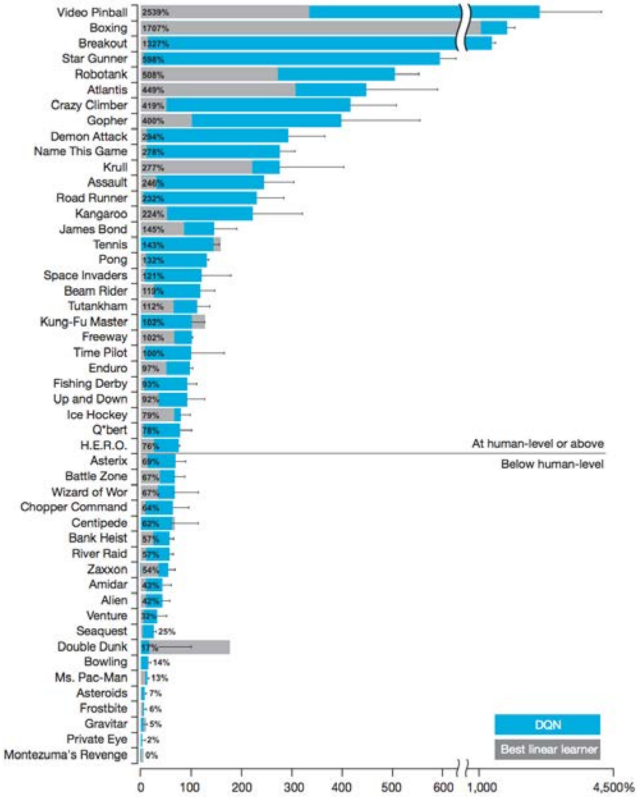
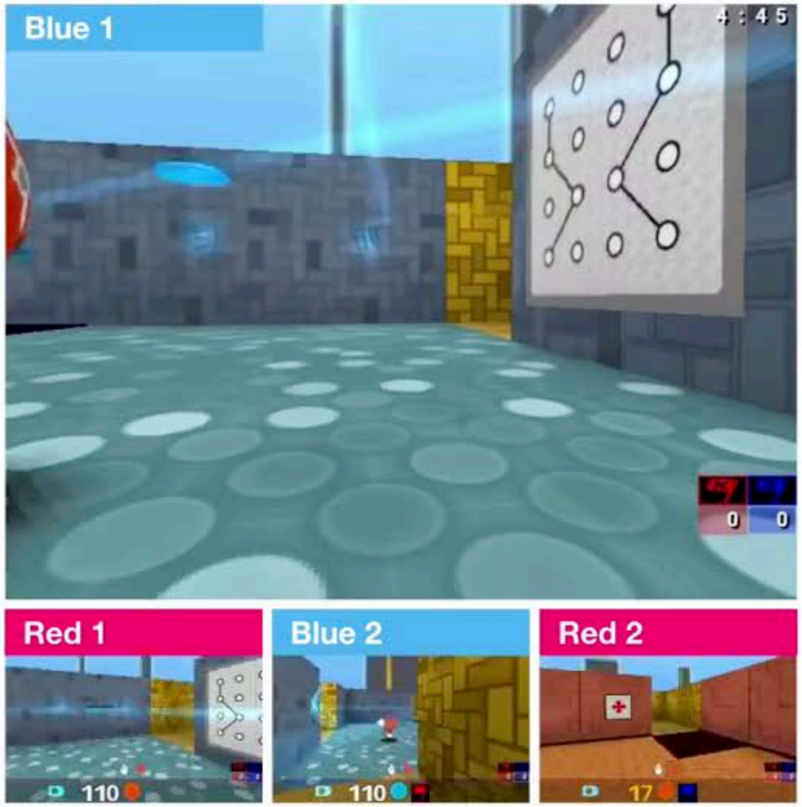
Toward object-oriented deep reinforcement learning

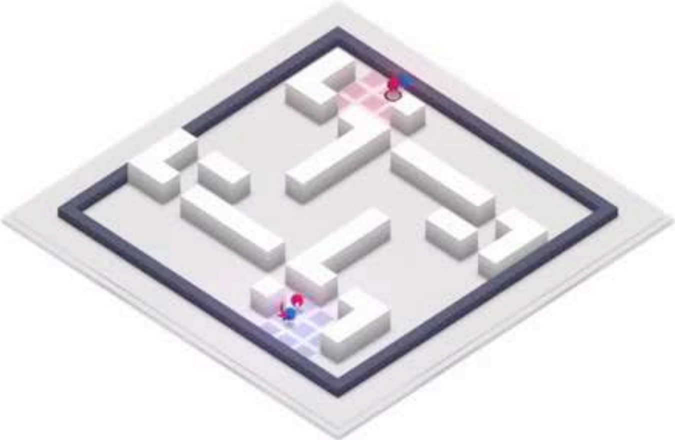
Matthew Botvinick
DeepMind, London UK
Gatsby Computational Neuroscience Unit, UCL



Mnih *et al, Nature* (2015)

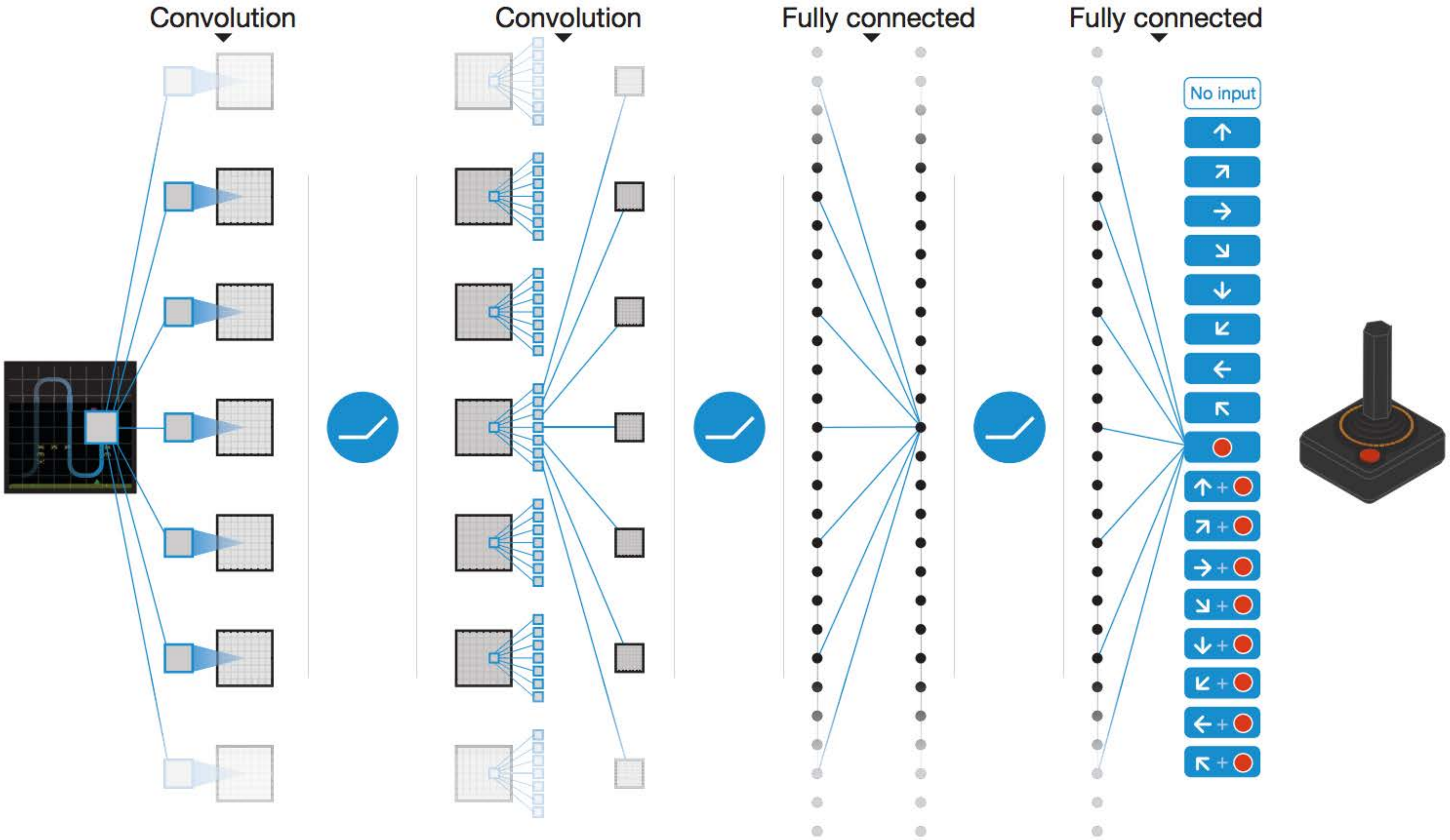
Agent observation raw pixels



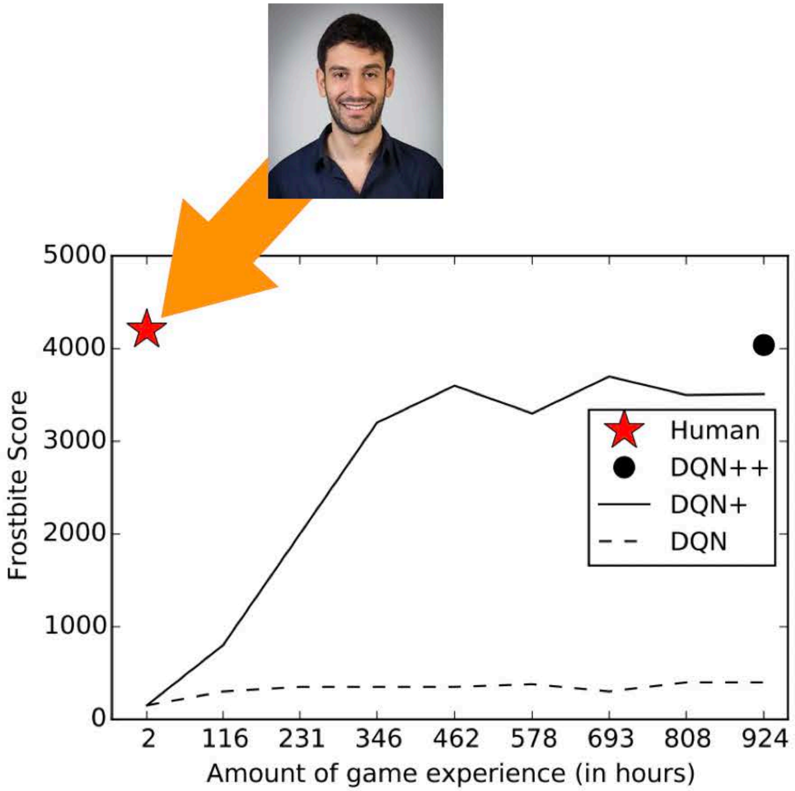


Indoor map overview

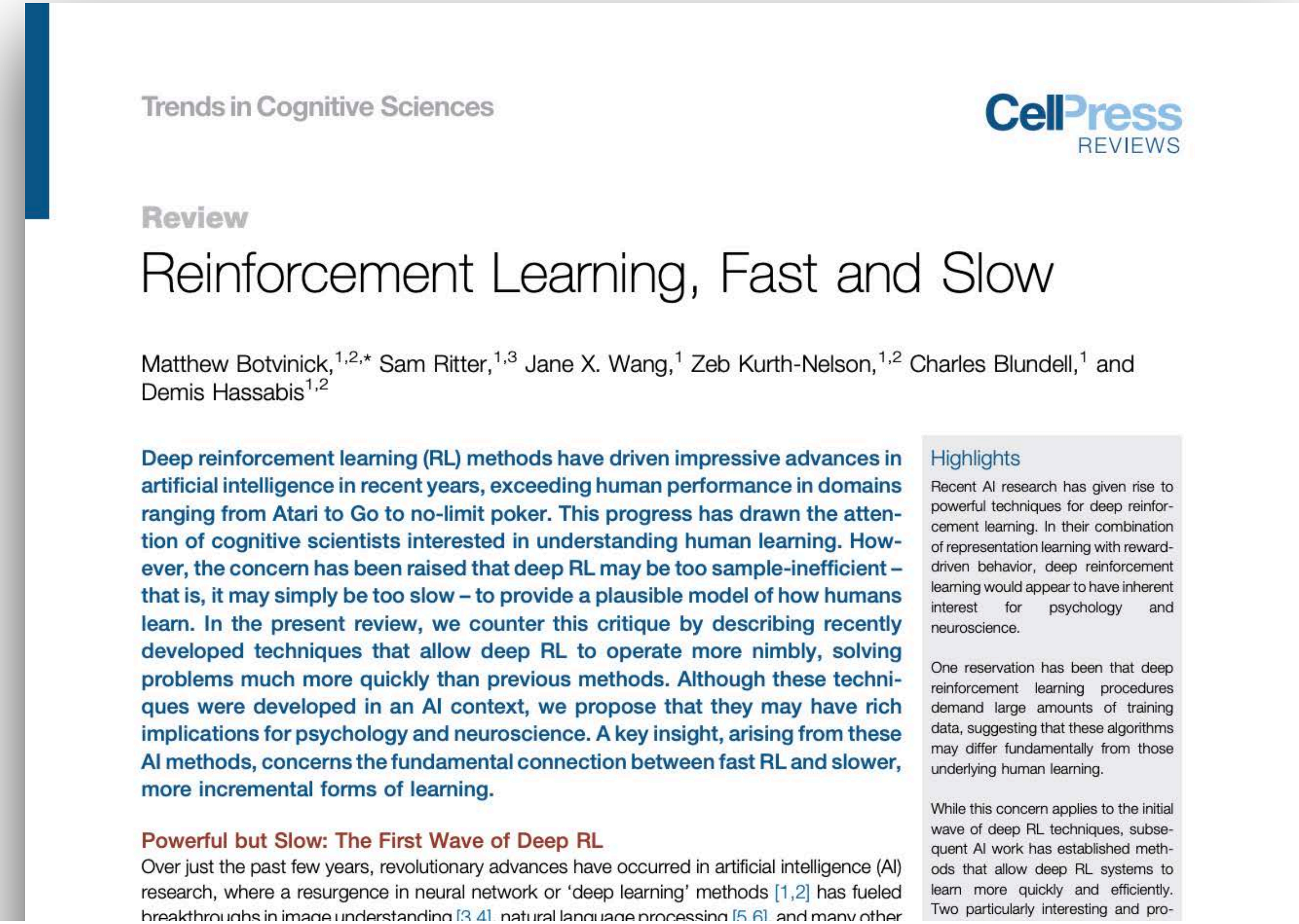
Jaderberg et al., *Science*, 2019



Mnih *et al, Nature* (2015)



Lake *et al, BBS* (2017)



Trends in Cognitive Sciences

CellPress
REVIEWS

Review

Reinforcement Learning, Fast and Slow

Matthew Botvinick,[1,2,*] Sam Ritter,[1,3] Jane X. Wang,[1] Zeb Kurth-Nelson,[1,2] Charles Blundell,[1] and Demis Hassabis[1,2]

Deep reinforcement learning (RL) methods have driven impressive advances in artificial intelligence in recent years, exceeding human performance in domains ranging from Atari to Go to no-limit poker. This progress has drawn the attention of cognitive scientists interested in understanding human learning. However, the concern has been raised that deep RL may be too sample-inefficient – that is, it may simply be too slow – to provide a plausible model of how humans learn. In the present review, we counter this critique by describing recently developed techniques that allow deep RL to operate more nimbly, solving problems much more quickly than previous methods. Although these techniques were developed in an AI context, we propose that they may have rich implications for psychology and neuroscience. A key insight, arising from these AI methods, concerns the fundamental connection between fast RL and slower, more incremental forms of learning.

