Neuron

CellPress
OPEN ACCESS

Perspective

# Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence

Martin Schrimpf,[1,2,3] Jonas Kubilius,[2,4,5] Michael J. Lee,[1,2] N. Apurva Ratan Murty,[1,2,3] Robert Ajemian,[1,2] and James J. DiCarlo[1,2,3,*]

[1]Department of Brain and Cognitive Sciences, MIT, Cambridge, MA, USA
[2]McGovern Institute for Brain Research, MIT, Cambridge, MA, USA
[3]Center for Brains, Minds and Machines, MIT, Cambridge, MA, USA
[4]Brain and Cognition, KU Leuven, Leuven, Belgium
[5]Three Thirds, Vilnius, Lithuania
*Correspondence: dicarlo@mit.edu
https://doi.org/10.1016/j.neuron.2020.07.040

## SUMMARY

A potentially organizing goal of the brain and cognitive sciences is to accurately explain domains of human intelligence as executable, neurally mechanistic models. Years of research have led to models that capture experimental results in individual behavioral tasks and individual brain regions. We here advocate for taking the next step: integrating experimental results from many laboratories into suites of benchmarks that, when considered together, push mechanistic models toward explaining entire domains of intelligence, such as vision, language, and motor control. Given recent successes of neurally mechanistic models and the surging availability of neural, anatomical, and behavioral data, we believe that now is the time to create integrative benchmarking platforms that incentivize ambitious, unified models. This perspective discusses the advantages and the challenges of this approach and proposes specific steps to achieve this goal in the domain of visual intelligence with the case study of an integrative benchmarking platform called Brain-Score.

## INTRODUCTION

Brain processing is an interplay of inter-connected networks of neurons, which, taken together, allow us to perceive the world around us, predict what will happen next, communicate with each other, manipulate objects, and much more. Since the ability to solve such problems evolved to enable survival and reproduction in the world, we envision that the neural mechanisms, together with the behavioral abilities corresponding to each such problem domain, offer a useful scale at which to study the brain. These "naturally intelligent" neural algorithms underlying each domain are not only scientifically fascinating, they are an untapped goldmine of next-generation human-intelligent machine systems and other applications.

However, it has been extremely challenging to explain even a single domain of human intelligence in terms of underlying neural mechanisms of the interacting sub-networks. For example, consider the domain of visual intelligence in primates. Even though we know that the cascade of neural impulses within the cortical areas V1, V2, V4, and inferior temporal (IT) cortex, in coordination with other subcortical and motor regions, underlie our ability to recognize and report objects in the world, we still lack a complete understanding of how neural signals mechanistically support complex visual behaviors in particular and visual intelligence more broadly.

Because of the daunting complexity of each domain of intelligence, much of neuroscience started with a "divide and conquer later" strategy: focus the scope of investigation on single brain regions (such as V1) or single psychophysical domains, and study each region or behavioral domain with simple models, often describing a small set of carefully parameterized experiments. The hope in this approach is that we will eventually be able to scale up the building blocks (or "principles") by combining them into an accurate unified model of the entire domain. Scientists, ourselves included, agree on both the continuous necessity of mapping out the space of building blocks but also the ultimate objective of unified modeling of entire domains of human intelligence. However, scientists disagree on when the time is right to start scaling up. Some of the field views it as currently premature to build unified models, because the underlying building blocks have not yet been sufficiently elaborated.

Here, we advocate for the alternative point of view that, based on recent developments, the time is now to build neurally mechanistic models of domains of intelligence and to test them against empirical data in order to develop better and better models by scaling up the right combinations of building blocks. To enable these large-scale modeling efforts, we argue that serious efforts must be spent on integrative benchmarking:

> *Specify a set of behaviors that together define a domain of intelligence, and assemble neural and behavioral data from across its supported domain, in order to guide and constrain neurally mechanistic models that are capable of generating that set of behaviors.*

Each such integrative, executable network model would be an account of how the brain might accomplish the domain of

Please cite this article in press as: Schrimpf et al., Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence, Neuron (2020), https://doi.org/10.1016/j.neuron.2020.07.040



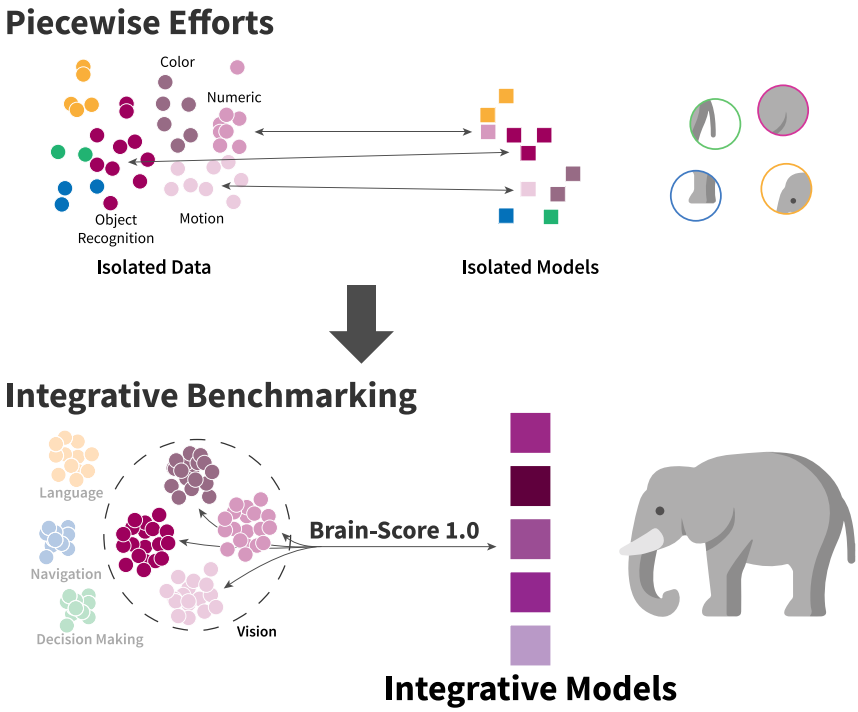


**Figure 1. Can We Build Integrative Models of Intelligence?**

Prior work has attempted to "divide and conquer later," creating models that are accurate but limited in their domain of explanation (describing only a part of the elephant but missing the integrative picture; upper panel). We believe the time is right to begin evaluating and developing models that simultaneously explain a broad set of empirical phenomena. Our belief is that models that aim to capture more of these benchmarks simultaneously will lead to faster progress toward models of an entire domain of intelligence. A single lab cannot accomplish this endeavor—communities need to share empirical benchmarks and candidate models through a common platform. Here we propose the Brain-Score 1.0 platform (see text).

intelligence in question in terms of basic mechanistic components (e.g., interconnected integrate and fire neurons). In that fundamental sense, such models are neurally mechanistic *hypotheses*, and, as such, they make unequivocal predictions that can be falsified by empirical neural and behavioral data. By not mandating these hypotheses to have additional qualifications—e.g., that they be simple and immediately explainable, which to us are not obvious criteria for something as complex as the brain—we are able to thoroughly test models with minimal assumptions, while reaping the benefits of having "merely" accurate models of brain function (see Looking Ahead).

