

CALIBRE@ICTS Interview Series

In Conversation with Prof. Vijay Balasubramanian

University of Pennsylvania (Visiting as Asha Gupta Chair Professor, CALIBRE, ICTS)

Host Dr. Sarthak Chandra, CALIBRE Faculty, ICTS

Dr. Chandra

At ICTS, you've had the opportunity to engage with researchers across several domains of physics, biology, AI, and complex systems. What motivated your engagement with CALIBRE at ICTS? What motivated you to come here and interact with people across all of these different domains?

Prof. Balasubramanian

Both ICTS — and I should mention NCBS too — have a substantial number of people who are excellent physicists in everything ranging from high energy physics to statistical physics to biophysics, and who have interests at the interface of machine learning, AI, and their fields. Sometimes because they would like to use AI to help their fields, and sometimes because they want to use the methods of physics to understand how these models manage to achieve the kind of wonders that they're doing.

At the same time, there's a very substantial landscape in biology. ICTS itself has several faculty working in biological physics who are deeply engaged with the actual phenomena of the living world. You yourself are an example in neuroscience, and there are people who work in cell biology, evolution, ecology, and things like this.

If you think about living systems, a characteristic feature of all of them is that they learn. They adapt and they learn. It's almost the difference between the kind of matter that makes chairs and the kind of matter that makes things alive — living things learn and do new things. So there's an acute awareness of that in the faculty here. It's a very good environment.

Also, the fact that NCBS is not so far away is really excellent. There's a great theory group in biology there, and they've been working for a long time on the Center for Study of Living Machines. So there's already a focus on the idea that machines and life have something shared. And there's a suite of fantastic experimentalists. In my own case, it's the experimental neuroscientists that I have most engagement with.

So it's a very fertile environment with very good people concentrated in a small space and a lovely environment to work in. What more can you ask?

Dr. Chandra

I'm glad you're having a good time — and a productive time, with several collaborations starting?

Prof. Balasubramanian

Yes, very productive. Several collaborations are starting, including with experimentalists over at NCBS.

Dr. Chandra

More broadly, what do you think is the importance of centers like this, particularly today, in trying to create this melting pot of physics, AI, and biology?

The academic world is fundamentally siloed. There are physics departments, biology departments, computer science departments, and nowadays machine learning and data science centers. The way these institutions have been conceived for more than 100 years is in a model where these are discrete subjects with discrete canons of learning and discrete subjects of inquiry. As a result, everything from the curricula that are taught to the hiring practices are channeled in that way.

Many institutions for a few decades now have mouthed the words that interdisciplinarity is important, because we understand that in our modern world, the division of knowledge conceived of in the late 19th century is not necessarily applicable today. Chemists do biology and physics; physicists do biology and chemistry; computer scientists do everything. There's lots of interdisciplinary work, but it's actually not that well supported institutionally.

Consider a physics department with a new hire slot. You need to cover astrophysics — somebody's got to think about stars and galaxies. You need particle physics. You need quantum materials. And then there's the interdisciplinary area at the interface of AI and neuroscience — well, why don't we ask the neuroscientists to do that hire so we don't lose our slot? The neuroscientists say, why don't we get the physicists to do that? The computer scientists say the same. It's a resource allocation problem.

So explicitly interdisciplinary ventures like CALIBRE help to craft the new landscape of knowledge of the 21st century, where things aren't divided in the same old way. By design, the postdoctoral program is intended to hire people at that interface, visitors are invited at that interface, and this helps create both the resources and the intellectual mood that facilitates new development. I think it's a really great thing, and I wish there were more of it at more institutions.

Dr. Chandra

That's wonderful. Now, people tend to think of things as being siloed, but when we say we're interested in interdisciplinarity, we also need to be clear about what that means — what the value proposition is on either end. And sometimes things can help in both directions and then produce something entirely new. Could you tell us a little about your own journey into this space? To most physicists, you're probably best known for work on black holes and quantum information. So what is someone like you doing at this intersection?