Powerful but Slow: The First Wave of Deep RL

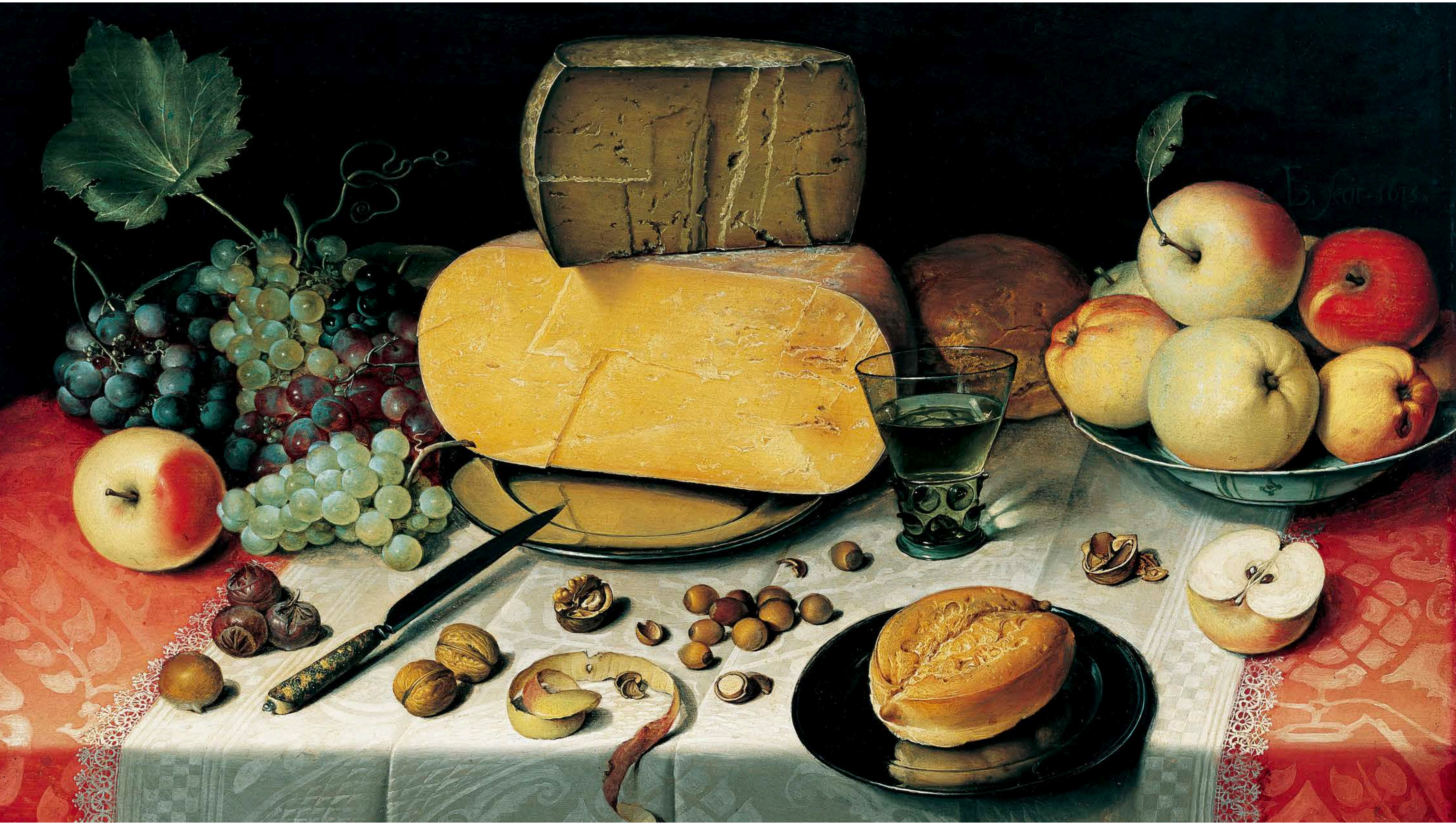
Over just the past few years, revolutionary advances have occurred in artificial intelligence (AI) research, where a resurgence in neural network or 'deep learning' methods [1,2] has fueled breakthroughs in image understanding [3,4], natural language processing [5,6], and many other

Highlights

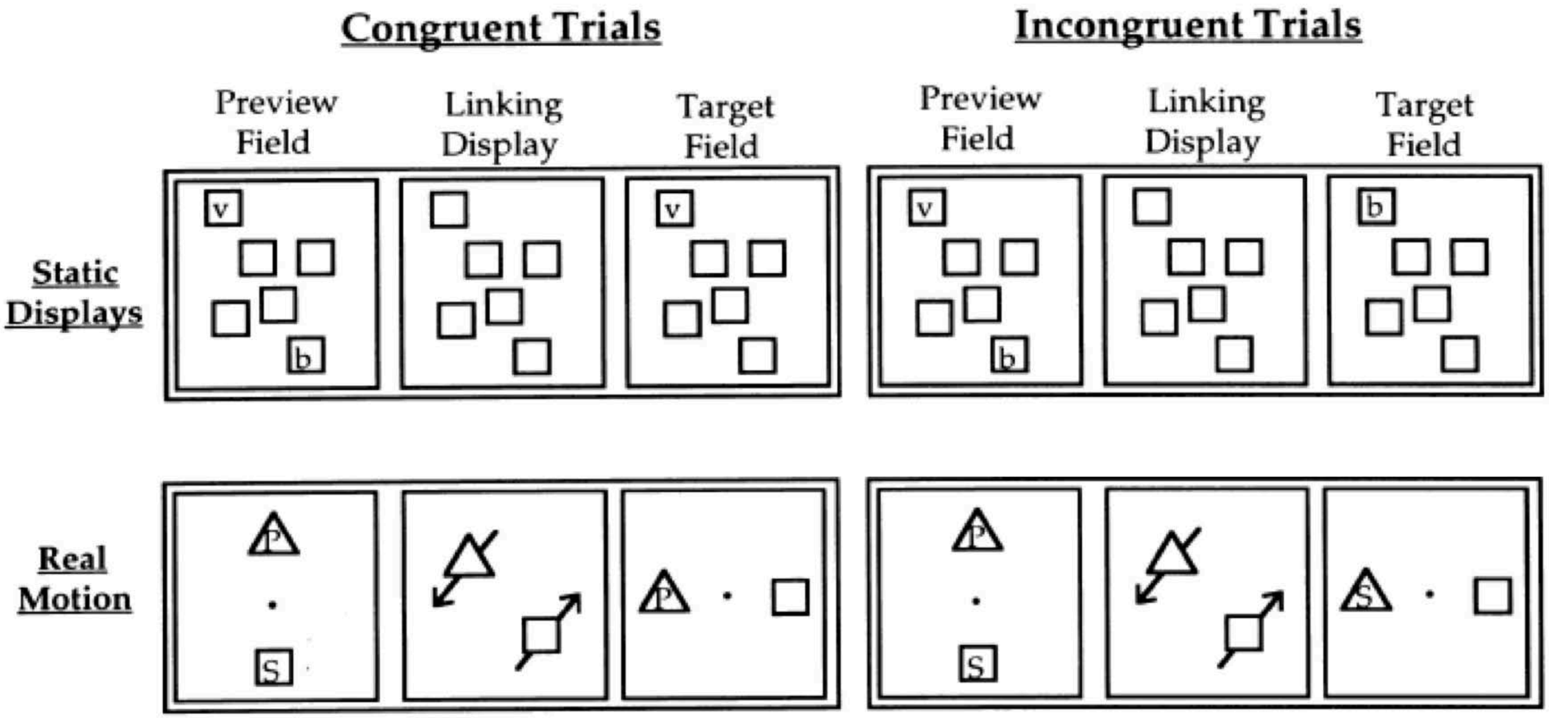
Recent AI research has given rise to powerful techniques for deep reinforcement learning. In their combination of representation learning with reward-driven behavior, deep reinforcement learning would appear to have inherent interest for psychology and neuroscience.

One reservation has been that deep reinforcement learning procedures demand large amounts of training data, suggesting that these algorithms may differ fundamentally from those underlying human learning.

While this concern applies to the initial wave of deep RL techniques, subsequent AI work has established methods that allow deep RL systems to learn more quickly and efficiently. Two particularly interesting and pro-

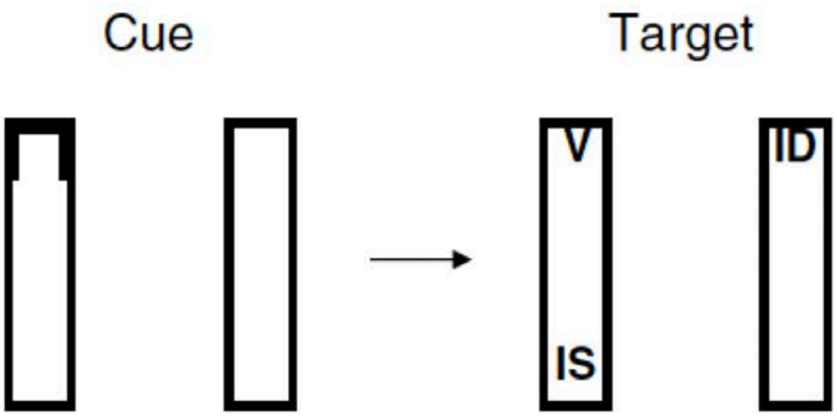

FS. fecit 1615

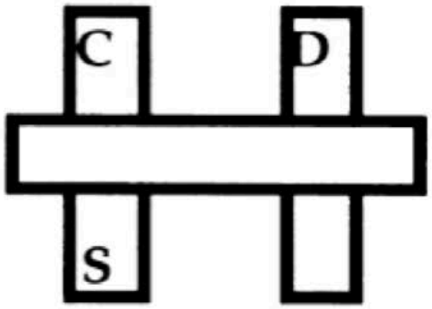
Object-specific Priming



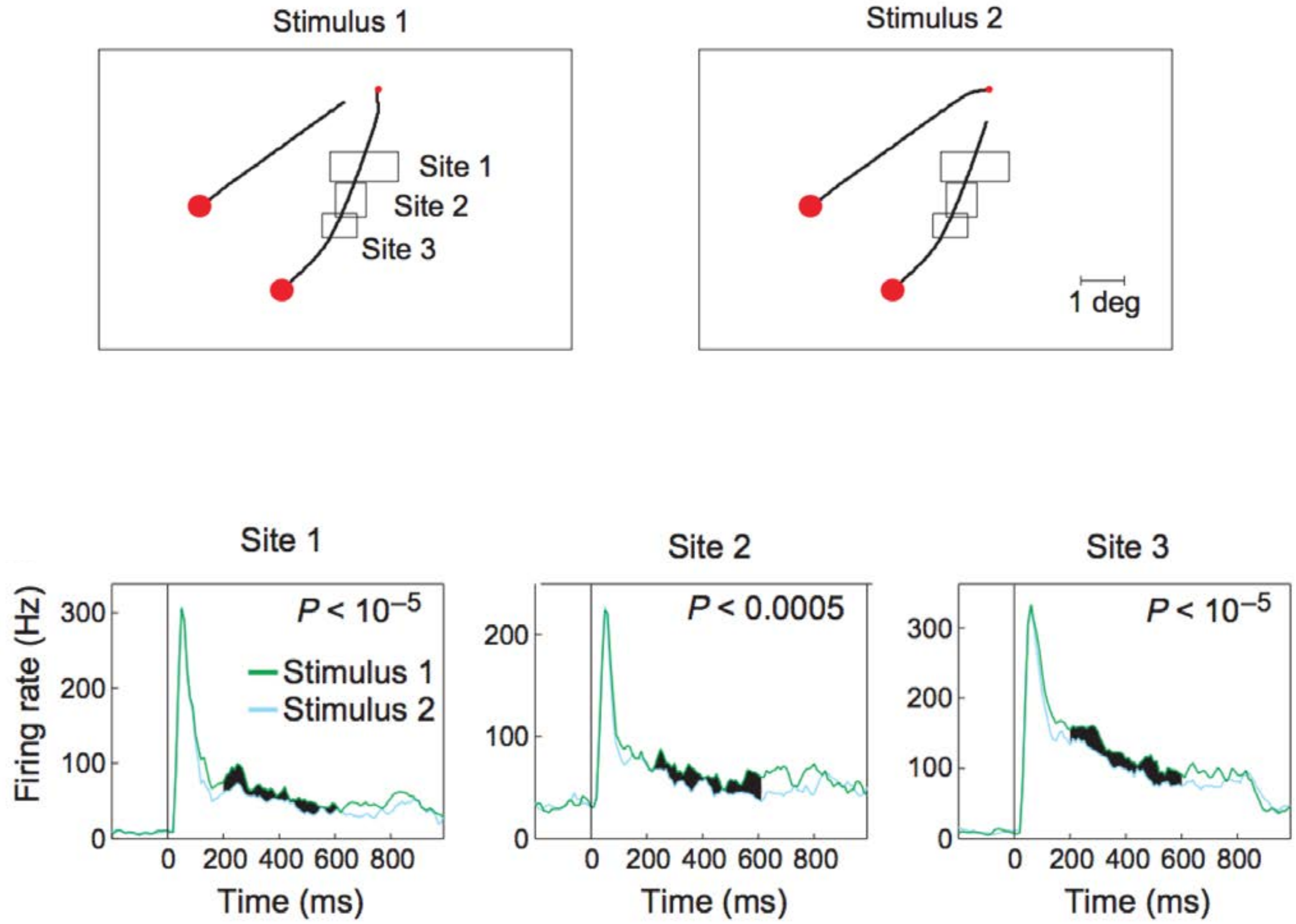
Kahneman et al., 1992

Automatic spread of attention

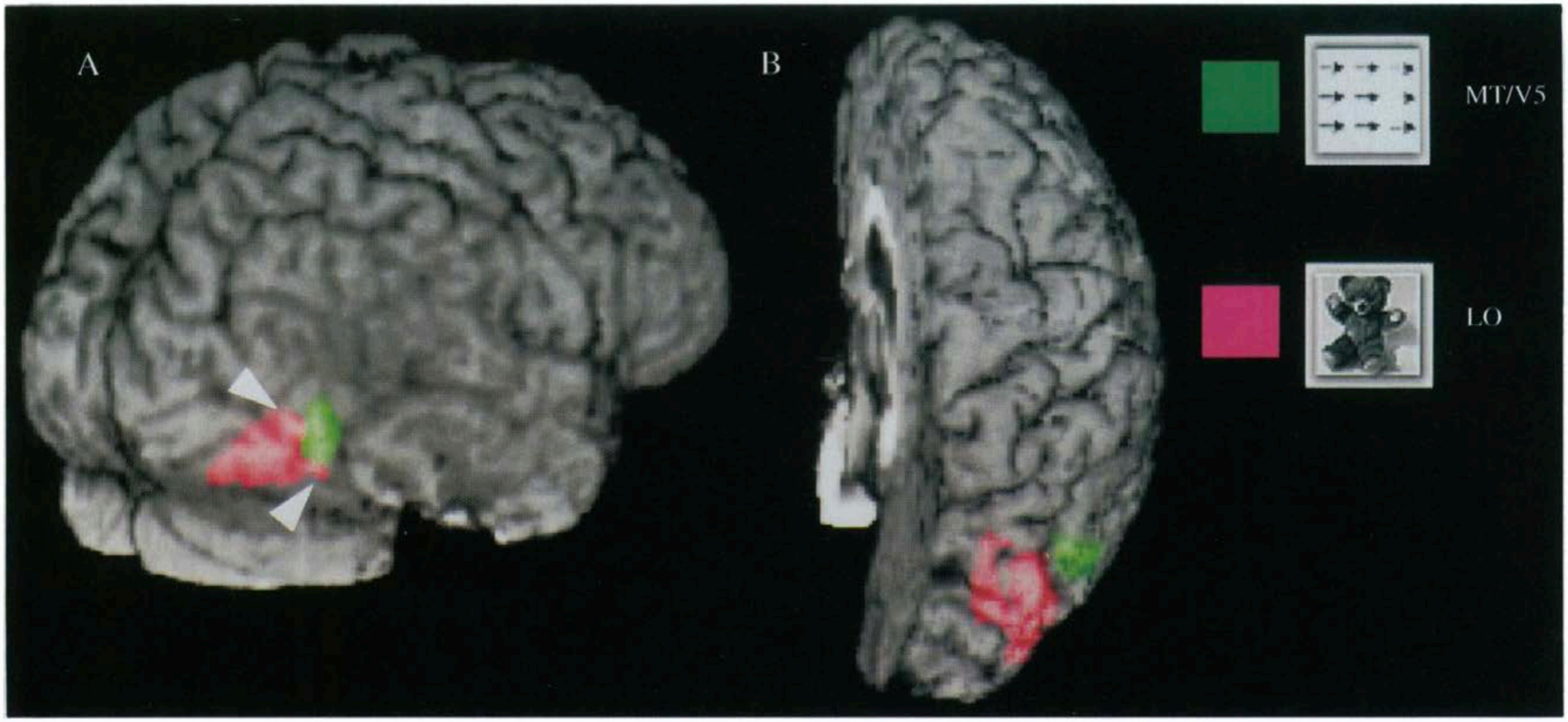




Egly, Driver, and Rafal (1994);
Moore, Yantis, and Vaughan (1998)

