011): Replication of eye movements in spatial reasoning, mental scanning, mental reinterpretation of spatial relations, and recall effects
- (Cognitive architecture-based—Sun & Zhang, 2004): Replication of human data in a minefield navigation task
- (Cognitive architecture-based—Madl et al., 2013): Replication of human performance in the traveling salesman problem and of map representation errors

Apart from psychological plausibility in terms of comparable behavior, the functional advantages of the models are also important aspects. Although all models represent spatial information in some form, there is a large difference in terms of the complexity of the environments they can handle, the accuracy of these representations, and the range of tasks they can be used for.

It should be noted that although all of these models can be said to create maps (of different kinds and different accuracies), only a few of them can be said to be modeling 'cognitive maps' in the sense of Tolman (1948), who has pointed out that cognitive maps can be used to plan novel routes such as shortcuts or detours (for known routes, no allocentric map would be necessary). In this sense, only 7 models are accounting for cognitive maps—those that can perform path planning (see also the 'Abilities' row in Figs. 2–9): Beeson et al. (2010), Byrne et al. (2007), Erdem and Hasselmo (2012), Gopal and Smith (1990), Madl et al. (2013), Schölkopf and Mallot (1995) and Voicu (2003).

In general, models capable of handling a higher environmental complexity in Table 1 should be regarded as functionally more powerful. Models capable of running in the real world face greater challenges and are more difficult to implement than simulated models, since they need to cope with noise and errors both in their sensory input and motor output, as well as with the usually greater complexity and unpredictability of real environments.

Table 1
Characteristics of the reviewed models. The table consists of seven columns, showing model name and citation, the elemental spatial position representation, reference frames accounted for (Ref.), learning mechanisms, if any (Learn.), structural similarity to corresponding brain areas (Sim.), complexity of the environment the model is able to operate in (C.), and tasks in which the model has been evaluated. The following further abbreviations are used:
• Superscripts in the model name denote whether they are symbolic (s), neural network-based (n), a combination of the two (hybrid—h) (necessary in the case of cognitive architectures). • In the reference frames accounted for: ego…egocentric, allo…allocentric, vs…visuo-spatial, +…ego + allo, *…all three (superscripts denote whether the maps are t…topological, m…metric, n…metric but non-Euclidean, o…containing orientation information)
• In the learning mechanism: prob…probabilistic, det…deterministic, SLAM…probabilistic Simultaneous Localization and Mapping, Hebb…Hebbian, RL…reinforcement learning
• In the structural similarity: numbers range from 5—strong similarity to 1—no clear similarity (5…biological ANN or SNN evaluated against neural data, 4…biological ANN or SNN, 3…non-biological ANN, 2…symbolic mechanism with clear correspondences to modeled biological mechanism, 1…no clear structural similarity)
• In the complexity of the environment: numbers range from 5—highly complex to 1—simple (5…large-scale real-world env., 4…small real-world env. mostly within agent's sensory horizon, 3…simulation with multiple objects or obstacles, 2…simulation with no objects or obstacles, 1…finite number of discrete states)
• In the tasks in which the model has been evaluated: brief task names are used; they are further described in the text (the superscript denotes the type of data compared against: h…human behavior, a…animal behavior, n…animal neural data, q…no quantitative, only qualitative comparison against human behavior data, and -…no biological or behavior data, only functional evaluation).

		Name/citation	Position repr.	Ref.	Learn.	Sim.	C.	Evaluation tasks
Symbolic	*Real-world*	Yeap et al. (2008) [based on ASRs by Yeap, 1988]	Boundary element triplets	$allo^{m}$	det	1	5	Mapping^{-} Homing^{-}
		Jefferies et al. (2008) [based on ASRs by Yeap, 1988]	Boundary element triplets	$allo^{m+t}$	det	1	5	Globalmapping^{-} Localmapping^{-}
		HSSH Beeson et al. (2010)	Occupancy grids, topological places & paths	$+^{m+t}$	SLAM	2	5	Global&localmaps^{-} Pathplanning^{-}
		Franz et al. (2008)	Disparity signatures	$+^{m+t}$	det	1	4	Mapping^{-}, homing^{-} Shortcuts^{-}
	Simulated	NAVIGATOR Gopal and Smith (1990)	Semantic networks (spatial + non-spatial info)	ego^{n}	det	1	1	Wayfinding errors^{q}
		perceptual wayfinding model Raubal (2001)	Topological graph (from fully observable env.)	$allo^{t}$	none	1	1	Wayfinding^{-}
		DP-model Brom et al. (2012)	Weighted ego + allo vectors	$+^{m}$	det	1	3	Pointing accuracy^{h}
Neural network-based	*Real-world*	Burgess et al. (2000) [later extended in simulation as the BVC model Barry et al., 2006]	Place & goal cells	$allo^{m}$	Hebbian Competitive	4 [5[a]]	4	Navigation$^{-[h]}$[a] Place fields$^{q[hn]}$[a]
		Strösslin et al. (2005) [based on Arleo & Gerstner, 2000]	'Step', place & HD cells	$allo^{m}$	Hebbian RL	4	4	Navigation^{-} Map learning^{-}
		Barrera et al. (2011) [based on Barrera & Weitzenfeld, 2008]	Place cells^{m} Actor units^{t}	$allo^{m+t}$	Hebbian RL	4	5	Navigation^{a} Map learning^{-}
	Simulated	Schölkopf and Mallot (1995)	Place specific SOM units	$allo^{t}$	Self-organized	3	1	Navigation^{-} Mapping (discrete maze)$^{-}$
		Voicu (2003) [extending Voicu & Schmajuk, 2000]	Hierarchical SOM units	$allo^{m}$	Self-organized	3	3	Map learning^{h} Distance judgments^{h}
		McNaughton et al. (1996) [evaluated by Samsonovich & McNaughton, 1997]	Place & HD cells	$allo^{m}$	Hebbian	5	2	Navigation & detour planning^{-} Path integration^{-} PF learning & stretching^{n}
		Byrne et al. (2007) [based on the BVC model Barry et al., 2006]	BVCs, place & HD cells	$+^{m}$	Hebbian	4	3	Slow rotation^{n} Mapping^{-}, PI^{-}, HD lesions^{q} Hemispheric neglect^{q}

(continued on next page)

Table 1 (*continued*)

	Name/citation	Position repr.	Ref.	Learn.	Sim.	C.	Evaluation tasks
Cog. architectures	Erdem and Hasselmo (2012) [based on Hasselmo, 2008]	Recency, topology, reward, Place, HD & grid cells	allo^{m+t}	Hebbian	5	3	Mapping^{-}, PI^{-}, navigation^{-} T-Maze, 'look-ahead'^{n}
	ACT-R/S^{s} Harrison et al. (2003)	Egocentric vectors in 'configural chunks'	ego^{m}	det (but prob. retrieval)	1	3	Map learning^{-}, PI^{-}, navigation^{-}
	Casimir^{h} Schultheis and Barkowsky (2011)	'Spatio-analogical fragments' (semantic network-like rep.)	$*^{m+t}$	det	1	3	Sp. reasoning^{h}, reinterpretation^{h} Mental scanning^{h}&recall effects^{h}
	CLARION^{h} Sun and Zhang (2004)	Chunks (explicit, symbolic) ANN (implicit)	ego^{m}	RL Backprop	3	3	Complex minefield navigation^{h}
	LIDA^{h} Madl et al. (2013)	'place nodes' in PAM	allo^{m}	det	3	3	Map learning^{-}, navigation (TSP)^{h} Map errors^{h}, WM capacity^{h}

[a] The later extension was substantiated against neural as well as behavior data (Barry et al., 2006), however, it was not implemented in a real-world robot.

Global mapping, i.e. correctly aligning multiple maps of local surroundings in the same reference frame, and loop closing, i.e. the problem of recognizing a place the agent has seen before (and correcting representation errors), are particularly difficult tasks in the real world. The main reason for this is that different places can look very similar (perceptual aliasing), and the same place can also look different at various times in dynamic environments. Only two of the reviewed models are able to perform both global mapping and loop closing in the real world (Beeson et al., 2010; Jefferies et al., 2008).

Looking at the structural similarities (which roughly translate to biological plausibility) and the environmental complexities in the table, it can be seen that in most cases there is a tradeoff between the two. Models with high biological realism (SNNs, e.g. McNaughton et al., 1996; or Erdem & Hasselmo, 2012) usually have trouble handling highly complex real-world environments (due mainly to their high computational demands, but also to the observation that it is easier to model high-level cognitive tasks such as planning with simpler – such as symbolic – models). In contrast, models built to work well on real-world robots (such as HSSH) usually cannot be called biologically realistic, and also have difficulties fitting human behavior data (due mainly to the abstractions and methodological shortcuts employed to quickly develop efficient algorithms that can tackle complex input, and also due to computational restrictions of robots).

It is very difficult to implement and run a model that incorporates both high psychological and biological plausibility and the ability to handle real-world environments. The model by Barrera et al. (2011) is notable because although it cannot close the loop and cannot perform global mapping, it can learn a real-world maze with a learning curve similar to rats, using a model that is highly structurally similar to rat brains.

The line of research attempting to implement real-world capable cognitive models can be expected to yield important insights in the cognitive sciences. First, because of the desirability of realistic input and output for accurate models of biological cognition (sticking to overly simplistic environments causes similar difficulties for a mechanistic understanding of cognition as studying spherical wooden balls or the solar system model would for nuclear physicists). Second, robotics and machine learning research has already provided significant insights and facilitated breakthroughs in cognitive neuroscience, and there is reason to believe it will continue to do so. Examples are the development of statistical methods to deal with sensory uncertainty (which later proved to help explain behavioral and neural data, starting the 'Bayesian brain' movement; Knill & Pouget, 2004), machine learning approaches for learning optimal action policies in unpredictable environments (reinforcement learning, which has contributed to understanding the neuroscientific study of conditioning; Maia, 2009), or dynamical systems and control theory (which have inspired dynamical systems approaches to cognition; Beer, 2000).

4.1. Open questions

It is interesting to note that the vast majority of the reviewed models incorporate allocentric representations (every reviewed real-world capable model does), and that a majority of the models capable of handling large-scale real-world environments represent both metric and topological spatial maps. The first point – the importance of allocentric spatial representations – has been known to cognitive science for many decades (Tolman, 1948). However, surprisingly little psychology and neuroscience research effort has been invested in identifying the mechanisms involved in topological mapping (for example, there is still no well-established neural correlate of topological maps in the brain—see Section 2; furthermore, the computational mechanism of how humans might partition space into topological maps is not well understood).

Models incorporating topological spatial representations such as the ones reviewed above might provide inspiration and insight for such research (unfortunately, none of them empirically validate their model with regard to topological mapping). Using empirically verified computational cognitive models to try out hypotheses regarding topological representation or the topology building mechanism in humans or animals would be an interesting and mostly unexplored line of research.

Along similar lines, it has long been suspected that the 'cognitive map' might be hierarchical (Derdikman & Moser, 2010; Hirtle & Jonides, 1985; McNamara, 1986), and multiple models incorporate hierarchies in their maps (such as HSSH, the model by Voicu, 2003, LIDA, and Casimir). Plausible neural correlates of hierarchical maps have also been identified in hippocampal and entorhinal cortical neurons with significantly varying firing field sizes (Derdikman & Moser, 2010). However, the mechanism which humans or animals use to cluster spatial representations into maps and sub-maps and organize them into a hierarchy is not yet understood (it is likely that the simple distance-based clustering mechanisms employed by most existing hierarchical models are insufficient to explain the error patterns caused by hierarchical maps; for example, perceptual or functional similarity almost certainly play a role in the mechanism organizing landmarks hierarchically in brains).

A further not fully understood part of spatial memory is the transformation process converting between egocentric and allocentric representations. Some of the reviewed models include both types of representations (Beeson et al., 2010; Brom et al., 2012; Byrne et al., 2007; Franz et al., 2008; Schultheis & Barkowsky, 2011). However, none of these models have evaluated their transformation mechanism against empirical data, with the exception of the neural network model by Byrne et al. (2007) (which seems to predict heavily coordinated and correlated activity in the neural correlates of transformation, i.e. the RSC; but such activity has been not observed).

A question that has yielded significant progress – but still no mature models explaining empirical data – regards the identification of 'landmarks' (how does the perceptual system identify landmarks, using them for orientation, as opposed to navigationally irrelevant stimuli?). Factors such as distance, stability, uniqueness, perceptual salience, and functional relevance seem to play a role. However, most existing spatial memory models either focus on localizing and navigating based on geometry, or are tested in sparse environments where a strategy of using every encountered object as a landmark is viable.

Finally, progress in the field of modeling spatial memory could be made by integrating the insights of individual models accounting for various phenomena (egocentric/allocentric, metric/topological, local/global, associative/reinforcement learning, geometric/landmark based, etc.) and tasks within the same model. Both the task of integrating these disparate processes, and evaluating them in a large number of tasks and settings, could yield new insights. Cognitive architectures would be in a uniquely suitable position to incorporate such an integration due to their generality and pre-existing non-spatial cognitive mechanisms.

4.2. Methods for verifying the biological plausibility of cognitive spatial memory models

The overview of Section 3 has outlined a number of qualitative and quantitative ways to evaluate computational models. In this section, we shall focus on describing recent methods for judging the biological plausibility of a model. Apart from qualitative evaluations of structural similarity to the underlying neurobiology

(such as the similarities in Table 1), it is also possible to empirically validate biological plausibility by comparing model predictions with neuroscientific data.

For biologically realistic neural network models, the most straightforward way of empirical verification is comparison with in-vivo electrophysiological single-unit recordings (in which microelectrodes are used to measure the action potentials of individual neurons in the brain of a live animal performing a task, preferably the same task in a similar environment as that of the model). For ANNs, a mapping function can be designed converting their numeric activation value to a spike rate; in the case of SNNs, the comparison is straightforward (spike trains or even voltage traces can be compared). The BVC model (Barry et al., 2006) is an example computational model successfully predicting the firing activity of spatially relevant neurons in single-unit recordings.

However, for most models, this is not viable; most often because they do not contain representations analogous to single biological neurons. In this case, higher-level brain-imaging data can be used for evaluation, which shows the time-dependent activity of brain areas involved in performing a task. Most frequently employed examples are fMRI (functional magnetic resonance imaging, a technique with high spatial but low temporal resolution) and EEG (electroencephalography, with low spatial but high temporal resolution). For models whose modules have been mapped to brain areas, it is possible to convert the activity of model parts into predicted brain area activations, and thus compare the model with neuroscientific data. Since the mapping function is arbitrary and does not place structural requirements on the underlying model, this procedure is possible even for models with little or no biological realism.

The ACT-R cognitive architecture is an example model that has used this approach successfully. ACT-R's major modules have been mapped to brain areas (such as the imaginal module to the posterior parietal region, or the central pattern matcher to the basal ganglia), and a suitable mapping function has been devised that converts activity in these modules into activation patterns resembling fMRI data (Qin et al., 2007), and more recently, EEG data (Motomura, Ojima, & Zhong, 2009), successfully predicting brain activity in novel circumstances (Anderson et al., 2008).

5. Conclusion

Having briefly summarized the basis of spatial memory in brains, we then reviewed a number of computational cognitive models of spatial memory, and presented a comparative table to help overview the major modeling directions taken within this large and highly fragmented topic. Although focusing on models concerned with human or spatial cognition, we have attempted to bring the fields of cognitive science, robotics, and neuroscience closer together by highlighting sources of overlap and interaction, and the modeling approaches most closely matched to each. We have pointed out what robotics and neuroscience can contribute to the field of cognitive modeling, and proposed some novel potential mappings between parts of existing models and relevant brain areas, in the hope of facilitating understanding, comparison, and evaluation. We have also outlined some open questions in the field, and how current (and future) models could address these questions. Computational cognitive modeling has much to offer spatial memory research (and cognitive science research in general), verifying existing hypotheses, yielding new ones, and guiding research.

Acknowledgments

We are grateful to Prof. Stan Franklin for his helpful comments. This work has been supported by EPSRC (Engineering and Physical Sciences Research Council) grant EP/I028099/1, and FWF (Austrian Science Fund) grant P25380-N23.

References

Allen, G. L. (2003). *Human spatial memory: remembering where*. Psychology Press.
Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*, 1036.
Anderson, J. R., Fincham, J. M., Qin, Y., & Stocco, A. (2008). A central circuit of the mind. *Trends in Cognitive Sciences*, *12*, 136–143.
Anderson, J. R., Matessa, M., & Lebiere, C. (1997). ACT-R: a theory of higher level cognition and its relation to visual attention. *Human–Computer Interactions*, *12*, 439–462.
Arleo, A., & Gerstner, W. (2000). Spatial cognition and neuro-mimetic navigation: a model of hippocampal place cell activity. *Biological Cybernetics*, *83*, 287–299.
Avraamides, M. N., & Kelly, J. W. (2008). Multiple systems of spatial memory and action. *Cognitive Processing*, *9*, 93–106.
Baars, B. J., & Franklin, S. (2009). Consciousness is computational: the LIDA model of global workspace theory. *International Journal of Machine Consciousness*, *1*, 23–32.
Baars, B. J., Franklin, S., & Ramsoy, T. Z. (2013). Global workspace dynamics: cortical 'binding and propagation' enables conscious contents. *Frontiers in Psychology*, *4*.
Baddeley, A. (1992). Working memory. *Science*, *255*, 556–559.
Bailey, T., & Durrant-Whyte, H. (2006). Simultaneous localization and mapping (SLAM): part II. *IEEE Robotics & Automation Magazine*, *13*, 108–117.
Barrera, A., Cáceres, A., Weitzenfeld, A., & Ramirez-Amaya, V. (2011). Comparative experimental studies on spatial memory and learning in rats and robots. *Journal of Intelligent & Robotic Systems*, *63*, 361–397.
Barrera, A., & Weitzenfeld, A. (2008). Biologically-inspired robot spatial cognition based on rat neurophysiological studies. *Autonomous Robots*, *25*, 147–169.
Barry, C., Lever, C., Hayman, R., Hartley, T., Burton, S., O'Keefe, J., et al. (2006). The boundary vector cell model of place cell firing and spatial memory. *Reviews in the Neurosciences*, *17*, 71–97.
Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*, 617–645.
Baumann, O., & Mattingley, J. B. (2010). Medial parietal cortex encodes perceived heading direction in humans. *Journal of Neuroscience*, *30*, 12897–12901.
Beer, R. D. (2000). Dynamical approaches to cognitive science. *Trends in Cognitive Sciences*, *4*, 91–99.
Beeson, P., Modayil, J., & Kuipers, B. (2010). Factoring the mapping problem: mobile robot map-building in the hybrid spatial semantic hierarchy. *The International Journal of Robotics Research*, *29*, 428–459.
Bhattacharyya, R., Musallam, S., & Andersen, R. A. (2009). Parietal reach region encodes reach depth using retinal disparity and vergence angle signals. *Journal of Neurophysiology*, *102*, 805–816.
Booij, O., Terwijn, B., Zivkovic, Z., & Krose, B. (2007). Navigation using an appearance based topological map. In *2007 IEEE international conference on robotics and automation* (pp. 3927–3932). IEEE.
Bringsjord, S. (2008). Declarative/logic-based computational cognitive modeling. In *The handbook of computational cognitive modeling*. Cambridge: Cambridge University Press.
Brom, C., Vyhnánek, J., Lukavský, J., Waller, D., & Kadlec, R. (2012). A computational model of the allocentric and egocentric spatial memory by means of virtual agents, or how simple virtual agents can help to build complex computational models. *Cognitive Systems Research*, *17–18*, 1–24.
Brown, M. W., & Aggleton, J. P. (2001). Recognition memory: what are the roles of the perirhinal cortex and hippocampus? *Nature Reviews Neuroscience*, *2*, 51–61.
Brown, E. N., Frank, L. M., Tang, D., Quirk, M. C., & Wilson, M. A. (1998). A statistical paradigm for neural spike train decoding applied to position prediction from ensemble firing patterns of rat hippocampal place cells. *The Journal of Neuroscience*, *18*, 7411–7425.
Burak, Y., & Fiete, I. R. (2009). Accurate path integration in continuous attractor network models of grid cells. *PLoS Computational Biology*, *5*, e1000291.
Burgess, N. (2008). Spatial cognition and the brain. *Annals of the New York Academy of Sciences*, *1124*, 77–97.
Burgess, N., Jackson, A., Hartley, T., & O'Keefe, J. (2000). Predictions derived from modelling the hippocampal role in navigation. *Biological Cybernetics*, *83*, 301–312.
Byrne, P., Becker, S., & Burgess, N. (2007). Remembering the past and imagining the future: a neural model of spatial memory and imagery. *Psychological Review*, *114*, 340.
Calton, J. L., & Taube, J. S. (2009). Where am i and how will i get there from here? A role for posterior parietal cortex in the integration of spatial information and route planning. *Neurobiology of Learning and Memory*, *91*, 186–196.
Cassimatis, N. L., Bello, P., & Langley, P. (2008). Ability, breadth, and parsimony in computational models of higher-order cognition. *Cognitive Science*, *32*, 1304–1322.
Chen, Z., Kloosterman, F., Brown, E. N., & Wilson, M. A. (2012). Uncovering spatial topology represented by rat hippocampal population neuronal codes. *Journal of Computational Neuroscience*, *33*, 227–255.
Cheng, K., Shettleworth, S. J., Huttenlocher, J., & Rieser, J. J. (2007). Bayesian integration of spatial information. *Psychological Bulletin*, *133*, 625–637.
Cheung, A., Ball, D., Milford, M., Wyeth, G., & Wiles, J. (2012). Maintaining a cognitive map in darkness: the need to fuse boundary knowledge with path integration. *PLoS Computational Biology*, *8*, e1002651.
Crowe, D. A., Averbeck, B. B., & Chafee, M. V. (2008). Neural ensemble decoding reveals a correlate of viewer-to object-centered spatial transformation in monkey parietal cortex. *The Journal of Neuroscience*, *28*, 5218–5228.

Dabaghian, Y., Cohn, A. G., & Frank, L. (2011). Topological coding in hippocampus. In *Computational modeling and simulation of intellect: current state and future prospectives* (pp. 293–320).

Dawson, M. R., & Shamanski, K. S. (1994). Connectionism, confusion and cognitive science. *Journal of Intelligent Systems*, *4*, 215–262.

Derdikman, D., & Moser, E. I. (2010). A manifold of spatial maps in the brain. *Trends in Cognitive Sciences*, *14*, 561–569.

Doeller, C. F., Barry, C., & Burgess, N. (2011). From cells to systems: grids and boundaries in spatial memory. *The Neuroscientist*,.

Duch, W., Oentaryo, R.J., & Pasquier, M. (2008). Cognitive architectures: where do we go from here? In *AGI, Volume 171* (pp. 122–136).

Duhamel, J.-R., Colby, C. L., & Goldberg, M. E. (1998). Ventral intraparietal area of the macaque: congruent visual and somatic response properties. *Journal of Neurophysiology*, *79*, 126–136.

Durrant-Whyte, H., & Bailey, T. (2006). Simultaneous localization and mapping: part I. *IEEE Robotics & Automation Magazine*, *13*, 99–110.

Ekstrom, A. D., Kahana, M. J., Caplan, J. B., Fields, T. A., Isham, E. A., Newman, E. L., et al. (2003). Cellular networks underlying human spatial navigation. *Nature*, *424*, 184–187.

Eliasmith, C. (2007). Attractor network. *Scholarpedia*, *2*(10), 1380.

Epstein, R. A. (2008). Parahippocampal and retrosplenial contributions to human spatial navigation. *Trends in Cognitive Sciences*, *12*, 388–396.

Erdem, U. M., & Hasselmo, M. (2012). A goal-directed spatial navigation model using forward trajectory planning based on grid cells. *European Journal of Neuroscience*, *35*, 916–931.

Etienne, A. S., Maurer, R., & Séguinot, V. (1996). Path integration in mammals and its interaction with visual landmarks. *Journal of Fish Biology*, *199*, 201–209.

Fox, C. W., & Prescott, T. J. (2010). Hippocampus as unitary coherent particle filter. In *IJCNN* (pp. 1–8). IEEE Press.

Franklin, S. (2003). IDA, a conscious artifact? *Journal of Consciousness Studies*, *10*, 4–5.

Franklin, S., Madl, T., D'Mello, S., & Snaider, J. (2014). LIDA: a systems-level architecture for cognition, emotion, and learning. *IEEE Transactions on Autonomous Mental Development*, *6*, 19–41.

Franz, M. O., & Mallot, H. A. (2000). Biomimetic robot navigation. *Robotics and Autonomous Systems*, *30*, 133–153.

Franz, M. O., Stürzl, W., Hübner, W., & Mallot, H. A. (2008). A robot system for biomimetic navigation—from snapshots to metric embeddings of view graphs. In *Robotics and cognitive approaches to spatial mapping* (pp. 297–314). Springer.

Fuster, J. M. (2002). Physiology of executive functions: the perception–action cycle. In D. T. Stuss, & R. T. Knight (Eds.), *Principles of frontal lobe function* (pp. 96–108). New York: Oxford University Press.

Gallistel, C. R. (2008). Dead reckoning, cognitive maps, animal navigation and the representation of space: an introduction. In *Robotics and cognitive approaches to spatial mapping* (pp. 137–143). Springer.

Ghosh-Dastidar, S., & Adeli, H. (2009). Spiking neural networks. *International Journal of Neural Systems*, *19*, 295–308.

Gibson, J. J. (1986). *The ecological approach to visual perception*. Psychology Press.

Godfrey-Smith, P. (2003). Theory and reality. *Science Education*, *88*, 236.

Goertzel, B., Lian, R., Arel, I., de Garis, H., & Chen, S. (2010). A world survey of artificial brain projects, part II: biologically inspired cognitive architectures. *Neurocomputing*, *74*, 30–49.

Gopal, S., Klatzky, R. L., & Smith, T. R. (1989). Navigator: a psychologically based model of environmental learning through navigation. *Journal of Environmental Psychology*, *9*, 309–331.

Gopal, S., & Smith, T. (1990). Human way-finding in an urban environment: a performance analysis of a computational process model. *Environment and Planning A*, *22*, 169–191.

Gorchetchnikov, A., & Hasselmo, M. (2005). A biophysical implementation of a bidirectional graph search algorithm to solve multiple goal navigation tasks. *Connection Science*, *17*, 145–164.

Graham, P., & Collett, T. S. (2002). View-based navigation in insects: how wood ants (Formica rufa L.) look at and are guided by extended landmarks. *Journal of Experimental Biology*, *205*, 2499–2509.

Grossberg, S. (1987). Competitive learning: from interactive activation to adaptive resonance. *Cognitive Science*, *11*, 23–63.

Gupta, K., Erdem, U., & Hasselmo, M. (2013). Modeling of grid cell activity demonstrates *in vivo* entorhinal 'look-ahead' properties. *Neuroscience*, *247*, 395–411.

Hafting, T., Fyhn, M., Molden, S., Moser, M., & Moser, E. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature*, *436*, 801–806.

Harrison, A.M., & Schunn, C.D. et al. (2003). ACT-R/S: look ma, no 'cognitive-map'. In *International conference on cognitive modeling* (pp. 129–134).

Hartley, T., Burgess, N., Lever, C., Cacucci, F., & O'Keefe, J. (2000). Modeling place fields in terms of the cortical inputs to the hippocampus. *Hippocampus*, *10*, 369–379.

Hartley, T., Maguire, E. A., Spiers, H. J., & Burgess, N. (2003). The well-worn route and the path less traveled: distinct neural bases of route following and wayfinding in humans. *Neuron*, *37*, 877–888.

Hasselmo, M. E. (2008). Grid cell mechanisms and function: contributions of entorhinal persistent spiking and phase resetting. *Hippocampus*, *18*, 1213–1229.

Hirtle, S., & Jonides, J. (1985). Evidence of hierarchies in cognitive maps. *Memory & Cognition*, *13*, 208–217.

Hok, V., Save, E., Lenck-Santini, P., & Poucet, B. (2005). Coding for spatial goals in the prelimbic/infralimbic area of the rat frontal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 4602–4607.

Holmes, M. C., & Sholl, M. J. (2005). Allocentric coding of object-to-object relations in overlearned and novel environments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 1069.

Husain, M. (2008). Hemineglect. *Scholarpedia*, *3*(2), 3681.

Jain, A. K., Mao, J., & Mohiuddin, K. M. (1996). Artificial neural networks: a tutorial. *IEEE Computer*, *29*, 31–44.

Jefferies, M., Baker, J., & Weng, W. (2008). Robot cognitive mapping—a role for a global metric map in a cognitive mapping process. In *Robotics and cognitive approaches to spatial mapping* (pp. 265–279).

Jefferies, M., & Yeap, W. (2008). *Robotics and cognitive approaches to spatial mapping. Vol. 38*. Springer Verlag.

Jensen, O., & Lisman, J. E. (2000). Position reconstruction from an ensemble of hippocampal place cells: contribution of theta phase coding. *Journal of Neurophysiology*, *83*, 2602–2609.

Kaski, S., & Kohonen, T. (1994). Winner-take-all networks for physiological models of competitive learning. *Neural Networks*, *7*, 973–984.

Kim, J., Delcasso, S., & Lee, I. (2011). Neural correlates of object-in-place learning in hippocampus and prefrontal cortex. *The Journal of Neuroscience*, *31*, 16991–17006.

Kjelstrup, K. B., Solstad, T., Brun, V. H., Hafting, T., Leutgeb, S., Witter, M. P., et al. (2008). Finite scale of spatial representation in the hippocampus. *Science*, *321*, 140–143.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*, 712–719.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*, 1464–1480.

Kravitz, D. J., Saleem, K. S., Baker, C. I., & Mishkin, M. (2011). A new neural framework for visuospatial processing. *Nature Reviews Neuroscience*, *12*, 217–230.

Kuipers, B. (2000). The spatial semantic hierarchy. *Artificial Intelligence*, *119*, 191–233.

Kuipers, B. (2008). An intellectual history of the spatial semantic hierarchy. In *Robotics and cognitive approaches to spatial mapping* (pp. 243–264). Springer.

Lever, C., Burton, S., Jeewajee, A., O'Keefe, J., & Burgess, N. (2009). Boundary vector cells in the subiculum of the hippocampal formation. *Journal of Neuroscience*, *29*, 9771–9777.

Madl, T., Franklin, S., Chen, K., Montaldi, D., & Trappl, R. (2014). Bayesian integration of information in hippocampal place cells. *PLoS One*, e89762.

Madl, T., Franklin, S., Chen, K., & Trappl, R. (2013). Spatial working memory in the LIDA cognitive architecture. In *Proc. international conference on cognitive modelling* (pp. 384–390).

Maia, T. V. (2009). Reinforcement learning, conditioning, and the brain: successes and challenges. *Cognitive, Affective, & Behavioral Neuroscience*, *9*, 343–364.

Mani, K., & Johnson-Laird, P. N. (1982). The mental representation of spatial descriptions. *Memory & Cognition*, *10*, 181–187.

Manns, J. R., & Eichenbaum, H. (2009). A cognitive map for object memory in the hippocampus. *Learning & Memory*, *16*, 616–624.

Mark, D. M., Freksa, C., Hirtle, S. C., Lloyd, R., & Tversky, B. (1999). Cognitive models of geographical space. *International Journal of Geographical Information Science*, *13*, 747–774.

Martin, A., & Chao, L. L. (2001). Semantic memory and the brain: structure and processes. *Current Opinion in Neurobiology*, *11*, 194–201.

Marzocchi, N., Breveglieri, R., Galletti, C., & Fattori, P. (2008). Reaching activity in parietal area V6A of macaque: eye influence on arm activity or retinocentric coding of reaching movements? *European Journal of Neuroscience*, *27*, 775–789.

McClelland, J. L. (2009). The place of modeling in cognitive science. *Topics in Cognitive Science*, *1*, 11–38.

McNamara, T. P. (1986). Mental representations of spatial relations. *Cognitive Psychology*, *18*, 87–121.

McNaughton, B., Barnes, C., Gerrard, J., Gothard, K., Jung, M., Knierim, J., et al. (1996). Deciphering the hippocampal polyglot: the hippocampus as a path integration system. *Journal of Fish Biology*, *199*, 173–185.

McNaughton, B. L., Battaglia, F. P., Jensen, O., Moser, E. I., & Moser, M.-B. (2006). Path integration and the neural basis of the 'cognitive map'. *Nature Reviews Neuroscience*, *7*, 663–678.

Milford, M., & Wyeth, G. (2010). Persistent navigation and mapping using a biologically inspired SLAM system. *The International Journal of Robotics Research*, *29*, 1131–1153.

Moser, E. I., Kropff, E., & Moser, M.-B. (2008). Place cells, grid cells, and the brain's spatial representation system. *Annual Review of Neuroscience*, *31*, 69–89.

Motomura, S., Ojima, Y., & Zhong, N. (2009). EEG/ERP meets ACT-R: a case study for investigating human computation mechanism. In *Brain informatics* (pp. 63–73). Springer.

Myung, I. J., Pitt, M. A., & Kim, W. (2005). Model evaluation, testing and selection. In *Handbook of cognition* (pp. 422–436).

Nardini, M., Jones, P., Bedford, R., & Braddick, O. (2008). Development of cue integration in human navigation. *Current Biology*, *18*, 689–693.

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: symbols and search. *Communications of the ACM*, *19*, 113–126.

O'Doherty, J. P. (2004). Reward representations and reward-related learning in the human brain: insights from neuroimaging. *Current Opinion in Neurobiology*, *14*, 769–776.

O'Keefe, J., & Burgess, N. (1996). Geometric determinants of the place fields of hippocampal neurons. *Nature*, *381*, 425–428.

O'Keefe, J., & Dostrovsky, J. (1971). The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Research*, *34*, 171–175.

O'Reilly, R. C. (1998). Six principles for biologically based computational models of cortical cognition. *Trends in Cognitive Sciences*, *2*, 455–462.

Packard, M. G., & McGaugh, J. L. (1996). Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects expression of place and response learning. *Neurobiology of Learning and Memory*, *65*, 65–72.

Pavlidis, T., & Horowitz, S. L. (1974). Segmentation of plane curves. *IEEE Transactions on Computers*, *23*, 860–870.

Pesaran, B., Nelson, M. J., & Andersen, R. A. (2006). Dorsal premotor neurons encode the relative position of the hand, eye, and goal during reach planning. *Neuron*, *51*, 125–134.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*, 472.

Plank, M. (2009). *Behavioral, electrocortical and neuroanatomical correlates of egocentric and allocentric reference frames during visual path integration*. (Ph.D. thesis), Ludwig-Maximilans-Universität München.

Poucet, B. (1993). Spatial cognitive maps in animals: new hypotheses on their structure and neural mechanisms. *Psychological Review*, *100*, 163.

Qin, Y., Bothell, D., & Anderson, J. R. (2007). ACT-R meets fMRI. In *Web intelligence meets brain informatics* (pp. 205–222). Springer.

Raubal, M. (2001). Human wayfinding in unfamiliar buildings: a simulation with a cognizing agent. *Cognitive Processing*, *2*, 363–388.

Rolls, E. T., & Xiang, J.-Z. (2006). Spatial view cells in the primate hippocampus and memory recall. *Reviews in the Neurosciences*, *17*, 175–200.

Rumelhart, D. E., & Zipser, D. (1985). Feature discovery by competitive learning*. *Cognitive Science*, *9*, 75–112.

Samsonovich, A. V. (2010). Toward a unified catalog of implemented cognitive architectures. In *BICA, 221* (pp. 195–244).

Samsonovich, A., & McNaughton, B. L. (1997). Path integration and cognitive mapping in a continuous attractor neural network model. *The Journal of Neuroscience*, *17*, 5900–5920.

Schölkopf, B., & Mallot, H. A. (1995). View-based cognitive mapping and path planning. *Adaptive Behavior*, *3*, 311–348.

Schultheis, H., & Barkowsky, T. (2011). Casimir: an architecture for mental spatial knowledge processing. *Topics in Cognitive Science*, *3*, 778–795.

Schultheis, H., Lile, S., & Barkowsky, T. (2007). Extending ACT-R's memory capabilities. In *Proc. of EuroCogSci, Vol. 7* (pp. 758–763).

Shepard, R., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science (New York, NY)*, *171*, 701.

Sima, J. F. (2011). The nature of mental images—an integrative computational theory. In *Proceedings of the 33rd annual conference of the cognitive science society* (pp. 2878–2883). Citeseer.

Sima, J. F., Lindner, M., Schultheis, H., & Barkowsky, T. (2010). Eye movements reflect reasoning with mental images but not with mental models in orientation knowledge tasks. In *Spatial cognition VII* (pp. 248–261). Springer.

Sloman, A. (1998). What sort of architecture is required for a human-like agent? In Charles Ling, & Ron Sun (Eds.), *Cognitive modeling workshop, at AAAI 1998* (pp. 1–8). AAAI.

Smolensky, P. (1987). Connectionist AI, symbolic AI, and the brain. *Artificial Intelligence Review*, *1*, 95–109.

Snyder, L. H., Grieve, K. L., Brotchie, P., & Andersen, R. A. (1998). Separate body-and world-referenced representations of visual space in parietal cortex. *Nature*, *394*, 887–891.

Solstad, T., Boccara, C. N., Kropff, E., Moser, M.-B., & Moser, E. I. (2008). Representation of geometric borders in the entorhinal cortex. *Science*, *322*, 1865–1868.

Solstad, T., Moser, E. I., & Einevoll, G. T. (2006). From grid cells to place cells: a mathematical model. *Hippocampus*, *1031*, 1026–1031.

Song, S., Miller, K. D., & Abbott, L. F. (2000). Competitive hebbian learning through spike-timing-dependent synaptic plasticity. *Nature Neuroscience*, *3*, 919–926.

Strösslin, T., Sheynikhovich, D., Chavarriaga, R., & Gerstner, W. (2005). Robust self-localisation and navigation based on hippocampal place cells. *Neural Networks*, *18*, 1125–1140.

Sun, R. (2007). The importance of cognitive architectures: an analysis based on clarion. *Journal of Experimental & Theoretical Artificial Intelligence*, *19*, 159–193.

Sun, R. (2008a). *The Cambridge handbook of computational psychology*. Cambridge: Cambridge University Press.

Sun, R. (2008b). Introduction to computational cognitive modeling. In *Cambridge handbook of computational psychology* (pp. 3–19).

Sun, R., Merrill, E., & Peterson, T. (2001). From implicit skills to explicit knowledge: a bottom–up model of skill learning. *Cognitive Science*, *25*, 203–244.

Sun, R., & Zhang, X. (2004). Top–down versus bottom–up learning in cognitive skill acquisition. *Cognitive Systems Research*, *5*, 63–89.

Taube, J. S. (2007). The head direction signal: origins and sensory-motor integration. *Annual Review of Neuroscience*, *30*, 181–207.

Thomas, M. S., & McClelland, J. L. (2008). Connectionist models of cognition. In *The Cambridge handbook of computational psychology* (pp. 23–58).

Thrun, S., & Leonard, J. J. (2008). Simultaneous localization and mapping. In *Springer handbook of robotics* (pp. 871–889).

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*, 189.

Tommasi, L., Chiandetti, C., Pecchia, T., Sovrano, V. A., & Vallortigara, G. (2012). From natural geometry to spatial cognition. *Neuroscience & Biobehavioral Reviews*, *36*, 799–824.

Tommasi, L., & Laeng, B. (2012). Psychology of spatial cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *3*, 565–580.

Trullier, O., Wiener, S. I., Berthoz, A., & Meyer, J.-A. (1997). Biologically based artificial navigation systems: review and prospects. *Progress in Neurobiology*, *51*, 483–544.

Tversky, B. (2005). Functional significance of visuospatial representations. In *Handbook of higher-level visuospatial thinking* (pp. 1–34).

Vann, S. D., Aggleton, J. P., & Maguire, E. A. (2009). What does the retrosplenial cortex do? *Nature Reviews Neuroscience*, *10*, 792–802.

Vogeley, K., May, M., Ritzl, A., Falkai, P., Zilles, K., & Fink, G. R. (2004). Neural correlates of first-person perspective as one constituent of human self-consciousness. *Journal of Cognitive Neuroscience*, *16*, 817–827.

Voicu, H. (2003). Hierarchical cognitive maps. *Neural Networks*, *16*, 569–576.

Voicu, H., & Schmajuk, N. (2000). Exploration, navigation and cognitive mapping. *Adaptive Behavior*, *8*, 207–223.

Waller, D., & Hodgson, E. (2006). Transient and enduring spatial representations under disorientation and self-rotation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 867.

Webb, B. (2000). What does robotics offer animal behaviour? *Animal Behaviour*, *60*, 545–558.

Webb, B. (2001). Can robots make good models of biological behaviour? *Behavioral and Brain Sciences*, *24*, 1033–1050.

Willshaw, D. J., & Von Der Malsburg, C. (1976). How patterned neural connections can be set up by self-organization. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, *194*, 431–445.

Woergoetter, F., & Porr, B. (2008). Reinforcement learning. *Scholarpedia*, *3*, 1448.

Yeap, W. K. (1988). Towards a computational theory of cognitive maps. *Artificial Intelligence*, *34*, 297–360.

Yeap, W. K., Wong, C. K., & Schmidt, J. (2008). Using a mobile robot to test a theory of cognitive mapping. In *Robotics and cognitive approaches to spatial mapping* (pp. 281–295). Springer.

Yonelinas, A. P., Otten, L. J., Shaw, K. N., & Rugg, M. D. (2005). Separating the brain regions involved in recollection and familiarity in recognition memory. *The Journal of Neuroscience*, *25*, 3002–3008.

Yoon, K., Buice, M. A., Barry, C., Hayman, R., Burgess, N., & Fiete, I. R. (2013). Specific evidence of low-dimensional continuous attractor dynamics in grid cells. *Nature Neuroscience*, *16*, 1077–1084.

Zaehle, T., Jordan, K., Wüstenberg, T., Baudewig, J., Dechent, P., & Mast, F. W. (2007). The neural basis of the egocentric and allocentric spatial frame of reference. *Brain Research*, *1137*, 92–103.

Chapter 4

Bayesian integration of information in hippocampal place cells

Publication 2 / 4. Madl T., Franklin S., Chen K., Montaldi D. & Trappl R., 2014. Bayesian Integration of Information in Hippocampal Place Cells. *PLoS ONE 9(3), e89762*

*N*ote: the manuscript was originally published with incorrect figure ordering. The correct figure order was published as a correction (doi: 10.1371/journal.pone. 0136128), but PLOS has decided to maintain the old manuscript with incorrect ordering online. The reprint below contains the corrected figure order. No other changes have been made to the online version.

OPEN ACCESS Freely available online

PLOS ONE

Bayesian Integration of Information in Hippocampal Place Cells

Tamas Madl[1,4]*, Stan Franklin[2], Ke Chen[1], Daniela Montaldi[3], Robert Trappl[4]

1 School of Computer Science, University of Manchester, Manchester, United Kingdom, **2** Institute for Intelligent Systems, University of Memphis, Memphis, Tennessee, United States of America, **3** School of Psychological Sciences, University of Manchester, Manchester, United Kingdom, **4** Austrian Research Institute for Artificial Intelligence, Vienna, Austria

Abstract

Accurate spatial localization requires a mechanism that corrects for errors, which might arise from inaccurate sensory information or neuronal noise. In this paper, we propose that Hippocampal place cells might implement such an error correction mechanism by integrating different sources of information in an approximately Bayes-optimal fashion. We compare the predictions of our model with physiological data from rats. Our results suggest that useful predictions regarding the firing fields of place cells can be made based on a single underlying principle, Bayesian cue integration, and that such predictions are possible using a remarkably small number of model parameters.

Citation: Madl T, Franklin S, Chen K, Montaldi D, Trappl R (2014) Bayesian Integration of Information in Hippocampal Place Cells. PLoS ONE 9(3): e89762. doi:10.1371/journal.pone.0089762

Editor: Gareth Robert Barnes, University College of London - Institute of Neurology, United Kingdom

Received May 21, 2013; **Accepted** January 24, 2014; **Published** March 6, 2014

Funding: This research was supported by EPSRC grant EP/I028099/1 (Engineering and Physical Sciences Research Council, http://www.epsrc.ac.uk), and FWF grant P25380-N23 (Austrian Science Fund, http://www.fwf.ac.at/en). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: tamas.madl@gmail.com

Introduction

For successful navigation, an organism needs to be able to localize itself (i.e. determine its position and orientation) as well as its goal, and it needs to be able to calculate a route between these locations. Since the first reports of physiological evidence for hippocampal 'place cells' [1] which exhibit increased firing only in specific locations in the environment, there have been a large number of empirical findings supporting the idea that the Hippocampal-Entorhinal Complex (HEC) is a major neuronal correlate underlying spatial localization and mapping [2].

To keep track of their location when they move, mammals must integrate self-motion signals, and use them to update their location estimate, using a process commonly referred to as path integration or dead reckoning. It has been suggested that self-motion information might be the primary constituent in the formation of the firing fields of place cells [3,4]. However, path integration alone is prone to accumulating errors (arising from the inaccuracy of sensory inputs and neuronal noise), which add up over time until the location estimate becomes too inaccurate to allow for efficient navigation [5,6]. Because path integration errors are cumulative, path integrators have to be corrected using allothetic sensory information from the environment in order to ensure that the estimated location will stay close to the true location.

It has also been suggested that place cells rely heavily on visual information [1,2,7]. However, the question of how exactly different sources of information are combined, from different boundaries or landmarks, has received little attention in the literature. This paper investigates how place cells in the Hippocampus might integrate information to provide an accurate location estimate. We propose that the integration of cues from different sources might occur in an approximately Bayesian fashion; i.e. that the information is weighted according to its accuracy when combined with a final estimate, with more precise information receiving a higher importance weight. We provide supporting evidence and theoretical arguments for this claim in the Results section. We will compare neuronal recordings of place cells with predictions of a Bayesian model, and present a possible explanation for how approximate Bayesian inference, although insufficient to fully explain firing fields, might provide a useful framework within which to understand cue integration. Finally, we will present a possible model of how Bayesian inference might be implemented at the neuronal level in the hippocampus.

Our results are consistent with the 'Bayesian brain hypothesis' [8]; the idea that the brain integrates information in a statistically optimal fashion. There is increasing behavioural evidence for Bayesian informational integration for different modalities, e.g. for visual and haptic [9], for force [10], but also for spatial information, e.g. [11] (see Discussion). Other models of statistically optimal or near-optimal spatial cue integration have been proposed previously [11–14], although mostly at Marr's computational or algorithmic level, rather than at a physical level. The latter, mechanistic Bayesian view, has been cautioned against due to lacking evidence on the single neuron level [15]. Our results partially account for three disparate single-cell electrophysiological data sets using a Bayesian framework, and suggest that although such models might be too simple to fully explain patterns of neuronal firing, they will still be highly valuable to our understanding of the relationship between neuronal activity and the environment.

Neuronal correlates of localization

Here we briefly summarize the neuroscientific literature concerning how mammalian brains represent space. Most of these results come from animal (rat, and to a lesser extent, monkey) cellular recording studies, although there is some recent evidence substantiating the existence of these cell types in humans.

Four types of cells play an important role for allocentric spatial representations in mammalian brains:

1. **Grid cells** in the medial entorhinal cortex show increased firing at multiple locations, regularly positioned in a grid across the environment consisting of equilateral triangles [16]. Grids from neighbouring cells share the same orientation, but have different and randomly distributed offsets, meaning that a small number of them can cover an entire environment. It has also been suggested that grid cells play a major role in path integration, their activation being updated depending on the animal's movement speed and direction [2,16–18]. There is evidence to suggest that they exist not only in mammals, but also in the human entorhinal cortex (EC) [19].
2. **Head-direction cells** fire whenever the animal's head is pointing in a certain direction. The primary circuit responsible for head direction signals projects from the dorsal tegmental nucleus to the lateral mammillary nucleus, anterior thalamus and postsubiculum, terminating in the entorhinal cortex [20]. There is evidence that head direction cells exist in the human brain within the medial parietal cortex [21].
3. **Border cells** and **boundary vector cells** (BVCs), which are cells with boundary related firing properties. The former [22,23] seem to fire in proximity to environment boundaries, while the firing of the latter [2,24] depends on boundary proximity as well as direction relative to the mammal's head. Cells with these properties have been found in the mammalian subiculum and entorhinal cortex [22,23], and there is also some behavioural evidence substantiating their existence in humans [24].
4. **Place cells** are pyramidal cells in the hippocampus which exhibit strongly increased firing in specific spatial locations, largely independent from orientation in open environments [2,25], thus providing a representation of an animal's (or human's [26]) location in the environment. A possible explanation for the formation of place fields (the areas of the environment in which place cells show increased firing) is that they emerge from a combination of grid cell inputs on different scales [3,4]. It has also been proposed that place fields might be mainly driven by environmental geometry, arising from a sum of boundary vector cell inputs [7,24]. This model has successfully accounted for a number of empirical observations, e.g. the effects of environment deformations [7], or of inserting a barrier into an environment, on place fields [24].

Hippocampal place cells play a prominent role in navigation, the association of episodic memories with places, and other important spatio-cognitive functions, which might be impaired if their place fields were inaccurate. However, neither of the outlined place field models fully explain how place cells combine different inputs for accurate localization. The grid cell input model is subject to corruption of the location estimate by accumulating errors which would eventually render the estimate useless unless corrected by observations (see Introduction). On the other hand, boundary vectors alone (if driven solely by geometry, not by features) do not always yield unambiguous location estimates [14]. Even given complex visual information (which border-related cells do not seem to respond to [22], and of which a rat might not see much, given its poor visual acuity [27]), localization without path integration is difficult (localization without odometry was solved in robotics only recently, and is still much more error-prone than combining observations with odometry [28]). For many place cells, both the path integration inputs from grid cells and observation inputs from border-related cells (and possibly others) seem to be required in order to ensure accuracy and certainty. This has been pointed out before (e.g. [29]), but the question of how exactly these inputs are combined has received little attention (but see the Discussion section for related work).

A further, as of yet unanswered, question is how exactly information from different sources (boundaries, landmarks, different senses etc.) might be combined. Although the BVC model made detailed predictions as to the kinds of inputs received by place cells, was fitted successfully to electrophysiological data, and matched empirical observations (such as what happens with place fields on barrier insertion), it does not propose a general principle of cue integration. In order for the model to accurately reflect place field location and size in a given environment, a number of weight and tuning parameters have to be adjusted for every single place cell [7,24]. In contrast, the Bayesian hypothesis that we investigate in this work implies a general underlying principle for how inputs into place cells are weighted; according to their precision and with more accurate inputs influencing the result stronger than less accurate inputs. The biggest advantage of such a general principle is that it significantly reduces the number of parameters required to account for large datasets (see Results).

Please note that we adopt a highly simplified and constrained view of HEC function and anatomy in this paper. Hippocampal cells play a role in many cognitive functions other than spatial localization; among others long-term episodic/declarative memory [30,31], memory based prediction [32], and possibly short-term memory [33] and perception [34]. Furthermore, place cells receive a broader array of inputs than just those transmitting visual and path integration information, such as odours and tactile information [35]. Finally, while cells from different parts of the hippocampus differ in their connectivity and in the information they receive, we believe that dealing with a small subset of functionality and anatomy suffices for investigating the existence of statistically near-optimal information integration in place cells.

Hypotheses

In this paper, we describe a Bayesian mechanism of information integration in place cells accounting for place field formation. This mechanism rests on the following hypotheses:

H1. Some Hippocampal place cells perform approximate Bayesian cue integration - they combine different sources of information in an approximately Bayes-optimal fashion, weighting inputs according to their precision. This means that when sensory inputs change, some place fields should shift and resize in a manner predictable by a Bayesian model.

H2. A Bayesian view requires that HEC neurons encode a mammal's uncertainty regarding its position, in addition to its actual location. We hypothesize that the sizes of place cell firing fields are correlated with this location uncertainty.

H3. The uncertainty of distance measurements to borders σ_b depends on the boundary distance d_b, and can be approximated by a linear relationship using some constant s (cf. Weber's law): $\sigma_b = s \cdot d_b$. There is some physiological evidence for this in border-related cells [22,23], as well as some behavioural evidence that Weber's law holds for spatial distance perception in rats [36] and mammals [37]. That the tuning breadths of BVCs should increase with distance is also a prerequisite of the Boundary Vector Cell

model [7,24], has been successfully fit to neuronal and behavioural data, and is supported by physiological evidence [22].

These hypotheses are interdependent, and will be investigated together. To generate verifiable predictions from the Bayesian hypothesis (H1) we need to assume how uncertainty is represented (H2) and how it can be derived from the geometry of the environment (H3). Together, these hypotheses allow the making of predictions about the sizes of place cell firing fields, given the distances of all boundaries, in some cases using just a single parameter specifying how uncertainty depends on distance. The Bayesian mechanism attempts to account for the sizes of single firing fields, deriving them from the distances of boundaries or obstacles (H3) - thus, place cells with multiple firing fields can be modelled by dealing with each firing field separately, even under Gaussian assumptions. In the Discussion section, we briefly describe how the model could be extended by relaxing some of its assumptions, and we report applications of the extended model in the Results section. We do not claim that place cells implement any statistical equation (especially not the simplistic ones described here), but we propose that investigating their firing fields within a statistical framework can yield useful insights about the way they combine information.

Methods

The hypothesis of approximate Bayesian integration of information in place cells (H1) yields verifiable electrophysiological predictions. Since we hypothesized that place cells can perform approximate Bayesian cue integration (H1), and place field sizes are correlated with uncertainty (H2), and that uncertainty depends on distance (H3), expected place field sizes can be predicted from the geometry of an environment using a Bayesian model. This section will outline such a Bayesian model.

Model assumptions

To simplify the mathematics, and because this assumption fits our data well, we will assume elliptical firing fields shaped like two-dimensional Gaussians. We do not claim that place cells encode exact Gaussian distributions (there are also asymmetric place fields in the hippocampus - see the Discussion for potential extensions of this simple model). However, investigating their firing fields in a Bayesian framework can yield useful insights about cue integration. The predictions in the Results section are generated from Bayesian models using Gaussian probability distributions to represent locations, in simplified two-dimensional environments, with sizes and distances adjusted to those of the respective in-vivo experiments.

Bayesian spatial cue integration

Bayesian inference under Gaussian assumptions implies that information from different observations should be weighted according to its accuracy. This claim can be formalized using Bayes' rule, according to which the probability distribution of the location given a number of observations can be calculated from

$$p(\boldsymbol{x}|\boldsymbol{O}) \propto p(\boldsymbol{x})p(\boldsymbol{O}|\boldsymbol{x}) \tag{1}$$

where $\boldsymbol{x}$ is the animals location in the environment and $\boldsymbol{O}=\{o_1,\cdots,o_N\}$ represents a set of N observations. $p(\boldsymbol{x}|\boldsymbol{O})$ is the posterior location belief, given all observations. $p(\boldsymbol{x})$ is a prior belief over the location (for example via path integration), and $p(\boldsymbol{O}|\boldsymbol{x})$ represents the probability of the current observations given $\boldsymbol{x}$ (such as boundaries or landmarks), characterized by the distance from $\boldsymbol{x}$ and their uncertainty (see below). Since observations can be assumed to be conditionally independent given the location (this is an assumption commonly made in robotics, see [38,39]), we can expand equation (1) to

$$p(\boldsymbol{x}|\boldsymbol{O}) \propto p(\boldsymbol{x}) \prod_{i=1}^{N} p(\boldsymbol{o}_i|\boldsymbol{x}). \tag{2}$$

In this simplified model, the probability distributions are assumed to be Gaussian. Thus, for multiple spatial dimensions, equation (2) can be written as

$$\mathcal{N}(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}}) = \gamma \mathcal{N}(\boldsymbol{\mu}_p,\boldsymbol{\Sigma}_p) \prod_{i=1}^{N} \mathcal{N}(\boldsymbol{\mu}_{o,i},\boldsymbol{\Sigma}_{o,i}). \tag{3}$$

In the case of a single spatial dimension, and in environments where spatial dimensions can be assumed to be independent and thus can be considered separately, equation (2) can be written as

$$\mathcal{N}(\hat{\mu}_x,\hat{\sigma}_x) = \gamma \mathcal{N}(\mu_p,\sigma_p) \prod_{i=1}^{N} \mathcal{N}(\mu_{o,i},\sigma_{o,i}). \tag{4}$$

Here, $\hat{\boldsymbol{\mu}}$ (or $\hat{\mu}_x$ in one dimension) is the mean of the posterior or the 'best guess' location, $\hat{\boldsymbol{\Sigma}}$ (or $\hat{\sigma}_x$ in one dimension) the uncertainty (covariance, or standard deviation) associated with this location, $\boldsymbol{\mu}_p$ (or μ_p) and $\boldsymbol{\Sigma}_p$ (or σ_p) are the mean and the uncertainty of the prior belief location, $\boldsymbol{\mu}_{o,i}$ (or $\mu_{o,i}$) and $\boldsymbol{\Sigma}_{o,i}$ (or $\sigma_{o,i}$) are the means and uncertainties of the individual observations, and γ is a constant for normalization.

Calculating the uncertainty $\hat{\sigma}_x$ (standard deviation) in one spatial dimension is sometimes sufficient in environments in which the width is negligible compared to the length (such as the first two environments in the Results section: the linear track in Figure 1, and the circular track in Figure 2). In the rectangular environments of Figure 3, the x and y dimensions were assumed to be independent, and the uncertainties were calculated independently - which is a reasonable approximation for this particular dataset, since the observations (the walls of the environment) were orthogonal. However, for more complex environments, the covariances $\hat{\boldsymbol{\Sigma}}$ would have to be calculated from equation (3) instead of individually calculating the standard deviations in each dimension (see Text S1 in the Supporting Information for the derivation of the covariance matrix from distance measurements, for two-dimensional environments in which the dimensions cannot be assumed to be independent).

