High Aspiration Meets Technical Opportunity

In this year's class, we are going to take on an ambitious challenge as a class project. This is not the first time we've done this and later I'll provide examples of class projects that resulted in papers co-authored by the students taking the class.

Google's company *mission* is to organize the world's information and make it universally accessible and useful. Larry Page encouraged us to create products that were useful to lots of people - a million dollars or a million new users, both were considered a measure of success.

Demis Hassabis and the research scientists at DeepMind pursue a similar mission by sharing technology, tools, and datasets, and by tackling – and solving – problems that were once considered out of reach for machines, e.g., grandmaster level Chess, Go and Starcraft, and, more recently, protein folding.

Sometimes it is enough to take on a seemingly intractable problem, pursue a novel approach, and demonstrate progress, in order to inspire optimism and provide direction for others to follow. DeepMind has done just that by demonstrating an approach for developing AI agents that learn to imitate, acquire language from, and interact with human agents.

This year we will conduct a thought experiment to explore how one might apply the strategies developed by DeepMind to build AI agents capable of interacting with software engineers to assist in software development, by providing the skills of a *programmer's apprentice*.

Some might think that tackling two AI-complete problems in a single challenge is ridiculously optimistic, but I believe that pursuing such an aspirational goal will provide insight into both natural language acquisition and automated programming.

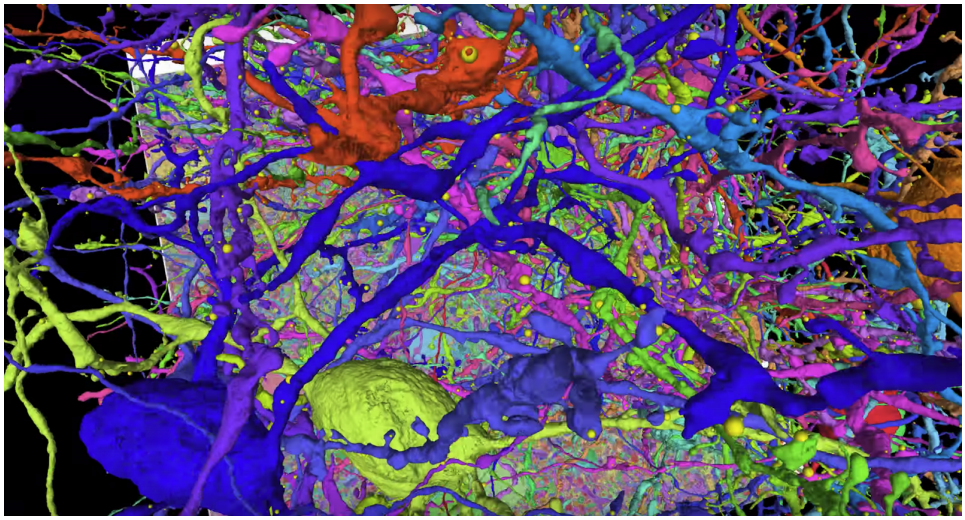
In the process, we will explore AI agents whose embodiment is in the form of a set of common tools that engineers employ in developing software, and whose use of language is grounded in both its interaction with computers and the common ground it shares with the programmer. We also take seriously the idea that analogical reasoning is an essential skill in both understanding and communicating algorithmic thinking and programming knowledge.

We will be joined in carrying out this experiment by a collection of experts including linguists, developmental psychologists, researchers who specialize in automated code synthesis, and members of the Interactive Agents Group at DeepMind who contributed to the research mentioned above and were co-authors on the recent paper describing their progress.

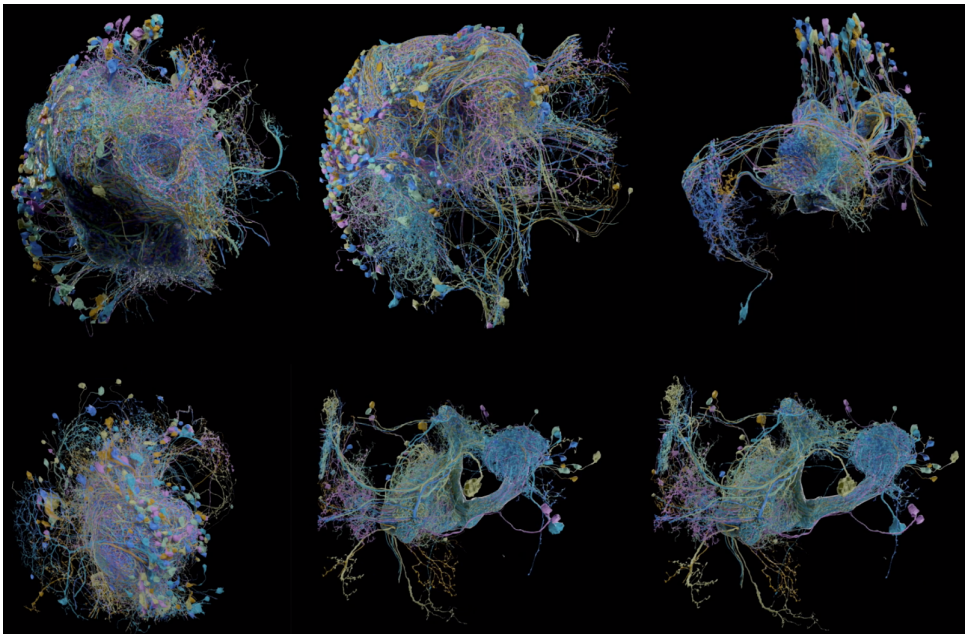
In the summer of 2013 following the end of classes, eight students and I wrote a white paper on the technology prospects and investment opportunities for scalable neuroscience. We studied a wide range of technologies including scanning electron microscopy (SEM), focused ion beam scanning electron microscopy (FIB-SEM), two-photon excitation imaging (2PE), optogenetics, functional magnetic-resonance imaging (fMRI), diffusion tensor imaging (DTI), and magneto electroencephalography (MEG) to name a few of the most promising technologies studied.