Under our view, this next phase of scientific progress is then driven by a continuous cycle of model (i.e., hypothesis) creation, model prediction, and model testing against all experimental results. Doing so will move the field forward by encouraging the creation of unified, neurally mechanistic models, culling those that are less accurate and drawing focus and further model engineering work to improving and extending the most accurate models (a.k.a. the leading scientific hypotheses).

We here refer to models that attempt to predict *all* relevant neural activity and behavioral data in a domain of human intelligence as "integrative models" or "unified models." We refer to models that are built only with artificial "neurons" and connections of those neurons as "neurally mechanistic models." And we refer to the currently assembled set of empirical behavioral and neural measurements as "integrative benchmarks" (see later).

Given the stated goal (explain a domain of intelligence) and scientific approach of hypothesis testing in the form of model testing (outlined above), the primary purpose of this perspective is to introduce a software platform called Brain-Score as a proposed method of implementing this approach in the domain of visual intelligence. Brain-Score has been built to enable shared community benchmarking, precisely to facilitate this cyclic process of using brain and cognitive science results to hone in on better and better neural network models of particular domains of natural intelligence (Figure 1) (Schrimpf et al., 2018). As of this writing, Brain-Score 1.0 focuses on visual object intelligence in primates, which is mediated, in large part, by the complex and highly interconnected circuit identified as the ventral visual stream (DiCarlo et al., 2012). Later, we discuss how the same integrative benchmarking approach could be extended to other domains of human intelligence.

Critically, model testing as implemented by Brain-Score is only one step of the cyclic process. The other step in the process—model building (i.e., hypothesis creation)—is just as important, and different strategies exist in our community to develop integrative models, which we also discuss in this perspective.

## Piecewise Efforts Are Only the First Step

Until very recently, it was impossible to study primate visual intelligence using integrative, neurally mechanistic models because no existing model (neurally mechanistic or otherwise) was capable of rivaling behavioral performance of humans or other primates in real-world object recognition. Because of this, the full-scale problem just seemed too hard to take on all at once.

Hence, rather than aiming to understand all of visual intelligence and the underlying neural mechanisms in their totality, the field necessarily had to begin with the study of different sub-parts of the problem. Psychophysicists studied elements of visually driven behaviors (e.g., orientation discrimination), cognitive neuroscientists aimed to elucidate brain sub-regions that might be differentially involved (e.g., in different categories of objects), and visual neurophysiologists provided partial characterizations of neural responses in different areas of the ventral (V1, V2, V4, IT, etc.) and dorsal visual processing streams. The implicit hope was that these segregated endeavors would someday, somehow result in a unified model of the neural mechanisms of primate visual processing and primate visual intelligence more broadly. In sum, our field strongly embraced the "divide and conquer later" approach.

Please cite this article in press as: Schrimpf et al., Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence, Neuron (2020), https://doi.org/10.1016/j.neuron.2020.07.040





Collectively, all these important lines of work and many others form the basis of our current knowledge of visual processing in the brain and visually intelligent behavior. But, for those of us working in this area over the last 20 years, this felt like we were grasping at only individual parts of an elephant, while simultaneously hoping to somehow grasp the bigger picture (Figure 1). In the analogy, each experiment probes a different part: one describes the sharp tusk, another the floppy ears, a third the tail, and so forth. Each individual experiment by itself is unable to obtain an integrative perspective on its own. While none of these experiments and resulting datasets is wrong, when considered in isolation, each is incomplete. In short, each domain of intelligence is just too big for one experiment or one laboratory, and our collective results seem much less than their sum.

We are not dismissing the discovery of building blocks, as this first step importantly shapes the types of hypotheses (models) that are built. But the "divide and conquer later" approach will clearly not be enough. Instead, we here advocate for *now* putting more focus on putting together and implementing at scale the existing principles that our field already has so that they can be collectively tested on an integrative set of benchmarks and determine if and where we are falling short.

## Why Don't We Already Have Unified Neurally Mechanistic Models of Intelligence?

Consider building a new model of visual intelligence in primates. For instance, this model might include a new type of circuit recurrence, more biologically correct single neurons, topographic organization of neural selectivity, a novel unsupervised learning strategy, or a proposed way to make an even simpler, more conceptually or theoretically intuitive model than those currently leading the field. To resolve whether any such new model is "better" than previous models, it must be evaluated on how well it agrees with all relevant empirical measurements.

This is no easy task. Which measurements are the most relevant? Where are those measurements and how does one gain access to them? Further, even if the measurements were available, will it be possible to use them in a way that can be fairly compared with the way that others have attempted to explain those same measurements with other models? Data alone are often not enough, as comparative metrics together with the correct experimental paradigm on models ("benchmarks"; see later) can be difficult to implement without knowing the experiment by heart.

All those hurdles explain why a modeler will usually choose to take the current standard approach: rather than collaborate on unifying models, they will take a more individualistic approach by defining a new, and thus necessarily more narrow, sub-problem that is more tractable. In taking this approach, the individual's only hope is that this sub-problem will reveal some yet-to-be-discovered brain principle, and the job of integrating this sought-after principle and all other principles into a unified model is left for some future day.

We believe that day has arrived for the domain of primate visual intelligence, and we built Brain-Score 1.0 to incentivize and facilitate this collaborative, integrative approach.

## Brain-Score 1.0

To incentivize integrative model building efforts, we propose Brain-Score 1.0: an integrative benchmarking platform for the primate visual ventral stream and its supported behaviors. Now that the ability exists to build end-to-end neurally mechanistic models capable of rivaling primate performance in visual object recognition tasks (see Why Now?), the primary purpose of Brain-Score 1.0 is to provide a central repository where (ideally) all brain data pertinent to the problem of ventral stream processing and object perception behaviors are maintained in a form adequate for model benchmarking, along with the currently leading models to date and their evaluation scores. We refer to this as version 1.0 to indicate that it focuses only on visual intelligence, and, even in that domain, it is still far from where it should ideally be (for example, it currently only includes ventral stream cortical processing, see later).

## Proposed Rules of the Road

Achieving this collaborative goal requires some consensus on standards of what a solution (i.e., a specific model, as motivated above) must minimally be able to do and explain. In the domain of visual intelligence (of which the ventral visual stream supports a still unknown portion), models must make the following three commitments ("rules of the road"):

(1) Take input that is the same as the input to the primate retina (spatiotemporal patterns of spectral luminance—i.e., images and movies). We decide in favor of pixel input over stimulus parameters because pixels are applicable to virtually any experiment whereas the set of possible parametrizations would only grow longer as more experiments are added.

(2) Be built using only internal parts that approximate those of biological neural networks, with mapping commitments from internals of the model to biological brain regions and actual neurons (that is, "black box" components are not allowed). Because the brain is a biological neural network (plus its support structures), the most accurate, neurally mechanistic model of any domain of intelligence must be within the class of all possible artificial neural networks. That model might look very different from the artificial neural networks in use today, but the general principle of networks of neurons (the neuron doctrine) will prevail.

