CRAIG: 00:01 Hi, I'm Craig Smith and this is Eye on AI. We've been slow getting started this year, so apologies to regular listeners. I've been working on an AI enabled audio editing tool to speed production of the podcast. I'll let you know if it works. This week I talked to Terry Sejnowski, one of the pioneers of deep learning who together with Geoff Hinton created Boltzmann machines, a deep learning network that has remarkable similarities to learning in the brain. I had recently read Terry's wonderful book Deep Learning Revolution, and so our conversation followed much of what I learned from his writing. Terry has a remarkable mind, focused now on the convergence between deep learning and neuroscience. We talked about whether machines dream, and the algorithms of the brain, whether Marvin Minsky was the devil, and how deep learning is shaping the future of education. Terry is also the chairman of the NeurIPS foundation, which puts on the premier AI conference each year. We ended by talking about what's in store for the conference, which has grown and grown in recent years. I hope you find the conversation as captivating as I did.

CRAIG: 01:21 Speaking about deep learning, you're one of the founders. So I thought I'd ask a little bit about how you got into machine learning and in particular around the time that you and Geoff Hinton developed Boltzmann machines and what Boltzmann machines meant for the development of deep learning.

TERRY: 01:41 Well, I was trained as a physicist, theoretical physicist, but got my PhD from Princeton university and was fascinated by the brain. And what do physicists do? They, they write down some equations and try to solve them. And these are particularly difficult for the brain because the complexity of the brain is so great and all of the equations are nonlinear and in mathematics, these are the toughest equations because there are, there are no methods really that can allow you to analytically solve them, predict what the behavior is. So at the time we, you know, we were doing simulations with computers that really were puny in comparison to today's computers. Machine learning was in its infancy back then. The particular way that Geoff and I got into it was through neural networks that were bio inspired from how the brain is organized. But these are much simpler in terms of the connectivity patterns and the complexity of the neuron itself in their model.

TERRY: 02:38 They're just simple functions, nonlinear functions. But they were sufficiently complex that we were trying to use them to solve complex problems in vision. You know, vision seems very simple and deceivingly so because of the fact that when we open our eyes in the morning, we see objects, we pick them up, you know, that's seems like it doesn't take any, any effort, any thought, and that's because nature has evolved our brains to see and to move over hundreds of billions of years. And that machinery is just opaque. We don't understand it. We had no concept back then for the degree of complexity. In other words, how much computing power the brain has. And so it was really a challenge and not just, you know, for us trying to understand the brain, but also for engineers who were trying to write programs that could run on a computer that would allow a computer to see as well as people.

TERRY: 03:32 But we started out, our goal was to try to solve a problem that had been thought to be impossible to solve, which is trying to take a network with multiple layers - They'll have an input layer, you have an output layer and you have layers in between in the real brain. Those would be the cerebral cortex over the surface of the brain that really is representing the world and all of your plans and actions and you know, it's the highest level of processing in your brain. There's an outstanding problem, which was how do you learn in a system that has that complexity with all those layers. And it was generally thought because of early work that was done in AI in the 60s that no one would ever find such a learning algorithm because it just mathematically it was it was too difficult. And that's where Geoff and I invented the Boltzmann machine.

TERRY: 04:20 And so what is the Boltzmann machine? The Boltzmann machine is an architecture that's inspired by physics. And what made it different from all the other architectures at the time that were being looked at was that it was probabilistic. See, normally if you have a system where the input comes in like into your retina and it goes through a series of stages. It does that in a deterministic way in the models that were available at the time. But what we tried to do was to say, look, maybe we can make progress if instead of automatically getting the same output, if the unit itself would have a probability to have an output that varied with the amount of input that you're giving it. So more input probability gets higher, it's going to produce an output. And if the input is low, you know you'll still produce an output with a very low profitability. And it introduces a degree of variability.