We predicted that the new multi-beam electron microscopes being developed by Zeiss would be powerful enough to make it possible to reconstruct the entire connectome of the common fruit fly, Drosophila melanogaster. Flies have on the order of 100,000 neurons and, at the time, the largest connectone of an entire organism was that of a nematode that goes by the taxonomic name of Caenorhabditis elegans, has exactly 302 neurons and took over 20 years and scores of graduate students to reconstruct.

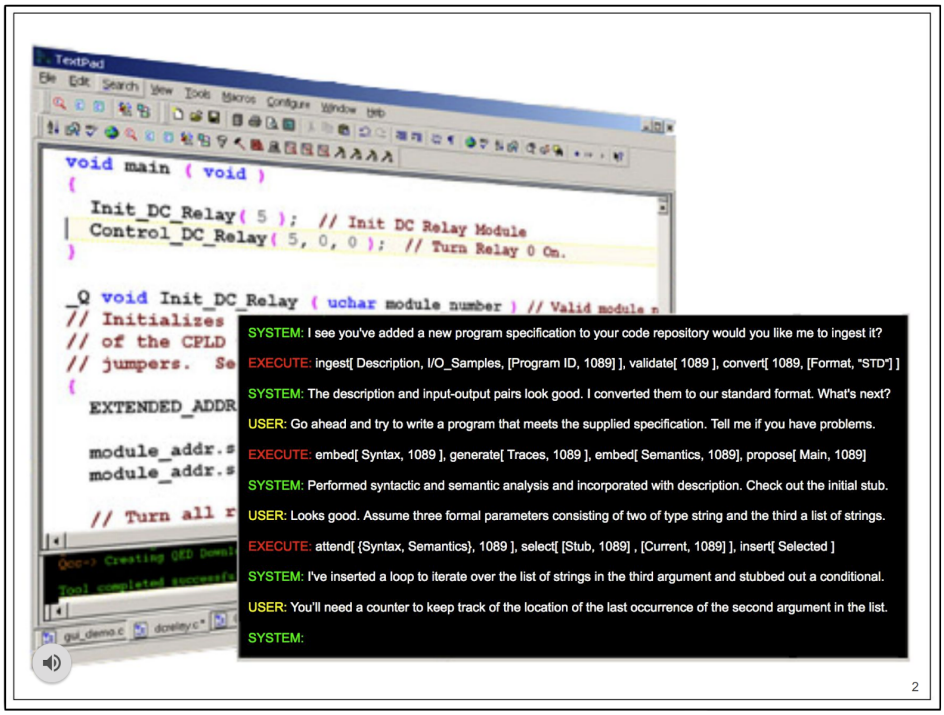
Three years later I had put together a team of Google engineers and hired Viren Jain from HHMI Janelia Campus. Viren and I started partnerships with Christof Koch at the Allen Institute for Brain Science, Sebastian Seung at Princeton, Jeff Lichtman at Harvard, Ed Boyden at MIT, and Winfried Denk at the Max Planck Institute of Neurobiology to supply us with tissue samples from a variety of organisms. In 2020 in collaboration with HHMI we announced the first fully automated reconstruction of the Drosophila hemibrain.



The technology involved advancements in scalable infrastructure and machine learning technology employing deep neural networks for tracing neural circuits, identifying neuron cell types, and characterizing properties of synapses relevant to predicting the likelihood of their transmitting neural signals. The team is now working on the mouse connectome — an undertaking that involves on the order of 100 million neurons, billions of synaptic connections, and exabytes of data.



In the summer of 2018, eleven students and I wrote another prospectus entitled, Amanuensis: The Programmer's Apprentice, spurred in part by a presentation I gave at the Kavli Futures Symposium on next-generation, open-source neurotechnology in October the previous year. The gist of that talk was the prediction that in the next decade AI technology will enable human-machine symbiotic relationships allowing neuroscientists working with large datasets to interact easily in natural language with AI systems capable of writing code to perform sophisticated data analyses as well as accelerate the development and evaluation of powerful explanatory models. The programmer's apprentice project was resurrected!



The Programmer's Apprentice was the name of a project started by Charles Rich and Richard Waters at the MIT AI lab in 1987. The goal of the project was to develop a theory of how expert programmers analyze, synthesize, modify, explain, specify, verify, and document programs, and, if possible, implement them. Their research plan was to build prototypes of the apprentice incrementally. Our research plan also involves making incremental steps. However, we will be able to make substantially larger steps by exploiting and contributing to the powerful AI technologies developed during the intervening 30 years with our primary focus on recent advances in applied machine learning and artificial neural networks.

