Research Statement

Pratik Chaudhari, University of Pennsylvania
Email: pratikac@seas.upenn.edu

I work on understanding how the geometric structure in the space of learnable tasks enables artificial neural networks to learn efficiently,[1–3] with recent breakthroughs.[4–9] This is a cross-disciplinary effort and spans techniques from statistical physics, information theory and differential geometry in addition to statistical learning theory and optimization.

I believe that understanding should improve practice. I have developed state-of-the-art algorithms for training deep networks that are faster and generalize better,[1,10] few-shot learning,[11] multi-task and continual learning[12] and reinforcement learning.[13,14] Some of these algorithms have been deployed by major companies.

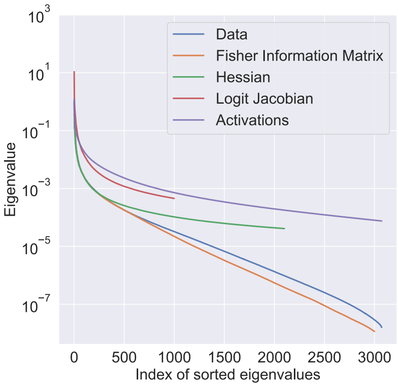
I aim for broad societal impact. My work[15] at NuTonomy[16] (now Hyundai-Aptiv Motional) helped launch the world's first autonomous taxi service in 2016. I have demonstrated techniques to obtain unbiased predictions using machine learning models for subjects in different gender, age and racial groups, clinical studies for neurological disorders.[17–19]

I will next describe my scientific accomplishments at Penn and articulate my future plans.

1 How does the structure of the task enable efficient learning?

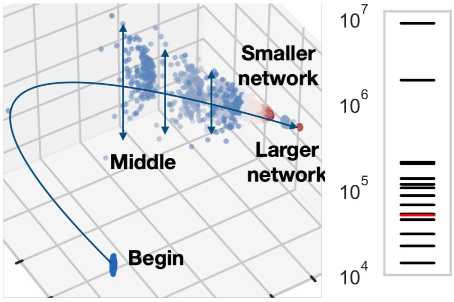
There are two stark paradoxes in deep learning today. Artificial neural networks have many more parameters than the number of training data. They can therefore overfit, i.e., make inaccurate predictions outside the training set. And yet, these networks predict remarkably accurately—defying accepted statistical wisdom. Training the network involves solving a high-dimensional, large-scale and non-convex optimization problem and should be prohibitively hard. And yet, training is tractable—even easy. I have made key advances in resolving these paradoxes.

Typical tasks are "sloppy" I showed that the ability of neural networks to avoid overfitting could be explained by a certain characteristic structure called "sloppiness".[6] The signature of sloppiness is a Fisher Information Matrix (FIM) with eigenvalues that are distributed uniformly on a logarithmic scale (see adjoining figure for a residual network on CIFAR-10). This indicates a large degree of redundancy in the learned parameters; there is one set of parameters that is tightly constrained by the data, another which can vary twice as much without affecting predictions, and so on. Sloppy models with such a FIM exhibit a range of sensitivities, they are neither insensitive to perturbations (difficult to adapt, akin to a low-rank FIM) nor exceedingly agile (sensitive to variability in the input, akin to a flat spectrum).



I have obtained a non-vacuous analytical bound on the generalization error of deep networks using this idea. This bound works for general models and does not rely on weight-compression or modeling choices such as two-layer networks or infinitely-wide networks. I have also developed numerical methods to optimize PAC-Bayes bounds using data-distribution dependent priors. I have shown how the peculiar eigenspectrum of the FIM (or the Hessian, correlations of activations, Jacobians etc.) is a consequence of a similar "sloppy" eigenspectrum for the input correlation matrix of typical tasks (images, text and tabular data). This theory therefore shows how the complexity of the functions learned by deep networks is strongly determined by the structure of the task. Even if the network is a universal approximator, the structure of the task ensures that less than 0.5% out of the millions of degrees of freedom in the learned network govern predictions.

For typical tasks, deep networks explore a low-dimensional manifold of functions during training The key hurdle in understanding deep networks comes from the complexity of the map between the weights w and the probabilistic model, i.e., $p_w(y \mid x)$ for input x and output y. I have developed new techniques to study such maps. Given N training samples $\{(x_i, y_i^*)\}_{i=1}^N$ and C classes, we can think of the network as a joint probability distribution $[0,1]^{C \times N} \ni (p_w(y = i \mid x_j)\ \forall i, j)$. This is a finite-dimensional slice of the infinite-dimensional object $p_w(y \mid x)$.