Prof. Balasubramanian

Let me clarify that. My undergraduate degrees were in computer science and in physics. The computer science degree was because, on one hand, I was interested in the foundations of the universe — that was the physics — but I was also convinced that the way we write down the laws of the universe are constrained by the computer inside our head. A cat can't do calculus because of the computer inside its head. So we might have constraints too. But I knew zero biology at the time.

My approach to trying to understand those constraints was to try to build artificially intelligent machines. That's what I did as an undergraduate. My first work was actually in machine learning — specifically, speech recognition. I have two patents from that period for algorithms in speech recognition. Most of the schemes, algorithms, and networks built to do things in those days are actually very similar to what's being done today. It's just that today you have an entire internet's worth of data and enormous compute. We worked on machines that were state of the art then but had less compute power than your phone.

I actually started as a graduate student at the AI lab at MIT. My master's thesis was on the foundations of statistical inference and hidden Markov models for speech recognition. Then I began to realize that the engineering approach to the brain wasn't really answering the questions I had about the nature of thought. I didn't know any biology, certainly didn't know neuroscience, and wasn't even aware of neuroscience as a field of inquiry.

So I switched back into a PhD in physics, moved to Princeton. When I started the PhD, there was a new professor then — Stanislas Leibler, who went on to essentially invent synthetic biology. I was his first student. In his lab, I built his first microscope to track individual bacteria as they swam around, connecting genes to behavior. I loved it. But then the duality revolution began to happen in string theory, and it was just too mathematically beautiful for me to not pursue — you have one shot in life to do these things.

So I switched back into that. But then towards the end of my PhD, I became aware of systems neuroscience as a field. Even while working in string theory, my first paper as a graduate student

was called "Occam's Razor, Statistical Inference, and Statistical Mechanics in the Space of Probability Distributions" — it was actually about statistical inference by machines, not string theory.

I then got a postdoc at the Harvard Society of Fellows, which made me an independent fellow. That allowed me to do my day job as a string theorist and moonlight in a neuroscience lab. In the evenings I would work with a friend named Michael Berry, who was a retinal neurophysiologist and a postdoc at the time. We started writing papers together. That was my entry into neuroscience.

Once I arrived as a professor, I thought I would switch fully into neuroscience and was getting ready to apply for a postdoc there — but then I got faculty jobs in string theory. Given that faculty jobs are very difficult to get, I took one and then started collaborating with people in neuroscience again, at Penn. That started a long series of papers and projects.

My approach arriving at the questions of learning was through early work on structural adaptation of neural circuits — evolutionary adaptation or developmental adaptation, which I view as a kind of learning. I was very interested in how neural circuits mold themselves to the tasks they need to do so that they can be efficient, given that they don't have very many resources.

Working on those kinds of problems led me gradually upward to high-level learning — language, spatial cognition, navigation. I felt that as a person without a background in biology, I had to work my way up to the problems. I knew plenty of machine learning, but not much about biological learning. I felt I had to start with sensory systems and really understand anatomy and cellular physiology.

The sequence of papers I worked on has been a process of learning my way through the field — pick a topic, understand it; pick a topic, understand it. In the end, I hope to now work on things that originally motivated me: biological learning, what kinds of computation is the brain capable of, how that informs the way you build machines that learn, and how machine learning informs how the brain learns. These are questions I originally cared about but didn't know how to answer as an undergraduate.

I feel we can do so now partly because I know more, and partly because both neuroscience and machine learning have progressed enormously. We just know so much more than when I was a graduate student, and I think these questions are now accessible.

Dr. Chandra

It's a complicated but very interesting trajectory! So what aspects of these problems weren't addressable earlier that are now?

Prof. Balasubramanian

The basic architecture of the brain is so different from the basic architecture of any engine people are doing learning with, even now. Back then it was even worse. The experimental data people could record was from one neuron or a handful — there's this enormous network in the brain, and researchers had only tiny bits of information about it. It wasn't at scale.