Specifically, we will build on the work of the Interactive Agents Group at DeepMind and borrow from a select set of papers that provide specific technologies relevant to building interactive agents that work closely with humans using natural language to communicate and tackle hard problems that can benefit from the strengths of both human experts and machines with more narrow expertise that is nonetheless useful in rendering humans more capable.

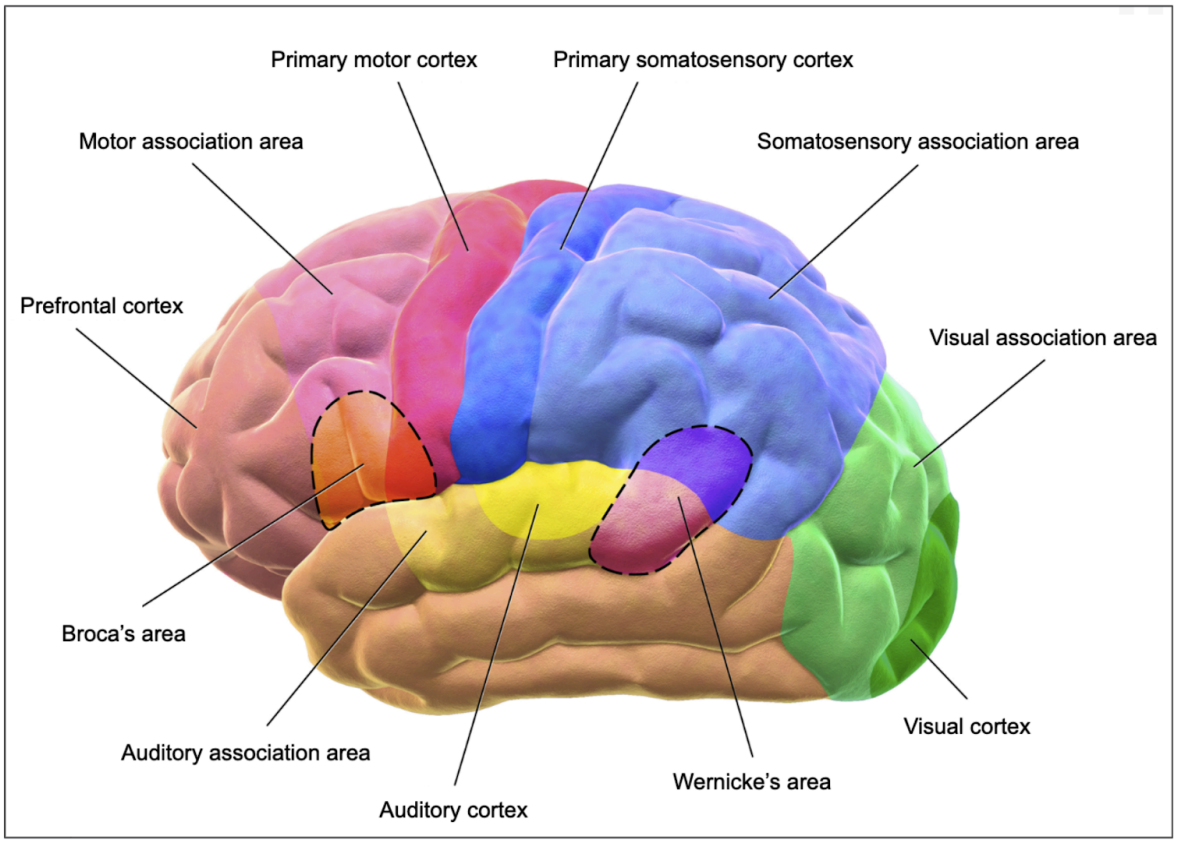
This is an aspirational endeavor. As in the case of our research investigating the prospects for developing advanced technology to spur innovation in neuroscience, the goal here is to examine the prospects for scaling and extending the technology developed by the Interactive Agents Group in order to build collaborative systems to assist humans in tackling challenging technical problems.

Architectural Inspiration from Neuroscience

This course is nominally about computational models of the neocortex. That said, in past years, we have focused on other parts of the primate brain as well as the brains of nematodes, sea slugs, fruit flies, mice, rats, and Etruscan shrews.