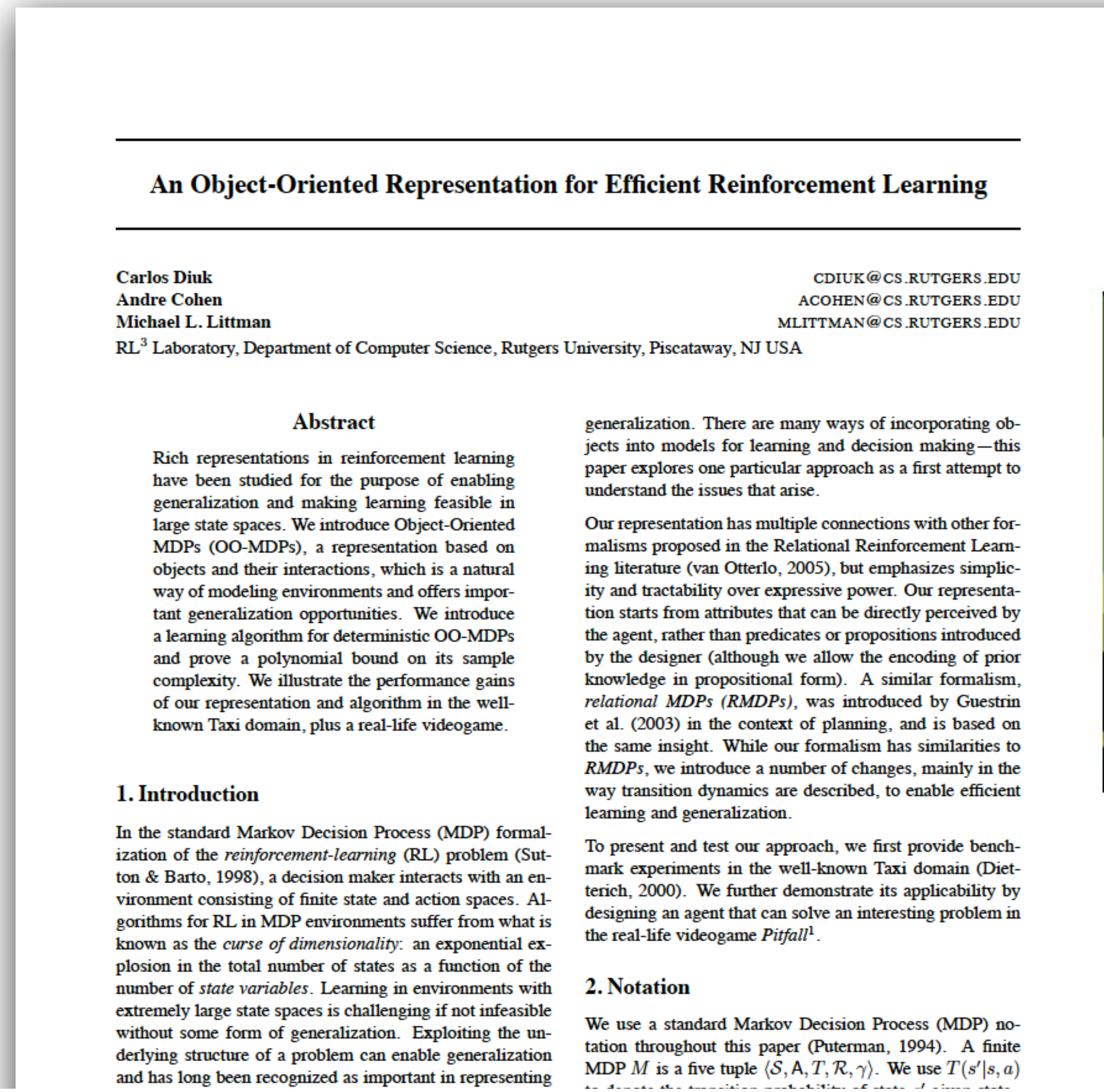


Roelfsema et al. *Nature*, 1998



Malach et al., *PNAS,* 1995

FB. fecit 1615



An Object-Oriented Representation for Efficient Reinforcement Learning

Carlos Diuk CDIUK@CS.RUTGERS.EDU
Andre Cohen ACOHEN@CS.RUTGERS.EDU
Michael L. Littman MLITTMAN@CS.RUTGERS.EDU
RL[3] Laboratory, Department of Computer Science, Rutgers University, Piscataway, NJ USA

Abstract

Rich representations in reinforcement learning have been studied for the purpose of enabling generalization and making learning feasible in large state spaces. We introduce Object-Oriented MDPs (OO-MDPs), a representation based on objects and their interactions, which is a natural way of modeling environments and offers important generalization opportunities. We introduce a learning algorithm for deterministic OO-MDPs and prove a polynomial bound on its sample complexity. We illustrate the performance gains of our representation and algorithm in the well-known Taxi domain, plus a real-life videogame.

1. Introduction

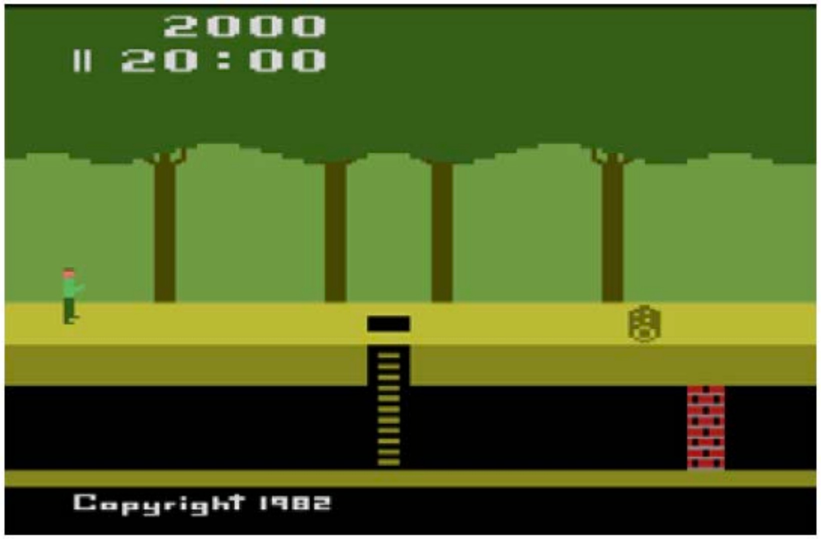
In the standard Markov Decision Process (MDP) formalization of the *reinforcement-learning* (RL) problem (Sutton & Barto, 1998), a decision maker interacts with an environment consisting of finite state and action spaces. Algorithms for RL in MDP environments suffer from what is known as the *curse of dimensionality*: an exponential explosion in the total number of states as a function of the number of *state variables*. Learning in environments with extremely large state spaces is challenging if not infeasible without some form of generalization. Exploiting the underlying structure of a problem can enable generalization and has long been recognized as important in representing generalization. There are many ways of incorporating objects into models for learning and decision making—this paper explores one particular approach as a first attempt to understand the issues that arise.

Our representation has multiple connections with other formalisms proposed in the Relational Reinforcement Learning literature (van Otterlo, 2005), but emphasizes simplicity and tractability over expressive power. Our representation starts from attributes that can be directly perceived by the agent, rather than predicates or propositions introduced by the designer (although we allow the encoding of prior knowledge in propositional form). A similar formalism, *relational MDPs (RMDPs)*, was introduced by Guestrin et al. (2003) in the context of planning, and is based on the same insight. While our formalism has similarities to *RMDPs*, we introduce a number of changes, mainly in the way transition dynamics are described, to enable efficient learning and generalization.

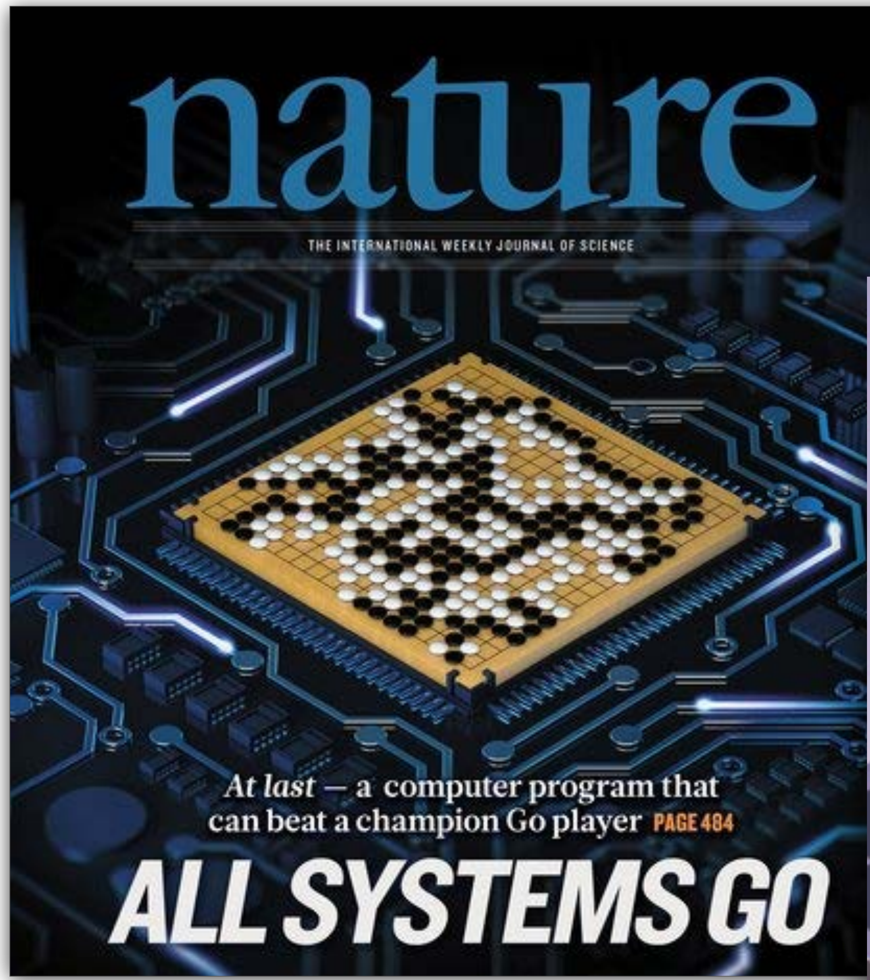
To present and test our approach, we first provide benchmark experiments in the well-known Taxi domain (Dietterich, 2000). We further demonstrate its applicability by designing an agent that can solve an interesting problem in the real-life videogame *Pitfall*[1].

2. Notation

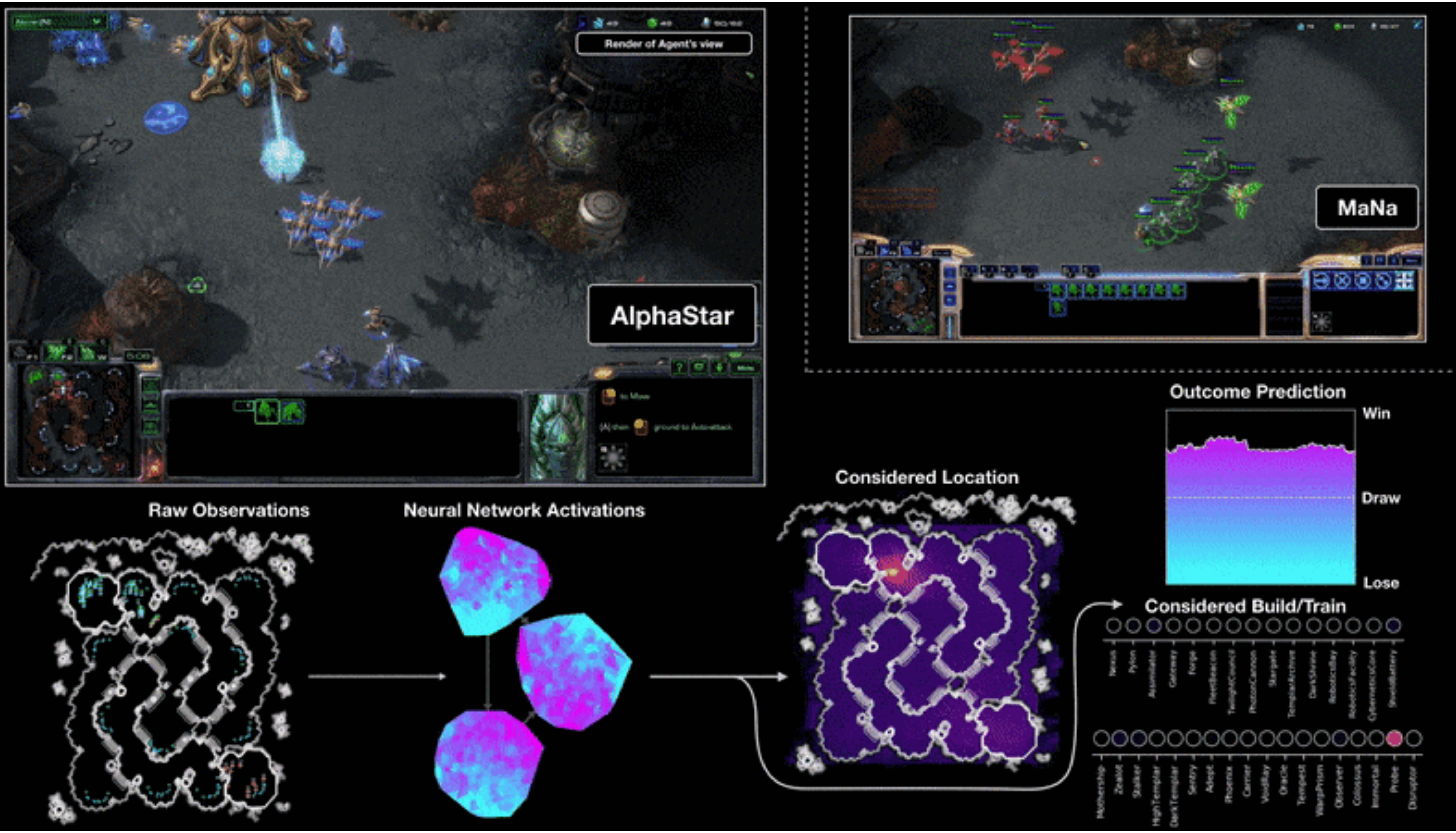
We use a standard Markov Decision Process (MDP) notation throughout this paper (Puterman, 1994). A finite MDP M is a five tuple $\langle \mathcal{S}, \mathrm{A}, T, \mathcal{R}, \gamma\rangle$. We use $T(s'|s,a)$

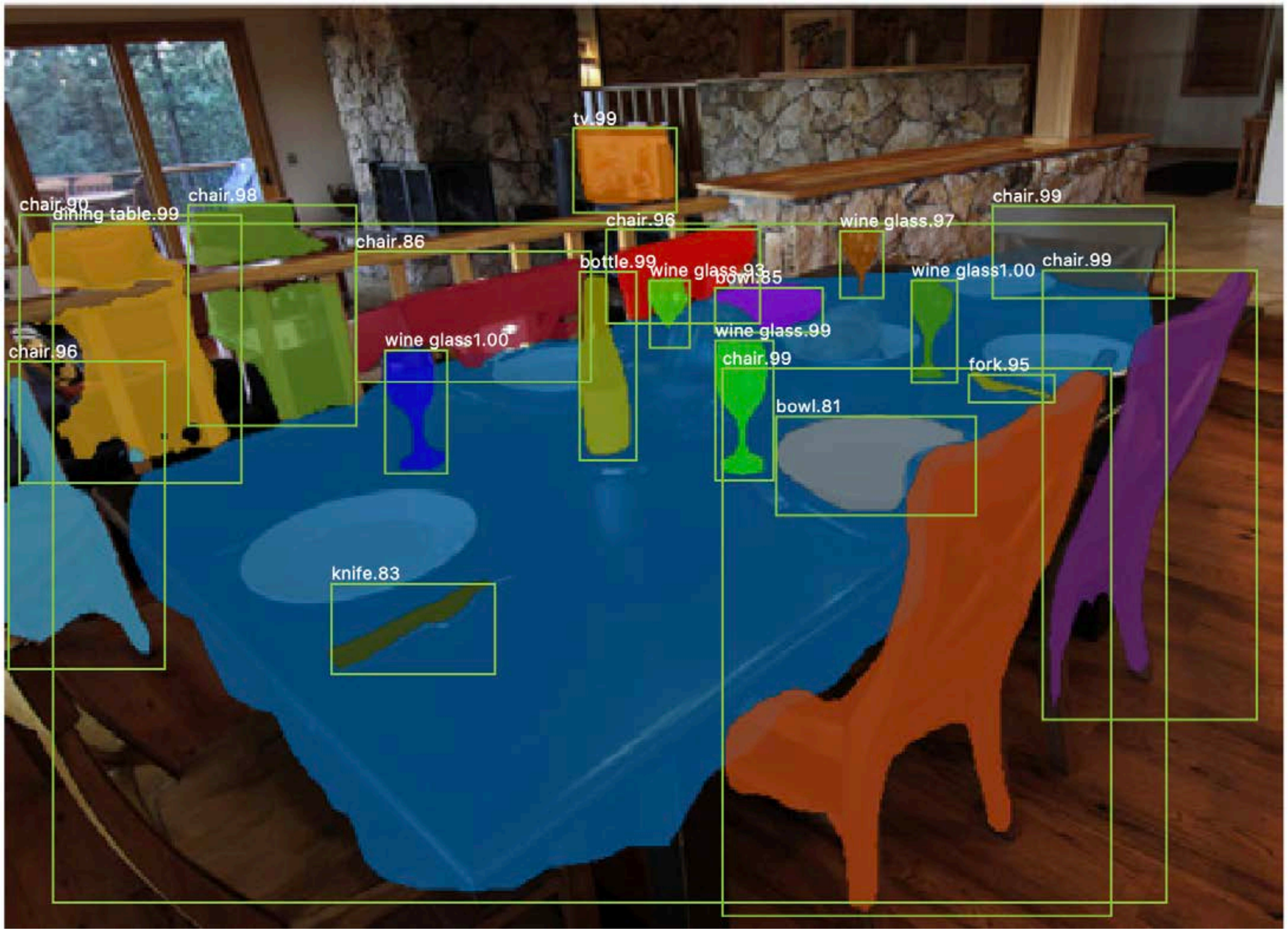


cf. Keramati et al., 2018; Cobo et al., 2013; Garnelo et al., 2016; Lazaro-Gradillo et al., 2019; Zambaldi, et al., 2018