We can compare any two networks using their finite-dimensional vectors, irrespective of their architectures or training and regularization methodology. To do so, I am using techniques to embed high-dimensional probabilistic models (typically $NC \approx 10^8$) isometrically into lower-dimensional Minkowski spaces,[20] have also developed new techniques to compute geodesics and study trajectories in such spaces. Historically, it has been difficult to use ideas from information geometry to understand mainstream questions in learning because of its rather abstract ideas. My computational approach to information geometry is an attempt to change this.

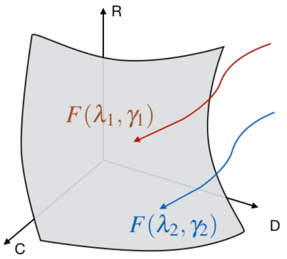
These analyses have revealed some surprises.[7] In the adjoining figure, 3,000 networks were trained on CIFAR-10 for different numbers of epochs; each point is the embedding of one network (0.5M dimensions). Dark blue indicates models at initialization, red indicates interpolating models at the end of training. There are two architectures here but they both explore very low-dimensional (eigenvalues shown on the right) and essentially indistinguishable manifolds during training (3 dimensions account for 77% variance). Larger models (small red cloud) progress further than small ones (larger red cloud) but essentially along the same trajectory. This result shows that models with very different architectures (fully-connected, convolutional, or self-attention-based networks), training regimen (SGD vs. Adam, large vs. small batch-sizes), and regularization (w and w/o batch-normalization, weight-decay, augmentation) explore a very similar manifold of functions during training. This suggests that different model classes learn the same kinds of functions for sloppy tasks.

Funding NSF CAREER award ($549K; PI), ONR ($406K; PI), Amazon ($70K; PI), Intel Rising Star faculty award ($50K; PI).

2 Characterizing the geometry in the space of learnable tasks

A deep network trained on one task can be adapted easily to new tasks. This is peculiar because training does not explicitly encourage such flexibility. Burgeoning fields such as transfer, multi-task, meta, few-shot, continual and self-supervised learning exploit this ability of deep networks. But key questions such as "when are tasks similar to each other" or "how to best transfer a model" are unanswered today. I have characterized the structure in the space of learnable tasks to understand the flexibility of neural representations. I have also developed algorithms that harness this understanding to reduce the amount of data required to learn a task, and to learn multiple tasks together.

A thermodynamics of representation learning The Information Bottleneck (IB) Principle is a generalization of rate-distortion theory: given data x and outputs y, the IB posits that good representations z are sufficient, e.g., the mutual information $I(z;y)$ is maximized, but minimal, e.g., irrelevant information in the input is discarded by say minimizing $I(x;z)$. The IB has been widely used both in neuroscience and deep learning.



I have argued how the IB cannot be a complete picture of representation learning because it does not capture the flexibility of representations that we see in practice.[4] To wit, information discarded while learning one task is precisely what will be useful for a different task. I have developed generalizations of the IB to rectify this. The key idea is to use an auxiliary task, e.g., reconstruction of the input, to force redundancy in the learned representation. What emerges is a thermodynamics where good representations minimize a certain free energy $F(\lambda,\gamma)$ that captures redundant information D, task-relevant information C and minimality R (see adjoining figure). Principles that are analogous to the first and second laws of thermodynamics, as well as statistical analogues of quantities like specific heat which capture how well an architecture is suited for a dataset, arise from this theoretical framework.

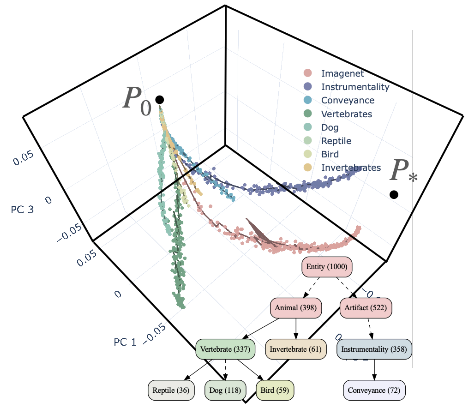
Using these ideas, I have developed an "iso-classification transfer process" that can guarantee that the accuracy of the network on the target task after transfer is as good as the accuracy on the source task before transfer. Such guarantees have never been given before—transfer learning is the most widely used technique in deep learning today and no other method can control the performance of the transferred model on the target task.