(3) Be capable of attempting the same behavioral tasks that are performed in visually guided experiments. Note that, taken together, this set of tasks and their natural generalizations is, in effect, Brain-Score's current operational definition of the "domain" of visual intelligence: an operational definition that is not closed, but can—and must—naturally expand as new tasks are tested and added.

## The Three Key Elements of Brain-Score

### *A Software Standard Implementation for Brain Models*

The first key element of the Brain-Score 1.0 platform is the adoption of the rules of the road (above) via software that helps modelers make these commitments on any proposed model and enforces these commitments in model evaluation. Specifically, the Brain-Score software platform defines a common interface that

Please cite this article in press as: Schrimpf et al., Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence, Neuron (2020), https://doi.org/10.1016/j.neuron.2020.07.040



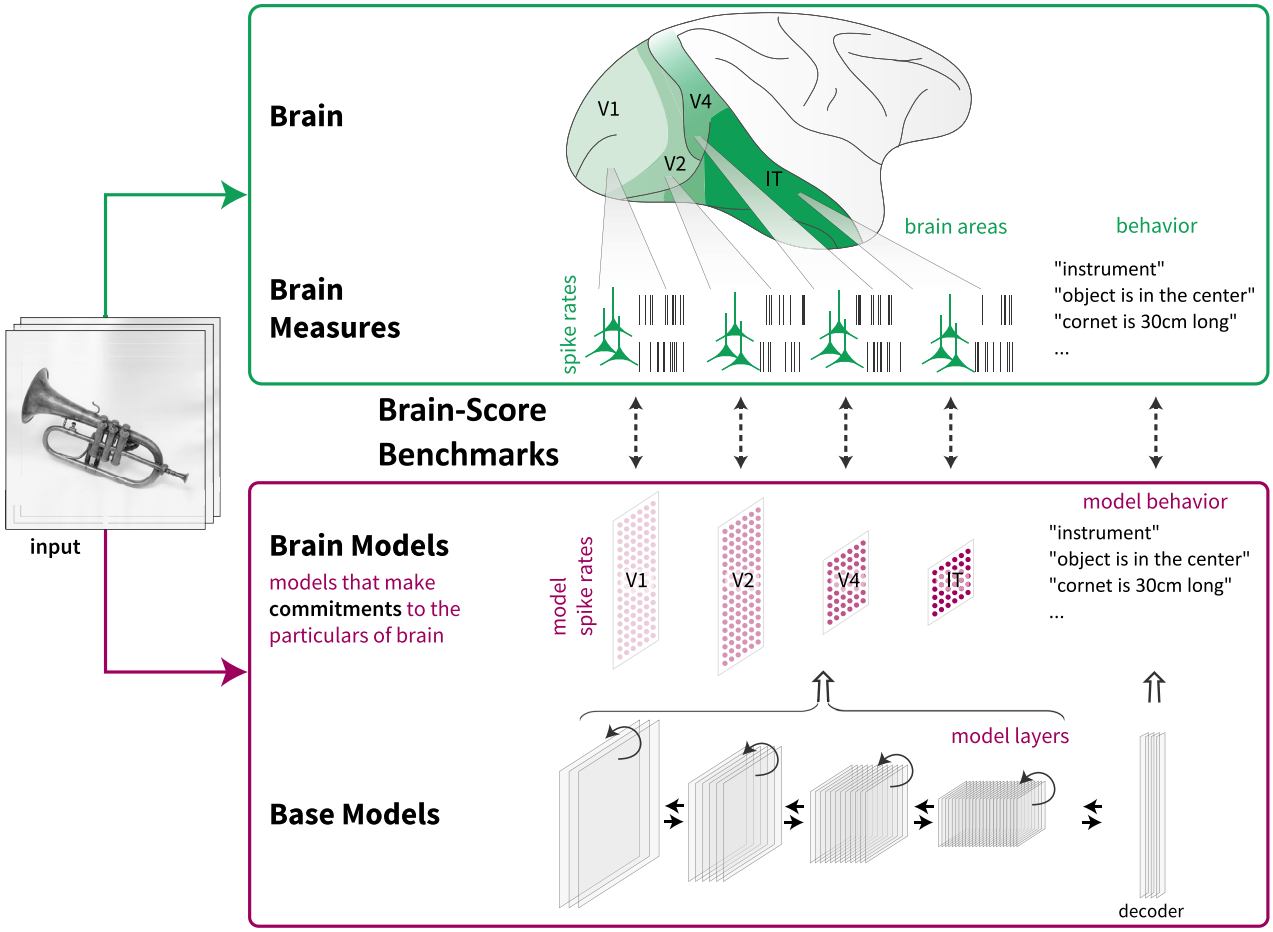


**Figure 2. Brain Models**

The set of Brain Score benchmarks evaluates the match of Brain Models on their match to Brain Measures. Brain Measures can entail, for example, neural spike rates and behavioral responses, which together represent the key internal and external functional measures of the underlying system of interest—here, the ventral visual stream (top green panel). The brain areas shown here are not exhaustive of the brain regions critical to visual intelligence but are simply a starting point. Brain Models are *in silico* brain hypotheses that can be experimented on like a natural brain, for instance, by "recording" from area V1 or IT or by probing it for behavioral responses. Since most machine learning models (termed Base Models) do not make the commitments necessary for such experimentation, Brain-Score provides an API and tools for conversion. These entail translating pixels into degrees visual angle, layers into brain regions, artificial neuron activations into spike rates, probabilities into behavioral output, and so on (bottom red panel).

all candidate models should follow to be evaluated. Models that follow that common API are referred to as Brain Models (Figure 2). This common API forces commitments on, for instance, how many of the model's first layer of neurons correspond to a degree of visual angle, which model layers correspond to which brain regions, and how model neuronal activations are scaled to biological neural firing rates, so that the Brain Model makes unequivocal predictions.

Because of this common API, candidate Brain Models in the Brain-Score platform can unequivocally be compared and become easy to experiment on as if they are biological subjects: showing stimuli, recording, perturbing, and so forth become straightforward. Since all models follow the same interface, this solves the technical differences and commitment ambiguities presented by the large pool of possible candidate models in the machine learning and computer vision communities.

Brain-Score is therefore agnostic to the exact implementation: it does not matter if it is a deep network from computer vision, or a hand-designed model, as long as it makes predictions on experiments by implementing the API.

One downside of this Brain Model standardization is that artificial neural network (ANN) models from other communities (e.g., computer vision) may not be instantly available as possible hypotheses about primate visual intelligence, but this (slight) cost of translation of the newest such models to Brain Models is unavoidable for the overarching goal of integrative and fair quantitative engagement with a wide range of experimental data and the resulting model-to-model comparisons. To help modelers overcome this energy barrier, the Brain-Score platform provides a standard implementation to convert their models (in standard machine learning frameworks, e.g., PyTorch, TensorFlow, Keras) to Brain Models.