Likewise, neural networks back then would have a couple of layers with a few units, would take forever to train, and also weren't at scale. The kinds of things you could do — recognizing the numbers one through ten — that's nowhere close to what I would call interesting aspects of animal learning.

What's changed, I think, is a matter of scale. In neuroscience, using everything from fMRI to optical recordings to electrical recordings from very dense electrode arrays, we can now get data from hundreds or perhaps thousands of neurons simultaneously. Optical recordings give you thousands of neurons. fMRI gives you whole-brain coverage at a coarser scale. The connectome projects are giving very fine-grained detail about the actual anatomy and circuitry. Transcriptomic methods have given us a much more detailed picture of the cellular components of neural circuits.

Likewise, machines are now being built at scale — with the kind of number of units you might compare to the brain, with enormous numbers of parameters. And the performance of chatbots is so astonishing as to be weird and requires an explanation.

That said, I still feel that even today, the actual machine architectures are nothing like the brain. The brain has an enormous diversity of cell types, wired in highly specific circuits at every scale. In something like a fly, it'll be exactly the same circuit in every brain. In humans, it'll be statistically the

same. But you recognize the architectures again and again in every mouse, every human, at every scale.

Interestingly, if you look at the evolution of the neural networks people are building nowadays, they're starting to add "harnesses" — special purpose components for mathematical reasoning or other specific tasks. To me, that resembles a sort of engineering-based evolution of these architectures, analogous to what evolution did over time. So I think there will be things to learn from that comparison.

Dr. Chandra

When we train these networks with tasks in mind, are those tasks similar to what biology is perhaps trying to optimize towards, or are we just coincidentally arriving at similar solutions?

Prof. Balasubramanian

That's such an interesting question — it gets to the heart of the practice of science. The way we do science is: you make a hypothesis, build a model (mathematical or neural network), fiddle parameters, and if it matches the phenomena you're trying to describe, you try to make predictions for new phenomena. If those work out, you've gotten something.

But what have you learned? You could ask the same question about a mathematical model as about a neural network model. Some weird mathematical object produces an effect that looks like something you saw in the world and makes a couple of predictions that work out. What have you actually learned?

In biology, to get an answer out from these things, you have to hypothesize what the system is for. Biology is soft matter, but soft matter with purpose — it's trying to do something. You've made some guess about what this piece of tissue or circuit is trying to do. You don't know, of course — evolution selects for survival of the organism, so you don't really know what goal this little piece has to achieve. But you can make some guesses, and that's the job of the scientist.

You've also usually made some assumptions about the constraints present in the system. Because if you don't constrain things, if you try to get the best possible circuit or equation, you'll get something mathematically optimal but with very little relationship to what happens in the world. For living things, it's a world of constraints — so many cells, so much time, so much energy. So when things work out and the model resembles in some structural way the thing you see in the world, one of the things you've learned is that perhaps your hypothesis for what the system was trying to do, and your hypothesis for what constraints mattered, weren't so bad. Notice I'm not saying they're correct — establishing correctness is a very high bar. One prefers to say they're not completely wrong. It's a modest way to put it, but it's good enough: if it's not completely wrong, you can work with it and improve.

No model is the world, but it can resemble or extract properties of the world. You hope you've extracted the correct universal aspect of the thing you're trying to describe, such that it confers some understanding — and understanding is itself a loaded word. I mean understanding in human terms, given the limitations of my own head. I want to construct models in a way that confers human understanding. At least that's what motivates me.

Dr. Chandra

There's this ongoing debate about whether building machine learning models that replicate biological phenomena — say, vision models or spatial navigation — actually gives us insight into the brain, or whether we're just building another model that can solve the task. Are we swapping one hard-to-understand model for another? And are there heuristics by which we can know if we've actually gotten useful insight?