TERRY: 05:11 Not only that, but it creates a different class of network, which is generative. And what do I mean by that? By that I mean in the traditional input output network, you know, no input, no output. Basically there's nothing going on inside. But in the Boltzmann machine, even without an input, the thing is chugging away because there's always some probability that there'll be an output from each unit and therein lies the secret that we discovered to how you are able to learn in a very complex network with many layers which we now call deep learning. And that was by giving the network the input and then keeping track of the activity patterns within the network and for each connection you kept track of the input and the output, the correlation. But then in order to be able to learn, and this is all mathematical analysis that we had done, you have to basically get rid of the inputs and the outputs and let it free run in a sense, put the network to sleep - but it's not of course, it's chugging away and you can do the same measurement for every pair of units with a connection.

TERRY: 06:14 You keep track of the correlation and we call it the sleep phase and the learning algorithm is very simple. You subtract the sleep phase correlation from the awake learning phase and that's how you change the strength of the weight. Either it goes up or it goes down. And we showed that if you do that and you have a big enough data set that you can learn arbitrary mappings. And not only that, and this is the generative part and this is a much more elegant architecture than the traditional back prop that we now have that is used routinely, is that once you've trained it up, you now can look at the output and say that you've trained it in order to be able to discriminate between, this is what we at the time, what we were doing, handwritten digits on zip codes, you know, so there's 10 output units. And so you give it some input, which is a little handwritten number 2, and then the unit at the very top, which represents 2, is going to be active at the highest level compared to the other units.

TERRY: 07:13 And so that was how you classified the digits. But now what you can do is clamp, we called it, which means that you fix the output of the 2 so that it's the only one that's active and the rest are off. And now that percolates down because this network had inputs and outputs going up and down. It was literally a very highly recurrent network. And what it would do is start creating inputs that looked like 2s, but they would be constantly changing. You know, the, the loop at the top would come and go and then the loop at the bottom would come and go and it would, they would wander around. And so it was basically dreaming. It was a dreaming about 2-ness and, and the network had created an internal representation of what it meant to be a 2.

CRAIG: And when you say put it to sleep, you mean stop with the inputs?

TERRY: 08:06 That's right. That's right. In other words, you prevent any input from coming in so that the network could express and input that represented this concept at the highest level. And so the information now instead of flowing from the input to the output is flowing from the output to the input. And that's what's called a generative network. And now we have even more powerful generative networks, the generative adversarial networks, which are amazing because not only can you generate 2s, but you can generate pictures of people's faces, you can generate, you know, you give that, you have to give it a bunch of examples, right? It's just give it a bunch of examples of rooms like the one we're in and it will start generating new rooms that don't exist with different kinds of tables and chairs and windows and they all look real, photo realistic. And that's what's really astonishing because we can create very high fidelity models of the world.

TERRY: 08:56 And in a sense that's what the brain does because when we fall asleep and we dream, I mean that's exactly what we're seeing. We're seeing the generated patterns that are based on our experience.

SHORT INTERLUDE: 09:04 CHORD

CRAIG: 09:05 That's a nice analogy, but you'd say in the book that a lot of times analogies are wrong. Do you think that it really is an analogy for brain learning during sleep?

TERRY: We thought so. Geoff and I were completely convinced we had figured out how the brain works. In other words, is it just a coincidence that in order to be able to learn in a multilayer network, you had to go to sleep? Humans go to sleep every night for eight hours. Why did we go to sleep? In fact, and this is a really fascinating area because one of the areas that now that I've helped to pioneer is trying to really understand what goes on in your brain when you fall asleep.

TERRY: 09:41 There are scientists and people who are doing computational models like me, have really made a tremendous amount of progress on understanding something about how experiences you have had during the day get integrated into your brain at night. It's called memory consolidation and there's an overwhelming amount of evidence now, both on the psychological side, but also recordings, that this is what's happening. There's something called replay that happens between a part of your brain that's important for memories, episodic memories, things that have happened to you, events, unique objects, things that - something happens to you during the day. This never happened before, right, and you remember it. You need the hippocampus for that, and during the night the hippocampus plays back literally those experiences to the cortex and the cortex then has to integrate that into the knowledge base, this semantic knowledge that you have about the world. It turns out that the Boltzmann machine analogy turned out to actually to be a really good insight into what's really going on during sleep. But now obviously, what's really going on during sleep is orders of magnitude more complex in terms of the numbers of neurons, the patterns of activity, which we have studied in great detail, but we really think that computationally, it's actually what's going on.