A distance on the space of tasks It is widely accepted that if source and target tasks are "close" then we can learn them together or transfer across them. But what we mean by "close" is not usually made precise, beyond methods that are specific to certain architectures or domains.

I have characterized the "optimal" distance between learnable tasks. This distance is the shortest trajectory (in terms of its Riemann length) that a model trained on the source task needs to take in order to predict well on the target task. I show how simply fitting the model on the target task (as is typically done) is sub-optimal; one must also modify the task gradually from the source to the target, e.g., using displacement interpolation in optimal transport. This leads to some remarkable properties, e.g., distance between the same pair of tasks is small if the model class is larger; tasks that were presumed to be far away are seen to be actually close. This theory also leads to an algorithm that can optimally adapt a trained model to a new task. The resultant algorithm are quite unusual because no existing transfer learning method interpolates the task.

I have also built upon these ideas to develop algorithms that can learn effectively when the training data consists of some tasks that aid each other when trained together (close-by ones) and dissimilar tasks that compete for the learning capacity (far-away ones).[12] These algorithms have significantly advanced the empirical state-of-the-art in multi-task and continual learning. As a consequence, they have received wide interest from the industry.

A picture of the space of typical learnable tasks I have characterized the geometry of the space of learnable tasks using my techniques to study high-dimensional probabilistic models (Section 1) with some surprising findings.[8] The manifold of probabilistic models trained on different tasks using different representation learning methods (e.g., supervised, transfer, multi-task, meta, or contrastive learning) is effectively low-dimensional. And this dimensionality is very small—for Imagenet, 3 dimensions of the embedding out of ~50M dimensions capture 80% of the variance. This suggests that there is a strong shared structure among different tasks.



Supervised learning on one task (e.g., all Dogs in Imagenet) results in a representation that is similar to the one obtained after ~63% of the way while training on the entire Imagenet. If the training task is diverse (e.g., 333 random classes), then this progress is much larger (~92%). In the adjoining figure, models learned while fitting tasks like Conveyance and Instrumentality are closer to each other than those of Animals. Within Animals, models learned while fitting tasks like Reptiles, Dogs and Birds are similar to each other than Invertebrates, etc. This is surprising because the phylogenetic tree of Wordnet (inset) was created using natural language-based semantics, not visual features. The tools that I have developed to study probabilistic models allow us to precisely quantify the above statements.

This work has interesting algorithmic implications. It shows that episodic meta-learning learns a similar representation eventually as that of standard supervised learning but via a longer, less efficient trajectory; the representation learned by contrastive learning is very similar to that of standard supervised learning but it does not progress towards the solution as much; techniques to resolve gradient conflicts in multi-task learning do not change the representation that is eventually learned; one should not fit the source task completely if the eventual goal is to fine-tune the representation on a target task, etc.

Funding NSF CISE Medium ($1.2M; PI), ONR ($406K; PI).

3 Exploiting the structure of the task to learn from small amounts of data

As deep learning permeates diverse spheres of our life, there are many problems where massive amounts of data required to train these models are difficult to obtain. Small data is the next frontier of machine learning. I have made theoretical and empirical advances that enable us to benefit from the superior modeling abilities of deep networks, but while working around the issues caused by their dimensionality and small sample sizes.

How to make optimal use of data? Suppose we want to estimate the bias of a coin $w \in [0, 1]$. Bayes' rule tells us to pick a prior $\pi(w)$ and calculate the posterior—we can correctly estimate the bias if the number of coin tosses N is large. If N is small then we should pick the prior differently. For example, if $N = 1$, with one bit of information, we can distinguish at most two models and should pick a prior $\pi(w) = 1/2$ for $w \in \{0, 1\}$ and zero elsewhere.

I have shown how to calculate such "reference priors" for high-dimensional models such as deep networks[9]—this is the first algorithmic instantiation after they were discovered in the late 70s. If N is small, reference priors automatically select a small hypothesis class but this can have complex models and so we do not lose any modeling power.[9] Calculating reference priors involves maximizing the mutual information between the weights and the data, and in this sense they are an optimal way to utilize data (unlabeled data or labeled data from another task). I have argued how such priors are tractable only if the task is sloppy—which typical tasks seem to be.