In the one-dimensional case, solving equation (4) for the standard deviations (see [40] for the derivation of the standard deviation of a product of Gaussians), we can calculate the uncertainty associated with the 'best guess' location, $\hat{\sigma}_x$, which for a single observation is

$$\hat{\sigma}_x = \sqrt{\frac{\sigma_p^2\sigma_o^2}{\sigma_p^2+\sigma_o^2}} = \sqrt{\left(\frac{1}{\sigma_p^2}+\frac{1}{\sigma_o^2}\right)^{-1}}. \tag{5}$$

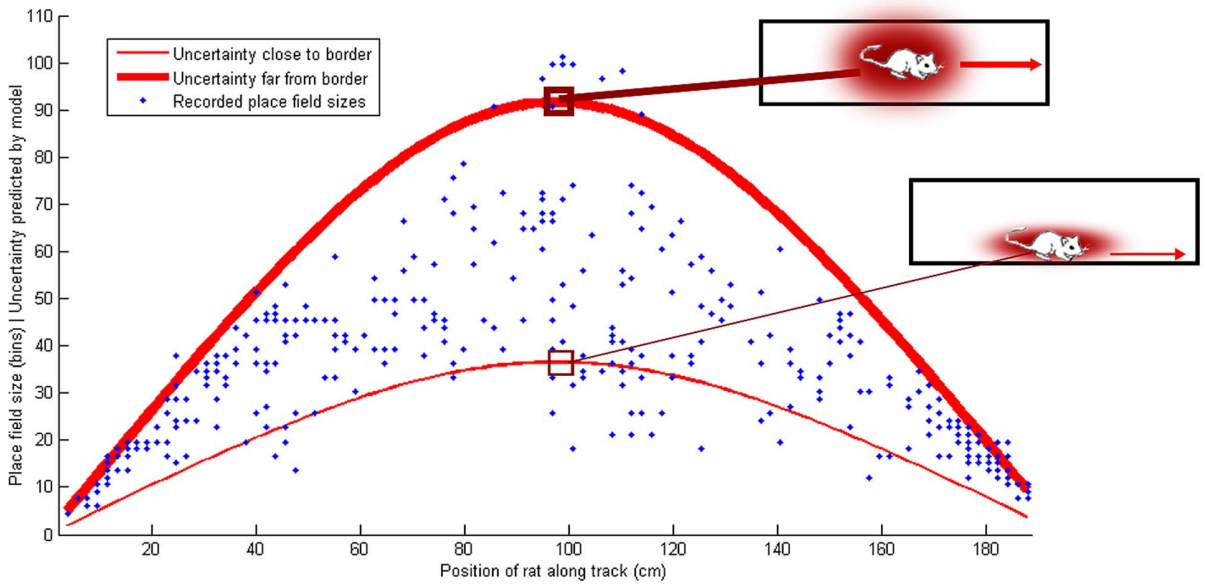


Figure 1. Place field sizes, and predicted uncertainty, on an empty rectangular track. The blue dots show the sizes of individual place fields in bins (one bin equals 1.9 cm). The red lines show the location uncertainty predicted by the Bayesian model - the thin red line (bottom) represents a trajectory very close to either the top or the bottom border (which means a small uncertainty in the y dimension), and the thick red line (top) shows a trajectory in the middle of the track, far from the borders (which means a large uncertainty in the y dimension). They account for 85% of the place fields between them and thus explain most of the variance. (Data from [68]).
doi:10.1371/journal.pone.0089762.g001

For N observations, the uncertainty is:

$$\hat{\sigma}_x = \sqrt{\left(\frac{1}{\sigma_p^2} + \sum_{i=1}^{N} \frac{1}{\sigma_{o,i}^2}\right)^{-1}} = \sqrt{\left(\frac{1}{\sigma_p^2} + \frac{1}{s^2}\sum_{i=1}^{N} \frac{1}{d_i^2}\right)^{-1}}. \quad (6)$$

According to hypothesis 3 (see Hypotheses section), the observation uncertainty is proportional to the distance d_i: $\sigma_{o,i} = s \cdot d_i$. Thus, $\frac{1}{\sigma_{o,i}^2} = \frac{1}{s^2}\frac{1}{d_i^2}$. Substituting the precision or accuracy of the prior belief $\frac{1}{\sigma_p^2}$ by a_p, and the factor $\frac{1}{s^2}$ influencing observation precision (i.e. how rapidly the accuracy of distance judgements decreases with increasing distance) by a_o, we arrive at equation (7), which can be used to calculate the resulting uncertainty given a prior belief accuracy (which might depend on the path integrator) and the distances and accuracies of all observations.

$$\hat{\sigma}_x = \sqrt{\left(a_p + a_o \sum_{i=1}^{N} \frac{1}{d_i^2}\right)^{-1}} \quad (7)$$

Equation (7) was used in the Results section to predict uncertainties (hypothesized to be correlated with place field sizes), given distances to boundaries or landmarks. Explained proportions of variance R^2 were calculated from $R^2 = 1 - SS_{err}/SS_{tot}$, where SS_{err} is the sum of squared differences between the model prediction and the recorded data, and SS_{tot} is the total sum of squares.

For the data analysed in the Results section, we assumed the parameter a_p to be negligible - a_o was the sole parameter fitted to the data. The single-unit place field data on the linear and circular tracks (see first two subsections in Results) has been obtained from electrodes in distal parts of area CA1 of the Hippocampus (closest to the subiculum), which receive few connections from the neural path integrator (MEC), as opposed to proximal CA1 [41]. These recorded place cells were probably mainly driven by sensory information (subiculum, LEC) instead of path integration information (MEC) [41,42], which is why we assumed a_p, the parameter accounting for path integration accuracy, to be negligible for these specific datasets.

Since the simplifying assumptions made by the model presented here are too strong for real-world environments, and since place cell firing is influenced by many more factors other than environmental geometry, such a simple model cannot yield highly accurate predictions of electrophysiological recordings. However, if place cells integrate information in a Bayesian fashion, and if the sizes of their place fields are correlated with uncertainty, then even this simple model should be able to approximately account for the distribution of place field sizes and their dependence on the distances to boundaries and landmarks in the environment. For example, place fields should be smaller close to boundaries and larger far from boundaries. In the Results section, we will compare these predictions of the Bayesian model to data recorded from rat place cells in different environments.

Equation (7) can be extended to only include subsets of observed objects (see Discussion) by introducing a set of binary variables $u_i \in \{0,1\}$ indicating whether a certain object observation is being used in the uncertainty estimation. If $u_i = 0$, then the probability of observation i does not influence the posterior probability. Thus, in the one-dimensional case, the observation probabilities will be

$$p(\boldsymbol{o}_i|\boldsymbol{x}) = \begin{cases} 1 & \text{if } u_i = 0 \\ \mathcal{N}(\mu_{o,i}, \sigma_{o,i}) & \text{if } u_i = 1 \end{cases} \quad (8)$$

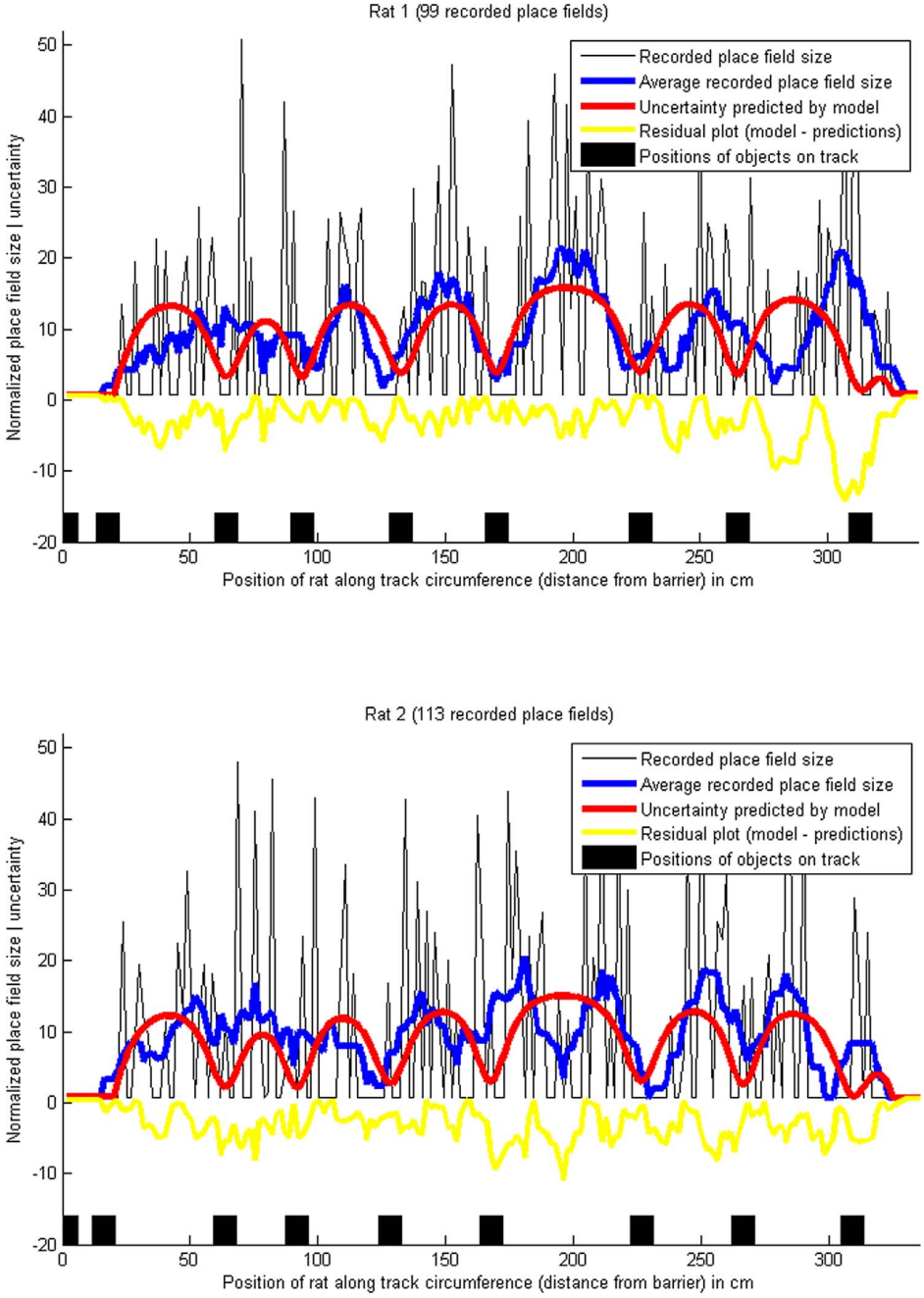


Figure 2. Place field sizes, and predicted uncertainty, on a circular track with objects. The blue lines show the smoothed place field sizes (10-point moving average), normalized to a mean of 0 and variance of 1, and the red lines show the location uncertainty predicted by the Bayesian model. The minima of the red lines correspond to the black squares marking the positions of the objects on the track, since the location uncertainty is lowest near to an object and highest when the rat is far from the objects. Pearson's correlation coefficient between the recorded place field sizes and the predicted uncertainty was $r = 0.56$ for rat 1 and $r = 0.55$ for rat 2. The proportions of explained variance were $R^2 = 0.22$ for rat 1 and $R^2 = 0.20$ for rat 2. (Data from [42]).
doi:10.1371/journal.pone.0089762.g002

If we insert equation (8) into equation (4) calculating the mean and uncertainty (standard deviation) of the 'best guess' location, and solve for the standard deviation (see [40]), we get the following extended expression representing the uncertainty depending on the distances of a subset of the observations:

$$\hat{\sigma}_x = \sqrt{\left(a_p + a_o \sum_{i=1}^{N} \frac{u_i}{d_i^2}\right)^{-1}} \tag{9}$$

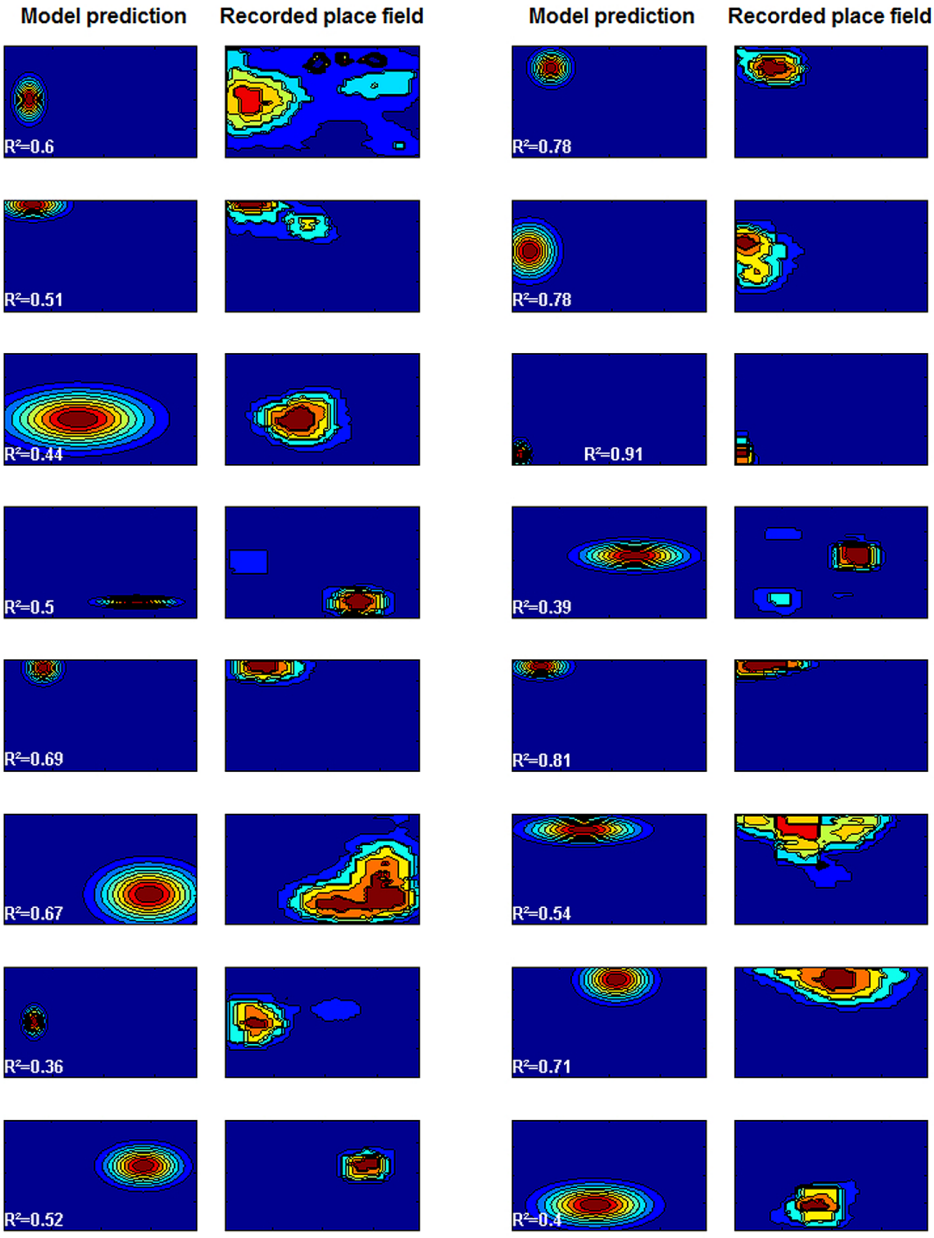


Figure 3. Predicted and recorded place fields in environment B. The squares represent firing rates at each point of the big square environment, with hot colors marking high firing rates, and cold colors low firing rates (the plots have been scaled to fit the page - see main text for the actual proportions of the environments). The model prediction was made based on parameters estimated from the other environments (environments A, C and D). The overall mean proportion of explained variance was $R^2 = 0.60$ (Data from [69]).
doi:10.1371/journal.pone.0089762.g003

where $i = 1, \cdots, N$ indexes one of N objects, boundaries, or obstacles. u_i can be fitted using e.g. a non-linear optimizer or a brute-force approach - trying all possibilities - if the number of obstacles is small enough (in the Results section, we have adopted the latter approach).

Bayesian inference on the neuronal level: a possible model

The hypotheses outlined in the Introduction section imply that, physiologically, the firing fields of place cells should shift and shrink in a statistically optimal fashion. This might be caused by a large number of possible mechanisms (see e.g. [43] for some

proposed implementations of Bayesian inference in brains, and the Discussion section). We have chosen to implement a different solution for Bayesian inference in spiking neuronal networks, based on coincidence detection. We report simulation results of this neuronal Bayesian inference model in the last subsection of Results. This neuronal model rests on the following assumptions:

Inference using coincidence detection. A mechanism for obtaining a Bayesian posterior requires a multiplication of probability distributions. In a network of spiking neurons, such multiplication might be implemented by coincidence detection [44], a mechanism that hippocampal CA1 neurons have been observed to exhibit [45–47]. This particular implementation of multiplication is a hypothesis that our proposed conclusion does not depend on, since multiplication could also be implemented neuronally in several other ways (e.g. [48]). However, we chose this one for its simplicity and computational efficiency. Furthermore, a number of neuronal network models capable of performing Bayesian inference have been proposed before [43,49–52]; nevertheless, none of these methods are fully compatible with the anatomical properties of the HEC and the physiological evidence from place cells (see Discussion). For this reason we chose to implement a novel solution for Bayesian inference in spiking neuronal networks, based on coincidence detection, and inspired by sampling-based approaches to represent probability distributions [53–55].

The temporal resolution of coincidence detection is in the right range to approximate multiplication. Bayesian inference requires multiplication. Multiplication by coincidence detection only works well within a certain range of temporal resolution of the coincidence detection. If the temporal resolution is too high, very few inputs, or even one input, can elicit output spikes, in effect leading to an addition of the inputs instead of a multiplication. Too low a temporal resolution on the other hand could lead to very sparse output spikes, leading to a displacement of the output firing field and destroying the statistical near-optimality (or, in the extreme case, to zero output spikes). The coincidence detection properties of noisy integrate-and-fire neurons have been analysed in two studies [56,57] (although their analyses are based on a simple spiking neuron model, recordings by [56] indicate that these expressions closely model the coincidence detection behaviour of biological neurons in vitro). According to [57], the temporal resolution can be approximated based on the standard deviation of the fluctuation of the membrane potential σ, the membrane time constant τ_m and the amplitude w of the postsynaptic potentials (PSPs) as follows:

$$T \approx 1.35 \frac{\sigma}{w} \tau_m \tag{10}$$

Inserting standard values observed in vivo in area CA1 of the Hippocampus into equation (10) ($\tau_m \approx 18ms$ [58,59], $\sigma \approx 6mV$ [60], and w just under the $24 \pm 9mV$ necessary to discharge a place cell [61]) yields around $T \approx 7 \pm 3ms$. The temporal resolution of the coincidence detection in hippocampal CA1 neurons has also been measured in vitro, and is of the same order of magnitude. For example, Jarsky et al. have found that CA1 neuron firing upon perforant path input spikes is strongly facilitated by synchronous spikes from Schaffer-collateral (SC) synapses arriving within 5–10 ms, but is otherwise unreliable if no synchronous SC input is present [45].

This temporal resolution constant T is small enough to approximate multiplication (see Results), but sufficiently large to allow enough coincidences to form a place field. Even with very sparse information, e.g. in rat experiments under total darkness [62,63] in which the place fields presumably arise mostly from grid cell input, place cells might receive up to 200–20,000 incoming spikes per second (based on around 100–1,000 connections between grid cells and a place cell [4,64,65], and a grid cell firing rate around 2–20 Hz [16,66]). Given the temporal resolution of $T \approx 7 \pm 3ms$, this spike rate is sufficient to elicit the empirically observed CA1 place cell firing rates of around 1–10 Hz (e.g. [42,67]) in locations where many grid cells firing fields overlap.

Approximating a Bayes-optimal location estimate. Place cells should approximate a Bayesian posterior according to hypothesis 2, as expressed in equations (1) and (2). Neuronally, each border cell could represent a boundary proximity probability distribution $p(\boldsymbol{o}_i|\boldsymbol{x})$, if we assume that firing rate distributions are correlated with probability distributions (cf. hypothesis 1). The MEC grid cell path integrator could provide the prior location distribution $p(\boldsymbol{x})$. Although a single grid cell cannot provide an unambiguous estimate, having many firing fields, an ensemble of multiple thresholded grid cell inputs yields a single firing field (or few firing fields) in small environments, as pointed out by grid cell-driven place field models [3,4]. This reduction to one or few firing fields works both with additive inputs, as in most rate-coded neural network models, and with multiplications of inputs.

Integrate-and-fire spiking neurons are able to approximate the multiplication of their inputs by making use of coincidence detection (see Figure 4). Thus, such neurons can represent a posterior (i.e. a product of probability distributions). If we represent the spike train of each neuron using a function $S(t)$, which at a given time t is $S(t)=1$ if the neuron has fired a spike within the time interval $[t,t+\tau)$, and 0 otherwise (τ being the time discretization parameter of the model, which we set to the temporal resolution of coincidence detection in place cells - see Text S2 in the Supporting Information), then the spike train of the place cell computing the posterior, S_{pc} can be expressed using the spike trains of its M input neurons, S_i:

$$S_{pc}(t) = \mathrm{H}\left(\frac{1}{M} \sum_{i=1}^{M} (S_i(t) - \alpha) \right) \tag{11}$$

Where $H(\cdot)$ is the Heaviside step function, and $\alpha = (0,1]$ is the proportion of input neurons required to spike within τ time in order to elicit an output spike in the place cell. See Text S2 in the Supporting Information for the derivation, and for arguments why this expression approximates multiplication. Using equation (11), we can express the probability P_{x_A,x_B} that the rat is on a path between the locations x_A and x_B during K time intervals of duration τ (represented by $T_{A,B}$), using the spike train of a place cell presumably representing the outcome of the Bayesian inference process S_{pc}, the spike trains representing of N grid cells $S_{gc,1} \ldots S_{gc,N}$, and the spike trains of M border cells $S_{bc,1} \ldots S_{bc,M}$:

$$P_{x_A,x_B} \propto \frac{1}{K} \sum_{t \in T_{A,B}} S_{pc}(t) \tag{12}$$

$$S_{pc} = \mathrm{H}\left(\frac{1}{N} \sum_{i=1}^{N} (S_{gc,i}(t) - \alpha) + \frac{1}{M} \sum_{i=1}^{M} (S_{bc,i}(t) - \alpha) \right) \tag{13}$$

Equations (12) and (13) describe how Bayesian inference can be implemented in a spiking neuronal network, approximating the

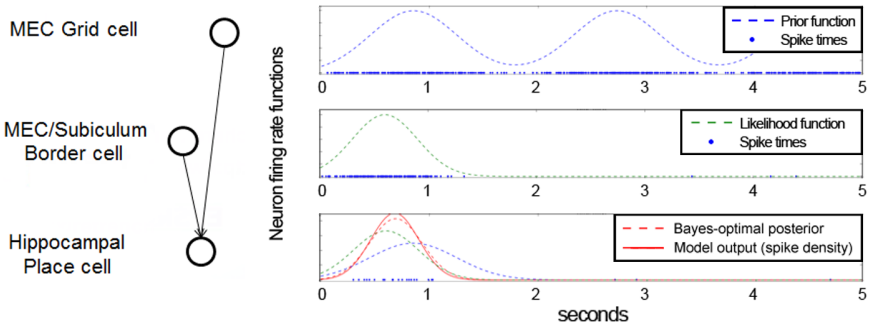


Figure 4. Neuronal implementation of Bayesian inference based on coincidence detection. This simple integrate-and-fire model contains only three spiking neurons, and shows their spikes over 5 simulated seconds. Each plot shows the spikes (blue dots in bottom rows), as well as the corresponding instantaneous firing rate or spike density. First row: a simulated grid cell (pre-defined firing rate function), used as the prior. Second row: simulated border cell (pre-defined firing rate function), used as the observation likelihood. Third row: simulated place cell, representing the posterior, firing only when all incoming inputs are coincident (i.e. they occur within a small time window). The Gaussian drawn over the mean and standard deviation of the noise-filtered spikes represents the place field, and approximates the Bayesian optimum. Bottom row: plot of the membrane potential of the place cell.
doi:10.1371/journal.pone.0089762.g004

posterior probability distributions with spikes of the place cell which are viewed as samples of that distribution (see Text S2 in the Supporting Information for the derivation, and for a formulation of coincidence detection as rejection sampling; and see Figure 4 for simulation results using integrate-and-fire spiking neurons).

Results

Place field sizes on a linear track

Figure 1 shows this prediction of the Bayesian model in a rectangular environment, and compares it to single-unit recordings of the place cells in area CA1 of the hippocampus of ten male Lister Hooded rats (data from [68]). The rats ran on a narrow rectangular track with food cups at both ends. These sizes were also used to generate the model predictions. In the following, x denotes the distance of the rat from the eastern boundary, y the distance from the southern boundary, and L and W the constant length and width of the environment ($L=254cm, W=10cm$ [68]). The model was instantiated with the four boundaries of the environment, and the uncertainty at each point of the track calculated by multiplying the separately calculated x and y uncertainties $\sigma=\hat{\sigma}_x\hat{\sigma}_y$, which are assumed to be independent on this track (see Methods).

$$\hat{\sigma}(x,y)=\sqrt{a_o^{-2}\left(\frac{1}{x^2}+\frac{1}{(L-x)^2}\right)^{-1}\left(\frac{1}{y^2}+\frac{1}{(W-y)^2}\right)^{-1}} \qquad (14)$$

The y-axis of Figure 1 shows the total place field area of the recorded place cells, in bins of 1.9 cm. Under the hypothesis that uncertainty is correlated with place field size (H2), equation (14) implies that the biggest place fields should be in the center of the track. Since both the distance from the east boundary and from the north boundary influence the uncertainty, it also implies that at each position along the length of the track, there should be multiple uncertainties, depending on whether the rat is close or far from the side borders (the south/north border), which is shown by the two red lines in Figure 1 (the thin red line corresponds to the rat running close to the south/north border, and the thick red line to it running in the center, far from those borders). The parameter a_o in equation (14) was adjusted using a coordinate descent algorithm. Using this single parameter, the model can explain why place fields were bigger when closer to the center of the track. Most of the recorded place field sizes (85%) fall between the boundaries of the model.

Place field sizes on a circular track with objects

Figure 2 shows the results of the model in a more complex environment, comparing the sizes of place fields of recorded place cells of two male Fischer-344 rats in an experiment performed by Burke et al. [42], in which the rats were running on a circular track with 106.7 cm diameter and 15 cm width. The track contained a barrier with food trays on each side to motivate the rats to run along the track, alternating between clockwise and counter-clockwise laps. It also contained 8 randomly distributed objects, and was otherwise featureless. The Bayesian model, equation (7), was fitted to the recorded data, using $N=9$ observations (the 8 objects, and the barrier). Uncertainty was calculated in one spatial dimension, which corresponds to the distance of the rat from the barrier along the track.

The single-parameter model achieved correlations of $r_{f1}=0.56$ for rat 1 and $r_{f2}=0.55$ for rat 2 between the smoothed place field sizes and the fitted model - see Figure 2 (the probabilities of getting correlations as large as these values by random chance are negligible: $p_{r1}=3*10^{-16}$ for rat 1 and $p_{r2}=2*10^{-17}$ for rat 2). The average place field sizes clearly have a non-random structure, with the minima corresponding to the locations of the 8 objects and the barrier, as predicted by the Bayesian model (the null hypothesis of the data being random can be rejected with high confidence, with $p_1=0.001, p_2=0.008$ for the two rats according to a chi-square goodness-of-fit test of the place field size data against a normal distribution).

On the other hand, it is plausible that the residual errors, i.e. the model subtracted from the average place field sizes, are randomly drawn from a normal distribution, implying that the model explains a significant part of the non-random structure (the null hypothesis of the errors being random cannot be rejected according to a chi-square goodness-of-fit test of the residual errors against a normal distribution, with $p_1=0.175, p_2=0.119$ for the two rats). Some recorded place cells had multiple place fields [42], in which case the predicted uncertainty was calculated for each place field separately.

In Figure 2, the x-axis shows the positions of the means (centroid) of the recorded spikes of each place field, and the y-axis shows the size of the fields, derived by calculating the standard deviations of the spike positions. This makes these place field sizes directly comparable to the uncertainties calculated by equation (7), provided that the place fields resemble Gaussians, being approximately symmetric, and having the highest spike density around the mean (centroid). If this was not the case - if the recorded place fields were not approximately Gaussian -, the spike densities would

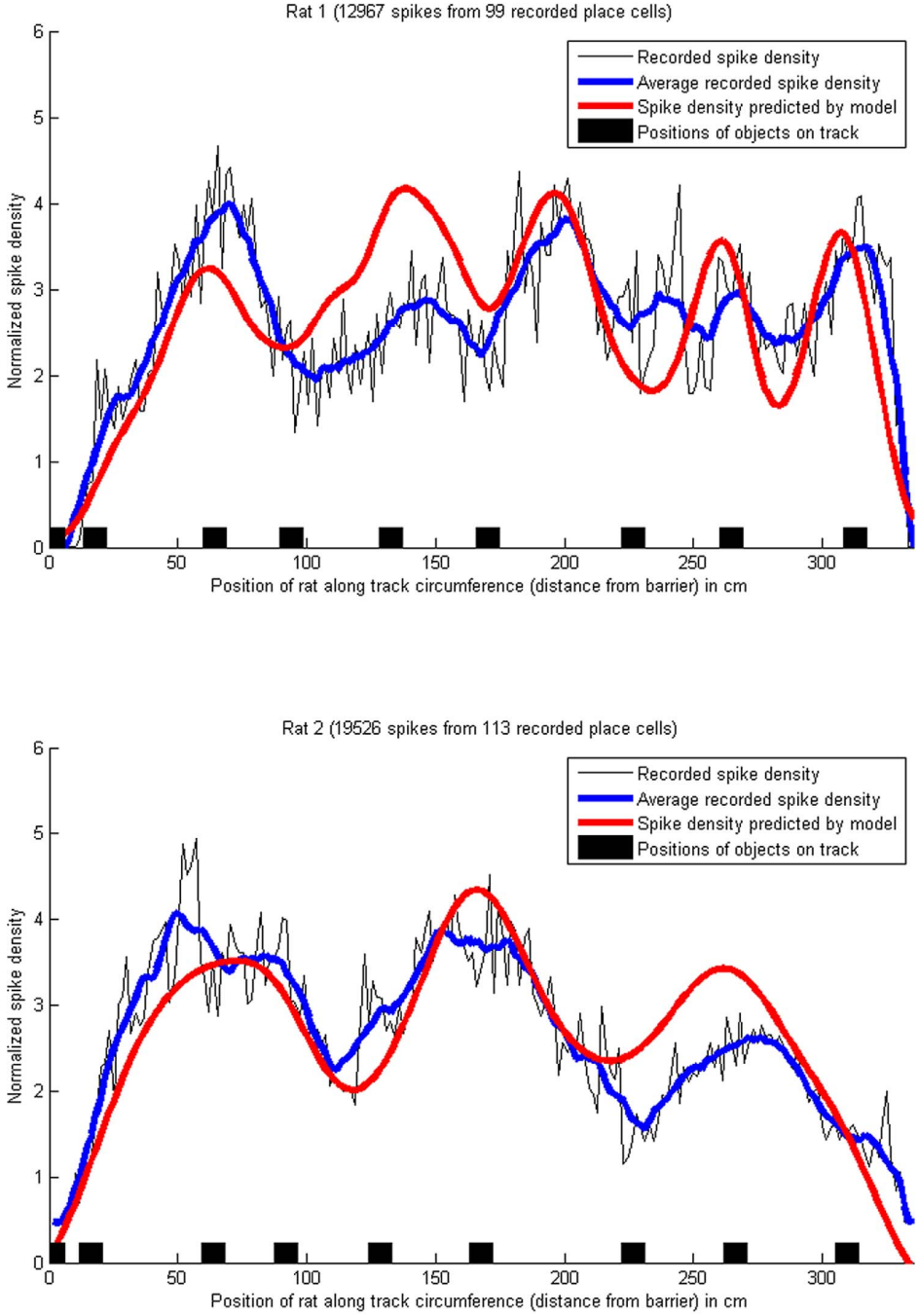


Figure 5. Density of place cell spikes, and predicted uncertainty, on a circular track with objects. The blue lines show the smoothed (averaged) density of place field spikes, i.e. the number of spikes across all recorded place cells for each centimetre of the track, normalized to a mean of 0 and variance of 1. The red lines have been obtained by summing Gaussian distributions, one for each place cell, with the means set to the center of each place field, and the standard deviations set to the location uncertainties (hypothesized to be correlated with place field sizes, see H2) as above. The exact amplitude of the spike density at each location depends on the place cells firing rate, which is influenced by many non-spatial factors such as running speed [67], but the shape of the curves is comparable. Pearson's correlation coefficient between the recorded place field sizes and the predicted uncertainty was $r=0.74$ for rat 1 and $r=0.86$ for rat 2. The proportions of explained variance were $R^2=0.38$ for rat 1 and $R^2=0.70$ for rat 2. (Data from [42]). doi:10.1371/journal.pone.0089762.g005

deviate from the prediction of the model. Figure 5 compares the spike densities of all recorded spikes to the densities predicted by the model, achieving correlations of $r_{s1}=0.74$ for rat 1 and $r_{s2}=0.86$ for rat 2 (the probabilities of getting correlations as large as these values by random chance are negligible: $p_{r1}=7*10^{-37}$ for rat 1 and $p_{r2}=1*10^{-60}$ for rat 2).

Place field sizes after changes in the environment size

Changes in the environment have been shown to influence place fields. In order to show that the Bayesian model does not violate the observed effects, and can predict place field size in novel environments, we have applied it to the data presented in [7] for evaluating the BVC model, and originally reported in [69]. The data was recorded from six rats foraging for food in four

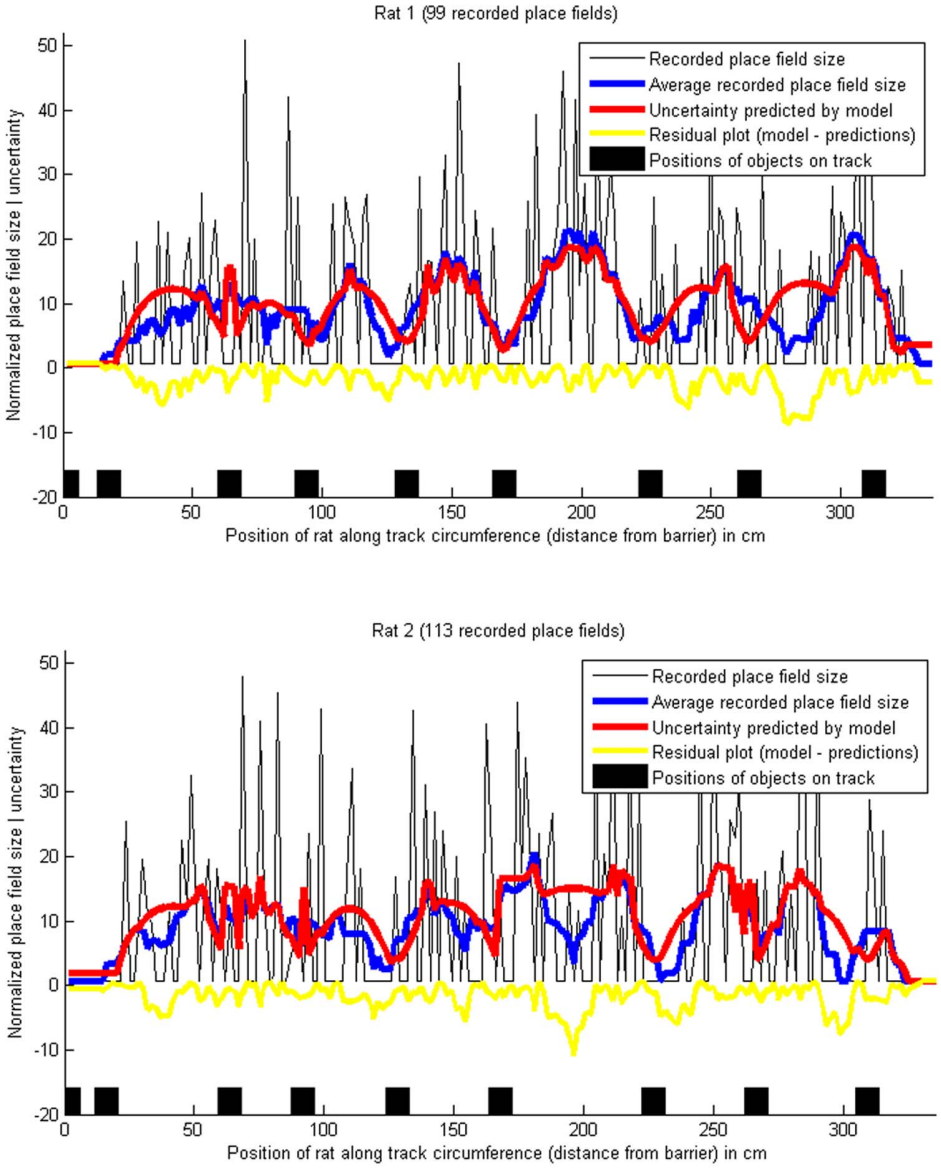


Figure 6. Place field sizes, and predicted uncertainty, on a circular track with objects, using the extended model. The blue lines show the smoothed place field sizes (10-point moving average), normalized to a mean of 0 and variance of 1, and the red lines show the location uncertainty predicted by the extended Bayesian model (which takes into account only a subset of the objects on the track at each point). Pearson's correlation coefficient between the recorded place field sizes and the predicted uncertainty was $r = 0.82$ both for rat 1 and rat 2. The proportions of explained variance were $R^2 = 0.66$ for rat 1 and $R^2 = 0.60$ for rat 2. (Data from [42]).
doi:10.1371/journal.pone.0089762.g006

different environments: a small square of size 61×61 cm (environment A), a large square of 122×122 cm (environment B), and a horizontal and vertical rectangle of 61×122 cm and 122×61 cm (environments C and D). 12 of the 28 recorded place fields were discarded from the dataset because they were asymmetric and did not fit a Gaussian distribution (see Discussion for possible model extensions). For the remaining 16 place fields, the parameters of the model were adjusted using the data from two of the four environments, C and D. The means and standard deviations of the Gaussians used to represent the place field in the x and y dimensions were obtained by using a least squares fitting procedure, and the parameter a_o calculated from the known distances and standard deviations using equation (7). This equation also allowed calculating the predicted place field size, i.e. the standard deviation of the representing Gaussian, in the remaining two environments, by using appropriately scaled distance relations.

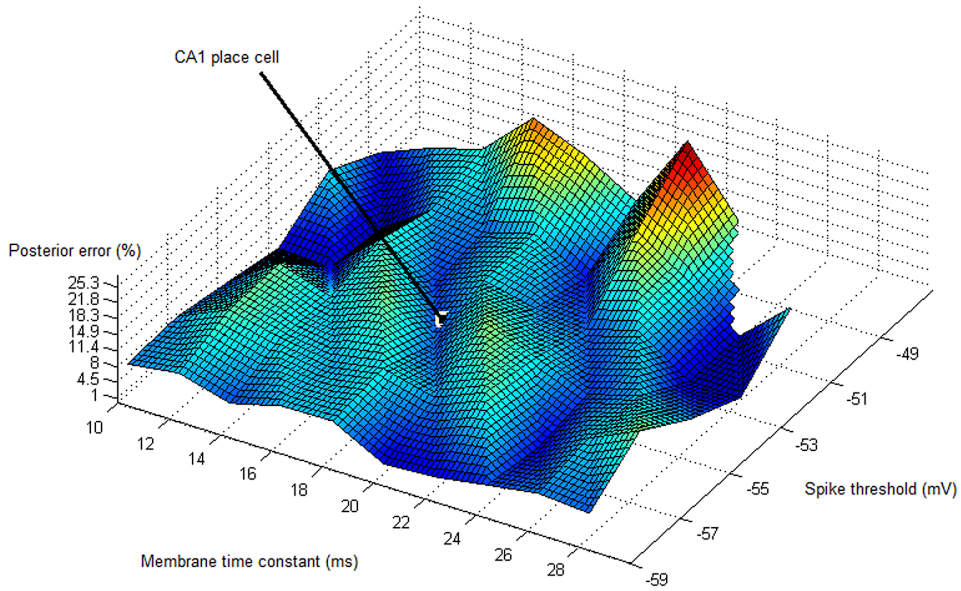


Figure 7. Errors of coincidence-based multiplication based on a simple integrate-and-fire model. The altitude shows the error (lowest point: 1%, highest point: 16%, error at CA1 place cell parameters: 5%), and the x and y axes show the dependence of the error on the membrane time constant τ and the spike threshold V respectively. Interestingly, the parameters of some CA1 place cells ($\tau = 17ms$, $V = -54mV$) fall into one of the local minima of error; and no hippocampal place cell reaches the area of maximum error.
doi:10.1371/journal.pone.0089762.g007

Then, the predicted and recorded place fields were compared at each point in the environment (see Figure 3). Figure 3 shows the results for environment B (the large square environments), achieving a mean proportion of explained variance of $R^2_{mean} = 0.60$. This fit of the model predictions was compared against the optimal fit possibly achievable by Gaussian functions, calculated by fitting Gaussians to each actual firing field in the B environments using a least square errors procedure. This optimal fitting procedure yielded $R^2_{optimal} = 0.68$ on average, which is not statistically different from the model fit ($p = 0.29$ on a paired t-test over all individual place field R^2 values). This shows that the Bayesian model can make predictions which fit the data almost as well as optimally fitted Gaussian functions. The difference between the fit of the model and of this optimal fit is statistically insignificant.

Place field sizes from subsets of observed objects

The model used so far makes a number of simplifying assumptions, which yield a very simple mathematical form with up to two parameters - see equation (7) - and already provides reasonable predictions of experimental data (see above). However, the accuracy of the model can be improved by relaxing some of these assumptions, at the expense of simplicity (see the Discussion section).

One way to improve the model accuracy is to allow place cells to be driven not by every single boundary and obstacle in the immediate environment, but only by a subset of these objects. Equation (9) allows the calculation of uncertainties taking into account a subset of observations (see Methods). This extension introduces N additional model parameters for the N binary variables u_i specifying whether or not observation i is being taken into account.

Fitting this extended model to the data recorded on the circular track with objects, significantly increases the model fit - instead of explained proportions of variance $R^2 = 0.22$ for rat 1 and $R^2 = 0.20$ for rat 2, the extended model achieves $R^2 = 0.66$ for rat 1 and $R^2 = 0.60$ for rat 2 (see Figure 6). However, this extended model uses $N = 9$ more parameters than the original model (8 objects on the track, plus the barrier with the adjacent food trays). To take into account the number of parameters relative to the number of data points, we also compare the adjusted R^2 values (which we denote by $\overline{R}^2$). Instead of $\overline{R}^2 = 0.21$ for rat 1 and $\overline{R}^2 = 0.19$ for rat 2, the extended model yields $\overline{R}^2 = 0.62$ for rat 1 and $\overline{R}^2 = 0.56$ for rat 2 after adjustment by the number of parameters.

Further possible extensions of the model, such as allowing skewed place fields, will be described in the Discussion section.

Bayesian inference on the neuronal level: a possible model

As argued above, the sizes of place fields should be dependent on incoming sensory information, in order to approximate the statistically optimal location of the animal, and the uncertainty associated with it. Mathematically, this means calculating a Bayesian posterior (see Methods). We have already presented some evidence that place cells might be able to approximate such Bayesian calculations in the previous sections. Here we extend this idea by suggesting a tentative model of how these calculations might be implemented on the neuronal level.

A spiking neuronal network could implement the multiplication operation required for calculating a Bayesian posterior by making use of coincidence detection. Figure 4 shows a simple example of a place cell receiving input from only one grid cell (path integration) and one border cell (observation). The place cell is modeled using a current-based integrate-and-fire neuron model [70] (membrane time constant $\tau_m = 17ms$, synaptic time constant $\tau_s = 5ms$, resting potential $V_r = -80mV$, spike threshold $V_t = -55mV$, synaptic weights $w = 26mV$). Synaptic inputs are modeled as spike trains drawn from non-homogeneous Poisson processes, with firing rates

controlled by Gaussian distributions (see dashed lines in the figure) to approximate the symmetric firing fields of grid cells and some border cells. The Brian simulator was used to simulate the place cell and to plot Figure 4 [71].

In the figure, the place cell only fires when both the grid cell and the border cell inputs arrive within a small time window. This leads to a shifting of the place field - the place cell combines both types of information, and forms the place field at a location specified by the weighted average of the grid field and border field location, the weighting depending on the uncertainties (field sizes) of the inputs. Thus, the place field is located between the grid and the border field, but closer to the border field because it is narrower (more accurate).

Figure 4 is intended to illustrate the concept of inference by coincidence detection. The model relies on the fact that if the threshold of the output neuron is set high enough to only allow output spikes on synchronous input spikes, then the output neuron performs approximate multiplication, as required by Bayesian inference. The approximation error mainly depends on two parameters of the output neuron: its membrane time constant, and its spike threshold voltage. For our purposes, we define the approximation error as the absolute difference between the posterior mean estimated by the model, and the mean of the exact posterior according to Bayes' rule (the error of the posterior mean is most relevant for a location model, since the statistically optimal location estimate is located at the mean of the posterior distribution under Gaussian assumptions). Figure 7 shows how this approximation error depends on these two parameters. For an analytical discussion of the coincidence detection properties of integrate-and-fire neurons, see [56,57].

Discussion

We have attempted to highlight the usefulness of Bayesian models in explaining information combination in place cells. Although such models are too simple to explain all firing properties, their predictions fit the data quite well given their simplicity (low numbers of parameters), which is an important property of good models [72–74]. We have compared such model predictions to three different datasets recorded from rat place cells in different environments in the Results section, using firing field size as a measure of uncertainty. Our results suggest that the 'Bayesian brain' hypothesis might be useful in trying to understand information processing in Hippocampal place cells, not just at a computational level as has been suggested many times before [8–11], but also at the neuronal level.

Bayesian spatial cue integration has been investigated before on the behavioural level. Nardini et al. [75] investigated cue integration in human children and adults, using a paradigm in which subjects had to return an object to its original place, either given only landmark information, only self-motion information, or both. Their results suggest that adults are able to reduce the variance (uncertainty) in their response by integrating different spatial cues in a statistically near-optimal fashion. Cheng et al. [11] reviewed animal experiments, arguing that the integration of different spatial cues might be partially explained by Bayes' rule - for example, pigeons seem to assign weights to information from different landmarks using Bayesian principles. Therefore, in contrast to previous work, this paper significantly extends these ideas by directly comparing the predictions of Bayesian spatial cue integration to physiological data recorded from rat place cells, and argues for the plausibility of this cue integration mechanism on the neuronal level.

The claim that perception (spatial or otherwise) is based on Bayesian inference, implemented physically as a neuronal mechanism, has been criticized for multiple reasons [15]: the lack of strong physiological evidence in favour of the Bayesian hypothesis (most existing evidence to date is behavioural, coming from 'Bayesian psychophysics' [15,76]), the arbitrary choice of prior functions in favour of simplicity in many of these models (instead of the choice being based on empirical data), and the ability to explain Bayes-optimal perception in cue integration in some paradigms *without* a Bayesian mechanism, by implementing reinforcement learning.

In this paper we have argued that firing field properties of single place cells resemble the outcomes of Bayesian inference processes. Following the advice of [77] we have generated quantitative experimentally testable predictions, and compared them with empirical results. Thus, in contrast with the view that *'Bayesian models do not provide mechanistic explanations currently, instead they are predictive instruments'* [15], we provide one of a few existing pieces of empirical evidence in favour of the idea that the brain might represent uncertainty at a neuronal level, and that there are some neuronal level mechanisms approximately conforming to Bayesian principles. Our results therefore contribute to the *'current challenge for these* [Bayesian] *models* [is] *to yield good, clear, and testable predictions at the neural level, a goal that has yet to be satisfactorily reached'* [15].

Bayesian localization

Bayesian cue integration might also play a role in the more complex problem of maintaining a near-optimal location estimate through time, despite noise and accumulating errors. In robotics, one popular family of solutions for maintaining statistically optimal location estimates is called Bayesian localization (an example algorithm from this family would be the Kalman filter) [38]. Given some simplifying assumptions, Bayesian localization can be performed by the following three computations at each time step, in order to maintain a statistically optimal, error corrected location estimate:

1. **Path integration.** Updates the prior location belief with (possibly erroneous) movement vectors using a motion model at each time step.
2. **Correction.** A Bayesian inference mechanism that corrects the location belief using observations.
3. **Update.** Finally, the path integrator's estimate is updated to the corrected estimate.

There is ample evidence in literature that the HEC is able to perform step 1 [17] - grid cells update their firing with each movement. We have presented evidence in the Results section for step 2, strongly suggesting that place cells might be able to perform approximate Bayesian computation. With respect to step 3, there is anatomical evidence that such an update could happen - place cells can project back to grid cells and influence their firing [78–80]. Such back-projections might serve the role of providing environmental stability for the grids [81], and prevent the accumulation of error during path integration [3,18,82]. They are also postulated in a model of grid-cell based error correction, which shows how the redundant modular coding in the entorhinal cortex might constitute an exponentially strong population code - it can *'produce exponentially small error at asymptotically finite information rates'* [83] (however, this model does not account for location correction using observations). The idea of back-projections from grid cells to place cells is supported by recording evidence showing that grid cell representations become erroneous, less gridlike, and expand in field size in novel environments [66]; and recent

evidence indicating that deactivating the hippocampus extinguishes grid fields [84].

Thus, the Hippocampal-Entorhinal Complex might be able to implement Bayesian localization and maintain approximately statistically optimal location estimates through time, despite accumulating errors. Entorhinal grid cells are able to integrate movement signals [17]. Bayesian cue integration in place cells (see Results section) might be the mechanism performing the correction step and then, after near-optimal cue integration, the corrected location estimate would update grid cells (the neuronal path integrator) through the place cell back-projections.

Phase resetting presents a plausible mechanism by which to perform this update step. It has previously been suggested that error correction in oscillatory interference models of grid cells might be implemented through phase reset, the resetting of the phase of intrinsic oscillations in MEC grid cells [81,85,86]. Therefore, when entering a new environment, connections might form between place cells and grid cells firing simultaneously (i.e. between cells with coinciding firing fields), to anchor the grid field representation to environmental features such as boundaries. These connections could induce a reset of the intrinsic oscillation phase of the grid cell when the grid field shifts (e.g. due to path integration errors) [85]. The changed oscillation phase would lead to a displacement of the grid field back to the center of the place field, because grid cell firing fields arise from the oscillatory interference patterns between background theta oscillations and the intrinsic oscillations in the grid cell in oscillatory interference models, with the grid cell firing rate being highest when the phases coincide [87].

There is some recording evidence showing that single incoming spikes can indeed reset intrinsic oscillation phases in cells of the entorhinal cortex [88–90]. Because a single postsynaptic potential suffices, the probability of phase reset occurring depends on the firing rate(s) of the place cell(s) connected through the back-projections. Thus, as the animal is running through the place field, the firing rate within the grid field might gradually adapt to the firing rate of the place field, and the fields would become aligned, completing the update step.

Possible extensions

There are some properties of place fields which the model presented here, in its simplest form, while not inconsistent with, cannot account for. The basic uncertainty estimation, equation (7), does not account for place cells driven by only a subset of the objects in the environment, instead of all of them, however, some place fields have been observed to be controlled by specific landmarks [91]. Equation (9) makes it possible to parametrize which subset of the object distances are taken into account for the uncertainty calculation, yielding a significantly better model-data fit on the track with multiple objects (see Results).

Although the equations used in the Results section use a single Gaussian distribution to model a place field, this model can be used to model place cells with multiple place fields in a straightforward fashion, by calculating a separate uncertainty value for each place field using the respective distances of objects from the place field centroids. Thus, multiple uncertainty values can be associated with each place cell, one for each place field - as in Figure 2 for example, in which many of the plotted place fields belong to multi-field place cells (see [42] for the distribution of single-field and multi-field place cells in this dataset).

Further phenomena not explained by the simple model include asymmetric place fields that are frequently found in area CA1 of the hippocampus, and the observation that place field sizes seem to increase along the dorso-ventral axis of the hippocampus [67]. Asymmetric place fields could potentially be modelled using skewed probability distributions such as the Skew-Normal Distribution [92] as observation likelihoods instead of Gaussians, using a similar approach to the one described in the Methods section. The grid cell input to a place cell is usually symmetric, but the firing fields of border-related cells can be skewed [22,23], which might give rise to asymmetric place fields. The skewness parameter of an asymmetric probability distribution (such as the Skew-Normal Distribution) in such an extended model might increase as a function of familiarity with the environment (time spent in the same environment), in order to model the experience-dependent asymmetry of some CA1 place fields [93]. The mean and variance of such a distribution could be estimated similarly to the approach proposed in the Methods section. Future work, and experimental data from place cells recorded over extended periods of time, will be needed to verify how well such an asymmetric model could account for skewed place fields.

It is interesting to note, with respect to the fact that the place field sizes increase along the dorso-ventral hippocampal axis, that the same field size increase has been observed in grid cells in the medial entorhinal cortex [94]. Since grid cells are hypothesized to play a role in driving place cell firing, both in our model and in previous models [3,4], this might account for the place field size gradient. In an extended model taking into account the spatial configuration of the hippocampal-entorhinal complex, if the dorsal grid cells are adjusted to have small firing fields and the ventral ones large firing fields (50 cm–3 m, see [94]), this will lead to a similar gradient in the resultant place fields, given that the grid cells at least partially drive the firing of the place cells. The role played by boundary-related inputs would mean that not every place field would fit this dorso-ventral size gradient, but on average a field size gradient could be observed in such a model.

Related work

The Boundary Vector Cell model [24] of place cell firing also explains place fields in terms of geometric relations to environment features, although it does not suggest statistical near-optimality and does not make use of Bayesian cue integration. The objective fit of the simple model presented in this paper is not as good as the fit achieved by the Boundary Vector Cell model ([7] describes the fit of the BVC model to the data in figure 3). The BVC model could, in principle, also be fitted to the first two datasets presented in the Results section, but would require the adjustment of a higher number of parameters than there are data points and thus would not have a unique solution (Hartley et al. [7] simulated 2–4 inputs per place cell, requiring up to 7 parameters to be adjusted for each place cell; and a few additional global parameters - over 700 fitting parameters for the data in Figure 2).

The model presented here serves a different purpose; not to present a more accurate model of place fields, but rather to highlight that the information integration in place cells approximately resembles simple Bayesian computation. Our results suggest that predictions resembling in-vivo recorded place field data can be made based on a *single underlying principle: the statistically optimal combination of information*. Because of its simplicity, this model cannot fully explain experimental data, and does not achieve a fit as good as previously suggested models such as the Boundary Vector Cell model [2,7,24] (since it only uses a single global parameter for the results illustrated in Figures 1, 2 and 3). It has been argued that in addition to quantitative fit, simplicity and parsimony are also important and desirable characteristics for potentially valid computational models [72–74]. Thus, we believe it is important to consider not only models that are capable of fitting data very well, but also models that offer simple

explanations, and we have described such a model, using a Bayesian framework and a single parameter.

It has been suggested earlier [29] that sensory information might be used to correct path integration error. Previous work building on this idea can be categorized into high-level models, suggesting correction mechanisms but unconcerned with the details of neuronal implementation, and neuronal-level models.

High-level models of hippocampal error correction have proposed a Bayesian information integration mechanism before [11–14]. Cheung et al. [14] show that featureless boundaries alone are insufficient for unambiguous localization, and propose a similar model of Bayesian localization to the one outlined here, based on the implementation of a particle filter, and replicate some experimental results on place and grid field stability using their high-level model. However, they do not account for single cell firing field data, and they do not suggest how the particle filter might be implemented in the brain. MacNeilage et al. [13] suggest Bayesian cue integration to estimate spatial orientation under uncertainty, suggesting Kalman filters (which use unimodal Gaussian probability distributions) or, alternatively, particle filters (which are capable of dealing with multimodal and non-Gaussian probability distributions) as the mechanistic implementation. Pfuhl et al. [12] also hypothesize spatial information integration to be Bayesian, choosing Kalman filters as their implementation. Finally, Cheng et al. [11] propose that spatial information is integrated in a Bayesian fashion, without suggesting a formal model or a neuronal implementation, and provide some behavioural evidence for this claim.

Kalman filters are possible to implement on biologically plausible attractor networks [51], although they have the disadvantage of being unable to deal with multimodal, non-Gaussian distributions. Taking a different approach, Samu et al. [82] have used a recurrently interconnected attractor network to correct path integration errors, using sensory information via hippocampal back-projections. Their model, like most attractor-based path integration models, relies on recurrent interconnections (which area CA1 of the hippocampus seems to lack [3]). Extending their ideas, Fox and Prescott [95] have attempted to map the hippocampal formation onto a temporal restricted Boltzmann machine (and argue that inference in their model resembles particle filtering), also modelling on a functional level but trying to adhere more closely to anatomical connectivity. However, like the previously mentioned concrete computational models, they do not model empirical data to substantiate their model. Using oscillatory interference theory instead of an attractor model as their theoretical basis, Monaco et al. [96] also use cue-driven feedback to correct location errors and to handle cue conflicts. They also reproduce partial remapping in an experiment, strengthening the mechanism the model uses to resolve cue conflicts. Cue-driven location correction is also employed in the model proposed by Sheynikhovich et al. [97], in the form of connections between view cells and grid cells, weighted using Hebbian learning.

Unlike many of these models, apart from presenting a high-level model of Bayesian cue integration, we have also attempted to suggest a tentative neuronal mechanism that might underly the implementation of approximate inference. Starting from mathematical theory, a number of implementations of Bayesian inference have already been proposed (e.g. [43,49–52]), although none known to the authors in the context of HEC error correction. We believe the inference mechanism described in the Results section offers a useful contribution, because most previously published spiking neuron inference mechanisms predict anatomical and firing properties inconsistent with some empirical observations if applied to place cells. For example, the distribution population coding method [98] assumes prespecific tuning functions and a sophisticated decoding operation with unclear neuronal implementation. Inference mechanisms based on a log probability population code [52] have more plausible decoding schemes, but require recurrent connectivity and global recurrent inhibition, which have only been observed in CA3, not in CA1 place cells [99], in contrast to physiological data from CA1 suggestive of Bayesian inference (see Results section). In addition, they assume specific weight matrices for statistical optimality - which could be learned in principle, but would require a non-Hebbian learning rule. Finally, probabilistic population codes (PPC) have been widely used in modelling inference [50], recently also supported by physiological data [100]. However, PPCs have no clear way to implement learning [101], and they also require recurrent connections [50]. Furthermore, the standard PPC inference scheme assumes Poisson-like variability to allow simple addition to implement inference [43,50], which implies a direct relationship between the absolute firing rates of neurons in a PPC and the uncertainty (standard deviation) of the encoded distribution - a relationship predicted by most inference schemes. However, it has been observed that place cell firing rates increase with the animals movement speed [67] - if place cells used a PPC with Poisson variability, or any other probabilistic encoding scheme predicting such a relationship, this would imply that the faster they would run, the more certain they would become of their location (location uncertainty would decrease with increasing running speed), which is counter-intuitive and contradicts the frequently observed trade-off between speed and accuracy [102].

The model we propose has its own shortcomings, but is simple and does not depend on specific weight matrices or variability distributions. Our aim was to show that even without additional assumptions regarding connectivity, weights, or learning, the anatomy of the Hippocampal-Entorhinal Complex might be able to implement approximate Bayesian inference. Although we were unable to substantiate this tentative model with physiological data as of yet, we hope that the reported results will encourage future research addressing the often sceptically regarded [6] mechanistic 'Bayesian brain'.

Supporting Information

Text S1 Location uncertainty in the two-dimensional case.
(PDF)

Text S2 Coincidence detection as rejection sampling and multiplication by coincidence detection.
(PDF)

Acknowledgments

The authors gratefully thank Carol A. Barnes and Sara N. Burke for kindly providing the place field dataset on the circular track, and also acknowledge the thought-provoking personal communications about the topic with Máte Tóth, Armin Basic, and the helpful comments of Steve Strain, who has commented on the manuscript.

Author Contributions

Conceived and designed the experiments: TM DM. Performed the experiments: TM. Analyzed the data: TM. Contributed reagents/materials/analysis tools: SF KC RT. Wrote the paper: TM DM. Critical revision of manuscript: SF KC DM.