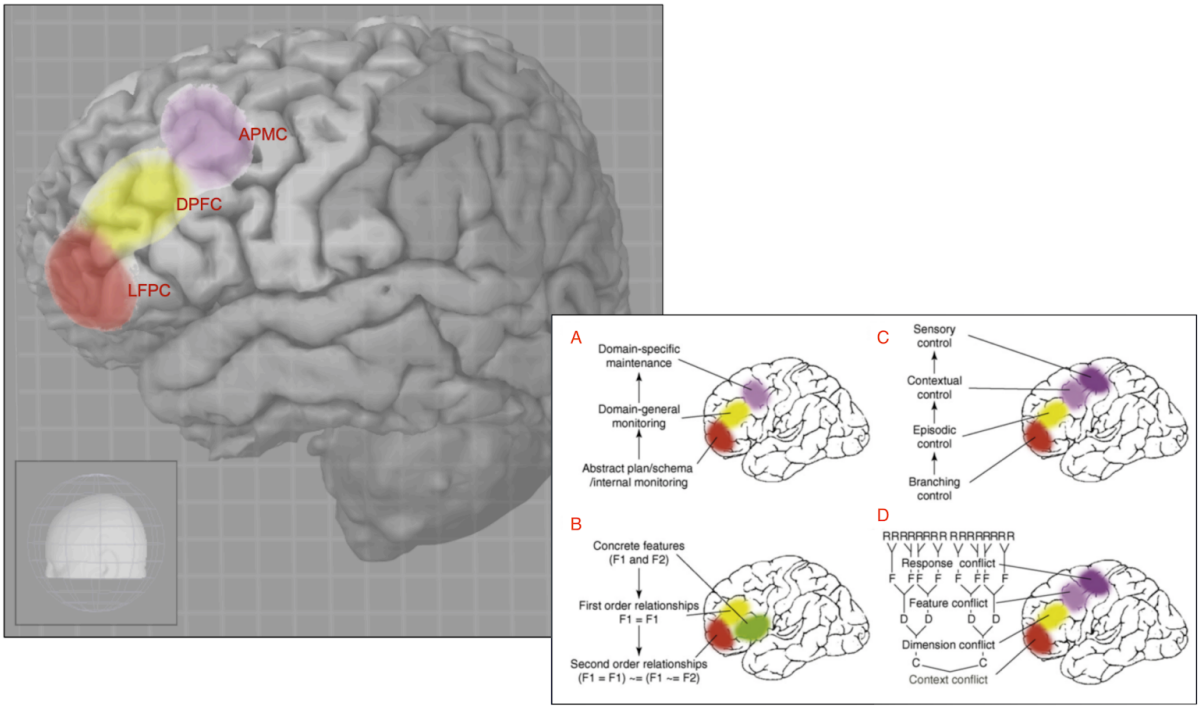
This year the neuroscience emphasis is more cognitive than cellular, and we draw more heavily on developmental psychology and linguistics than we have in the past. However, much of the neural network architecture that we will discuss in this year's class draws its inspiration from our admittedly still-evolving understanding of the primate brain.

To strike an appropriate balance in this year's class, this lecture explores some of the research in neuroscience that has inspired the technologies we will employ in this year's class project. No attempt will be made to slavishly follow the biological precedents, but the basic principles will substantially influence our design.



This is not an introductory course in neuroscience and you won't be expected to demonstrate your prowess in rattling off the names of cortical regions. That said, many of the concepts and technologies discussed in class were inspired by our current understanding of the human brain and so in this lecture, I'll explain the basic concepts relevant to the primary focus of the class and in case you are interested in learning more I've provided some basic resources that I will include in the class discussion list. The above graphic labels the primary landmarks on the human brain; most of the names are self-explanatory, Broca's and Wernicke's areas correspond to regions associated with language production and language understanding respectively.

Rostro-Caudal Gradient: Hierarchical & Relational Complexity

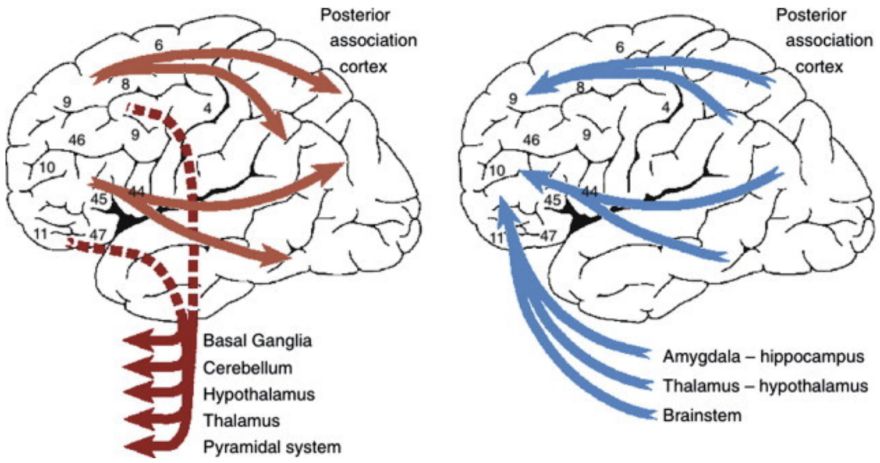


Koechlin, Ody, and Frederique Kouneiher. The architecture of cognitive control in the human prefrontal cortex. *Science*, 302:1181-1185, 2003.

David Badre. Cognitive control, hierarchy, and rostro-caudal organization of the frontal lobes. *Trends in Cognitive Sciences*, 12(5):193-200, 2008.

The mammalian brain is organized hierarchically. This hierarchical structure is constructed in stages over our long development beginning prenatally and extending well into early adulthood. The constructive process depends on both our genetic heritage and the physical and social environment in which we are raised. Our goal today is to understand some of the information processing implications of our experience of the environment on the structure and function of the primate brain and the neocortex in particular. The graphic on the left depicts three regions in the prefrontal cortex implicated in higher-order processing and the inset on the right shows four different interpretations of their function.