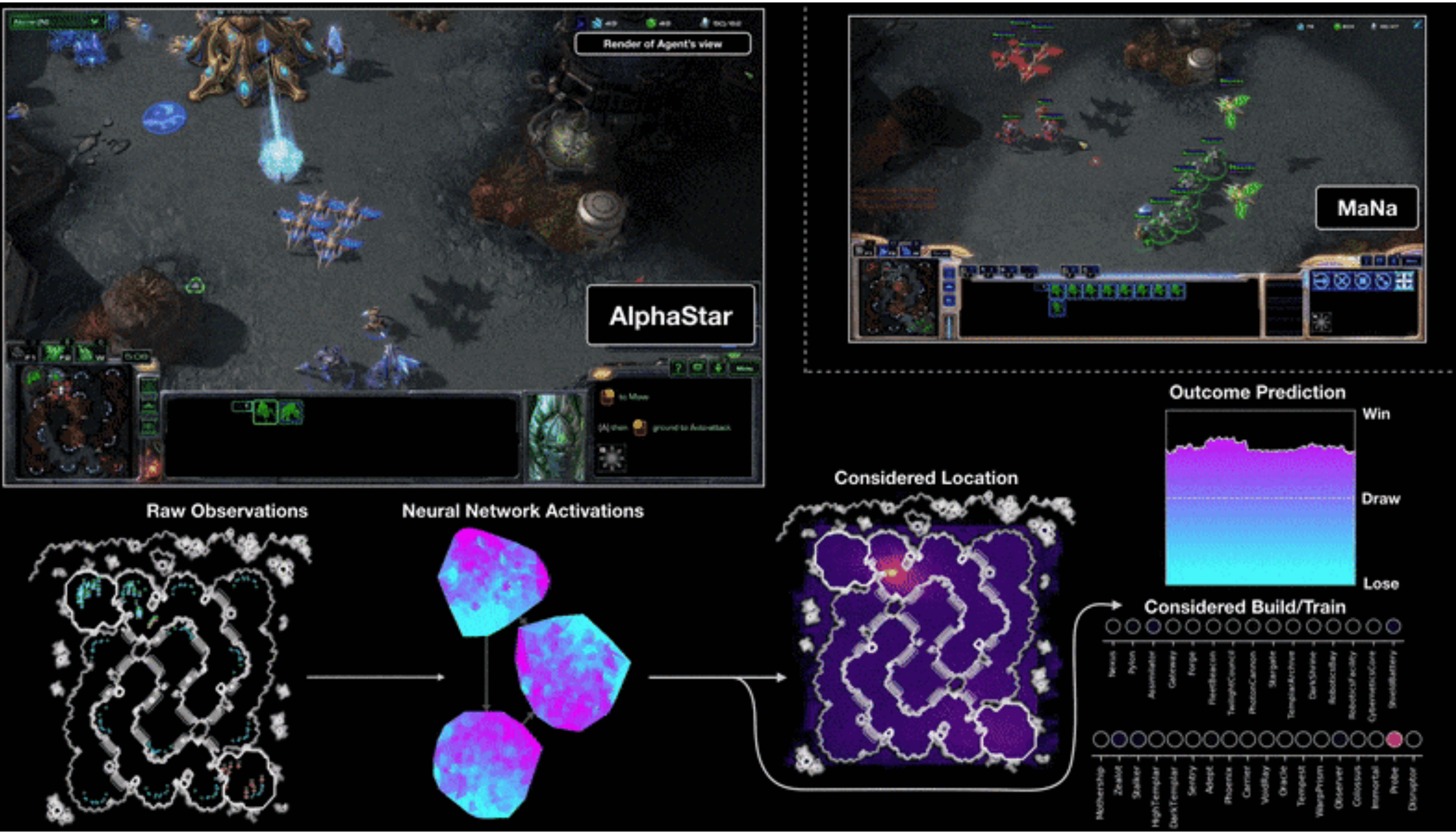

nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
At last — a computer program that can beat a champion Go player PAGE 484
ALL SYSTEMS GO

How research funders profit from hidden investments p. 1100
New books for budding scientists p. 1104
Drug leads for malaria pp. 1122 & 1129
Science
$15
7 DECEMBER 2018
sciencemag.org
AAAS
92.1%
67.0%


Render of Agent's view
AlphaStar
to Move
[A] then ground to Auto-attack
MaNa
Outcome Prediction
Win
Draw
Lose
Considered Location
Raw Observations
Neural Network Activations
Considered Build/Train
Nexus
Pylon
Assimilator
Gateway
Forge
FleetBeacon
TwilightCouncil
PhotonCannon
Stargate
TemplarArchive
DarkShrine
RoboticsBay
RoboticsFacility
CyberneticsCore
ShieldBattery
Mothership
Zealot
Stalker
HighTemplar
DarkTemplar
Sentry
Adept
Phoenix
Carrier
VoidRay
Oracle
Tempest
WarpPrism
Observer
Colossus
Immortal
Probe
Disruptor



E.g., Girshick, 2015; He et al., 2017; Redmon & Farhadi, 2018


Render of Agent's view
AlphaStar
to Move
[A] then ground to Auto-attack
MaNa
Outcome Prediction
Win
Draw
Lose
Raw Observations
Neural Network Activations
Considered Location
Considered Build/Train
Nexus
Pylon
Assimilator
Gateway
Forge
FleetBeacon
TwilightCouncil
PhotonCannon
Stargate
TemplarArchive
DarkShrine
RoboticsBay
RoboticsFacility
CyberneticsCore
ShieldBattery
Mothership
Zealot
Stalker
HighTemplar
DarkTemplar
Sentry
Adept
Phoenix
Carrier
VoidRay
Oracle
Tempest
WarpPrism
Observer
Colossus
Immortal
Probe
Disruptor

Alex Lerchner

Chris Burgess

Loic Matthey

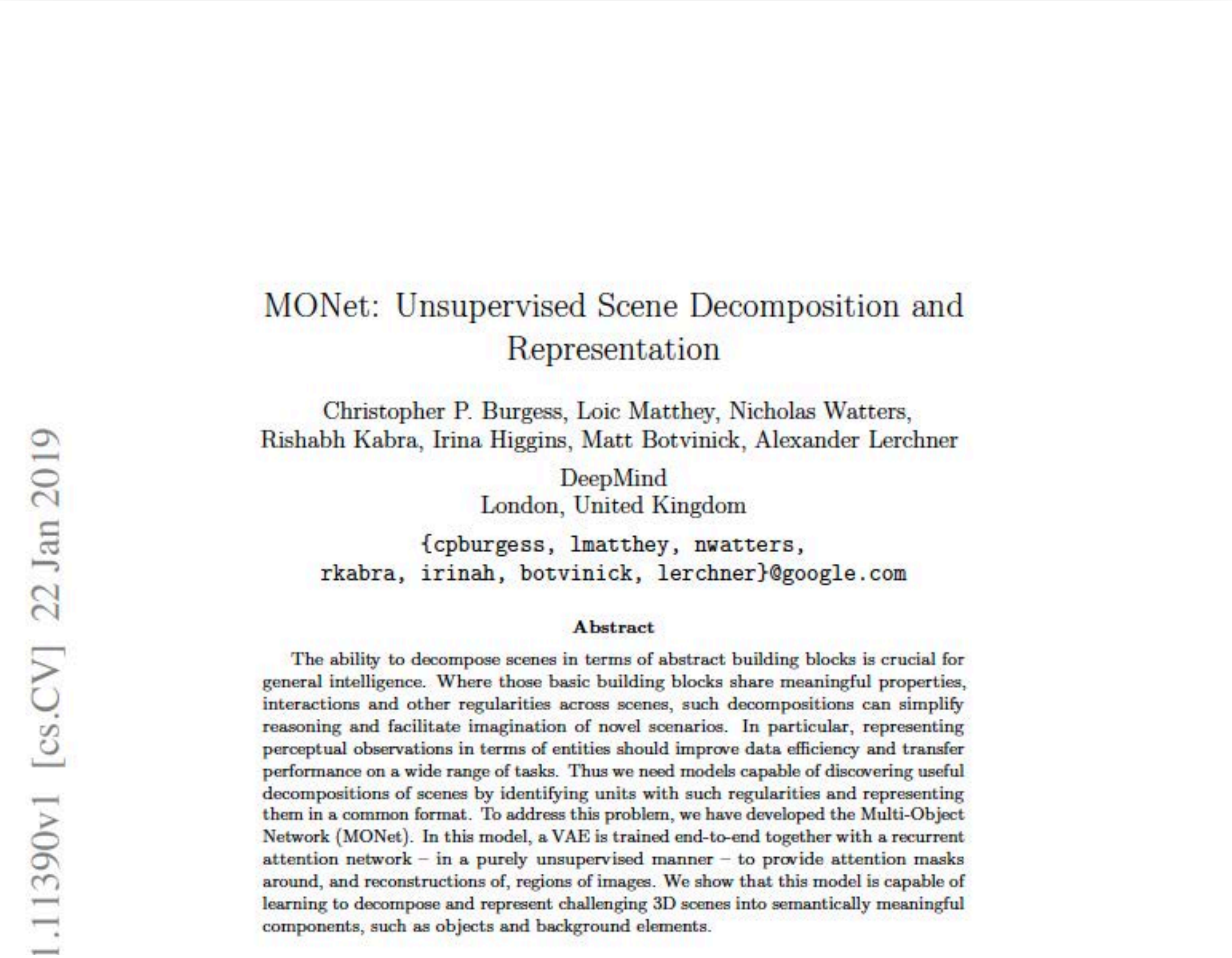
Klaus Greff

Nick Watters

Irina Higgins

Rishabh Kabra

Malcolm Reynolds



MONet: Unsupervised Scene Decomposition and Representation

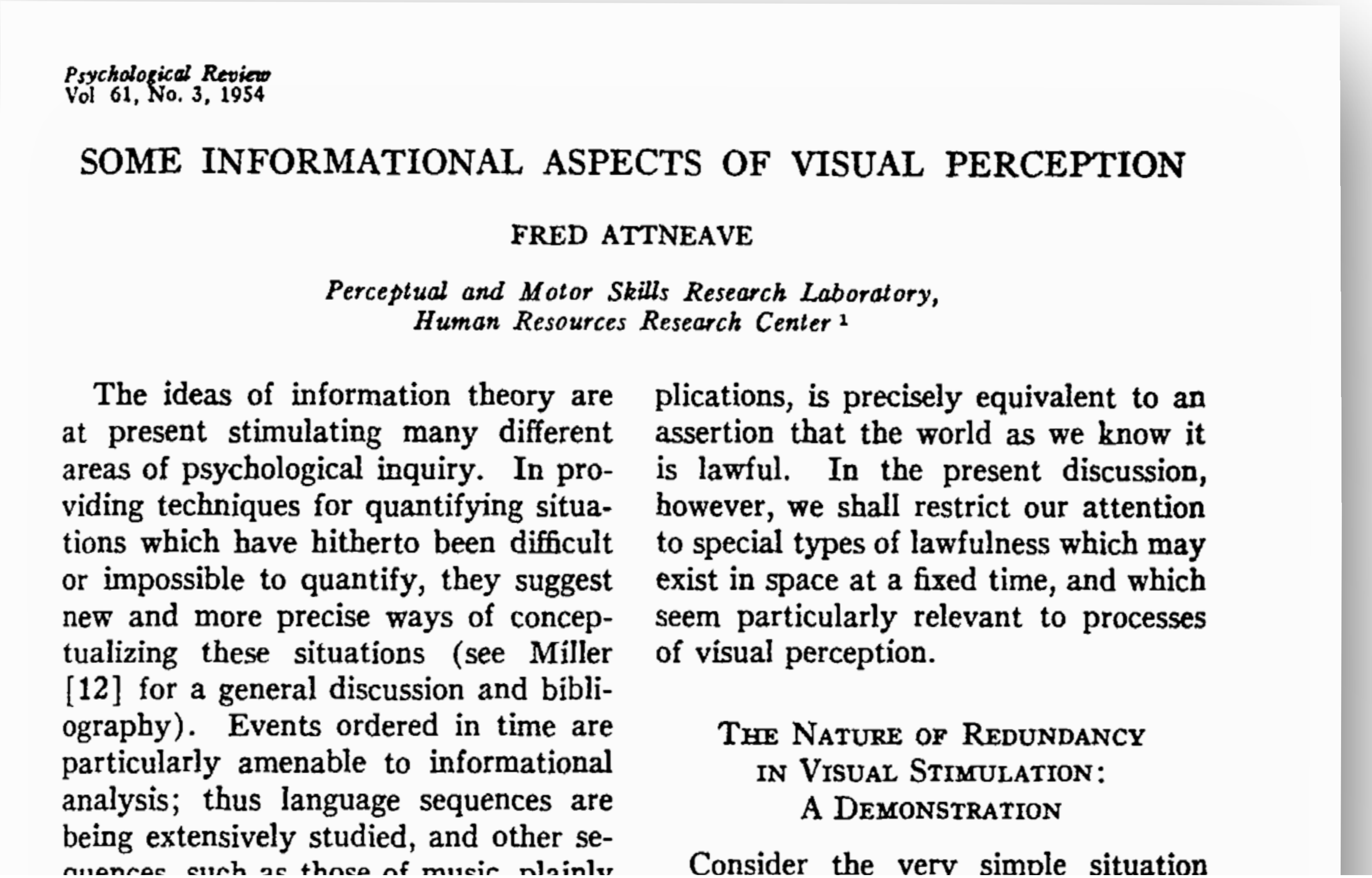
Christopher P. Burgess, Loic Matthey, Nicholas Watters,
Rishabh Kabra, Irina Higgins, Matt Botvinick, Alexander Lerchner

DeepMind
London, United Kingdom

{cpburgess, lmatthey, nwatters,
rkabra, irinah, botvinick, lerchner}@google.com

Abstract

The ability to decompose scenes in terms of abstract building blocks is crucial for general intelligence. Where those basic building blocks share meaningful properties, interactions and other regularities across scenes, such decompositions can simplify reasoning and facilitate imagination of novel scenarios. In particular, representing perceptual observations in terms of entities should improve data efficiency and transfer performance on a wide range of tasks. Thus we need models capable of discovering useful decompositions of scenes by identifying units with such regularities and representing them in a common format. To address this problem, we have developed the Multi-Object Network (MONet). In this model, a VAE is trained end-to-end together with a recurrent attention network – in a purely unsupervised manner – to provide attention masks around, and reconstructions of, regions of images. We show that this model is capable of learning to decompose and represent challenging 3D scenes into semantically meaningful components, such as objects and background elements.