References

1. O'Keefe J, Dostrovsky J (1971) The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. Brain Research 34: 171–175.
2. Burgess N (2008) Spatial cognition and the brain. Annals of the New York Academy of Sciences 1124: 77–97.
3. Moser EI, Kropff E, Moser MB (2008) Place cells, grid cells, and the brain's spatial representation system. Annual Review of Neuroscience 31: 69–89.
4. Solstad T, Moser EI, Einevoll GT (2006) From grid cells to place cells : a mathematical model. Hippocampus 1031: 1026–1031.
5. Etienne AS, Maurer R, Sguinot V (1996) Path integration in mammals and its interaction with visual landmarks. Journal of Experimental Biology 199: 201–9.
6. Jeffery KJ (2007) Self-localization and the entorhinal-hippocampal system. Current Opinion in Neurobiology 17: 684–91.
7. Hartley T, Burgess N, Lever C, Cacucci F, O'Keefe J (2000) Modeling place fields in terms of the cortical inputs to the hippocampus. Hippocampus 10: 369–79.
8. Knill DC, Pouget A (2004) The Bayesian brain: the role of uncertainty in neural coding and computation. Trends in Neurosciences 27: 712–9.
9. Ernst MO, Banks MS (2002) Humans integrate visual and haptic information in a statistically optimal fashion. Nature 415: 429–33.
10. Körding KP, Ku Sp, Wolpert DM (2004) Bayesian integration in force estimation. Journal of Neurophysiology 92: 3161–3165.
11. Cheng K, Shettleworth SJ, Huttenlocher J, Rieser JJ (2007) Bayesian integration of spatial information. Psychological Bulletin 133: 625–37.
12. Pfuhl G, Tjelmeland H, Biegler R (2011) Precision and reliability in animal navigation. Bulletin of Mathematical Biology 73: 951–77.
13. MacNeilage PR, Ganesan N, Angelaki DE (2008) Computational approaches to spatial orientation: from transfer functions to dynamic Bayesian inference. Journal of Neurophysiology 100: 2981–96.
14. Cheung A, Ball D, Milford M, Wyeth G, Wiles J (2012) Maintaining a cognitive map in darkness: the need to fuse boundary knowledge with path integration. PLoS Computational Biology 8: e1002651.
15. Colombo M, Seriès P (2012) Bayes in the brain - on Bayesian modelling in neuroscience. The British Journal for the Philosophy of Science 63: 697–723.
16. Hafting T, Fyhn M, Molden S, Moser M, Moser E (2005) Microstructure of a spatial map in the entorhinal cortex. Nature 436: 801–806.
17. McNaughton BL, Battaglia FP, Jensen O, Moser EI, Moser MB (2006) Path integration and the neural basis of the 'cognitive map'. Nature Reviews Neuroscience 7: 663–78.
18. O'Keefe J, Burgess N (2005) Dual phase and rate coding in hippocampal place cells: theoretical significance and relationship to entorhinal grid cells. Hippocampus 15: 853–866.
19. Doeller CF, Barry C, Burgess N (2012) From cells to systems : grids and boundaries in spatial memory. The Neuroscientist 18: 556–566.
20. Taube JS (2007) The head direction signal: origins and sensory-motor integration. Annual Review of Neuroscience 30: 181–207.
21. Baumann O, Mattingley JB (2010) Medial parietal cortex encodes perceived heading direction in humans. Journal of Neuroscience 30: 12897–12901.
22. Lever C, Burton S, Jeewajee A, O Keefe J, Burgess N (2009) Boundary Vector Cells in the subiculum of the hippocampal formation. Journal of Neuroscience 29: 9771–7.
23. Solstad T, Boccara CN, Kropff E, Moser MB, Moser EI (2008) Representation of geometric borders in the entorhinal cortex. Science 322: 1865–8.
24. Barry C, Lever C, Hayman R, Hartley T, Burton S, et al. (2006) The boundary vector cell model of place cell firing and spatial memory. Reviews in the Neurosciences 17: 71–97.
25. O'Keefe J, Dostrovsky J (1971) The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. Brain Research 34: 171–175.
26. Ekstrom AD, Kahana MJ, Caplan JB, Fields TA, Isham EA, et al. (2003) Cellular networks underlying human spatial navigation. Nature 424: 184–187.
27. Prusky GT, West PW, Douglas RM (2000) Behavioral assessment of visual acuity in mice and rats. Vision Research 40: 2201–2209.
28. Okada K, Fujimoto Y (2011) Grid-based localization and mapping method without odometry information. In: IECON 2011-37th Annual Conference on IEEE Industrial Electronics Society. IEEE, pp. 159–164.
29. McNaughton BL, Barnes CA, Gerrard JL, Gothard K, Jung MW, et al. (1996) Deciphering the hippocampal polyglot: the hippocampus as a path integration system. Journal of Experimental Biology 199: 173–185.
30. Squire LR, Stark CEL, Clark RE (2004) The medial temporal lobe. Annual Review of Neuroscience 27: 279–306.
31. Montaldi D, Mayes AR (2010) The role of recollection and familiarity in the functional differentiation of the medial temporal lobes. Hippocampus 20: 1291–1314.
32. Lisman J, Redish AD (2009) Prediction, sequences and the hippocampus. Philosophical transactions of the Royal Society of London Series B, Biological Sciences 364: 1193–201.
33. Bird CM, Burgess N (2008) The hippocampus and memory: insights from spatial processing. Nature reviews Neuroscience 9: 182–94.
34. Lee SA, Sovrano VA, Spelke ES (2012) Navigation as a source of geometric knowledge: Young children's use of length, angle, distance, and direction in a reorientation task. Cognition 123: 144–61.
35. Young BJ, Fox GD, Eichenbaum H (1994) Correlates of hippocampal complex-spike cell activity in rats performing a nonspatial radial maze task. The Journal of neuroscience 14: 6553–6563.
36. Yoshioka JG (1929)Weber's law in the discrimination of maze distance by the white rat. University of California Publications in Psychology 4: 155–184.
37. Cheng K, Spetch ML (1998) Landmark-based spatial memory in birds and mammals. In: Healy S, editor, Spatial Representation in Animals, New York: Oxford University Press. pp. 1–17.
38. Negenborn R (2003) Robot localization and Kalman filters. Ph.D. thesis, Utrecht University.
39. Durrant-Whyte H, Bailey T (2006) Simultaneous localization and mapping: Part 1. IEEE Robotics Automation Magazine 13: 9–110.
40. Bromiley P (2003) Products and convolutions of Gaussian distributions. Medical School, Univ Manchester, Manchester, UK, Tech Rep 3: 2003.
41. Ahmed O, Mehta M (2009) The hippocampal rate code: anatomy, physiology and theory. Trends in neurosciences 32: 329–338.
42. Burke SN, Maurer AP, Nematollahi S, Uprety AR, Wallace JL, et al. (2011) The inuence of objects on place field expression and size in distal hippocampal CA1. Hippocampus 21: 783–801.
43. Ma WJ, Beck JM, Pouget A (2008) Spiking networks for Bayesian inference and choice. Current Opinion in Neurobiology 18: 217–22.
44. Koch C, Segev I (2000) The role of single neurons in information processing. Nature Neuroscience 3 Suppl: 1171–1177.
45. Jarsky T, Roxin A, Kath WL, Spruston N (2005) Conditional dendritic spike propagation following distal synaptic activation of hippocampal CA1 pyramidal neurons. Nature Neuroscience 8: 1667–1676.
46. Takahashi H, Magee JC (2009) Pathway interactions and synaptic plasticity in the dendritic tuft regions of CA1 pyramidal neurons. Neuron 62: 102–111.
47. Katz Y, Kath WL, Spruston N, Hasselmo ME (2007) Coincidence detection of place and temporal context in a network model of spiking hippocampal neurons. PLoS Computational Biology 3: e234.
48. Nezis P, Van Rossum MCW (2011) Accurate multiplication with noisy spiking neurons. Journal of Neural Engineering 8: 034005.
49. Deneve S (2008) Bayesian spiking neurons I: inference. Neural computation 20: 91–117.
50. Ma WJ, Beck JM, Latham PE, Pouget A (2006) Bayesian inference with probabilistic population codes. Nature Neuroscience 9: 1432–1438.
51. Deneve S, Duhamel JR, Pouget A (2007) Optimal sensorimotor integration in recurrent cortical networks: a neural implementation of Kalman filters. The Journal of neuroscience 27: 5744–5756.
52. Rao RPN (2004) Bayesian computation in recurrent neural circuits. Neural Computation 16: 1–38.
53. Hoyer PO, Hyvärinen A (2003) Interpreting neural response variability as Monte Carlo sampling of the posterior, MIT Press, volume 15. p. 293.
54. Büsing L, Bill J, Nessler B, Maass W (2011) Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons. PLoS Computational Biology 7: e1002211.
55. Lee TS, Mumford D (2003) Hierarchical Bayesian inference in the visual cortex. Journal of the Optical Society of America A 20: 1434–1448.
56. Rossant C, Leijon S, Magnusson A, Brette R (2011) Sensitivity of noisy neurons to coincident inputs. The Journal of Neuroscience 31: 17193–17206.
57. Brette R (2012) Computing with neural synchrony. PLoS Computational Biology 8: e1002561.
58. Szilagyi E, Halasy K, Somogyi P (1996) Physiological properties of anatomically identified basket and bistratified cells in the CAl area of the rat hippocampus in vitro. Hippocampus 6: 294–305.
59. Zemankovics R, Káli S, Paulsen O, Freund T, Hájos N (2010) Differences in subthreshold resonance of hippocampal pyramidal cells and interneurons: the role of h-current and passive membrane characteristics. The Journal of physiology 588: 2109–2132.
60. Harvey C, Collman F, Dombeck D, Tank D (2009) Intracellular dynamics of hippocampal place cells during virtual navigation. Nature 461: 941–946.
61. Hoppensteadt F, Izhikevich E (1997) Weakly connected neural networks, volume 126. Springer.
62. Markus E, Barnes C, McNaughton B, Gladden V, Skaggs W (2004) Spatial information content and reliability of hippocampal CA1 neurons: effects of visual input. Hippocampus 4: 410–421.
63. Quirk G, Muller R, Kubie J (1990) The firing of hippocampal place cells in the dark depends on the rat's recent experience. The Journal of Neuroscience 10: 2008–2017.
64. Amaral DG, Ishizuka N, Claiborne B (1990) Neurons, numbers and the hippocampal network. Progress in Brain Research 83: 1–11.
65. Rapp P, Gallagher M (1996) Preserved neuron number in the hippocampus of aged rats with spatial learning deficits. Proceedings of the National Academy of Sciences 93: 9926–9930.
66. Barry C, Bush D (2012) From A to Z: A potential role for grid cells in spatial navigation. Neural systems & circuits 2: 6.

67. Maurer AP, Vanrhoads SR, Sutherland GR, Lipa P, McNaughton BL (2005) Self-motion and the origin of differential spatial scaling along the septo-temporal axis of the hippocampus. Hippocampus 15: 841–52.
68. Odobescu R (2010) Exteroceptive and interoceptive cue control of hippocampal place cells. Ph.D. thesis, UCL (University College London).
69. O'Keefe J, Burgess N (1996) Geometric determinants of the place fields of hippocampal neurons. Nature 381: 425–428.
70. Brette R, Rudolph M, Carnevale T, Hines M, Beeman D, et al. (2007) Simulation of networks of spiking neurons: a review of tools and strategies. Journal of computational neuroscience 23: 349–398.
71. Goodman DF, Brette R (2009) The brian simulator. Frontiers in neuroscience 3: 192.
72. Myung IJ, Pitt MA (1997) Applying Occam's razor in modeling cognition: A Bayesian approach. Psychonomic Bulletin & Review 4: 79–95.
73. Myung IJ, Pitt MA, Kim W(2005) Model evaluation, testing and selection. Handbook of cognition : 422–436.
74. Regier T (2003) Constraining computational models of cognition. In: Nadel L, editor, Encyclopedia of Cognitive Science, London: Macmillan. pp. 611–615.
75. Nardini M, Jones P, Bedford R, Braddick O (2008) Development of cue integration in human navigation. Current Biology 18: 689–93.
76. Shadlen M, Britten K, Newsome W, Movshon J (1996) A computational analysis of the relationship between neuronal and behavioral responses to visual motion. The Journal of Neuroscience 16: 1486–1510.
77. Stocker AA, Simoncelli EP (2006) Noise characteristics and prior expectations in human visual speed perception. Nature Neuroscience 9: 578–585.
78. Canto C,Wouterlood F, Witter M (2008)What does the anatomical organization of the entorhinal cortex tell us? Neural plasticity 2008.
79. Kajiwara R, Wouterlood FG, Sah A, Boekel AJ, Baks-te Bulte LT, et al. (2008) Convergence of entorhinal and CA3 inputs onto pyramidal neurons and interneurons in hippocampal area CA1 - an anatomical study in the rat. Hippocampus 18: 266–280.
80. Witter M (2011) Entorhinal cortex. Scholarpedia 6: 4380.
81. Burgess N, OKeefe J (2011) Models of place and grid cell firing and theta rhythmicity. Current opinion in neurobiology 21: 734–744.
82. Samu D, Eros P, Ujfalussy B, Kiss T (2009) Robust path integration in the entorhinal grid cell system with hippocampal feed-back. Biological Cybernetics 101: 19–34.
83. Sreenivasan S, Fiete I (2011) Grid cells generate an analog error-correcting code for singularly precise neural computation. Nature neuroscience 14: 1330–1337.
84. Bonnevie T, Dunn B, Fyhn M, Hafting T, Derdikman D, et al. (2013) Grid cells require excitatory drive from the hippocampus. Nature neuroscience 16: 309–317.
85. Burgess N, Barry C, O'Keefe J (2007) An oscillatory interference model of grid cell firing. Hippocampus 17: 801–812.
86. Hasselmo ME (2008) Grid cell mechanisms and function: contributions of entorhinal persistent spiking and phase resetting. Hippocampus 18: 1213–1229.
87. Zilli Ea (2012) Models of grid cell spatial firing published 2005–2011. Frontiers in Neural Circuits 6: 1–17.
88. Engel TA, Schimansky-Geier L, Herz AV, Schreiber S, Erchova I (2008) Subthreshold membranepotential resonances shape spike-train patterns in the entorhinal cortex. Journal of neurophysiology 100: 1576–1589.
89. Dickson CT, Magistretti J, Shalinsky M, Hamam B, Alonso A (2000) Oscillatory activity in entorhinal neurons and circuits: Mechanisms and function. Annals of the New York Academy of Sciences 911: 127–150.
90. Dickson CT, de Curtis M (2002) Enhancement of temporal and spatial synchronization of entorhinal gamma activity by phase reset. Hippocampus 12: 447–456.
91. Deshmukh SS, Knierim JJ (2013) Inuence of local objects on hippocampal representations: Landmark vectors and memory. Hippocampus 23: 253–267.
92. Azzalini A (2005) The Skew-normal Distribution and Related Multivariate Families*. Scandinavian Journal of Statistics 32: 159–188.
93. Mehta MR, Quirk MC, Wilson MA (2000) Experience-dependent asymmetric shape of hippocampal receptive fields. Neuron 25: 707–715.
94. Brun VH, Solstad T, Kjelstrup KB, Fyhn M, Witter MP, et al. (2008) Progressive increase in grid scale from dorsal to ventral medial entorhinal cortex. Hippocampus 18: 1200–1212.
95. Fox CW, Prescott TJ (2010) Hippocampus as unitary coherent particle filter. In: IJCNN. IEEE Press, pp. 1–8.
96. Joseph D Monaco JJK, Zhang K (2011) Sensory feedback, error correction, and remapping in a multiple oscillator model of place cell activity. Frontiers in Computational Neuroscience 5.
97. Sheynikhovich D, Chavarriaga R, Strosslin T, Arleo A, Gerstner W (2009) Is there a geometric module for spatial orientation? Insights from a rodent navigation model. Psychological review 116: 540.
98. Zemel R, Dayan P, Pouget A (1998) Probabilistic interpretation of population codes. Neural Computation 10: 403–430.
99. Lee I, Yoganarasimha D, Rao G, Knierim JJ (2004) Comparison of population coherence of place cells in hippocampal subfields CA1 and CA3. Nature 430: 456–459.
100. Yang T, Shadlen MN (2007) Probabilistic reasoning by neurons. Nature 447: 1075–1080.
101. Fiser J, Berkes P, Orbn G, Lengyel M (2010) Statistically optimal perception and learning: from behavior to neural representations. Trends in Cognitive Sciences 14: 119–130.
102. Hancock PA, Newell KM (1985) The movement speed-accuracy relationship in space-time. In: Motor Behavior, Springer. pp. 153–188.

Chapter 5

The structure of spatial representations

Publication 3 / 4. Madl T., Franklin S., Chen K., Trappl R. & Montaldi D., submitted. Exploring the structure of spatial representations. *Cognitive Processing*

Exploring the structure of spatial representations

Tamas Madl[a,b,*], Stan Franklin[c], Ke Chen[a], Robert Trappl[b], Daniela Montaldi[d]

[a]*School of Computer Science, University of Manchester, Manchester M13 9PL, UK*
[b]*Austrian Research Institute for Artificial Intelligence, Vienna A-1010, Austria*
[c]*Institute for Intelligent Systems, University of Memphis, Memphis TN 38152, USA*
[d]*School of Psychological Sciences, University of Manchester, Manchester M13 9PL, UK*

Abstract

It has been suggested that the map-like representations that support human spatial memory are fragmented into sub-maps with local reference frames, rather than being unitary and global. However, the principles underlying the proposed structure of these 'cognitive maps' are not well understood.

We propose that the structure of the representations of navigation space arises from clustering, i.e. from a process that groups together objects that are close in a given psychological space, and we present evidence for this claim based on participants' long-term spatial memories regarding buildings in real-world, as well as virtual reality, environments. We compare plausible dimensions of this psychological space, including spatial distance, visual similarity and functional similarity, and report strong correlations between these dimensions and the grouping probability in participants' spatial map structures, which empirically support the clustering hypothesis.

In addition, we also present the first formal predictive model of human navigation-scale spatial representation structure, based on the Bayesian cognition paradigm, and show that this probabilistic model of clustering, when provided with information regarding psychological spaces, learned from subjects, allows the prediction of their cognitive map structures for the first time.

Keywords:
Spatial representations, cognitive maps, hierarchical cognitive maps, spatial structure, spatial memory, computational cognitive modeling

1. Introduction

There has been considerable research on spatial representations facilitating navigation since Tolman coined the term 'cognitive map' (Tolman, 1948). Since then, the neural bases of such allocentric (world-centered) representations of space have been identified in rats (O'Keefe & Nadel, 1978; McNaughton et al., 2006) and humans (Ekstrom et al., 2003; Barry et al., 2006) and have been shown to play a vital role in representing locations within the environment in long-term memory. Instead of learning a single spatial map with a global reference frame, as proposed originally (Tolman, 1948; O'Keefe & Nadel, 1978), humans (as well as some non-human animals) seem to form structured spatial maps, consisting of multiple 'sub-maps', i.e. multiple representations containing spatial information about sub-sets of objects in the environment, with separate local frames of reference.

Behavioural evidence has suggested that human spatial maps are structured, and has been interpreted as comprising multi-level hierarchies (Hirtle & Jonides, 1985; McNamara, 1986; McNamara et al., 1989; Holding, 1994; Wiener & Mallot, 2003), or at least as having multiple local reference frames (Meilinger et al., 2014; Greenauer & Waller, 2010). These hierarchies, extracted from recall sequences, can be observed even in the case of randomly distributed objects with no boundaries (McNamara et al., 1989), with participants' response

*tamas.madl@gmail.com

times and accuracies being affected by this structure (subjects overestimated distances between objects on different branches of the hierarchy and underestimated distances within branches, and showed shorter response times for within-branch judgements). Further evidence for the existence of multiple representations in different spatial reference frames (Greenauer & Waller, 2010; Shelton & McNamara, 2001; Meilinger et al., 2014) has been derived from the accuracies of judgements of relative direction, which are heavily affected by subjects' frames of reference.

In addition to behavioural data, there is also strong neuroscientific evidence for hierarchical spatial representations (Brun et al., 2008; Kjelstrup et al., 2008), and for fragmentation into sub-maps (Derdikman & Moser, 2010) in mammalian brains. Finally, organized and structured maps (instead of a single representation) are consistent with 'chunking' long-term memory (Gobet et al., 2001) and with hierarchical models of cognition (Cohen, 2000), and have multiple information processing advantages, including the increased speed and efficiency of retrieval search, and economical storage.

The rate at which results about structured cognitive maps (navigation-space allocentric representations) in humans have been published has declined since the pioneering work of the eighties and nineties, partly because of some controversy surrounding the term 'cognitive map' [1]. The methodological difficulties plaguing behavioural research into the organization of cognitive maps are additional likely reasons for this decline. Unfortunately, humans do not have introspective insight into the structure of their cognitive maps. Thus, map structure can only be inferred indirectly, with a small set of possible behavioural paradigms such as those tapping recall patterns or priming effects, which are prone to noise (see Section 4, General Discussion, for a comparison of advantages and disadvantages of different methods).

Although map structure is not introspectively accessible nor immediately apparent, it does play an important role in spatial cognition. It has been shown in experiments involving priming, distance and angle estimations, and sketch maps, that the speed and accuracy of subjects at various spatially relevant tasks are significantly influenced by how they represent space (Hirtle & Jonides, 1985; McNamara et al., 1989; Han & Becker, 2014; Hommel et al., 2000). In addition to helping us understand the influence on cognitive performance, a model of cognitive map structure could facilitate several neighbouring fields, including human-robot interaction (allowing robots to use human-like spatial concepts), artificial intelligence (use insights from human memory to improve artificial memory), and geographic information science (present spatial information in a more easily comprehensible and memorable fashion) - see Section 4.1.

Despite the importance of this question, and perhaps because of the above-mentioned difficulties, no formal theories or models concerning the organizational principles of cognitive maps, able to account for empirical data, have been published since cognitive maps were first proposed to be structured. Little progress has been made on explaining how these representations might be structured in non-trivial, open environments. Multiple features influencing map structure have been suggested, including boundaries in the environment (Wang & Spelke, 2002; Barry et al., 2006), spatial distance and familiarity (Hirtle & Jonides, 1985), action-based and perception-based similarity (Hommel et al., 2000; Hurts, 2008), and functional / semantic similarity (Holding, 1994). However, to the authors' best knowledge, these influences have never been compared based on behavioural data.

A few formal models of map structure do exist - which are predominantly empirically untested -, e.g. the graph-based model by (Thomas & Donikian, 2007) for outdoor virtual reality environments, or based on predicate logic (Reineking et al., 2008), for indoor environments (neither of these have been evaluated against human data); as well as more neuronally plausible but functionally simpler models of place cells such as (Sato & Yamaguchi, 2009) (also empirically untested), or the model by (Byrne et al., 2007) (which can account for lesion effects in humans, but not for large-scale cognitive map structure). Voicu (2003) has published

[1] Some researchers have argued that humans depend on landmark-based instead of map-based navigation whenever they can (Foo et al., 2005), and that most animal behaviour can be explained without the cognitive map hypothesis (Bennett, 1996). However, the well-established body of neuroscientific evidence for dedicated brain regions containing allocentric spatial representations (Moser et al., 2008; Derdikman & Moser, 2010) - both in human and non-human mammals -, together with the ability of human subjects to plan complex novel shortcuts or detours or produce sketch maps, render the idea of allocentric, map-like representations - at least in humans - difficult to dismiss. On the other hand, 'cognitive maps' might well be different from geographical maps in several respects, including being limited in scope, detail, and accuracy, being dynamic, and possibly using metrics that are not (or not exclusively) Euclidean (Spelke et al., 2010; Jeffery, 2015).

the modelling work closest in spirit to the predictive models reported below, utilizing self-organizing maps to model hierarchical cognitive map structure, and reporting that on average, the model exhibits similar distance estimation error patterns to the estimation biases (averaged over all subjects) reported by Hirtle & Jonides (1985). However, this model has not been compared to individual subject maps; and is unable to account for per-subject data in Hirtle's dataset, since it uses only Euclidean spatial distance and no other features (whereas many of Hirtle's subjects do not cluster exclusively based on spatial distance). To date, no empirically tested, formally defined model exists that would be able to predict, or even quantitatively explain, the structure of the individual spatial maps constructed by humans in unconstrained large-scale environments; and the features of the psychological spaces [2] underlying such a model have not yet been explored empirically.

Formulating models and testable hypotheses precisely and unambiguously, is important for efficiently driving research, especially in interdisciplinary areas such as spatial memory (which is of interest in psychology, neuroscience, and artificial intelligence, among other fields). Computational cognitive models are well suited to this challenge as such unambiguous formal descriptions, and provide a common language across disciplines, as well as the additional advantage of very fast prediction generation and hypothesis testing (once the data has been collected, such models can be rapidly run and verified on computers). Thus, they play an important role in the cognitive science of spatial memory, helping to integrate findings, to generate, define, formalize and test hypotheses, and to guide research.

In order to develop and validate a computational model of cognitive map structure, it is necessary - but not sufficient - to tackle the methodological difficulties associated with indirectly inferring consciously inaccessible spatial representation structure from noisy data. In addition, there are also computational challenges. Just like brains can be said to create object representations based on perceived and remembered properties of objects, a computational cognitive model also needs such representations, capturing relevant features. Furthermore, a method is needed that helps to decide which representations should be grouped together on sub-maps. While many low-level features of these representations can be neglected for simplicity in a model on Marr's computational level (Marr & Poggio, 1976), an appropriate metric[3] for capturing similarities between objects is crucially important for exploring how object representations are grouped together onto sub-maps by the brain. Various features, with different levels of importance, can influence whether objects belong together; and defining a metric is a way to formally account for these 'feature importances' (Figure 1). Although the entorhinal cortex has been argued to contain two-dimensional metric grids analogous to graph paper (Jeffery & Burgess, 2006), recent evidence implies the brain's distance metric to be locally distorted (Jeffery, 2015), non-Euclidean (Spelke et al., 2010), and dependent on the above-mentioned features of familiarity, functional similarity, and perceptual similarity (Hommel et al., 2000; Hurts, 2008; Holding, 1994). These features might not be equally important, and their relative importance might not be the same across individuals.

Thus, finding a metric under which objects grouped together by human subjects are 'closer', or more similar, than those not grouped together, is vital for modelling an individual's spatial representation structure. Using this metric, a model can generate predictions regarding which group or sub-map a new object might belong to. Alternatively, objects can be represented in a metric space (which models a subject's psychological space), spanned by relevant features constituting the axes of that space, weighted by their importance. In fact, learning a projection into such a space, in which objects which belong together have a smaller Euclidean distance than those which do not, and learning a metric under which this grouping

[2] By a 'psychological space' we mean a metric space within which the similarities of objects can be represented as distances between points in that space; consistent with the pioneering models of stimulus identification (Shepard, 1957) or categorization (Nosofsky, 1986) or, most recently, conceptual spaces (Gärdenfors, 2004). By a 'feature' of the psychological space, we mean one of the dimensions of this space, which allows measuring similarity along a single aspect, such as the functional similarity of buildings.

[3] A metric (or distance function) is a function that defines a non-negative 'distance' between pairs of objects. Two well-known examples include the Euclidean (geodesic) distance, and the taxicab (Manhattan) distance. In this paper, we model the dissimilarity between two building representations by means of their 'distance' according to a learned metric, which operates on multiple features including but not limited to spatial position (with a distance/dissimilarity value of 0 meaning that the representations are equivalent).

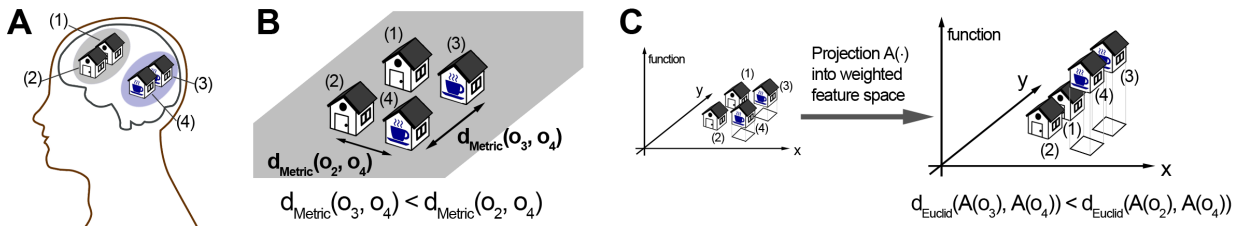


Figure 1: Formalizing relative feature importances for grouping objects. Panel A: A subject might group (represent) the two coffee shops together (buildings 3 and 4), even if they are spatially farther apart from each other than to other houses; i.e. (3) and (4) are psychologically closer (more similar) for that individual than (2) and (4). The idea of some features being more important than others when grouping objects can be formally captured either by defining a metric d_{Metric} reflecting the subject's psychological similarity by weighting features appropriately (panel B), or by projecting objects into a feature space (psychological space) spanned by weighted features, in which the Euclidean distances d_{Euclid} are consistent with the subjects' psychological similarity (panel C). The central challenge for a predictive model of spatial groupings is learning these feature importances or 'weights' which parametrize d_{Metric} (or the projection function A) from subjects. Representing subjects' psychological similarities by applying these weights in a distance metric is equivalent in outcome to representing them by functions projecting into a weighted 'psychological space' (similar objects will appear closer under the learned metric / in the learned space).

relationships hold, are two views on the same problem, and the solutions are mathematically analogous (see Figure 1 for an informal and footnote 21 in Section 3.5.1 for a formal argument). In both cases, the solution involves learning parameters corresponding to the relative importances of the features for a given subject. Although traditionally, cognitive psychology has mostly used the former approach, using multidimensional scaling (MDS) to project into a psychological space correctly reflecting similarities (Shepard, 1957), this method is not applicable in our case (the reasons for this are outlined in Section 2.4).

This paper aims to tackle the above-mentioned challenges associated with exploring the structure of spatial representations, and to take a first step towards establishing a formal and empirically substantiated model of this structure. Our core hypothesis is that **the structure of spatial representations in humans arises from a process of clustering** of the represented objects, in a psychological space characterised by multiple relevant information types (features) including the ones mentioned above. Clustering extracts groups or clusters by assuming that objects belonging to the same group (in our case, sub-map) are closer to each other within psychological space than objects belonging to different groups (sub-maps). The characteristics of the psychological space within which this clustering takes place, i.e. which features are relevant and how important they are, has to be learned from participants' responses. The main contributions of this paper are as follows.

1. We present evidence for the clustering hypothesis both in virtual reality and in real-world environments, and compare the influence of several information types (features) on cognitive map structure, and the stabilities of these feature influences across environments and subjects.
2. We show that the structure of spatial representations, far from being a confounding effect of the recall process or a minor mechanistic detail of memory, has an important role in, and influence on, multiple cognitive phenomena, including (but not limited to) planning, distance estimation, memory accuracy, and response times.
3. We propose and evaluate three computational methods to learn models of subject-specific psychological spaces (either in the form of weighted feature spaces or as distance metrics), even if only small amounts of training data are available
4. We present the first (to our best knowledge) quantitative model able to predict individual cognitive map structures in navigation space, and evidence supporting it.

We only use a few simple types of features for modelling and prediction. Nevertheless, and despite the large amounts of unreliability and noise both in these features and in the participant responses, we show that **spatial map structures can be predicted** for human subjects, in a large number of real and virtual environments. We adopt the behavioural methodology used by (Hirtle & Jonides, 1985; McNamara et al.,

1989; McNamara, 1986; Holding, 1994) among others, which infers subjects' representation structure based on recall sequences, and assumes that objects recalled together belong to the same sub-map. Despite of some shortcomings (see Sections 2.1 and 4), the clear and significant influence of the resulting map structures on several kinds of cognitive phenomena (as well as its prior success at showing the influence of hierarchical cognitive maps) lend credence to this method.

Finally, we also make freely available as a web application the experiment software developed to investigate cognitive map structure, at `https://github.com/tmadl/Cognitive-Map-Structure-Experiment`, with the aim to encourage future work on this important but neglected research area.

2. Experimental paradigm

We investigated the structure of spatial representations in navigation space in three experiments. All of the experiments were concerned with the representations of buildings and their relation to each other. In Experiments 1 and 3, subjects recalled real-world buildings that they were already highly familiar with (see Figure 2). In Experiment 2, subjects were presented with three-dimensional virtual reality environments - containing buildings with automatically generated properties - which they had to memorize prior to the recall task from which the representation structure was inferred (see Figure 3).

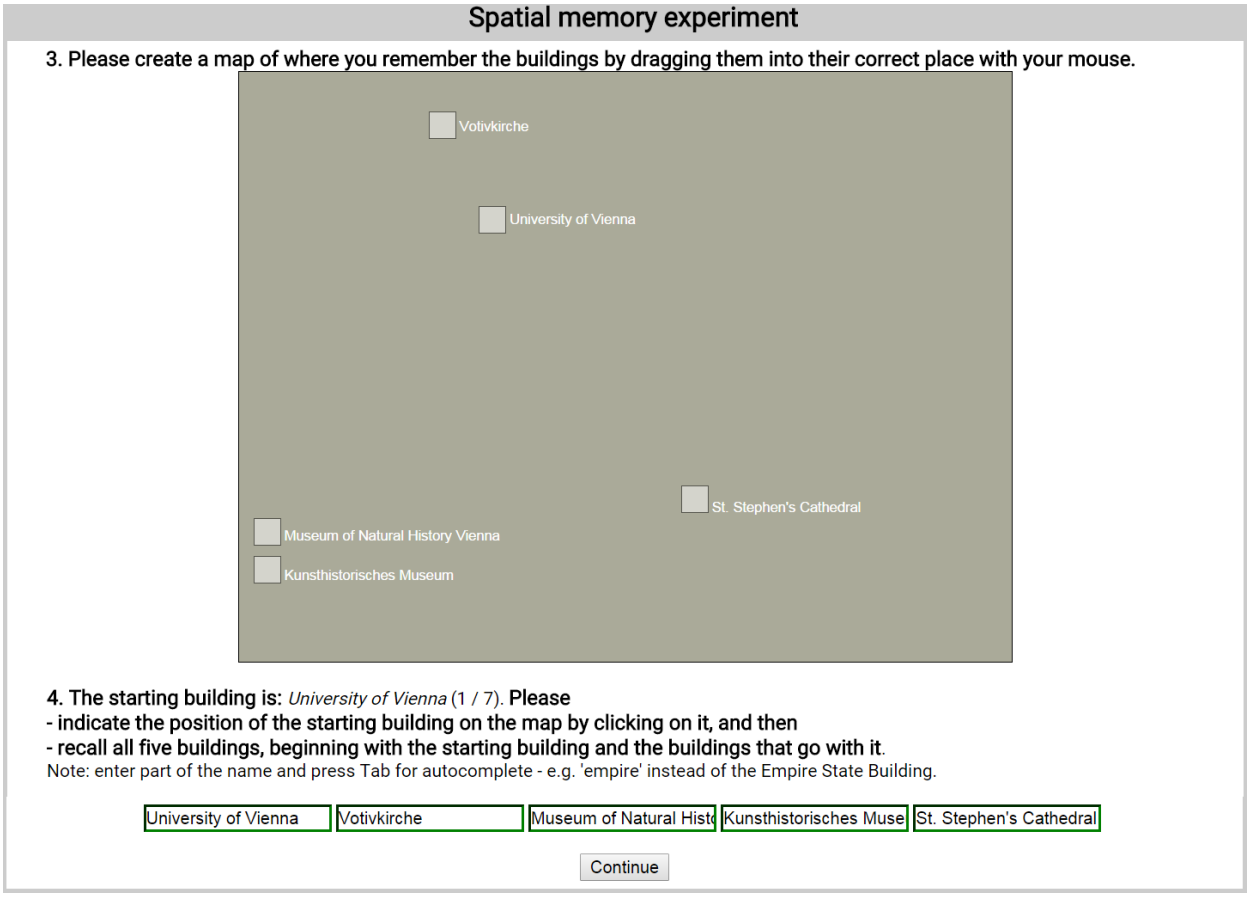


Figure 2: A part of the real-world memories experiment interface of Experiments 1 and 3, with the sketch map question for verifying that subjects have indeed formed allocentric cognitive maps (top), and the recall sequence question requiring them to recall every single building name multiple times (bottom). During this recall question the labelled sketch map was not visible to subjects.

Figure 3: A part of the virtual reality experiment interface of Experiment 2 (the recall sequence interface was equivalent to the real-world experiments; see Figure 2)

2.1. Extraction of spatial representation structure

To extract the structure of spatial representations, we use a variant of ordered tree analysis on subjects' recall sequences, a behavioural methodology used by (Hirtle & Jonides, 1985; McNamara et al., 1989; McNamara, 1986; Holding, 1994) among others for extracting hierarchies in spatial representations, and by (Naveh-Benjamin et al., 1986; Reitman & Rueter, 1980) for verbal stimuli. The core assumption behind this methodology is that objects recalled together belong to the same representation; i.e. that on the whole, subjects recall every object within a representation (or sub-map) before moving on to the next representation (see Figure 4). Tree analysis operates on a set of recall sequences (with each sequence consisting of all object names, recalled with a particular ordering - usually different from the other recall sequences -, as exemplified in Figure 2A). Variety among these recall sequences is encouraged by cueing subjects with the object they are required to start with (and only uncued parts of the sequence are analysed to avoid the interference of the cue) (Hirtle & Jonides, 1985).

To briefly summarize the collection of these recall sequences (for details, see Section 3.2): in each trial, subjects were first asked to pick a few buildings (5 or 8) within walking distance of each other, which they were very familiar with, and where they knew how to walk from any one building to any other. Subsequently, they were asked to recall the complete list (i.e. recall sequence) of their chosen buildings, starting with a cue building (except for two interspersed uncued trials), multiple times. If building names were missing or incorrect, subjects were prompted again, until they got all of them right. Thus, the ordering within the individual sequences was their only variable aspect.

After obtaining the recall sequences, for each subject, the algorithm simply iterates through all possible combinations of subsets of object names in each recall sequence, finds those subsets which consistently appear together in all sequences (regardless of order), and constructs a hierarchy based on containment relationships from the subsets of items occurring together. The original algorithm also extracts directionality information for each group (whether the items within that group have always been recalled using a consistent ordering). We do not use the order information in the recall sequences in this work (see Supplementary Information for the algorithm we have used). Figure 4 A shows example abbreviated recall sequences, and the resulting tree structure, where each branch or sub-map consists of items which always occur together in the sequences. Unambiguous sub-map memberships are obtained at the level just above the leaf nodes, defining sub-maps

as elementary sets of co-occurring items, i.e. those which do not themselves contain further co-occurring items. This procedure partitions buildings into one or two sub-maps in Experiments 1, 2 and 3A, and up to four sub-maps in Experiment 3B.

Since this tree analysis algorithm requires buildings to be recalled together in every single recall sequence in order to infer subjects' sub-maps, it is very sensitive to individual inconsistencies that may result from lapses of attention, task interruptions, and other kinds of noise within participant response (see Section 4 for a discussion and comparison with other approaches of inferring cognitive map structure). To mitigate this, we have eliminated 'outlier' recall sequences, defined as sequences which would have statistically significantly altered the structure if they were included (whereas all others would not). As proposed in previous work on hierarchical cognitive maps (Hirtle & Jonides, 1985; McNamara, 1986; McNamara et al., 1989), we used jackknifing to eliminate outliers. For each sequence, this procedure calculates how the inferred tree structure would change if the sequence were omitted. Trees were quantified using two statistics, tree height and log-cardinality (the logarithm of the number of possible recall sequences consistent with that tree). These statistics were calculated for the tree resulting from all sequences of one trial, as well as for all trees that would result from possible sequence omissions (i.e. if only sequences excluding the omitted one had been entered by the participant). If any of the sequence omissions lead to a statistically significant change in the tree statistics, at a significance level of $\alpha = 0.05$[4], then that sequence was deemed an outlier and was omitted, and the tree resulting from the other sequences of that trial was used for further analysis. All sequences except for outliers were consistent with the same tree structure. Thus, outlier sequences, which significantly changed the tree structure, were likely to arise from the above-mentioned sources of noise (lapses of attention, interruptions, etc.). The outlier sequences detected and removed by the jackknifing procedure comprised 8.5% in Experiment 1, 10.0% in Experiment 2 and 9.5% in Experiment 3, corresponding to less than one omission per subject (across the 7 recall sequences produced per subject and trial).

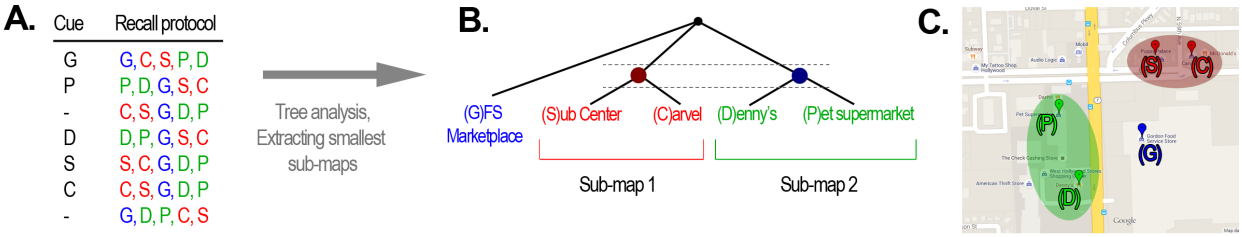


Figure 4: The recall sequence-based method used to extract cognitive map structure. A: Example recall sequences of one of the participants of Experiment 3. Each building was cued once, with two uncued recall trials interspersed (full building names abbreviated by their first character). B: Hierarchical tree structures were constructed by tree analysis, based on the assumption that buildings always recalled together belong to the same sub-map. C: Geographic map of the buildings recalled by this participant. Sub-maps shown in colour, according to the extracted structure.

To simplify the analysis, we subsequently extract the elementary sub-maps (those not containing smaller sub-maps) from the constructed tree - this allows us to model sub-maps, as opposed to full hierarchies. These elementary sub-maps must contain at least two buildings. If a sub-map only contains a single object, then this object is excluded from subsequent analysis. The main reason being that our hypothesis implies sub-maps to be clusters or groups of objects; however, there is no way to verify the plausibility of a single-object cluster (as opposed to clusters containing multiple buildings, for which performance consequences such as between/within-cluster distance biases, priming effects etc. can be investigated - see Section 3.2 for evidence). A further reason for the exclusion of single-object sub-maps is that they were likely to actually be parts of bigger sub-maps in subjects' spatial memories, together with additional buildings not captured

[4]We used a less conservative significance criterion than prior work due to the simpler structures and smaller numbers of objects used (using the extremely conservative significance level of $\alpha = 0.001$ used in the prior work cited above would have led to zero outliers being detected - presumably incorrectly, since it is unlikely that not a single participant would have had any interruptions or lapses of attention).

due to the necessarily limited number of recalled items per trial in our experiments. The exclusion of these single buildings did not have an impact on the plausibility of our claims, since two sub-maps containing pairs of buildings suffice for comparing within sub-map and across sub-map estimations in order to investigate whether map structure has an effect on spatial cognition (see Section 3.2. Experiment 3B collected map structures with eight buildings and up to four sub-maps to show that the model is not limited to two).

A final difference between our methodology and prior uses of the recall order paradigm is the repetition with several different geographic environments for each subject in Experiment 3. Repeatedly extracting cognitive map structures from the same participants is not only interesting, e.g. to compare the variability of the features in subjects' psychological spaces, but also of vital importance for producing and validating a predictive model of the structure of spatial representations. Given the large inter-subject variability in terms of features and feature importances influencing map structure, parametrizing such a predictive model necessitates gathering multiple different cognitive map structures from separate environments (and not just one structure), both for training the model, and for subsequently testing it. The main differences between a repeated and a single-trial paradigm include possible effects of fatigue due to the increased length of experiments, as well as declining accuracy of representations towards the later stages (participants started struggling to cue readily available buildings which they could accurately draw on a map beyond 20 buildings, as evidenced by much slower progress, higher error rates, and much higher rate of participants abandoning the experiment as compared to Experiment 1 which used single trials).

An attempt to mitigate these effects - as well as practical limitations - motivated the decision to use a smaller number of buildings (five in Experiments 1 and 2, five and eight in 3 A and B) compared to the single-trial setup of (Hirtle & Jonides, 1985; McNamara et al., 1989), who used 32 and 28 objects, respectively. Using their dozens buildings for each of the five or three map structures of Experiment 3 would have required participants to recall (and accurately localize) around one hundred buildings or more - as well as judging all of their pairwise similarities, the number of which increases quadratically with the number of buildings (in the case of 32 buildings, they would amount to 496 similarity judgements each for visual and functional similarities, and for each trial, which is nowhere near feasible).

2.2. Experimental platforms and participants

Participants in two of the three Experiments (1 and 3) were recruited from the online survey website Amazon Mechanical Turk (MTurk)[5]. Multiple psychological findings have been replicated before, using subjects from MTurk (Crump et al., 2013), showing the breadth of this platform for psychological experimentation. MTurk offers a participant pool that is significantly more diverse than samples of university students, containing subjects from many countries worldwide and of different age groups; as well being several orders of magnitude larger than most universities' subject pools. But the most important advantage offered by this platform lay in facilitating the collection of information about spatial representations of many, very different geographic environments. Such variety is critical for two main reasons:

- To facilitate generalizable observations (for example, insights from inflexibly planned city areas such as the grid layout of Manhattan might not have been generalizable to other street layouts), and
- To avoid local biases (for example, using exclusively local maps in the same city for each participant might have led to conclusions about the spatial structure of the local city, reflected in subjects' representations, as opposed to insights into the way subjects structure space in general).

Our objective of collecting cognitive map structures from a large variety of different geographical environments was indeed successful - we collected data and analysed spatial representations from several environments within **149 different cities** across multiple continents (see Figure 5 - a list of these cities can be found in the Supplementary Information).

[5] https://www.mturk.com

Figure 5: Overview over the 149 cities in which participants' spatial memory structures were extracted (and predicted by the computational model) in the real-world experiments (full list in Supplementary Information)

2.3. Exclusion of participant maps not significantly better than random chance

Throughout this paper, we have only analysed participants' data if their sketch maps were significantly better than random chance, in order to avoid false conclusions about cognitive maps being made on the basis of non-allocentric representations. Since route knowledge suffices for navigating between buildings, participants might have lacked survey knowledge about some of the buildings in these experiments. To rule out participant data not showing evidence of allocentric cognitive maps, we first performed a test of participants' sketch maps against randomness, before carrying out the subsequent analyses described in Section 3.

We compared the sum of squared errors (SSE) calculated by subtracting the positions of buildings on participants' sketch maps from those on the correct geographical map (obtained from Google Maps), with the SSEs of 10,000 randomly generated maps containing 5 buildings against the correct map. Since subjects' sketch maps were produced on empty surfaces without any position, orientation or scale cues, as seen in Figure 2, they were first aligned (translated, rotated, and scaled) with the correct map using Procrustes analysis (Gower, 1975) without reflection. The randomly generated maps were also aligned in the same fashion. The distribution of the SSEs of 10,000 [6] Procrustes-aligned random maps was then used to test whether subject maps were better than random. Specifically, subject SSEs were tested against the null hypothesis that they were drawn from the distribution of uniformly random map SSEs. Two different significance tests were applied at $\alpha = 0.05$ significance level, and found to largely agree (in all but 3% of the cases in Experiment 1, 1.3% in Exp. 2 and 4% in Exp. 3): a Z-test assuming normal distributions of SSEs, and a non-parametric Bootstrap hypothesis test (MacKinnon, 2009), which requires subject maps to be better than a proportion of $1 - \alpha$ of the random maps.

Because the former test makes the assumption of normally distributed data, which is incorrect for the vast majority of distributions of random map SSEs according to Shapiro-Wilk tests, we use the latter, non-parametric hypothesis testing method throughout this paper to test participant maps against randomness.

2.4. Data analysis

Subject data collected using the recall order paradigm was analysed as follows. Maps not significantly better than random (and corresponding recall lists), and recall sequences not containing structure (where no items consistently occur together), were not analysed further, since the former show no evidence of allocentric representations of the buildings on that map being present in the subjects' spatial memory, and the latter

[6] We have also tried higher numbers or randomly generated maps, and found that 10,000 samples suffice for an approximation of the distribution under the null hypothesis, since increasing this number to 15,000 or 20,000 did not make a difference.

shows no structure to be analysed. Next, map structures (sub-map memberships) were derived using tree analysis, and pairwise distances in all features were calculated (see Section 3.3). This data allows reporting the influence of features on map structure, and the inter- and intra-subject variability of this influence (Section 2.2). It also allows inferring subject-specific models reflecting the individual feature importances of a subject, in the form of a metric or a 'psychological space' (subject-specific feature space) - see Figure 1. A clustering algorithm (described in Section 3.4), operating on the learned model, allows prediction of sub-map structure, if the clustering hypothesis is plausible (assuming that cluster memberships correspond to sub-map memberships).

Simple clustering based on a Euclidean distance metric, in the original feature space, fails to account for participant map structures. The main reason for this is that different participants might not rely on the same set of features; and the relative importance of these features might also differ across subjects (see below, particularly Figure 7, for evidence). For this reason, learning an appropriate subject-specific model (metric, or feature space) is crucial in order for our computational model to provide accurate predictions. Since learning human representation similarity metrics (or psychological spaces) from highly noisy and sparse data is a largely unexplored problem in the cognitive sciences[7], we turn to machine learning for possible solutions. We describe and empirically test three computational methods for tackling this problem below (see Sections 3.4.1 and 3.5.1, as well as the Supplementary Information for details).

Together, these learned subject-specific models, and the clustering algorithm, constitute a computational model of cognitive map structure learning able to predict sub-map structure in advance, and allowing the verification of such predictions (Sections 3.4 and 3.5 compare the model predictions against human data).

3. Experiments

3.1. Overview of the experiments

This section reports the results of four experiments investigating the principles underlying cognitive map structure. Experiment 1 (Section 3.2) is concerned with the question of whether this kind of structure uncovered by the recall order paradigm is relevant - whether it impacts cognitive performance in other ways than recall sequences -, investigating effects on distance estimation biases, sketch map accuracies, estimated walking times, and planning times in real-world environments well known to subjects.

The plausibility of our central hypothesis - that cognitive map structure arises from clustering - is investigated in the subsequent section (3.3), also in real-world environments chosen by subjects themselves. This claim requires buildings that are more similar (closer in psychological space) to be more likely to be grouped together in long-term memory representations, and thus more likely to be recalled together. We report correlations between the probability of buildings being represented together, and proximity in various features relevant to cognitive mapping. We also make between- and across-subject comparisons with regard to feature importances.

Since a good model should be able to make predictions, we proceed to report the predictability of spatial representation structure. We use a clustering model in order to predict map structure, assuming that cluster membership in an appropriate 'psychological space' (i.e. weighted feature space) corresponds to sub-map memberships.

However, a model clustering buildings in a static feature space fails to produce accurate predictions, simply because there exists no feature space generalizable across participants (see Section 3.3). In order to learn subject-specific feature spaces, we utilize three methods to uncover the features and feature importances spanning the psychological space hypothesized to underlie spatial representation grouping in Experiments 2 and 3. We collect map structures of several different environments from the same subjects in these experiments, using a subset of them to learn a model, and testing its predictions on the remaining subset.

[7] There has been an approach used in cognitive psychology for projecting data into a space in which distances reflect subject similarities, called multidimensional scaling (MDS) (Shepard, 1957). This method is not applicable in our case, because it requires a full pairwise distance matrix. However, our training data comes from several different environments; and pairwise distances and similarities are only known within, and not across, those environments.

In Experiment 2 (Section 3.4), subjects are asked to learn spatial memories of 3D virtual reality environments. Unlike the other experiments, this approach allows full control over all properties of the stimuli being memorized. Utilizing this flexibility of virtual reality, we report prediction results using clustering, and the decision hyperplane method for learning subject-specific models, which tackles the challenge of inferring multiple feature importances from few data points by generating the environments such that participants' subsequent responses minimize the uncertainty of the model regarding the feature importances, inspired by active learning in machine learning (Settles, 2010). After subjects have been queried on a reasonable number of environments, and the model's uncertainty regarding their psychological space has decreased, they are presented with completely random environments, on which the trained models are tested. We report prediction accuracies both on environments generated such that they minimize model uncertainty (using active learning), and on random environments.

Although virtual reality allows fine-grained control over memorized environments, it is necessarily composed of strongly simplified stimuli and less complex surroundings. To show that the approach of inferring subject-specific models and subsequently clustering objects can also successfully predict cognitive map structure in the much more complex real world, we once again collect data from subjects' spatial memories of real environments freely chosen by them in Experiment 3 (Section 3.5). Since the approach of optimally minimizing model uncertainty is infeasible when using uncontrolled real-world memories, we use two more general methods to infer subjects' psychological spaces, global optimization and Gaussian Discriminant Analysis (see Section 3.5.1). Of these, the latter is novel, and to our knowledge the only metric learning approach applicable to our data. We report prediction results on data excluded from the model training process, substantiating our central hypothesis, and showing, for the first time, the predictability of spatial representation structures on the individual level.

3.2. Experiment 1 - Relevance of cognitive map structure extracted from recall sequences

This experiment was conducted to substantiate the recall order paradigm used throughout this paper to infer cognitive map structure. If this paradigm infers something about actual representation structures in spatial memory, then the uncovered structures should have a significant impact on both the speed and accuracy of memory recall for spatially relevant information. To avoid possibly confounding effects of stimulus presentation and memorization, the stimuli used were ones participants had already committed to their long-term spatial memory - the experiment used buildings subjects were already very familiar with and could easily recall information about[8].

Although data consistent with two of the results presented in this section (the effects of map structure on distance estimation biases and sketch map accuracies) have been observed and published before, this prior work had used significantly fewer subjects than our experiment, and exclusively university students, unlike our participants. (Hirtle & Jonides, 1985) had six participants, reporting distance biases and sketch map accuracies; and (McNamara et al., 1989) had twenty eight, reporting only the former.

3.2.1. Participants

One hundred and fifty participants were recruited, consented, and compensated through the Amazon Mechanical Turk (MTurk) online survey system (78 females, 74 males). Participants were required to have at least 95% approval rating on previous MTurk jobs to ensure higher data quality, and all of them were over 18 years of age (as required by the website).

3.2.2. Procedure

The experiment was conducted on a website participants could access through MTurk after giving their consent. In the first two questions, subjects were asked to enter the name of a city they were very familiar

[8]The possible objection that the structures might be induced by the experimental paradigm, and learned by participants during the trials, can be excluded, because of the approximately uniform distribution of the outlier sequences (the first few sequences were not more likely to be outliers than the last few sequences, and no evidence for any learning of map structures during the real-world experiments could be found in the data - see Supplementary Information for details).

with, and, subsequently, to pick five buildings they know well. Thus, well-established long-term memories were tested instead of novel stimuli. Subjects were instructed to make sure that they knew where in the city these buildings were located, how to walk from any one building to any of the others, what each building looked like, and what purpose it served. They were only able to proceed past this stage if the website was able to locate all five of the buildings on a geographical map (Google Maps API[9] was used to look up the latitude and longitude of each building).

To verify that subjects had indeed formed allocentric spatial representations of the area of the city they had selected, and to allow the analysis of the accuracy of their representations, they were also asked to produce a 'sketch map', by dragging and dropping five labelled squares representing their buildings into their correct place using their mouse (Figure 4A, top). No cues or information was provided on the sketch map canvas, just an empty gray surface with five squares labelled according to subjects' entered building names. Thus, only the relative configuration of the buildings was analysed in this research, after optimal translation, rotation and scaling to fit the placement and size of the correct map as well as possible, by using Procrustes transformation (Gower, 1975).

After the sketch map, subjects performed a seven-trial recall test. In five of the seven trials, they were given a cue or starting building, and were instructed to *'recall all five buildings, beginning with the starting buildings and the buildings that you think go with it'*, encouraging recall of building names in the order they came to subjects' mind, closely following the instructions given by (Hirtle & Jonides, 1985; McNamara et al., 1989) and others. In the remaining two, uncued trials, subjects were asked to start with any building they wished. If subjects omitted or incorrectly recalled any of the buildings, they had to repeat the trial (thus, only the ordering changed across trials).

The recall test allowed the experiment software to immediately infer subjects' map structure using the tree analysis algorithm (see Section 2.1. Smallest sub-maps - those not containing further sub-maps - were extracted). The next stage of the experiment proceeded based on this structure. Participants were first asked to estimate the time required to walk between four pairs of buildings. Unbeknownst to them, two of the estimations concerned within-, and two of them across-sub-map pairs, in randomized order, and were generated such that the Euclidean distances in the within-cluster trials were as close as possible to the distances in the across-cluster trials, to mitigate effects of simple distance, as opposed to map structure. After reading the instructions in their own time, subjects were told to estimate and enter the walking time in minutes (the time required to walk from one of these buildings to another) as rapidly as possible. Their responses, as well as their response times (time elapsed between presentation of the pair of buildings for walking time estimation and subjects entering a number and clicking a button) were recorded.

In a subsequent stage, also based on the uncovered map structure, participants had to estimate the distance between four pairs of buildings (Euclidean distance - 'as the crow flies' - as opposed to the walking times of the previous stage). Once again, two within-cluster and two across-cluster pairs were selected such that within- and across-cluster trials differed as little as possible from each other in terms of spatial distance.

Finally, once again in an untimed fashion, subjects were asked to judge the similarities of all pairs of buildings, i.e. $\binom{5}{2} = 10$ pairs, as well as a control pair of one of the buildings to itself, both in terms of visual similarity, and similarity of purpose/function - thus, they had to enter 2x11 similarity judgements. Similarities were judged with the help of 1-10 rating scales, with 1 standing for not similar and 10 for very similar. The two self-similarity judgements were randomly interspersed and verified to avoid subjects rushing the process or entering random values.

Ground truth geographical maps containing participants' self-chosen buildings were constructed by obtaining latitude and longitude coordinates from Google Maps API, and utilizing an elliptical Mercator projection to obtain x and y coordinates suitable for comparison with subjects' sketch maps. Euclidean distances between buildings were also calculated based on this projection (as this procedure is more accurate than most alternatives such as the Haversine formula). Finally, path distances as well as ground truth walking times were obtained from Google Directions API [10], which plans the shortest possible walking route

[9] https://developers.google.com/maps/

[10] https://developers.google.com/maps/documentation/directions/

	Actual distance (m)	**Estimated distance (m)**	**Distance bias (Estimated-Actual)**	**Estimated walking time (min:sec)**	**Response time when estimating walking time (s)**
Within mean Within std	$\mu = 1242$, $\sigma = 1508$	$\mu = 676$, $\sigma = 1036$	$\mu = -574$, $\sigma = 1825$	$\mu = 8:43$, $\sigma = 8:23$	$\mu = 8.4$, $\sigma = 6.0$
Across mean Across std	$\mu = 1245$, $\sigma = 1931$	$\mu = 1139$, $\sigma = 1739$	$\mu = -146$, $\sigma = 1703$	$\mu = 12:45$, $\sigma = 11:36$	$\mu = 18.0$, $\sigma = 92.3$
Significance of difference	$p = 0.109$ (nonsignificant), $U = 12594$	$p = 0.019$ (significant), $U = 12502$	$p = 0.047$ (significant), $U = 11900$	$p = 0.001$ (significant), $U = 11009$	$p = 0.030$ (significant), $U = 13097$

Table 1: Effects of spatial representation structure on distance estimation, walking time estimation, and response times. All of these estimated magnitudes, as well as response times, are significantly smaller when both buildings are on the same sub-map (i.e. on the same representation) compared to when they are not. Data from 380 pairs of buildings were compared (269 across sub-maps, and 111 within sub-map). Apart from the representation-dependent biases, subject estimations were reasonably accurate (correlation of $r = 0.40$ between estimated and actual Euclidean distance, and $r = 0.48$ between estimated and actual walking time as calculated by Google Maps)

between two buildings along pedestrian paths (which is usually distinct from, and longer than, Euclidean or 'beeline' distance).

3.2.3. Results

Participants with sketch maps not significantly better than random chance were excluded (using the procedure described in Section 2.3). 86 participants with reasonably accurate survey knowledge of their chosen environments remained (40 female, 46 male). Of these participants, 53 had structure apparent in their recall sequences (20 female, 33 male). The difference in the ratio of structured representations between male (72%) and female (50%) participants is statistically significant at $p = 0.04$ ($U = 4.39$) according to a Mann-Whitney U test. We employed this test here and for a majority of our other significance tests (unless otherwise specified), because the tested variables were not normally distributed according to a Shapiro-Wilk normality test ($p = 0.00$, $W = 0.63$), violating the assumptions behind ANOVA or t-testing. The Mann-Whitney test is a nonparametric test which has greater efficiency than the t-test on non-normal distributions (and is comparably efficient to the t-test even on normal distributions) (Nachar, 2008).