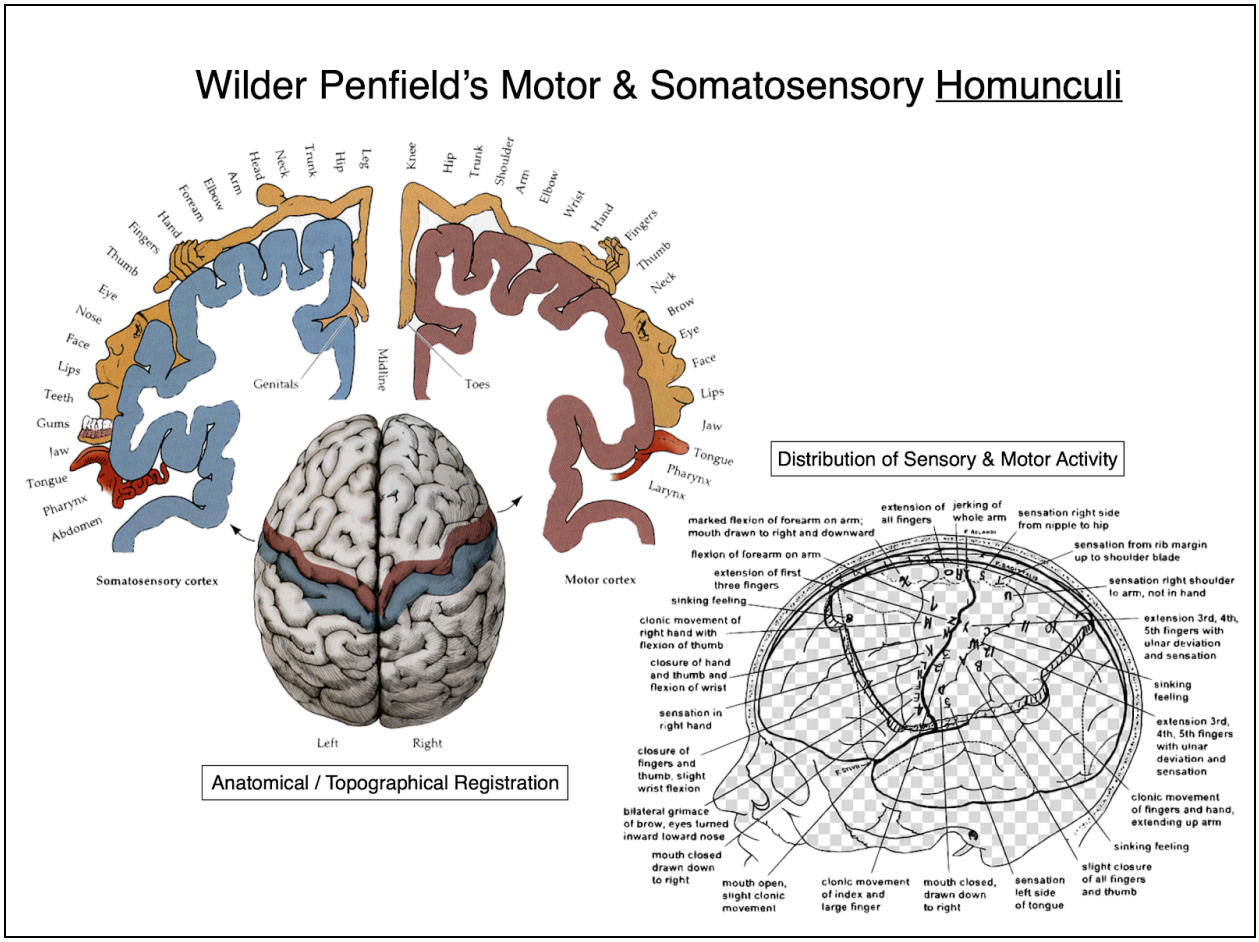
Vladimir Betz's Anterior / Posterior Functional Dichotomy



Betz, W. (1874) Anatomischer Nachweis Zweier Gehirn Centra. *Central-blatt für die medizinische Wissenschaften*. 12, 578–580.

Sergiy V. Kushchayev, Vitaliy F. Moskalenko, Philip C. Wiener, Vitaliy I. Tsymbaliuk, Viktor G. Cherkasov, Irina V. Dzyavulska, Oleksander I. Kovalchuk, Volker K. H. Sonntag, Robert F. Spetzler, and Mark C. Preul. The discovery of the pyramidal neurons: Vladimir Betz and a new era of neuroscience. *Brain*, 135(1):285-300, 2011.

Our brains are also organized along the anterior-posterior axis with the back of the brain largely responsible for sensory processing and the front responsible for action selection broadly speaking. Sensory areas range from single modality primary sensing areas responsible for relatively primitive features such as adjacent areas of high contrast in the visual field to secondary areas featuring more abstract features to multimodal association areas that combine sensory modalities to construct highly abstract features that allow us to represent the interdependence between sensory stimuli, for example, enabling us to combine an auditory signal such as a bird call with a visual signal such as a flash of bright color in a tree.



The frontal cortex and in particular the motor cortex is similarly organized, ranging from input and output from the spinal cord and muscles in the periphery to sensorimotor areas that map the origin of sensations in the periphery, proprioceptive areas that enable awareness of body position, and, closer to the front of the brain, circuits responsible for generating action plans and higher-level executive planning, regulating emotions and controlling impulses in the prefrontal cortex. Individual maps retain the topographical structure of their respective stimuli, e.g., the primary visual cortex maintains the spatial layout of the retina. Multiple maps such as the visual and auditory association areas are spatially aligned to support multisensory integration in order to facilitate, for example, quickly figuring out where a sound is coming from.