Psychological Review
Vol 61, No. 3, 1954

SOME INFORMATIONAL ASPECTS OF VISUAL PERCEPTION

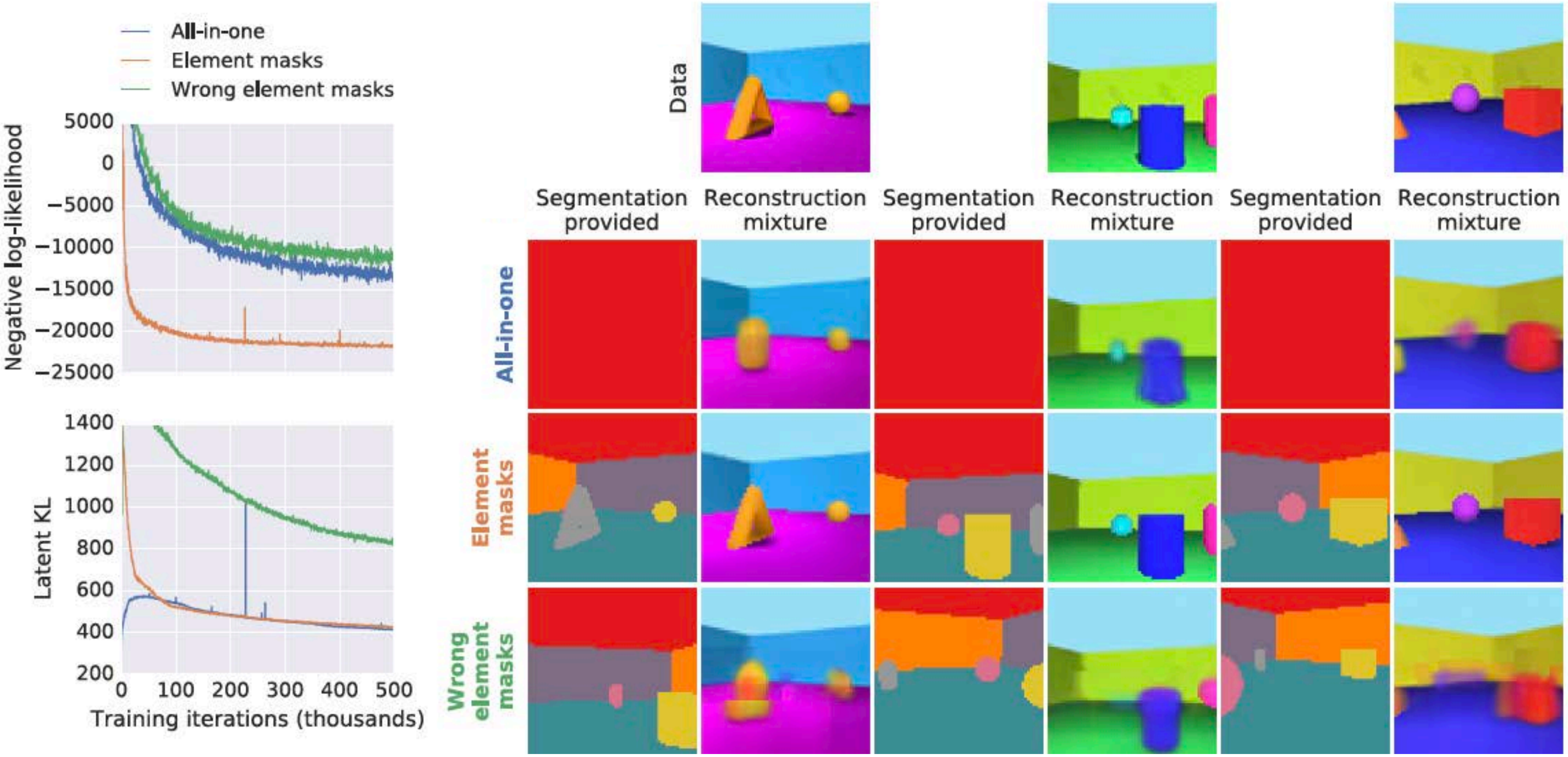
FRED ATTNEAVE

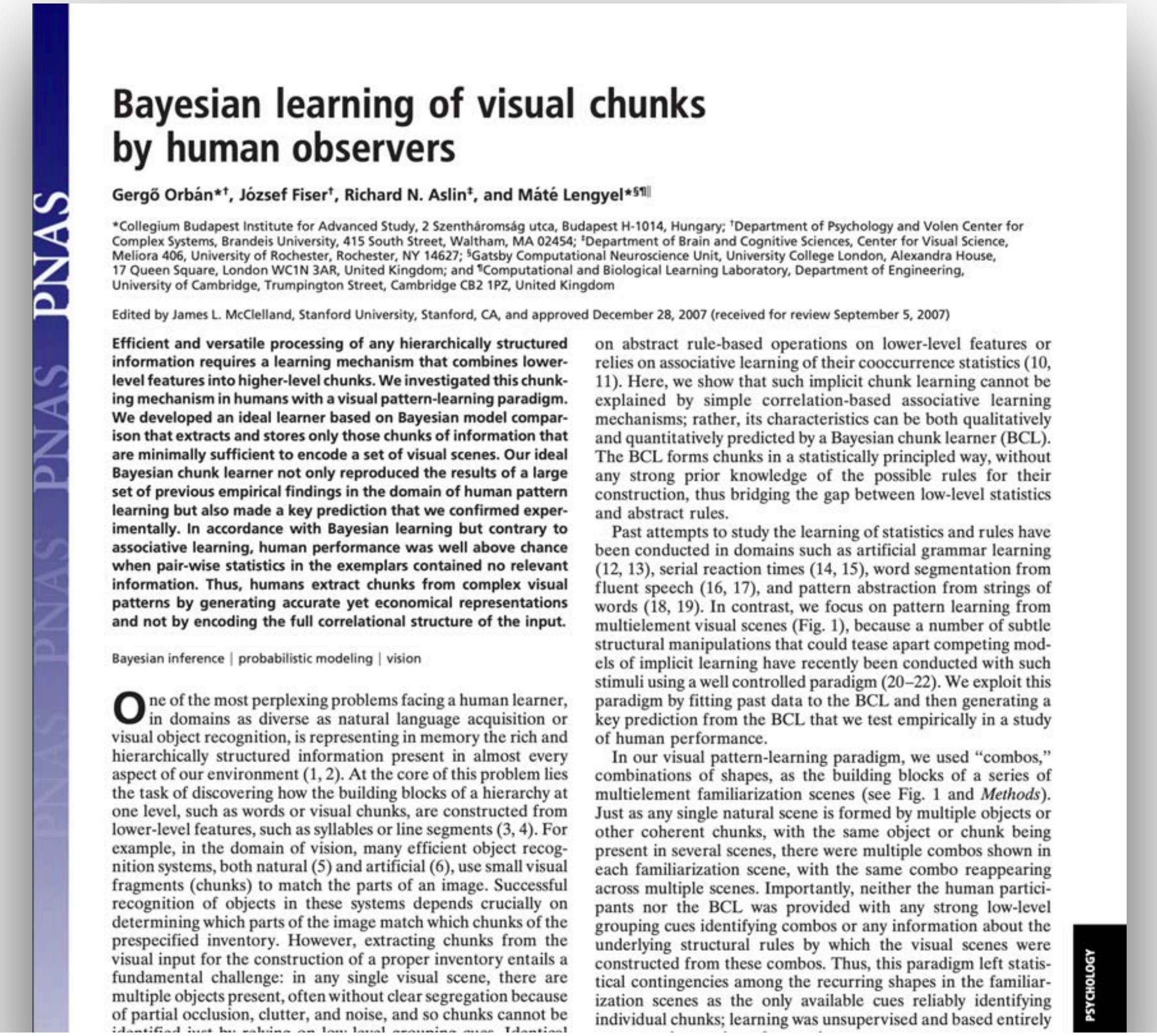
Perceptual and Motor Skills Research Laboratory,
Human Resources Research Center [1]

The ideas of information theory are at present stimulating many different areas of psychological inquiry. In providing techniques for quantifying situations which have hitherto been difficult or impossible to quantify, they suggest new and more precise ways of conceptualizing these situations (see Miller [12] for a general discussion and bibliography). Events ordered in time are particularly amenable to informational analysis; thus language sequences are being extensively studied, and other sequences, such as those of music, plainly plications, is precisely equivalent to an assertion that the world as we know it is lawful. In the present discussion, however, we shall restrict our attention to special types of lawfulness which may exist in space at a fixed time, and which seem particularly relevant to processes of visual perception.

The Nature of Redundancy in Visual Stimulation: A Demonstration

Consider the very simple situation


All-in-one
Element masks
Wrong element masks
Negative log-likelihood
5000
0
−5000
−10000
−15000
−20000
−25000
Latent KL
1400
1200
1000
800
600
400
200
0
100
200
300
400
500
Training iterations (thousands)
Data
Segmentation provided
Reconstruction mixture
Segmentation provided
Reconstruction mixture
Segmentation provided
Reconstruction mixture
All-in-one
Element masks
Wrong element masks



PNAS

Bayesian learning of visual chunks by human observers

Gergő Orbán*[†], József Fiser[†], Richard N. Aslin[‡], and Máté Lengyel*[§¶‖]

*Collegium Budapest Institute for Advanced Study, 2 Szentháromság utca, Budapest H-1014, Hungary; [†]Department of Psychology and Volen Center for Complex Systems, Brandeis University, 415 South Street, Waltham, MA 02454; [‡]Department of Brain and Cognitive Sciences, Center for Visual Science, Meliora 406, University of Rochester, Rochester, NY 14627; [§]Gatsby Computational Neuroscience Unit, University College London, Alexandra House, 17 Queen Square, London WC1N 3AR, United Kingdom; and [¶]Computational and Biological Learning Laboratory, Department of Engineering, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, United Kingdom

Edited by James L. McClelland, Stanford University, Stanford, CA, and approved December 28, 2007 (received for review September 5, 2007)

Efficient and versatile processing of any hierarchically structured information requires a learning mechanism that combines lower-level features into higher-level chunks. We investigated this chunking mechanism in humans with a visual pattern-learning paradigm. We developed an ideal learner based on Bayesian model comparison that extracts and stores only those chunks of information that are minimally sufficient to encode a set of visual scenes. Our ideal Bayesian chunk learner not only reproduced the results of a large set of previous empirical findings in the domain of human pattern learning but also made a key prediction that we confirmed experimentally. In accordance with Bayesian learning but contrary to associative learning, human performance was well above chance when pair-wise statistics in the exemplars contained no relevant information. Thus, humans extract chunks from complex visual patterns by generating accurate yet economical representations and not by encoding the full correlational structure of the input.

Bayesian inference | probabilistic modeling | vision

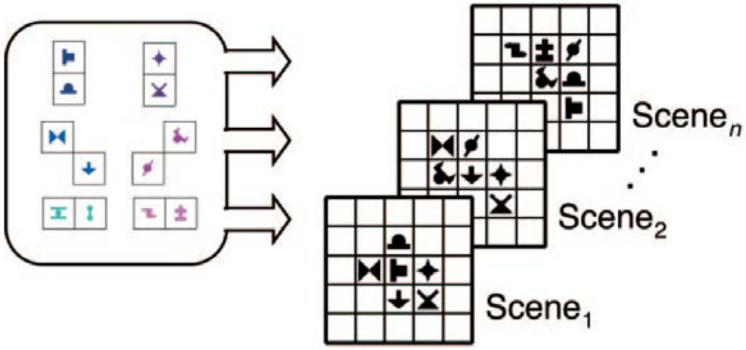
One of the most perplexing problems facing a human learner, in domains as diverse as natural language acquisition or visual object recognition, is representing in memory the rich and hierarchically structured information present in almost every aspect of our environment (1, 2). At the core of this problem lies the task of discovering how the building blocks of a hierarchy at one level, such as words or visual chunks, are constructed from lower-level features, such as syllables or line segments (3, 4). For example, in the domain of vision, many efficient object recognition systems, both natural (5) and artificial (6), use small visual fragments (chunks) to match the parts of an image. Successful recognition of objects in these systems depends crucially on determining which parts of the image match which chunks of the prespecified inventory. However, extracting chunks from the visual input for the construction of a proper inventory entails a fundamental challenge: in any single visual scene, there are multiple objects present, often without clear segregation because of partial occlusion, clutter, and noise, and so chunks cannot be

on abstract rule-based operations on lower-level features or relies on associative learning of their cooccurrence statistics (10, 11). Here, we show that such implicit chunk learning cannot be explained by simple correlation-based associative learning mechanisms; rather, its characteristics can be both qualitatively and quantitatively predicted by a Bayesian chunk learner (BCL). The BCL forms chunks in a statistically principled way, without any strong prior knowledge of the possible rules for their construction, thus bridging the gap between low-level statistics and abstract rules.

Past attempts to study the learning of statistics and rules have been conducted in domains such as artificial grammar learning (12, 13), serial reaction times (14, 15), word segmentation from fluent speech (16, 17), and pattern abstraction from strings of words (18, 19). In contrast, we focus on pattern learning from multielement visual scenes (Fig. 1), because a number of subtle structural manipulations that could tease apart competing models of implicit learning have recently been conducted with such stimuli using a well controlled paradigm (20–22). We exploit this paradigm by fitting past data to the BCL and then generating a key prediction from the BCL that we test empirically in a study of human performance.

In our visual pattern-learning paradigm, we used "combos," combinations of shapes, as the building blocks of a series of multielement familiarization scenes (see Fig. 1 and *Methods*). Just as any single natural scene is formed by multiple objects or other coherent chunks, with the same object or chunk being present in several scenes, there were multiple combos shown in each familiarization scene, with the same combo reappearing across multiple scenes. Importantly, neither the human participants nor the BCL was provided with any strong low-level grouping cues identifying combos or any information about the underlying structural rules by which the visual scenes were constructed from these combos. Thus, this paradigm left statistical contingencies among the recurring shapes in the familiarization scenes as the only available cues reliably identifying individual chunks; learning was unsupervised and based entirely