Semi/self-supervised learning algorithms implement a number of heuristics, e.g., enforcing similarity of features across different augmentations, minimizing the entropy of predictions, pseudo-labeling/disagreement-based losses etc. Although these techniques work well, e.g., they can work with ~1000× fewer data than supervised learning, they are quite ad hoc. I have shown how these different heuristics implement the reference prior objective in parts. Reference priors are therefore a theoretical formalization of semi/self-supervised learning. When implemented for image classification tasks, reference priors match the performance of the best semi-supervised learning methods.

Applications to problems in clinical neuroscience Clinical data is highly heterogeneous, e.g., for Alzheimer's disease, the heterogeneity stems from diverse anatomies, overlapping clinical phenotypes, and genomic traits, but due to social and operational aspects such as demographic/racial groups and data acquisition protocols of different hospitals. Heterogeneity manifests as a small sample size for different sub-groups. As a consequence, machine learning models often do not predict accurately for the entire population.

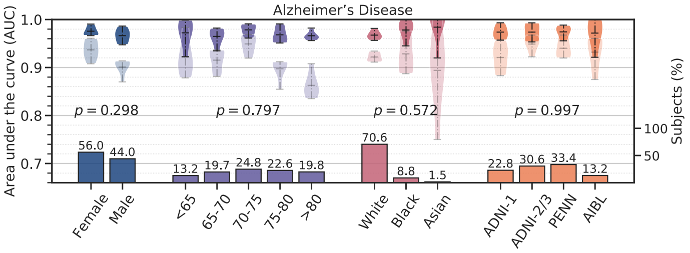


Figure 1: Solid violin plots denote a model trained on using structural, demographic, clinical, etc. features from 5 different clinical studies while translucent ones denote a baseline deep network using the same features with standard preprocessing. Bar plots denote the size of the subgroup/study (%); in many cases, there is strong data imbalance. The p-values indicate that we cannot reject the null hypothesis that the accuracy for different subgroups has the same mean (at significance level < 0.01). This is not the case for the baseline deep network. Results for schizophrenia and autism spectrum disorder classification are similar.

Specifically, the bias of machine learning models, for subjects from different demographic, racial and clinical study cohorts has received wide attention.[21] My work provides a positive datapoint to this debate. For Alzheimer's disease, schizophrenia and autism spectrum disorder, I have shown evidence that predictions need not be biased if some simple safeguards such as adequate data pre-processing techniques, hyper-parameter tuning and rigorous model selection are adopted.[17] This work also demonstrates that somewhat old-fashioned features (e.g., brain volumes in anatomical regions of interest) are more effective for these challenging heterogeneous problems than directly using high-dimensional MRI data with deep networks.

When the degree of heterogeneity is large, I have argued how a single model that predicts well on average does not predict accurately for any sub-group.[12,18] In such cases, one must embrace the heterogeneity and adapt the model. I have built upon my work on few-shot learning[11] to adapt a trained model to different sub-groups using small amounts of data. This gives the best of both worlds: the pretrained model discovers salient features across the entire population which are then refined for the specific sub-group. I have shown significantly better performance compared to existing methods using such adaptation. Typical domain adaptation techniques need labeled data from the target sub-group, but I have shown how auxiliary tasks (e.g., labels for age/gender are easily available) can also be used to adapt effectively.

Funding NSF CAREER award ($549K; PI), NIH R01 (~$145K; Senior Investigator; PI Rahim Rizi Radiology), Amazon ($45K; PI).

4 Future research directions

I will next discuss some new research directions in my group and their relationship to my existing work.