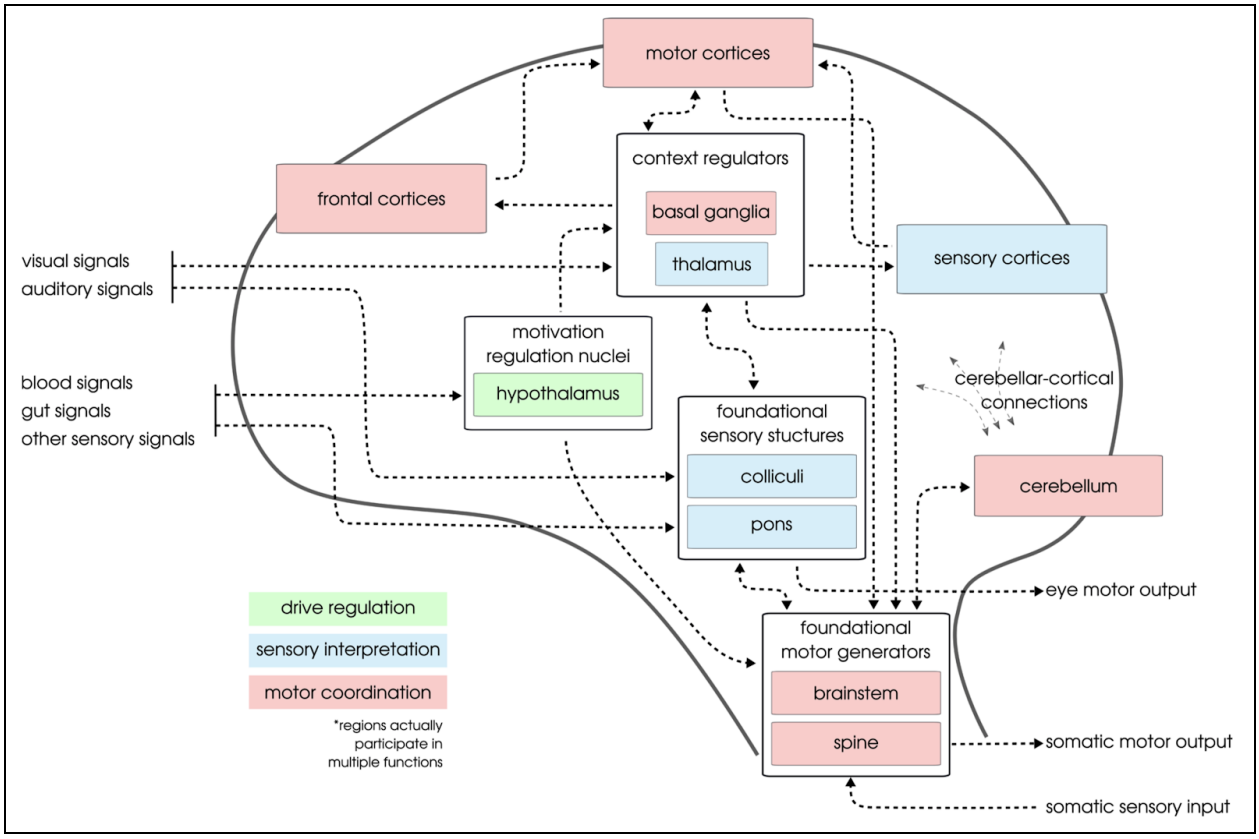


FIGURE: Diagram from Merel et al 2019 *Hierarchical Motor Control in Mammals and Machines*

I don't want you to infer from my high-level description of information processing in the brain, that the circuits governing our behavior are simple by any stretch of the imagination. Quite the contrary, we are a consequence of natural selection and our often baroque and puzzling brain circuitry reflects the exigencies of our species surviving for hundreds of thousands of years in a constantly changing and unforgiving environment.

The above diagram from a paper by Josh Merel, Matt Botvinick, and Greg Wayne provides a high-level architectural view of the brain regions involved in motor control that still doesn't do justice to the complexity of the primate brain, but does illustrate some of the neural circuits that engineers have drawn inspiration from in designing control systems. We'll revisit this paper later when we consider how their model can be adapted to cognitive reasoning of the sort required for understanding computer programs. If you are interested in the details of their paper, see Josh's presentation in last year's class here.