PSYCHOLOGY



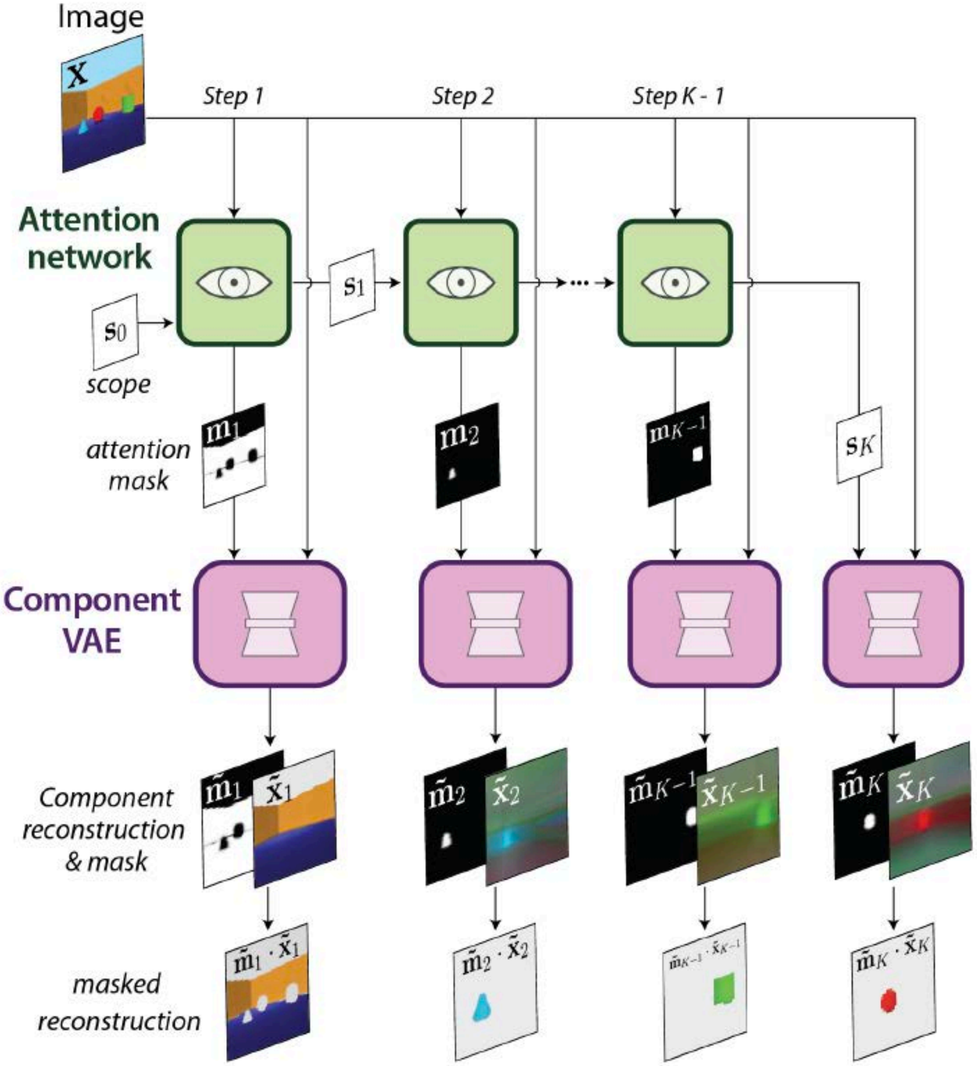
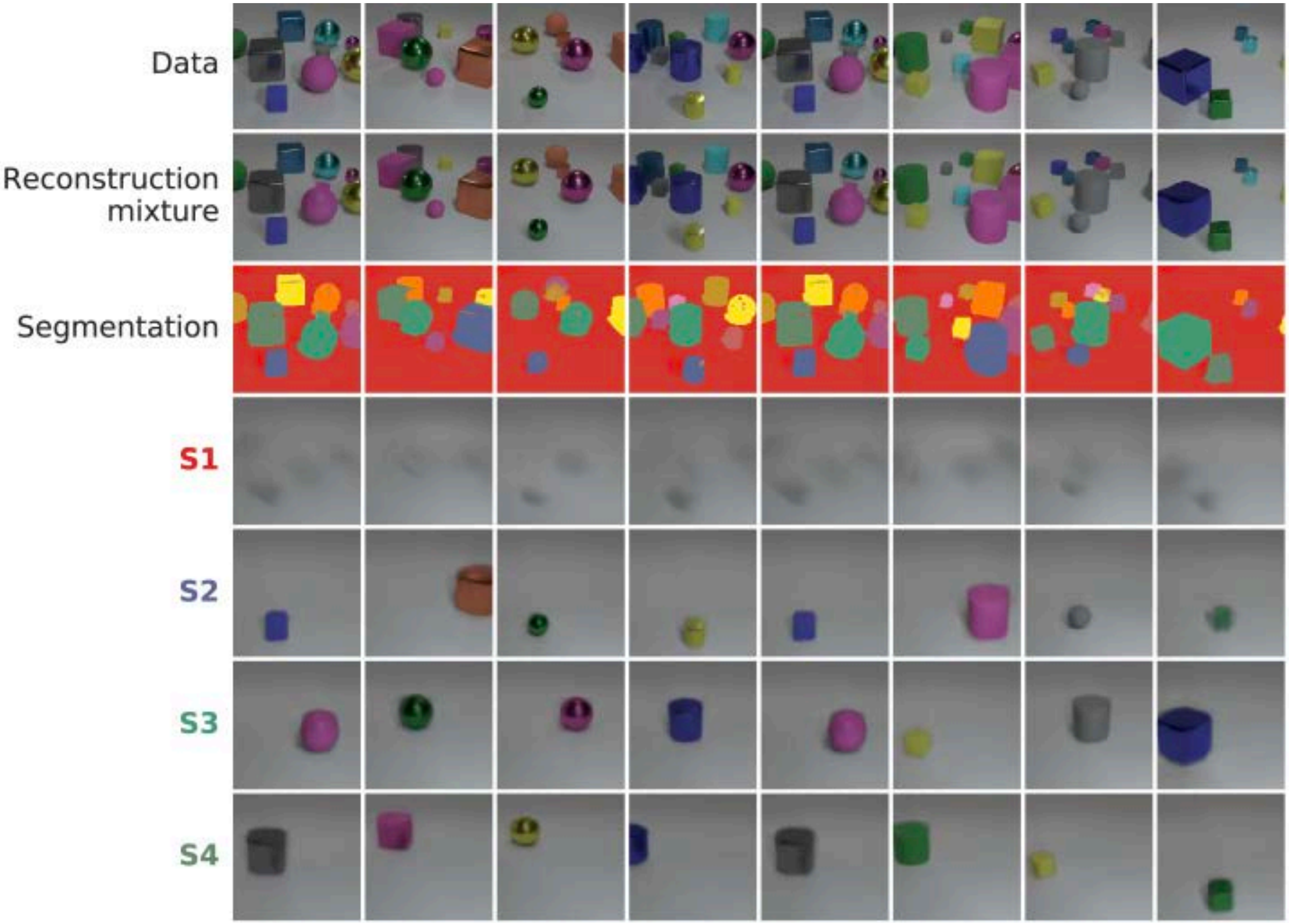
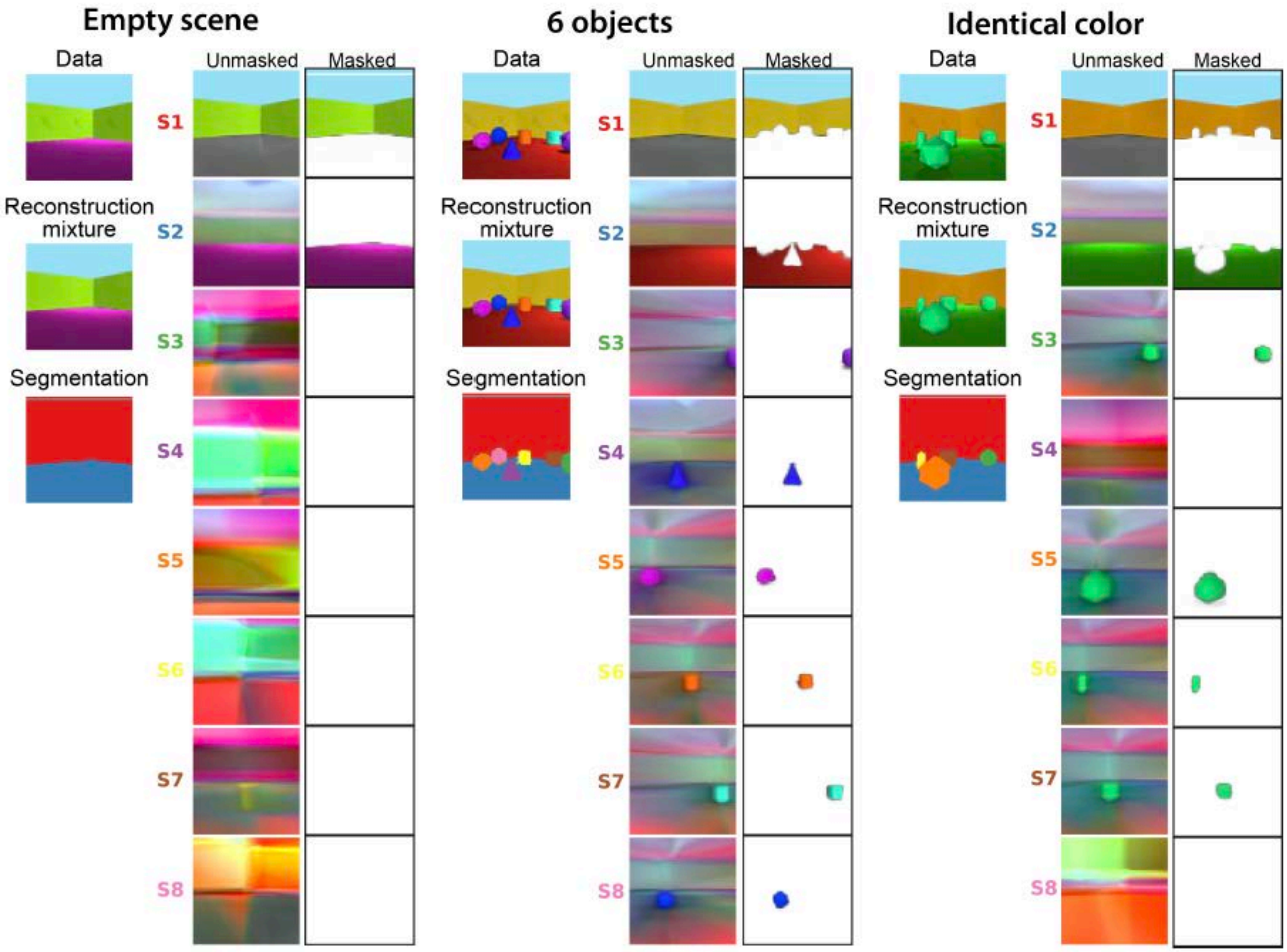
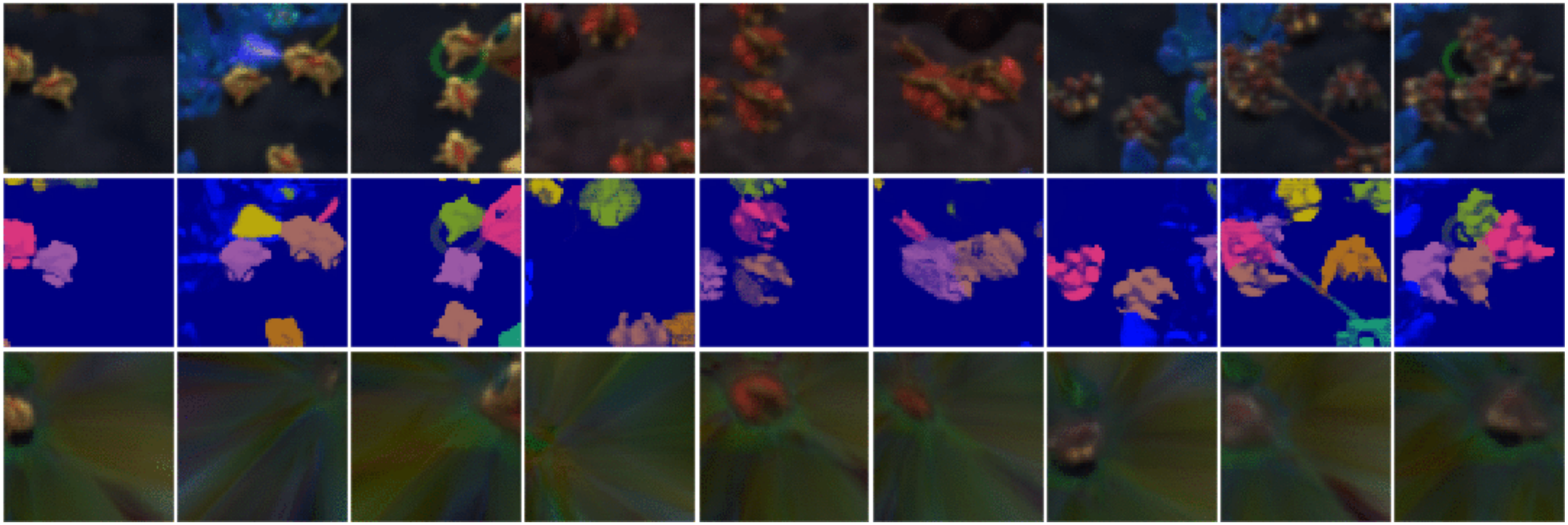


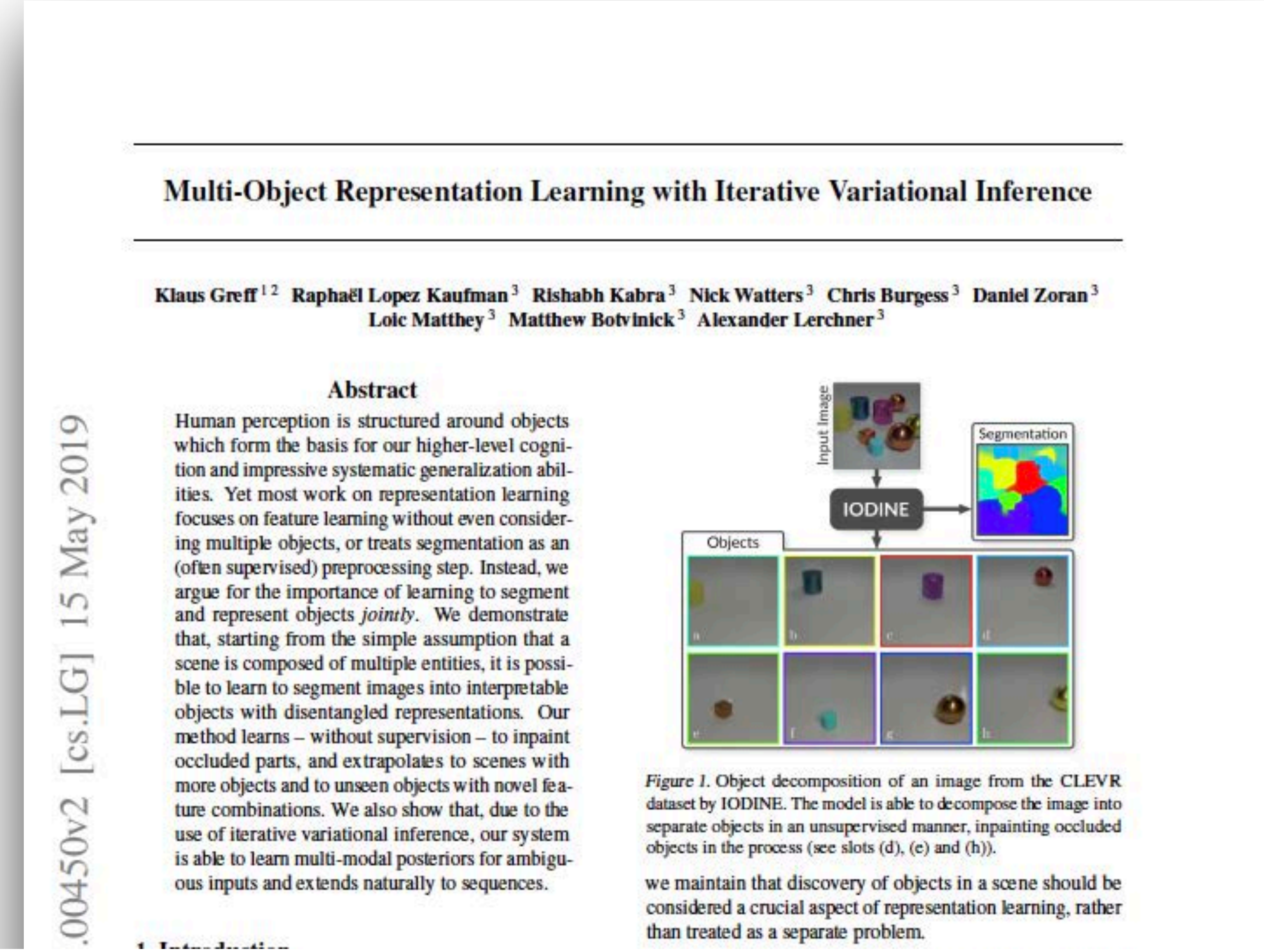
Image
$\mathbf{x}$
Step 1
Step 2
Step K - 1
Attention network
$\mathbf{s}_0$
scope
$\mathbf{s}_1$
$\mathbf{s}_K$
attention mask
$\mathbf{m}_1$
$\mathbf{m}_2$
$\mathbf{m}_{K-1}$
Component VAE
Component reconstruction & mask
$\tilde{\mathbf{m}}_1$
$\tilde{\mathbf{x}}_1$
$\tilde{\mathbf{m}}_2$
$\tilde{\mathbf{x}}_2$
$\tilde{\mathbf{m}}_{K-1}$
$\tilde{\mathbf{x}}_{K-1}$
$\tilde{\mathbf{m}}_K$
$\tilde{\mathbf{x}}_K$
masked reconstruction
$\tilde{\mathbf{m}}_1 \cdot \tilde{\mathbf{x}}_1$
$\tilde{\mathbf{m}}_2 \cdot \tilde{\mathbf{x}}_2$
$\tilde{\mathbf{m}}_{K-1} \cdot \tilde{\mathbf{x}}_{K-1}$
$\tilde{\mathbf{m}}_K \cdot \tilde{\mathbf{x}}_K$

$$\mathcal{L}(\phi;\theta;\psi;\mathbf{x}) = -\log\sum_{k=1}^{K}\mathbf{m}_k p_\theta(\mathbf{x}|\mathbf{z}_k) + \beta D_{KL}\big(\prod_{k=1}^{K} q_\phi(\mathbf{z}_k|\mathbf{x},\mathbf{m}_k) \parallel p(\mathbf{z})\big)$$
$$+ \gamma D_{KL}\big(q_\psi(\mathbf{c}|\mathbf{x}) \parallel p_\theta(\mathbf{c}|\{\mathbf{z}_k\})\big)$$


Data
Reconstruction mixture
Segmentation
S1
S2
S3
S4


Empty scene
Data
Unmasked
Masked
Reconstruction mixture
Segmentation
S1
S2
S3
S4
S5
S6
S7
S8
6 objects
Data
Unmasked
Masked
Reconstruction mixture
Segmentation
S1
S2
S3
S4
S5
S6
S7
S8
Identical color
Data
Unmasked
Masked
Reconstruction mixture
Segmentation
S1
S2
S3
S4
S5
S6
S7
S8





Multi-Object Representation Learning with Iterative Variational Inference

Klaus Greff [1 2] Raphaël Lopez Kaufman [3] Rishabh Kabra [3] Nick Watters [3] Chris Burgess [3] Daniel Zoran [3]
Loic Matthey [3] Matthew Botvinick [3] Alexander Lerchner [3]

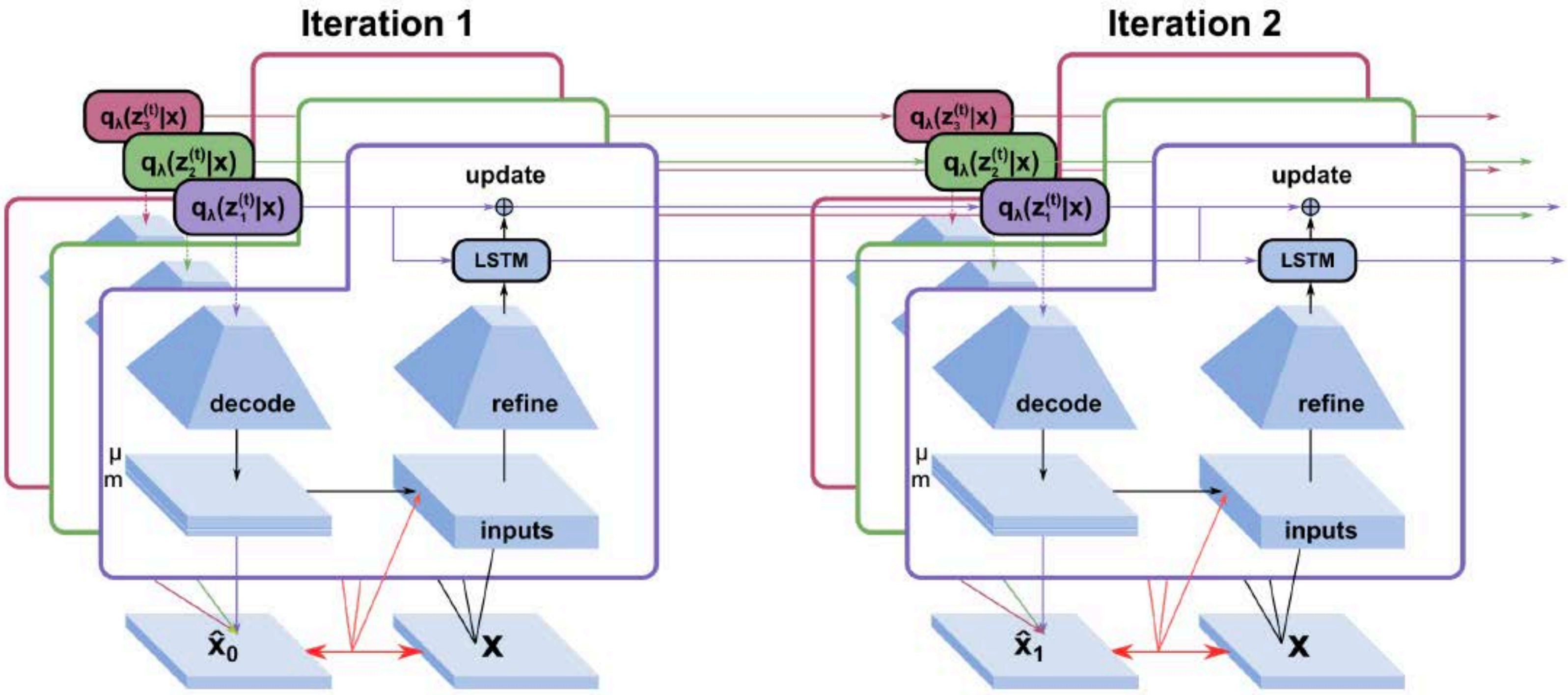
Abstract

Human perception is structured around objects which form the basis for our higher-level cognition and impressive systematic generalization abilities. Yet most work on representation learning focuses on feature learning without even considering multiple objects, or treats segmentation as an (often supervised) preprocessing step. Instead, we argue for the importance of learning to segment and represent objects *jointly*. We demonstrate that, starting from the simple assumption that a scene is composed of multiple entities, it is possible to learn to segment images into interpretable objects with disentangled representations. Our method learns – without supervision – to inpaint occluded parts, and extrapolates to scenes with more objects and to unseen objects with novel feature combinations. We also show that, due to the use of iterative variational inference, our system is able to learn multi-modal posteriors for ambiguous inputs and extends naturally to sequences.