TERRY: 10:55 So this is really, this is a theme that has come out in which, which is really the central theme of my book, which is that there's a convergence going on right now between our knowledge of the brain, on the one hand, and our ability to now create these large scale networks in the image of the brain. Not precisely, we're not trying to duplicate the brain, but rather take the principles from the brain and try to build up systems that have some of the capabilities of the brain, like vision, like speech recognition, like language processing and the, and this is really going back and forth now because now neuroscientists are watching what's happening with deep learning and getting inspired and coming up with hypotheses and now going back and testing it and the brain. And as we learn more about the brain, how it solves these problems, we can take that. And I'll give you some examples like attention. You know, one of the things while we're looking around is, is that we were not just trying to process everything that's out there. We focus, right? You focus on a particular object, you want to pick up, focus on reading, you're reading a sentence and you're, you're looking for something, right? And that means you have to redirect your attention around. Well, it turns out that if you add attention to these deep learning networks, you vastly improve their performance.

CRAIG: 12:06 When you add attention in, in, in what way? By focusing the units?

TERRY: 12:10 By having the network decide what's important in a scene like this, in other word, salience, what's, what's important, where should you be looking? Or if you're trying to do language translation, you know, you may want to have a word at the beginning of a sentence, may have a strong relationship with a word later in the sentence. And so you want to be able to hold onto that information, attend to it while the inputs are coming in in sequence. And now another word shows up and those two words have to link together, right? So that's why attention is a way of marking and saying this is important. Keep it in mind. And then after you've linked up all these words in into a semantic, it's now a meaningful representation. You then begin to output words in another language. Again, respecting those relationships between the words, how they're ordered and what their clauses look like.

TERRY: 12:58 And in German you have to wait till the end of the sentence in order to put the verb, right? The network has to understand that, it has to keep track of what the verb is, know what the verb is and know where to put it. And this is all something we take for granted, right? That's what our brains are really good at. And so as we learn more about the mechanisms that the brain uses for processing words and also, you know, speech, vision and so forth, these will get incorporated and improve the performance of the networks. And now, especially with natural language processing, this has really reached a point as you probably know from your cell phone where it's really good. I mean, you know, speech recognition has gotten amazingly good for, you know, speaker independence speech, right? This is just an amazing, you know, ability now even in noisy environments to be able to detect people, you know, and, and now even, you know, voicemail is getting transcribed on your phone. So this is like, you know, it's a whole new era.

SHORT INTERLUDE: 13:49 CHORD

CRAIG: 13:49 Yeah. Two things on that, on the attention part. I read a paper that they were doing some psychometric testing of networks and the network could solve certain problems as well as or better than humans, but they really broke down on a, it was called a concentric circle test or something where there were dots and you had to identify which pattern was in a concentric circle even though the lines were not…

TERRY: Yes, I'm familiar with that test.

CRAIG: Okay. And the network performed very poorly. And they realized that the sensors were evenly dispersed in the field of vision. And just intuitively they thought, well, let's try like focusing the sensors and giving the computer vision foveal vision focus and bang it, it outperformed the humans. I mean that to me was really interesting about, I don't know if that's related at all to the idea of attention.

TERRY: 14:52 Well, what you're describing is overt attention. So because we have a fovea and can move it around, we can automatically attend to, with high resolution, a particular object in front of us and then jump to another object or on the face. You can, we're not aware of it but we are. We're jumping around three times a second and that's means you're taking in input and combining them across - they're called [saccades](https://en.wikipedia.org/wiki/Saccade) - very fast movements. But then there's covert attention. In other words, I could be looking at this and attending to you and that means that I have the ability in my mind's eye to - you'll switch information channels around and both of those are important. And you're describing going from a camera which has a uniform resolution to a foveal representation which you have very high res at the very central few degrees and then falls off in our eyes very, very rapidly.