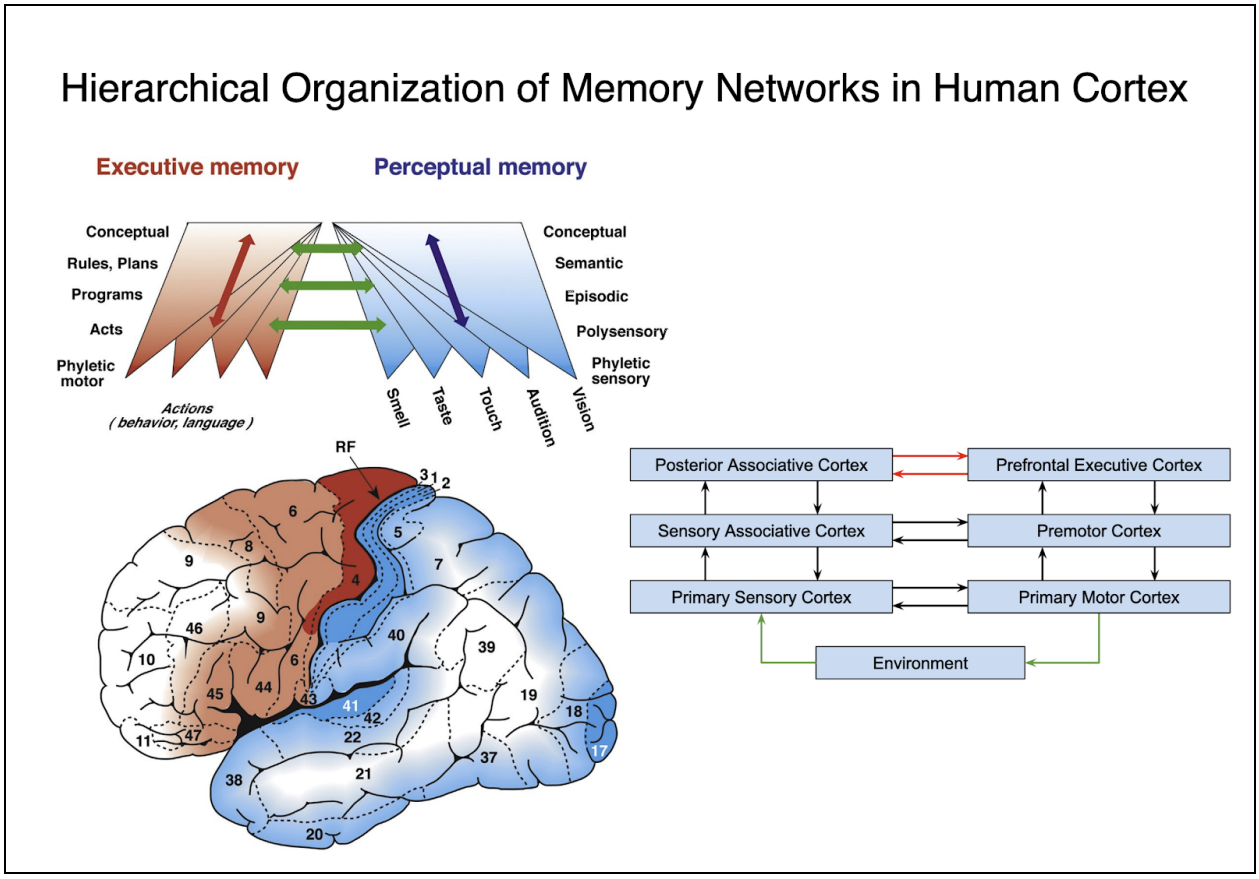


FIGURE: Diagram from The *Prefrontal Cortex*, Chapter 8 by Joaquin Fuster

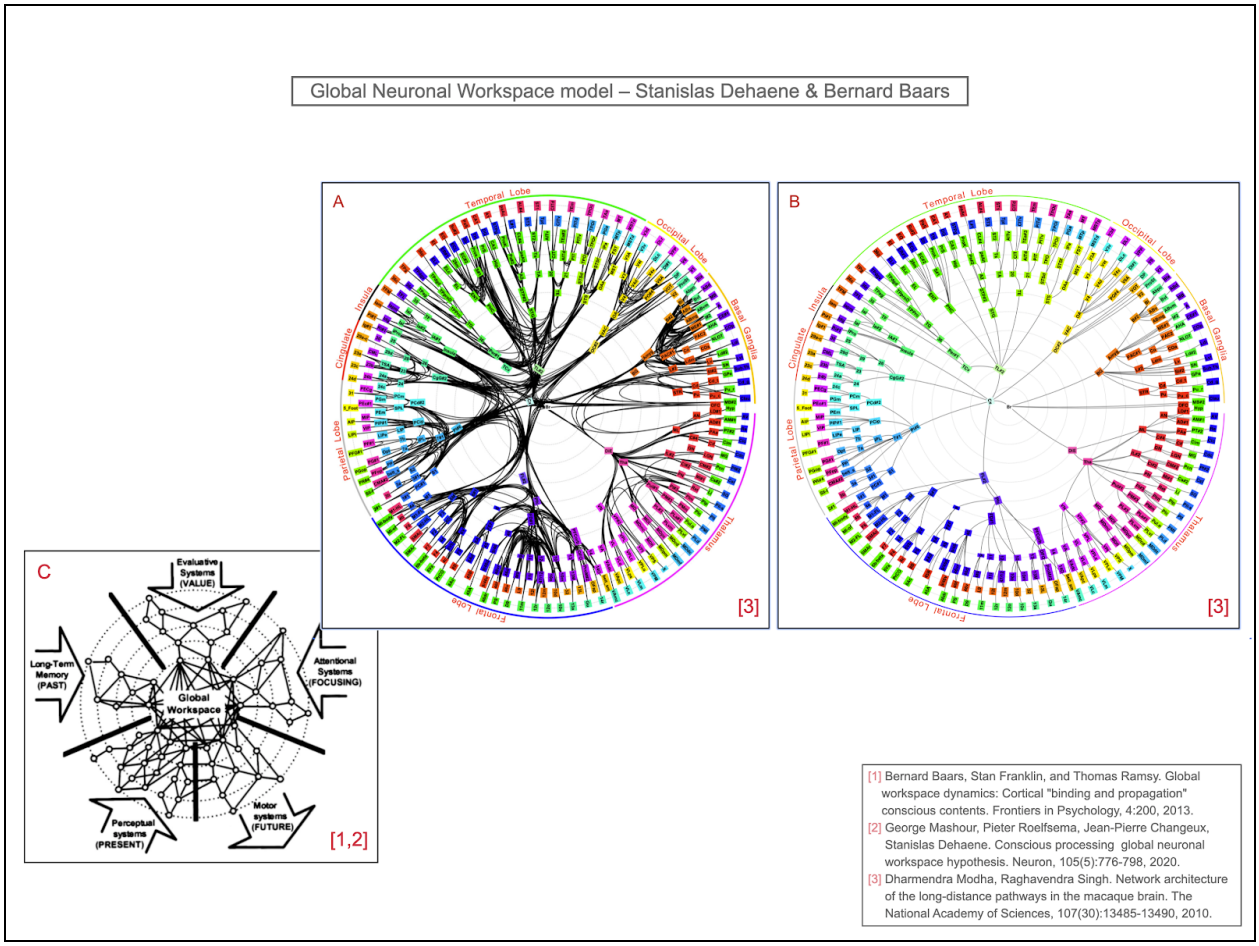
This organizational structure is captured in a model that is often referred to as Fuster's hierarchy – a phrase introduced by Matt Botvinick in his 2007 paper in the Transactions of the Royal Society and featured in Joaquin Fuster's highly influential book entitled "The Prefrontal Cortex" now in its fifth edition. In addition to its hierarchical structure, Fuster's model is also notable for its explicit recognition of the reciprocal connections between the sensory and motor areas at multiple levels of abstraction, and its explicit depiction of how we influence the environment by our actions and how we observe the consequences of our actions on the environment through our sensations.