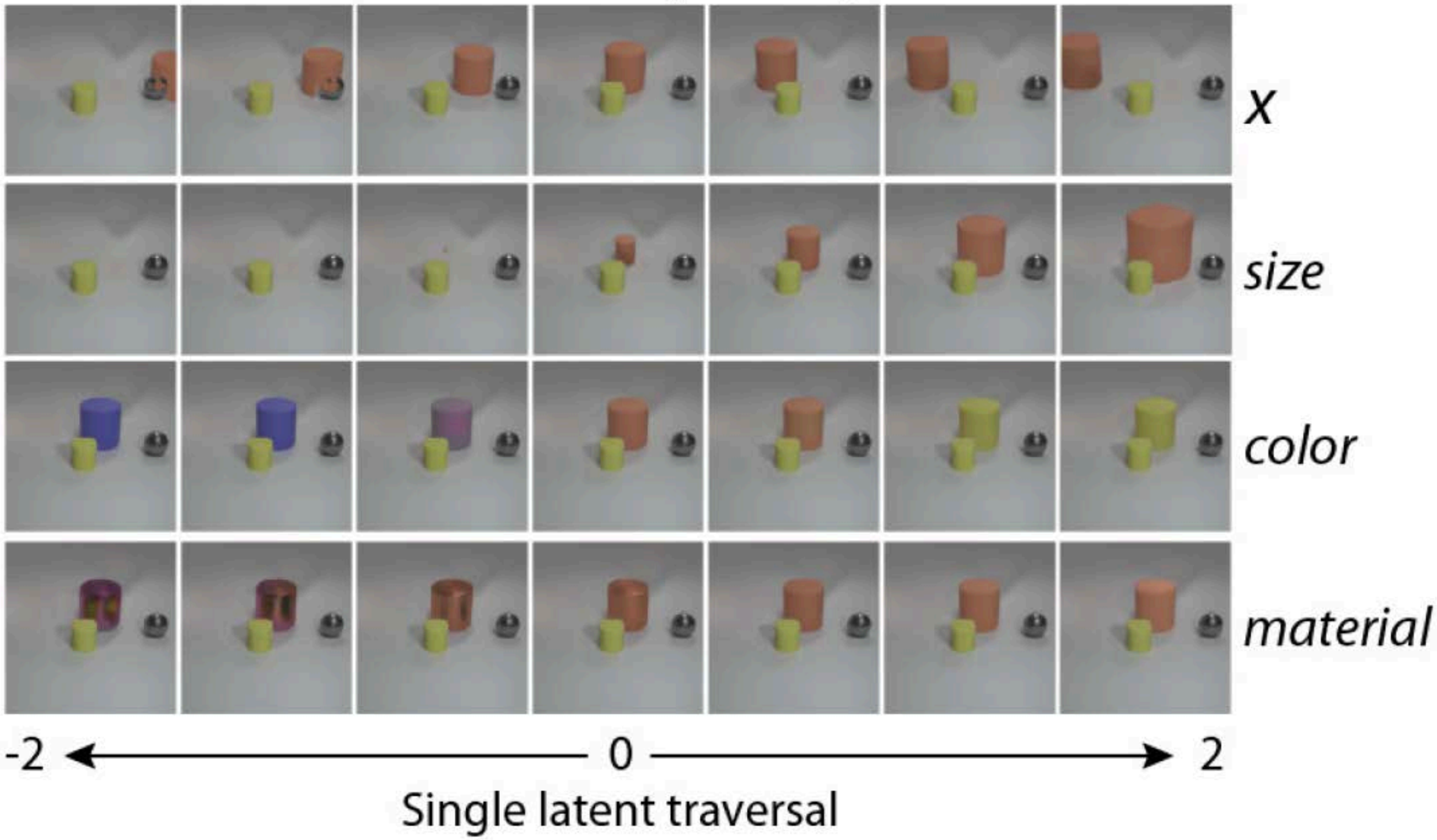
1 Introduction

Figure 1. Object decomposition of an image from the CLEVR dataset by IODINE. The model is able to decompose the image into separate objects in an unsupervised manner, inpainting occluded objects in the process (see slots (d), (e) and (h)).

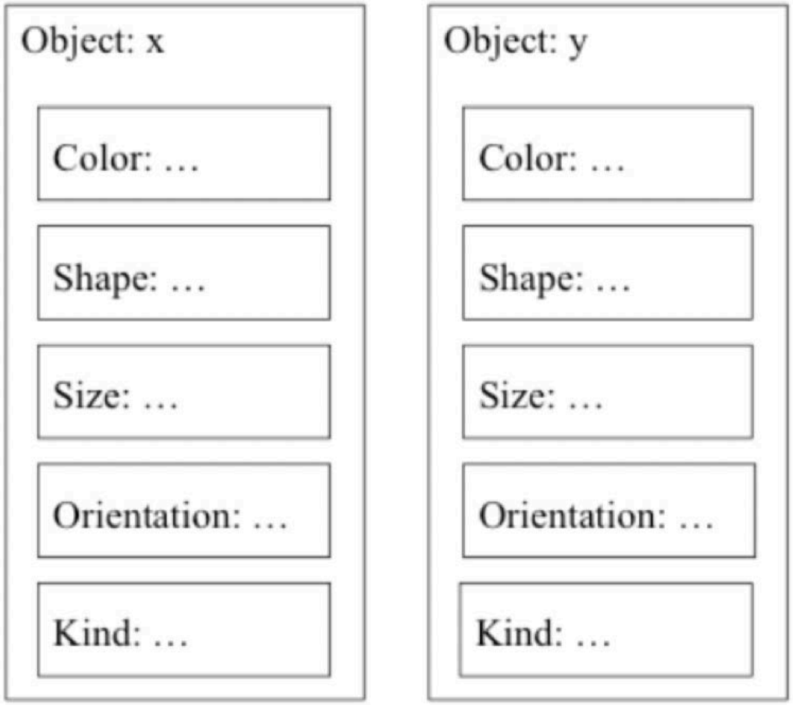
we maintain that discovery of objects in a scene should be considered a crucial aspect of representation learning, rather than treated as a separate problem.



$$\mathcal{L}^{(t)} \leftarrow D_{KL}(q_\lambda(\mathbf{z}^{(t)}|\mathbf{x})||p(\mathbf{z})) - \log p(\mathbf{x}|\mathbf{z}^{(t)})$$


x
size
color
material
-2
0
2
Single latent traversal

Kahneman & Treisman, 1984: Object Files



Object: x

Color: …

Shape: …

Size: …

Orientation: …

Kind: …

Object: y

Color: …

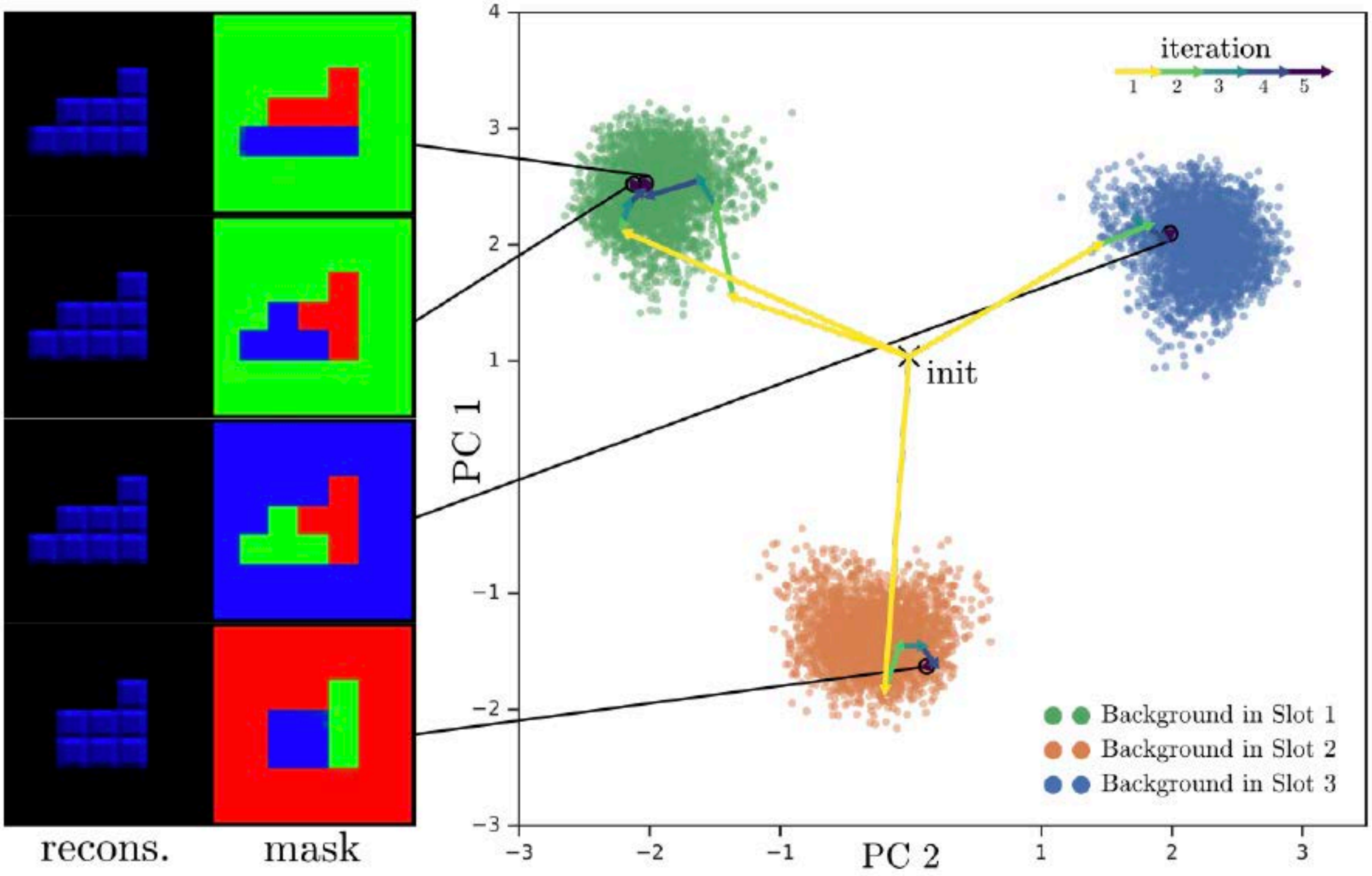
Shape: …

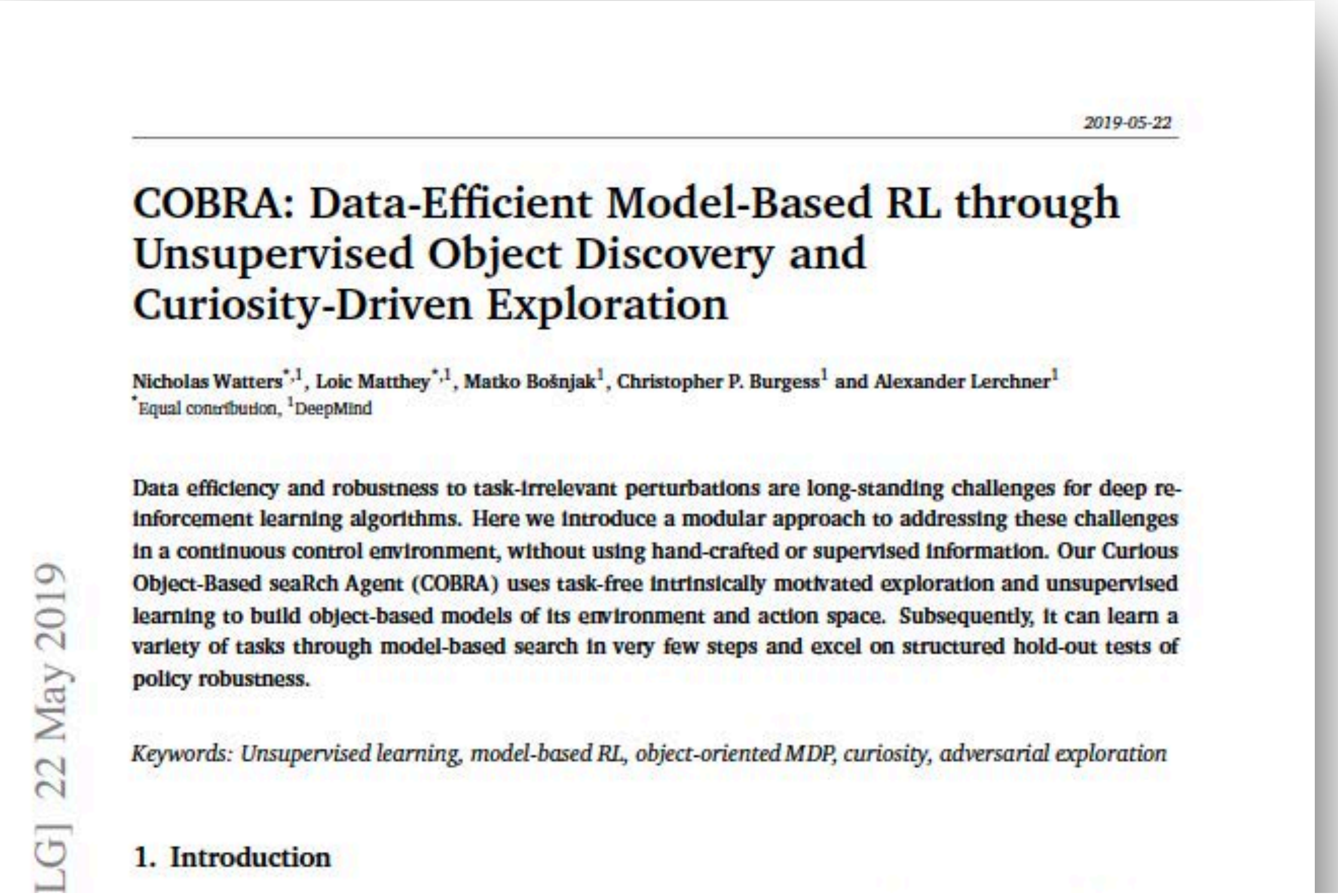
Size: …

Orientation: …

Kind: …

Green, Edwin James, and Jake Quilty-Dunn. "what is an object file?." *The British Journal for the Philosophy of Science* (2017).


recons.
mask
iteration
1
2
3
4
5
init
PC 1
PC 2
4
3
2
1
−1
−2
−3
−3
−2
−1
1
2
3
Background in Slot 1
Background in Slot 2
Background in Slot 3