Prof. Balasubramanian

I think we're in early days yet, and the methodology for doing that has not yet been fully established. You might draw an analogy to mathematics giving insight into physics. We now have, developed over a century, ways of thinking about Lagrangians, actions, fundamental fields, Feynman diagrams — a whole mental toolkit that allows us to conceptualize and understand what's going on in terms of very complicated equations.

Take Bessel functions. A Bessel function is a box — you put numbers on one side, numbers come out the other. It's no different from a neural network with multiple layers. But over a century or two, people worked out properties of this object. It recurred as a solution to many different kinds of problems. People worked out: when its argument is big, it acts like this; when small, it acts like that; there's a series expansion; there are interesting relationships. So we understand many qualitative properties, and that allows us to reason about consequences.

I think what will happen — what needs to happen — is we'll develop a similar understanding of computational building blocks. We'll be able to say: certain kinds of networks have certain kinds of properties, and we can reason things out from that. Some of that is already happening. Fifteen years ago, if you wanted to build a deep network, you had to program everything by hand. Now you bring up TensorFlow and say: put down a convolutional layer, an attention head — and you just do it, because you understand what those things do. After training, a convolutional layer will extract features from the input. You understand that at a qualitative level.

Similarly, if I build a recurrent layer with balanced excitation and inhibition and train it with the right objective, I know it'll produce various kinds of attractor wells — something like a Hopfield model of memory. Knowing that, I can reason my way into what layers I should stack to achieve some goal. The process is starting where we have a logic of these little pieces — like functions. I stick a Bessel function here, I do this, and I know what it will do. Likewise, I know these layers will do things. Some tweaking of parameters is necessary, but there's a growing logic.

What will have to happen is that kind of mental understanding at various scales of organization — within the units, across layers connected in different ways. The challenge now, compared to before, is that all of the systems we're talking about are vast, with many, many small interactions. And we know from example after example in physics, chemistry, and biology that you can have emergent phenomena — many elementary objects that you understand perfectly well, interacting in weak ways, and yet a completely unanticipated phenomenon can emerge that you'd never predict from the elementary behavior of the constituents.

I would say the language produced by chatbots has that character. And I would say my own behavior, produced out of whatever's inside my head, has that character. Who would think this motion in my head could produce this? So the deep challenge in everything I've said is emergence: new phenomena that are not easily implicit in the constituents, but rather in their interactions.

We do not, in any field, have a systematic way of thinking about the logic of emergence. We have examples, and every time it happens we say, "Isn't that marvelous?" But to generalize and understand the logic of how things emerge, you first need many examples, because there will be different ways these things happen. The logic of emergence is a very long-term goal — it's relevant in sociology, in physics, in every field. But in the shorter term, I think coming to some logic of computational models in neural networks can happen, and that is related to understanding the computational logic of brain areas and why brain areas are made of little other modules.

I view all of these — both in brains and in machines — as breaking big jobs into little pieces, each with a well-defined logic, a well-defined interface: you know what it's going to do, you give it structured inputs, and you read out structured outputs. That's a common organizational principle arising in both brain-like and computer-like systems that are supposed to do some thinking.

Dr. Chandra

The fundamental question I still grapple with is: how do I know I've arrived at a solution that is actually relevant to biology? What makes for a good model in these spaces? How do I know I'm finding the same kind of solution as what biology is doing, rather than just arriving at something completely independent that happens to solve the same task?

Prof. Balasubramanian

It depends upon what you want to use the model for. I always tell my students: all models are wrong, but some are useful. If you're a particle physicist, you sometimes have the delusion that the model is reality. But I think that's probably fiction. We're limited beings and most models are kind of wrong, but useful — and we want them to be wrong in the right way.

The traditional way of checking that your model has some range of validity is to get it to predict new phenomena. When you build a model, it has free parameters that you fit so it reproduces the things you started trying to describe. Then you check: does it have additional consequences? Are they interesting? Are they testable? Usually you'll find they're kind of right, but a little bit off — and that

"little bit off" gives you the guide for refinement. Sometimes that refinement means adding bells and whistles; sometimes it means completely rethinking what's going on.