Please cite this article in press as: Schrimpf et al., Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence, Neuron (2020), https://doi.org/10.1016/j.neuron.2020.07.040





Given these tools, if a modeler is still not (yet) interested in making such API commitments for their models, then their models cannot (yet) be evaluated as a model candidate by Brain-Score. This is the platform's reflection of one of its core scientific values: Hypotheses of the neural mechanisms of any domain of intelligence must be falsifiable, and the Brain Model API formalizes this for the primate ventral stream.

### Benchmarks for Comparison to Experimental Data

The second key element of Brain-Score is its set of experimental benchmarks that can be—and must be—continually expanded (see Figure 1). Each such benchmark relies on experimental data (e.g., neural activity, behavioral responses, etc.), but each is importantly more than those data. Specifically, each benchmark specifies the steps to reproduce the experiment on a candidate Brain Model and applies a particular unambiguous metric to compare resulting model data with experimental data. Thus, all models can be directly compared as they are tested under the same conditions and with the same set of benchmarks. We emphasize that data alone are not enough and that benchmarks are necessary to engage with integrative models.

Examples of benchmarks include single neural predictions in particular visual areas to particular image sets, distributions of tuning functions in each visual area, patterns of behavioral performance in particular visually guided tasks, and so on. The intent of the platform is to (1) adopt many (or ideally all) of the experimental benchmarks that are already implicit in the field but have not yet been formalized, standardized, and aggregated, and (2) to establish a platform for rapid deployment of new benchmarks as new experimental data become available. Multiple benchmarks might derive from the same underlying experimental data (for example, each using a different metric and/or focusing on a particular subset of the data). Depending on the experimental data in question, the computational construction of each benchmark will usually require some data processing to render into standard formats, equate measurement conventions, clarify in detail experimental conditions, expose and account for the uncertainty range in each measurement, etc.

For many brain scientists, the term "benchmark" may sound unusual or inappropriate. However, the *motivation* of benchmarks is not new to brain science: the operations that are applied to experimental data to make a benchmark are exactly analogous to the analyses that experimental brain scientists have decided are important to extracting meaning from data. This has two important consequences: first, it means that Brain-Score implicitly adopts the prior beliefs of experimentalists in the field, thus respecting and incorporating the added value of those experimental efforts rather than simply asking for experimentalists' raw data and starting from scratch. Examples include choices of firing rate analysis windows (e.g., mean rates in a latency-adjusted time window), aggregating particular conditions (e.g., all face images) and contrasting particular conditions (e.g., tuning sharpness over visual gratings), summarizing over population measures (e.g., RDMs), and many others. Second, it means that Brain-Score does not always need to start with new experiments (Marques and DiCarlo, 2019) or with entirely raw experimental data buried deep in the bowels of individual laboratories but can instead harvest many benchmarks from the published literature.

However, what is almost entirely new to brain science is the formal *practice* of benchmarking. Specifically, once the data and metrics have been defined and standardized in the Brain-Score platform, each resulting benchmark is a valuable model test (a.k.a. model constraint) that can now be accessed by any interested modeler via checking their model's Brain-Score on that particular benchmark (below). Facilitating the actual practice of benchmarking in brain science is a major goal of the Brain-Score platform.

### Scores to Quantify and Guide Progress

For each and every candidate Brain Model, the Brain-Score platform produces a concise summary evaluation on each and every benchmark in the form of an explicit score on each benchmark (e.g., a score for predicting IT neural responses to a set of images), an explicit aggregated score over related sets of benchmarks (e.g., an overall IT score), and an explicit overall score (for Brain-Score 1.0, this is an overall ventral stream score). These scores precisely communicate up-to-date model accomplishments, enable direct comparison between models, and thus can be used to guide new models.

To reproducibly score models on benchmarks, we share Brain Models and benchmarks in the form of executable code that removes any ambiguities.

## Brain-Score Challenges and Proposed Solutions

To take these goals seriously, we need to implement the three key elements of Brain-Score (above), and in doing so, choices need to be made and challenges need to be overcome:

### How Do We Choose the Experimental Benchmarks?

Within a domain of intelligence, many benchmarks (i.e., a particular set of experimental data with a particular choice of metric) might be considered. Choosing the right benchmarks to score models is a difficult question, and the choices here might easily begin to reflect the slightly different goals that our community has previously been interested in, and thus there is the danger of taking us back toward piecewise efforts (Figure 1). Because of this, we adopt an inclusive view: for any given domain of intelligence, we endorse all benchmarks that seem even distantly relevant to that domain because we believe that, taken together, these will converge to a set of benchmarks that push our field toward an accurate unified model. In practice, placing benchmarks on the Brain-Score platform requires effort; thus, those researchers that are willing to participate and help in this effort will tend to implicitly shape benchmark priorities. We focus most on the benchmarks that will most differentiate between models, but generally try to incorporate all benchmarks our field produces because it is still useful to know if all or no models capture a particular benchmark.

In Brain-Score 1.0, which focuses on the primate ventral visual stream, we include benchmarks with both neural and behavioral datasets in the context of the ventral visual stream because we are ultimately interested in how neurons mechanistically implement visual behaviors. We believe that the metrics and conceptual setup can be naturally transferred to other domains of intelligence (see Looking Ahead). At the moment, the neural measurements are electrode recordings from non-human

Please cite this article in press as: Schrimpf et al., Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence, Neuron (2020), https://doi.org/10.1016/j.neuron.2020.07.040



primates—the ventral stream of human primates has been shown to be near identical (Kriegeskorte et al., 2008), and we believe electrode recordings are the least filtered and most direct measurements and thus added them first. Brain-Score, however, is not restricted to this recording modality, and measurements such as fMRI are good candidates for the next set of benchmarks that the community could implement. The current behavioral datasets are match-to-sample tasks performed by human subjects, but we again see this as only a small beginning and are hoping the community will add benchmarks based on a variety of visual tasks.

The difficulty in being inclusive of many different kinds of neural (e.g., recording modalities) and behavioral (e.g., tasks) benchmarks is that models need to implement all of them in order to be scored. As stated in The Three Key Elements of Brain-Score, we provide software to help modelers as much as possible, but the choices we make may be sub-optimal as we are not domain experts in all types of ventral stream measurements. To keep things manageable, we currently restrict neural data to neural recordings at the anatomical specificity of visual areas (e.g., V1, IT) from both human and non-human primates. In the future, we might also include neural perturbation benchmarks (e.g., stimulation or lesion studies). Due to the slightly different anatomy between humans and macaques, we anticipate models might have to provide per-species commitments to anatomy. Behavioral data of any visual task are accepted, but we hope behavioral benchmark submissions will provide software that enables models to attempt the same task that the biological subjects were required to perform.

Further, our current belief is that for many—but not all—applications, we may be able to abstract away measurements below the level of spikes (see later). Thus, the initial Brain-Score neural benchmarks are all derived from spiking measures. Once models at this level of abstraction are powerful enough, expanding them into even more spatially precise regimes would allow new progress on connecting to measures of sub-cellular and molecular processes, which would in turn enable new model-guided applications at the molecular and genetic levels.