2019-05-22

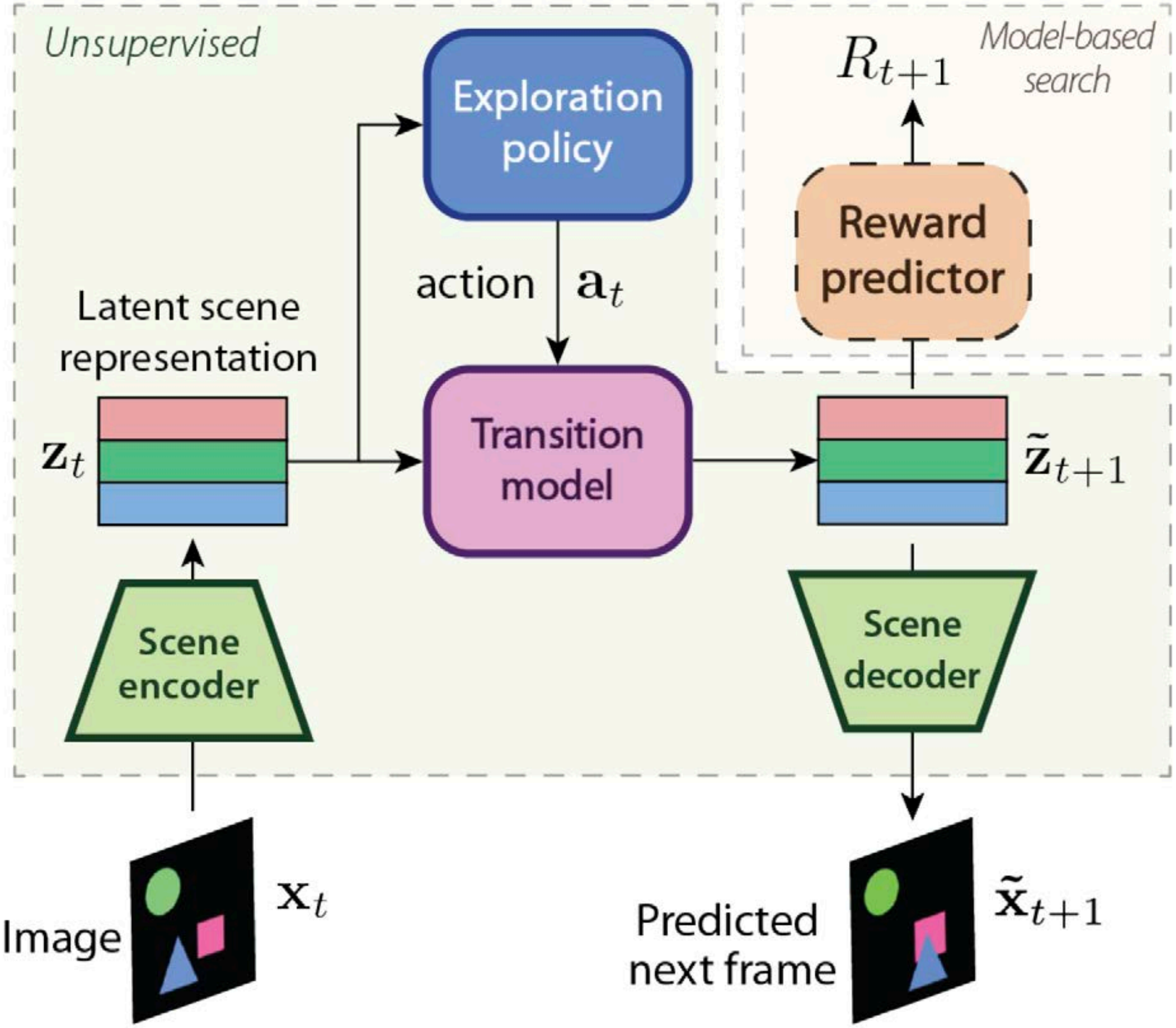
COBRA: Data-Efficient Model-Based RL through Unsupervised Object Discovery and Curiosity-Driven Exploration

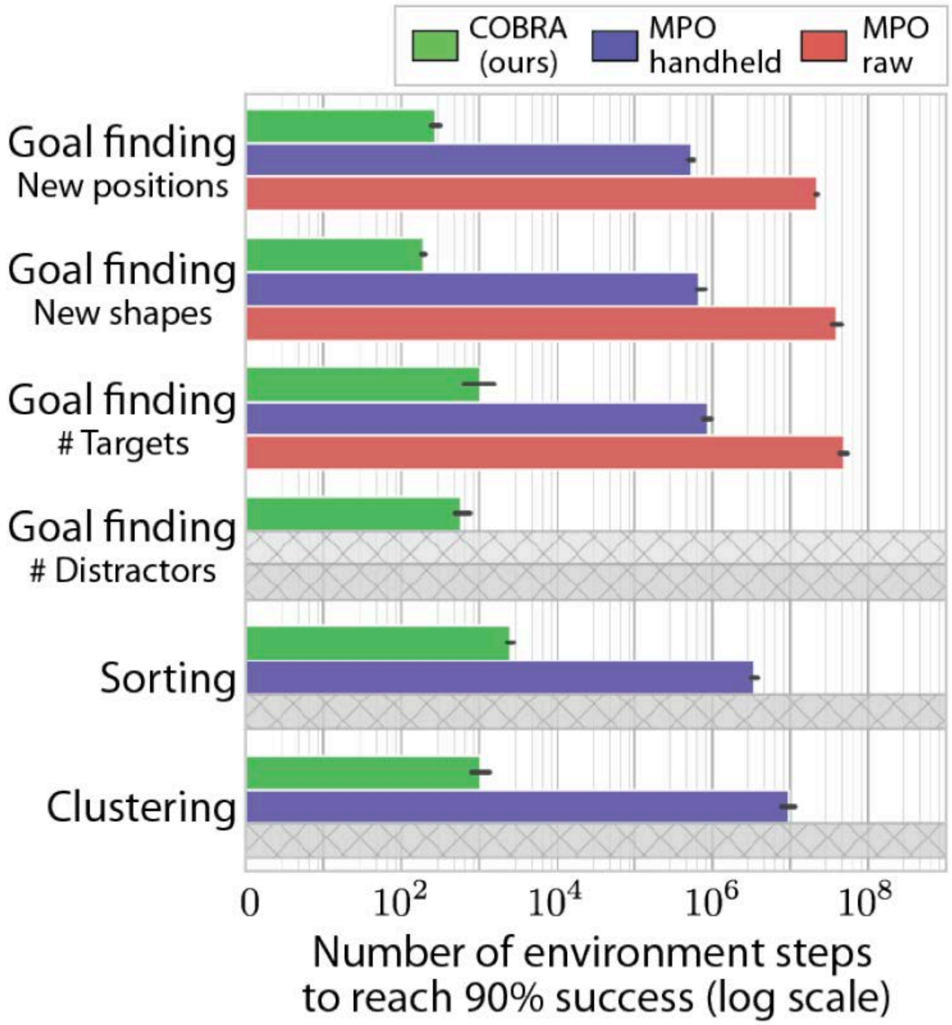
Nicholas Watters[*,1], Loic Matthey[*,1], Matko Bošnjak[1], Christopher P. Burgess[1] and Alexander Lerchner[1]
[*]Equal contribution, [1]DeepMind

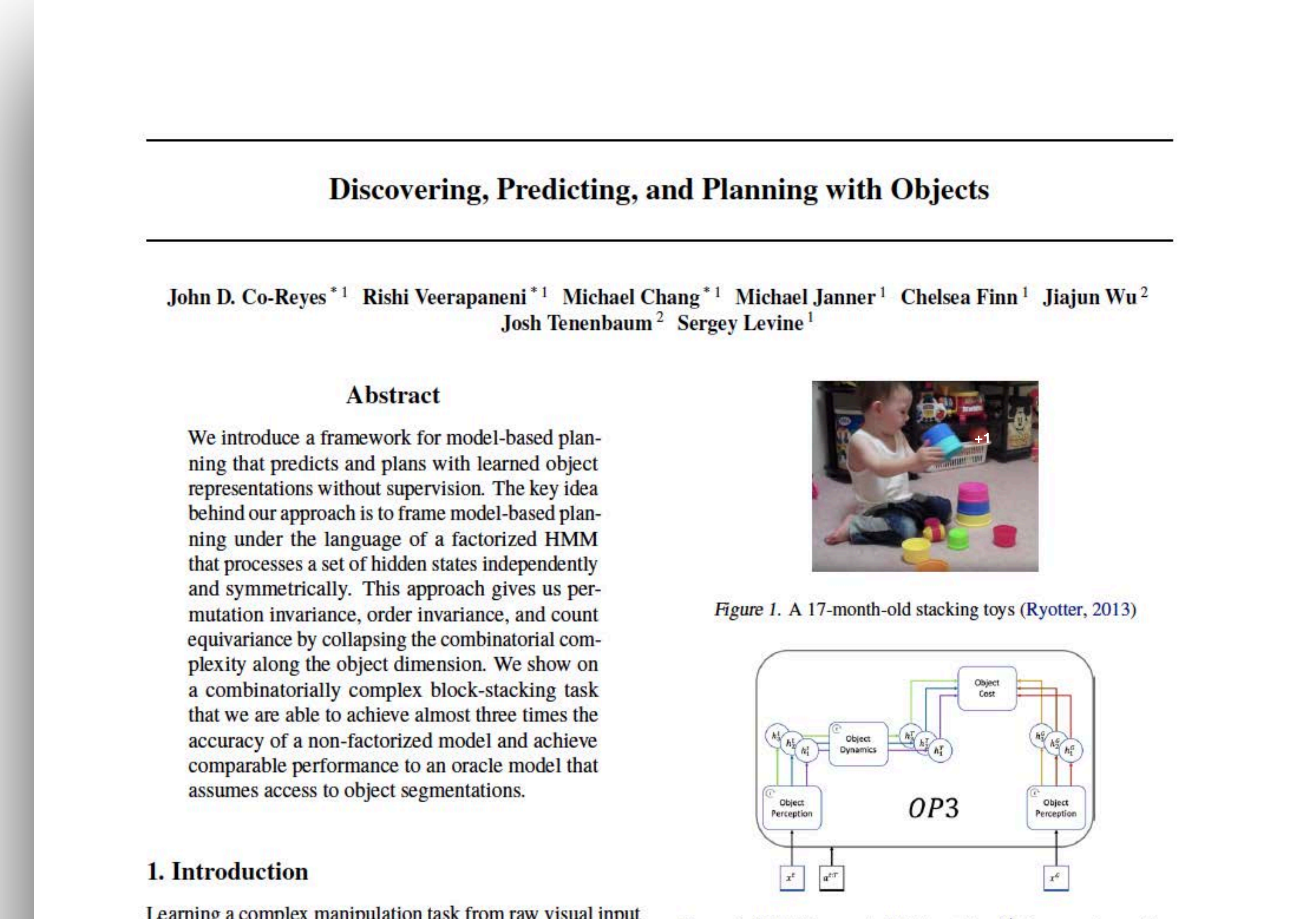
Data efficiency and robustness to task-irrelevant perturbations are long-standing challenges for deep reinforcement learning algorithms. Here we introduce a modular approach to addressing these challenges in a continuous control environment, without using hand-crafted or supervised information. Our Curious Object-Based seaRch Agent (COBRA) uses task-free intrinsically motivated exploration and unsupervised learning to build object-based models of its environment and action space. Subsequently, it can learn a variety of tasks through model-based search in very few steps and excel on structured hold-out tests of policy robustness.

Keywords: Unsupervised learning, model-based RL, object-oriented MDP, curiosity, adversarial exploration

1. Introduction


Unsupervised
Exploration policy
Model-based search
R_{t+1}
Reward predictor
action
$\mathbf{a}_t$
Latent scene representation
$\mathbf{z}_t$
Transition model
$\tilde{\mathbf{z}}_{t+1}$
Scene encoder
Scene decoder
Image
$\mathbf{x}_t$
Predicted next frame
$\tilde{\mathbf{x}}_{t+1}$


🟩 COBRA (ours)
🟦 MPO handheld
🟥 MPO raw
Goal finding
New positions
Goal finding
New shapes
Goal finding
# Targets
Goal finding
# Distractors
Sorting
Clustering
0
10^2
10^4
10^6
10^8
Number of environment steps
to reach 90% success (log scale)



Discovering, Predicting, and Planning with Objects

John D. Co-Reyes * 1 Rishi Veerapaneni * 1 Michael Chang * 1 Michael Janner 1 Chelsea Finn 1 Jiajun Wu 2
Josh Tenenbaum 2 Sergey Levine 1

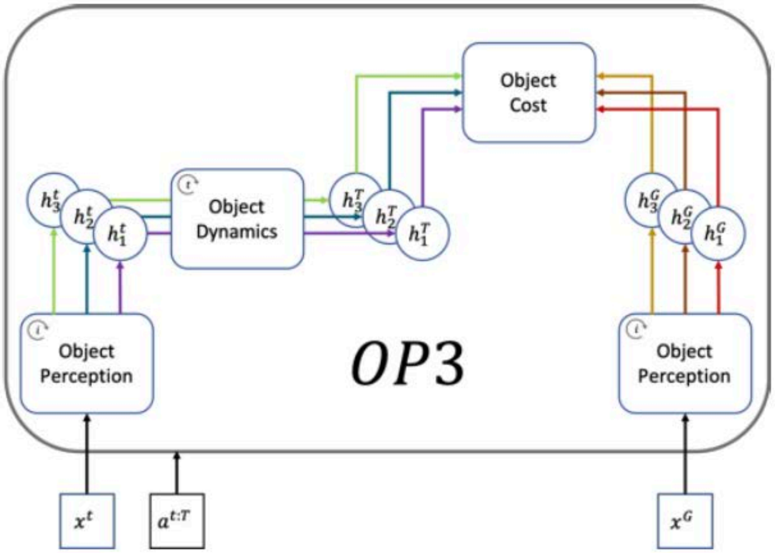
Abstract

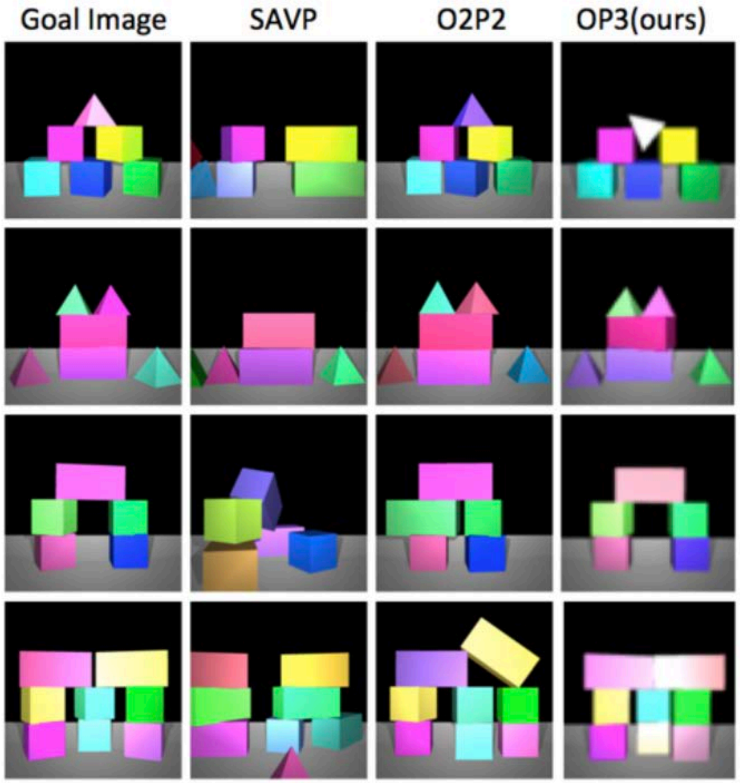
We introduce a framework for model-based planning that predicts and plans with learned object representations without supervision. The key idea behind our approach is to frame model-based planning under the language of a factorized HMM that processes a set of hidden states independently and symmetrically. This approach gives us permutation invariance, order invariance, and count equivariance by collapsing the combinatorial complexity along the object dimension. We show on a combinatorially complex block-stacking task that we are able to achieve almost three times the accuracy of a non-factorized model and achieve comparable performance to an oracle model that assumes access to object segmentations.

Figure 1. A 17-month-old stacking toys (Ryotter, 2013)

1. Introduction

Learning a complex manipulation task from raw visual input


Object Cost
Object Dynamics
Object Perception
Object Perception
OP3
h_3^t h_2^t h_1^t
h_3^T h_2^T h_1^T
h_3^G h_2^G h_1^G
x^t
$a^{t:T}$
x^G


Goal Image
SAVP
O2P2
OP3(ours)

Alex Lerchner

Chris Burgess

Loic Matthey

Klaus Greff

Nick Watters

Irina Higgins

Rishabh Kabra

Malcolm Reynolds