To test whether map structure has an impact on other cognitive phenomena, we compared estimations of distance, walking times, and planning times, between pairs of buildings lying on the same representation (within sub-map estimations), and pairs of buildings on different representations (across sub-map estimations). Table 1 reports the results (6 across sub-map and 1 within sub-map distance estimations were excluded, because they exceeded $10km$, clearly violating the instruction of being within walking distance). Reported correlations are Spearman's correlation coefficients, here as well as throughout the paper.

In order to avoid effects arising purely from differences in spatial distance, we have queried subjects on the pairs of their buildings (among all possible pairs) which were the least different in spatial distance. In these comparisons, effects purely of spatial distance are unlikely, since distances were not significantly different between within sub-map and across sub-maps estimations ($1242m$ and $1245m$ on average) - according to a Mann-Whitney test ($U = 12594$, $p = 0.11$), the difference is not significant.

We have also examined the effect of whether maps were structured on sketch map accuracies. The sum of squared errors (SSE) between the resulting sketch map building positions and the geographical building positions were calculated, and SSEs for all maps with structure ($\mu = 0.305$, $\sigma = 0.276$) were compared to the SSEs for maps without structure ($\mu = 0.370$, $\sigma = 0.307$). SSEs were found to be significantly smaller for structured than for unstructured maps ($p = 0.019$, $U = 2325$), hinting at a correlation between map accuracy and structuredness which can indeed be observed ($r = -0.17$, $p = 0.04$).

Finally, the SSEs between sketch map and geographic map distances were compared for pairs of within

sub-maps and pairs across sub-maps, after alignment and normalization. The sketch map distance SSEs within sub-maps ($\mu = 0.607$, $\sigma = 1.677$) were significantly smaller than those across sub-maps ($\mu = 0.916$, $\sigma = 1.53$) according to a Mann-Whitney U test ($p = 0.023$, $U = 6304000$).

3.2.4. Discussion

The highly significant differences in the accuracies of sketch maps, distance and walking time estimations, which all depend on whether or not the buildings involved in the estimation are on the same sub-map or on different sub-maps, substantiate the claim that the structures uncovered by this method are indeed relevant, and play a significant role in multiple cognitive mechanisms.

The trends in the distance error biases - distances generally being underestimated within sub-maps compared to across sub-map estimates - match previously made observations using smaller numbers of subjects (Hirtle & Jonides, 1985). The main difference is that this previous work has found underestimation within- and overestimation across sub-maps, whereas our results suggest underestimation in both cases. The negativity (underestimation) of the across sub-map distance estimation errors is statistically significant compared to the null hypothesis that there is zero or positive bias ($p = 0.03$, $U = 4937$).

Both the difference in estimated walking times, and the differences in the response time in this question, are novel results. As opposed to Euclidean distance estimation or sketch map drawing, which can be done by glancing at or recalling a geographical map, accurate walking times are difficult to estimate without actually having explored this environment and being able to plan the routes in question. Subjects need to mentally plan the route and simulate the walk to estimate the time (or to recall the duration of the walk from long-term memory, should the durations of all walks between all possible building pairs be readily memorized by subjects, which is unlikely). The observation that the mean time required to do so more than doubles across sub-maps, compared to within (and that the variance in RTs increases by an order of magnitude) provides additional, substantial evidence for the relevance of map structures - as inferred from recall sequences - to spatial cognitive processes.

3.3. Clustering and features determining map structure

In the Introduction, we have hypothesized that the structure of spatial representations in humans arises from clustering within some psychological space. In this section, we investigate the plausibility of this hypothesis. If this was the case, we would expect the probability of a pair of buildings being co-represented (i.e. represented on the same sub-map) to strongly depend on their 'similarity' or distance across various features including spatial distance, with stronger dependencies for spatially relevant features compared to semantic or visual features. We would also expect several such features to play a role, since distance alone is insufficient to explain previous results (Hirtle & Jonides, 1985; McNamara et al., 1989). We would expect the relevance of each feature to be apparent from its influence on map structure, measurable by the correlation between co-representation probability (the probability that two buildings are co-represented on the same sub-map) and the distance in this feature. Finally, we would expect large inter- but small intra-subject variability in these correlations, i.e. stable feature relevances within subjects which are not necessarily generalizable across subjects, analogously to psychological spaces for concept representation (Nosofsky, 1986; Gärdenfors, 2004).

We investigate several features listed below, motivated by hints in the literature that they might play a role in the representation structure of object-location memory.

1. Remembered distance, i.e. the distance on subjects' sketch maps
2. True Euclidean distance based on geographical maps
3. Path distance (or 'city-block' / 'Manhattan' distance), since recent brain imaging evidence suggests that the hippocampus - a spatially relevant brain region - represents both Euclidean and path distances (Howard et al., 2014)
4. Boundaries in the environment (such as rivers, cliffs, city walls, etc.) - based on neuroscientific and behavioral evidence that boundaries play an important role in spatial memories (Wang & Spelke, 2002; Barry et al., 2006)

5. The number of streets separating a pair of buildings (intersecting with a straight line connecting these buildings)
6. The sizes of separating streets; that is, whether these streets could easily be crossed (whether or not they were highways/motorways/primary roads which are difficult for pedestrians to cross)
7. Visual similarity (as indicated by participants), since clustering by perceptual properties has been reported (Hommel et al., 2000), and vision has been suggested to be vital to spatial representation (Ekstrom, 2015),
8. Functional similarity, or similarity of purpose, as indicated by participants - because action-based similarity has been claimed to have an effect on spatial memory (Hurts, 2008), and also because of the importance of action-related roles within the influential grounded cognition paradigm (Barsalou, 2008).
9. Phonetic [11] and morphological [12] similarity of building names. The main motivation behind including these features was to investigate any possible interference on the structures inferred from recall sequences caused by verbal short-term representations. Subjects might employ some short-term representation strategy to complete the recall trials more rapidly - instead of recalling from long-term spatial memory -, for example subvocal rehearsal loops (articulatory loops). Including phonetic and morphological similarity features helps measure the effect of such verbal strategies.

The first six of these features - remembered, Euclidean and path distance, and boundaries, separating streets, and crossable streets - were obtained based on geographical data available online. Most such geospatial ground truth data used was obtained using Google's publicly available Maps API, with the exception of boundaries in the environment, and crossable streets (whether separating streets were difficult to cross) - these two features were obtained from Open Street Maps (OSM) through their publicly available API called Overpass [13]. As in all experiments in this paper, ground truth maps and distances are based on an elliptical Mercator projection of latitudes and longitudes obtained from Google Maps API, except for path distances and walking times which were queried from Google Directions API.

All features were converted into distances / dissimilarities before subsequent analysis. Similarity features, such as visual, functional, phonetic and morphological similarities, were subtracted from the maximum value possible for that feature to obtain corresponding dissimilarities.

3.3.1. Participants, Materials, and Procedure

Data from Experiment 1 (Section 3.2 above) as well as Exp. 2 (see Section 3.4) and Exp. 3 A and B (see Section 3.5) were analysed with regard to the plausibility of the clustering hypothesis, as well as the underlying features determining map structure. Thus, the participants, materials and procedures for data collection were exactly the same as in those experiments, following the recall order paradigm described in Section 2.

All Figures in this Section are split into four parts, for Experiment 1, Exp. 2, and conditions A and B of Experiment 3. We report results separately, since there were slight changes in procedure. Briefly, Exp. 2 was conducted in three-dimensional virtual reality environments, whereas the other experiments used subjects' established real-world spatial memories. Furthermore, cues were presented verbally in Exp. 1 and 2 and spatially, highlighted on sketch maps, in Exp. 3 (Exp. 3B also used 8 buildings, unlike the 5 used in the other experiments). Finally, Experiments 2 and 3 tested spatial memories of several different environments, in order to facilitate learning a model and testing predictions, whereas Exp. 1 did not.

[11] Phonetic similarities have been determined using the Double Metaphone (Philips, 2000) phonetic encoding algorithm, since it is more accurate than older alternatives such as Soundex, and also accounts for a large number of irregularities in multiple languages, not just English (vital since many participants from several non-English speaking countries participated in this experiment - see Supplementary Information for a detailed list). For building names consisting of multiple words, the sum of the phonetic similarities of the constituent words was used.

[12] Morphological similarities were calculated based on recent work by (Khorsi, 2013), implemented by the PhonologicalCorpusTools library accessible at http://kchall.github.io/CorpusTools/. For building names consisting of multiple words, the sum of the morphological similarities of the constituent words was used.

[13] http://overpass-api.de/

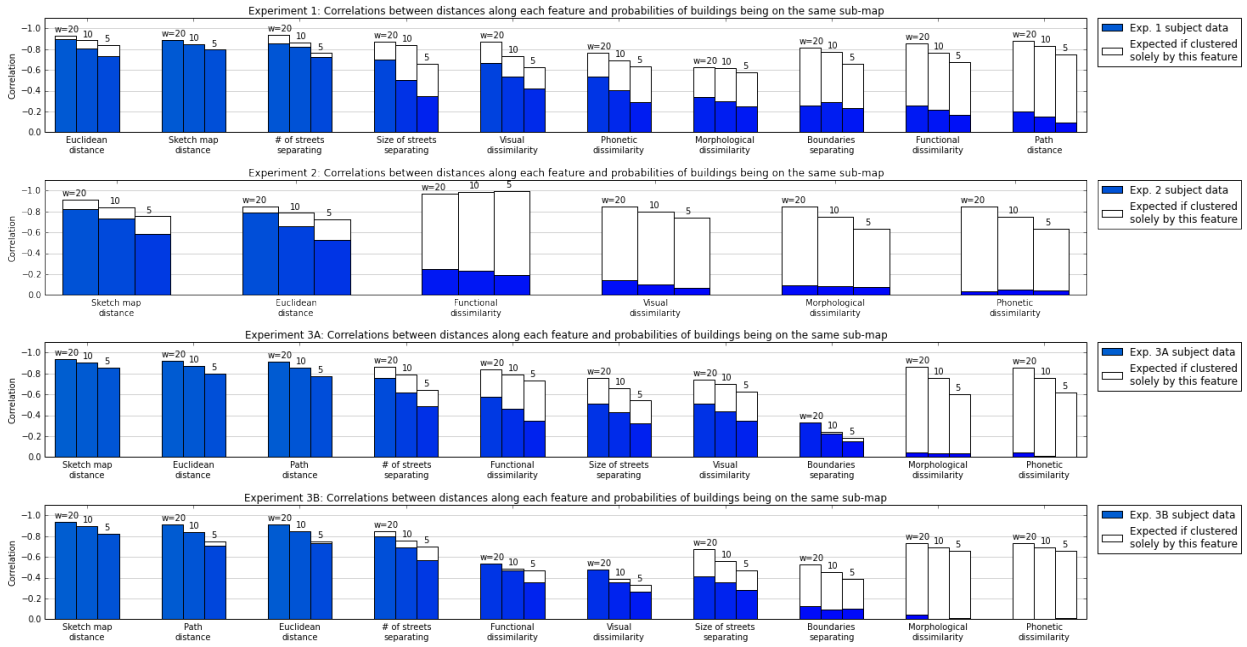


Figure 6: Correlations between probabilities of being on the same sub-map, and distances along each feature, for pairs of buildings in Experiments (from top to bottom): 1, Experiment 2 in virtual reality (therefore lacking geospatial features), and 3A, 3B. Correlations are reported separately for each feature. The three bars per feature show results at three different window sizes w used for calculating co-representation probabilities (higher w lead to less noisy probability estimates through smoothing, resulting in higher correlations). Empty bars show levels of correlation that would be expected if maps were clustered according to the single respective feature only.

3.3.2. Results

The clustering hypothesis introduced in Section 1 implies that buildings closer together in psychological space are more likely to be represented on the same sub-map in participants' spatial memory. To test this hypothesis, we investigated the correlations between the probabilities of pairs of buildings belonging on the same sub-maps and between the distance between them, along the various features listed above.

Figure 6 provides an overview of the correlations of these features with the probabilities of co-representation on the same sub-map. These probabilities were calculated using a moving average with window w of the binary vector indicating whether or not pairs were stored on the same sub-map - simply put, the likelihood of co-representation at a specific distance equals the ratio of the number of co-represented pairs divided by the number of all pairs within some small window w close to this distance (for example, if $w = 3$ and out of three building pairs with distances $95m$, $100m$ and $105m$ two were represented on the same sub-map, then the probability of co-representation at $100m$ would equal $p = 2/3$).

The Figure also shows the correlations that could be expected if participant's map structure had arisen from clustering by just that one feature (empty bars in Figure 6) - i.e. the correlations that would have been observed had participants 1) used clustering to structure their maps, and 2) used only distances within one respective feature for this clustering. These expected correlations were calculated using the same participant data; but artificially structuring the subject map - using clustering along one respective feature - instead of using subjects' sub-map memberships. Gaussian mixture models (GMMs) (Redner & Walker, 1984) were used for the artificial structuring, just like for prediction in the computational models described below, since they are more psychologically motivated than other clustering algorithms (see Section 3.4.1).

Next, we have investigated the variability of the reliance of these features within and across subjects; i.e. whether the same features were used by - and whether they were similarly important for - all subjects, and whether they were the same for individual participants in different environments. Figure 7 shows the standard deviations of the co-representation correlations of these features, across subjects (left panels)

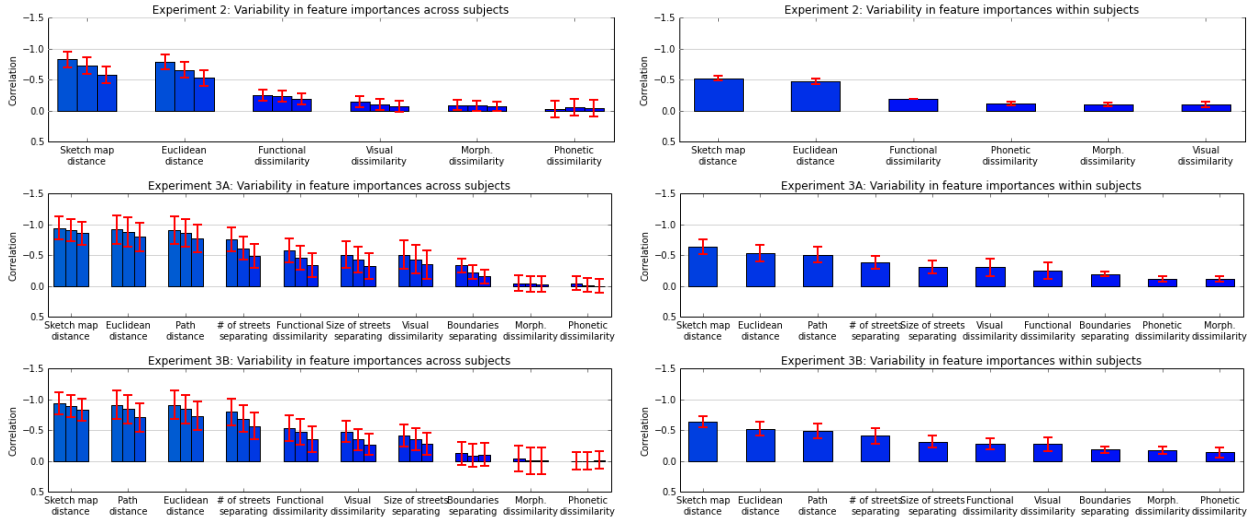


Figure 7: Variability of features influencing cognitive map structure. Feature variabilities across all subjects (left) and across map structures of individual subjects (right) are shown, plotted as error bars on each average feature correlation. Top: Bottom: Feature variabilities in the test trials of Experiment 2. Middle: Feature variabilities in Experiment 3A. Bottom: Feature variabilities in Experiment 3B.

and within subjects, i.e. across the maps of individual subjects (right panels), averaged over all subjects. Specifically, the standard deviations of the point biserial correlation coefficients [14] between the feature distances and co-representation probabilities are reported for all error bars in the plot. For the within-subject plots (right panels), the magnitude of the bars is also calculated using point biserial correlation, for the same reason - there being too few within-subject building pairs for the moving average-based probability calculation.

Finally, according to Shapiro-Wilk normality tests, none of the distributions of feature correlations are normally distributed (all p values for all features are many orders of magnitude less than 0.01). Rather than most subject structures arising from these features weighted in the same fashion, or from feature correlations concentrated around a most common value, the particular strengths of the influences on participants' representation structures seem to be irregularly distributed. However, they are significantly more consistent across individual participants' map structures (Figure 7 right) than across all participants (Figure 7 left).

3.3.3. Discussion

The strong dependence of co-representation probability on distance along various features (Figure 6) provides strong evidence for the plausibility of the clustering hypothesis. Furthermore, confirming intuitive expectation, spatial features show a much stronger influence on map structure than other, non-spatial dimensions.

Figure 7 shows that there is a large amount of variability in the importance of different features to various subjects. This spread is significantly less across the map structures of individual participants (Figure 7 right) than across all participants. Thus, although collecting a high enough number of map structures to reliably infer subject-specific feature importances presents several practical challenges (see next section), doing so is unavoidable for predicting spatial representation structures.

It is important to point out that correlation with co-representation probability alone is not a sufficient metric for describing the influence exerted by a feature on cognitive maps. There might be indirect causation

[14] Biserial correlation with the binary vector indicating same or different sub-map pairs was used, instead of calculating probabilities and using continuous correlation, because the numbers of available within and across sub-map pairs of buildings for a specific map of a specific participant were frequently below the window sizes used for estimating co-representation probabilities in Figure 6.

or a common cause, or deceptively low correlations due to sparse data (for example, very few natural boundaries are present in most cities, which causes low correlations despite their importance according to the results below), or other reasons for correlation not translating to causation.

For this reason, to what extent different features facilitate the prediction of individual map structures is a more meaningful measure of their importance in the cognitive map structuring process. The following sections report prediction results, both in automatically generated virtual reality environments (Section 2) and in real-world environments freely chosen by subjects (Experiment 3).

3.4. Experiment 2 - Predictability of map structure in virtual reality environments

This experiment investigated the question whether the clustering hypothesis allows robust advance prediction of participant map structures. Because of the observation that feature importances vary greatly across subjects, but less for individuals (Figure 7), it was designed to first learn these per-subject importances, before producing predictions using a clustering mechanism. This process was inspired by *active learning* (Settles, 2010), a field in machine learning which allows algorithms to choose the data from which they learn, thus facilitating better performance with less training data. This latter point is crucial for our experimental paradigm - as inferring the representation structure of even small environments with few buildings requires several full recall sequences, there is a practical limit on how many structures per participant can be produced - thus, this limited budget of data should be used in a fashion close to the statistical optimum. Optimally reducing model uncertainty using active learning is one possible approach towards this objective.

3.4.1. Computational methods

As described in the Introduction, a computational model of cognitive map structure requires learning subject-specific models reflecting feature importances, as well as a clustering algorithm. For this experiment, which allows full control over the memorized environments, we have used the decision hyperplane method to infer learn feature importances. We constructed a training environment for each trial such that 1) they contained two clusters (shop buildings and house buildings), 2) only the features of a single building, which lay somewhere between the two clusters, were varied (see Figure 8). We trained a linear classifier to assign the middle buildings of all trials of a participant to one or the other cluster in feature space. The class label (dependent variable) y was derived from that participant's recall sequences in each trial ($y = 1$ if the middle building was co-represented - i.e. recalled together - with the shop buildings, and 0 if it was co-represented with the house buildings). The distances of the middle building from the shop buildings along all features (in unweighted feature space) served as predictor (independent) variables x.

Based on these variables, a linear 'decision hyperplane' was calculated, which separated the set of all data points characterizing the middle buildings of a participant's trials into two sets: into middle buildings which were represented together with shops (if below the decision hyperplane) and into those which were co-represented with houses (if above the decision hyperplane) - see Figure 8. The slope of this 'decision hyperplane' in each feature dimension (distance, visual similarity / colour similarity, functional similarity) thus indicated the importance of each feature to this participant (for example, if the decision hyperplane in Figure 8 was horizontal, that would mean that the y-axis - spatial distance - would be the only feature of relevance for this subject. Conversely, if the plane was almost vertical, spatial distance would be unimportant).

The decision hyperplane was calculated using logistic regression (Hosmer & Lemeshow, 2004), formulating the question whether to group the middle building on the shop sub-map or the house sub-map as a binary classification problem. Thus, the probability $P(S|D)$ of clustering the middle building to the shop sub-map, given a set of distances $D = (d_s, d_f, d_p, ..., d_n)$ from the shop buildings along a number of features, including spatial (d_s), functional (d_f) and perceptual (d_p) distance (difference in colour), was modelled using the logistic regression equation

$$P(S|D) = \frac{1}{1 + e^{-W^T D}}, \tag{1}$$

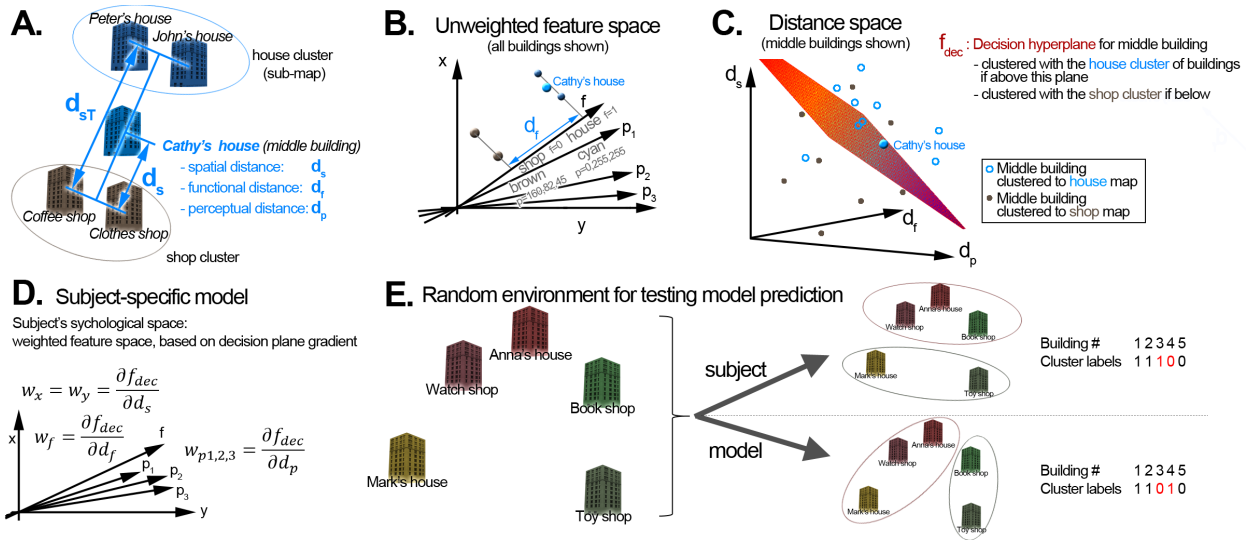


Figure 8: The decision hyperplane method for inferring feature importances and generating environments in Experiment 2. A: General layout of training trials, which consisted of two groups of two buildings (with equal colour and function), and a middle building, the parameters of which could be varied (distance, similarity in colour and in function to the shop group). B: Feature space representation - each building can be represented as a single point in a space spanned by the features (position, colour, function). C: Distance space representation - middle buildings can be represented in terms of their distance to the shop group along each feature. According to the clustering hypothesis, there has to be a 'decision hyperplane' calculable from these middle buildings, such that those below the plane (i.e. those closer to the shop group) are most likely clustered with the shop group, and those above the plane (i.e. farther away from the shop group) are most likely clustered with the house group. D. Subject-specific models consist of a weighted feature space and a clustering algorithm. The weights of the feature space can be calculated from the decision boundary - the importance of each feature is proportional to the derivative (slope) of the decision boundary by that feature. E. Randomly generated testing environments, and comparison procedure. Subjects impose a grouping even on random, unstructured environments, as shown by previous research (McNamara et al., 1989). The clustering model also produces a grouping, based on the learned subject-specific model and the clustering algorithm. Subsequently, cluster labels are compared (and, in this example, found to be incorrect).

in which the model parameters $W = (w_s, w_f, w_p, ...)$ control the slope of the decision hyperplane, and thus represent participants' feature importances in this model. They were used to construct the participant's 'psychological space', i.e. a feature space weighted by these parameters (as illustrated in Figure 1B), which lead to attenuated differences along features unimportant to the participant.

Subsequently, we used clustering in this weighted feature space for prediction. We employed the DP-GMM (Dirichlet Process Gaussian Mixture Model), from the family of Bayesian nonparametric models, for clustering (see Supplementary Information for the mathematical formulation and (Gershman & Blei, 2012) for a tutorial review). Bayesian nonparametric models were successfully employed in categorization models (Sanborn et al., 2006) and shown to be psychologically plausible, unifying previously proposed models of category learning (Griffiths et al., 2007) and accounting for several cognitive mechanisms including category learning and causal learning (Tenenbaum et al., 2011), transfer learning (Canini et al., 2010), and human semi-supervised learning (Gibson et al., 2013). Given that such models give a good account of how humans acquire novel concepts (subsuming prototype, exemplar, and rational models of category learning, among others), and given that they can be seen as probabilistic clustering models, we hypothesized that they might also account for sub-map learning.

DP-GMMs are extensions of Gaussian Mixture Models (GMMs) for an unlimited number of clusters. GMMs are statistical models which aim to partition a set of data points in some space into a number of clusters C by fitting C Gaussian probability distributions to the data, i.e. adjusting the parameters of these C Gaussians such that the probability that the data was drawn from these distributions is maximized. DP-GMMs have the same aim, but also allow inferring the number of distributions (and thus the number of clusters C), not just their parameters. In this lies their key advantage compared to most other clustering

models: they can be used without prior knowledge of the correct number of clusters (and they can expand by adding new points either to the most likely existing cluster, or to a novel cluster, when observing new data). This process of assigning new data points to clusters by calculating probabilities from distributions optimally fitted to previous data has a lot in common with the basic problem of categorization, which is to identify the category of a new object based on its observed properties and previously observed objects, which is why Bayesian nonparametric models are similar to (in fact, if parametrized accordingly, mathematically equivalent to) multiple psychological models of category learning proposed in the past (Griffiths et al., 2007).

The final sub-map membership predictions were generated by performing clustering, using a DP-GMM[15], in the weighted feature space learned from the subject. These predictions were evaluated by calculating prediction accuracies and Rand indices (Rand, 1971). The former is simply the ratio of perfectly predicted sub-map structures to all subject structures - however, this strict accuracy metric penalizes 'near misses' equally to completely wrong structure predictions (e.g. if seven building sub-map memberships are correct, but a single one incorrect, the entire prediction is counted as incorrect; just like completely wrong structures). Average Rand indices are reported as more fair metrics which provide a continuum between flawlessly correct ($R = 1$) and completely incorrect ($R = 0$) predictions. The Rand index is a measure of the amount of correctly assigned pairs among all pairs, and is defined as $R = (s + d)/\binom{B}{2}$, where B is the total number of buildings on a map structure, s is the number of building pairs on the same sub-map both in the predicted and actual map structure, and d the pairs on different sub-maps both in prediction and in subject data.

3.4.2. Participants

Participants were students at the University of Manchester (compensated by vouchers). Subjects who did not produce sketch maps significantly better than random chance in at least 50% of all training trials were excluded, leaving 12 subjects whose data was analysed. Participants were told they need prior experience with either virtual realities or three-dimensional computer games. These participants were recruited and tested at the University of Manchester (instead of online) primarily because the setup required a modern PC equipped with a graphics card to run the experiment smoothly. Further reasons were the requirement of prior 3D gaming experience (difficult to verify online), and the need to ensure that the setup was equivalent across subjects (e.g. screens were of the same size and quality, all subjects used a mouse and not a touchpad, etc.).

3.4.3. Procedure

After giving their consent and reading instructions, participants completed 20 trials - 15 'training' trials which were used for training the model, and 5 'testing' trials which were used for verifying the predictions of the computational model. In total, the experiment took about 1.5 hours on average. Each trial was set in a unique environment consisting of a horizontal ground plane, featureless sky, and 5 buildings. All buildings used the same 3D model and thus had equal measurements, but could vary in colour, in function (being labelled as either shops or houses) and in distance; and could have different labels (e.g. coffee shop, John's house).

Both trial types followed the same sequence. First, participants could freely explore the environment, and were asked to memorize the positions and names of all buildings in it. In this memorization phase, they were also asked to deliver a package from one of the shops to one of the houses. This task served the dual purpose of forcing subjects to do a minimum amount of exploration, and, additionally, to make the functional distinction between shops and houses more meaningful. After the memorization phase and the delivery task, the environment vanished, and participants' spatial memory was tested, by asking them for 1) a sketch map, produced by dragging and dropping labelled squares into their correct places, and 2) seven recall sequences, 5 cued, and 2 uncued.

The first 15 'training trials' each contained two distant groups of two buildings in close proximity, and a 'middle' building somewhere between these two groups. Both buildings in each group always had the same

[15] With variational inference to infer the most likely cluster memberships and parameters. We have used the *bnpy* Python library for inference (Hughes & Sudderth, 2013)

colour and function, and there was always one group containing two shops and a second group containing two houses. The middle building was intended to be represented together with one or the other group by subjects, depending on its distance and similarity to the groups. In the first 7 of these trials, the colours, functions, and distances of the groups and the middle building were generated randomly, ensuring only within-group consistency of colour and function and that buildings within groups were closer than the distance between the groups, such that they unambiguously formed clusters.

After the 7th[16] training trial and all subsequent training trials, a 'decision hyperplane' was calculated using logistic regression, which separated all middle buildings into two groups, those belonging to the shop cluster, and those belonging to the house cluster. This decision hyperplane facilitated the generation of the remaining 8 training trial environments. For each trial after the 7th, the two groups were again generated randomly, but the middle building was parametrized such that the uncertainty regarding subjects' feature importances was minimized. To achieve this, the parameters of the new middle building were drawn from the region of the currently calculated decision hyperplane, since this is the region in which the model is least certain as to where buildings should be assigned[17]. Formally, this is equivalent to active learning (Settles, 2010) with uncertainty sampling (Lewis & Gale, 1994) in machine learning. Each of these remaining 8 training trials maximally reduced the model uncertainty regarding feature importances.

Finally, participants completed 5 'testing' trials, going through the same procedure of memorization, delivery task, and producing a sketch map and recall sequences. These testing trials were generated completely randomly, without any restrictions on building parameters, not even the restriction of there needing to be clusters defined in any way along any of the features. They were used to test the predictions of the computational model.

3.4.4. Results

We included all features described in Section 3.3 in the following analysis, except for the four geospatial features not relevant in our simple virtual reality environment (path distance, natural boundaries, number of intersecting streets, whether they could be easily crossed). For the correlations of these features with co-representation probabilities (Figure 6), as well as across- and within-subject variances of these correlations (Figure 7), see Section 3.3.

Above, we have introduced a method to infer participants' feature importances for clustering, based on the inference of a decision hyperplane describing at which point in feature space subjects stop assigning a middle building to one sub-map and start assigning it to another. With this method, we have both components of a predictive model of cognitive map structure: 1) subjects' psychological spaces, spanned by a set of features and feature importances, as inferred by the decision hyperplane approach, and 2) a clustering algorithm. We chose DP-GMM as the clustering algorithm, given its substantial advantage of being able to infer the number of sub-maps automatically, and motivated by its success in other psychological models.

Figure 9 shows the results of this predictive model on all participant cognitive map structures (20 per subject; 15 training maps used to infer feature importances, and 5 testing maps used to verify model predictions). Prediction can be incorrect on training trials, because feature importances are being inferred using the decision hyperplane approach without taking into account the clustering algorithm and its idiosyncrasies (see red cells of the first 3 rows). After inference of feature importances and running the clustering model within this feature space, 73.5% of the training map structures could be predicted.

The interesting part of Figure 9 is the bottom row of each sub-plot, which contains the advance predictions of the model on randomly generated environments it was not trained on and not confronted with prior to

[16] Due to the noise inherent in the inference of building map memberships, using active learning from the start yielded bad results, the model hypothesizing a highly sub-optimal decision hyperplane it could not recover from using the limited number of subsequent data points. A few randomly initialized trials were used at the start to avoid this and to allow the inference of an approximate decision hyperplane before starting the active learning process. Empirical experimentation using artificially generated maps, and an amount of outliers comparable to subjects, suggested 7 random and 8 uncertainty sampling trials, when given 15 datapoints (from the 15 trials).

[17] As the region of least certainty, or greatest uncertainty, comprises the points with a classification probability of 0.5 to either class, these points can be defined as: $D_{LC} = \underset{D}{\operatorname{argmin}} |0.5 - P(S|D)|$. From this and eq. (1), it follows that $W^T D_{LC} = 0$, i.e. that points of least certainty lie on the hyperplane described by W, confirming the informal argument in the text above.

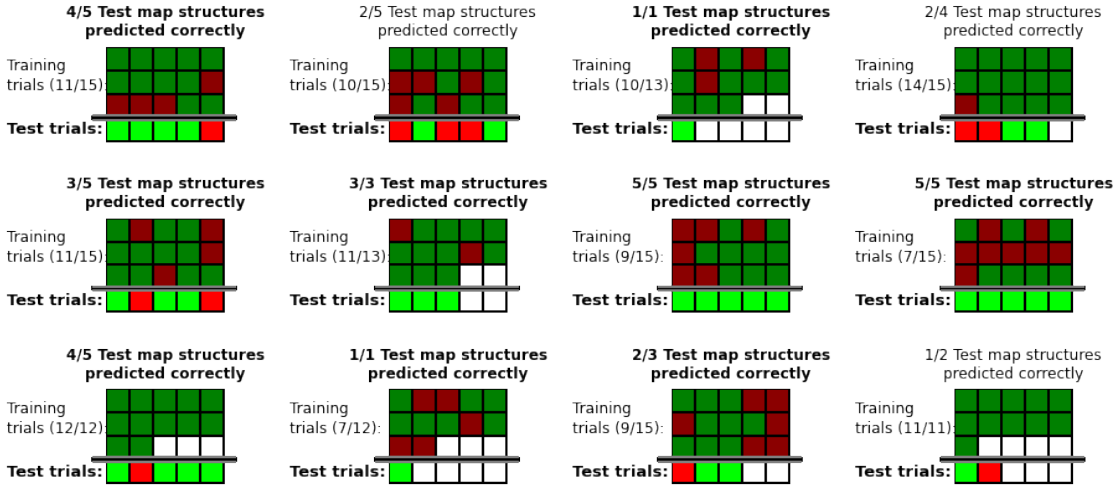


Figure 9: Results of a predictive clustering model using subjects' feature importances, learned using the decision hyperplane approach. Each sub-plot reports all prediction results for one subject, using green cells for correct predictions, red cells for incorrect predictions (one or more buildings grouped to the wrong sub-map), and white cells for subject maps either not better than random chance or without apparent structure. Top 3 rows in each subplot show results on the training trials (dark colours), and the 4th, bottom row shows the prediction accuracies on the test trials (bright colours). On average, 75% of all test map structures could be predicted correctly (green cells). For comparison, the probability of prediction by random chance is 0.4% for two sub-map and 3.1% for one sub-map structures.

making the prediction. On average, **75.0% of all test map structures could be predicted correctly in advance** using the decision hyperplane method and DP-GMM for clustering; and the majority of map structures could be predicted for all subjects except for one.

Note that this is a strict accuracy metric - if the model predicts four out of five building sub-map memberships correctly, but a single one incorrectly, the entire prediction is counted as incorrect. The Rand index (Rand, 1971) is a more comprehensive metric, providing a number between 1 (flawless clustering) and 0 (all cluster memberships incorrect). The **average Rand index of predicted vs. actual test map structures was 0.83** in this experiment, meaning that for 83% of the pairs of buildings, it could be correctly predicted whether or not they belong to the same sub-map in participants' spatial memory (according to their recall sequences).

If using the same DP-GMM model with feature importances inferred from co-representation correlations instead, the prediction accuracy drops to 59.1% on the testing maps, with an average test-map Rand index of 0.75, indicating that the decision hyperplane approach is better suited to uncovering feature importances than just using correlations.

Since each environment contained five buildings, there could be up to two sub-maps, and the clustering process could be framed as assigning one of three values to each building - member of sub-map #1, or of sub-map #2, or a single-building cluster (sub-maps with only a single building were excluded from participant data, for reasons explained in Section 2; however, if the model produced single-building clusters, these were not excluded from the model predictions, but instead counted as mistakes). Thus, the baseline probability of randomly coming up with the correct clustering is, on average, $(1/3)^5 = 0.4\%$ for map structures with two sub-maps, and $(1/2)^5 = 3.1\%$ for structures with one sub-map of unknown size. In this experiment, 14 subject test map structures contained two sub-maps, and 30 structures one sub-map.

3.4.5. Discussion

The observation that a large majority of subject map structures can be predicted in advance using a clustering model, together with an appropriately scaled feature space, provides further support for the clustering hypothesis. The improvement of prediction accuracy from 59.1% to 75.0% (and Rand index

from 0.75 to 0.83) when using the decision hyperplane approach to infer feature importances, instead of just using co-representation correlations, suggests that this approach is more suitable to uncover the psychological spaces in which the clustering takes place.

However, the present approach has several shortcomings. First, it is only applicable to controlled environments - thus, investigating participants' past long-term memory structures requires different methods (see next section). Second, the fact that calculating feature weights from a decision hyperplane does not take into account the actual model generating the predictions (in this case, the DP-GMM). Finally, the approach assumes linearity, i.e. that the surface separating buildings co-represented with one or the other sub-map is a linear hyperplane (as opposed to a non-linear surface). These shortcomings are reflected in the sub-optimal performance of the model on the training trials in Figure 9. Although a model should be able to fit its training data well, the performance on training trials (73.5) and testing trials (75%) is not statistically significantly different.

Thus, it is likely that more powerful models to learn subjects' feature spaces are needed. The next section introduces two such approaches addressing these shortcomings, one learning the optimal feature weights for the employed clustering model using global optimization, and the second lifting the linearity assumption. Both of them have the additional advantage that they do not require controlled environments.

3.5. Experiment 3 - Predictability of cognitive map structure in the real world

In this experiment, real-world buildings well known to participants were used (similarly to Exp. 1). Apart from providing additional evidence for the clustering hypothesis by showing that cognitive map structures in real-world environments can be predicted using a clustering model, this section also introduces and validates two generally applicable ways of learning subject-specific models.

3.5.1. Computational methods

Unlike in the previous section, where participants' feature importances were inferred using the decision hyperplane method - which requires controlled environments (Section 3.4), we use two generally applicable methods in this section (see Figure 10):

1. Global optimization (Jones et al., 1993)[18] - among all possible feature weights (between 0 and 1), select the features and weights best explaining the groupings of the 'training' structures obtained from the participants (a part of each participant's data was used for training, and the rest for 'testing', i.e. prediction verification). Use clustering in this weighted feature space for prediction.
2. GDA (Gaussian Discriminant Analysis) (Bensmail & Celeux, 1996) - using the set of all training building pairs, learn a probabilistic (Gaussian-based) model capable of calculating the probability of whether any given pair of buildings are co-represented on the same sub-map, given the distances of this pair along various features. Use this probabilistic model as a distance metric[19] (such that building pairs which are likely to be on the same representation are close, and those which are not are distant, under this metric). Predict subject map structures by clustering under this learned, subject-specific metric. See Supplementary Information for the mathematical formulation.

The first of these two, as well as the hyperplane approach, are both linear methods, whereas the latter method (GDA) allows non-linear solutions. Linear methods project data into psychological space by linearly

[18] We used the locally biased variant (Gablonsky & Kelley, 2001) of DIRECT (DIviding RECTangles) (Jones et al., 1993), a global, deterministic, derivative-free optimization method based on Lipschitzian optimization, which can handle the kinds of non-linear and non-convex functions which clustering accuracy inevitably entails. DIRECT finds global optima by systematically dividing the feature space into smaller and smaller hyperrectangles, returning the one yielding the best results upon convergence.

[19] For a pair of buildings represented by feature vectors $\mathbf{x}_1$ and $\mathbf{x}_2$, given their absolute difference $\Delta\mathbf{x} = |\mathbf{x}_1 - \mathbf{x}_2|$ as well as a trained GDA model which is able to calculate the probability $p(c = 1|\Delta\mathbf{x})$ that these buildings are co-represented on the same sub-map, based on appropriately fitted Gaussian distributions (Bensmail & Celeux, 1996), we simply define the metric as $d_{Metric}(\mathbf{x}_1, \mathbf{x}_2) = 1 - p(c = 1|\Delta\mathbf{x})$, where the probability of co-representation is derived using Bayes rule, $p(c = 1|\Delta\mathbf{x}) \propto p(\Delta\mathbf{x}|c = 1)p(c = 1)$, and the generative densities are modelled using multivariate Normal distributions $p(\Delta\mathbf{x}|c = 1; \mu_i, \Sigma_i) = (2\pi)^{-\frac{D}{2}} |\mathbf{\Sigma}_i|^{-\frac{1}{2}} e^{-\frac{1}{2}(\Delta\mathbf{x}-\boldsymbol{\mu}_i)^\intercal \mathbf{\Sigma}_i^{-1}(\Delta\mathbf{x}-\boldsymbol{\mu}_i)}$ (see Supplementary Information for details).

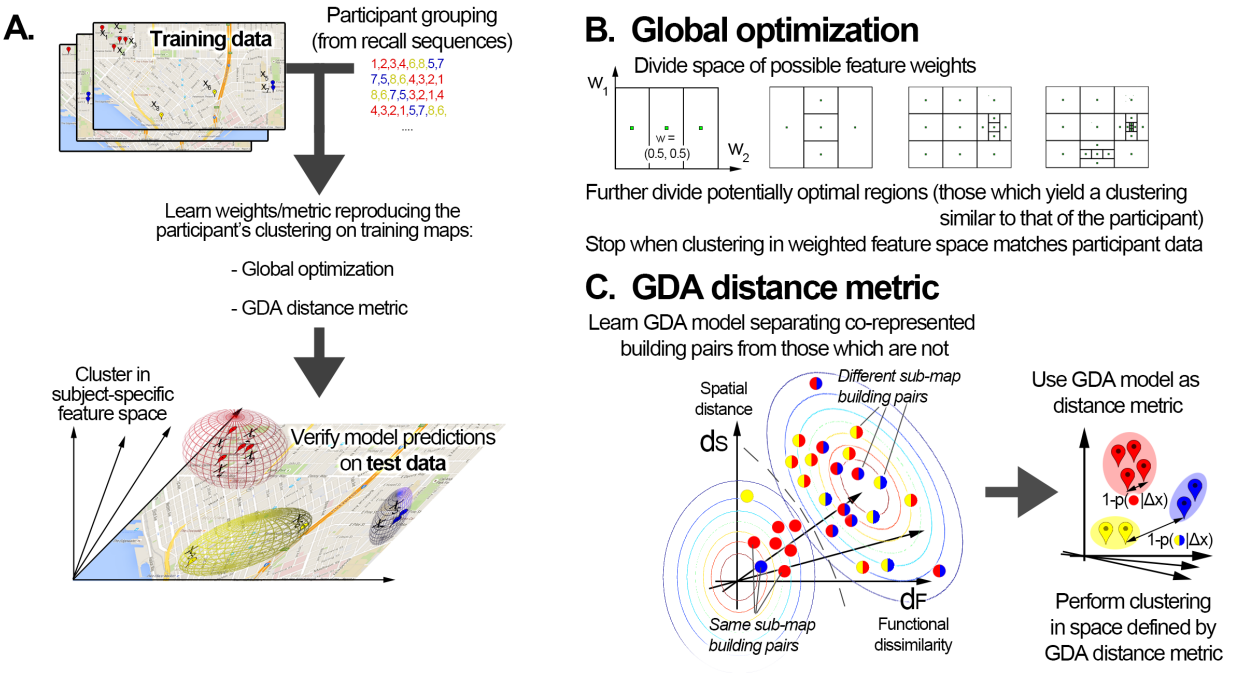


Figure 10: Learning subject-specific models for predicting cognitive map structure. A: General modelling procedure. Several sets of buildings and their groupings are obtained from different environments using the recall sequence paradigm, and these are split into two parts, training data and test data. A subject-specific model is learnt such that clustering under this model (e.g. in an appropriately weighted feature space) reproduces the training data as well as possible. Finally, model predictions are generated by clustering test buildings not seen at training time, under the learned models, and these predictions are compared to the actual participant map structures (groupings) in the test data. B: A weighted feature space (modelling participants' 'psychological space') can be obtained by searching for the optimal weights using global optimization. This method keeps dividing the space of possible feature weights into thirds, and further divides potentially optimal regions (those which correspond to feature weights under which clustering yields a grouping close to participant's map structure), until the weights best matching participant training data are found. C: A metric space modelling participants' psychological space can also be defined by a non-linear metric instead of linear feature weights. Such a metric can be learned by fitting a GDA model to separate building pairs that belong to the same sub-map from those that do not. In order to make predictions, building representations are projected into a space where their distances are dictated by this GDA model (such that they are close if they are likely to belong to the same sub-map); and clustering is performed in this space.

weighting the features, and try to find weights such that buildings co-represented on the same sub-maps are closer to each other in this weighted feature space than other buildings (note that the problem of projecting the data into a subject-specific feature space with some learned weights is equivalent to finding a distance metric with those feature weights[20]). In addition to these linear methods, we wanted to test a more powerful method that can capture non-linearities as well as interactions between the features (e.g. situations where the importance of one feature depends on the magnitude of another). We implemented a novel method, instead of using existing metric learning approaches, see (Yang & Jin, 2006) for a review, for the following reasons. First, our method can naturally incorporate the hypothesis that same sub-map building pair differences should be small, and thus located close to the origin, and should be separable from different-map building pair differences (these two distributions of pair differences can be naturally modelled using Gaussian distributions) - see Figure 10C. Second, our data violates some of the assumptions of existing methods[21]. Third, most existing machine learning solutions - as well as MDS, used in cognitive

[20]This equivalence is easy to see from rewriting the equation of the Mahalanobis distance metric, $d_M(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{(\mathbf{x}_2 - \mathbf{x}_1)^\intercal W (\mathbf{x}_2 - \mathbf{x}_1)}$, to the following form: $d^2_M(\mathbf{x}_1, \mathbf{x}_2) = ||A(\mathbf{x}_2 - \mathbf{x}_1)||$, where $W = A^\intercal A$, and A is a projection map that can transform data x into weighted feature space $\mathbf{y} = A\mathbf{x}$

[21]Metric learning is concerned with finding a distance metric - such as linear, Mahalanobis distance metrics, and their associated parameters, e.g. (Xing et al., 2002), or non-linear metrics by projecting the data into kernel space using e.g.

psychology to model similarities as distances (Shepard, 1957) - need to embed both training map and test map buildings into the same space for model training and testing. This is not possible in our case, because 1) for the features of functional and perceptual similarity, the pairwise similarities across environments are unknown (since subjects only indicate these within each map, not across maps), and 2) spatial distances might not be comparable across cities or countries (whether two buildings belong to the same representation strongly depends on their geographical distance; but this dependence likely becomes weak or non-existent if they are very far apart).

After a subject-specific model has been learned, sub-map memberships can be predicted by performing clustering based on this model (i.e. within the feature space / under the metric learned from the subject). Just like in the previous section, we used the DP-GMM clustering algorithm for this purpose.

Before reporting prediction results, we should point out that there are theoretical as well as practical limits on the predictability of cognitive map structures. Section 4.3 discusses these in more detail and suggests some solutions. Here, we shall focus on the main issue concerning data analysis, namely detecting and removing outliers caused by distractions or lapses of attention. If a set of buildings that are actually co-represented on a sub-map in a subjects' spatial memory is recalled together most of the time, but the subject is distracted during one of the recall sequences, and recalls a different (not co-represented) building instead, the subsequently extracted structure will be incorrect (since tree analysis requires items to occur together in *every* recall sequence in order to identify a sub-map). Even a single distraction during the 7 or 10 (in Experiment 3 A or B) recall sequences per trial can yield substantially different structures (see example in Figure 13 in Section 4.3, in which a distraction causes a drop of 0.6 in the Rand index to the correct structure).

The jackknifing procedure we use to eliminate outliers was suggested by the authors pioneering the recall order paradigm (Hirtle & Jonides, 1985; McNamara et al., 1989) to mitigate this issue, but relies on statistical significance testing to identify those outliers, and thus frequently fails to do so due to the small number of recall sequences collected in our experiments (a necessary limit arising from the need to collect multiple different map structures for training and testing a predictive model - subjects already took up to 3.5 hours for these experiments even with this small number of sequences).

It is possible to estimate the effectiveness of jackknifing in our data - and the percentage of incorrectly inferred and thus unpredictable map structures resulting from it (see Figure 11). To do this, we simulated distractions by randomly swapping two items in one of the sequences in each trial. This is a reasonable model of distractions, since the only way subjects can make mistakes is by changing the order of their input (they are forced to repeat the trial if they omit or incorrectly recall an item).

The *number* of simulated distractions (frequency of swapped items) makes no difference to the estimated *percentage* of outliers that are not caught and excluded by jackknifing. We used one distraction per trial (however, the following results stayed the same with 0.5 or 2 distractions per trial). For the 5 buildings maps (and 7 recall sequences), and averaging over 100 runs, each with a single random non-cue lapse for all subjects, simulated distractions cause changes in map structure (relevant outliers) in $\mu_n = 65.4\%$, $\sigma_n = 3.7\%$, and within these, outlier removal is effective in $\mu_e = 59.4\%$, $\sigma_e = 5.0\%$. The situation is somewhat better on the 8 building maps, due to the larger numbers of sequences collected and thus higher statistical power - here, outlier removal is effective in $\mu_e = 56.0\%$, $\sigma_e = 8.1\%$ of the cases (and necessary only in $\mu_n = 33.2\%$, $\sigma_n = 6.0\%$). This leaves on average $\mu_u = 26.6\%$ ($\sigma_u = \sqrt{\sigma_e^2 * \sigma_n^2 + \mu_e^2 * \sigma_n^2 + \mu_n^2 * \sigma_e^2} = 3.9\%$) of disruptive simulated lapses of attention for condition A, and $\mu_u = 15.0\%$ ($\sigma_u = 4.3\%$) for condition B, which cannot be mitigated by jackknifing.

If we assume this uniform random swapping to be a reasonable approximation of subject distractions, this would mean that apart from the approximately $o = 9.5\%$ of sequences which were successfully removed

a Radial Basis Function (RBF) kernel Φ in a distance function $d_{RBF}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{(\Phi(\mathbf{x}_1) - \Phi(\mathbf{x}_2))^\intercal (\Phi(\mathbf{x}_1) - \Phi(\mathbf{x}_2))}$, e.g. Baghshah & Shouraki (2010); Chitta et al. (2011). However, the former makes the assumption of *linear separability*, and the latter require *variances to be isotropic*, i.e. to not differ much across features (since the RBF kernel uses a diagonal covariance matrix, it cannot fit non-isotropic data well - see Ong et al. (2005)). As both of these assumptions are occasionally violated in our subject data, these metric learning approaches are not applicable. In contrast to existing metric learning, our proposed approach learns a probabilistic model in the space of pairwise differences (instead of learning from scalar distance values, it learns from difference vectors), and thus can fit non-isotropic and non-linear data.

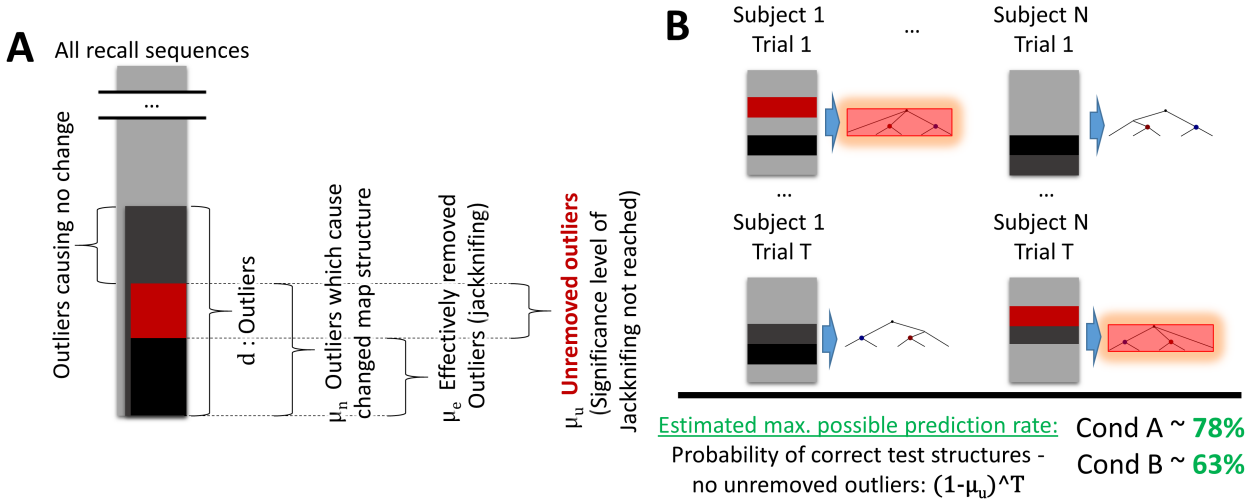


Figure 11: Estimated maximum possible prediction rate using the data in Experiment 3. A: Assuming that distractions / lapses of attention manifest as randomly swapped items in recall sequences (and cause changes in the inferred tree structures), a substantial number of them cannot be detected using the outlier detection procedure (jackknifing) proposed in the seminal work on hierarchical cognitive maps and employed in this paper. B: Undetected outliers in recall sequences cause a number of inferred map structures to be incorrect. This results in a percentage of map structures not predictable even by good models.

as outliers using jackknifing (and thus part of the effective 59.4% or 56% for cond. A and B), there would be an expected additional $\mu_o = 6.5\%$ ($\sigma_o = 0.1\%$) of sequences for condition A, and expected $\mu_o = 2.5\%$ ($\sigma_o = 0.7\%$) for condition B, which would likely be outliers causing structure changes which have not been removed by jackknifing because of the lack of statistical significance. It follows that the expected probability of extracting correct map structures under these assumptions - and thus the maximum possible prediction rate - is around $(1 - 0.065)^7 \simeq 63\%$ for condition A (since there are 7 sequences per trial), and around $(1 - 0.025)^{10} \simeq 78\%$ for condition B (since there are 10 sequences per trial).

To summarize, the observation that not all simulated distractions (outliers) can be identified and omitted by the jackknifing procedure strongly suggests that the data collected from human subjects also contains outliers not caught by jackknifing. Thus, these outliers prevent perfect prediction of subject map structures. Figure 11 summarizes this reasoning and the maximum possible prediction rates estimated based on it for both conditions.

3.5.2. Participants

Data from 71 participants was analysed in this section, 54 in Experiment 3A (asked for 5 environments with 5 buildings each), and 19 in Exp 3B (asked for 3 environments with 8 buildings). Subjects unable to produce at least two sketch maps significantly better than random chance (see Section 2.3), with structure apparent from their recall sequences for at least two maps, were excluded, as at least two map structures were required to have both a training and testing map. Participants were recruited, consented, and compensated through the Amazon Mechanical Turk online survey system, and were required to have at least 95% approval rating on previous jobs to ensure higher data quality.

3.5.3. Procedure

The procedure was similar to the one used in Experiment 1. This experiment was also conducted on a website participants could access through MTurk after giving their consent. Unlike 1, this experiment consisted of multiple trials (5 in condition A, 3 in condition B), each trial following an equivalent procedure but asking for a completely different set of buildings, possibly in a different city. Subjects took between one and 3.5 hours to complete this repeated trial experiment (this includes possible breaks, since the experiment was performed online in participants' homes, unsupervised, and the experiment was not timed).

In the first questions of each trial, subjects were asked to pick a number of buildings they know well - 5 in condition A, and 8 in condition B (thus, in total, 25 buildings had to be recalled for the 5 trials of condition A, and 24 for the 3 trials of condition B). Thus, well-memorized long-term memories of real-world environments were tested instead of novel stimuli in virtual reality. Subjects were instructed to make sure that they know where in the city these buildings are located, how to walk from any one building to any of the others, what each building looks like, and what purpose it serves.

The subsequent questions of each trial required subjects to produce a sketch map, and to perform a recall test consisting of 7 recall sequences in condition A, and 10 in condition B (in both cases, as many cued sequences as there were buildings on the maps, and two additional uncued sequences). Subjects followed the same instructions as in Experiment 1 ; the crucial difference being that instead of presenting cues verbally by writing out the name of the cue building, cues were presented visually (cue modality was changed to mitigate the strong effects of phonetic and morphological similarity in the prior experiments, presumably due to articulatory rehearsal strategies). Participants were shown building positions on their own sketch maps prior to each recall sequence question, excluding the labels - only the uniform gray squares symbolizing the buildings were shown. For each cued recall question, the cue (starting building) was indicated by highlighting the cue building in green colour and with a thick border.

In the final question, subjects were asked to judge the similarities of all pairs of buildings, i.e. $\binom{5}{2} = 10$ pairs in condition A and $\binom{8}{2} = 28$ pairs in condition B, as well as a control pair of one of the buildings to itself, both in terms of visual similarity, and similarity of purpose/function (using 1-10 rating scales as before).

3.5.4. Results

Figure 12 shows prediction accuracies (the ratio of perfectly predicted map structures to all subject map structures) using DP-GMM clustering and GDA subject-specific model learning. Using the best possible set of features shown to the model[22], **68.6% of the 185 subject map structures with 5 buildings of Experiment 3A** (with up to two sub-maps per structure), and **79.2% of the 48 subject map structures with 8 buildings of Experiment 3B** (with up to four sub-maps per structure) **can be predicted accurately**, such that every single predicted sub-map membership is correct for these percentages of test maps. **Average Rand indices for these models are 0.87 for condition A and 0.95 for condition B**, which means that even the structures which are imperfectly predicted, causing a lower than optimal prediction accuracy, are highly similar to the correct structures (co-represented building pairs are predicted correctly in 87% in condition A and 95% in B). Note that the prediction accuracy of the best model is statistically indistinguishable from the estimated maximum possible prediction rate (calculated above based on simulating distractions by random swapping). This suggests that the proposed novel GDA-based method does well at learning subject feature spaces, and that the subsequent clustering model, based on a previously proposed Bayesian model of category learning, can infer the sub-map memberships and numbers accurately.

Figure 12 also shows the numbers of sub-maps contained in participants' structures. In general, the prediction task can be seen as assigning one of $K + 1$ values to each building, where K is the maximal number of possible sub-maps (single-building clusters are also possible, hence the increment by one). Thus, the baseline probability of randomly coming up with the correct clustering is, for condition A, $(1/3)^5 = 0.4\%$ for map structures with two sub-maps, and $(1/2)^5 = 3.1\%$ for structures with one sub-map. For condition B, this baseline expected random clustering accuracies are several orders of magnitude lower ($2.5 * 10^{-4}\%, 1.5 * 10^{-3}\%, 1.5 * 10^{-2}\%$ and 0.3% respectively for $K = 4, 3, 2$ and 1).

The model accuracies when successively removing particular features (bars from left to right in Figure 12) provide an additional measure for how important these features were, aggregated over all subjects, and measuring importance in a causal fashion, since this is a predictive model. The most important features were those which caused the greatest drops in accuracy upon their removal. In condition A, two features

[22]Estimated from the training data, using a greedy search approach - starting with a single feature (Euclidean distance) and then iteratively adding the feature which brings the clustering prediction closest to participants' actual groupings; repeated until either all features are included or the clustering prediction accuracy stops increasing.