### Will Experimentalists Submit Benchmarks to Brain-Score?

In our experience thus far, a lack of experimental data is not the most pressing problem. A successful strategy has been to simply ask labs if they would be willing to share data (Majaj et al., 2015; Rajalingham et al., 2018; Cadena et al., 2019; David et al., 2004; Freeman et al., 2013; Kar et al., 2019), and modelers interested in evaluating their models (i.e., the "second person" working with the data) then do the work of building benchmarks (Schrimpf et al., 2018; Nayebi et al., 2018; Kubilius et al., 2019). We further give the option of using data only as a private benchmark, that is, models can be run on the benchmark and can obtain a score, but the data itself are not released. In practice, collaborating with other groups to develop new benchmarks has worked well so far, and many researchers have contributed their data.

### Should We Summarize All Benchmark Scores in a Single Brain-Score?

The spirit of integrative benchmarking is that models we most value are those that score well on all of the available benchmarks (see Introduction). Incentivizing this requires some kind of aggregation of scores across *all* benchmarks (e.g., the mean of all scores, the worst of all scores, etc.). Of course, we could choose to weight the scores differently, which makes explicit which benchmarks we (the field) think are most important. Because that seems arbitrary at the moment, Brain-Score 1.0 computes a neutral (equal) weighting of all scores. Going forward, one idea could be to hierarchically aggregate scores based on associated region and to then equally weigh those regions; that is, there will be an overall V1 score, which is the mean of a spatial frequency score, a center surround score, etc., which in turn are again the mean of specific benchmark scores.

As more benchmarks are accumulated, we hope this issue will initiate active community-wide discussions and meetings as to which benchmarks matter most and why. Researchers in each domain of intelligence may not all agree about the relative importance of each benchmark, but not yet having a common discussion ground such as Brain-Score signals that we do not even care that we may not agree. And divided we fail. In short, figuring out how to best aggregate (i.e., weight) the individual benchmarks is part of the scientific discussion Brain-Score intends to facilitate.

### Will Anyone Submit Models to Brain-Score?

Brain-Score offers a straightforward way to empirically demonstrate model match-to-brain, minimizing the need for implementing metrics and gathering data (see Why Don't We Already Have Unified Neurally Mechanistic Models of Intelligence?). While empirical benchmarks to compare models are abundant in machine learning, the same is not true in neuroscience (but see, e.g., Neural Prediction Challenge, Algonauts [Cichy et al., 2019]), and to our knowledge, Brain-Score is currently the largest-scale benchmarking platform to do so.

Such benchmarks allow modelers to be rewarded for the most brain-like model, and our hope is that this recognition will draw engineering talent and effort toward model building for neuroscience, rather than only competing on machine learning benchmarks. For the modeling community, successful models on Brain-Score can also be used as good starting points for new models and changes to the model evaluated by tracking changes in the scores.

### How Can We Deal with the Possibility that Models May Overfit Some or All Brain-Score Benchmarks?

One should rightly worry that a high benchmark score might indicate the ability of the model to re-express the results that align with the benchmark, rather than the ability of the model to also capture a similar benchmark if new data were collected (i.e., new subjects, new neurons, new images, etc.). In machine learning, this situation is referred to as "overfitting." In particular, by Brain-Score making public the test scores of models on a benchmark, over successive model development cycles, newer models might start overfitting that benchmark. We propose to guard against this in at least four ways: (1) restrict the number of test scores a model submitter can obtain (for instance, only weekly); (2) estimate generalization errors in the benchmark scores to ensure new data are likely to be captured equally well, e.g., by cross-validation; (3) dynamically add more and more benchmarks to strongly constrain the model and discourage overfitting to single experiments—in other words,

Please cite this article in press as: Schrimpf et al., Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence, Neuron (2020), https://doi.org/10.1016/j.neuron.2020.07.040





incentivize generalization to new benchmarks; (4) embrace the idea that if some model in the future has "fit" enough benchmarks and generalizes well to the newest benchmark, it might have "overfit" the *entire* domain in question. At that point we can declare success for the goal stated at the outset of the perspective—finding accurate neurally mechanistic models of the domain.

### Who Is Going to Maintain the Brain-Score Platform?

While most programming libraries' core developers are paid for their work by their respective companies, comparable grants in academia for software engineers are only just starting. In practice, a similar setup can be deployed by a group of labs: each diverts some of their resources to pay for a developer maintaining the platform; in return, their researchers can more efficiently compare models or test them on benchmarks and benefit from premium support. We further hope the community will actively contribute to the open-source code base by adding metrics, data, benchmarks, and models and by improving the technical infrastructure.

## Why Now?

Putting together the principles we have already learned into large-scale, integrative models has arguably been the next step in computational neuroscience for some time. In the domain of visual intelligence, we see two reasons that now is the right time to take this step.

### Recent Availability of Mechanistic Models of Visual Processing

Visual neuroscience modeling traditionally comprises proposals of many different kinds of model functions: scientists studying LGN have, for instance, emphasized gain control models through lateral interactions, while others studying V1 have focused on edge detection and sparsity constraints, and others studying V2, V4, and IT have put forward texture, curvature, and face models, respectively. None of these perspectives is wrong, though all are incomplete. Attempts at integrative models of vision such as HMAX (Riesenhuber and Poggio, 1999) were initially promising, but at the time swayed by shortcomings in the model's behavioral performance (relative to humans).

However, the situation has changed drastically in recent years. Thanks to the accumulating advances in the ability to scale deep neural networks, we now have "end-to-end" models that rival human performance over the entire domain of visual object recognition tasks. Indeed, several neuroscience labs (Yamins et al., 2014; Cichy et al., 2016; Nayebi et al., 2018; Kubilius et al., 2019) have been able to develop end-to-end models of visual object recognition that (1) take pixels as input, (2) produce behavioral categorizations as outputs, (3) are generally consistent with the ventral stream architecture, and (4) demonstrate internal "unit" response properties at each level of the network that are individually fairly similar to actual neurophysiological unit responses and other brain measurements at the corresponding levels.

Crucially, these models, while still far from complete, are the first neurally mechanistic models that are built at scale (that is, they can, within limits, take any retinal image as input) and that can rival the behavioral performance of primates in this subdomain of visual intelligence (object recognition). By now, there are dozens of such models (e.g., Krizhevsky et al., 2012; He et al., 2015; Huang et al., 2017; Howard et al., 2017; Zoph and Le, 2017; Bassett et al., 2018; Richards et al., 2019; Cichy and Kaiser, 2019). How can we determine which of these models best captures brain processing? Even more importantly, because all of the existing models are almost surely incorrect in some way, how do we organize our data and ongoing experiments to incentivize the next generation of these at-scale, neurally mechanistic models? Brain-Score is our answer.

### Increasing Availability of Brain Data Benchmarks

In addition to the availability of neurally mechanistic end-to-end models, we are also starting to see an increasing availability of brain data and an increase in ways to compare data with neurally mechanistic models (referred to as "metrics"). Years of research in our community has resulted in many datasets, and with advanced recording techniques, the pace of acquiring new data is gaining momentum (Hong and Lieber, 2019): more neurons can be recorded simultaneously, at multiple cortical depths; recording durations are increasing from days to months; and upcoming wireless recording techniques will allow large amounts of data to be collected rapidly. Similarly, low-cost, large-scale behavioral experiments utilizing online platforms have already become ubiquitous, allowing collection of behavioral measurements from hundreds of subjects. Given the availability of data, combining individual experiments into an integrative set of benchmarks that constrain model candidates is now a possibility.