4.1 Principles of learning that cut across artificial and biological systems

An astonishing variety of systems in physics and biology are sloppy.[22] The fact that deep networks are too may therefore hint towards an underlying structure in the learnable task. Some evidence of this structure is seen in classic work in neuroscience[23,24] which suggests that typical inputs (images, language etc.) lie on low-dimensional manifolds. But the nature of semantics upon these inputs (object categories, phonemes, etc.) is less studied. In ongoing work, I am investigating the hypothesis that such semantics are highly redundant functions of inputs and therefore a network can predict accurately using essentially any low-dimensional statistic of the input. I hope to make advances on long-standing questions of representation learning where the prevalent conundrum has been that such sufficient statistics must be difficult to find in high-dimensional data with lots of irrelevant, nuisance variables.[25–27]

Biological systems also exhibit a huge diversity of functional blocks that work in highly redundant ways,[28] e.g., there are ~60 types of neurons in the mammalian retina. Why would biology create such redundancy? I am investigating the hypothesis that it is a manifestation of adaptation. If a system, e.g., the retina, contains many unconstrained degrees of freedom, e.g., over the course of evolution, then it can preserve its essential function learned using few constrained degrees of freedom (stiff subspace of the FIM) and use the sloppy subspace to adapt to the changes in the environment. As a first step towards understanding such commonalities between artificial and biological learning systems, I hope to obtain normative principles that govern the design of neural circuits. My larger goal is to emulate these principles into artificial systems that match the efficiency of biological learning.

Funding NSF CISE Medium ($1.2M; PI), Intel Rising Star faculty award ($50K; PI).

4.2 Neural networks on optical computing devices

During my doctoral studies, I showed that stochastic gradient descent, when it generalizes well for deep networks, converges to “wide regions” in the energy landscape.[1] I devised a modified objective called Local Entropy motivated from statistical physics that biases optimization towards such regions that enables 2-4× faster training while improving generalization.[10] Local Entropy-like objectives are closely related to survey propagation-based techniques for solving combinatorial optimization problems (SAT, LDPC decoding, or a neural network with binary weights) and have powerful connections to quantum annealing. I am working on using these methods to decompose large combinatorial optimization problems into smaller ones which will be solved on an Ising machine built using optical solitons.

Funding DARPA Quantum Inspired Classical Computing (~$735K; co-PI; PI Firooz Aflatooni ESE).

4.3 Active semantic scene understanding

In robotics, the notion of a task is less clear and may change depending upon the context, e.g., object localization, scene understanding, motion planning, etc. Consider a problem where multiple unmanned aerial vehicles (UAVs) explore an unknown and unstructured region such as a forest. The task here can range from estimating the volume of timber in a given region, estimating the amount of biomass/carbon captured, building a catalog of biodiversity, to estimating models of tree growth across seasons. These tasks require complex decision making (e.g., should the UAV estimate the diameter of this tree trunk accurately, or move on to the next tree). There is also a large variability in the data (the same tree looks very different across seasons). The key feature of these problems that is absent in machine learning is that the UAVs can move to investigate their surroundings and gather new kinds of data to improve the representation. I am adapting ideas from the Information Bottleneck principle and active perception to formalize representation learning for such problems. My goal is to build semantic representations of the scene to answer questions such as “how should the multi-UAV team move to make progress on a task?” and instantiate these principles on actual hardware.

Funding NSF National Robotics Initiative ($1M; PI).