Sometimes models can explain what's happening in the biology at a conceptual level, but not the mechanistic level. Suppose the goal is navigation. There's no guarantee there's only one way to navigate. You might find completely different solutions that achieve the same functional goal. Just because there's some scale separation doesn't mean you have to build a navigational machine out of cells that have genes. On a computer, you certainly wouldn't.

In some circumstances, the details at the lower level do matter. For example, in the brain, adaptation of a circuit or growing new synapses involves gene expression. And there are phenomena in the brain that we don't include in machines — for instance, neuromodulators in the brain are global knobs that change how things learn. We don't usually put those in machines, but we could.

Let me give a concrete example. I'm working on something called Dravet syndrome with colleagues at the Children's Hospital of Philadelphia — a form of childhood epilepsy where children are fine until their temperature rises above a certain point, at which point they can have a seizure. This can be fatal. If you made a model that just said, "I'm modeling the overall pattern of firing of a neuron," there's no way you could describe what's going on. Because it turns out the problem arises from a particular ion channel in a particular inhibitory neuron in the brain. You have to know that detail.

And knowing that, you can propose a solution: this ion channel modifies the conductivity of the neuron through the sodium channel at high temperatures. One way of getting the neuron to not do this is to use a drug that modifies the conductivity of the potassium channel — a drug already approved for other treatments. So you can use it to prevent seizures. For that you need to know the detail. But that detail is not relevant for understanding the overall computations the brain seeks to do.

So it's a question of scale and the kind of question you want to ask. Even in physics — Newton's model of gravity works beautifully for most purposes, but if you want to describe the precession of Mercury's perihelion, it doesn't. You need general relativity, a much more complicated model. And that will surely turn out to be wrong too, for some sufficiently intricate phenomenon. It's the same set of issues, just writ large because these systems are so vast and interacting.

Dr. Chandra

In much of what you've been talking about, you've been referring to emergent phenomena. You've also looked at emergence in non-biological contexts — such as how aspects of spacetime and geometry are emergent phenomena. Do you think there are conceptual parallels that help you relate these kinds of emergence across different domains?

Prof. Balasubramanian

That's a question I'm asked often, because the fields I've worked in seem pretty distant from each other.

At the methodological level, I've used information theory in basically everything I've worked on. I actually learned information theory first by working in machine learning, then extensively in neuroscience. Only later did I come to quantum information theory, which I've been using in the other domain. And I seem to understand most of the concepts in a more classical way because of the long prior usage in these other fields. So there's a methodological similarity and definitely a similarity in the way you think, even though the phenomena are so different.

My first paper as a graduate student — on statistical inference by machines — occurred because when I arrived in grad school, I hadn't taken some of the courses other physics theory students had taken, like general relativity. I needed to know the subject to pass my qualifiers, so I learned it in a hurry. And suddenly I knew all the methods of differential geometry. That paper arose by applying the methods of differential geometry to manifolds of probability distributions and thinking about what their structure is — how inference works on the surface of distributions in the space of distributions. It came directly out of having just learned general relativity. That's an example of how interdisciplinarity helps: you have a different toolkit, and because you have it, you can say, "This thing over here looks like a screwdriver."

More broadly, I'd like to put to you that there are actually not that many different ideas across all the fields of learning. I think this comes from the limitations of the human mind. We are animals embedded in the world, evolved to act mechanically in it. What we understand is: push something, it moves. There's something in the object. There's momentum. Bouncy balls go boing. Springs

oscillate. And if you look at all the ways scientists across all different fields describe what they're studying — whether neural circuits, individual neurons, quantum gravity, the emergence of space and time — people basically use the same language: this goes from here to here, these two things meet and bounce or merge, this thing wobbles back and forth. It's the same language everybody uses.