The ImageNet competition—a large set of annotated images and associated performance benchmarks (Deng et al., 2009)—has paved the way for better and better computer vision models, and we hope Brain-Score will incentivize similar modeling efforts in computational neuroscience that result in better and better Brain Models (Figure 3). Even with only the initial set of data currently in Brain-Score, the first success of guided model building has already been shown with a new model, CORnet (Kubilius et al., 2019): a compact hierarchical recurrent neural network that more closely follows brain anatomy (few layers and recurrent). But because of Brain-Score, we also know that this model is not yet meeting all existing benchmarks. Even more importantly, with the availability of the Brain-Score platform, we are looking forward to CORnet soon being surpassed by the next generation of neurally mechanistic models from another group in the field.

## Building the Next Generation of Integrative Models

As stated in Proposed Rules of the Road, neurally mechanistic models of intelligence must be within the class of all possible artificial neural networks, although they might look very different from today's models. Engineering accurate such models currently involves a lot of guesswork, and we propose to make use of the history of scores across an integrative set of benchmarks to provide more guidance for this process.

Many early models emerging from the "divide and conquer later" approach tended to be relatively simple and low in the number of parameters (e.g., models of V1 neural responses were simple functions of a fitted, preferred orientation angle)—

Please cite this article in press as: Schrimpf et al., Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence, Neuron (2020), https://doi.org/10.1016/j.neuron.2020.07.040



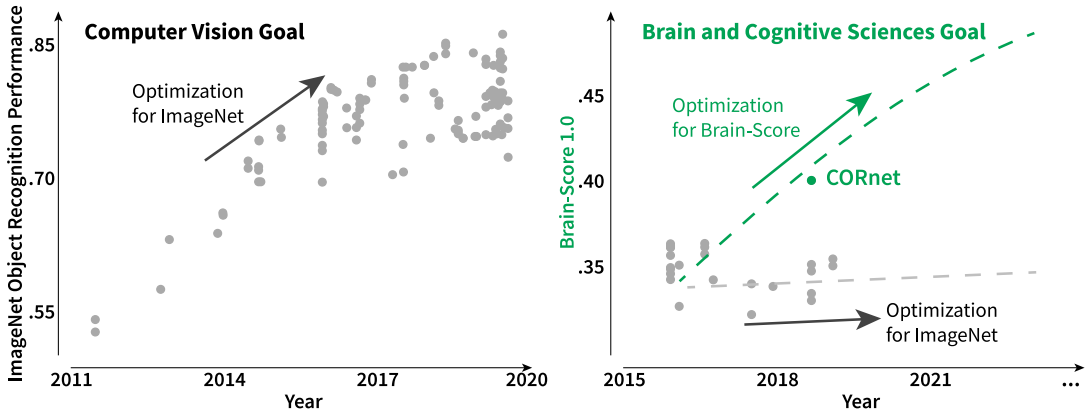


**Figure 3. Incentivizing Integrative Brain Modeling**

In engineering and machine learning, a common evaluation of each model's ability to perform well on a large set of recognition tasks (ImageNet) helped incentivize computer vision out of toy problems into more and more powerful computer vision systems that generalized to real-world challenges (today's "AI"). We believe that Brain-Score will have a similar effect on computational neuroscience—leading our field to higher-fidelity, more generalizable models in neuroscience. Our first hint of success of this strategy is with the model CORnet, which was built with guidance from neuroscience (recurrence), and because its Brain-Score is higher than prior models, that model building effort is thus rewarded. While we do not presume to know the best next modeling steps, we can—and must—create incentives (Brain-Score) that nourish and reward the best models, regardless of where they originate.

perhaps as a consequence of the relatively narrow empirical scope of each sub-problem. But additionally, developing simple models borne out of narrow empirical domains might have been preferred *a priori* because of the perception that science has a long history of success of this approach. Under this approach, mathematically or conceptually elegant descriptions of nature extracted from sub-parts of some natural domain become the basis for capturing the full domain (e.g., planetary motion from Newton's laws of mechanics). We refer to this as a "principles-driven" approach to modeling and note that any such simple model must eventually be scaled up to engage with the full extent of a domain of human intelligence.

We view as an alternative approach "reverse-engineering" (DiCarlo, 2018): effectively searching through the very large class of all possible neural network models by iteratively improving the current best model, based on guidance from benchmarks. Since the delta in benchmark scores reveal model changes that are most effective at improving model accuracy, these model changes can then be iteratively fleshed out and improved upon in a virtuous cycle of model engineering and model testing against experimental benchmarks. But even though Brain-Score can provide these delta "gradients" over model generations, it does not reveal the *specific* next steps to improve models. Should a new normalization be added? Does topography improve behavioral match? Whether and how detailed changes to the models can be inferred from benchmarks and past model scores is an open challenge. The specific improvements to a model in the reverse-engineering approach are thus a combination of manual and automated search of new architectural components, training objectives, loss functions, and so on—but, importantly, all under the guidance and constraint of myriad experimental benchmarks.

Ultimately, the best models might come out of a combination of principles-driven and reverse-engineering approaches where model components are first roughly defined (principles-driven) and then scaled up and their hyper-parameters iteratively optimized (reverse-engineering). For any newly proposed unified model from either combination of approaches, an integrative benchmarking platform such as Brain-Score is vital in order to evaluate which models are the most accurate on a wide range of benchmarks.

## Looking Ahead

We see integrative benchmarking as the next step to building neurally mechanistic models of domains of human visual intelligence. Summarizing the many advantages stated throughout the text, we argue that integrative benchmarking can inform model development with the history of comparable and reproducible scores over generations of models. These scores will provide a useful gradient that will point out which changes have led to which kinds of improvements, and the availability of the benchmarking platform will rally more human talent to the cause, resulting in new, more accurate models (Figure 3). Some experimental benchmarks will turn out to be harder for models to meet than others, and we believe that Brain-Score will help create a future in which experimentalists are lauded for their ability to produce new benchmarks that separate competing models or that show that all current models are inadequate. Most importantly, this friendly competition between models and experimental benchmarks will drive the scientific cycle of "strong inference" (Platt, 1964), wherein new benchmarks dissociate models while new models aim to integratively capture all benchmarks. Each new model will be thus be a closer and closer approximation of the true neural mechanisms of the domain of intelligence, with all the attendant benefits of discovering that truth (see Are Accurate In Silico Models Useful on Their Own?).

### *Next Steps for Brain-Score*

The Brain-Score platform has seen many updates since its inception in late 2018, such as more data, public benchmarks, and an automated submission system. While we are excited

Please cite this article in press as: Schrimpf et al., Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence, Neuron (2020), https://doi.org/10.1016/j.neuron.2020.07.040





about what has been accomplished, we were hoping to have accomplished more since then. The major hurdles have been of a technical nature: we first had to conceptually define a model interface that would allow for many different interactions while being straightforward to use, set up infrastructure to, e.g., automatically score models, and develop code that would be easy to use and adapt for the community. These things took a lot of time but were necessary to scale to many models and many benchmarks.