Condition	All features, subject-specific GDA model	A-priori features subject-specific GDA model	No subject-specific model
Condition 3A	**79.2 % (RI=0.94)**	70.8% (RI=0.88)	41.7% (RI=0.76)
Condition 3B	**68.6% (RI=0.89)**	63.4% (RI=0.83)	60.2% (RI=0.78)

Table 2: Prediction accuracies (and Rand indices) in Experiment 3, for all features and subject-specific GDA+DP-GMM model (second column), for features known a-priori, without having to ask subjects to rate similarities or draw sketch maps (third column), and finally using a subject-general model, without learning subject-specific feature weights. Rows: Condition 3A (19 subjects, 48 map structures from as many distinct environments, 112 sub-maps), and condition B (54 subjects, 185 map structures from as many distinct environments, 310 sub-maps).

are significantly more important than the rest - sketch map distance and the product of path distance and visual similarity -, whereas the importances are similar in condition B, with a slightly larger accuracy drop caused if omitting sketch map distance. In both conditions, about 2 out of 5 map structures can still be predicted when using solely Euclidean distance.

The strong influence of sketch map distances raises an additional question regarding predictability of cognitive map structures - is advance prediction possible without asking the subject anything (other than a list of buildings he knows)? To investigate this question, we have run the predictive model on data from which visual similarities and sketch map distances were removed, i.e. solely on data which can be derived from the list of subjects' buildings (see Section 3.3 for geospatial data sources). Subjects' functional similarities were also removed from this data, and replaced by an objectively calculated measure of functional relatedness. Specifically, we used the Jaccard similarity metric on lists of building types from Google Places API [23]. The objective functional similarity metric thus obtained does reflect subjects' own judgements - the correlation between them is $r = 0.66$ - but is somewhat different, since it does not reflect subject idiosyncrasies, and is also free of noise or biases.

Using GDA for subject-specific model inference, and using these features which are all known a priori - derivable from the subject building lists and public geospatial databases -, 75% of map structures can be predicted in advance for condition A (Rand index: 0.91), and 68.8% in condition B (Rand index: 0.88).

Finally, we have attempted to predict subjects' cognitive map structures without learning subject-specific models at all, by trying to infer a psychological space common to all subjects, and clustering within this space. Inferring someone's spatial representation structure without knowing anything about them would have great advantages for robotics applications and geographical planning and map design, among other fields (see Section 4.1). The resulting prediction accuracies (and Rand indices) for condition A and B were 41.7% and 60.2% (and $RI = 0.76$ and 0.78) respectively. In accordance with the results in Section 3.3, the model performs significantly worse when not allowed to learn subject-specific feature spaces. However, even these impoverished models can predict whether or not two buildings are co-represented on the same sub-map in more than 3 out of 4 cases.

3.5.5. Discussion

The model prediction accuracies reported above are close to the estimated maximum possible prediction rates from noisy map structures (based on simulating participant distractions using random swapping), calculated at the beginning of this Section: 62.5% for condition A, and 78.0% for condition B. This shows that the model accounts well for this noisy data, despite not being able to predict 100% of subject map structures.

[23] Places API can return a list of known building types when queried - see `https://developers.google.com/places/supported_types` for a list. Usually buildings have several applicable types, ranging from specific to general, e.g. 'meal takeaway', 'restaurant' and 'food' for McDonalds. The Jaccard index (JI), defined as the ratio of the size of the intersection to the size of the union of two sets, measures how many items in these type lists match between two buildings, as a proxy for their functional similarity. For example, $JI = 0.5$ between the McDonalds example and a building with types 'bakery', 'restaurant' and 'food'. The type 'establishment' was present for almost all buildings and was thus excluded from the computation of JIs, being uninformative.

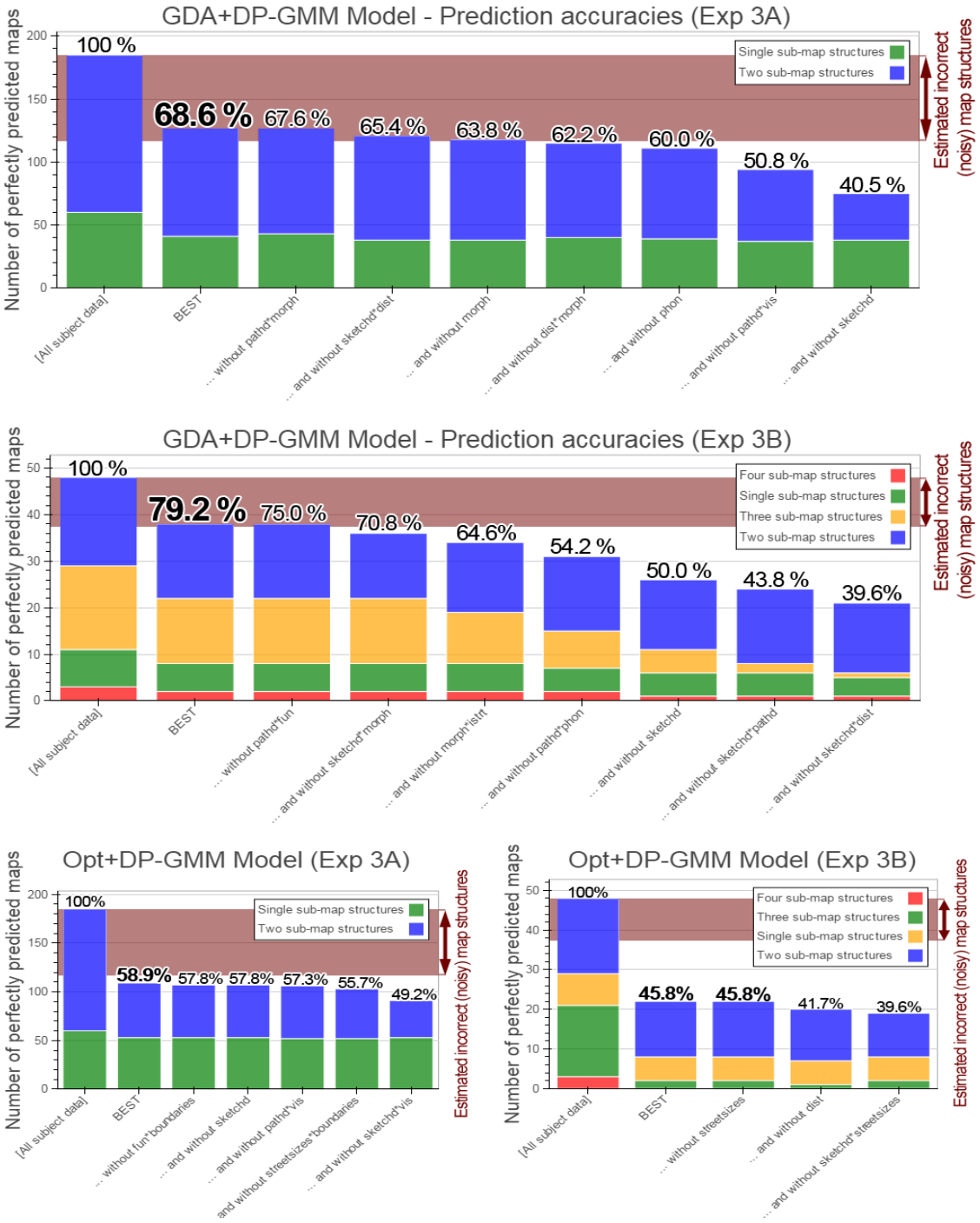


Figure 12: Accuracies obtained by predicting participant's map structures using DP-GMM clustering under the learned subject-specific models. The first bar shows the number of all subject map structures (and, within them, the numbers of structures containing the specified numbers of sub-maps). Top: results in condition A. The first bar shows all subject maps, the second the prediction accuracy for the best feature set (Euclidean, path*morphological, sketch map, path*visual, phonetic, Euclidean*morphological, morphological, sketch map*Euclidean distances); bars 3-8. shows accuracies when successively removing the last feature. Middle: results in condition B. The second bar shows the best prediction accuracy (Euclidean, path*functional, sketch map*morphological, morphological*separating streets, path*phonetic, sketch map, sketch map*path and sketch map*Euclidean distances); bars 3-8. accuracies when successively removing the last feature. Bottom: Prediction accuracies of the optimization-based subject-specific model. Best model accuracies for condition A and B are shown in the second bar of the left and right bottom plot.

Even using solely features which can be objectively derived from geospatial (and linguistic) information from participants' specified buildings, without collecting any subjective data such as sketch maps or visual similarity judgements for the test maps, solid prediction is still possible - 70.8% for condition A and 68.8% in B (although recall sequences still need to be collected in order to learn subject-specific models). This makes a subject model, once learned, applicable to any environment encountered by that subject.

These results further substantiate the plausibility of the clustering hypothesis, and in particular, provide evidence that nonparametric Bayesian clustering is a suitable model not only for human category learning (Griffiths et al., 2007), but also for cognitive map structure learning; and fit in well with the growing body of evidence for 'Bayesian cognition' (Tenenbaum et al., 2011).

4. General Discussion

A growing body of evidence suggests that rather than storing spatial information within some global reference frame, human spatial memory employs local, object-centered representations (Marchette & Shelton, 2010; Chen & McNamara, 2011; Greenauer & Waller, 2010; Meilinger et al., 2014). This is consistent with the much earlier proposal that spatial memories are organized according to hierarchies (Hirtle & Jonides, 1985; McNamara, 1986; McNamara et al., 1989; Holding, 1994; Wiener & Mallot, 2003), as well as with recent neuronal evidence (Derdikman & Moser, 2010; Han & Becker, 2014).

In this paper, we made the first attempt to quantitatively explain and predict the local structure of spatial representations. We have found strong correlations between the probability that two buildings are co-represented[24] and features such as Euclidean distance, path distance, and visual and functional similarity. These correlations suggest that clustering based on proximity along these features is likely to give rise to the observed representation structure. We have developed multiple methods for exploring how important these features are for individual subjects (i.e. learning their 'psychological spaces'), even if only small amounts of data are available, and have developed and evaluated a predictive model of cognitive map structure based on Bayesian nonparametric clustering in these learned psychological spaces. We have shown that our model can successfully predict spatial representation structures in advance in the majority of cases.

The results from our model are very promising, but their plausibility depends on the empirical method used to expose spatial representation structure. Although the structures identified by our recall order paradigm are substantiated by their significant influence on several cognitive phenomena (Section 3.2), there is clearly room for improving the experimental methodology. After briefly outlining the implications of models of cognitive map structure, the discussion below outlines some alternative approaches, and suggests reasons for the imperfect prediction rates.

4.1. Implications of modelling cognitive map structure

We have reported significant effects exerted by cognitive map structure on spatial memory-related performance in Section 3.2. Together with prior evidence on priming, map distortion, distance estimation biases, and related effects, it seems clear that representation structure is relevant to spatial memory.

Apart from psychology, its investigation is also of interest for neuroscience. Strong evidence exists for hierarchies in the neural correlates of rodent spatial memory, place cells and grid cells, specialized neuron types discovered in mammalian - and, more recently, human - brains (Ekstrom et al., 2003; Jacobs et al., 2013), and is shown to play a key role in representing space (Moser et al., 2008). Place cells show increased activity in small, spatially localized areas, encoding spatial locations within particular spaces - with firing patterns changing significantly upon switching or changing immediate surroundings (the set of active place cells is completely different in separate environments). Grid cell firing shows a highly regular, triangular grid spanning the surface of an environment, independently of its configuration of landmarks, thus encoding a direction and distance metric.

Both of these spatially relevant neuron types have been observed to show natural hierarchies, with the granularity of representations (the sizes of the firing fields of individual cells) increasing from dorsal to ventral

[24]Stored on the same representation, as indicated by the recall order paradigm (i.e. always recalled together)

poles of the relevant brain areas (Brun et al., 2008; Kjelstrup et al., 2008). Furthermore, fragmentation in separate parts of an environment has also been observed in electrophysiological recordings of grid cells (Derdikman et al., 2009; Frank et al., 2000), indicating that instead of a single 'cognitive map', there a manifold of sub-maps are represented in brains (Derdikman & Moser, 2010).

However, the connection between these hierarchical and/or fragmented neural representations, and cognitive representations of map structure, remains largely unexplored. The predictive modelling approach presented in this paper could facilitate and accelerate research into this connection - after a subject-specific model has been learned from a small number of environments, subjects do not need to be subjected to arduous recall sequences (or large numbers of estimations), and can quickly be tested in large numbers of virtual reality environments in an fMRI.

Models of cognitive map structure could be of interest not only to the cognitive sciences but also to neighbouring fields. For example, in geographic information science, the insight that both planning times and estimation accuracies are improved within sub-maps compared to across, together with a subject-general model (which is good enough for this purpose - see Section 3.5), could help design schematics or transit maps which are cognitively easy to use for a majority of subjects.

Furthermore, models of human spatial representation are relevant for robotics for the purpose of communicating and interacting with humans. This is a rapidly growing area, with over three million[25] personal (non-industrial) service robots sold in 2012; a figure that can be expected to grow with the increasing demands on care robotics due to the rapid ageing of the world population. A model of spatial representation structure could allow artificial agents to use and understand human-like concepts (for example, translating latitudes and longitudes to easily understandable expressions like 'between the shopping area and the university buildings'). Approaches to conceptualize spatial representations exist only for indoor robots (Zender et al., 2008). The present approach, in contrast, is applicable to unconstrained outdoor environments (and is demonstrated by our results to work in a human-like fashion in over a hundred cities).

Finally, the particular way individual subjects structure their commonly encountered environments depending on past experience and task demands could give insight into computationally more efficient spatial representations for artificial intelligence (AI). With only around 40 million principal neurons in the human Hippocampus (Andersen et al., 2006), adults seem to be able to effortlessly store and recall navigation-relevant spatial details of many dozens of cities and hundreds of square kilometers. Storing a comparable amount on a trivial AI map representation such as an occupancy grid (Elfes, 1989), with the accuracy relevant for navigation, and including rich perceptual information, is not possible using today's hardware (let alone searching through such a vast database in split seconds, as humans are able to do). Human spatial representation structure could give inspiration for more efficient computational structures for representing space.

4.2. Alternative empirical approaches to uncovering cognitive map structure

Since humans do not have introspective insight into their own memory structure, uncovering organization principles of spatial memory is challenging. Several methods have been proposed in the literature to investigate which reference frames, or imposed structure, might be employed by participants. Of these, the recall order paradigm was used here, and described in Section 2. Its main shortcomings are the lack of robustness to outliers due to e.g. lapses of attention (mitigated by the jackknifing procedure), and the influences of phonological and morphological features of verbally cued items (mitigated by spatial cueing, as in Experiment 3). Despite these shortcomings, the structures extracted by this method have substantial influence on various cognitive phenomena, as reported in Section 3.2.

Other experimental approaches for investigating representation structure include judgements of relative direction (JRD), in which subjects imagine standing at some specified location and heading, are asked to point to specified objects they have memorized previously. The angular error in JRD seems to be strongly affected by interobject spatial relations (rather than only depending on a global reference frame), with better accuracy for judgements aligned with the intrinsic reference frame of an array of objects both in

[25] According to the World Robotics 2013 Service Robot Statistics, http://www.ifr.org/service-robots/statistics/

navigation space (McNamara et al., 2003; Meilinger et al., 2014) and in small-scale environments in a room (Mou & McNamara, 2002). These experiments have utilized object arrays with clear axes of alignment, either employing a grid-like array (mainly used in small-scale experiments) or making use of single major roads or paths as intrinsic axes in large-scale surroundings. This setup limits the applicability to general environments. However, the idea of direction judgement errors induced by changes of reference frame is generalizable, and has also been used to investigate reference frames of arrays without enforced intrinsic structure (Han & Becker, 2014; Chen & McNamara, 2011). Because direction errors are smaller within reference frames than across (Han & Becker, 2014), they could in principle be used to infer representation structure. The main disadvantage of this approach is the large number of direction estimations required to distinguish reference frames reliably, due to the large variance of direction errors. Furthermore, the number of estimations needed for pairwise comparison grows quadratically with the number of objects and / or frames (none of the cited papers compare more than two frames).

Cognitive map structure impinges on behavioural performance in several ways, most notably including biasing direction estimation (see above), distance estimation - overestimated across- and underestimated within representations (Hirtle & Jonides, 1985) -, and priming, i.e. accelerated recognition latencies (McNamara et al., 1989), direction estimation latencies (Han & Becker, 2014), and verifications of spatial relations (Hommel et al., 2000). All of these biases in errors or response times cause the same difficulties when trying to infer the exact representation structure for a particular participant - due to large variances, a very large number of judgements is required to obtain acceptable statistical significance (and the number grows quadratically with the number of objects). How to mitigate this problem, and which of these metrics have the smallest variance and thus highest reliability for map structure extraction, as well as whether they all yield consistent structures as would be expected, remain important questions for future research on cognitive map structuring.

Assuming either no distractions, or that jackknifing can successfully eliminate the majority of outliers caused by distractions, the recall order paradigm is able to provide the most deterministic way of inferring map structure, since it does not rely on comparing distributions of errors (or response times) using significance testing. It is also deterministic over time, resulting in very similar structures to the original hierarchies when re-testing subjects several weeks later (Hirtle & Jonides, 1985). These advantages, together with the difficulty of obtaining statistically significant results from error / RT patterns with high variances, have motivated our choice for the recall order paradigm for uncovering the structures modelled in this work.

4.3. Obstacles to predicting cognitive map structure

Our results indicate strong correlations of co-representation probability with distance (Section 3.3), suggesting that a clustering mechanism underlies map structures, and substantiating the plausibility of our computational model. However, these conclusions are based on a number of assumptions; and it is possible that some of them might not be correct. Below, we list some possible obstacles to a predictive model based on these assumptions.

First, it might be the case that subjects did not learn allocentric spatial representations of their chosen buildings at all. They might have painstakingly constructed the sketch maps in these experiments from egocentric representations, for example by imagining egocentric vectors from a particular vantage point, and estimating distances. If subjects can accurately estimate distances, then this procedure might yield sketch maps that are better than random, despite the absence of a metric cognitive map (subjects might well do this, for example, if they have only ever visited their chosen buildings by underground public transport). However, note that 1) in this case they would be violating the experiment instructions, which state that they need to know how to walk from any of the buildings to any of the others, and 2) it is much harder to draw accurate sketch maps when estimating from only one (or few) egocentric vantage points, as opposed to when a full 'map' is accessible allowing the choice of any building or points between buildings as vantage points.

Second, subject cognitive maps might be unstructured. However, according to the recall order paradigm, structure is evident from the recalls of a majority of subjects and subject maps. There is also the independent evidence of several distinct local reference frames, and of local neural representations (see above).

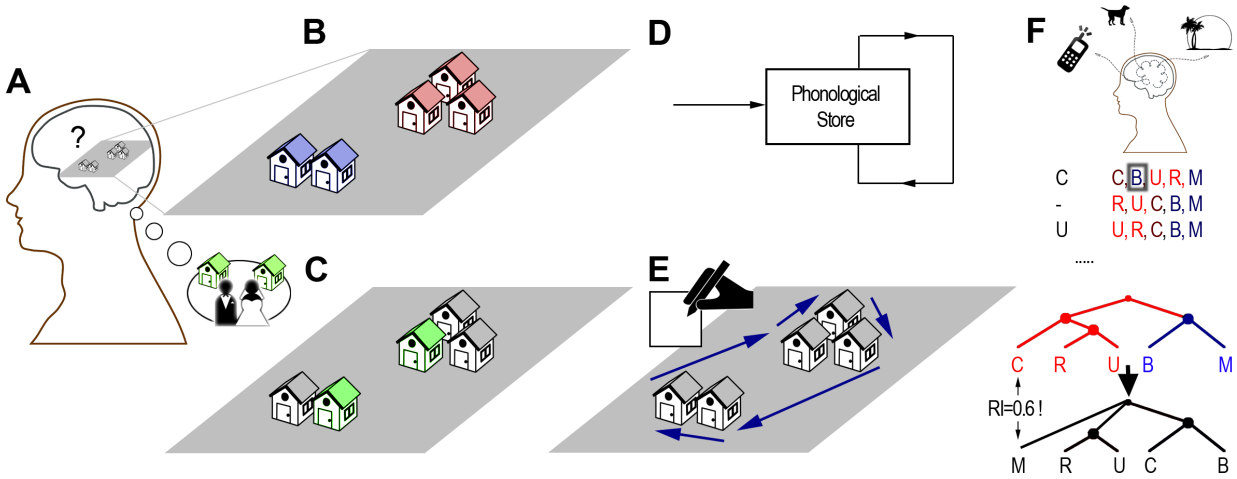


Figure 13: Possible obstacles to predicting subject cognitive map structures. A: Subjects may not have formed allocentric cognitive maps. B: Their maps may not have been structured. C: The apparent structure might be due to episodic memories, emotionally significant events, or other types of non-spatial long-term memory. D: Spurious structure might arise from articulatory rehearsal or other working memory strategies, instead of LTM. E: Subjects can list or sketch their buildings on paper, instead of recalling them from memory, to make the task faster and easier; usually resulting in circular recall sequences. F: Mind wandering or lapses in attention during recall sequences can cause tree analysis to reconstruct incorrect map structures.

Third, apparent structure might actually arise from non-spatial context effects or long-term memory events which happened at, or are relevant to, a sub-set of buildings or locations on a subject's cognitive map. For example, a subject might cluster together multiple restaurants after having had dinner at all of them with her significant other. When filling out the recall sequences, she might employ her salient episodic memories of these dinners to quickly recall these restaurants (and recall them together, which would lead to the tree analysis algorithm to assume that they are clustered together). It is difficult to exclude such influences in the real-world experiments, as most buildings familiar to subjects will have some sort of episodic memories associated with them. How frequent such influences are, and to what extent they distort apparent map structure, remain questions for future research (one approach might be trying to induce meaningful episodic memories in the virtual reality experiment, and measure their effects). However, if a majority of subject map structures had been affected by such context effects (which naturally cannot be modelled with the described features), reliable prediction would not be possible at all. The observation that a majority of structures *can* be predicted suggests that these influences affect a minority of recalled structures.

Fourth, spurious structures could appear in the recall sequences from phonetic or morphological name similarity in case subjects use articulatory rehearsal to facilitate quick recall; in which case it is a natural strategy to rehearse and recall similarly sounding object names together. This was indeed a significant influence in the verbally cued experiments (Experiment 1 and 2), although much weaker than the dominating influence of spatial distance. However, it seems that the effect can be mitigated substantially by changing the cue modality from verbal to visuospatial cues, which reduces the correlations between phonetic/morphological similarity and co-representation probability to insignificant levels (see Figure 6). A further possible objection related to working memory, that the uncovered structures might be learned during the experiment (instead of arising from long-term spatial memory), can be ruled out based on the approximately uniform distribution of outlier positions (the first few sequences were not more likely to be outliers than the last few sequences, and no evidence for any learning of map structures during the real-world experiments could be found in the data - see Supplementary Information for details).

Fifth, in the real-world experiments during which subjects were not observed, they could have lightened the cognitive load and speeded up the process by either writing down the list of buildings, or sketching a map on paper, and then reading instead of recalling. Although they were explicitly instructed to do everything from memory, without looking anything up, an unfortunate side effect of the monetary re-compensation

is that they have financial incentive to speed up the task (however, (Goodman et al., 2013) have found no significant difference between the ratio of correct answers between Mechanical Turk participants and supervised subject from a middle-class urban neighbourhood; although there was a significant difference to student participants). The proportion of subjects ignoring task instructions can be reduced by ensuring that most of their other tasks were accepted by requesters on MTurk (in these experiments, they were required to have at least 95% approval rating on previous jobs to ensure higher data quality). Furthermore, since the easiest strategy when using a list or a sketch on paper is to always use the same ordering, this should cause recall sequences to be circular, which can be detected in the data. As would be expected, the rate of circular recalls is significantly higher for the MTurk subjects (Experiment 3) - 12.6% - than for the student participants of Experiment 2 - 5.3%. However, they are still a minority of the data, and have been excluded in the reported analyses (as they lead to a lack of apparent structure).

An additional obstacle to predicting cognitive map structure is the rigidity of the tree analysis algorithm. Sub-maps are only recognized as such if they occur together, without interruption, in *every single recall sequence*. Figure 13 F illustrates an example (revisiting the example from Figure 4) where a distraction, which interrupts the sequence cued with 'C' and causes the participant to continue with 'B', for example because the distraction has reminded him of 'B'. This causes a substantially different extracted map structure - were a well-trained predictive model to predict the correct (CRU) - (BM) sub-map structure, it would show up as an incorrect prediction, and to have a Rand index of 0.6 instead of 1.0. Section 3.5 suggests a calculation of how many such such incorrectly inferred map structures there might be in our data, based on the percentage of recognized outliers using jackknifing.

Apart from devising a less simplistic outlier detection method, one possibility to reduce the occurrence of distractions - for future work - would be timing all recall sequences, and discarding those that exceed a temporal threshold, forcing participants to re-do the recall.

5. Conclusion

The way spatial memories of open, large-scale environments are structured has remained an unanswered question. In this paper, we have provided the first attempt at a quantitative answer, hypothesizing that cognitive map structure arises from clustering in some subject-specific psychological space, including (but not necessarily limited to) a list of features such as spatial distance, separating boundaries and streets, and visual and functional similarity, which we have proposed based on past empirical results. As this claim implies a strong dependence between whether or not objects are stored on the same representations, and these features, we have examined this dependence using subjects from over a hundred cities worldwide. We have found that there is a strong correlation between the probability of co-representation of buildings and their distance in these features (including, perhaps surprisingly, their visual similarities). Furthermore, we report that despite the noisy inference of subject map structures, they can be predicted correctly in a majority of cases, after learning subjects' psychological spaces and applying clustering, using a novel computational model of cognitive map structuring based on Bayesian models of cognition. Together, these results provide strong support for the clustering hypothesis, and for the plausibility of a Bayesian model of cognitive map structuring.

Acknowledgements

This work has been supported by EPSRC (Engineering and Physical Sciences Research Council) grant EP/I028099/1, and FWF (Austrian Science Fund) grant P25380-N23.

References

Andersen, P., Morris, R., Amaral, D., Bliss, T., & O'Keefe, J. (2006). *The hippocampus book*. Oxford University Press.

Baghshah, M. S., & Shouraki, S. B. (2010). Kernel-based metric learning for semi-supervised clustering. *Neurocomputing*, *73*, 1352–1361.

Barry, C., Lever, C., Hayman, R., Hartley, T., Burton, S., O'Keefe, J., Jeffery, K., & Burgess, N. (2006). The boundary vector cell model of place cell firing and spatial memory. *Reviews in the Neurosciences*, *17*, 71–97.

Barsalou, L. W. (2008). Grounded cognition. *Annu. Rev. Psychol.*, *59*, 617–645.

Bennett, A. T. (1996). Do animals have cognitive maps? *The journal of experimental biology*, *199*, 219–224.

Bensmail, H., & Celeux, G. (1996). Regularized gaussian discriminant analysis through eigenvalue decomposition. *Journal of the American statistical Association*, *91*, 1743–1748.

Brun, V. H., Solstad, T., Kjelstrup, K. B., Fyhn, M., Witter, M. P., Moser, E. I., & Moser, M.-B. (2008). Progressive increase in grid scale from dorsal to ventral medial entorhinal cortex. *Hippocampus*, *18*, 1200–1212.

Byrne, P., Becker, S., & Burgess, N. (2007). Remembering the past and imagining the future: a neural model of spatial memory and imagery. *Psychological review*, *114*, 340.

Canini, K. R., Shashkov, M. M., & Griffiths, T. L. (2010). Modeling transfer learning in human categorization with the hierarchical dirichlet process. In *ICML* (pp. 151–158).

Chen, X., & McNamara, T. (2011). Object-centered reference systems and human spatial memory. *Psychonomic bulletin & review*, *18*, 985–991.

Chitta, R., Jin, R., Havens, T. C., & Jain, A. K. (2011). Approximate kernel k-means: Solution to large scale kernel clustering. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 895–903). ACM.

Cohen, G. (2000). Hierarchical models in cognition: Do they have psychological reality? *European Journal of Cognitive Psychology*, *12*, 1–36.

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating amazon's mechanical turk as a tool for experimental behavioral research. *PloS one*, *8*, e57410.

Derdikman, D., & Moser, E. I. (2010). A manifold of spatial maps in the brain. *Trends in cognitive sciences*, *14*, 561–569.

Derdikman, D., Whitlock, J., Tsao, A., Fyhn, M., Hafting, T., Moser, M., & Moser, E. (2009). Fragmentation of grid cell maps in a multicompartment environment. *Nature neuroscience*, *12*, 1325–1332.

Ekstrom, A. D. (2015). Why vision is important to how we navigate. *Hippocampus*, .

Ekstrom, A. D., Kahana, M. J., Caplan, J. B., Fields, T. A., Isham, E. A., Newman, E. L., & Fried, I. (2003). Cellular networks underlying human spatial navigation. *Nature*, *424*, 184–187.

Elfes, A. (1989). Using occupancy grids for mobile robot perception and navigation. *Computer*, *22*, 46–57.

Foo, P., Warren, W. H., Duchon, A., & Tarr, M. J. (2005). Do humans integrate routes into a cognitive map? map-versus landmark-based navigation of novel shortcuts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 195.

Frank, L. M., Brown, E. N., & Wilson, M. (2000). Trajectory encoding in the hippocampus and entorhinal cortex. *Neuron*, *27*, 169–178.

Gablonsky, J. M., & Kelley, C. T. (2001). A locally-biased form of the direct algorithm. *Journal of Global Optimization*, *21*, 27–37.

Gärdenfors, P. (2004). *Conceptual spaces: The geometry of thought*. MIT press.

Gershman, S. J., & Blei, D. M. (2012). A tutorial on bayesian nonparametric models. *Journal of Mathematical Psychology*, *56*, 1–12.

Gibson, B. R., Rogers, T. T., & Zhu, X. (2013). Human semi-supervised learning. *Topics in cognitive science*, *5*, 132–172.

Gobet, F., Lane, P. C., Croker, S., Cheng, P. C., Jones, G., Oliver, I., & Pine, J. M. (2001). Chunking mechanisms in human learning. *Trends in cognitive sciences*, *5*, 236–243.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of mechanical turk samples. *Journal of Behavioral Decision Making*, *26*, 213–224.

Gower, J. C. (1975). Generalized procrustes analysis. *Psychometrika*, *40*, 33–51.

Greenauer, N., & Waller, D. (2010). Micro-and macroreference frames: Specifying the relations between spatial categories in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 938.

Griffiths, T. L., Canini, K. R., Sanborn, A. N., & Navarro, D. J. (2007). Unifying rational models of categorization via the hierarchical dirichlet process. In *Proceedings of the 29th annual conference of the cognitive science society* (pp. 323–328).

Han, X., & Becker, S. (2014). One spatial map or many? spatial coding of connected environments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*, 511.

Hirtle, S., & Jonides, J. (1985). Evidence of hierarchies in cognitive maps. *Memory & Cognition*, *13*, 208–217.

Holding, C. S. (1994). Further evidence for the hierarchical representation of spatial information. *Journal of Environmental Psychology*, *14*, 137–147.

Hommel, B., Gehrke, J., & Knuf, L. (2000). Hierarchical coding in the perception and memory of spatial layouts. *Psychological Research*, *64*, 1–10.

Hosmer, D. W., & Lemeshow, S. (2004). *Applied logistic regression*. John Wiley & Sons.

Howard, L. R., Javadi, A. H., Yu, Y., Mill, R. D., Morrison, L. C., Knight, R., Loftus, M. M., Staskute, L., & Spiers, H. J. (2014). The hippocampus and entorhinal cortex encode the path and euclidean distances to goals during navigation. *Current Biology*, *24*, 1331–1340.

Hughes, M. C., & Sudderth, E. (2013). Memoized online variational inference for dirichlet process mixture models. In *Advances in Neural Information Processing Systems* (pp. 1133–1141).

Hurts, K. (2008). Spatial memory as a function of action-based and perception-based similarity. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 1165–1169). SAGE Publications volume 52.

Jacobs, J., Weidemann, C. T., Miller, J. F., Solway, A., Burke, J. F., Wei, X.-X., Suthana, N., Sperling, M. R., Sharan, A. D., Fried, I. et al. (2013). Direct recordings of grid-like neuronal activity in human spatial navigation. *Nature neuroscience*, *16*,

1188–1190.
Jeffery, K. J. (2015). Distorting the metric fabric of the cognitive map. *Trends in Cognitive Sciences*, .
Jeffery, K. J., & Burgess, N. (2006). A metric for the cognitive map: found at last? *Trends in cognitive sciences*, *10*, 1–3.
Jones, D. R., Perttunen, C. D., & Stuckman, B. E. (1993). Lipschitzian optimization without the lipschitz constant. *Journal of Optimization Theory and Applications*, *79*, 157–181.
Khorsi, A. (2013). On morphological relatedness. *Natural Language Engineering*, *19*, 537–555.
Kjelstrup, K. B., Solstad, T., Brun, V. H., Hafting, T., Leutgeb, S., Witter, M. P., Moser, E. I., & Moser, M.-B. (2008). Finite scale of spatial representation in the hippocampus. *Science*, *321*, 140–143.
Lewis, D. D., & Gale, W. A. (1994). A sequential algorithm for training text classifiers. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 3–12). Springer-Verlag New York, Inc.
MacKinnon, J. G. (2009). Bootstrap hypothesis testing. *Handbook of Computational Econometrics*, (pp. 183–213).
Marchette, S. A., & Shelton, A. L. (2010). Object properties and frame of reference in spatial memory representations. *Spatial Cognition & Computation*, *10*, 1–27.
Marr, D., & Poggio, T. (1976). From understanding computation to understanding neural circuitry. *AI Memo*, . URL: `http://mit.dspace.org/handle/1721.1/5782`.
McNamara, T. P. (1986). Mental representations of spatial relations. *Cognitive psychology*, *18*, 87–121.
McNamara, T. P., Hardy, J. K., & Hirtle, S. C. (1989). Subjective hierarchies in spatial memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 211.
McNamara, T. P., Rump, B., & Werner, S. (2003). Egocentric and geocentric frames of reference in memory of large-scale space. *Psychonomic Bulletin & Review*, *10*, 589–595.
McNaughton, B. L., Battaglia, F. P., Jensen, O., Moser, E. I., & Moser, M.-B. (2006). Path integration and the neural basis of the 'cognitive map'. *Nature Reviews. Neuroscience*, *7*, 663–78. doi:`10.1038/nrn1932`.
Meilinger, T., Riecke, B. E., & Bülthoff, H. H. (2014). Local and global reference frames for environmental spaces. *The Quarterly Journal of Experimental Psychology*, *67*, 542–569.
Moser, E. I., Kropff, E., & Moser, M.-B. (2008). Place cells, grid cells, and the brain's spatial representation system. *Annual review of neuroscience*, *31*, 69–89. doi:`10.1146/annurev.neuro.31.061307.090723`.
Mou, W., & McNamara, T. P. (2002). Intrinsic frames of reference in spatial memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 162.
Nachar, N. (2008). The mann-whitney u: a test for assessing whether two independent samples come from the same distribution. *Tutorials in Quantitative Methods for Psychology*, *4*, 13–20.
Naveh-Benjamin, M., McKeachie, W. J., Lin, Y.-G., & Tucker, D. G. (1986). Inferring students' cognitive structures and their development using the "ordered tree technique". *Journal of Educational Psychology*, *78*, 130.
Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General*, *115*, 39.
O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map* volume 3. Clarendon Press Oxford.
Ong, C. S., Williamson, R. C., & Smola, A. J. (2005). Learning the kernel with hyperkernels. In *Journal of Machine Learning Research* (pp. 1043–1071).
Philips, L. (2000). The double metaphone search algorithm. *C/C++ users journal*, *18*, 38–43.
Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, *66*, 846–850.
Redner, R. A., & Walker, H. F. (1984). Mixture densities, maximum likelihood and the em algorithm. *SIAM review*, *26*, 195–239.
Reineking, T., Kohlhagen, C., & Zetzsche, C. (2008). Efficient wayfinding in hierarchically regionalized spatial environments. In *Spatial Cognition VI. Learning, Reasoning, and Talking about Space* (pp. 56–70). Springer.
Reitman, J. S., & Rueter, H. H. (1980). Organization revealed by recall orders and confirmed by pauses. *Cognitive Psychology*, *12*, 554–581.
Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2006). A more rational model of categorization. In *Proceedings of the 28th annual conference of the cognitive science society* (pp. 726–731).
Sato, N., & Yamaguchi, Y. (2009). Spatial-area selective retrieval of multiple object–place associations in a hierarchical cognitive map formed by theta phase coding. *Cognitive neurodynamics*, *3*, 131–140.
Settles, B. (2010). Active learning literature survey. *Computer Sciences Technical Report 1648*, .
Shelton, A. L., & McNamara, T. P. (2001). Systems of spatial reference in human memory. *Cognitive psychology*, *43*, 274–310.
Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*, 325–345.
Spelke, E., Lee, S. A., & Izard, V. (2010). Beyond core knowledge: Natural geometry. *Cognitive Science*, *34*, 863–884.
Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *science*, *331*, 1279–1285.
Thomas, R., & Donikian, S. (2007). A spatial cognitive map and a human-like memory model dedicated to pedestrian navigation in virtual urban environments. In *Spatial Cognition V Reasoning, Action, Interaction* (pp. 421–438). Springer.
Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological review*, *55*, 189.
Voicu, H. (2003). Hierarchical cognitive maps. *Neural Networks*, *16*, 569–576.
Wang, R. F., & Spelke, E. S. (2002). Human spatial representation: Insights from animals. *Trends in cognitive sciences*, *6*, 376–382.
Wiener, J. M., & Mallot, H. A. (2003). 'fine-to-coarse'route planning and navigation in regionalized environments. *Spatial*

cognition and computation, *3*, 331–358.

Xing, E. P., Jordan, M. I., Russell, S., & Ng, A. Y. (2002). Distance metric learning with application to clustering with side-information. In *Advances in neural information processing systems* (pp. 505–512).

Yang, L., & Jin, R. (2006). Distance metric learning: A comprehensive survey. *Michigan State Universiy*, *2*.

Zender, H., Mozos, O. M., Jensfelt, P., Kruijff, G.-J., & Burgard, W. (2008). Conceptual spatial representations for indoor mobile robots. *Robotics and Autonomous Systems*, *56*, 493–502.

Chapter 6

Towards real-world capable spatial memory in the LIDA cognitive architecture

Publication 4 / 4. Madl T, Franklin S, Chen K, Montaldi D & Trappl R, submitted. Towards real-world capable spatial memory in the LIDA cognitive architecture. *Biologically Inspired Cognitive Architectures*

Towards real-world capable spatial memory in the LIDA cognitive architecture

Tamas Madl[a,c,*], Stan Franklin[d], Ke Chen[a], Daniela Montaldi[b], Robert Trappl[c]

[a] *School of Computer Science, University of Manchester, Manchester M13 9PL, UK*
[b] *School of Psychological Sciences, University of Manchester, Manchester M13 9PL, UK*
[c] *Austrian Research Institute for Artificial Intelligence, Vienna A-1010, Austria*
[d] *Institute for Intelligent Systems, University of Memphis, Memphis TN 38152, USA*

Abstract

The ability to represent and utilize spatial information relevant to their goals is vital for intelligent agents. Doing so in the real world presents significant challenges, which have so far mostly been addressed by robotics approaches neglecting cognitive plausibility; whereas existing cognitive models mostly implement spatial abilities in simplistic environments, neglecting uncertainty and complexity.

Here, we take a step towards computational software agents capable to form spatial memories in realistic environments, based on the biologically inspired LIDA cognitive architecture. We identify and address challenges faced by agents operating with noisy sensors and actuators in a complex physical world, including near-optimal integration of spatial cues from different modalities for localization and mapping, correcting cognitive maps when revisiting locations, the structuring of complex maps for computational efficiency, and multi-goal route planning on hierarchical cognitive maps. We also describe computational mechanisms addressing these challenges based on LIDA, and demonstrate their functionality by replicating several psychological experiments.

Keywords:
spatial memory, LIDA, cognitive architecture, computational cognitive modeling

1. Introduction

Spatial representations are important for biological as well as artificial agents, in order for these agents to be able to localize, and navigate to, important objects and places (such as food sources or shelters). Current computer models for learning spatial representations either neglect cognitive plausibility in favour of performance, such as Simultaneous Localization and Mapping (SLAM) in robotics, or are incapable of running in large-scale, complex, uncertain environments perceived through noisy sensors.

Since biological cognition has been shaped by the structure, constraints, and challenges of the physical world, we argue cognitive architectures should take these into account as well. This argument is in accordance with the roadmap for the BICA challenge, which also places importance on real-life capability (Samsonovich, 2012). This paper describes an effort to take the LIDA (Learning Intelligent Distribution Agent) cognitive architecture (Franklin et al., 2014) closer to this goal. We introduce a novel conceptual and partially implemented, hierarchical spatial memory model, inspired by the neural basis of spatial cognition in brains, and provide a preliminary interface to realistic environments via the Robot Operating System (ROS) (Quigley et al., 2009). We demonstrate these extensions in three-dimensional simulated environments which include simulated physics and high-quality graphics, based on the Player/Stage/Gazebo simulator[1]. This simulator presents the same interface to the agent as real devices, and an agent able to control a robot in Gazebo is also able to control the same robot in similar environments in the real world, without any

*tamas.madl@gmail.com

[1] http://www.gazebosim.org/

changes to the control code (Rusu et al., 2007).

This paper describes an effort to extend the LIDA cognitive architecture by cognitively and biologically plausible spatial mechanisms, which are capable of handling the challenges of the real world, associated with noisy sensors and large-scale environments. We hypothesize and implement approaches to tackle the sensory noise, uncertainty, and complexity of realistic environments.

We build on and integrate our previous work investigating biologically and cognitively plausible implementations of Bayesian localization (Madl et al., 2014), Bayesian nonparametric clustering for map structuring (Madl et al., submitted), and route planning based on activation gradients[2] (Madl et al., 2013). The method for cognitive map correction (loop closing) is presented for the first time below. Although based on established mathematical tools from robotics, it is - to our knowledge - the first mechanism for large-scale cognitive map correction implementable in brains, and consistent with the replay phenomena observed in the rodent hippocampus (Carr et al., 2011).

The present work is also (to our knowledge) the first to provide implementations of these mechanisms in a both cognitively and biologically plausible fashion (fitting behaviour data and implementable in brains), and integrated within the same cognitive architecture. Further contributions include concrete implementations of some features listed by the BICA Table (Samsonovich, 2010) which until now were only part of conceptual LIDA, including basic stereo colour vision, a cognitive map, spatial learning, and fusing information from multiple types of sensors and modalities via Bayesian update.

1.1. Related work

Apart from the complex perception problem, the most challenging problems for building spatial representations in realistic environments include localization and mapping under sensory noise, and correcting incorrect representations when revisiting known locations (loop closing). The robotics community has developed several solutions to these problems - see (Thrun and Leonard, 2008; Durrant-Whyte and Bailey, 2006; Bailey and Durrant-Whyte, 2006; Williams et al., 2009). They have been designed to be accurate, not cognitively or biologically plausible, and rely on mechanisms which are difficult to implement in brains (e.g. many iterations performing operations on large matrices).

An exception is the partially connectionist RatSLAM system (Milford et al., 2004) which can learn robust maps in outdoor environments (Prasser et al., 2006), and close large loops successfully, if extended by a sophisticated data association method (Glover et al., 2010). Parts of it have been argued to be biologically plausible (Milford et al., 2010). However, RatSLAM has two disadvantages in the context of a cognitive model with long-term learning aiming for plausibility: 1) route planning only works along established routes (novel detours or shortcuts have not been demonstrated), 2) learned spatial information is mapped to a finite structure of fixed size (a continuous attractor network with wrapping connectivity) which cannot be expanded.

On the other hand, models which emphasize plausibility - cognitive architectures and plausible spatial memory models - mostly focus on simplistic simulated environments, usually with no sensory noise and limited size/complexity. There are a few neurally inspired spatial memory models which can deal with a limited amount of uncertainty and noise (Burgess et al., 2000; Strösslin et al., 2005; Barrera et al., 2011); but have only been tested in small indoor environments. See Madl et al. (2015) for a review.

2. Hypotheses

The LIDA cognitive architecture is based on Global Workspace Theory (GWT) (Baars, 2002; Baars and Franklin, 2009), an empirically supported theory of consciousness (Baars et al., 2013), and has been argued to be biologically plausible (Franklin et al., 2014, 2012). Just as the rest of LIDA can be mapped on to the underlying neuroscience (Franklin et al., 2012) (although not always in a one-to-one fashion), it is also the aim of the model proposed here to have parts which functionally correspond to the relevant areas in the brain representing space. This imposes some functional and connectivity constraints.

Apart from well-established implications of the neural representations in these brain areas, including the existence of a neural path integrator (Mc-

[2]Route planning in navigation space based on activation gradients has been proposed before (Schölkopf and Mallot, 1995; Burgess et al., 2000), but not on a hierarchy - as it is in this work - which significantly improves its performance on multigoal problems.

Naughton et al., 2006) and of cells representing current location (hippocampal 'place cells' (Moser et al., 2008)), the spatial memory model presented here also proposes and requires the following hypotheses. They are motivated by computational challenges facing agents operating in the real world - the ability to represent uncertainty, to estimate locations based on uncertain data, and to represent large amounts of spatial information efficiently are all essential for a real-life, embodied cognitive agent. Our choice of computational approaches (among all possible mechanisms) directly follow from these hypotheses.

1. Spatial uncertainty is encoded in brains, and spatial cues are integrated in an approximately Bayes-optimal fashion. The representation of uncertainty is a computational requirement for localization in the real world, given the unavoidable sensory inaccuracies and noise; and implies the existence of a mechanism for combining modalities with different accuracies. Apart from behavioural evidence substantiating such a mechanism (Cheng et al., 2007), we have found neural evidence based on single-cell recordings of rat hippocampal place cells in previous work, implying that these cells are able not only to represent, but also to combine, information from different modalities and the associated uncertainties (Madl et al., 2014).
2. Hippocampal replay (Carr et al., 2011) in awake mammals aids correcting cognitive maps based on revisited places (see Section 3). Despite local error correction by integrating spatial information, residual errors are still accumulating. This can lead to incorrect maps and to duplicate representations of the same places. Thus, a mechanism is required that can close loops, and correct maps, when revisiting places.
3. Instead of a single unitary and global map, cognitive maps are fragmented (Derdikman and Moser, 2010) and hierarchical (Hirtle and Jonides, 1985), and their structure arises from clustering, i.e. from a process grouping together objects which are 'close' in some psychological space. Hierarchical representations are ubiquitous in computer science and robotics, given their efficiency in terms of access and search time and memory use. These advantages are important for storing and accessing large-scale cognitive maps. We found evidence for hierarchies and for a clustering mechanism accounting for them in (Madl et al., submitted).
4. Human multi-goal route planning is consistent with a simple navigation strategy based on spreading activation on a recurrently interconnected, hierarchical, grid-like network of nodes representing locations (see Section 3 and (Madl et al., 2013)).

3. Spatial memory in brains

Spatial Memory encodes, stores and recalls spatial information about the environment and the self location of agents (biological or artificial), which they need to keep track of to successfully navigate. In most mammals, keeping track of position is achieved by path integration, which refers to updating the agent's position based on a fixed point and the movement trajectory (based on information from proprioceptive and vestibular systems as well as sensory flow (Mittelstaedt and Mittelstaedt, 1980; Fortin, 2008)), and is a noisy process accumulating large errors if uncorrected (Etienne et al., 1996).

Spatial information can be encoded in an egocentric fashion - relative to the agents body and head direction - or as allocentric representations, relative to environmental landmarks/boundaries. Here, we will describe major brain areas associated with these representations, and their correspondences in LIDA. For reasons of space, these descriptions will be very brief. More detail can be found in Madl et al. (2015).

The ability to recognize objects (e.g. landmarks, shelters, food sources, ...) is a prerequisite for encoding useful spatial memories. The brain areas involved in this complex functionality include the sensory cortices and the areas marked (1a) and (1b) in Figure 1 (Kiani et al., 2007; Davachi et al., 2003; Winters and Bussey, 2005; Wilson et al., 2013). The recognition of places is associated with its own area in the parahippocampal cortex, often called the Parahippocampal Place Area (PPA) (Epstein, 2008).

Representations of allocentric (world centered) in mammalian brains include place cells in the hippocampus, which represent spatial locations, firing only in small spatially constrained areas in an environment (ranging from 20cm or less to several meters in diameter, O'Keefe and Burgess, 1996; Kjel-

strup et al., 2008). They also participate in associating objects with specific places (Kim et al., 2011; Manns and Eichenbaum, 2009). In these cells, 'hippocampal replay' has been observed - a sequence of place cells associated with visited locations is frequently re-activated either in the same order or in reverse, on rapid (sub-second) timescales; suggested to aid memory consolidation (Carr et al., 2011). Replay often occurs in reverse at the end of a run, and forward when anticipating a run, and contains distance information between the firing fields (Diba and Buzsáki, 2007). Head direction is encoded by cells in a network including the Anterior thalamic nuclei, mamillary body, subiculum and EC (Taube, 2007). Border cells (Lever et al., 2009; Solstad et al., 2008) and boundary vector cells (BVCs) (Burgess, 2008a; Barry et al., 2006) in the subiculum play a role in representing the distance (and, for BVCs, the direction) to boundaries in the environment. Path integration, i.e. maintaining a location estimate by integrating self-motion signals, is performed by grid cells in the medial EC (Hafting et al., 2005; McNaughton et al., 2006).

Together, these cell types form a core part of the 'cognitive map', i.e. a map-like allocentric representation of the environment (McNaughton et al., 2006; Burgess, 2008b); and allow animals to keep track of where they are (place cells and grid cells), which direction they are facing (head direction cells), and where boundaries (border cells / BVCs) and objects (place cells) might be in their vicinity (see markers 2a-2c in Figure 1).

In addition to allocentric representations, there are multiple egocentric brain areas encoding spatial information relative to the animal. This includes the visual and auditory systems, and the precuneus ((3) in Figure 1), which is the main brain area concerned with egocentric representations and their use in controlling body and limb-centered actions (Zaehle et al., 2007; Kravitz et al., 2011; Vogeley et al., 2004) (for example, area 5d within the precuneus encodes 'reach vectors' between hand and target). The retrosplenial cortex (RSC) is involved with converting between egocentric and allocentric representations (Epstein, 2008). Finally, the basal ganglia encode guidance behaviours by means of associating spatial relations relative to the animal with actions (e.g. turn right at the rock). This is an effective strategy for well-known routes (Hartley et al., 2003); however, allocentric representations ('cognitive maps') are required in order to be able to plan novel routes or shortcuts.

The LIDA cognitive architecture

Here we will briefly introduce LIDA - see (Franklin et al., 2014, 2012) for a more detailed description of LIDA, and its relationship to the brain. The LIDA cognitive architecture is based on prevalent cognitive science and neuroscience theories (e.g. Global Workspace Theory, situated cognition, perceptual symbol systems, ... (Baars and Franklin, 2009)), and is one of the few cognitive models which are biologically plausible and to provide a plausible account for consciousness (Baars and Franklin, 2009; Baars et al., 2013), attention, feelings and emotions; and has been partially implemented (Franklin et al., 2014; Goertzel et al., 2010; Snaider et al., 2011).

Similarly to the action-perception cycle in neuroscience (Freeman, 2002; Fuster, 2002), LIDA's cognitive cycle has the purpose of selecting an action based on percepts (Figure 1 bottom). During each cycle the LIDA agent senses its environment, stores information in Sensory Memory and tries to recognize familiar objects, which are represented as nodes in Perceptual Associative Memory (PAM). It associates percepts with memories (declarative, episodic, spatial) recalled from a Sparse Distributed Memory (SDM) instance, creating models of the current situation (CSM) in the Workspace, which consist of the relevant PAM nodes copied to the Workspace. Several Structure Building Codelets[3] (SBC) - specialized 'processors' - operate on the pre-conscious representations in the Workspace. Subsequently, the agent decides what part is most in need of attention (Attention Codelets), which is moved to Global Workspace. Broadcasting the most salient[4] portion of CSM (bringing it to consciousness) enables the agent to choose actions applicable in the current situation from Procedural Memory and to select the action best serving its goals (Action Selection).

Figure 1 contains a tentative mapping to spatially relevant modules and mechanisms in LIDA to those in the brain, described below. It is intended to provide a starting point for the implementation of these mechanisms (taking inspiration from the

[3]In LIDA, the term codelet refers to small, special purpose processors or running pieces of software code; and corresponds to 'processors' in Global Workspace Theory (Baars and Franklin, 2009)

[4]We use 'salient' as an umbrella term for percepts which are important, urgent, insistent, novel, threatening, promising, arousing, unexpected etc.

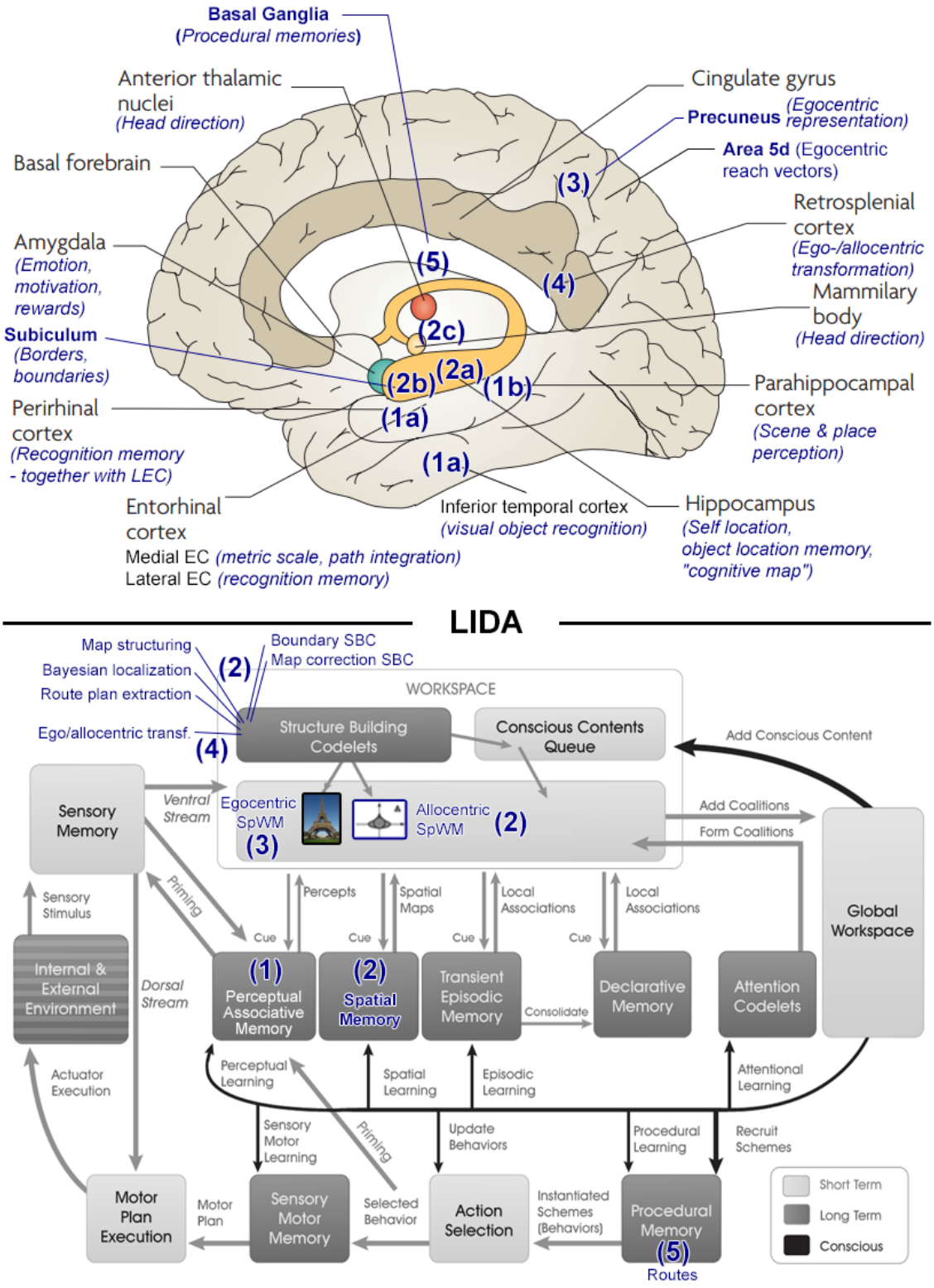


Figure 1: **Spatially relevant brain areas and LIDA modules**. Top: Neural correlates involved in spatial processing. Modified from (Bird and Burgess, 2008) with permission. Bottom: functionally corresponding modules and processes in LIDA. Only spatially relevant correspondences are marked here; see Franklin et al. (2014, 2012) for others.

underlying neural correlates), as well as to clarify LIDA's functionality to readers with relevant neuroscience knowledge by pointing out functional correspondences. This tentative mapping is by no means intended to suggest that LIDA implements exact neural mechanisms. Although heavily inspired by and resting on results from cognitive neuroscience and psychology, LIDA is a model of minds, not of brains (Franklin et al., 2012).

Towards real-world capable spatial memory in LIDA

The following subsections describe computational extensions made to LIDA in order to allow it to encode, store and recall spatial information obtained from real-world environments. Figure 2 provides an overview of these extensions. Note that some of these, such as the LIDA-ROS interface and the visual recognition mechanism in EPAM (Extended PAM), do not have correspondents in conceptual LIDA, and are not claimed to plausibly model minds. Rather, they make extensive use of already existing technologies for solving low-level problems (mainly vision and motor control), which are outside the scope of this work. Although efforts are under way to implement these mechanisms in a cognitively plausible fashion (see e.g. (McCall and Franklin, 2013; Agrawal and Franklin, 2014) for perceptual learning via cortical learning algorithms, and (Dong and Franklin, 2015) for action execution), they are not yet mature enough to facilitate the present application scenario.

Perception, object and uncertainty representation

LIDA's PAM contains nodes and links which are the building blocks of 'node structures', which are similar to and inspired by Barsalou's perceptual symbols (Franklin et al., 2014; Barsalou, 1999). PAM nodes represent higher-level features, such as objects, categories, relations, events, situations, feelings/emotions, etc; and are connected by PAM links, which are weighted and allow passing activation between the nodes. In the implementations in this paper, we have extended LIDA's PAM by an object recognition system based on a convolutional neural network (CNN), yielding EPAM (Extended PAM).

CNNs are a kind of deep learning architecture designed to process 2D or 3D data such as images - on which they have led to several breakthroughs (LeCun et al., 2015) -, and are usually trained by a gradient descent procedure called backpropagation. This algorithm has been criticized as not biologically realistic (Stork, 1989) (although there are versions of deep learning that can be implemented by biological neurons (Bengio et al., 2015)). However, despite these arguments concerning implementation, the representations found by state of the art CNNs trained on real-world images are highly similar to those recorded in the inferior temporal (IT) cortex of human and nonhuman primates (Khaligh-Razavi and Kriegeskorte, 2014; Yamins et al., 2013).

We have extended PAM by pre-trained CNNs [5] for object recognition (Szegedy et al., 2014) and road detection (Brust et al., 2015) - see Figure 3. The top layer (softmax layer) of the former was replaced by a classifier trained offline using a dataset of the buildings used in the Gazebo simulation, rendered from different perspectives and distances. (Learning should happen in a development fashion in LIDA, not offline; but this exceeds the scope of the current work). Since CNNs perform best on images containing a single object in the foreground, having difficulties with clutter, camera images were first segmented, and object recognition performed on the individual segments.

Allocentric spatial memory and localization

As described in Section 3, in brains, hippocampal place cells encode animals' current location in the environment, as well as providing object-place associations. Their equivalent in LIDA is implemented via a special type of PAM nodes, 'place nodes', each of which represent a specific region in the environment, and which reside in the Workspace (as part of the Current Situational Model). Place nodes can be associated with objects perceived to be at that particular location via PAM links - for example, agents' self-representation ('self' PAM node) can be associated with the place node representing their most likely location (which needs to be updated regularly). They are also initially connected recurrently to all their neighbours via PAM links. This has been argued to be a plausible connectivity pattern of the hippocampus (Moser et al., 2008; Csizmadia and Muller, 2008; Samsonovich and McNaughton, 1997).