References

1. Chaudhari, P., Choromanska, A., Soatto, S., LeCun, Y., Baldassi, C., Borgs, C., Chayes, J., Sagun, L. & Zecchina, R. Entropy-SGD: Biasing Gradient Descent into Wide Valleys. in *Proc. of International Conference of Learning and Representations (ICLR)* (2017).
2. Chaudhari, P. & Soatto, S. Stochastic Gradient Descent Performs Variational Inference, Converges to Limit Cycles for Deep Networks. in *Proc. of International Conference of Learning and Representations (ICLR)* (2018).
3. Chaudhari, P., Oberman, A., Osher, S., Soatto, S. & Carlier, G. Deep Relaxation: Partial Differential Equations for Optimizing Deep Neural Networks. *Journal of Research in the Mathematical Sciences (RMS)* **5,** 1–30 (2018).
4. Gao, Y. & Chaudhari, P. A Free-Energy Principle for Representation Learning. in *Proc. of International Conference of Machine Learning (ICML)* (2020).
5. Gao, Y. & Chaudhari, P. An Information-Geometric Distance on the Space of Tasks. in *Proc. of International Conference of Machine Learning (ICML)* (2021).
6. Yang, R., Mao, J. & Chaudhari, P. Does the Data Induce Capacity Control in Deep Learning? in *Proc. of International Conference of Machine Learning (ICML)* (2022).
7. Mao, J., Griniasty, I., Yang, R., Teoh, H. K., Ramesh, R., Transtrum, M., Sethna, J. & Chaudhari, P. A Picture of the Prediction Space of Deep Neural Networks. *(In preparation)* (2022).
8. Ramesh, R., Mao, J., Griniasty, I., Yang, R., Teoh, H. K., Transtrum, M., Sethna, J. & Chaudhari, P. A Picture of the Space of Typical Learnable Tasks. *arXiv preprint arXiv:2210.17011* (2022).
9. *Gao, Y., *Ramesh, R. & Chaudhari, P. Deep Reference Priors: What Is the Best Way to Pretrain a Model? in *Proc. of International Conference of Machine Learning (ICML)* (2022).
10. Chaudhari, P., Baldassi, C., Zecchina, R., Soatto, S., Talwalkar, A. & Oberman, A. Parle: Parallelizing Stochastic Gradient Descent. in *Conference on Machine Learning and Systems (MLSys)* (2018).
11. Dhillon, G. S., Chaudhari, P., Ravichandran, A. & Soatto, S. A Baseline for Few-Shot Image Classification. in *Proc. of International Conference of Learning and Representations (ICLR)* (2020).
12. Ramesh, R. & Chaudhari, P. Model Zoo: A Growing "Brain" That Learns Continually. in *Proc. of International Conference of Learning and Representations (ICLR)* (2022).
13. Fakoor, R., Chaudhari, P., Soatto, S. & Smola, A. J. Meta-Q-Learning. in *Proc. of International Conference of Learning and Representations (ICLR)* (2020).
14. Fakoor, R., Mueller, J., Chaudhari, P. & Smola, A. J. Continuous Doubly Constrained Batch Reinforcement Learning. in *Proc. of Conference on Neural Information Processing Systems (NeurIPS)* (2021).
15. Castro, L. I. R., Chaudhari, P., Tumova, J., Karaman, S., Frazzoli, E. & Rus, D. Incremental Sampling-Based Algorithms for Minimum-Violation Motion Planning. in *Proc. of Conference on Decision and Control (CDC) Dec 10-13, 2012* (2013).
16. NuTonomy. https://en.wikipedia.org/wiki/NuTonomy#Technology. Oct. 2022.
17. Wang, R., *Chaudhari, P. & *Davatzikos, C. Machine Learning Models Are Not Necessarily Biased When Constructed Properly: Evidence from Neuroimaging Studies. *arXiv preprint arXiv:2205.13421 (under review)* (2022).
18. Wang, R., Chaudhari, P. & Davatzikos, C. Embracing the Disharmony in Medical Imaging: A Simple and Effective Framework for Domain Adaptation. *Medical Image Analysis* (2021).
19. Wang, R., Chaudhari, P. & Davatzikos, C. Harmonization with Flow-Based Causal Inference. in *Proc. of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)* (2021).

20. Quinn, K. N., Clement, C. B., De Bernardis, F., Niemack, M. D. & Sethna, J. P. Visualizing Probabilistic Models and Data with Intensive Principal Component Analysis. *Proceedings of the National Academy of Sciences* **116,** 13762–13767 (2019).

21. Davatzikos, C. Machine Learning in Neuroimaging: Progress and Challenges. *NeuroImage* **197,** 652 (2019).

22. Quinn, K. N., Abbott, M. C., Transtrum, M. K., Machta, B. B. & Sethna, J. P. Information Geometry for Multiparameter Models: New Perspectives on the Origin of Simplicity, arXiv:2111.07176 (2021).

23. Olshausen, B. A. & Field, D. J. Emergence of Simple-Cell Receptive Field Properties by Learning a Sparse Code for Natural Images. *Nature* **381,** 607–609 (1996).

24. Field, D. J. What Is the Goal of Sensory Coding? *Neural computation* **6,** 559–601 (1994).

25. Mitter, S. K. Towards a Unified View of Inference, Communication and Control. in *2010 International Conference on Signal Processing and Communications (SPCOM)* (July 2010), 1–2.

26. Burns, J. B., Weiss, R. S. & Riseman, E. M. in *Geometric Invariance in Computer Vision* 120–131 (1992).

27. Soatto, S. Steps towards a Theory of Visual Information: Active Perception, Signal-to-Symbol Conversion and the Interplay between Sensing and Control. *arXiv preprint arXiv:1110.2053* (2011).

28. Balasubramanian, V. Heterogeneity and Efficiency in the Brain. *Proceedings of the IEEE* **103,** 1346–1358 (2015).