The six scored benchmarks currently live on Brain-Score are way too few of what we think we need—yet they are the biggest set of benchmarks that we are aware of. And each has hundreds to thousands of individual comparison points (image level comparisons). To add many more benchmarks, we can also turn to published results with a recently developed framework to digest papers into quantitative benchmarks (Marques and DiCarlo, 2019). Applying this framework together with community contributions makes for 23 new benchmarks that are in the pipeline and will be released soon. We hope the community will add many more.

### *Can Integrative Benchmarking and Modeling Be Applied to Other Domains of Intelligence besides Vision?*

While we believe that all domains of human intelligence must ultimately be captured as unified models discovered via integrative benchmarking, some domains are more ready than others to do this. Specifically, domains that fulfill the following criteria are, in our view, ready to take this next step:

(1) Sensory inputs to the system need to be clearly defined (e.g., in visual object recognition [OR], patterns of photons striking the central retina, approximated as pixels).
(2) Important outputs need to be clearly defined in the form of behaviors (e.g., in OR, choice among alternative names, visual search for target object, etc.).
(3) The parts of the brain primarily involved in performing these behaviors in response to inputs need to be reasonably well established (e.g., in OR, the ventral stream).
(4) Initial neural and behavioral data exist, and techniques exist to easily collect more (e.g., in OR, chronic array recordings along the ventral stream, Amazon Mechanical Turk behavioral experiments, etc.).
(5) First models exist that (1) accept the defined system input, (2) perform at least some of the behaviors, (3) use internal parts that can be physically mapped to parts in the involved brain areas (i.e., the models are neurally mechanistic in that sense), and (4) explain and predict some non-trivial portion of the data.

### *Why Emphasize* In Silico *Models over Principles and Theory?*

One core belief in this perspective is that new *in silico* models of visual intelligence can be built based on the ideas, concepts, principles, and small-scale models our field has already produced. That is, we propose that the next important phase in our field may not depend on first discovering new conceptual models, new principles underlying neural processing, or a unified theory of brain function. Indeed, models (engineered systems) have often preceded theoretical understanding in the history of science and technology (LeCun, 2017): for instance, in physics, Kepler's empirical laws of planetary motion building on Brahe's careful measurements preceded Newton's more general laws of physics by over 80 years (Ajemian and Hogan, 2010). More recently, certain deep neural networks have proven to be surprisingly accurate models of each and every layer of the ventral visual stream and visual object recognition behavior, despite lacking a unifying theory for those models. Following the practical success of these models, they are now the subject of much theoretical work (Banburski et al., 2019; Golowich et al., 2018; Belkin et al., 2019; Casper et al., 2019). Similarly, Brain Models are already starting to inform theory: for instance, with model reduction methods of retina models (Tanaka et al., 2019).

A practical argument is even more compelling: we (scientists) have no way to possibly know if our field's current ideas, concepts, principles, and theories are sufficient to explain the phenomena of interest (neural and behavioral data) in a domain of intelligence until our existing principles are combined into *in silico* models that are scaled up for that domain and a wide-ranging collection of neural and behavioral benchmarks is brought to bear to evaluate each model's successes and failures. That is, we see no practical way forward on testing our field's current concepts and principles without something akin to Brain-Score.

Our overarching belief is that a science of any domain of natural intelligence will almost surely move through—but likely not end with—accurate neurally mechanistic models of that domain. This next movement—accurate integrative *in silico* models—is what Brain-Score aims to incentivize and accomplish, which will lay the foundation for theorists to later build upon.

### *Are Accurate* In Silico *Models Useful on Their Own?*

Even without a complete set of principles and theories of brain processing (above), accurate simulations of the brain's working (i.e., *in silico* unified models) have far-reaching practical applications. For example, focusing on vision, such possible applications include:

*Research:* In the ventral visual stream, models of this type are already being used to focus experimental resources on the most interesting aspects of brain function that are not yet accurately described (Tang et al., 2018; Kar et al., 2019; Hénaff et al., 2019; Kietzmann et al., 2019; Golan et al., 2019). In addition, by drawing on the predictive accuracy of these models, neuroscientists can now use them to control individual neurons and entire populations of neurons deep in the visual system via model-synthesized patterns of light applied to retinae (Bashivan et al., 2019; Ponce et al., 2019).

*AI:* As our field discovers models that are ever more closely aligned to primate brains and human behavior, we will in fact be discovering machine systems that (e.g.) successfully generalize more like humans, are less susceptible to adversarial attacks, and potentially more energetically efficient.

*Brain-Machine Interfaces:* Sufficiently accurate integrative models of visual processing can be used to determine complex, non-intuitive microstimulation patterns that should be applied in mid- and high-level visual areas to replicate visual percepts (e.g., in blind individuals).

*Brain Disorders:* For most disorders, the treatment goal is to precisely modulate brain activity in a helpful way. While it is commonly assumed that such interventions will be best delivered via new pharmaceuticals (difficult to target precisely) or

perhaps with inserted probes (dangerous and still not precise), accurate *in silico* models might reveal entirely new treatment possibilities: for instance, by directing the synthesis of patterns of light delivered to the retina that predictably and precisely modulate entire populations of neurons deep in the brain at single-neuron resolution.

The key overall point is that all of the above applications—and myriad others not yet imagined—may not require new principles, theories or even "understanding" (however that is defined for each of us). But each potential application area will get ever better with ever more accurate *in silico*, neurally mechanistic models. Brain-Score incentivizes the discovery of those models in domains of intelligence that are ripe to do so.

## ACKNOWLEDGMENTS

We thank Tiago Marques, Kohitij Kar, Rishi Rajalingham, Kamila Jozwik, Daniel Yamins, Pouya Bashivan, Elias Issa, and Arash Afraz for useful discussions. We thank Najib Majaj, Ha Hong, Rishi Rajalingham, Kohitij Kar, Stephen David, Jack Gallant, Corey Ziemba, Anthony Movshon, Santiago Cadena, Andreas Tolias, Shashi Kant, Mengmi Zhang, Gabriel Kreiman, and Ilya Kuzovkin for being among the first to share their data through Brain-Score.

This work was supported in part by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 705498 (J.K.), the SRC Semiconductor Research Corporation (M.S., M.J.L.) and DARPA, Simons Foundation grant SCGB-542965 (J.J.D.), Office of Naval Research MURI-114407 (to J.J.D.), and McGovern Institute for Brain Research.

## REFERENCES

Ajemian, R., and Hogan, N. (2010). Experimenting with theoretical motor neuroscience. J. Mot. Behav. *42*, 333–342.

Banburski, A., Liao, Q., Miranda, B., Rosasco, L., Hidary, J., and Poggio, T. (2019). Theory III: Dynamics and Generalization in Deep Networks – a simple solution. arXiv, 1903.04991.

Bashivan, P., Kar, K., and DiCarlo, J.J. (2019). Neural population control via deep image synthesis. Science *364*, eaav9436.