[5] These CNNs were available from https://github.com/BVLC/caffe/wiki/Model-Zoo and https://github.com/cvjena/cn24

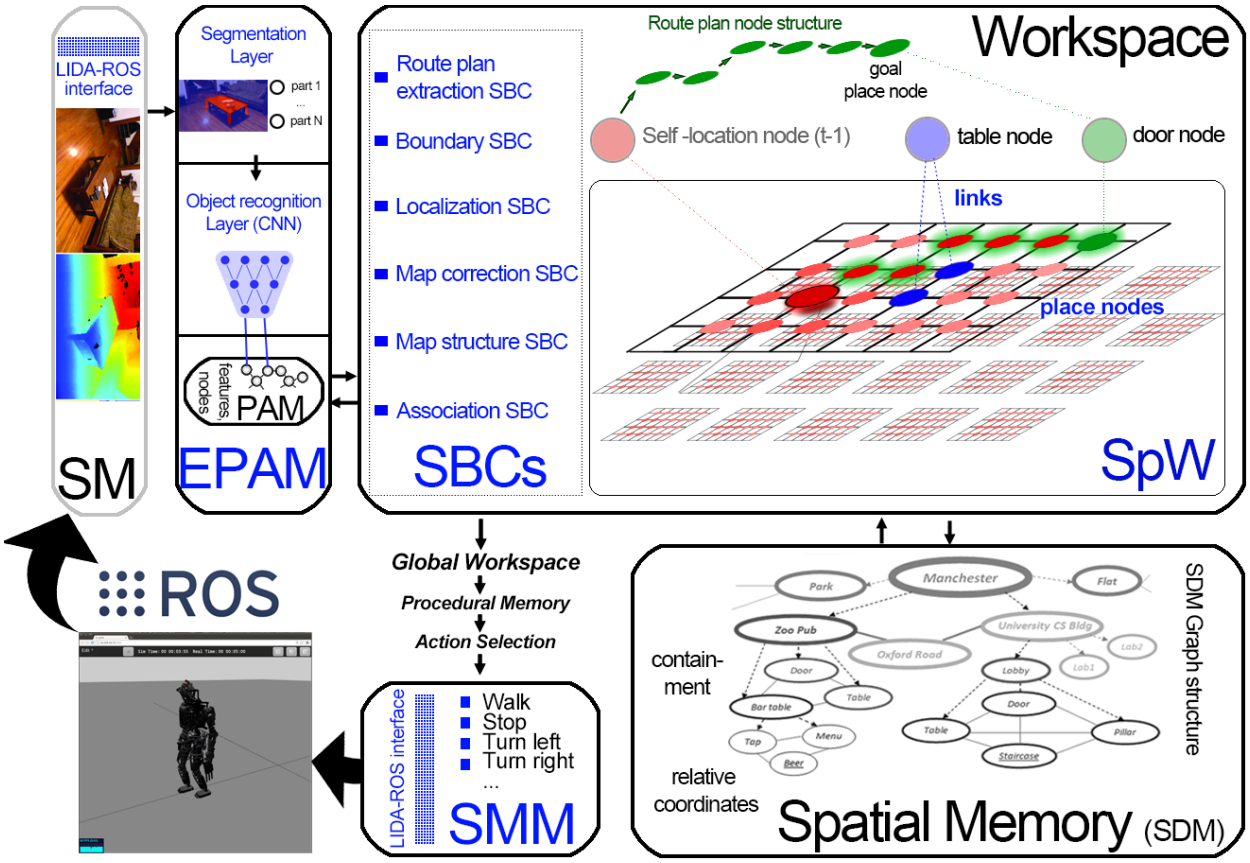


Figure 2: **Extensions to add spatial abilities to LIDA**. From the top left, clockwise: the LIDA-ROS interface transmits image and depth information (from stereo disparity) from the robot's visual sensors to Sensory Memory (SM). Object recognition is performed by CNNs in EPAM (Extended PAM), which pass activation to recognized PAM nodes representing objects. These can be associated with the place nodes corresponding to their most likely location in SpW (Spatial Workspace) in the Workspace. These place nodes, links between them, and object associations constitute 'cognitive maps', and are constructed, updated, and organized by various Structure Building Codelets (SBCs). Those with enough activation to be broadcast consciously can be learned as long-term SDM representations; and can also recruit route-following behaviours in Procedural Memory and Action Selection, leading to the execution of a low-level action in Sensory-Motor Memory (SMM), which is transferred to the robot via the LIDA-ROS interface.

Any PAM node in the Workspace representing currently or recently perceived objects (obstacles, landmarks, goals, etc.) in LIDA's Workspace can be associated via PAM links with spatial locations represented by place nodes. A node structure comprised of such object nodes, association links, and place nodes together constitute a 'cognitive map'. Multiple 'cognitive maps' can be used within the same environment in a hierarchical fashion (there can be maps and sub-maps on different scales and resolutions, and relative position and containment relations between them). This is consistent with neural and behavioural evidence that the human cognitive map structured (Derdikman and Moser, 2010) and hierarchical (Hirtle and Jonides, 1985) (see (Madl et al., submitted) for more extensive literature and evidence). It should be mentioned that the regular grid-like pattern of these place nodes, imposed for computational simplicity, is not biologically realistic, as no regularities have been found in the distribution of firing fields of place cells (however, a regular grid has been observed in the EC).

Although these maps are temporary, created and updated in the Workspace, they can be stored in the Spatial Memory module (which can encode trees and sequences (Snaider and Franklin, 2014)) as long-term memories if they are salient enough to be broadcast consciously. This long-term memory

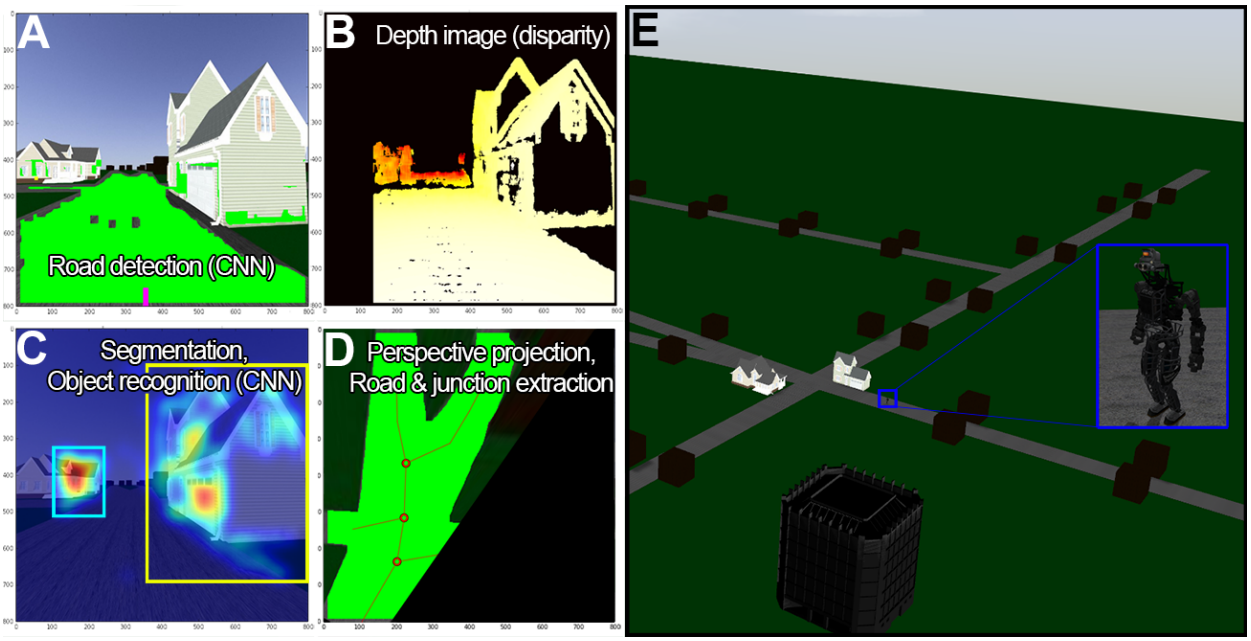


Figure 3: **Representations in Extended PAM (A-D) in one of the environments recreated in the Gazebo simulator (E).** A: Camera image with detected road. B: Depth image from binocular disparity. C. Likely objects from segmentation (hot colours), recognized by a CNN. D: Perceived road after denoising and projection based on the depth image.

storage mechanism has not been implemented yet.

Cognitive maps are assembled and updated by structure-building codelets (SBC) in the Workspace (LIDA's pre-conscious working memory). Each of these SBCs addresses a computational challenge associated with endowing an autonomous agent with spatial capabilities (see Figure 2):

- the 'Association SBC' associates large objects recognized by EPAM with place nodes, making use of distance information from stereo disparity to infer their approximate position and size,
- the 'Boundary SBC' detects boundaries in the Workspace, removing links at the locations of these boundaries (currently performed at the boundaries of recognized roads), only leaving links between traversable places (facilitating planning),
- the 'Localization SBC' is responsible for updating the link between the Self PAM node and the place node representing the agents most likely current position in the environment, using Bayesian inference to combine spatial cues,
- the 'Map correction SBC' corrects the map (closes the loop) based on revisited locations (see next section),
- the 'Map structure SBC' spawns new cognitive maps from parts of the current map, based on the proximity of objects represented on a map, in a process resembling clustering, and
- the 'Route plan extraction SBC' extracts shortest routes if a goal representation is present in the Workspace.

The Map structure SBC processes all place nodes which have associated objects, and clusters these objects based on 1) their spatial location, 2) their functional similarity, and 3) boundaries separating them; using Bayesian nonparametric clustering (as described and substantiated experimentally in (Madl et al., submitted)). Apart from accounting for the structure of cognitive maps, Bayesian nonparametric models has also been successful at accounting for category learning (Sanborn et al., 2006) and unifying rational models of categorization (Griffiths et al., 2007). This SBC groups together objects that are close to each other along the given features (in our case, spatial distance and functional similarity). The Map structure SBC spawns a new cognitive map (sub-map) for each

identified cluster, consisting of the objects in that cluster and their place nodes; and adjusts the density of place nodes depending on the area of this cognitive map (so that large-scale maps contain a low resolution and small-scale maps a high resolution place node grid). This process leads to a hierarchy of cognitive maps, a structure suggested to be employed by human spatial memory (Hirtle and Jonides, 1985; McNamara et al., 1989; Madl et al., submitted).

The Localization SBC is responsible for updating the agents estimated location after each movement, by linking its Self PAM node with the place node representing this location. Simply using path integration (odometry) to add up self-motion signals keeps accumulating errors (Etienne et al., 1996; Jeffery, 2007). This problem has been tackled in robotics in the framework of Bayesian inference, integrating information from odometry with sensory observations (the remembered vs. perceived locations of landmarks / boundaries in the environment) in a statistically optimal fashion (Thrun and Leonard, 2008). In its simplest form, this entails Bayesian integration of spatial cues. It has been argued that brains might employ a similar mechanism (Cheung et al., 2012; Madl et al., 2014).

Spatial uncertainty regarding the location of the agent, as well as the locations of recognized objects, is encoded as the parameters of multivariate Gaussians (i.e. means and covariances) attached to EPAM nodes in the Workspace and manipulated by the Localization SBC. After every movement, this SBC performs three steps. First, the location estimate of the agent is moved based on the self-motion signal. Second, the self-location estimate, and the landmark location estimates, are corrected in a Bayesian fashion (see equations (1) and (2) below). Finally, the links of the nodes representing them are updated to the place node corresponding to the best estimate. Correction takes into account the mean and uncertainty of the positions $\boldsymbol{x}$ updated by path integration of all movement signals $\boldsymbol{m}_1, ..., \boldsymbol{m}_i \in M$, as well as the current observations $\boldsymbol{o}_1, ..., \boldsymbol{o}_N \in O$ of landmark positions $\boldsymbol{l}_{i,1}, ..., \boldsymbol{l}_{i,N} \in L$, expressed as probability distributions. It calculates the Bayesian posterior of the location given the measurements at each step i, and the previous movement signals: $p(\boldsymbol{x}_i|L, M) = \eta p(\boldsymbol{x}_i|\boldsymbol{u}_{1:t})p(L|\boldsymbol{x}_i)$. Since conditioning on the position renders landmark positions independent (Montemerlo et al., 2002), the distributions representing landmarks can simply be multiplied:

$$p(\boldsymbol{x}_i|O, L, M) = \gamma p(\boldsymbol{x}_i|M) \prod_{j=1}^{N} p(\boldsymbol{l}_{i,j}|\boldsymbol{o}_j, \boldsymbol{x}_i). \quad (1)$$

In the above, $p(\boldsymbol{x}_i|M)$ represents the current location estimate from the path integration system (which adds up all movement signals $\boldsymbol{m} \in M$), $p(\boldsymbol{l}_j|\boldsymbol{o}_j, \boldsymbol{x}_i)$ represents estimated landmark positions, and γ is a constant normalization factors. After correcting the location estimate, the estimated landmark positions are also corrected:

$$p(\boldsymbol{l}_{i,j}|\boldsymbol{o}_j, \boldsymbol{x}_i) = \eta p(\boldsymbol{o}_j|\boldsymbol{l}_{i,j}, \boldsymbol{x}_i)p(\boldsymbol{l}_{i-1,j}) \quad (2)$$

In our case, $p(\boldsymbol{o}_j|\boldsymbol{l}_{i,j}, \boldsymbol{x}_i)$ is a Gaussian centered at the location of the observed landmark (measured from stereo disparity). $p(\boldsymbol{l}_{i-1,j})$ is simply the previous estimate of the same landmark, and η a normalization constant. The repeated steps of Bayesian localization which implement equations (1) and (2), expressed in this simple form[6], are implementable in brains and in spiking neural networks (Figure 4).

Path integration has been shown to be performed by grid cells in the entorhinal cortex (McNaughton et al., 2006). We recently presented evidence that hippocampal place cells are able to perform the correction step, based on neuronal recordings of several hundred place cells and multiple different environments, in which the firing fields of these cells corresponded to the predictions of a Bayesian integration model (Madl et al., 2014). In the same paper, we have also suggested how coincidence detection, observed in place cells (Jarsky et al., 2005; Takahashi and Magee, 2009; Katz et al., 2007), can implement the multiplication in Equation (1). One can interpret each input spike as a sample from a probability distribution, in which case coincidence detection (which only leads to output spikes if the input spikes arrive in close temporal proximity) performs multiplication (Koch and Segev, 2000) of these distributions (see Figure 4). Finally, we have argued that phase resetting observed in grid cells can implement the update step, completing the localiza-

[6]Our formulation contains heavy simplifications compared to robotics solutions, which are far more accurate - for example, only taking into a account local landmarks for the correction step, and omitting angles from the state representation. We sacrifice accuracy for plausibility, noting that humans are not very accurate at localization either (see Results).

tion cycle. Based on the observation that it partially accounts for single-cell recording data in multiple environments (Madl et al., 2014), and that it can be implemented as a biological neural network in a straightforward fashion, we think this kind of Bayesian correction constitutes a plausible model of local spatial error correction.

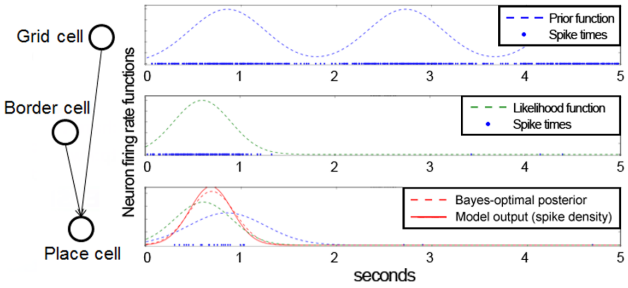


Figure 4: **Approximate Bayesian cue integration in spiking neurons**. Calculating the posterior probability distribution of the current location (equation (1)) requires multiplying a prior location distribution from path integration (represented by grid cells) with likelihood distributions from measurements of objects or boundaries (here represented by a border cell). If the place cell receiving input from the grid and border cells performs coincidence detection, it can multiply its inputs, yielding an approximate Bayesian posterior, and representing the associated uncertainty via the size of its firing field. A Bayesian model can account for hippocampal place field sizes in behaving rats. (Figure adapted from (Madl et al., 2014)).

The Route plan extraction SBC creates PAM node structures representing the shortest path to the agents current goal, if such a goal is currently present in the Workspace; leveraging the interconnected place node network constituting cognitive maps. The kind of recurrently interconnected network constituted by place nodes facilitates a very simple path planning mechanism (Figure 5). Assuming that every goal location G passes activation through the network, the distance to the goal can be decreased by moving to the adjacent neighbour node with the highest activation. If the nodes representing the locations of possible obstacles are connected with zero or near-zero weights, this mechanism can implement obstacle avoidance as well as path planning. Crucially, this activation-based planning mechanism operates on a hierarchy of 'cognitive maps', rather than on a single level. We argue that this allows better solutions of multi-goal navigation problems such as the travelling salesman problem. The evaluation of this planning mechanism against human data was briefly described in (Madl et al., 2013) (for details see the Supplementary Information).

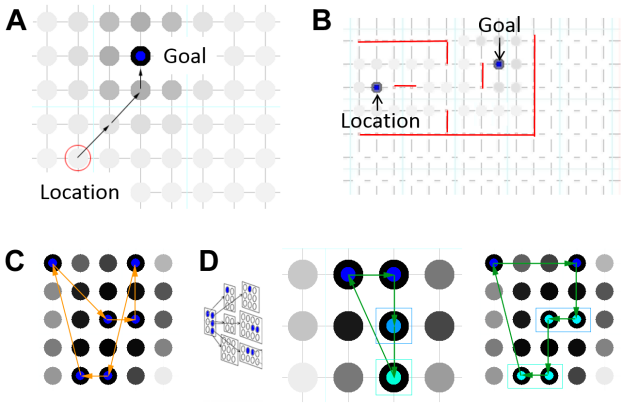


Figure 5: **Route planning on recurrently interconnected place nodes**. A: Single goal routes can be obtained by following an activation gradient to a goal. B: Obstacle avoidance can be implemented by setting connection weights to zero near boundaries (red lines). C: On a flat grid, following activation gradients can lead to sub-optimal paths for multi-goal navigation. D: However, when operating on a hierarchy - planning rough, low-resolution routes first, and then refining them on higher resolution maps -, this mechanism can yield near-optimal solutions.

The Map correction SBC serves the role of making large-scale corrections (as opposed to the local corrections made by the Localization SBC) when revisiting a known location. It is described in detail in the next subsection.

Loop closing - fixing previously learned maps

If uncorrected, accumulating path integration errors eventually render learned spatial representations useless; a problem necessitating the use of other modalities for map learning. Integrating spatial information in an approximately statistically optimal (Bayesian) fashion, as described above, helps correct local maps. However, only the agent's current location and the locations of currently perceived objects are updated with our procedure. When traversing large cycles (loops) in an environment and returning to a previously visited location, the remaining errors still accumulate and prevent this loop to be represented correctly, causing multiple representations of the same places (of subsequently revisited places) - see Figure 6.

Therefore, a mechanism is needed to correct the representation of locations encountered during

loops (such a correction is called 'loop closure' or 'closing the loop' in robotics literature (Williams et al., 2009)). This section outlines a biologically plausible solution to this problem, and its relation to phenomena observed in hippocampal neurons. This solution is also used by the Map correction SBC to correct errors in learned cognitive maps.

Although the problem of accumulating errors and the resulting need to correct maps with sensory information has been identified early in spatial modelling literature (McNaughton et al., 1996), the question how brains might 'close the loop' has received very little attention, and no plausible mechanisms have been proposed to the authors' knowledge. The large majority of robotics solutions to this problem require many iterations over huge matrices containing information regarding every position ever visited (Thrun and Leonard, 2008; Durrant-Whyte and Bailey, 2006; Bailey and Durrant-Whyte, 2006; Williams et al., 2009), and are thus neurally implausible. However, a probabilistic perspective on this problem can still help find a plausible candidate algorithm, consistent with hippocampal replay as the correction mechanism, which we summarize below.

First, let us assume that it is sufficient to correct the route taken during the loop. Local, currently perceived landmark positions are corrected separately as described above. When performing large-scale loop closing, our scheme applies the same correction to a position and the local landmarks around it[7]. We also make the assumption that correction only concerns position representations and not angular representations, since there is neuronal evidence for the former but not the latter (replay of encountered information happens in place cells, but has not been observed for direction-sensitive neurons such as head-direction cells in the post-subiculum (Brandon et al., 2012)).

The available information includes the path X consisting of estimated, recently visited locations $\boldsymbol{x}_0, ..., \boldsymbol{x}_m \in X$, and a set of constraints $\boldsymbol{c}_1, ...\boldsymbol{c}_m \in C$ specifying how far two locations should be from each other - this includes distances from the path integration system for subsequent locations, and equivalence constraints (with zero distance) when revisited places are recognized. We will temporarily assume simultaneous access to all path integration constraints, and will drop this implausible requirement later. Each constraint between two locations is represented as a Gaussian with the measured distance $\boldsymbol{c}_i$ as the mean, and the associated uncertainty represented by the covariance S_i (e.g. path integration is inexact - high uncertainty; but a recognized revisited place is at the same location - low uncertainty). The correct path is the one that is most consistent with all known constraints (known distances between the locations); or, from a probabilistic perspective, the one that maximizes the conditional probability of the locations constituting the path, given the constraints[8]:

$$P(X|C) \propto \prod_{i=1}^{m} P(\boldsymbol{c}_i|X). \tag{3}$$

Since each constraint is represented as a Gaussian over the distance between a pair of locations a_i and b_i, $P(\boldsymbol{c}_i|X) \propto \mathcal{N}(\boldsymbol{x}_a - \boldsymbol{x}_b; \boldsymbol{c}_i, S_i)$, and the conditional probability is

$$P(X|C) \propto \prod_{i=1}^{m} \exp -\Big(\frac{1}{2}||\boldsymbol{x}_a - \boldsymbol{x}_b - \boldsymbol{c}_i||_{S_i}\Big). \tag{4}$$

We will denote the discrepancy between the constraint i and the difference between corrected locations a_i and b_i as $\boldsymbol{d}_i = \boldsymbol{x}_a - \boldsymbol{x}_b - \boldsymbol{c}_i$. Under ideal conditions without noise and errors, all d_i would be zero; but in realistic environment there will be discrepancies between estimated and measured differences. The 'best' path estimate maximizes $P(X|C)$, or equivalently minimizes its negative logarithm $-logP(X|C)$ (minimizes the discrepancies):

$$X_{ML} = \arg\max_X P(X|C) = \arg\min_X \sum_{i=1}^{m} ||\boldsymbol{d}_i||_{S_i^{-1}} \tag{5}$$

The constraints can be represented as the edges of a graph which has the locations along the path

[7] Unlike the strong evidence for hippocampal replay concerning place cells representing recently visited locations, it is unclear whether cells associated with landmarks are also 'replayed'. Therefore, we forgo separate landmark correction in loops for now.

[8] In robotic SLAM solutions, the path likelihood would also depend on all landmark observations. We omit them here because our loop closing procedure updates each position along with the path together with its local landmarks, applying the same translation to both, which renders the observation conditionals constant; once again sacrificing accuracy for plausibility.

as vertices. Let us denote the incidence matrix H of this graph, such that $H_{i,j}$ is 1 if there is a link between i and j, -1 for links between j and i, and 0 if there is no link. We can express the distances between the pairs of locations for which there is a constraint, given any hypothetical locations X, as $C' = HX$. The discrepancy between actual and hypothetical distances becomes $C - C'$, and the associated uncertainty a matrix S containing the S_i as sub-matrices, which yields the ML equation in matrix form - expressing that the best hypothetical locations are the ones minimizing the discrepancy.

$$X_{ML} = \arg\min_X (C - HX)^T S^{-1} (C - HX) \quad (6)$$

This formulation is called linear least squares SLAM (Frese, 2006) in robotics literature (where the solution is made more complicated - but more accurate - by also representing and correcting angles along the path, making the equations non-linear). The minimum in equation (6) is given by $X = (H^T S^{-1} H)^{-1} H^T S^{-1} C$, and can be re-written to a form $Ax = b$ and approximated using the Gauss-Seidel method - which has been shown to be implementable as a neural network (Delgado and Fausett, 1995). Unfortunately, this requires accessing every visited location in X and all uncertainties S_i, which cannot be implemented in a biological neural network within the sub-second durations of hippocampal replay (Carr et al., 2011). Obtaining uncertainty estimates from the encoding suggested in Figure 4 would take seconds for every place cell.

Fortunately, there is a more plausible solution which can be implemented neurally. It has been argued that Spike-Time Dependent Plasticity (STDP) can implement gradient descent in biological neurons (Bengio et al., 2015). Our starting point is the stochastic gradient descent-based maximization of $P(X|C)$ described in (Olson et al., 2006), which suggests the following gradient with respect to constraint i:

$$\Delta X \approx \alpha (J S^{-1} J)^{-1} J_i^T S_i^{-1} \boldsymbol{d}_i, \quad (7)$$

where α is a learning rate, J is the full Jacobian of all constraints with respect to the locations, and J_i the Jacobian of constraint i. Because constraints apply to locations incrementally (with zero sensory errors, the correct current location would be $\boldsymbol{x}_c = \sum_i \boldsymbol{c}_i$), the Jacobian is also incremental, spreading out the discrepancy $\boldsymbol{d}_i = (\boldsymbol{x}_a - \boldsymbol{x}_b - \boldsymbol{c}_i)$ over an entire loop (by means of having a structure similar to the incidence matrix). This means the Jacobian need not be explicitly computed or represented. For a given loop closed by $\boldsymbol{c}_i$ with uncertainty S_i, let us assume unchanging path integration uncertainties S_P for each movement within the loop, and introduce a loop precision parameter A_i specifying the uncertainty of the current loop closure in relation to that of path integration, $A_i = S_i / S_P$. The correction applied to any single location $\boldsymbol{x}_j$ visited after the recognized previous location $\boldsymbol{a}_i$ (i.e. if $j > a_i$) thus becomes:

$$\Delta \boldsymbol{x}_j \approx \alpha \boldsymbol{d}_i \frac{\sum_{k=a+1}^{j} S_i^{-1}}{\sum_{k=a+1}^{min(j,b)} S_P^{-1}} = \alpha A_i \boldsymbol{d}_i p_j, \quad (8)$$

where $p_j = (min(j, b_i) - a_i - 1)/(b_i - a_i - 1)$ denotes how far $\boldsymbol{x}_j$ lies along the loop, with $0 \leq p_j \leq 1$.

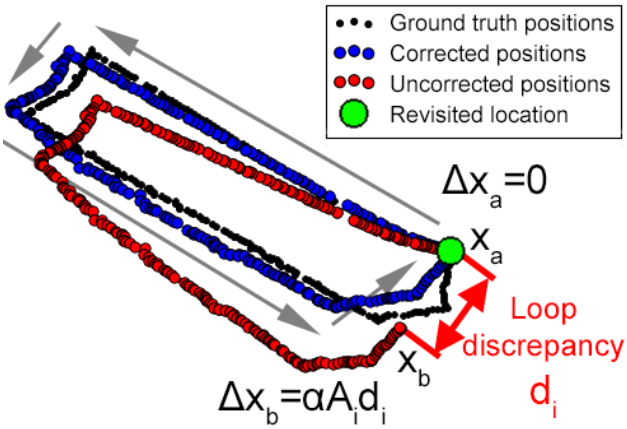


Figure 6: **Loop closing performed by the Map correction SBC**. Correcting estimated positions along a path when re-visiting a known place (large green dot), after traversing a large loop. Recognizing this place yields the knowledge that current estimated location $\boldsymbol{x}_b$ should equal $\boldsymbol{x}_a$; and the correction $\boldsymbol{d}_i$ based on the discrepancy is applied proportionally to all visited places along the loop. This backward correction is consistent with hippocampal replay.

Conveniently, we can neglect path integration constraints - they are already included in the path X, and, since they concern subsequent locations with $b = a + 1$, they lead to $\Delta x = 0$ according to equation (8). The updates only concern loop closing constraints. Given that the distance to the same place when re-visiting it is zero, $\boldsymbol{d}_i = \boldsymbol{x}_a - \boldsymbol{x}_b$. Furthermore, we don't have to re-activate all locations ever visited; only those in the loop. The ensuing

correction mechanism is simple (and easily implementable with neurons): when a loop closure is detected, the locations along the loop are iteratively corrected with the discrepancy between estimated and observed location according to equation (8). All locations subsequent to a re-visit are corrected by as much as the last point in the loop. The iteration proceeds backwards, starting at the estimated location at the re-visited place, and has to run several times to approximate a near-optimal solution. This is consistent with backward replay of visited locations in hippocampal place cells (Carr et al., 2011), with the presence of distances between locations encoded in such replays (Diba and Buzsáki, 2007), and with the observation that replay happens significantly more often than the number of times the animal re-visits places.

The described procedure carried out regularly by the Map correction SBC after a loop closure has been detected. It simply spreads out the discrepancy $\boldsymbol{d}_i$ proportionally along the place nodes representing the traversed loop, according to Equation (8) (see Figure 6). The Map correction SBC also corrects the positions of encountered buildings, and of the traversed road, stored on the cognitive map (i.e. the same correction is applied to building nodes and road nodes as to the $\boldsymbol{x}_j$ closest to them).

Apart from behavioural predictions regarding cognitive map accuracy, validated in the next subsection, and the prediction that hippocampal replay (Carr et al., 2011) might (also) serve the purpose of correcting cognitive maps, this suggested mechanism also yields a quantitative prediction on a cellular level, assuming that the synaptic strength place cells depends on the distance $\boldsymbol{d}_{pf}$ between their place fields. For example, (Csizmadia and Muller, 2008) suggest that the synaptic weight converges to $S = exp(-kd_{pf})$. which for small kd_{PF} can be approximated by $S = 1 - kd_{pf}$. Furthermore, STDP implies a weight change proportional to the change in post-synaptic voltage potential (Bengio et al., 2015). Under these assumptions, our suggested cognitive map correction mechanism implies that after re-visiting a location, during subsequent hippocampal replay, for a pair of place cells which are sufficiently close together for the approximation to hold, changes in post-synaptic voltage potential will be approximately proportional to the correction magnitude $\Delta\boldsymbol{x}$, i.e. to the amount the place field has shifted during replay. It is clear from empirical data that place fields shift after re-visiting locations in an environment (Mehta et al., 2000), and that backward replay contains distance information between place fields (Diba and Buzsáki, 2007). We leave the verification of the mentioned prediction for future work.

Results

This section reports results obtained by LIDA agents with the extensions described above, reproducing data from psychological experiments. These experiments were chosen to compare the agent's spatial estimation accuracies, and cognitive map structures, with human subjects.

Instead of free exploration, the routes in the experiments below were pre-programmed into the agents' long term memory, by storing the turns to be taken in the form of schemes (percept-action mappings) in Procedural Memory, for the following reasons. In Experiment 1, closely reproducing the participant trajectories (as opposed to exploration behaviour) was crucial to modelling accumulating uncertainty. In Experiments 2 and 3, subjects' exploration trajectories in their home towns were not known (having happened years or decades before the experiment). Furthermore, exploring environments on the scale of the participant cities modelled in Experiment 2 in tractable timeframes would have required an intelligent exploration strategy, which we have not implemented yet in LIDA. Therefore, the agent was given the turns it should take.

All other information came from noisy sensors, and no ground truth information was provided to the agents, which makes the experiments suitable for evaluating spatial representation accuracy.

Experiment 1 - Localization and cue integration

In order to substantiate the Bayesian localization and cue integration mechanism, we have replicated a behavioural experiment Nardini et al. (2008) investigating the integration of self-motion and sensory information in location estimation. In this experiment, subjects were asked to pick up a series of glowing objects in a dark room and to subsequently return the first object to its original location. In the self-motion+landmarks condition, there were three landmarks available for orientation, and subjects were not disoriented - both sources of information were available. In the landmarks condition, subjects were disoriented by turning to deprive them of orientation information. In the self-motion condition, subjects were not disoriented but the landmarks were turned off.

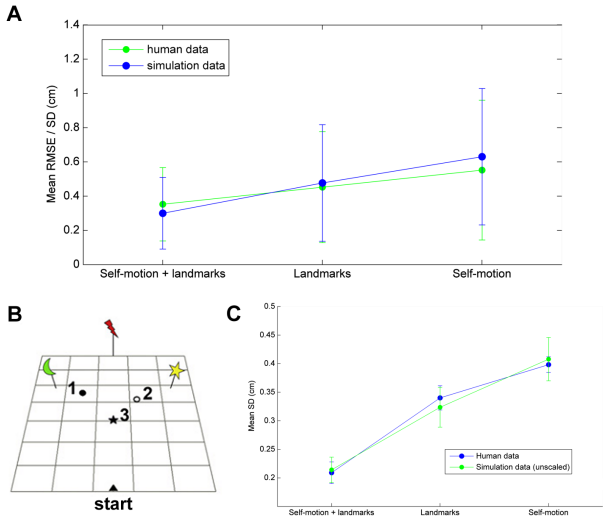


Figure 7: **Position errors and standard deviations in the cue integration experiment by Nardini et al. (2008)**. A. Mean RMSE (root mean squared errors) of participants, and mean SD (standard deviation), for the responses of human subjects (green) and the agent (blue), respectively. B. The experiment environment. Participants had to pick up objects 1-3 in order, and then replace object 1. The colored objects (moon, star, lightning) are the landmarks. (From Nardini et al. (2008)). C. Mean SD of participants (green) and the agents (blue)

To simulate this experiment, the same environmental layout (with accurate object distances) was reproduced in a simulation. The agent performed Bayesian localization as described above. Distance estimation inaccuracies were set to 3%, which is a frequently observed distance estimation error in virtual (Waller, 1999; Murgia et al., 2009) and real environments (Plumert et al., 2005; Grechkin et al., 2010). The two remaining noise parameters (linear and angular self-motion estimation inaccuracies) were adjusted to fit the data using coordinate descent. Path integration errors were modelled by multiplicative 1-mean Gaussian noise, since variability in human odometry is proportional to magnitude (Durgin et al., 2009). Figure 7 shows the simulation results, which are consistent with the empirical data for adult subjects.

Experiment 2 - Cognitive map accuracy (real environments)

Here we replicate map accuracies of Experiment 3B in (Madl et al., submitted), in which participants were asked to pick 8 very familiar buildings (such that they knew how to walk from any one to the other), and to create a sketch map by indicating their positions on a featureless canvas on a computer. Sketch maps were linearly translated, rotated and scaled to fit the correct map best using Procrustes analysis (Gower, 1975). Each subject produced three sketch maps, of which those not significantly better than random guessing were excluded. Sketch maps spanning an area larger than $4km^2$ were also excluded to reduce computational load. This left 19 participants, and a total of 28 different maps (environments) in 21 different cities. To reduce computational load, only the roads (and adjacent buildings) were modelled which allowed getting from one of these buildings to the other, i.e. which lay along one of the $\binom{8}{2} = 28$ shortest routes between two respective buildings for each map. These roads and buildings were recreated with the correct real-world distances in the simulation (geospatial information was obtained via Google Maps API[9]), yielding multiple routes several kilometers long.

Figure 8 compares the errors of the maps learned by the agent with human sketch maps, after adjustment of the linear and angular path integration noise parameters by coordinate descent. Map errors are measured as the sum of squared errors (SSE) between the correct geographical building locations, and the locations estimated by the participants / by the model. Unlike the model predictions, which are already in the correct reference frame, human data is linearly translated, rotated and scaled first to fit the correct map. Errors averaged over all maps are $1.07km^2$ ($\sigma = 0.85$) for humans, and $1.08km^2$ ($\sigma = 1.39$) for the model, and the model errors correlate with human errors with $r_{m,h} = 0.80$ ($p = 2.42 * 10^{-7}$), with a coefficient of determination (proportion of explained variance) of $R^2 = 0.60$ which suggests that the model explains the majority of the variance in human map error data.

Note that this model only uses the eight buildings the participant indicated as being very familiar to recognize having revisited a place and to correct maps. Along routes of this size, humans can presumably re-identify more than these eight places. Even in areas without salient landmarks, a matching visual sequence while walking can trigger a feeling of familiarity. We will implement this kind of episodic sequence-based place recognition in future work.

[9] https://developers.google.com/maps/

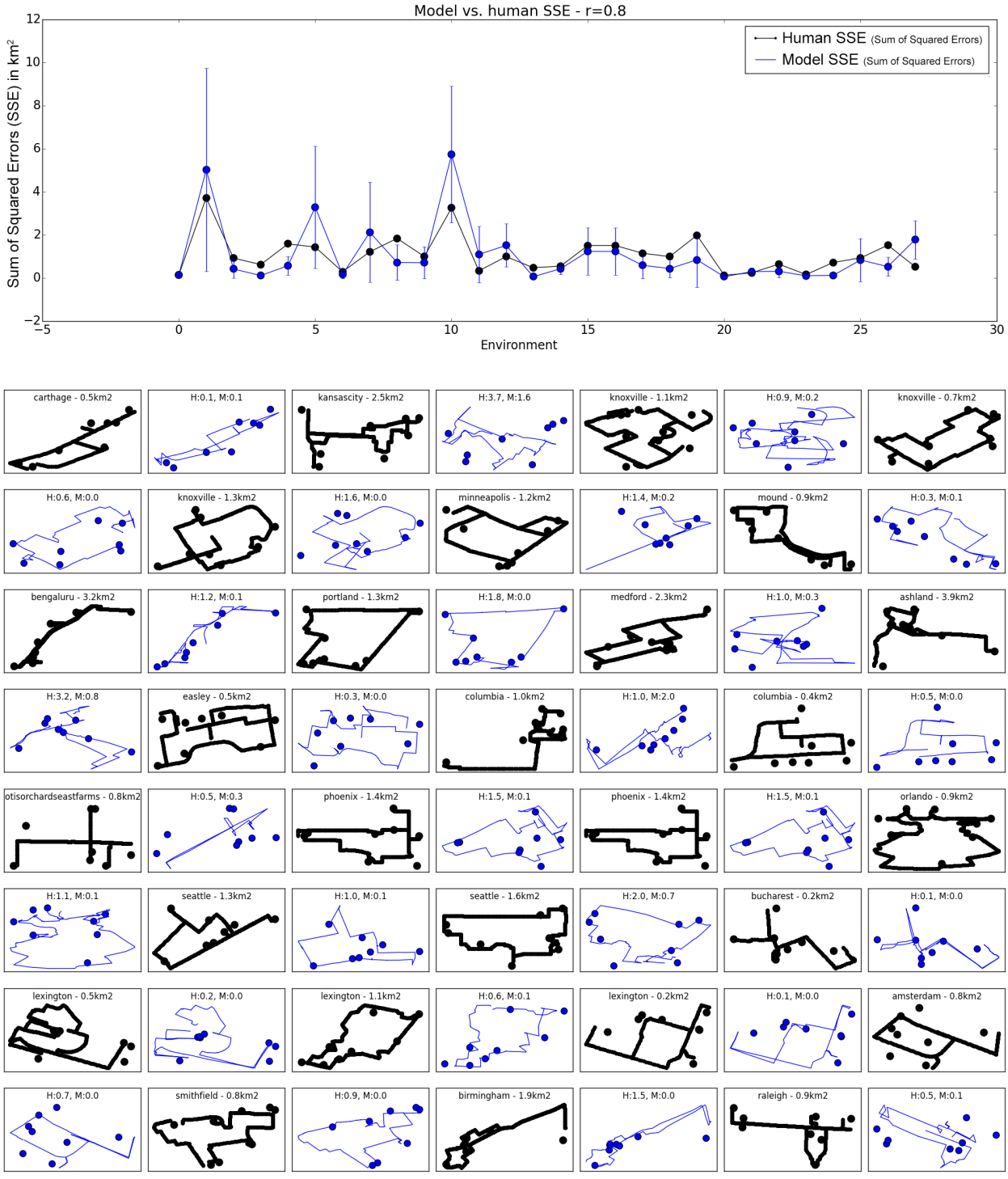


Figure 8: **Comparison with human and model errors over all environments (top), and plots containing the ground truth (black) and learned (blue) street and salient building locations.** Titles indicate the city name and region area for the ground truth, and (H)uman and (M)odel errors for the model subplots.

Conclusion

In order to tackle challenges posed by noisy sensors and complex, uncertain environments, we have extended LIDA by CNN-based perception, and by mechanisms for learning and correcting cognitive maps facilitating navigation. These include novel reinterpretations of coincidence detection in place cells as approximate Bayesian cue integration, and hippocampal replay as cognitive map correction; and suggested computational and algorithmic models of these phenomena, consistent with the 'Bayesian brain' paradigm (Knill and Pouget, 2004). We have also compared spatial representation accuracies to human subjects. Although a large number of issues remain to be solved for real-world-capable autonomous agents (including developmental learning of perceptual representations and affordances, visual place recognition in the absence of salient landmarks, long-term spatial and episodic memories, transferring learned spatial knowledge and expectations between environments, and spatial reasoning, to name just a few), we believe these extensions provide a first step towards a cognitive architecture combining biological plausibility and real-world functionality.

Acknowledgements

This work has been supported by EPSRC (Engineering and Physical Sciences Research Council) grant EP/I028099/1, and FWF (Austrian Science Fund) grant P25380-N23.

References

Agrawal, P., Franklin, S., 2014. Multi-layer cortical learning algorithms, in: Computational Intelligence, Cognitive Algorithms, Mind, and Brain (CCMB), 2014 IEEE Symposium on, IEEE. pp. 141–147.

Baars, B.J., 2002. The conscious access hypothesis: origins and recent evidence. Trends in cognitive sciences 6, 47–52.

Baars, B.J., Franklin, S., 2009. Consciousness is computational: The LIDA model of global workspace theory. International Journal of Machine Consciousness 1, 23–32.

Baars, B.J., Franklin, S., Ramsoy, T.Z., 2013. Global workspace dynamics: cortical 'binding and propagation' enables conscious contents. Frontiers in psychology 4.

Bailey, T., Durrant-Whyte, H., 2006. Simultaneous localization and mapping (slam): Part ii. IEEE Robotics & Automation Magazine 13, 108–117.

Barrera, A., Cáceres, A., Weitzenfeld, A., Ramirez-Amaya, V., 2011. Comparative experimental studies on spatial memory and learning in rats and robots. Journal of Intelligent & Robotic Systems 63, 361–397.

Barry, C., Lever, C., Hayman, R., Hartley, T., Burton, S., O'Keefe, J., Jeffery, K., Burgess, N., 2006. The boundary vector cell model of place cell firing and spatial memory. Reviews in the Neurosciences 17, 71–97.

Barsalou, L.W., 1999. Perceptual symbol systems. Behavioral and brain sciences 22, 577–660.

Bengio, Y., Lee, D.H., Bornschein, J., Lin, Z., 2015. Towards biologically plausible deep learning. arXiv preprint arXiv:1502.04156 .

Bird, C.M., Burgess, N., 2008. The hippocampus and memory: insights from spatial processing. Nature Reviews Neuroscience 9, 182–194.

Brandon, M.P., Bogaard, A.R., Andrews, C.M., Hasselmo, M.E., 2012. Head direction cells in the postsubiculum do not show replay of prior waking sequences during sleep. Hippocampus 22, 604–618.

Brust, C.A., Sickert, S., Simon, M., Rodner, E., Denzler, J., 2015. Convolutional patch networks with spatial prior for road detection and urban scene understanding. arXiv preprint arXiv:1502.06344 .

Burgess, N., 2008a. Spatial cognition and the brain. Annals of the New York Academy of Sciences 1124, 77–97. doi:10.1196/annals.1440.002.

Burgess, N., 2008b. Spatial cognition and the brain. Annals of the New York Academy of Sciences 1124, 77–97. doi:10.1196/annals.1440.002.

Burgess, N., Jackson, A., Hartley, T., O'keefe, J., 2000. Predictions derived from modelling the hippocampal role in navigation. Biological cybernetics 83, 301–312.

Carr, M.F., Jadhav, S.P., Frank, L.M., 2011. Hippocampal replay in the awake state: a potential substrate for memory consolidation and retrieval. Nature neuroscience 14, 147–153.

Cheng, K., Shettleworth, S.J., Huttenlocher, J., Rieser, J.J., 2007. Bayesian integration of spatial information. Psychological Bulletin 133, 625–37. URL: `http://www.ncbi.nlm.nih.gov/pubmed/17592958`, doi:10.1037/0033-2909.133.4.625.

Cheung, A., Ball, D., Milford, M., Wyeth, G., Wiles, J., 2012. Maintaining a cognitive map in darkness: the need to fuse boundary knowledge with path integration. PLoS Comput. Biol 8, e1002651.

Csizmadia, G., Muller, R.U., 2008. Storage of the distance between place cell firing fields in the strength of plastic synapses with a novel learning rule. Hippocampal Place Fields: Relevance to Learning and Memory: Relevance to Learning and Memory , 343.

Davachi, L., Mitchell, J.P., Wagner, A.D., 2003. Multiple routes to memory: distinct medial temporal lobe processes build item and source memories. Proceedings of the National Academy of Sciences 100, 2157–2162.

Delgado, H.J., Fausett, L.V., 1995. Solution of a linear system using a neural network based on iterative techniques, in: SPIE's 1995 Symposium on OE/Aerospace Sensing and Dual Use Photonics, International Society for Optics and Photonics. pp. 798–809.

Derdikman, D., Moser, E.I., 2010. A manifold of spatial maps in the brain. Trends in cognitive sciences 14, 561–569.

Diba, K., Buzsáki, G., 2007. Forward and reverse hippocampal place-cell sequences during ripples. Nature neuroscience 10, 1241–1242.

Dong, D., Franklin, S., 2015. A new action execution module for the learning intelligent distribution agent (lida): The sensory motor system. Cognitive Computation , 1–17.

Durgin, F.H., Akagi, M., Gallistel, C.R., Haiken, W., 2009. The precision of locomotor odometry in humans. Experimental brain research 193, 429–436.

Durrant-Whyte, H., Bailey, T., 2006. Simultaneous localization and mapping: part i. Robotics & Automation Magazine, IEEE 13, 99–110.

Epstein, R.A., 2008. Parahippocampal and retrosplenial contributions to human spatial navigation. Trends in cognitive sciences 12, 388–396.

Etienne, A.S., Maurer, R., Sguinot, V., 1996. Path integration in mammals and its interaction with visual landmarks. Journal of Experimental Biology 199, 201–9.

Fortin, N., 2008. Navigation and episodic-like memory in mammals. Elsevier. volume 1. pp. 385–418.

Franklin, S., Madl, T., D'Mello, S., Snaider, J., 2014. Lida: A systems-level architecture for cognition, emotion, and learning. Autonomous Mental Development, IEEE Transactions on 6, 19–41. doi:10.1109/TAMD.2013.2277589.

Franklin, S., Strain, S., Snaider, J., McCall, R., Faghihi, U., 2012. Global workspace theory, its lida model and the underlying neuroscience. Biologically Inspired Cognitive Architectures .

Freeman, W.J., 2002. The limbic action-perception cycle controlling goal-directed animal behavior. Neural Networks 3, 2249–2254.

Frese, U., 2006. A discussion of simultaneous localization and mapping. Autonomous Robots 20, 25–42.

Fuster, J.M., 2002. Physiology of executive functions: The perception-action cycle. Principles of frontal lobe function , 96–108.

Glover, A.J., Maddern, W.P., Milford, M.J., Wyeth, G.F., 2010. Fab-map+ ratslam: appearance-based slam for multiple times of day, in: Robotics and Automation (ICRA), 2010 IEEE International Conference on, IEEE. pp. 3507–3512.

Goertzel, B., Lian, R., Arel, I., de Garis, H., Chen, S., 2010. A world survey of artificial brain projects, part ii: Biologically inspired cognitive architectures. Neurocomputing 74, 30–49.

Gower, J.C., 1975. Generalized procrustes analysis. Psychometrika 40, 33–51.

Grechkin, T.Y., Nguyen, T.D., Plumert, J.M., Cremer, J.F., Kearney, J.K., 2010. How does presentation method and measurement protocol affect distance estimation in real and virtual environments? ACM Transactions on Applied Perception (TAP) 7, 26.

Griffiths, T.L., Canini, K.R., Sanborn, A.N., Navarro, D.J., 2007. Unifying rational models of categorization via the hierarchical dirichlet process, in: Proceedings of the 29th annual conference of the cognitive science society, pp. 323–328.

Hafting, T., Fyhn, M., Molden, S., Moser, M., Moser, E., 2005. Microstructure of a spatial map in the entorhinal cortex. Nature 436, 801–806.

Hartley, T., Maguire, E.A., Spiers, H.J., Burgess, N., 2003. The well-worn route and the path less traveled: distinct neural bases of route following and wayfinding in humans. Neuron 37, 877–888.

Hirtle, S., Jonides, J., 1985. Evidence of hierarchies in cognitive maps. Memory & Cognition 13, 208–217.

Jarsky, T., Roxin, A., Kath, W.L., Spruston, N., 2005. Conditional dendritic spike propagation following distal synaptic activation of hippocampal CA1 pyramidal neurons. Nature Neuroscience 8, 1667–1676. URL: http://www.ncbi.nlm.nih.gov/pubmed/16299501.

Jeffery, K.J., 2007. Self-localization and the entorhinal-hippocampal system. Current Opinion in Neurobiology 17, 684–91. URL: http://www.ncbi.nlm.nih.gov/pubmed/18249109, doi:10.1016/j.conb.2007.11.008.

Katz, Y., Kath, W.L., Spruston, N., Hasselmo, M.E., 2007. Coincidence detection of place and temporal context in a network model of spiking hippocampal neurons. PLoS Computational Biology 3, e234.

Khaligh-Razavi, S., Kriegeskorte, N., 2014. Deep supervised, but not unsupervised, models may explain it cortical representation. PLoS computational biology 10, e1003915.

Kiani, R., Esteky, H., Mirpour, K., Tanaka, K., 2007. Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. Journal of neurophysiology 97, 4296–4309.

Kim, J., Delcasso, S., Lee, I., 2011. Neural correlates of object-in-place learning in hippocampus and prefrontal cortex. The Journal of Neuroscience 31, 16991–17006.

Kjelstrup, K.B., Solstad, T., Brun, V.H., Hafting, T., Leutgeb, S., Witter, M.P., Moser, E.I., Moser, M.B., 2008. Finite scale of spatial representation in the hippocampus. Science 321, 140–143.

Knill, D.C., Pouget, A., 2004. The Bayesian brain: the role of uncertainty in neural coding and computation. Trends in Neurosciences 27, 712–9. URL: http://www.ncbi.nlm.nih.gov/pubmed/15541511, doi:10.1016/j.tins.2004.10.007.

Koch, C., Segev, I., 2000. The role of single neurons in information processing. nature neuroscience 3, 1171–1177.

Kravitz, D.J., Saleem, K.S., Baker, C.I., Mishkin, M., 2011. A new neural framework for visuospatial processing. Nature Reviews Neuroscience 12, 217–230.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.

Lever, C., Burton, S., Jeewajee, A., O Keefe, J., Burgess, N., 2009. Boundary vector cells in the subiculum of the hippocampal formation. Journal of Neuroscience 29, 9771–7.

Madl, T., Chen, K., Montaldi, D., Trappl, R., 2015. Computational cognitive models of spatial memory in navigation space: A review. Neural Networks 65, 18–43.

Madl, T., Franklin, S., Chen, K., Montaldi, D., Trappl, R., 2014. Bayesian integration of information in hippocampal place cells. PLoS ONE , e89762doi:10.1371/journal.pone.0089762.

Madl, T., Franklin, S., Chen, K., Trappl, R., 2013. Spatial working memory in the lida cognitive architecture, in: Proc. international conference on cognitive modelling.

Madl, T., Franklin, S., Chen, K., Trappl, R., Montaldi, D., submitted. Exploring the structure of spatial representations. Cognitive Processing .

Manns, J.R., Eichenbaum, H., 2009. A cognitive map for object memory in the hippocampus. Learning & Memory 16, 616–624.

McCall, R., Franklin, S., 2013. Cortical learning algorithms with predictive coding for a systems-level cognitive architecture, in: Second Annual Conference on Advances in Cognitive Systems Poster Collection, pp. 149–66.

McNamara, T.P., Hardy, J.K., Hirtle, S.C., 1989. Subjective hierarchies in spatial memory. Journal of Experimental Psychology: Learning, Memory, and Cognition 15, 211.

McNaughton, B., Barnes, C., Gerrard, J., Gothard, K., Jung, M., Knierim, J., Kudrimoti, H., Qin, Y., Skaggs, W., Suster, M., et al., 1996. Deciphering the hippocampal polyglot: the hippocampus as a path integration system.

The Journal of Experimental Biology 199, 173–185.

McNaughton, B.L., Battaglia, F.P., Jensen, O., Moser, E.I., Moser, M.B., 2006. Path integration and the neural basis of the 'cognitive map'. Nature Reviews. Neuroscience 7, 663–78. doi:10.1038/nrn1932.

Mehta, M.R., Quirk, M.C., Wilson, M.A., 2000. Experience-dependent asymmetric shape of hippocampal receptive fields. Neuron 25, 707–715.

Milford, M.J., Wiles, J., Wyeth, G.F., 2010. Solving navigational uncertainty using grid cells on robots. PLoS Computational Biology 6, e1000995–1.

Milford, M.J., Wyeth, G.F., Rasser, D., 2004. Ratslam: a hippocampal model for simultaneous localization and mapping, in: Robotics and Automation, 2004. Proceedings. ICRA'04. 2004 IEEE International Conference on, IEEE. pp. 403–408.

Mittelstaedt, M., Mittelstaedt, H., 1980. Homing by path integration in a mammal. Naturwissenschaften 67, 566–567.

Montemerlo, M., Thrun, S., Koller, D., Wegbreit, B., et al., 2002. Fastslam: A factored solution to the simultaneous localization and mapping problem, in: AAAI/IAAI, pp. 593–598.

Moser, E.I., Kropff, E., Moser, M.B., 2008. Place cells, grid cells, and the brain's spatial representation system. Annual review of neuroscience 31, 69–89. doi:10.1146/annurev.neuro.31.061307.090723.

Murgia, A., Sharkey, P.M., et al., 2009. Estimation of distances in virtual environments using size constancy. The International Journal of Virtual Reality 8, 67–74.

Nardini, M., Jones, P., Bedford, R., Braddick, O., 2008. Development of cue integration in human navigation. Current biology 18, 689–693.

O'Keefe, J., Burgess, N., 1996. Geometric determinants of the place fields of hippocampal neurons. Nature 381, 425–428.

Olson, E., Leonard, J., Teller, S., 2006. Fast iterative alignment of pose graphs with poor initial estimates, in: Robotics and Automation, 2006. ICRA 2006. Proceedings 2006 IEEE International Conference on, IEEE. pp. 2262–2269.

Plumert, J.M., Kearney, J.K., Cremer, J.F., Recker, K., 2005. Distance perception in real and virtual environments. ACM Transactions on Applied Perception (TAP) 2, 216–233.

Prasser, D., Milford, M., Wyeth, G., 2006. Outdoor simultaneous localisation and mapping using ratslam, in: Field and Service Robotics, Springer. pp. 143–154.

Quigley, M., Conley, K., Gerkey, B., Faust, J., Foote, T., Leibs, J., Wheeler, R., Ng, A.Y., 2009. Ros: an open-source robot operating system, in: ICRA Workshop on Open Source Software, p. 5.

Rusu, R.B., Maldonado, A., Beetz, M., Gerkey, B., 2007. Extending player/stage/gazebo towards cognitive robots acting in ubiquitous sensor-equipped environments, in: ICRA Workshop for Networked Robot Systems.

Samsonovich, A., McNaughton, B.L., 1997. Path integration and cognitive mapping in a continuous attractor neural network model. The Journal of neuroscience 17, 5900–5920.

Samsonovich, A.V., 2010. Toward a unified catalog of implemented cognitive architectures. BICA 221, 195–244.

Samsonovich, A.V., 2012. On a roadmap for the bica challenge. Biologically Inspired Cognitive Architectures 1, 100–107.

Sanborn, A.N., Griffiths, T.L., Navarro, D.J., 2006. A more rational model of categorization, in: Proceedings of the 28th annual conference of the cognitive science society, pp. 726–731.

Schölkopf, B., Mallot, H.A., 1995. View-based cognitive mapping and path planning. Adaptive Behavior 3, 311–348.

Snaider, J., Franklin, S., 2014. Modular composite representation. Cognitive Computation 6, 510–527.

Snaider, J., McCall, R., Franklin, S., 2011. The lida framework as a general tool for agi, in: Artificial General Intelligence. Springer, pp. 133–142.

Solstad, T., Boccara, C.N., Kropff, E., Moser, M.B., Moser, E.I., 2008. Representation of geometric borders in the entorhinal cortex. Science 322, 1865–8. doi:10.1126/science.1166466.

Stork, D.G., 1989. Is backpropagation biologically plausible?, in: Neural Networks, 1989. IJCNN., International Joint Conference on, IEEE. pp. 241–246.

Strösslin, T., Sheynikhovich, D., Chavarriaga, R., Gerstner, W., 2005. Robust self-localisation and navigation based on hippocampal place cells. Neural networks 18, 1125–1140.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2014. Going deeper with convolutions. arXiv preprint arXiv:1409.4842 .

Takahashi, H., Magee, J.C., 2009. Pathway interactions and synaptic plasticity in the dendritic tuft regions of CA1 pyramidal neurons. Neuron 62, 102–111. URL: `http://www.ncbi.nlm.nih.gov/pubmed/19376070`.

Taube, J.S., 2007. The head direction signal: origins and sensory-motor integration. Annual Review of Neuroscience 30, 181–207.

Thrun, S., Leonard, J.J., 2008. Simultaneous localization and mapping. Springer handbook of robotics , 871–889.

Vogeley, K., May, M., Ritzl, A., Falkai, P., Zilles, K., Fink, G.R., 2004. Neural correlates of first-person perspective as one constituent of human self-consciousness. Journal of cognitive neuroscience 16, 817–827.

Waller, D., 1999. Factors affecting the perception of interobject distances in virtual environments. Presence: Teleoperators and Virtual Environments 8, 657–670.

Williams, B., Cummins, M., Neira, J., Newman, P., Reid, I., Tardós, J., 2009. A comparison of loop closing techniques in monocular slam. Robotics and Autonomous Systems 57, 1188–1197.

Wilson, D.I., Langston, R.F., Schlesiger, M.I., Wagner, M., Watanabe, S., Ainge, J.A., 2013. Lateral entorhinal cortex is critical for novel object-context recognition. Hippocampus 23, 352–366.

Winters, B.D., Bussey, T.J., 2005. Transient inactivation of perirhinal cortex disrupts encoding, retrieval, and consolidation of object recognition memory. The Journal of neuroscience 25, 52–61.

Yamins, D.L., Hong, H., Cadieu, C., DiCarlo, J.J., 2013. Hierarchical modular optimization of convolutional networks achieves representations similar to macaque it and human ventral stream, in: Advances in Neural Information Processing Systems, pp. 3093–3101.

Zaehle, T., Jordan, K., Wüstenberg, T., Baudewig, J., Dechent, P., Mast, F.W., 2007. The neural basis of the egocentric and allocentric spatial frame of reference. Brain research 1137, 92–103.

Chapter 7

Discussion

In Chapters 4-6 above, we have argued for the necessity of probabilistic mechanisms in spatial cognition when faced with a complex, uncertain environment perceived through noisy sensors. Although by no means conclusive, we have presented evidence that

1. hippocampal place cells represent spatial uncertainty,
2. they can perform approximate Bayesian inference,
3. the representations by recently active place cells can be corrected near-optimally through reverse replay when revisiting a place, and
4. spatial representation structure arises from clustering under a metric defined over features including distance and visual and functional similarity.