So there's a certain commonality of reasoning no matter what the phenomena are. I think that's coming from underlying limitations of the human brain and mind. And somehow those limited modes of thought, assembled in sufficient numbers, seem to have been sufficient so far to describe most phenomena — including emergent phenomena — because there seems to be a kind of universal vocabulary for describing complex phenomena.

I don't understand why that's the case, and understanding why is a long-term intellectual project I've had. Why are you able to explain all this stuff with such simple tools? I think it's a bit of a mystery. We have no right to be able to do that, but we are. There's something extremely deep about it, but I don't quite know how to answer it. Something I keep thinking about.

Dr. Chandra

Do you think the flip side is also true — have you found that your intuition from physics sometimes leads you astray when thinking about biology, or that there are problems in biology that just don't seem to fit into the kind of physicist's elegance we often try to think about?

Prof. Balasubramanian

Your question presupposes a distinction between physics and biology that I don't grant. Newton would not have granted that distinction either. He conceived of himself as working in natural philosophy — all phenomena of nature can be studied.

It just happens, in my view, that the phenomena of biology are the most complex we know. They contain enormously many components interacting in ways that produce extraordinary emergent phenomena. Everything is out of equilibrium, so you can't think statically. Everything's changing all the time. And there are long-term processes where systems act as if they have a goal, because of selection — the dynamics of an individual like you or me have been shaped over a long period of time to act as if we have a goal: to keep ourselves going, produce offspring, and so on. We're essentially bags of chemistry — extremely weird chemistry — that's been selected to do this.

Physicists have mostly worked in extremely simple systems. They have a tendency to think they've done all kinds of complicated things, but they've actually been working in extraordinarily simple systems. The Standard Model of particle physics is trivial in how complicated it is compared to a single cell. That's why physicists were able to make progress — there were amazing emergent phenomena arising from simple things. The theory of gravity is, in the end, pretty simple, and what's amazing is that it's universal.

In biology, physicists need to understand that if they wish to work in that area, they have to take the domain seriously. It won't do to say, "Oh, I have this simple model, I love simple models, I'll pretend this complicated biological system is my simple model." You may capture some aspect of the phenomenon — but you'll capture a very narrow, infinitesimal piece of the phenomena that exists in one of these complex systems.

The study of very complex systems has only really just gotten going. In the 1980s, people first started talking about chaos as a thing to focus on, networks as a thing to focus on. The Santa Fe Institute was heavily involved. Then in the 1990s, condensed matter theorists began talking about complex phenomena in electronic systems; soft matter got well off the ground; non-equilibrium phenomena and statistical physics became important. Put all of this together and you begin to realize that in active systems — where power is running through the system — components can do all kinds of crazy things.

Those of us who have moved on to work on living systems understand, I hope, that we have to work at a completely different scale and that the emergence of phenomena is among the most interesting things.

Now, let me address the issue of elegance, because part of what you're asking about is physicists' cultural belief that there exist perhaps simple, but certainly beautiful, explanations of phenomena. We have a tendency to use those words together. "Beautiful" is like the most used word in physics. And sometimes, when you look at a biological system, one has a doubt: is there any kind of simple or

elegant theory here? Biologists, long experienced in the sheer diversity of living systems, are often very skeptical of this.

So you often run into a tension. As a physicist, you'll say: I know the system has a bazillion details, I'm going to ignore almost all of them, keep three or four, build a minimal description, and it reproduces the effect. Then you submit this to a biology journal, and the biologists will say: but so-and-so's experiment had this detail, these two experiments contradict each other — don't all these details matter? Go back and build a model that contains all of them.

I've personally found myself many times in long-running arguments with referees, trying to explain a mode of doing science where all those details matter but may not matter for the particular phenomenon in question. When looking at smooth laminar water flow, all the little bumpy molecular details may not be important if I just want to describe the laminar flow. In physics, the experience has been that if you're trying to describe one or two simple things, you can simplify the system because most of the details weren't relevant for that. To describe the pressure of the gas in this room, I don't need to know where every single gas molecule is.