Bassett, D.S., Zurn, P., and Gold, J.I. (2018). On the nature and use of models in network neuroscience. Nat. Rev. Neurosci. *19*, 566–578.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias-variance trade-off. Proc. Natl. Acad. Sci. USA *116*, 15849–15854.

Cadena, S.A., Denfield, G.H., Walker, E.Y., Gatys, L.A., Tolias, A.S., Bethge, M., and Ecker, A.S. (2019). Deep convolutional models improve predictions of macaque V1 responses to natural images. PLoS Comput. Biol. *15*, e1006897.

Casper, S., Boix, X., D'Amario, V., Guo, L., Schrimpf, M., Vinken, K., and Kreiman, G. (2019). Removable and/or Repeated Units Emerge in Overparametrized Deep Neural Networks. arXiv, 1912.04783.

Cichy, R.M., and Kaiser, D. (2019). Deep Neural Networks as Scientific Models. Trends Cogn. Sci. *23*, 305–317.

Cichy, R.M., Khosla, A., Pantazis, D., Torralba, A., and Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. Sci. Rep. *6*, 27755, https://doi.org/10.1038/srep27755.

Cichy, R.M., Roig, G., Andonian, A., Dwivedi, K., Lahner, B., Lascelles, A., Mohsenzadeh, Y., Ramakrishnan, K., and Oliva, A. (2019). The Algonauts Project: A Platform for Communication between the Sciences of Biological and Artificial Intelligence. arXiv, 1905.05675.

David, S.V., Vinje, W.E., and Gallant, J.L. (2004). Natural stimulus statistics alter the receptive field structure of v1 neurons. J. Neurosci. *24*, 6991–7006.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. IEEE Conference on Computer Vision and Pattern Recognition (IEEE), 248–255.

DiCarlo, J.J. (2018). To Advance Artificial Intelligence, Reverse-Engineer the Brain. https://www.wired.com/story/to-advance-artificial-intelligence-reverse-engineer-the-brain/.

DiCarlo, J.J., Zoccolan, D., and Rust, N.C. (2012). How does the brain solve visual object recognition? Neuron *73*, 415–434.

Freeman, J., Ziemba, C.M., Heeger, D.J., Simoncelli, E.P., and Movshon, J.A. (2013). A functional and perceptual signature of the second visual area in primates. Nat. Neurosci. *16*, 974–981.

Golan, T., Raju, P.C., and Kriegeskorte, N. (2019). Controversial stimuli: pitting neural networks against each other as models of human recognition. arXiv, 1911.09288.

Golowich, N., Rakhlin, A., and Shamir, O. (2018). Size-independent sample complexity of neural networks. arXiv, 1712.06541.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep Residual Learning for Image Recognition. arXiv, 1512.03385.

Hénaff, O.J., Goris, R.L.T., and Simoncelli, E.P. (2019). Perceptual straightening of natural videos. Nat. Neurosci. *22*, 984–991.

Hong, G., and Lieber, C.M. (2019). Novel electrode technologies for neural recordings. Nat. Rev. Neurosci. *20*, 330–345.

Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv, 1704.04861.

Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K.Q. (2017). Densely connected convolutional networks. Computer Vision and Pattern Recognition, 2261–2269.

Kar, K., Kubilius, J., Schmidt, K., Issa, E.B., and DiCarlo, J.J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. Nat. Neurosci. *22*, 974–983.

Kietzmann, T.C., Spoerer, C.J., Sörensen, L.K.A., Cichy, R.M., Hauk, O., and Kriegeskorte, N.; Proceedings of the National Academy of Sciences (2019). Recurrence is required to capture the representational dynamics of the human visual system. Proc. Natl. Acad. Sci. USA *116*, 21854–21863.

Kriegeskorte, N., Mur, M., Ruff, D.A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., and Bandettini, P.A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron *60*, 1126–1141.

Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Neural Information Processing Systems (NIPS). arXiv, 1102.0183.

Kubilius, J., Schrimpf, M., Hong, H., Majaj, N.J., Rajalingham, R., Issa, E.B., Kar, K., Bashivan, P., Prescott-Roy, J., Schmidt, K., et al. (2019). Brain-like object recognition with high-performing shallow recurrent ANNs. arXiv, 1909.06161.

LeCun, Y. (2017). My take on Ali Rahimi's "Test of Time" award talk at NIPS. https://www2.isye.gatech.edu/~tzhao80/Yann_Response.pdf.

Majaj, N.J., Hong, H., Solomon, E.A., and DiCarlo, J.J. (2015). Simple Learned Weighted Sums of Inferior Temporal Neuronal Firing Rates Accurately Predict Human Core Object Recognition Performance. J. Neurosci. *35*, 13402–13418.

Marques, T., and DiCarlo, J.J. (2019). A meta-analysis of current DNNs as models of low-level visual processing. Bernstein Conference. https://doi.org/10.12751/nncn.bc2019.0088.

Nayebi, A., Bear, D., Kubilius, J., Kar, K., Ganguli, S., Sussillo, D., DiCarlo, J.J., and Yamins, D.L. (2018). Task-driven convolutional recurrent models of the visual system. Advances in Neural Information Processing Systems, 5295–5306.

Platt, J.R. (1964). Strong Inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others. Science *146*, 347–353.

Ponce, C.R., Xiao, W., Schade, P.F., Hartmann, T.S., Kreiman, G., and Livingstone, M.S. (2019). Evolving Images for Visual Neurons Using a Deep

Generative Network Reveals Coding Principles and Neuronal Preferences. Cell *177*, 999–1009.e10.

Rajalingham, R., Issa, E.B., Bashivan, P., Kar, K., Schmidt, K., and DiCarlo, J.J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. J. Neurosci. *38*, 7255–7269.

Richards, B.A., Lillicrap, T.P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R.P., de Berker, A., Ganguli, S., et al. (2019). A deep learning framework for neuroscience. Nat. Neurosci. *22*, 1761–1770.

Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. Nat. Neurosci. *2*, 1019–1025.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N.J., Rajalingham, R., Issa, E.B., Kar, K., Bashivan, P., Prescott-Roy, J., Schmidt, K., et al. (2018). Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? bioRxiv. https://doi.org/10.1101/407007.

Tanaka, H., Nayebi, A., Maheswaranathan, N., McIntosh, L., Baccus, S., and Ganguli, S. (2019). From deep learning to mechanistic understanding in neuroscience: the structure of retinal prediction. Neural Information Processing Systems (NeurIPS), 8535–8545.

Tang, H., Schrimpf, M., Lotter, W., Moerman, C., Paredes, A., Ortega Caro, J., Hardesty, W., Cox, D., and Kreiman, G. (2018). Recurrent computations for visual pattern completion. Proc. Natl. Acad. Sci. USA *115*, 8835–8840.

Yamins, D.L., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., and DiCarlo, J.J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc. Natl. Acad. Sci. USA *111*, 8619–8624.

Zoph, B., and Le, Q.V. (2017). Neural Architecture Search with Reinforcement Learning. In International Conference on Learning Representations (ICLR). arXiv, 1611.01578.