We have also integrated these suggested probabilistic mechanisms into LIDA, and embodied the resulting cognitive architecture in a robotic simulation. In this Chapter, we discuss the abilities, shortcomings, and missing functionalities of our models, and their consistency with related empirical findings, from a cognitive science perspective.

7.1 Other mechanisms and representations involved in spatial navigation

Tables 7.1 and 7.2 summarize the processes and representations involved in spatial navigation in biological cognition. The first columns provide overviews of these mechanisms and representations, based on Figure 1 in (Wolbers & Hegarty, 2010). The second column indicates the corresponding mechanism in our final LIDA-based model, as

described in Chapter 6. The rightmost column highlights some major elements missing from the models presented here but required for spatial navigation.

↓ Mechanism	In our model	Not implemented
Spatial computations		
Space perception	Limited (depth from stereo disparity*)	Estimating size, shape, movement, orientation, ...
Self-motion perception	Surrogate: odometry*	Motor efference, proprioceptive & vestibular senses
Translation btw. ego- and allocentric reference frames	Limited: Perspective projection via homography*	Plausible translation mechanism
Computing directions and distances to unseen goals	Route plan SBC (following gradient on a hierarchical grid)	Explicit direction estimation, systematic errors in estimation
Imagining shifts in spatial perspective	-	Sensory imagery
Executive processes		
Novelty detection	-	Perceptual recognition of known or novel places
Selection and maintenance of navigational goals	Attention codelets* & global broadcast* in LIDA's cognitive cycle	Reward representations, reinforcement learning
Route planning or selection	Route plan SBC (following gradient on a hierarchical grid)	Expectation violation / confirmation monitoring, re-planning, homing...
Uncertainty/Conflict resolution	Partial: Bayesian integration	Conflicting cues, cues other than odometry & estimated distance
Resetting mechanisms	Partial: maximum likelihood correction	Kidnapped robot problem

Table 7.1: Cognitive mechanisms involved in spatial navigation, based on (Wolbers & Hegarty, 2010). *: an ability of our model making use of existing implementations (in the LIDA cognitive architecture or the Robot Operating System).

↓ Representation	In our model	Not implemented
Online representations		
Self-position and orientation	'Self' PAM node	-
Egocentric self-to-object directions and distances	Limited (depth from stereo disparity*)	Egocentric vectors (e.g. 'reach vectors' in area 5a)
Allocentric object-to-object directions and distances	Indirect (on map representation, but not perceptually)	Allocentric visuo-spatial representations
Route progression	'Route' PAM nodes	Expectations
Navigation goals	'Goal' PAM nodes	Rewards
Offline representations		
Memories of local views and places	Partial (in pre-conscious working memory, not yet in long-term memory)	Long-term memory representations
Enduring, hierarchical representations of an environment (ego-/allocentric)	Hierarchical maps consisting of 'place nodes'	Hierarchical egocentric representations
Networks of habitual routes	Context-action-result chains in Procedural Memory*	-

Table 7.2: Representations involved in spatial navigation, based on (Wolbers & Hegarty, 2010)

7.2 Limitations and shortcomings

In addition to mechanisms and representations playing an important role in spatial navigation but not yet implemented in our model (Tables 7.1 and 7.2), there are several shortcomings of our models, which we outline in this Section. They can roughly be grouped into three categories: computational shortcomings, psychological implausibilities, and neural implausibilities.

7.2.1 Computational shortcomings

We have pointed out in Chapters 1 and 2 that the goal of this work was not to optimize for performance (but rather computational cognitive modelling), and that these problems can be solved more optimally and accurately, given enough computational resources. Accuracy and performance of spatial representations are the goals of Simultaneous Localization and Mapping (SLAM) in mobile robotics (Thrun & Leonard, 2008).

State of the art solutions to the SLAM problem can infer robot and landmark locations down to a few centimetres accuracy or better, but usually require $5-25\%$ of the processing power of a current Intel Core i7-3630QM CPU to do so (Santos et al., 2013), even when just mapping a small room, which amounts to $4-20$ billion floating point operations per second[1]. Achieving the same in large-scale outdoor environments would require even more computational resources.

Figure 7.1 shows the structure of modern end-to-end SLAM systems (Wang, 2015), such as e.g. (Newman et al., 2011). Components depending on the specific sensors and actuators ('front-end') are usually separated from the sensor-independent optimization part ('back-end'). In our final model described in Chapter 6, the 'front-end' roughly corresponds to the functionality of the Bayesian localization SBC, and the 'back-end' to that of the Map correction SBC. Both functionally correspond to hippocampal place cells, with the former mechanism partially implemented by coincidence detection, and the latter through reverse replay.

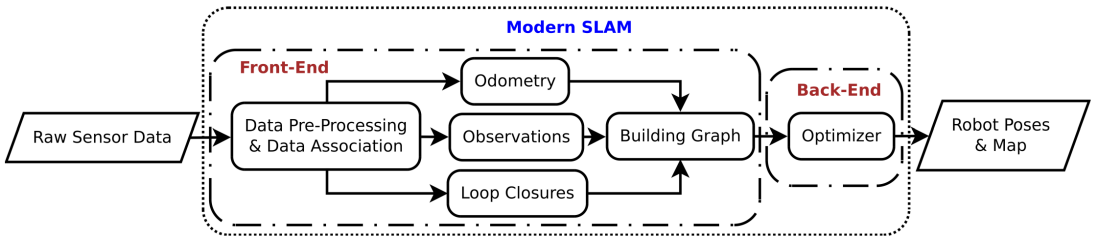


Figure 7.1: Components of a modern end-to-end SLAM system. From (Wang, 2015)

The two main computational shortcomings compared to modern SLAM include 1) not explicitly modelling rotations (thus avoiding non-linearity caused by robots which can turn), and 2) not explicitly optimizing landmark constraints (only path integration and loop closure constraints). These cause inferior localization and mapping accuracy compared to modern SLAM. However, they have allowed us to map Bayesian mechanisms to well-known neural correlates and mechanisms, and to implement simple models successfully replicating behaviour data, while still retaining the ability to tackle the uncertainty and noise problem in a realistic robotic simulation.

Although brains may well be capable of the processing power required by a SLAM

[1]Based on Intel i7 specifications, retrieved from `http://download.intel.com/support/processors/corei7/sb/core_i7-3600_m.pdf`

system, it is unlikely that they work the way modern SLAM solutions do (performing thousands of linear algebra operations serially) (Thrun & Leonard, 2008). Furthermore, human long-term memories are far from being as accurate as these SLAM systems, as shown e.g. in Figures 6.7 and 6.8 in Chapter 6, or by research regarding sketch maps, e.g. (Rovine & Weisman, 1989; Wang & Schwering, 2009). Nevertheless, there is value in looking at information processing in brains through the lens of normative models, of mathematical formulations of the problem to be solved; and of their implementability in brains and minds.

7.2.2 Psychological implausibilities

Apart from implementation details (in brains and in LIDA), on Marr's (1976) algorithmic level, three major mechanisms were suggested in this thesis: 1) a cue integration mechanism for localization, 2) correction of cognitive maps when re-visiting places, and 3) cognitive map structuring through clustering. Despite of their ability to fit behaviour data as described in Chapters 4-6, there are some psychological findings which are inconsistent with these mechanisms.

First, our models have focused on adult cognition, and have ignored developmental findings. Visual spatial integration progressively improves in children between 5 and 14 years of age (Kovacs et al., 1999). Spatial cue integration, while close to the Bayesian optimum in adults, seems to require a long developmental process; and children do not seem to integrate spatial cues, instead switching between exclusively using path integration or landmark information from trial to trial (Nardini et al., 2008). It is difficult to model this behaviour in our Bayesian framework.

In terms of adult spatial cognition, there are shortcomings in how landmarks are recognized. In the current model, any objects recognized by the CNN briefly described in 6 constitutes a landmark. However, in human (and animal) cognition, landmarks have to be reliable, salient, stable (unmoving), and possibly distal (Lew, 2011). These criteria defining landmarks for biological spatial cognition are not accounted for in the model. Neither are cues in the form of landmark arrays (e.g. humans use the natural axes of regular arrays of objects as a reference frame) (Lew, 2011; Burgess, 2006).

Phenomena observed in environments with competing cues (e.g. landmarks), where the information from the cues is not integrated, are also difficult to model in our probabilistic framework. Examples include 'overshadowing' (where the effect of a cue on an animal's behaviour may be reduced or eliminated when another, more salient cue is introduced) and 'blocking' (where a second cue is added after an animal has

been trained with the first, but the animal cannot use the second cue without the first) (Chamizo, 2003). Some evidence of landmark overshadowing and blocking in humans exists, e.g. (Spetch, 1995; Prados, 2011), and it has been argued that unlike the role of boundaries, associative reinforcement (and not a map-like representation) may be a better explanation for landmark learning (Doeller & Burgess, 2008).

Navigation based on two complementary systems running in parallel (a cognitive mapping system using the described mechanisms, and a reward-based associative learning system based on LIDA's procedural memory) is conceptually consistent with blocking and overshadowing, and may be able to explain these findings. We have not implemented this computationally, however; and the extent of cooperation / competition between these systems is not yet clear, even on a theoretical level (Lew, 2011; Cheng et al., 2013).

In addition to the role of landmarks, a 'geometric module' for navigation has been proposed, originally to explain errors which would have been avoidable if perceptual as opposed to geometric cues had been used (such as rats learning there is food in the corner of a rectangular environment, but often searching in the diagonally opposite corner of the environment, which was geometrically - but not perceptually - equivalent) (Cheng, 1986). Similar geometry-based behaviour has been observed in young children, e.g. by Huttenlocher et al. (1999) (see also (Cheng et al., 2013)). Recent findings cast the existence of a dedicated geometric module for orientation and navigation in doubt (Cheng, 2008). Nevertheless, empirical observations of such errors (which are consistent with geometry-based orientation, but could be avoided by making use of perceptual features/landmarks) are inconsistent with our model, which does not make such errors.

Other types of systematic errors in spatial representations have been pointed out in the literature which our model does not account for in its current form. Distortions result from the hierarchical organization in cognitive maps (Tversky, 1992; Hirtle & Jonides, 1985) - which, however, could easily be incorporated into the model, given that it already learns these hierarchies (all that is required is implementing an error function/mechanism). However, there are also systematic distortions of spatial representations which are not easily accounted for in this framework. They include effects of perspective (where participants are asked to imagine themselves when asked to estimate spatial relations), of cognitive reference points (distance judgements made from landmark A to building B usually differ from those made from building B to landmark A), and of detours or barriers (the length of circuitous routes is usually overestimated)

- see (Tversky, 1992, 2003). Differences in viewpoints used when learning spatial representations and when having to use them also cause systematic errors (e.g. (Shelton & McNamara, 2001, 2004; Burgess, 2006)) which have been neglected by the current models.

Finally, the current model, when forced to explore very large regions without being allowed to ever revisit known places, can incur catastrophically large errors to its learned representations, making the learned map largely useless (we know of no such effect observed in humans). It is likely that in very large scale environments, humans make use of several parallel mechanisms including spatial reasoning, as well as of prior knowledge of the structure of the environment (e.g. the usual shapes of roads), none of which have been included in the model.

We note that to our knowledge, no current computational cognitive model of spatial memory achieves full consistency with every empirical finding, while being capable of running in realistic environments at the same time (see review in Chapter 3). We have argued that our approach is a step in the direction of such models, which can be the case even if it does not support modelling some known aspects of spatial cognition. As long as the basic premises hold (that brains can represent uncertainty, and can perform approximate Bayesian inference), and if the shortcomings can be corrected in future models in a cognitively plausible fashion, the probabilistic approach to spatial cognition remains viable.

7.2.3 Neural implausibilities

In terms of consistency with neuroscientific findings, we have to distinguish between the final computational cognitive model based on the LIDA cognitive architecture (Chapter 6), and the suggested neural mechanisms regarding uncertainty representation and error correction in the hippocampus. We omit discussing the neural plausibility of the map structuring model introduced in Chapter 3, since we have not described any neural implementation of this mechanism, and have only validated it behaviourally (but see e.g. (Shi & Griffiths, 2009) or (Sanborn, 2015) for possible neural implementations of hierarchical Bayesian models, to which the DP-GMM belongs). It is to our knowledge the first model able to predict spatial representation structure on the individual level; and developing a biologically plausible implementation in addition to a normative and algorithmic model would have exceeded the time available for this PhD.

Regarding the final model (Chapter 6), LIDA intends to be a model of minds, not brains (it is a model on Marr's algorithmic and not implementation level) - see

(Franklin et al., 2012, 2014) for discussions of the relationship between LIDA and the underlying neuroscience. Nevertheless, we briefly point out a few mechanisms of the model in Chapter 6 (LIDA extended by the described probabilistic spatial mechanisms and embodied on a robot) which do not directly correspond to known neural processes.

The first salient difference is the visual recognition system, for which we used existing implementations to make this work tractable within one PhD. Specifically, we used convolutional neural networks for recognizing objects (Szegedy et al., 2014) and roads (Brust et al., 2015), which have been designed for maximizing recognition performance, not for neural plausibility. Curiously, they do seem to learn representations that are very similar to those recorded in human and primate inferior temporal cortex (Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2013). But their conventional training algorithms are not implementable in biological neurons (Stork, 1989; Bengio et al., 2015). Developing a plausible recognition system would have exceeded the scope of this PhD. The same is true for motor control, for which we used existing drivers of the Robot Operating System[2] (which are by no means brain-like).

In terms of the spatial extensions to LIDA, the biggest discrepancy is the regular grid formed by the 'place nodes' (Chapter 6). Place cells do not seem to map the surface of an environment in any systematic fashion (O'Keefe et al., 1998). It would be more accurate to think of 'place nodes' as combining several underlying spatially relevant cell types, including entorhinal grid cells, which do form regular grids (although triangular and not rectangular) (Moser et al., 2008). Grid cells also facilitate estimating directions and distances (Bush et al., 2015). However, the simple route planning strategy (based on spreading activation on hierarchical grids of place nodes) is not a faithful model of navigation in the hippocampal entorhinal complex, as it heavily relies on a regular structure and on specific link weights depending on distances and obstacles. Bush et al. (2015) reviews four more biologically plausible network models on Marr's implementation level. However, LIDA is concerned with the algorithmic level - and there is published behavioural evidence for such a mechanism (Mueller et al., 2013). We have also succeeded in replicating two multi-goal route planning datasets using our simple model (in virtual as well as real environments - see Appendix C), which substantiates its cognitive plausibility.

On the other hand, the plausibility of the probabilistic framework for cognitive modelling does require, at the very least, the possibility of neurally implementing

[2]See `http://wiki.ros.org/pid` and `http://gazebosim.org/tutorials?tut=drcsim_fakewalking&cat=drcsim`

Bayesian inference. To show evidence of this possibility, we have compared the firing of hippocampal place cells to predictions of a Bayesian model, and have suggested they might be able to represent uncertainty and perform approximately optimal inference (see Chapter 4). These are hypotheses on the neuronal level. As such, they can be compared to neuroscientific findings - and they do seem to be inconsistent with some, as summarized below.

First, humans with hippocampal lesions, although spatially impaired, do seem to be capable of spatial navigation. For example, (Teng & Squire, 1999) report a patient with damaged medial temporal areas who was able to describe routes, detours, and directions between landmarks in an environment he has learned early, before the damage. The authors suggest that the role of the hippocampus is time-limited, mostly concerning consolidation, and that long-term spatial memories are available after consolidation even with a lesioned hippocampus. Similar observations of largely unimpaired topographical abilities in patients with hippocampal damage were found by (Rosenbaum et al., 2000, 2005); although these patients did show some types of impairments (few recalled landmarks on sketch maps, no detailed geographical knowledge, impaired landmark recognition).

A later study by Maguire et al. (2006) reinforced the implication that although accessing long-established spatial memories is still possible with a damaged hippocampus, topographical knowledge of landmarks and of the relationships between them is impaired. Naturally, the ability to learn new spatial representations is also heavily impaired. Nevertheless, some functionalities requiring allocentric representations seem to be available to patients with hippocampal lesions, which is problematic for the 'cognitive map' hypothesis in general, as well as for our model.

Second, the firing fields of place cells do not behave like unique, one-to-one representations of location. Some place cells (a minority) have more than one firing field (Burke et al., 2011). Although usually there are geometric similarities between the locations of these firing fields (Barry et al., 2006), there are also cases where there seem to be no systematic commonalities (Park et al., 2011) between them (e.g. similar distances to surroundings) as would be predicted by a model using these firing fields as probability distributions. Place fields are also not always regular and elliptic, as prescribed by the simplest Gaussian model in Chapter 4 (although this is not an issue for the particle filter-based formulation in Chapter 6, which can represent multimodal distributions).

Furthermore, it is not always the case that place fields close to boundaries have to

be smaller than those further away, as would be predicted if they solely represented uncertainty. For example, firing fields of cells in dorsal hippocampus are generally smaller than those of cells in more ventral areas (Kjelstrup et al., 2008). There are also some other phenomena observed in recordings from place cells of behaving animals which do not easily fit into a probabilistic model. These include remapping (Colgin et al., 2008) and theta phase precession (Skaggs & McNaughton, 1996).

However, these inconsistencies do not falsify the possibility of an approximate Bayesian inference mechanism operating in the hippocampus in parallel to several other mechanisms not accounted for (and in some cases inconsistent with) such a mechanism. Brains exhibit a high degree of redundancy, and there is no reason to assume that one cell type only performs one function.

Over-reliance on only a single or few place cells inconsistent with the statistical optimum could destroy the models functionality. But a larger ensemble of place cells, a majority of which do represent location estimates and their associated approximate uncertainty, can still facilitate approximately optimal localization if the contradicting information in the ensemble (representing other things, such as an episodic memories (Tulving & Markowitsch, 1998)) is a minority. The approximate Bayesian place cell hypothesis could be falsified if the number of place cells used for localization, and having firing fields inconsistent with Bayesian uncertainty predictions, could be shown to be a majority. This does not seem to be the case in the recordings and environments investigated here (see Chapter 4).

We can further support the claim of multiple parallel hippocampal mechanisms, one of which may be approximate Bayesian inference, using three observations. First, the reasonably good fit of Bayesian predictions with empirical place field sizes reported in Chapter 4 would be extremely unlikely to occur by chance, given that hundreds of place fields were included in the comparison. Second, our particle filter localization model is largely robust to artificially increasing or decreasing the variance of the samples at some places[3], which is a rudimentary way of simulating some place fields having a different size than prescribed by a Bayesian model. Third, the uncertainties predicted by a sampling-based localization model can also successfully explain the frequency distribution of place field sizes, even when corrupted by location-unrelated samples (see comparison in Appendix D).

[3]In fact, adding random samples, independently from the Bayesian prediction, was one of the early methods used in robotics to combat 'particle depletion' and to increase the chances of the robot being able to recover its correct location in particle filter-based SLAM (Thrun et al., 2005).

Finally, in its current formulation, our model depends on approximate multiplication of incoming signals (e.g. from cells with border-related firing). We have shown that coincidence detection can implement this multiplication (Chapter 4), and have argued that the biophysical parameters of CA1 place cells seem to be in the right range to facilitate multiplication up to an estimated 5% error. However, a number of influential theories of place cell firing propose thresholded summation instead of multiplication in place cells. Notable and empirically well-supported examples include grid field summation models (Solstad et al., 2006), and the Boundary Vector Cell (BVC) model of place cell firing (Hartley et al., 2000; Barry et al., 2006). The former does not solve the accumulating path integration error problem (Etienne et al., 1996), and is thus not suitable for real-world navigation in its original form.

The BVC model serves a different purpose than our model: it is an explanatory model relying on a large number of parameters to achieve very good fit to a dataset (e.g. hundreds of parameters for the data in Figure 2 of Chapter 4 - several for each place cell), whereas our model is normative, arising from a single computational principle and requiring very few parameters (only path integration and measurement accuracies), at the cost of less-than-perfect fit to the data. In terms of implementation, the key difference is that the BVC model suggests place cell firing to depend on a thresholded sum of BVC firing fields; whereas our model proposes approximate multiplication.

Any function can be approximated by summing a sufficient number of parametrized Gaussians (Parzen, 1962), so it is unsurprising that the BVC model can fit any firing field; but it is less obvious that it can also successfully predict the responses of these fields to topographic changes in the environment (Barry et al., 2006). Our model can frequently make similar predictions with much less parameters (Chapter 4), but there are a number of empirically observed place field responses to such changes which are inconsistent with our model. Specifically, there is a small number of place cell firing fields which become bi-modal in larger environments (O'Keefe & Burgess, 1996). This is easy to explain using summation of two Gaussians anchored to opposite walls in the environment, but contradicts a multiplicative, strictly Bayesian framework.

It is of course possible for a subset of place cells to have a low membrane time and implement multiplication by coincidence detection, as suggested in Chapter 4, and for another subset with a higher membrane time to implement summation as suggested by the BVC model. In this way, the models could be complementary (with our model treating the minority of secondary firing fields as correctable noise). There is indeed more than 40% variation across place cells membrane time constants, suggested to lie

around $18.6 + / - 8.1ms$ (Szilagyi et al., 1996), with other observations ranging from $16.6ms$ (Zemankovics et al., 2010) to $23.2ms$ or $23.6ms$ (Johnston, 1981).

We have shown that these time constants facilitate calculating Bayesian posteriors using approximate multiplication, with just 5% (at $16.6ms$) to 17% (at $23.6ms$) error compared to the mathematically correct posterior in a leaky integrate-and-fire spiking neuron model of place cells (Figure 7 in Chapter 4). Of course, this does not prove that real place cells multiply their inputs, but it shows that they could. This is backed by some empirical evidence, e.g. the observation that CA1 cells only exhibit stable firing when synchronously receiving spikes from perforant path and Schaffer-collateral synapses, within $5 - 10ms$ (Jarsky et al., 2005). This empirically observed requirement of synchrony supports our coincidence detection model, and is inconsistent with summation.

Yet another possibility is that the calculation of an approximate location posterior is performed in a brain area other than the hippocampus, such as the entorhinal cortex, and that place cell firing simply constitutes the output. A similar suggestion has recently been made by Hardcastle et al. (2015), who suggest error correction occurs in grid cells based on border cell input.

Based on the near impossibility of the strong correlations between Bayesian predictions and recorded firing field sizes arising merely by random chance across hundreds of place cells (Chapter 4), and on the mathematical necessity of a correction mechanism for accumulating location estimate errors, we have argued for a probabilistic framework to model localization in biological cognition. We think this view has merit despite of some empirical phenomena inconsistent with it. More future experimental work will be necessary to isolate the exact computational mechanism implemented by place cells, to distinguish to what extent some or all of them may sum or multiply their inputs, and to better understand the role of multi-field place cells in spatial navigation.

Chapter 8

Conclusion

Humans live and act in a world they can only partially observe through imperfect sensors and process with an inherently noisy information processing system. In mathematics, probability theory has provided a framework for the representation and manipulation of uncertainty (Jaynes, 1996). In this thesis, we have argued for the necessity of such a framework within the field of computational cognitive modelling as well. We have modelled and interpreted neuroscientific evidence in a probabilistic framework, providing one of the first examples of Bayesian inference on a single-neuron level, in order to provide the foundation of this argument (Chapter 4).

Simply using existing algorithmic solutions of probabilistic localization, mapping, and clustering does not yield viable models of cognition, since these differ from biological cognitive processes in behaviour, computational requirements, and available information. However, most existing cognitive models of spatial memory, while plausibly modelling cognition, are unable to deal with sensory noise and uncertainty. We have provided a detailed review and comparison of such models in Chapter 3, and have suggested the ability to function in realistic environments as one of the main gaps in the literature.

In order to take a first step towards filling this gap, we have proposed probabilistic computational models on Marr's (1976) algorithmic level for the following mechanisms:

- self-localization ('*where am I?*'),
- object localization ('*where is this object?*'),
- map correction after revisiting a place ('*I've been here before - now how do I fix my map?*'),

- multi-goal route planning (‘*how do I get to these places?*’), and
- map structuring (‘*which map does this object belong to?*’).

Although these problems, with the exception of the last, are well-known in robotics, we have provided the - to our knowledge - first computational cognitive models which 1) are implementable in brains, 2) can reproduce behaviour data, 3) are part of a cognitive architecture, integrated with other cognitive processes, and 4) are able to function in realistic environments with noise and uncertainty (in a robotic simulation providing the exact same interfaces as a real robot (Rusu et al., 2007)).

We have also shown, for the first time since the discovery of hierarchical structure in human spatial representations (Hirtle & Jonides, 1985), that such structures are predictable based on geospatial, perceptual, and functional properties of the environment. We have provided evidence that Bayesian nonparametric clustering under a subject-specific distance metric accounts for a large majority of buildings belonging together in participants’ established spatial memories.

Our models extend the ‘Bayesian brain’ (Knill & Pouget, 2004) and ‘Bayesian cognition’ (Chater et al., 2010) paradigms one step towards navigation-space cognitive representations and processes. We hope they will encourage further research on coping with the challenges posed by the real world in computational cognitive models of spatial memory.

8.1 Future Work

The work done during this PhD paves the way for several new directions for computational models of brains and minds. The first and most straightforward extension would be to implement the proposed mechanisms as a biological neural network, in order to make their predictions more tangible and directly comparable with neuroscience data. Apart from several minor implementation details, this would require designing a neural model of how hippocampal reverse replay (the suggested mechanism for map correction) could shift place cell firing fields in the correct direction. Phase resetting may be a plausible mechanism for shifting firing fields, but its implementation is unclear, as is the propagation of the discrepancy between the remembered and revisited location estimate when performing a loop.

The proposed localization and mapping mechanisms could also be made significantly more accurate, by including orientation information in the gradient descent-based map correction mechanism. This would make the equations in Chapter 6 non-linear, and their biologically plausible solution a lot more difficult. Robotics solutions prescribe geometrical tricks such as the use of rotation matrices to deal with orientation information (Olson et al., 2006). It would be interesting to investigate whether there is reason to believe that brains are able to do something similar. Some evidence for angular information directly encoded in hippocampal representations has been found recently. For example, Huxter et al. (2008) have recently succeeded in decoding both position and heading direction from just two place cell spikes. Of course, the question of how this direction information can be utilized and corrected is still unanswered, and difficult to tackle.

Noise and uncertainty affect not only navigation-space representations, but also the space of and around the body. It is likely that evidence for statistically near-optimal integration of information can be found for tasks such as reaching, and that they can be modelled in a probabilistic framework similar to the one presented in this thesis. The strong behavioural evidence for Bayesian cue integration of haptic and visual modalities (Ernst & Banks, 2002) has been one of the key findings precipitating the 'Bayesian brain' hypothesis (Knill & Pouget, 2004), and it is likely that future work can produce models explaining such observations not only on the behavioural but also on the neural level.

Other interesting avenues of research are opened up by the evidence that human spatial representation structures are predictable (Chapter 5). We have only modelled a simple two-layer structure, which can be extended to a full hierarchy (e.g. using nested Bayesian nonparametric models (Blei et al., 2010)), or to allow transferring learned spatial information to new environments (e.g. using hierarchical Dirichlet processes (Teh et al., 2006)).

Outside of spatial memory research, our results open up the possibility to facilitate human-robot interaction by designing new robotic representations corresponding to human-like spatial concepts (we have shown that even without a subject-specific model, a general model can predict whether or not objects belong together in people's spatial representations in 3 out of 4 cases - see Table 2 in Chapter 5). Our proposed model of human spatial representation structure could also inspire work in geographical information science (e.g. new ways of presenting spatial information in a more easily comprehensible and memorable fashion).

Bibliography

Allen, K., Rawlins, J. N. P., Bannerman, D. M., & Csicsvari, J. (2012). Hippocampal place cells can encode multiple trial-dependent features through rate remapping. *The Journal of Neuroscience*, *32*, 14752–14766.

Baghshah, M. S., & Shouraki, S. B. (2010). Kernel-based metric learning for semi-supervised clustering. *Neurocomputing*, *73*, 1352–1361.

Bailey, T., & Durrant-Whyte, H. (2006). Simultaneous localization and mapping (slam): Part ii. *IEEE Robotics & Automation Magazine*, *13*, 108–117.

Barber, M. J., Clark, J., & Anderson, C. H. (2003). Neural representation of probabilistic information. *Neural Computation*, *15*, 1843–1864.

Barbieri, R., Quirk, M. C., Frank, L. M., Wilson, M. A., & Brown, E. N. (2001). Construction and analysis of non-poisson stimulus-response models of neural spiking activity. *Journal of Neuroscience Methods*, *105*, 25–37.

Barry, C., Lever, C., Hayman, R., Hartley, T., Burton, S., O'Keefe, J., Jeffery, K., & Burgess, N. (2006). The boundary vector cell model of place cell firing and spatial memory. *Reviews in the Neurosciences*, *17*, 71.

Bengio, Y., Lee, D.-H., Bornschein, J., & Lin, Z. (2015). Towards biologically plausible deep learning. *arXiv preprint arXiv:1502.04156*, .

Bensmail, H., & Celeux, G. (1996). Regularized gaussian discriminant analysis through eigenvalue decomposition. *Journal of the American Statistical Association*, *91*, 1743–1748.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Blei, D. M., Griffiths, T. L., & Jordan, M. I. (2010). The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, *57*, 7.

Blei, D. M., Jordan, M. I. et al. (2006). Variational inference for dirichlet process mixtures. *Bayesian Analysis*, *1*, 121–143.

Borg, I., & Groenen, P. J. (2005). *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media.

Bousquet, O., Balakrishnan, K., & Honavar, V. (1997). Is the hippocampus a kalman filter? In *Proceedings of the Pacific Symposium on Biocomputing* (pp. 655–666).

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Brette, R. (2012). Computing with neural synchrony. *PLoS Computational Biology*, *8*, e1002561. doi:10.1371/journal.pcbi.1002561.

Brust, C.-A., Sickert, S., Simon, M., Rodner, E., & Denzler, J. (2015). Convolutional patch networks with spatial prior for road detection and urban scene understanding. *arXiv preprint arXiv:1502.06344*, .

Burgess, N. (2006). Spatial memory: how egocentric and allocentric combine. *Trends in Cognitive Sciences*, *10*, 551–557.

Burgess, N. (2014). The 2014 nobel prize in physiology or medicine: A spatial model for cognitive neuroscience. *Neuron*, *84*, 1120–1125.

Burke, S. N., Maurer, A. P., Nematollahi, S., Uprety, A. R., Wallace, J. L., & Barnes, C. A. (2011). The influence of objects on place field expression and size in distal hippocampal CA1. *Hippocampus*, *21*, 783–801. doi:10.1002/hipo.20929.

Bush, D., Barry, C., Manson, D., & Burgess, N. (2015). Using grid cells for navigation. *Neuron*, *87*, 507–520.

Büsing, L., Bill, J., Nessler, B., & Maass, W. (2011). Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons. *PLoS Computational Biology*, *7*, e1002211.

Calabrese, E., Johnson, G. A., & Watson, C. (2013). An ontology-based segmentation scheme for tracking postnatal changes in the developing rodent brain with mri. *NeuroImage*, *67*, 375–384.

Canini, K. R., Shashkov, M. M., & Griffiths, T. L. (2010). Modeling transfer learning in human categorization with the hierarchical dirichlet process. In *International Conference on Machine Learning* (pp. 151–158).

Carr, M. F., Jadhav, S. P., & Frank, L. M. (2011). Hippocampal replay in the awake state: a potential substrate for memory consolidation and retrieval. *Nature Neuroscience*, *14*, 147–153.

Chamizo, V. (2003). Acquisition of knowledge about spatial location: Assessing the generality of the mechanism of learning. *The Quarterly Journal of Experimental Psychology: Section B*, *56*, 102–113.

Chater, N., Oaksford, M., Hahn, U., & Heit, E. (2010). Bayesian models of cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*, 811–823.

Cheng, K. (1986). A purely geometric module in the rat's spatial representation. *Cognition*, *23*, 149–178.

Cheng, K. (2008). Whither geometry? troubles of the geometric module. *Trends in Cognitive Sciences*, *12*, 355–361.

Cheng, K., Huttenlocher, J., & Newcombe, N. S. (2013). 25 years of research on the use of geometry in spatial reorientation: a current theoretical perspective. *Psychonomic Bulletin & Review*, *20*, 1033–1054.

Cheng, K., Shettleworth, S. J., Huttenlocher, J., & Rieser, J. J. (2007). Bayesian integration of spatial information. *Psychological Bulletin*, *133*, 625.

Cheung, A., Ball, D., Milford, M., Wyeth, G., & Wiles, J. (2012). Maintaining a cognitive map in darkness: the need to fuse boundary knowledge with path integration. *PLoS Computational Biology*, *8*, e1002651.

Chitta, R., Jin, R., Havens, T. C., & Jain, A. K. (2011). Approximate kernel k-means: Solution to large scale kernel clustering. In *Proceedings of the 17th International Conference on Knowledge Discovery and Data Mining* (pp. 895–903). ACM.

Cohen, G. (2000). Hierarchical models in cognition: Do they have psychological reality? *European Journal of Cognitive Psychology*, *12*, 1–36.

Colgin, L. L., Moser, E. I., & Moser, M.-B. (2008). Understanding memory through hippocampal remapping. *Trends in Neurosciences*, *31*, 469–477.

Derdikman, D., & Moser, E. I. (2010). A manifold of spatial maps in the brain. *Trends in Cognitive Sciences*, *14*, 561–569.

Deshmukh, S. S., & Knierim, J. J. (2013). Influence of local objects on hippocampal representations: landmark vectors and memory. *Hippocampus*, *23*, 253–267.

Doeller, C. F., & Burgess, N. (2008). Distinct error-correcting and incidental learning of location relative to landmarks and boundaries. *Proceedings of the National Academy of Sciences*, *105*, 5909–5914.

Doucet, A., Godsill, S., & Andrieu, C. (2000). On sequential monte carlo sampling methods for bayesian filtering. *Statistics and Computing*, *10*, 197–208.

Durrant-Whyte, H., & Bailey, T. (2006). Simultaneous localization and mapping: part i. *Robotics & Automation Magazine, IEEE*, *13*, 99–110.

Ernst, M. O. (2006). A bayesian view on multimodal cue integration. *Human Body Perception from the Inside Out*, (pp. 105–131).

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*, 429–433.

Etienne, A. S., Maurer, R., & Séguinot, V. (1996). Path integration in mammals and its interaction with visual landmarks. *The Journal of Experimental Biology*, *199*, 201–209.

Fenton, A. A., & Muller, R. U. (1998). Place cell discharge is extremely variable during individual passes of the rat through the firing field. *Proceedings of the National Academy of Sciences*, *95*, 3182–3187.

Ferbinteanu, J., & Shapiro, M. L. (2003). Prospective and retrospective memory coding in the hippocampus. *Neuron*, *40*, 1227–1239.

Fiser, J., Berkes, P., Orbán, G., & Lengyel, M. (2010). Statistically optimal perception and learning: from behavior to neural representations. *Trends in Cognitive Sciences*, *14*, 119–130.

Fox, C., & Prescott, T. (2010). Hippocampus as unitary coherent particle filter. In *The 2010 International Joint Conference on Neural Networks* (pp. 1–8). IEEE.

Franklin, S., Madl, T., D'Mello, S., & Snaider, J. (2014). Lida: A systems-level architecture for cognition, emotion, and learning. *IEEE Transactions on Autonomous Mental Development*, *6*, 19–41. doi:10.1109/TAMD.2013.2277589.

Franklin, S., Strain, S., Snaider, J., McCall, R., & Faghihi, U. (2012). Global workspace theory, its lida model and the underlying neuroscience. *Biologically Inspired Cognitive Architectures*, *1*, 32–43.

Friston, K., Kilner, J., & Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology-Paris*, *100*, 70–87.

Friston, K., Mattout, J., & Kilner, J. (2011). Action understanding and active inference. *Biological Cybernetics*, *104*, 137–160.

Gershman, S. J., & Blei, D. M. (2012). A tutorial on bayesian nonparametric models. *Journal of Mathematical Psychology*, *56*, 1–12.

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*, *521*, 452–459.

Gibson, B. R., Rogers, T. T., & Zhu, X. (2013). Human semi-supervised learning. *Topics in Cognitive Science*, *5*, 132–172.

Gobet, F., Lane, P. C., Croker, S., Cheng, P. C., Jones, G., Oliver, I., & Pine, J. M. (2001). Chunking mechanisms in human learning. *Trends in Cognitive Sciences*, *5*, 236–243.

Greenauer, N., & Waller, D. (2010). Micro-and macroreference frames: Specifying the relations between spatial categories in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 938.

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.), *Cambridge Handbook of Computational Cognitive Modeling* (pp. 59–100). Cambridge: Cambridge University Press.

Hardcastle, K., Ganguli, S., & Giocomo, L. M. (2015). Environmental boundaries as an error correction mechanism for grid cells. *Neuron*, *86*, 827–839.

Harrison, A. M., Schunn, C. D. et al. (2003). ACT-R/S: Look ma, no" cognitive-map. In *International Conference on Cognitive Modeling* (pp. 129–134).

Hartley, T., Burgess, N., Lever, C., Cacucci, F., & O'keefe, J. (2000). Modeling place fields in terms of the cortical inputs to the hippocampus. *Hippocampus*, *10*, 369–379.

Hartley, T., Lever, C., Burgess, N., & O'Keefe, J. (2014). Space in the brain: how the hippocampal formation supports spatial cognition. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *369*, 20120510.

Hirtle, S., & Jonides, J. (1985). Evidence of hierarchies in cognitive maps. *Memory & Cognition*, *13*, 208–217.

Hoyer, P. O., & Hyvärinen, A. (2003). Interpreting neural response variability as monte carlo sampling of the posterior. In S. T. S Becker, & K. Obermayer (Eds.), *Advances in Neural Information Processing Systems 15* (p. 293). MIT Press volume 15.

Hughes, M. C., & Sudderth, E. (2013). Memoized online variational inference for dirichlet process mixture models. In *Advances in Neural Information Processing Systems* (pp. 1133–1141).

Huttenlocher, J., Newcombe, N., & Vasilyeva, M. (1999). Spatial scaling in young children. *Psychological Science*, *10*, 393–398.

Huxter, J. R., Senior, T. J., Allen, K., & Csicsvari, J. (2008). Theta phase–specific codes for two-dimensional position, trajectory and heading in the hippocampus. *Nature Neuroscience*, *11*, 587–594.

Jarsky, T., Roxin, A., Kath, W. L., & Spruston, N. (2005). Conditional dendritic spike propagation following distal synaptic activation of hippocampal ca1 pyramidal neurons. *Nature Neuroscience*, *8*, 1667–1676.

Jaynes, E. T. (1996). *Probability theory: the logic of science*. Washington University St. Louis, MO.

Jensen, O., & Lisman, J. E. (2000). Position reconstruction from an ensemble of hippocampal place cells: contribution of theta phase coding. *Journal of Neurophysiology*, *83*, 2602–2609.

Johnston, D. (1981). Passive cable properties of hippocampal ca3 pyramidal neurons. *Cellular and Molecular Neurobiology*, *1*, 41–55.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Fluids Engineering*, *82*, 35–45.

Khaligh-Razavi, S., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Computational Biology*, *10*, e1003915.

Kjelstrup, K. B., Solstad, T., Brun, V. H., Hafting, T., Leutgeb, S., Witter, M. P., Moser, E. I., & Moser, M.-B. (2008). Finite scale of spatial representation in the hippocampus. *Science*, *321*, 140–143.

Knill, D. C., & Pouget, A. (2004). The bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, *27*, 712–719.

Koechlin, E., Anton, J. L., & Burnod, Y. (1999). Bayesian inference in populations of cortical neurons: a model of motion integration and segmentation in area mt. *Biological Cybernetics*, *80*, 25–44.

Körding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*, 244–247.

Kovacs, I., Kozma, P., Feher, A., & Benedek, G. (1999). Late maturation of visual spatial integration in humans. *Proceedings of the National Academy of Sciences*, *96*, 12204–12209.

Kuipers, B. (2000). The spatial semantic hierarchy. *Artificial Intelligence*, *119*, 191–233.

Langley, P., Laird, J. E., & Rogers, S. (2009). Cognitive architectures: Research issues and challenges. *Cognitive Systems Research*, *10*, 141–160.

Leutgeb, S., Leutgeb, J. K., Barnes, C. A., Moser, E. I., McNaughton, B. L., & Moser, M.-B. (2005). Independent codes for spatial and episodic memory in hippocampal neuronal ensembles. *Science*, *309*, 619–623.

Lew, A. R. (2011). Looking beyond the boundaries: time to put landmarks back on the cognitive map? *Psychological Bulletin*, *137*, 484.

Liu, J. S. (1996). Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Statistics and Computing*, *6*, 113–119.

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*, 1432–1438.

MacNeilage, P. R., Ganesan, N., & Angelaki, D. E. (2008). Computational approaches to spatial orientation: from transfer functions to dynamic bayesian inference. *Journal of Neurophysiology*, *100*, 2981–2996.

Madl, T., Franklin, S., Chen, K., & Trappl, R. (). Spatial working memory in the lida cognitive architecture. In *Proceedings of the International Conference on Cognitive Modelling* (pp. 384–389).

Maguire, E. A., Nannery, R., & Spiers, H. J. (2006). Navigation around london by a taxi driver with bilateral hippocampal lesions. *Brain*, *129*, 2894–2907.

Marr, D., & Poggio, T. (1976). *From Understanding Computation to Understanding Neural Circuitry.*. Technical Report DTIC Document.

Maurer, A. P., VanRhoads, S. R., Sutherland, G. R., Lipa, P., & McNaughton, B. L. (2005). Self-motion and the origin of differential spatial scaling along the septo-temporal axis of the hippocampus. *Hippocampus*, *15*, 841–852.

McNamara, T. P., Hardy, J. K., & Hirtle, S. C. (1989). Subjective hierarchies in spatial memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 211.

Mehta, M. R., Quirk, M. C., & Wilson, M. A. (2000). Experience-dependent asymmetric shape of hippocampal receptive fields. *Neuron*, *25*, 707–715.

Montemerlo, M., & Thrun, S. (2007). *FastSLAM: A Scalable Method for the Simultaneous Localization and Mapping Problem in Robotics (Springer Tracts in Advanced Robotics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Moser, E. I., Kropff, E., & Moser, M.-B. (2008). Place cells, grid cells, and the brain's spatial representation system. *Annual Review of Neuroscience*, *31*, 69–89.

Mueller, S. T., Perelman, B. S., & Simpkins, B. G. (2013). Pathfinding in the cognitive map: Network models of mechanisms for search and planning. *Biologically Inspired Cognitive Architectures*, *5*, 94–111.

Nardini, M., Jones, P., Bedford, R., & Braddick, O. (2008). Development of cue integration in human navigation. *Current Biology*, *18*, 689–693.

Neal, R. M. (2000). Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, *9*, 249–265.

Negenborn, R. (2003). *Robot localization and Kalman filters*. Ph.D. thesis Utrecht University.

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In W. G. Chase (Ed.), *Visual Information Processing*. New York: Academic Press.

Newman, P., Chandran-Ramesh, M., Cole, D., Cummins, M., Harrison, A., Posner, I., & Schroeter, D. (2011). Describing, navigating and recognising urban spaces-building an end-to-end slam system. In *Robotics Research* (pp. 237–253). Springer.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.

O'Keefe, J., & Burgess, N. (1996). Geometric determinants of the place fields of hippocampal neurons. *Nature*, *381*, 425–428.

O'Keefe, J., Burgess, N., Donnett, J. G., Jeffery, K. J., & Maguire, E. A. (1998). Place cells, navigational accuracy, and the human hippocampus. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *353*, 1333–1340.

Olson, E., Leonard, J., & Teller, S. (2006). Fast iterative alignment of pose graphs with poor initial estimates. In *Proceedings 2006 IEEE International Conference on Robotics and Automation* (pp. 2262–2269). IEEE.

Ong, C. S., Williamson, R. C., & Smola, A. J. (2005). Learning the kernel with hyperkernels. In *Journal of Machine Learning Research* (pp. 1043–1071).

Osborn, G. W. (2010). A kalman filtering approach to the representation of kinematic quantities by the hippocampal-entorhinal complex. *Cognitive Neurodynamics*, *4*, 315–335.

Park, E., Dvorak, D., & Fenton, A. A. (2011). Ensemble place codes in hippocampus: Ca1, ca3, and dentate gyrus place cells have multiple place fields in large environments. *PLoS One*, *6*, e22349–e22349.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, (pp. 1065–1076).

Paulin, M. G. (2005). Evolution of the cerebellum as a neuronal machine for bayesian state estimation. *Journal of Neural Engineering*, *2*, S219–S234.

Paulin, M. G., & Hoffman, L. F. (2011). Bayesian head state prediction: Computing the dynamic prior with spiking neurons. *2011 Seventh International Conference on Natural Computation*, (pp. 445–449). doi:10.1109/ICNC.2011.6022088.

Penny, W., Zeidman, P., & Burgess, N. (2013). Forward and backward inference in spatial cognition. *PLoS Computational Biology*, *9*, e1003383.

Platt, J. et al. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, *10*, 61–74.

Poggio, T., & Marr, D. (1977). From understanding computation to understanding neural circuitry. *Neurosciences Research Program Bulletin*, *15*, 470–488.

Pouget, A., Beck, J. M., Ma, W. J., & Latham, P. E. (2013). Probabilistic brains: knowns and unknowns. *Nature Neuroscience*, *16*, 1170–1178.

Prados, J. (2011). Blocking and overshadowing in human geometry learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *37*, 121.

Rapp, P. R., & Gallagher, M. (1996). Preserved neuron number in the hippocampus of aged rats with spatial learning deficits. *Proceedings of the National Academy of Sciences*, *93*, 9926–9930.

Rasmussen, C. E. (1999). The infinite gaussian mixture model. In *Advances in Neural Information Processing Systems* (pp. 554–560). volume 12.

Reid, C. R., Latty, T., Dussutour, A., & Beekman, M. (2012). Slime mold uses an externalized spatial memory to navigate in complex environments. *Proceedings of the National Academy of Sciences*, *109*, 17490–17494.

Robert, C. P., & Casella, G. (1999). *Monte Carlo Statistical Methods*. (1st ed.). Springer-Verlag.

Rosenbaum, R. S., Gao, F., Richards, B., Black, S. E., & Moscovitch, M. (2005). where to? remote memory for spatial relations and landmark identity in former taxi drivers with alzheimer's disease and encephalitis. *Journal of Cognitive Neuroscience*, *17*, 446–462.

Rosenbaum, R. S., Priselac, S., Köhler, S., Black, S. E., Gao, F., Nadel, L., & Moscovitch, M. (2000). Remote spatial memory in an amnesic person with extensive bilateral hippocampal lesions. *Nature Neuroscience*, *3*, 1044–1048.

Rossant, C., Leijon, S., Magnusson, A., & Brette, R. (2011). Sensitivity of noisy neurons to coincident inputs. *The Journal of Neuroscience*, *31*, 17193–17206.

Rovine, M. J., & Weisman, G. D. (1989). Sketch-map variables as predictors of wayfinding performance. *Journal of Environmental Psychology*, *9*, 217–232.

Russell, S., & Norvig, P. (2009). *Artificial Intelligence: A Modern Approach (3rd Edition)*. (3rd ed.). Prentice Hall.

Rusu, R. B., Maldonado, A., Beetz, M., & Gerkey, B. (2007). Extending player/stage/gazebo towards cognitive robots acting in ubiquitous sensor-equipped environments. In *ICRA Workshop for Networked Robot Systems*.

Samsonovich, A. V. (2011). Comparative analysis of implemented cognitive architectures. In *Biologically Inspired Cognitive Architectures* (pp. 469–479).

Sanborn, A. N. (2015). Types of approximation for probabilistic cognition: Sampling and variational. *Brain and Cognition*, . doi:http://dx.doi.org/10.1016/j.bandc.2015.06.008.

Santos, J. M., Portugal, D., & Rocha, R. P. (2013). An evaluation of 2d slam techniques available in robot operating system. In *2013 IEEE International Symposium on Safety, Security, and Rescue Robotics* (pp. 1–6). IEEE.

Schultheis, H., & Barkowsky, T. (2011). Casimir: an architecture for mental spatial knowledge processing. *Topics in Cognitive Science*, *3*, 778–795.

Shelton, A. L., & McNamara, T. P. (2001). Systems of spatial reference in human memory. *Cognitive Psychology*, *43*, 274–310.

Shelton, A. L., & McNamara, T. P. (2004). Spatial memory and perspective taking. *Memory & Cognition*, *32*, 416–26.

Shi, L., & Griffiths, T. L. (2009). Neural implementation of hierarchical bayesian inference by importance sampling. In *Advances in Neural Information Processing Systems* (pp. 1669–1677).

Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010). Exemplar models as a mechanism for performing bayesian inference. *Psychonomic Bulletin & Review*, *17*, 443–464.

Šimić, G., & Bogdanović, N. (1997). Volume and number of neurons of the human hippocampal formation in normal aging and alzheimer's disease. *Journal of Comparative Neurology*, *379*, 482–494.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, (pp. 99–118).

Skaggs, W. E., & McNaughton, B. L. (1996). Theta phase precession in hippocampal. *Hippocampus*, *6*, 149–172.

Solstad, T., Moser, E. I., & Einevoll, G. T. (2006). From grid cells to place cells: a mathematical model. *Hippocampus*, *16*, 1026–1031.

Spetch, M. L. (1995). Overshadowing in landmark learning: touch-screen studies with pigeons and humans. *Journal of Experimental Psychology: Animal Behavior Processes*, *21*, 166.

Stork, D. G. (1989). Is backpropagation biologically plausible? In *Neural Networks, 1989. IJCNN., International Joint Conference on* (pp. 241–246). IEEE.

Sun, R. (2008). Introduction to computational cognitive modeling. *Cambridge Handbook of Computational Psychology*, (pp. 3–19).

Sun, R., & Zhang, X. (2004). Top-down versus bottom-up learning in cognitive skill acquisition. *Cognitive Systems Research*, *5*, 63–89.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2014). Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*, .

Szilagyi, E., Halasy, K., & Somogyi, P. (1996). Physiological properties of anatomically identified basket and bistratified cells in the cal area of the rat hippocampus in vitro. *Hippocampus*, *6*, 294–305.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association*, *101*.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*, 1279–1285.

Teng, E., & Squire, L. R. (1999). Memory for places learned long ago is intact after hippocampal damage. *Nature*, *400*, 675–677.

Thrun, S., Burgard, W., & Fox, D. (2005). *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*. The MIT Press.

Thrun, S., & Leonard, J. J. (2008). Simultaneous localization and mapping. In *Springer Handbook of Robotics* (pp. 871–889). Springer.

Tulving, E., & Markowitsch, H. J. (1998). Episodic and declarative memory: role of the hippocampus. *Hippocampus*, *8*.

Tuna, G., Gulez, K., Gungor, V. C., & Mumcu, T. V. (2012). Evaluations of different simultaneous localization and mapping (slam) algorithms. In *38th Annual Conference on IEEE Industrial Electronics Society* (pp. 2693–2698). IEEE.

Tversky, B. (1992). Distortions in cognitive maps. *Geoforum*, (pp. 131–138).

Tversky, B. (2003). Navigating by mind and by body. In *Spatial Cognition III* (pp. 1–10). Springer.

Van Rooij, I. (2008). The tractable cognition thesis. *Cognitive Science*, *32*, 939–984.

Vilares, I., & Kording, K. (2011). Bayesian models: the structure of the world, uncertainty, behavior, and the brain. *Annals of the New York Academy of Sciences*, *1224*, 22–39.

Wang, J., & Schwering, A. (2009). The accuracy of sketched spatial relations: How cognitive errors affect sketch representation. *Presenting Spatial Information: Granularity, Relevance, and Integration*, (p. 40).

Wang, Y. (2015). *Motion segmentation based robust RGB-D SLAM*. Ph.D. thesis University of Technology Sydney.

Wiener, J. M., Ehbauer, N. N., & Mallot, H. a. (2009). Planning paths to multiple targets: memory involvement and planning heuristics in spatial problem solving. *Psychological Research*, *73*, 644–58. doi:10.1007/s00426-008-0181-3.

Williams, B., Cummins, M., Neira, J., Newman, P., Reid, I., & Tardós, J. (2009). A comparison of loop closing techniques in monocular slam. *Robotics and Autonomous Systems*, *57*, 1188–1197.

Wolbers, T., & Hegarty, M. (2010). What determines our navigational abilities? *Trends in Cognitive Sciences*, *14*, 138–146.

Wu, J. (2004). Some properties of the gaussian distribution.

Wurm, K. M., Hornung, A., Bennewitz, M., Stachniss, C., & Burgard, W. (2010). Octomap: A probabilistic, flexible, and compact 3d map representation for robotic systems. In *Proceedings of the ICRA 2010 Workshop on Best Practice in 3D Perception and Modeling for Mobile Manipulation*. volume 2.

Xing, E. P., Jordan, M. I., Russell, S., & Ng, A. Y. (2002). Distance metric learning with application to clustering with side-information. In *Advances in Neural Information Processing Systems* (pp. 505–512).

Yamins, D. L., Hong, H., Cadieu, C., & DiCarlo, J. J. (2013). Hierarchical modular optimization of convolutional networks achieves representations similar to macaque it and human ventral stream. In *Advances in Neural Information Processing Systems* (pp. 3093–3101).

Yuille, A., & Kersten, D. (2006). Vision as bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences*, *10*, 301–308.

Zemankovics, R., Káli, S., Paulsen, O., Freund, T. F., & Hájos, N. (2010). Differences in subthreshold resonance of hippocampal pyramidal cells and interneurons: the role of h-current and passive membrane characteristics. *The Journal of Physiology*, *588*, 2109–2132.

Zheng, W.-S., Gong, S., & Xiang, T. (2011). Person re-identification by probabilistic relative distance comparison. In *2011 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 649–656). IEEE.

Appendix A

Supplementary Information for Chapter 4

A.1 Location uncertainty in the two-dimensional case

As described in the Methods section (see Equation (3) in the main text), under Gaussian assumptions, the probability distribution of the location given a number of observations can be calculated from

$$\mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\Sigma}) = \gamma \mathcal{N}(\boldsymbol{\mu}_p, \Sigma_p) \prod_{i=1}^{N} \mathcal{N}(\boldsymbol{\mu}_{o,i}, \Sigma_{o,i}) \tag{A.1}$$

Where $\hat{\boldsymbol{\mu}}$ is the mean of the posterior or the 'best guess' location, $\hat{\Sigma}$ the uncertainty (covariance) associated with this location, $\boldsymbol{\mu}_p$ and Σ_p are the mean and the uncertainty of the prior belief location, $\boldsymbol{\mu}_{o,i}$ and $\Sigma_{o,i}$ are the means and uncertainties of the individual observations, and γ is a constant for normalization.

Analogously to univariate Gaussians (Bromiley, 2003), the product of a number of multivariate Gaussians is also a multivariate Gaussian. The covariance of the product in equation (A.1), which contains the uncertainty of the 'best guess' location, can be calculated as follows (Wu, 2004):

$$\hat{\Sigma} = (\Sigma_p^{-1} + \sum_{i=1}^{N} \Sigma_{o,i}^{-1})^{-1} \tag{A.2}$$

According to hypothesis 3 (see Hypotheses section), the observation uncertainty is proportional to the distance d_i: $\sigma_{o,i} = s \cdot d_i$ in the one-dimensional case (where s is a factor modelling how sensory uncertainty depends on distance). In the two-dimensional

case, the uncertainty depends on the distances to the landmark in the x and y dimensions, $d_{x,i}$ and $d_{y,i}$, as well as the factors s_x and s_y controlling the dependences of the sensory uncertainties in the x and y dimensions, and the correlation ρ between x and y (see Negenborn (2003) or Thrun et al. (2005) for more complex sensory uncertainty models).

$$\Sigma_{o,i} = \begin{bmatrix} (s_x d_{x,i})^2 & (\rho s_x s_y d_{x,i} d_{y,i}) \\ (\rho s_y s_x d_{y,i} d_{x,i}) & (s_y d_{y,i})^2 \end{bmatrix} \tag{A.3}$$

Thus, the covariance matrix for the 'best guess' location estimate can be calculated from the distance measurements $d_{x,i}$ and $d_{y,i}$ to each landmark from equations (A.2) and (A.3). The covariance matrix modelling path integration uncertainty, Σ_p, and the factors modelling the sensory uncertainty, s_x and s_y (i.e. controlling how rapidly the accuracy of distance judgements decrease with increasing distances in the x and y dimensions) and the correlation ρ are adjustable parameters.

A.2 Coincidence detection as rejection sampling and multiplication by coincidence detection

Coincidence detection as rejection sampling

In the simple three-neuron example shown in Figure 6, the computation performed by the posterior neuron (place cell), taking as inputs a prior (grid cell) and an observation (BVC), can be shown to approximate Bayesian inference (i.e. to implement equation (1) of the main text). Let us consider a temporally constrained spike train, and view each spike within this spike train as a sample taken from a probability distribution - either the spikes of the place cell, sampling the posterior location distribution $p(x|o)$, or those of the grid cell, sampling the prior location distribution $p(x)$, or the spikes of the BVC, sampling the observation distribution $p(o|x)$. In this case, the computation performed by the place cell is equivalent to the rejection sampling technique (Liu, 1996; Bishop et al., 2006) used to approximate an unknown distribution. In rejection sampling, in order to approximate an unknown distribution p_u, a known distribution q is sampled which satisfies

$$\forall z : p_u(z) < Mq(z) \tag{A.4}$$

Where $M > 1$ bounds $p_u(z)/q(z)$. To ensure that the samples approximate $p(z)$, the known $q(z)$ is iteratively sampled, and the samples are accepted with a probability proportional to the ratio

$$p_A = p(accept|z) = \frac{p_u(z)}{Mq(z)} \tag{A.5}$$

If each sample is randomly drawn from q, and is accepted with probability p_A, it is straightforward to show that these samples will approximate $p(z)$ (Liu, 1996; Bishop et al., 2006). Briefly, the probability distribution over the accepted samples $p(z|accept)$ has to equal the unknown distribution p_u when using an infinite number of samples for the following reason. Using Bayes' theorem,

$$p(z|accept) = \frac{p(accept|z)p(z)}{p(accept)} \tag{A.6}$$

Where

$$p(accept|z) = \frac{p_u(z)}{Mq(z)} \tag{A.7}$$

And the prior probability of a sample z is given by q: $p(z) = q(z)$. The prior probability of acceptance is given by marginalizing

$$p(accept) = \int_z p(accept|z)p(z)dz = \int_z \frac{p_u(z)}{Mq(z)}q(z)dz = \frac{1}{M}\int_z p_u(z)dz = \frac{1}{M} \tag{A.8}$$

Thus, substituting into equation (A.6), we obtain the required equality.

$$p(z|accept) = \frac{\frac{p_u(z)}{Mq(z)}q(z)}{\frac{1}{M}} = p_u(z) \tag{A.9}$$

In our coincidence detection model (Figure 6), the acceptance probability of spikes generated by the grid cell is proportional to the spiking probability of the BVC. This is just the acceptance probability required in order to approximate a Bayesian posterior by rejection sampling, which can be shown as follows. We use the Bayesian posterior as the unknown distribution that is to be approximated (cf. equation (1) in the main text)

$$p_u = p(x|o) = \gamma p(x)p(o|x) \tag{A.10}$$

Where γ is a constant for normalization. We also assume that the grid cell in Figure

6 represents the prior $p(x)$ and that the BVC represents the observation likelihood $p(o|x)$. Substituting these expressions into equation (A.5), the acceptance probability then becomes

$$p_A = \frac{p_u(z)}{Mq(z)} = \frac{\gamma p(x)p(o|x)}{Mp(x)} = kp(o|x) \tag{A.11}$$

Where $k = \frac{\gamma}{M}$. Accepting samples drawn from the prior with this probability p_A ensures that the accepted samples will approximate the Bayesian posterior. Since in the simple network of Figure 6, the acceptance probability of spikes generated by the prior neuron (grid cell) is proportional to the spiking probability of the observation neuron (BVC) because of coincidence detection (Rossant et al., 2011), as in equation (A.11) - with spiking probability 1, every grid cell spike would be coincident with a BVC spike and thus accepted, and with spiking probability 0, no grid spike would be accepted - the network approximates a Bayesian posterior, just like a rejection sampling algorithm. With infinite firing rates, the network would yield the exact posterior. In realistic cases the errors depend on the membrane time constant and firing threshold (they lie around 5% for the parameters of CA1 place cells - see Results section in the main text). Figure 7 in the main text shows the error rates of an integrate-and-fire spiking neuron with different parameters.