The danger — the potential threat to that mode of analysis — is an absence of separations of scales. If small phenomena and big phenomena are more intimately tied together and there's no way of coarsening the small into effective big phenomena, then that will threaten this kind of analysis. But personally, I don't think that's generally going to be the case.

That said, there are cases where you do need the detail. As I mentioned with Dravet syndrome — there, you need to know about a specific ion channel in a specific cell type. So the question is always: what scale and kind of question do you want to ask? Even with gravity, Newton's model works beautifully — until it doesn't. It's the same set of issues, just writ large.

Dr. Chandra

You were talking about the boundary between physics and biology being blurry — perhaps nonexistent. But certainly there are students and researchers who think of this boundary as sharp. What would you say to them about the value of thinking about these apparently messy biological systems?

Prof. Balasubramanian

I don't actually try to convince anybody that there is value to this. I do what I do, and if they hear me give a talk, they realize it's extremely interesting. I think everyone has to show by example.

I'm a trained string theorist, so I absolutely understand why one might be attracted to the pristine beauty of mathematical descriptions of quantum gravity. It's a wonderful thing. But there's a completely different kind of pleasure in being able to explain phenomena in exceedingly complicated biological systems. There are so many components, and yet you can explain why things happen. That's really wonderful. And it has the additional joy that there's just an overflowing supply of experimental data, and experimentalists will talk to you and work with you if you make friends with them. It's a different kind of joy.

Not everybody will like both kinds of things. Some people motivated by the beauty of mathematics may not find working with messy data that interesting. And people who really love working with data may not find the dry and complicated mathematics compelling. That's fine — it takes all kinds to make the scientific landscape.

My approach for students is: people vote with their feet. I just do what I do. And if you look at what students are actually doing, increasing numbers are going into biophysics, because even physicists, when they see the papers and talks, realize they could actually work on this — it's interesting, and they never thought they could. That was the case for me growing up. I just assumed the biological stuff, the living world, wouldn't be amenable to the kinds of analyses I liked to do. As it turns out, it is — and that message propagates.

In the United States, the fastest growing section of the APS March Meeting is the Division of Biological Physics, and there's a reason for it. There's just so much to do, so many phenomena. Departments will move more slowly — institutional systems are a bit sclerotic, which is why, again, things like CALIBRE are very useful. Sometimes it's useful just to create a new space that's not taking away resources from the old space, so nobody's opposing it. It's just a new thing. And then after it's established, either the whole system expands, or if some fields are not as exciting to people anymore, young people don't go into them and they shrink. That's not bad — it means the earlier

generation did such a good job that they answered the key questions, and people move on to new things.

Personally, I would prefer a model of institutional growth that wasn't so siloed — physics departments where physicists can do anything. But that's easier said than done in a world of finite resources.

Dr. Chandra

What are the directions you're most excited about at this intersection of AI and biology? What do you think are the most promising things?

Prof. Balasubramanian

I would really like to know why these AI systems are able to work the way they do. Why are they able to produce such excellent simulacra of language — or maybe it isn't a simulacrum; maybe this is what language is. Why does it work, how does it work? Most of the AI companies aren't focused on that because they're engaged in a race to make the best engine quickly and corner the market. That's a different motivation. But I think the more interesting question is: why does it work? That's what I'm thinking about.

My own work gets partly driven by my students — I actually don't tell my students what to work on. I tell them to tell me what they want to work on, because that helps me move in new directions. Young people know what's hot on the ground. I have at least one student now, a physicist, whose primary motivation is to understand why and how these systems work. He's gotten me going, and we have a couple of papers at machine learning conferences.

I was previously quite skeptical about the approach to neuroscience and biology where you say, "Here's the task you want the system to perform, put a network in, press go, open it up and see what's inside." A lot of that work was done in a slightly shoddy way — you do something, you get what you get, and people tend to see what they want to see. They say, "Oh, it looks just like the brain." Well, it doesn't really. The brain has six processing stages between the eye and recognizing a face; that model has 100 layers. That doesn't look the same.