Fuster refers to this cycle of our initiating action, observing its impact on the environment, and interpreting our perception of the consequences of our actions as the *action-perception cycle*. This cycle enables us to ground our understanding of the cause-and-effect relationships between our actions and their consequences in the manifestation of those consequences in the physical world. It serves the role of a supervisor in facilitating our supervised learning how to act in order to further our goals by training the features we require to verify the consequences of our actions and provide the context for our selecting appropriate actions.

Importantly, the reciprocal horizontal connections between the levels in the sensory stack and the corresponding levels in the motor stack ensure that all of the features that we learn are

selected on the basis of their utility in pursuing our goals. In particular, this implies that we don't learn features in order to reconstruct, say, a detailed image of a cat unless those features are relevant to interactions with cats that serve our needs. A corollary of this constraint is that it is generally a poor idea for an engineer designing a computer vision system for a robot to select features based upon his or her understanding of what the robot will need to carry out its intended purpose.

The vertical connections between adjacent blocks in each of the sensory and motor stacks are intended to illustrate that features at one level are used to construct features at higher levels. These connections are also reciprocal and not shown are connections that enable dependencies between features that span multiple levels so higher levels that combine information from multiple modalities to adjust their posterior distribution can provide feedback to the lower levels to refine their features taking into account the higher-level processing. In general, these feature hierarchies are said to be compositional in that the features at one level are composed of those at lower levels.



There are other architectural features of primate brains that relate to the flexibility and plasticity of human thought. Consider how faced with a novel situation and a problem that needs solving we are able to draw on a large store of episodic and declarative knowledge and apply a wide

variety of cognitive strategies to analyze the situation and come up with multiple strategies for exploring our options and coming up with a solution to the problem posed. Bernard Baars global workspace theory of working memory and Donald Hebb's model of dynamic cell assemblies are two examples of tantalizingly attractive theories that are rife with speculation and short on models that have any convincing experimental evidence. Ideas from AI and machine learning offer possible interpretations of the relevant phenomena borne out of the technical requirements of engineers in solving specific problems.

For example, how do we learn new skills without inadvertently and negatively impacting existing skills? This is generally referred to as *catastrophic interference* and is often dealt with by using gating in LSTM (Long-short Term Memory) models or external memory in the case of NTM (Neural Turing Machine) and DNC (Differentiable Neural Computers). As another example, how do we quickly accommodate new information as in the case of learning a new word, integrating it with the rest of our vocabulary and regularly adjusting its precise meaning to account for changes in usage, or consolidating new episodic memories and reconsolidating old memories?

Rapid replay during slow-wave or non-REM (NREM) sleep is thought to be responsible for learning declarative memories, but the exact method for doing so is still largely a mystery. Techniques like *fast weights* allow you to store short term memory with a weight matrix, [Hinton and Plaut [1987], but there is still the issue of when and how you might subsequently store a carefully selected subset of those memories more long term and doing so with corrupting other memories. In working on the programmer's apprentice project during this quarter we will have the need for technology that supports novel learning strategies, and so return to these issues at that time.

The take-home message is as follows: the same reciprocal dependencies and reliance on feedback from the environment to anticipate the consequences of acting apply at all levels within the hierarchy from the lowest to highest, and in the case of the highest levels involving our overt behavior toward one another are clearly manifest in our social interactions and in particular in how we learn language and make use of it in everyday life.