Multiplication by coincidence detection

This section describes a mathematical model of how coincidence detection in spiking neurons can implement multiplication with any number of inputs, and thus approximate a Bayesian posterior from multiple observations (i.e. implement equation (2) from the main text). In the following we will assume that the spatially localized firing behaviour of place cells can be approximated and modelled by probability distributions, cf. hypothesis 1 in the Introduction of the main paper.

Specifically, we will assume that place cell instantaneous firing rates are proportional to the probability density function representing the probability that the rat is in a particular location: $r \propto p$. In the one-dimensional case, the probability P_{x_A,x_B} that a rat is located on a path lying between the points x_A and x_B in the environment, which it traverses between times t_A and t_B (at which it is at locations x_A and x_B respectively), is proportional to

$$P_{x_A,x_B} = \int_{x_A}^{x_B} p(x)\,\mathrm{d}x \propto \int_{t_A}^{t_B} r(t)\,\mathrm{d}t \tag{A.12}$$

If the place cell does not fire, the integral on the right yields zero, which under our assumption means that the probability that the rat is in the location represented by the place cell is zero. On the other hand, if the firing rate of the place cell is very high in the time interval during which the rat crosses the place cells represented location, the probability that the rat is in that location is also very high.

One way to approximate integrals with a number of samples randomly drawn from the function to be integrated is called Monte Carlo integration (Robert & Casella, 1999). If we view individual spikes of a neuron as such samples, then the density of a spike train can be viewed as approximating an integral of the form above (Monte Carlo approximation of probability distributions by spiking neurons has been suggested before, see e.g. (Hoyer & Hyvrinen, 2003; Paulin, 2005; Paulin & Hoffman, 2011; Bsing et al., 2011)).

Using a binary function S to represent a spike train, which at time t is $S(t) = 1$ if a neuron has fired a spike within the time interval $[t, t+\tau)$, and 0 otherwise, equation (A.12) can be approximated by the spike train S as follows.

$$\int_{t_A}^{t_B} r \, \mathrm{d}t \approx \frac{1}{N} \sum_{t \in T_{A,B}} S(t) \tag{A.13}$$

$$P_{x_A, x_B} \propto \frac{1}{N} \sum_{t \in T_{A,B}} S(t) \tag{A.14}$$

Where $T_{A,B}$ denotes the interval between t_A and t_B (during which the rat was located between x_A and x_B) in τ time steps, and $N = \frac{t_b - t_a}{\tau}$ is the number of time steps of duration τ within the interval $T_{A,B}$. In the context of this paper, we can neglect multiplicative constants and work with the proportionality relations, because the most important task of localization is to find the 'best guess' location $\boldsymbol{x}_b$ in the environment, i.e. the expected value of the location $\boldsymbol{x}$, for which the amplitude of the firing rate distribution is unimportant. Finding the maximum can be expressed as in equation (A.15). There is an alternative and possibly more accurate (Jensen & Lisman, 2000) way of deriving a 'best guess' location from place cell firing, based on theta phase instead of the maximum firing rate (see Discussion), but that way of estimating location does not depend on absolute firing rates either. The 'best guess' location $\boldsymbol{x}_b$ based on the expected value of the location distribution can be calculated from the function $S(t)$ representing the spike train, the locations $\boldsymbol{x}(t)$ at the time of each spike, and the total number of spikes

K in this interval, which is the sum of $S(t)$ over $T_{A,B}$:

$$\boldsymbol{x}_b = \mathbb{E}[\boldsymbol{x}] \approx \frac{1}{K} \sum_{t \in T_{A,B}} S(t)\boldsymbol{x}(t) = \frac{\sum_{t \in T_{A,B}} S(t)\boldsymbol{x}(t)}{\sum_{t \in T_{A,B}} S(t)} \tag{A.15}$$

The number of spikes is $K = \sum_{t \in T_{A,B}} S(t)$. Using standard deviation as a measure of uncertainty, the location uncertainty can be described as

$$\Sigma_b = \sqrt{\mathrm{Var}(p)} \approx \sqrt{\frac{1}{K} \sum_{t \in T_{A,B}} (S(t)\boldsymbol{x}(t) - \boldsymbol{x}_b)^2} \tag{A.16}$$

Using this way of approximating probability distribution functions with spike train densities, coincidence detection in spiking neurons can implement multiplication. Looking at two neurons A and B providing input to a third neuron C which performs coincidence detection, it can be shown that the function represented by C's spike train will approximate the product of the functions approximated by A and B. If we set the time discretization parameter τ, to the temporal resolution of coincidence detection (which mainly depends on the membrane potential and signal-to-noise ratio, see (Brette, 2012) or Figure 7 in the Results section of Chapter 4 for an error analysis), and ensure that C's spike threshold is high enough so that C only fires if input spikes arrive from both A and B within τ, then the function represented by C's spike train S_C will depend on the product of S_A and S_B:

$$S_C(t) = S_A(t)S_B(t) \tag{A.17}$$

Equation (A.17) is also extensible to the multiplication of a larger number M of input neurons Ni. If the threshold of the output neuron N is so high as to require synchronous spikes from *all* inputs within a time τ, it can simply be extended to calculate the function S_O representing the spike train of the output neuron:

$$S_O(t) = \prod_{i}^{M} S_{Ni}(t) \tag{A.18}$$

In general, the threshold will not be so high as equation (A.18) assumes. Hence, the output neuron will generally have a threshold such that a proportion $\alpha = (0,1]$ of all input neurons M spiking synchronously can elicit an output spike (i.e. there is an output spike only if $m > M\alpha$ input spikes arrive within τ). Then the approximation of a product of multiple input spike trains becomes

$$S_O(t) = \mathrm{H}\left(\frac{1}{M}\sum_{i=1}^{M}(S_i(t) - \alpha)\right) \tag{A.19}$$

Where $\mathrm{H}(a) = \begin{cases} 0 & \text{if } a < 0 \\ 1 & \text{if } a \geq 0 \end{cases}$ is the Heaviside step function. This equation equals (A.18) if $\alpha = 1$, and approximates it otherwise. It is a simplified formulation of how coincidence detection can perform multiplication in a spiking neuron. We can insert it into equation (A.14) to obtain the approximation of a probability distribution as follows.

$$P_{x_A,x_B} \propto \frac{1}{N}\sum_{t\in T_{A,B}} S_O(t) \tag{A.20}$$

The expected value of this function, i.e. the represented 'best guess' location, will approximate the mean of the product of input functions - see equation (A.15).

$$\boldsymbol{x}_b = \mathbb{E}[\boldsymbol{x}] \approx \frac{1}{K}\sum_{t\in T_{A,B}} S_O(t)\boldsymbol{x}(t) = \frac{\sum_{t\in T_{A,B}} S_O(t)\boldsymbol{x}(t)}{\sum_{t\in T_{A,B}} S_O(t)} \tag{A.21}$$

The particular threshold of the output neuron plays a large role in determining the accuracy of this approximation, as do the number of samples (spikes). This computation requires the neuronal parameters influencing temporal resolution and the threshold to be within a certain range to allow for reasonably accurate localization. Our computational simulations indicate that the empirically observed parameters of hippocampal place cells are indeed within a range to allow for statistically near-optimal localization (see the Results section in the main text).

Appendix B

Supplementary Information for Chapter 5

B.1 Tree analysis algorithm

We used an algorithm to extract map structure from recall orders which is functionally equivalent to the ordered tree algorithm used in prior work (Hirtle & Jonides, 1985; McNamara, 1986; McNamara et al., 1989), with the exception that we disregard order information (whether or not the leaves were always recalled in a particular ordering). The algorithm takes a list of recall protocols, as well as cues, and all possible buildings, and returns the map structure (all sets of buildings which always occur together).

Algorithm B.1.1: EXTRACTMAPSTRUCTURE($Protocols, Cues, Buildings$)

```
1 : submaps ← {}
2 : for each tuplelength ∈ (1, |Buildings| − 1)
3 :     for each C ∈ Combinations(Buildings, tuplelength)
4 :         occurseverywhere ← True
5 :         for each p ∈ (0, |Protocols|)
6 :             perm ← Permutations(C)
7 :             if Cues[p] ∉ C and ∀(PC ∈ perm : PC ∉ Protocols[p])
8 :                 occurseverywhere ← False
9 :                 break
10 :        if occurseverywhere
11 :            submaps ← submaps ∪ C
```

The algorithm iterates through all possible tuple lengths, and generates all possible combinations at the current tuple length. For these combinations C, it checks whether any permutation of C occurs uninterrupted in all protocols (i.e. whether all buildings in C have been recalled together); if so, C is added to the list of submaps. Notably, this check is only performed if C is not cued (line 7). It was argued in previous literature (Hirtle & Jonides, 1985; McNamara et al., 1989) that cueing can disrupt the re-call process. Therefore containment in all protocols is only tested for combinations which do not contain a cue, in order to avoid erroneously disregarding sub-maps which consistently occur together in all recall protocols except in those in which the cue has disrupted the natural recall order.

B.2 Full list of cities chosen by included subjects

The map in Figure B.1 provides a visual overview over all cities within which spatial memory data has been collected from the participants.

Figure B.1: Overview over the 149 cities chosen by subjects in Experiments 1, 3A and 3B.

List of cities in Experiment 1: Albany, Albuquerque, Ames, Ann Arbor, Austin, Baltimore, Belgrade, Belmopan, Buffalo, Chennai, Chicago, Chico, Cincinnati, Corvallis, Cupertino, Denton, Denver, Dunmore, Fort Collins, Gobichettipalayam, Hampton, Klamath Falls, Kochi, Lakeland, Las Vegas, Los Angeles, Madurai, Miami, Minneapolis, Miramar, Mount Kisco, New York, Orange (FL), Pittsburgh, Pittsfield Charter Township, Potterville, Reno, Rome, San Angelo, San Bernardino, San Diego, Somerville, Springfield, Strasbourg, White Salmon, Williamsburg, Wilmington.

List of cities in Experiment 3A: Alameda, Austin, Beacon, Bedford, Belleville, Bellingham, Bengaluru, Berkeley, Bloomington, Boston, Bowie, Bowling Green (KY), Brooksville, Brownsville, Buffalo, Burlington, Cambridge, Camden, Cape Girardeau, Castlerock, Chicago, Cincinnati, College Station (TX), Colombo, Columbia, Denver, Desert Hot Springs (CA), Desoto, Duluth, Eastbrunswick,

Edinburgh, Fairway, Farmersville, Fayetteville, Franklin, Germantown, Gettysburg, Glasgow, Goleta, Harwoodheights, Hemet, Highridge, Hollywood, Holt, Houston, Islavista, Jaipur, Karur, Keller, Lackawanna, Lake Oswego, Land O' Lakes, Lindsay, Little River-Academy (TX), Live Oak (TX), London, Lubbock, Marthandam, Mayfield, Minneapolis, Mission, Nagercoil, New York, Norridge, Orange, Overland Park (KA), Owensville, Palmsprings, Perryville, Pigeonforge, Poplarbluff, Portland, Poughkeepsie, Princeton, Provo, Revere, Rochester, Rochester Hills, Roeland Park, Salem, San Antonio, San Diego, Sanger, Savage, South Bend, Southport, Springboro, Springhill, St. Charles, Stony Brook (NY), St. Peters, Temple, Tirunelveli, Towson, Visakhapatnam, Warren, Weatherford, Wilmington, Xenia, Ypsilanti.

List of cities in Experiment 3B: Algonquin, Ashland, Chicago, Columbia, Jefferson City, Kansas City, Knoxville, Lexington, Linden, Medford, Minneapolis, Missoula, Mound, Overland Park, Portland, Seattle, Stara Zagora.

B.3 Exclusion of learning effects

A possible criticism of our results could be the claim that the structure apparent from the recall protocol orderings is being learned by the subjects during the recall process, as opposed to being an inherent property of their long-term memory (LTM). Our analysis procedure assumes one consistent structure in LTM underlying the recall protocols; and excludes possible 'outliers' using the jackknifing procedure (i.e. protocols which, when included, would statistically significantly change the resulting structure, are excluded from analysis).

If this assumption was incorrect, and subjects learned the structure during the experiment - or, alternatively, re-learned a different structure, then this would be apparent from the pattern of omitted recall protocols. Specifically, it would mean a significantly larger number of omitted early protocols compared to late protocols (the first few protocols would be inconsistent with the learned structure more often than the last few).

To test whether this learning effect can be observed, we have tested the distributions of omitted recall protocols against the null hypothesis that the likelihood of omissions was uniform (just as likely to occur for the first few as for the last few protocols), using a chi-square test. The table below shows the results.

For the real-world experiments, the null hypothesis cannot be rejected; thus, it is likely that there is no learning effect, and that our recall order paradigm indeed measures structures which have already been committed to LTM before the experiment. For the virtual reality experiment (Exp. 2), there seems to be some small non-uniformity, although not significant at $\alpha = 0.01$. However, contrary to the objection that the structure arises from learning during the recall trials, early protocols were less

Exp. 1	**Exp. 2**	**Exp. 3A**	**Exp. 3B**
$p = 0.886 > 0.01$	$p = 0.015 > 0.01$	$p = 0.146 > 0.01$	$p = 0.495 > 0.01$
$c = 2.339$	$c = 15.698$	$c = 9.538$	$c = 8.393$

Table B.1: Results of chi-squared tests against the null hypothesis that there is no learning effect in the recall protocol data, i.e. that early recall protocols are as likely to be outliers than late recall protocols (p is the p-value of the test; c denotes the chi square test statistic). The non-significance of the results suggests that our recall order paradigm measures a property of long-term memory, and not something learned during the recall trials.

likely [1], instead of more likely, to be excluded as outliers compared to late protocols.

B.4 Separability of co-represented and not co-represented building pairs

The co-representation correlations reported in Section 3.3 of the main text raise hopes of straightforward predictability - what if a simple distance thresholding or linear decision boundary in the reported feature space is capable of fully explaining cognitive map structure, even for the random testing environments? Unfortunately, within sub-map and across sub-map building pairs are not linearly separable; and difficult to separate in general, even with complex state of the art classifiers.

Figure B.2 shows the distances of all pairs of buildings in all features in Experiment 2, normalized by dividing each feature by its standard deviation for each participant map, and compressed down to two-dimensional space for visualization using t-SNE (without normalization across buildings of each map, classifiers are unable to perform above chance). Apart from the building pairs (which concentrate into two groups according to function - shops and houses), decision boundaries obtained with three different classifiers are also plotted. Although there is a trend of building pairs being more likely to be on the same sub-map when closer together (higher concentration of same-map pairs towards the lower left), the data is clearly not well-separable. As can be seen from this Figure, accurate prediction of full subject map structures - or even whether single building pairs belong to the same sub-map - using simple classification is impossible using naive approaches. More complex machine learning

[1] The frequency of omissions in Experiment 2 were: 1.1% for the protocols presented first, 0.7% for those at position 2, 1.3% at position 3, 1.6% at position 4, 1.7% at position 5, 1.9% at position 6, and 1.8% at position 7

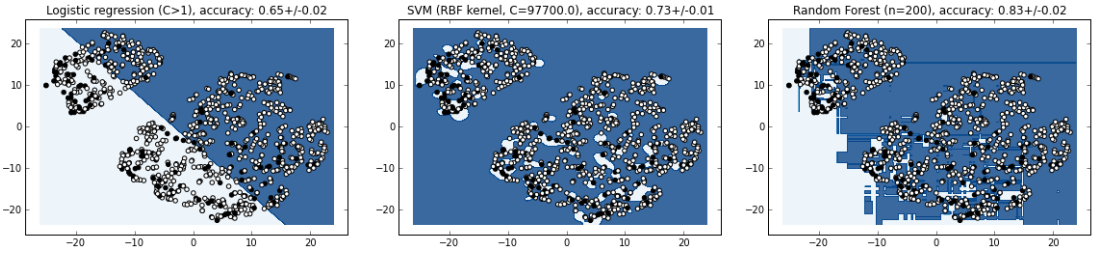


Figure B.2: Pairs of buildings in the space of all features, and separability according to co-representation, in Experiment 2 using 3 different classifiers: logistic regression (left), Support Vector Machine with RBF-kernel (middle) and Random Forest (right). Each point represents a building pair (filled black if both buildings lie on the same sub-map, and white if they do not), with its position being a two-dimensional projection of the full six-dimensional feature space using t-SNE. Although 2D decision boundaries are visualized, the reported classifier accuracies were obtained in the original feature space, using 10-fold cross validation and after hyperparameter optimization.

algorithms such as random forests (state-of-the art classifiers based on ensembles of decision trees) (Breiman, 2001) can predict for around 83% of building pairs whether they belong to the same representation (note that the accuracies were obtained by classifying the full high-dimensional data set, not just the 2D projection plotted in Figures B.2 and B.3). However, the map structures collected in our experiments contain 10 and 28 pairs (in the 5-building and 8-building maps), which would make the probability of full map structures - all pairs - being predicted correctly using this approach 15.5% and 0.5% respectively (the situation is even worse in real-world environments, as can be seen in the next section).

In the more complex real-world setting of Experiment 3, separating pairs of buildings which belong to the same sub-map and those belonging to different sub-maps is even more difficult than in virtual reality environments, as shown by Figure B.3 and evidenced by the lower accuracies obtained after 10-fold cross validation. This Figure shows the distances of all pairs of buildings in all features, normalized by dividing each feature by its standard deviation for each participant map, and compressed down to two-dimensional space for visualization using t-SNE. Note that the Figure shows prediction accuracies of pairs of buildings (whether or not a pair was represented on the same sub-map), and not of entire map structures. To correctly predict a map structure, all pairs within would need to be predicted correctly. Given the 77% accuracy of the best classifier in Figure B.3, correct predictions based on classification are even more unlikely than in Experiment 2 ($0.77^{\binom{5}{2}} = 7.3\%$ in condition A, and $0.77^{\binom{8}{2}} = 0.0\%$ in condition B).

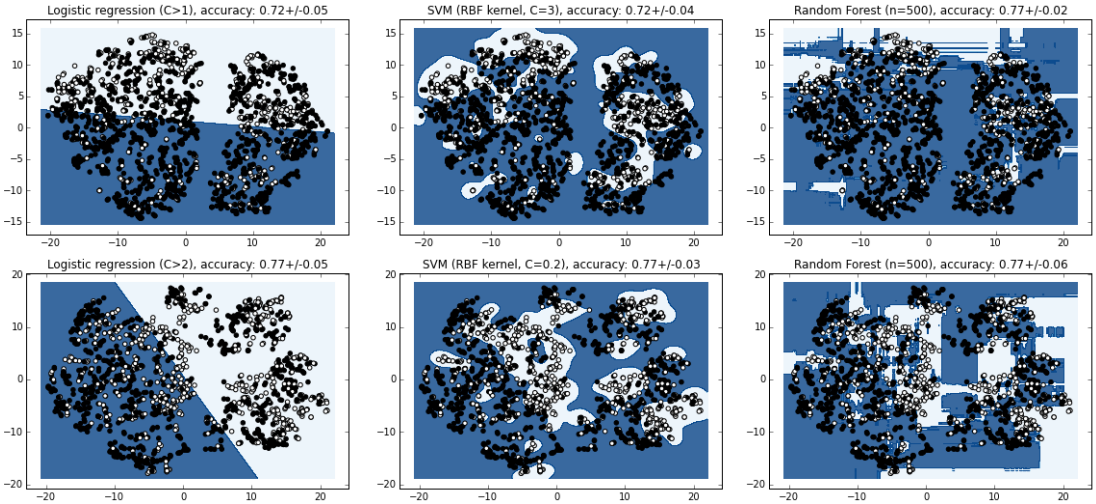


Figure B.3: Pairs of buildings in the space of all features, and their separability according to co-representation, in Experiment 3 using three different classifiers: logistic regression (left), Support Vector Machine with RBF-kernel (middle) and Random Forest (right). Top: Condition A. Bottom: Condition B

Appendix C

Supplementary Information for Chapter 6

C.1 Comparison of hierarchical activation gradient-based route planning with human performance on the TSP task

In order to evaluate the plausibility of gradient-based multi-goal route planning, as described in the main text (see Figure 5), we have used data collected from participants recruited on Amazon Mechanical Turk, tasked with solving the travelling salesperson problem (TSP) in virtual reality environments. Data from 46 participants was analysed here. They were asked to mark all buildings in the 3D environment and then return to the building they started out with, using the shortest path possible (see (Madl et al., 2013) for details).

Each participant performed 5 trials in each of three types of environments: random (in which buildings were randomly distributed), clustered by looks (in which buildings of the same type, e.g. churches, were ensured to be grouped, close to each other), and clustered by distance (in which some buildings were placed close to build groups, regardless of their visual similarity).

Figure C.1 shows participant performance, compared with a gradient-based route planner operating on a flat (single-level) representation. Interestingly, this non-hierarchical model seems to explain the human data well. As a caveat, it should be mentioned that participants were not checked for prior experience with 3D environments (an unknown

percentage may have had trouble with the controls, falling back to the simplest strategy). Furthermore, this task is inherently more difficult in virtual reality, where cues important in real-world navigation are not available (e.g. depth information from stereo disparity, path integration / self movement information, etc.).

To avoid these caveats, we have replicated a real-world TSP experiment by (Wiener et al., 2009). This experiment was performed in a 6.0m x 8.4m room, with 25 different locations marked by boxes with symbols on them, as illustrated in Figure C.2A. Subjects were given a 'shopping list' containing a number of different symbols, each of which denoted a location that they had to visit, and they subsequently had to plan the shortest route visiting all of these locations. Figure C.2 shows subjects' performance at this task, and compares it with the simulated performance of an agent using the gradient climbing heuristic on two-level hierarchical cognitive map. The models performance closely accounts for human data, as can be seen from this figure, which substantiates the models cognitive plausibility.

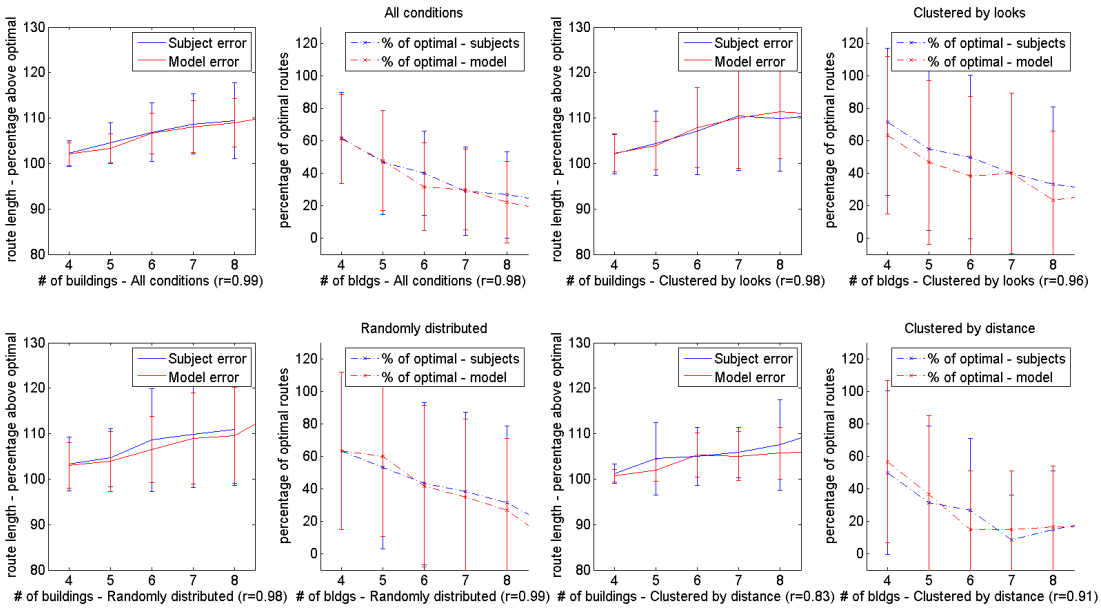


Figure C.1: Human performance in virtual reality, compared to gradient-based planning on a flat grid of place nodes.

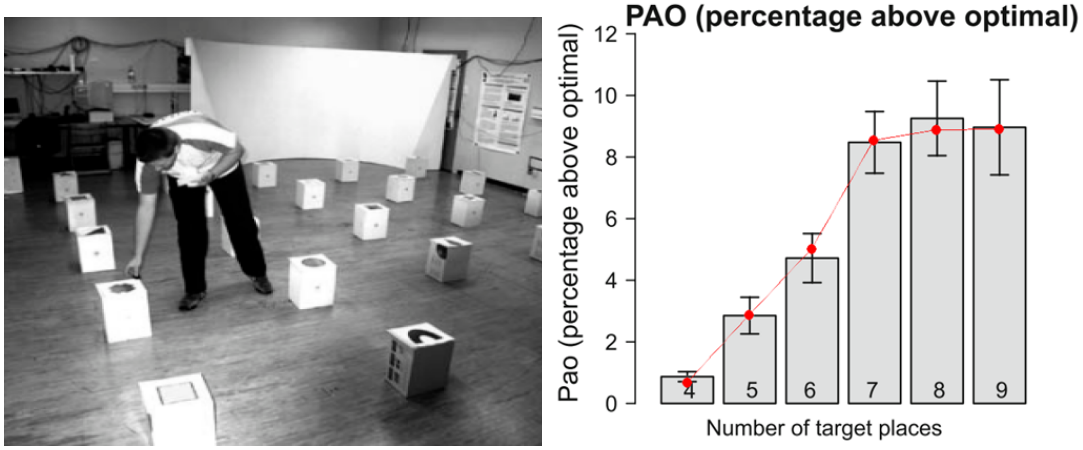


Figure C.2: Human performance in a real-world experiment (Wiener et al., 2009), compared to gradient-based planning, on a hierarchical grid of place nodes. Figures modified from (Wiener et al., 2009) with permission.

Appendix D

Additional evidence for sampling-based Bayesian localization

In Chapter 7, we have addressed the observation that there is a subset of neurons in the hippocampus with firing fields which seem to contradict the uncertainty predictions of a Bayesian model. The reasonably god fit to place fields sizes reported in Chapter 4, although imperfect, suggest that some place cells do approximate Bayesian posteriors. However, it is very likely that only a subset does so, and that part of the hippocampus performs different computations altogether.

To show that our sampling-based Bayesian localization model can function even under conditions where a proportion of the samples is corrupted by non-Bayesian processes, we have compared the distribution of uncertainties predicted by the sampling-based model, corrupted by 20% uniformly random samples, to the distribution of place field sizes recorded in an equivalent environment.

We have used the implemented sampling-based cognitive model (described in Chapter 6) to replicate the experiment (Burke et al., 2011). In the experiment, rats were running in circles on a circular track with 106cm diameter, which contained 2 food trays to motivate the rats, and no objects in one condition and 8 pseudorandomly distributed, different objects in another condition. The model performed random trajectories in an environment of the same proportions, with the same landmarks.

Figure D.1 shows the resulting distribution of uncertainties along the track, measured as the standard deviation of all posterior samples (location hypotheses) in the model. These uncertainties are compared to the distribution of place field sizes along the track. Note the matching ratio of the most frequent place field sizes and uncertainties at $21cm$ (objects) and $38cm$ (no objects) respectively, and the roughly matching

ratio between all normalized frequencies of occurrence between objects and no objects conditions.

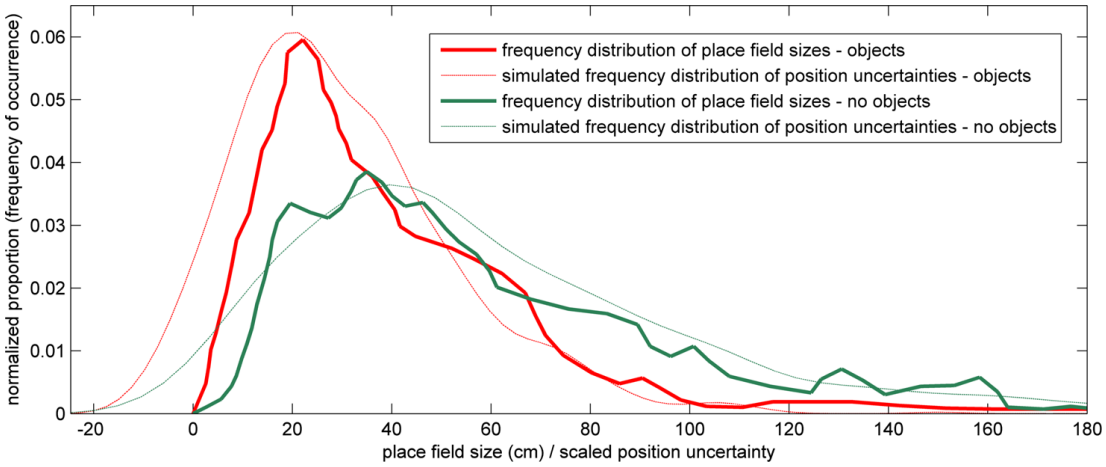


Figure D.1: Occurrence frequencies of different place field sizes in all measured neurons in the no objects (blue) and 8 objects (red) conditions, compared to the frequencies of position uncertainties in the simulation. Data from (Burke et al., 2011).

Appendix E

Metric learning in pairwise difference space

Introduction

Constraints on which data points should or should not be grouped together (must-link or cannot-link constraints) can significantly improve the performance of clustering. Most conventional semi-supervised approaches try to minimize must-link distances, maximize cannot-link distances, or to enforce a constant limit on either and optimize the other. They do not make explicit use of the probability distribution of must-link and cannot-link distances.

Here, we propose a framework for semi-supervised clustering based on projecting the data into a distance space in which distances reflect the linkage probabilities of belonging to the same cluster, using simple probabilistic classifiers trained on the available constraints. The framework can be seen as a novel approach to perform non-linear metric learning using weak supervision in the form of pairwise constraints, in order to improve clustering performance, as pioneered by (Xing et al., 2002). Although this problem is very similar to metric learning in general, the criterion of interest is often somewhat different: as opposed to optimizing the performance of some classifier as in e.g. (Xing et al., 2012)), or for a large margin as in e.g. (Weinberger et al., 2005), the goal is ensuring that all instances of a cluster are closer under the learned metric than those of different clusters. Furthermore, semi-supervised methods requiring partially labelled instances are not applicable if only pairwise constraints are available.

We have used this framework to model the structure of spatial representations in human participants in Chapter 5 above, using the information which buildings have

or have not been co-represented as the must-link and cannot-link constraints to train a Gaussian Discriminant Analysis (GDA) model in absolute pairwise difference space (see also Chapter 2.5 for the mathematical formulation, and Chapter 5 for the results).

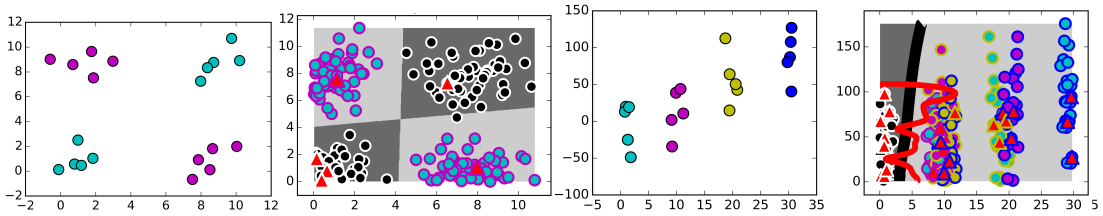
Motivation

For general applicability, non-linear metrics are vital, because of the problem of *multi-modality*. When data points forming several groups or 'modes' in unweighted feature space actually belong to the same cluster semantically, as indicated by ML constraints, there exists no linear projection able to separate them, and linear methods must inevitably fail. A traditional example is the XOR dataset, consisting of four groups, connected by diagonal ML constraints, such that there exists no linear separating hyperplane - see Figure E.1a.

Although kernel-based methods can deal with multimodal clustering problems (or any complicated data distribution) according to the Representer Theorem, in theory, given the optimal kernel and suitable parameters, in practice it is often difficult to find such a kernel. Most non-linear metric learning methods able to learn suitable kernels are sensitive to multiple hyperparameters, and, being nonconvex optimization problems, frequently get stuck in local minima.

A further issue with popular isotropic kernels such as the Radial Basis Function (RBF), frequently used in non-linear metric learning (Baghshah & Shouraki, 2010; Chitta et al., 2011), is that they are ill-suited for data with features on very different scales, since the optimal regularization parameters in one dimension can be suboptimal in other dimensions in this case, as pointed out by (Ong et al., 2005) (who propose a solution only in the supervised setting).

Our motivations for proposing a novel non-linear method are threefold. First, we would like to sidestep the difficulty of robustly finding a good non-linear metric for a particular dataset, in a probabilistic framework, without hyperparameter tuning. Second, instead of learning a metric using an objective function based on Lp-distance, which collapses the differences along the individual dimensions into one value, we would like to let the model directly access these individual differences, and thus to learn their importance, allowing it to easily deal with non-isotropic data (Figure E.1c-d). Third, we would like to make use of prior information regarding what constitutes a 'good' metric for clustering. In particular, unlike using a weighted Lp-metric which



(a) XOR dataset **(b)** Pairwise differences **(c)** Non-isotropic variance **(d)** Pairwise differences

Figure E.1: Motivation for the proposed metric learning approach. (a) Example not linearly separable data requiring non-linear metrics. (b) Visualization of the distribution of corresponding absolute pairwise differences (APD), containing the element-wise differences in all dimensions between all possible objects, within (black) and across (coloured) clusters. The background contour shows the probability of a pair with a given distance belonging to the same cluster, learned by Gaussian Discriminant Analysis, and used as the distance pseudometric. Red triangles show the labelled ML and CL constraints. (c) Example data with non-isotropic variance (three orders of magnitude larger in the y-axis direction). (d) Corresponding APD space coloured as in (a). In addition to the modelled probability of belonging to the same cluster, the models decision boundary (black), as well as the decision boundary of a Support Vector Machine with an isotropic RBF kernel (optimal parameters set by grid search), which overfits along the low-variance dimension.

attempts to summarize the structure of within-cluster and across-cluster pairwise distances as scalars, we define *a pseudo-metric based on the vector space of absolute pairwise differences* (APD).

This not only allows straightforward learning of the importance of each feature and thus fitting non-isotropic distributions better than isotropic kernels (Figure E.1c-d), but also makes explicit the structure in the distribution of constraints. It has been observed before that for data containing clusters, the probability density function of pairwise Lp distances shows two peaks (one for within- and one for across-cluster pairs), e.g. by (Brin, 1995). However, in the case of multiple clusters with different shapes and variances, a bimodal distribution is insufficient to reflect the true distributions of the instance differences within or across clusters. Clearly, within-cluster variances in one cluster do not have to equal those in another cluster, and the same is true for across-cluster variances (as illustrated by the variances of the groups of data in APD space in Figure E.1b and d). Learning in pairwise difference vector space (instead of collapsing these distances into scalars) allows our model to adapt locally to within- and across-cluster variances of different clusters, and therefore to better approximate the true pairwise distance distribution.

E.1 Supervised learning in absolute pairwise difference space

The formulation of (weakly) supervised learning in absolute pairwise difference space was given in Chapter 2.5 above. We briefly summarize it as follows.

Let $\mathcal{X} = (\boldsymbol{x}_i, ..., \boldsymbol{x}_n)$ be the feature vector representation of n objects which are to be clustered, where $x_i \in \mathbb{R}^D$ are vectors with D dimensions. Let the set of m given pairwise linkage constraints be denoted by $\mathcal{L}$, where $|\mathcal{L}| = m$, and $l_{i,j} \in \mathcal{L}$ is

$$l_{i,j} = \begin{cases} 1, & \text{if } i \text{ and } j \text{ belong to the same cluster (ML constraint)} \\ 0, & \text{if } i \text{ and } j \text{ belong to different clusters (CL constraint)} \end{cases} \tag{E.1}$$

Then, a distance metric d_l between two instances can be defined based on the probability that these instances belong to the same cluster, making use of a generative classifier (constituent model) trained on the given constraints:

$$d_l(\boldsymbol{x}_1, \boldsymbol{x}_2; \boldsymbol{\theta}) = 1 - p(l = 1 | \Delta\boldsymbol{x}, \boldsymbol{\theta}) = p(l = 0 | \Delta\boldsymbol{x}, \boldsymbol{\theta}) \tag{E.2}$$

There are three advantages of using generative classifiers instead of just the classification output as the pseudometric. First, this makes the model well-suited to learning in the inductive setting (where the examples from the test set are not seen at training time). Many existing approaches are designed for and evaluated in the transductive setting. Second, on more complex, unseparable data, utilizing the models confidence of whether instances should be linked, in addition to a binary prediction, greatly improves the resulting clustering. This choice also allows choosing suitable priors and likelihoods and thus tailoring the model to fit the data at hand.

The learned metric in (E.2) can subsequently be embedded into a new feature space using multi-dimensional scaling (MDS), and then used for clustering or classification.

E.2 Extension of the framework to semi-supervised learning

Given the large number of pairwise differences, $\binom{n}{2} = \frac{n^2-n}{2}$, the following question arises: why only use the tiny set of given pairwise constraints for learning, instead of making use of the entire distribution in APD space?

An extension of a recently proposed framework for semi-supervised learning, called Contrastive Pessimistic Likelihood Estimation (CPLE) (Loog, 2015), can facilitate such an approach for using the entire APD space. Loog (2015) points out that most current semi-supervised approaches are not 'safe' (do not guarantee better performance when including the unlabelled data than without), and often perform sub-optimally due to model misspecification (making assumptions violated by data). Instead, he suggests only using the intrinsic assumptions of a given classifier, and training it in a pessimistic framework: using optimization to find hypothetical labels for the unlabelled data corresponding to the worst case. This pessimism ensures that the inclusion of unlabelled data point cannot make the model accuracy worse than just using the labelled data.

In practice, his procedure pessimistically assign soft labels to the unlabelled data, such that the improvement over the supervised version is minimal; at the same time maximize log likelihood over labelled data. That is, the parameters θ_{semi} of a semi-supervised model in the CPLE framework are given by

$$\theta_{semi} = \arg\max_{\theta} \arg\min_{q} L(\theta|X,U,q) - L(\theta_{sup}|X,U,q), \tag{E.3}$$

where $X = (x_i, y_i)_{i=1}^{N}$ denotes the labelled data, U the unlabelled data, and q the hypothetical soft labels. (Loog, 2015) only provides a solution for Linear Discriminant Analysis (LDA), and his framework requires an explicit generative likelihood $L(\theta|X,U,q)$.

However, his framework can be modified to use discriminative likelihoods instead. This allows using both generative and discriminative classifiers in this framework, provided that they can output a prediction probability (such probability estimates can also be provided by Platt scaling (Platt et al., 1999) for classifiers which don't support them).

Figure E.2 shows the objective function this extended CPLE model, supporting pessimistic semi-supervised learning with any constituent model. Instead of optimizing the generative likelihood L, the negative log loss $J(y,r) = logp(y|r) = \frac{1}{N}\sum_{i=1}^{N}(y_i \log(r_i) + (1-y_i)\log(1-r_i))$ is used in the optimization objective, where y_i is the predicted class of the i'th instance, and r_i is the probability associated with this prediction.

The modified CPLE objective function in Figure E.2 can be used with any global optimization approach to train a model to fit the unlabelled data distribution in a pessimistic fashion, while maximizing performance over the labelled data. We used the locally biased variant (Gablonsky & Kelley, 2001) of DIRECT (DIviding RECTangles)

Algorithm E.2.1: CPLEOBJECTIVEFUNCTION($model, \theta, q, X, y, U$)

$$1 : \text{set } z_i \leftarrow \begin{cases} 1 & \text{if } q_i \geq 0.5 \\ 0 & \text{otherwise} \end{cases} \text{ for all } i \in [1, |U|]$$
$$2 : X' \leftarrow X \cup U$$
$$3 : y' \leftarrow y \cup z$$
$$4 : \text{set } w_i \leftarrow \begin{cases} 1 & \text{if } X'_i \in X \\ q_i & \text{otherwise} \end{cases} \text{ for all } i \in [1, |X'|]$$
$$5 : model \leftarrow train(model, X', y', w)$$
$$6 : r \leftarrow predictionprobability(model, X)$$
$$7 : r' \leftarrow predictionprobability(model, U)$$
$$8 : J_{labelled} \leftarrow \frac{1}{|X|} \sum_{i=1}^{|X|} (y_i \log(r_i) + (1 - y_i) \log(1 - r_i))$$
$$9 : J_{unlabelled} \leftarrow \frac{1}{|U|} \sum_{i=1}^{|U|} (z_i \log(r'_i) + (1 - z_i) \log(1 - r'_i))$$
$$10 : return(J_{unlabelled} - J_{labelled})$$

Figure E.2: A general pessimistic semi-supervised learning framework, based on a generalization of (Loog, 2015) to use discriminative likelihoods and to be usable with any classifier supporting prediction probabilities. It requires a constituent model, current hypothetical labels q of the unlabelled data points and model parameters θ which are to be optimized, the unlabelled data U, labelled data X, labels y of X, and the negative log loss $J_{labelled}$ of the model trained only on X. Minimizing this objective increases accuracy over the labelled data while assuming worst-case labels for the unlabelled data.

(Jones et al., 1993), a global, deterministic, derivative-free optimization method based on Lipschitzian optimization, which can handle the kinds of non-linear and non-convex functions constituted by Figure E.2.

The resulting semi-supervised learning framework is highly computationally expensive, but has the advantages of being a generally applicable framework, needing low memory, and making no additional assumptions except for the ones made by the chosen classifier. It has been made freely available online at `https://github.com/tmadl/semisup-learn`, together with a more detailed description.

E.3 Constituent models

In principle, model able to produce probability estimates could be used as the constituent model in (E.2) to model the linkage probability distribution. Here, we focus

on three models of this class: Gaussian Discriminant Analysis (GDA), Gaussian Process (GP), and Random Forest (RF). The first (GDA) has a closed form solution, thus constituting the - to our knowledge - first probabilistically motivated approach to learning a non-linear distance metric in closed form.

In the case of the **GDA** (Bensmail & Celeux, 1996), the likelihoods of a pair of instances belonging to the same cluster $p(\Delta\mathbf{x}|l=1;\mu_1,\Sigma_1)$ or to different clusters $p(\Delta\mathbf{x}|l=0;\mu_0,\Sigma_0)$, are modelled using multivariate Gaussians:

$$p(\Delta\mathbf{x}|l=i;\mu_i,\Sigma_i) = (2\pi)^{-\frac{D}{2}}|\mathbf{\Sigma}_i|^{-\frac{1}{2}}\, e^{-\frac{1}{2}(\Delta\mathbf{x}-\boldsymbol{\mu}_i)^\intercal\mathbf{\Sigma}_i^{-1}(\Delta\mathbf{x}-\boldsymbol{\mu}_i)}, \tag{E.4}$$

where $i \in \{0,1\}$. (μ_1,Σ_1) are the means and covariances of the APD distances of pairs in the same cluster, and (μ_0,Σ_0) those in different clusters. These parameters can be easily estimated from the instances corresponding to the must-link and cannot-link constraints, respectively, by calculating their means and covariances.

In the case of **GP**, we model the linkage probability distribution using a Gaussian Process with mean function $m(\cdot)$ and covariance function $k(\cdot)$:

$$p(\Delta\mathbf{x}|l=i;M,K) = \mathcal{GP}(m(\cdot),k(\cdot)) = \mathcal{N}\left(\begin{bmatrix} m(x_1) \\ \vdots \\ m(x_m) \end{bmatrix}, \begin{bmatrix} k(x_1,x_1),\ldots,k(x_1,x_m) \\ \vdots \quad \ddots \quad \vdots \\ k(x_m,x_1),\ldots,k(x_m,x_m) \end{bmatrix}\right) \tag{E.5}$$

In the experiments below, we have used a zero mean function and the square exponential kernel as the covariance function. See (Rasmussen & Williams, 2005) for an extensive book on Gaussian Processes.

Finally, in the case of **RF**, an ensemble of B decision trees is learned. Each decision tree is trained on a bootstrap resample of data (drawn with replacement), and only considering a randomly selected subset of the available features. See (Breiman, 2001) for details. There are several ways to obtain class probabilities from trained random forests (Boström, 2007). We used the relative class frequency, i.e. the averaged fractions of samples falling on that same class in the leaves of individual trees:

$$p(\Delta\mathbf{x}|l=i;M,K) = \frac{1}{B}\sum_{j=1}^{B}\frac{l(t_j,\Delta\mathbf{x},i)}{l(t_j,\Delta\mathbf{x},i)+l(t_j,\Delta\mathbf{x},1-i)}, \tag{E.6}$$

where t_j denotes the individual trained trees, and $l(t_j,\Delta\mathbf{x},i)$ is a function returning the number of training examples falling into the same leaf as $\Delta\mathbf{x}$ in tree t_j.

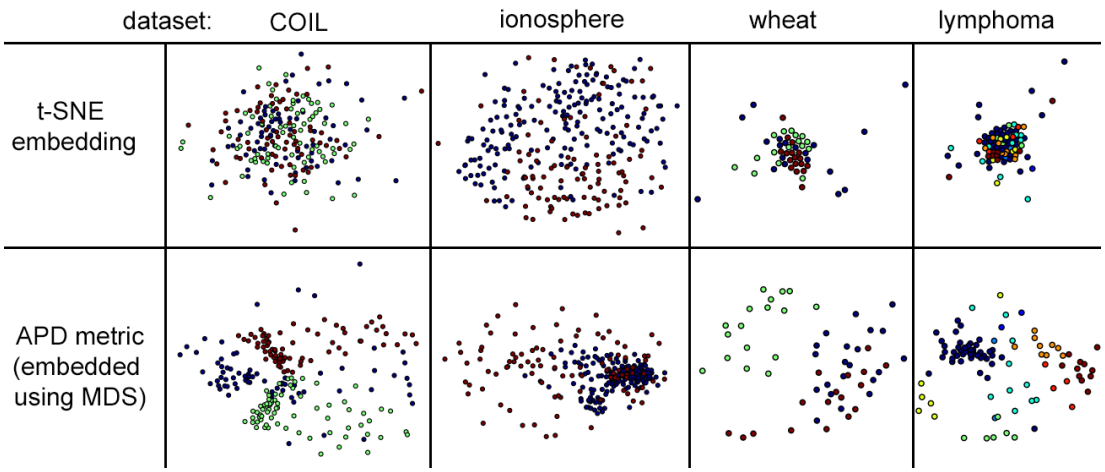


Figure E.3: Embedding of pairwise distances as suggested by our metric learning approach (top) and by t-SNE (bottom) on the COIL-3 dataset (128x128px images of 3 similar-looking toy cars) used by (Zeng & Cheung, 2012), the ionosphere and wheat datasets from the UCI machine learning repository, and the lymphoma microarray dataset obtained from the Broad Institute. The APD plot was created using supervised GDA as the constituent model.

E.4 Preliminary results

We will first assume a purely supervised case - plugging in a GDA model (Equation (E.4)) into the APD metric framework (Equation (E.2)), as used to model human spatial representation structure in Chapters 2.5 and 5. We have claimed in Chapter 2.5 that our metric is well-suited for clustering, in the sense that for most within-cluster differences $\Delta \boldsymbol{x}_r$ and across-cluster differences $\Delta \boldsymbol{x}_n$ it holds that $d_r(\boldsymbol{x}_{r,1}, \mathbf{x}_{r,2}; \boldsymbol{\theta}) < d_n(\boldsymbol{x}_{n,1}, \mathbf{x}_{n,2}; \boldsymbol{\theta})$.

Figure E.3 shows distances under our metric in three example datasets, embedded into two dimensions for visualization using MDS. It contrasts these distances with embeddings by t-SNE (Van der Maaten & Hinton, 2008), a state of the art method for visualization, designed to represent high-dimensional data points in such a way that similar objects are modelled by nearby points, and dissimilar objects by distant points. Figure E.3 shows that the APD metric 'pulls together' different clusters (shown using different colours) more effectively than t-SNE.

The Figure is intended to show that datasets with no clear intrinsic cluster structure can be made separable by learning in APD space given only a few pairwise constraints (20 in these examples, which is less than 0.1% of all pairwise constraints), even with a fully supervised constituent model. It is not intended as a fair comparison, as t-SNE is an unsupervised method.

Clustering results on benchmark datasets

We have compared the described metric learning framework in APD space against several state of the art semi-supervised clustering models on multiple datasets. The constituent models in this section were trained in a semi-supervised fashion. In the case of GDA, we applied the extended CPLE learning principle shown in Figure E.2 in APD space. In the case of the LS (label spreading) constituent model, we applied the learning algorithm by (Zhou et al., 2004), which can be described as a spreading activation model, labelling unlabelled data points based on their closest neighbours.

Tables E.1 and E.2 summarize the performance of semi-supervised clustering with APD metrics using these constituent models, compared to other recent approaches. In all cases, our model learned the metric in APD space, inferred and embedded all pairwise distances based on this metric, and then performed K-Means clustering.

dataset	cons-traints	DCA+ KMeans	Kmeans	MPC KMeans	SS-MMC	**APD-GDA**	**APD-LS**
pendigit	40	0.328	0.545	0.556	0.743	0.554	**0.874**
	60	0.553	0.545	0.535	0.831	0.717	**0.855**
	80	0.677	0.545	0.58	0.833	0.769	**0.901**
	100	0.721	0.545	0.537	0.861	0.8	**0.878**
letterIJL	40	0.207	0.266	0.223	**0.546**	0.194	0.467
	60	0.352	0.266	0.226	**0.691**	0.3	0.431
	80	0.4	0.266	0.167	**0.568**	0.417	0.329
	100	0.582	0.266	0.211	**0.655**	0.517	0.571
vehicle	100	0.233	0.186	0.122	**0.277**	0.161	0.208
	150	0.331	0.186	0.122	**0.365**	0.248	0.215
	200	0.418	0.186	0.102	**0.411**	0.355	0.288
	250	0.459	0.186	0.113	**0.458**	0.351	0.26
ionosphere	40	0.205	0.135	0.101	**0.242**	0.181	0.174
	60	0.167	0.135	0.094	**0.304**	0.293	0.148
	80	0.197	0.135	0.078	0.297	**0.332**	0.216
	100	0.199	0.135	0.094	0.313	**0.419**	0.282

Table E.1: Semi-supervised clustering comparison - many constraints ($10^1 - 10^3$)**.** The table shows results of the APD-space metric learning framework applied to clustering, with two constituent models, one using GDA and one using label spreading (LS) (Zhou et al., 2004). Compared methods: Discriminative Component Analysis (DCA)+KMeans (Hoi et al., 2006), KMeans, MPCKmeans (Bilenko et al., 2004), and Semi-Supervised Maximum Margin Clustering (Zeng & Cheung, 2012). Non-APD model performances were taken from (Zeng & Cheung, 2012).

Dataset	*MPC*-Kmeans	CP-Kmeans	CNP-Kmeans	PCA-Kmeans	GP-Kmeans	LDA-Kmeans	APD+QDA	APD+LS
pima	0.6021±0.0028	0.6406±0.0018	0.6602±0.0014	0.6602±0.0000	0.6602±0.0026	**0.6848**±0.0041	0.6391±0.0088	OOM
iris	0.9000±0.0014	0.9133±0.0022	0.9200±0.0026	0.9133±0.0008	0.9000±0.0042	**0.9400**±0.0026	0.6133±0.0430	0.8687
wdbc	**0.8541**±0.0000	**0.8541**±0.0000	**0.8541**±0.0000	**0.8541**±0.0000	**0.8541**±0.0000	0.7326±0.0012	0.7976±0.0300	OOM
wine	**0.9438**±0.0027	0.6651±0.0012	0.7047±0.0019	0.7022±0.0008	0.9123±0.0021	0.9123±0.0028	0.6413±0.0156	0.5729
glass	0.5286±0.0202	0.5467±0.0010	**0.5561**±0.0122	0.5187±0.0091	0.5324±0.0216	0.5066±0.0082	0.4610±0.0127	**0.9109**
bupa	0.5449±0.0032	**0.5739**±0.0022	0.5536±0.0041	0.5536±0.0021	0.5521±0.0025	0.5159±0.0126	**0.5540**±0.0045	0.3785
balance	0.5136±0.0046	0.6160±0.0076	**0.6192**±0.0080	0.5088±0.0094	0.5246±0.0062	0.5376±0.0026	0.4956±0.0129	0.5884
spectheart	**0.6142**±0.0180	0.6067±0.0032	0.6060±0.0026	0.5543±0.0049	0.6043±0.0061	0.5032±0.0082	**0.6250**±0.0165	**0.7378**
Hepatitis	0.5935±0.0021	0.6129±0.0042	**0.6226**±0.0061	0.5935±0.0082	0.6210±0.0077	0.6181±0.0024	**0.6219**±0.0160	0.651

Table E.2: Semi-supervised clustering comparison - few constraints ($10^0 - 10^1$)**.** The table shows results of the APD-space metric learning framework applied to clustering, with two constituent models, one using GDA and one using label spreading (LS) (Zhou et al., 2004). Compared methods: MPC-Kmeans (Bilenko et al., 2004), Constraint Projection (CP)-Kmeans (Tang et al., 2007), Constraint Neighbourhood Projection (CNP)-Kmeans (Wang et al., 2014), Principal Component Analysis (PCA)-Kmeans, GP-Kmeans (Eaton, 2005) and linear discriminant analysis (LDA)-Kmeans (Ding & Li, 2007). Non-APD model performances were taken from (Wang et al., 2014). 'OOM' stands for a model running out of memory.

Clustering results on cancer microarray data

Finally, we tested the other described constituent models on a bioinformatics dataset, to show its utility on non-benchmark problems with real-world relevance. It was recently proposed that semi-supervised consensus clustering (SSCC) outperforms the state of the art on cancer microarray data (Wang & Pan, 2014). We compared our APD-space based clustering with this ensemble clustering method, using the GP and RF constituent models.

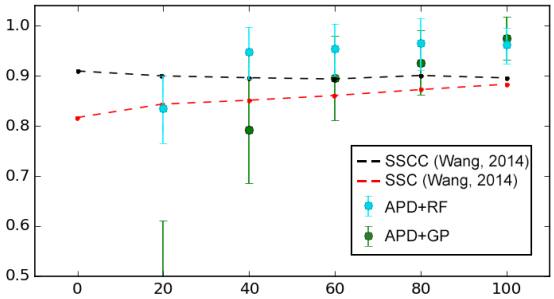


Figure E.4: Semi-supervised clustering results on cancer microarray data from (Golub et al., 1999), compared to the state of the art semi-supervised consensus clustering method by (Wang & Pan, 2014).

Figure E.4 shows the results on the Leukemia dataset, published and made available by (Golub et al., 1999). Despite the SSCC being an ensemble method, the APD-space methods significantly outperform it when given a sufficient number of constraints.

References for Appendix E

Baghshah, M. S., & Shouraki, S. B. (2010). Kernel-based metric learning for semi-supervised clustering. *Neurocomputing*, *73*, 1352–1361.

Bensmail, H., & Celeux, G. (1996). Regularized gaussian discriminant analysis through eigenvalue decomposition. *Journal of the American statistical Association*, *91*, 1743–1748.

Bilenko, M., Basu, S., & Mooney, R. J. (2004). Integrating constraints and metric learning in semi-supervised clustering. In *Proceedings of the twenty-first international conference on Machine learning* (p. 11). ACM.

Boström, H. (2007). Estimating class probabilities in random forests. In *Machine Learning and Applications, 2007. ICMLA 2007. Sixth International Conference on* (pp. 211–216). IEEE.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Brin, S. (1995). Near neighbor search in large metric spaces. In U. Dayal, P. M. D. Gray, & S. Nishio (Eds.), *VLDB '95: proceedings of the 21st International Conference on Very Large Data Bases, Zurich, Switzerland, Sept. 11–15, 1995* (pp. 574–584). Los Altos, CA 94022, USA: Morgan Kaufmann Publishers.

Chitta, R., Jin, R., Havens, T. C., & Jain, A. K. (2011). Approximate kernel k-means: Solution to large scale kernel clustering. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 895–903). ACM.

Ding, C., & Li, T. (2007). Adaptive dimension reduction using discriminant analysis and k-means clustering. In *Proceedings of the 24th international conference on Machine learning* (pp. 521–528). ACM.

Eaton, E. R. (2005). *Clustering with Propagated Constraints*. Ph.D. thesis Citeseer.

Gablonsky, J. M., & Kelley, C. T. (2001). A locally-biased form of the direct algorithm. *Journal of Global Optimization*, *21*, 27–37.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A. et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, *286*, 531–537.

Hoi, S. C., Liu, W., Lyu, M. R., & Ma, W.-Y. (2006). Learning distance metrics with contextual constraints for image retrieval. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on* (pp. 2072–2078). IEEE volume 2.

Jones, D. R., Perttunen, C. D., & Stuckman, B. E. (1993). Lipschitzian optimization without the lipschitz constant. *Journal of Optimization Theory and Applications*, *79*, 157–181.

Loog, M. (2015). Contrastive pessimistic likelihood estimation for semi-supervised classification. *arXiv preprint arXiv:1503.00269*, .

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, *9*, 85.

Ong, C. S., Williamson, R. C., & Smola, A. J. (2005). Learning the kernel with hyperkernels. In *Journal of Machine Learning Research* (pp. 1043–1071).

Platt, J. et al. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, *10*, 61–74.

Rasmussen, C. E., & Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning*. The MIT Press.

Tang, W., Xiong, H., Zhong, S., & Wu, J. (2007). Enhancing semi-supervised clustering: a feature projection perspective. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 707–716). ACM.

Wang, H., Li, T., Li, T., & Yang, Y. (2014). Constraint neighborhood projections for semi-supervised clustering. *Cybernetics, IEEE Transactions on*, *44*, 636–643.

Wang, Y., & Pan, Y. (2014). Semi-supervised consensus clustering for gene expression data analysis. *BioData mining*, *7*, 1–13.

Weinberger, K. Q., Blitzer, J., & Saul, L. K. (2005). Distance metric learning for large margin nearest neighbor classification. In *Advances in neural information processing systems* (pp. 1473–1480).

Xing, E. P., Jordan, M. I., Russell, S., & Ng, A. Y. (2002). Distance metric learning with application to clustering with side-information. In *Advances in neural information processing systems* (pp. 505–512).

Xing, E. P., Jordan, M. I., Russell, S., & Ng, A. Y. (2012). Similarity learning for provably accurate sparse linear classication. In *Proceedings of the 29th International Conference on Machine Learning (ICML)* (pp. 1871–1878).

Zeng, H., & Cheung, Y.-m. (2012). Semi-supervised maximum margin clustering with pairwise constraints. *Knowledge and Data Engineering, IEEE Transactions on*, *24*, 926–939.

Zhou, D., Bousquet, O., Lal, T. N., Weston, J., & Schölkopf, B. (2004). Learning with local and global consistency. *Advances in neural information processing systems*, *16*, 321–328.