But I've come to think it may actually be possible to do this in a more informative and interesting way. And again, that's been driven by students from the computer science side who said: let's build networks that are constrained in ways related to the brain and see what happens when we train them to perform some task — do they produce circuit architectures or unit behaviors that have the phenomena we see? And can we predict anything about the system from that? That's now more interesting to me than it was a year or two ago.

I've also begun to work on problems in cell biology — trying to use ideas from neuroscience and machine learning to think about the logic of biochemical circuits in the cell. Similar algorithmic thinking about systems, for both machines and the brain.

I have a class of questions regarding the repertoire of cell types and repeated circuit architectures. Why do we need so many cell types? Your retina has at least 65 types of cells — that's ridiculous. The brain itself has thousands of types. Why do you need more than one kind of cell? You could build an arbitrary computer using only a NAND gate. Maybe it's an efficiency criterion; maybe it's just historical contingency. These are all askable questions, and I think we can and should answer them.

I tend to think of the brain the way a classically trained programmer thinks about a large program: you break the problem up into a bunch of procedures or functions, you write each of them carefully, running fast, with good interfaces, and you hook them up together. The brain is a bunch of things that were selected over time for doing jobs really well. Why do you make these particular parts and not different ones? That's a long-standing interest of mine.

Thinking in computational terms may help me with that, and vice versa — thinking about the brain in these terms may help with the design of learning engines that are far more data- and energy-efficient than they currently are. I happen to think that part of the reason the brain is organized the way it is, is because it's efficient with respect to resources.

My laptop is running at 120W right now. My brain runs at 20W, with a tiny fraction of that devoted to learning. And I'm powered on mangoes. Just so much more efficient. In the future, that's the kind of machine we will want — because we can't afford, as a planet, the incredible amount of resources being spent on current AI systems. The power, water, and financial resources being sucked up into this enterprise — there are other productive uses for all of that.

I really think that the current approach uses an overabundance of resources because of the architecture — we're using matrix multiplication because that's what we have with GPUs, which were originally optimized for computer game graphics and ray tracing. It turns out that's also what you need to do backpropagation in feedforward neural networks. But the brain doesn't do matrix multiplication. The brain also doesn't globally optimize anything — there would be no way to tell every single synapse in the brain how to change at the same time in relationship to every other synapse. What the brain does instead is use local learning rules that adjust the whole system.

There's a field called neuromorphic computing that's supposed to be about more brain-like computing, which is far more efficient. I personally think hardware architectures inspired by this are going to become very important on a 10 to 20 year horizon. I'm actively working on how to design such things — not with any expectation of short-term impact on anyone's stock price, but with the belief that 20 years from now, these will be the important things.

The way these things work is: there's some kind of breakthrough moment where something gets demonstrated that couldn't be done before. Then people put attention to it, there's investment, clever engineers start thinking about it, and they figure out how to do it at scale. The scientists design a small system and show it can work — and once you see the opening, people are very clever. I think these things will be built at scale in due course.

I remember the 1990s, because that's when I worked in AI. Many things were tried, including most of the things being done now. It took a couple of decades to figure out what was necessary and how to build things. From around 1994 to AlexNet in 2012 or 2013 — that's roughly a 20-year span. And AlexNet was basically doing the kinds of architectures that had been done before, except more, better, and using GPUs. So I think it's a process. Twenty years from now, the landscape will be different — and we might as well get into it right now.

Dr. Chandra

That's absolutely wonderful. There's clearly a lot that biology has to offer to AI and machine learning. By having you here at ICTS, we're really glad to be able to discuss these problems at the intersection of AI, physics, and biology, and to think about new ways of approaching them. Thank you very much for being here and for sharing your insights. We look forward to continued dialogue at ICTS about all these emerging frontiers.

Prof. Balasubramanian

Thank you.