TERRY: 15:49 It's still very sensitive to motion and to other things that you need for alerting you. If something is coming at you from the side you, you want to respond to that quickly but you may not be able to detect what it is with the low res but then you could look at it and - it's interesting. This is another case where the model that we have for computer vision is based on the camera, which is frame based. So when you're taking a video, it's really a sequence of frames with images and your brain then puts them together into a sequence and so you can see motion and recognize things that are moving. There's a whole new generation now of cameras that are based on how your retina works. Your retina is actually a part of the brain, it's a little pouch in your eye - back surface - and what it does is it through several layers of processing It then converts the image first into electrical signals and then into spikes.

TERRY: 16:38 The information that's flowing into their brain has coded information about things having to do with color, motion and other properties. For example in time. How are things changing in time and the relative strengths, for example, on an edge where you have a change in contrast that's coded in spikes, so you have all of that information. Now this train of spikes is asynchronous. What I mean by that, unlike a frame where you collect information over 30 or 40 milliseconds, you can send a spike at any time. And that means you can send out spikes as something occurs in the world within a millisecond or less, you know, microsecond precision and the relative timing of the spikes carries a lot of information about where things are going. Much more information than if you use a frame based camera.

TERRY: 17:34 So it turns out that a lot of computer vision is simplified. If you use the spike based representation, it's call it a dynamic vision sensor. And what's nice about them is that they're very low power cause they're only putting out these spikes in this very sparse sense that if nothing's moving you actually don't get anything. You have to have motion. And it's a very lightweight, it's the perfect thing that you can, you can use, for example, if you want a robot because you know, you know power and the robot is just like, you know, very, very expensive. You know. And so if you can do vision with spikes instead of supercomputers, the GPU, which is what is being used for deep learning, now it's easier to be autonomous. And that's where we're headed. That is to say edge devices like your cell phone and your watches now, they are computers and they're soon going to have chips in it, which are deep learning chips, which are very power hungry.

TERRY: 18:32 So you know you have to have better batteries. But ultimately if you could replace the digital circuitry with some of these analog VLSI circuits, like a DVS camera, that is going to revolutionize, you know, the amount of computing you can do on board. And you know, in your hand.

CRAIG: The spike model is analog as opposed to digital?

TERRY: Well the spikes are interesting. So the neurons in the brain, emit these spikes and they're all or none, they last about a millisecond. So, and they're very relatively slow compared to digital electronics. It's in that sense that they, they are analog. Ultimately, the difference between a digital circuit, digital chip has a clock and every cycle, every transistor is updated, right? So you have to have a synchrony across the whole chip. Whereas in one of these analog VLSI chips, it's asynchronous. So every single model neuron can send a spike whenever it wants.

TERRY: 19:30 And these are then transferred up the road to the other chips, this route through a digital line. So it's a hybrid chip, right? It has analog processing, which is really cheap and not very accurate by the way, but that's okay. It turns out if you do a lot of parallel computations with a lot of elements and then integrate that information, it turns out that you're, you're better off. But to communicate between chips just like the way the brain does, you have to, you have to use these, these long distance connections. And, and in the case of these analog VLSI chips, so you can basically convert it into a digital bus, send the information over, you know, using some protocol.

INTERLUDE: 20:07 MUSIC

TERRY: 20:08 That was the 80s and that was a very exciting time. But once we realized that learning is possible in multi-layer networks, then a bunch of other learning algorithms were discovered literally within years.

TERRY: 20:20 And the one that has been the most popular is the backpropagation of error, which requires that you take information about how well you're doing and by comparing it to a teacher or labeled input and then using that error to go backwards and update the weights as you go down. And that was very efficient cause you know stochastic gradient descent is basically, you're always reducing the error and you can do that very efficiently, very quickly. And so because it's so efficient, it's now the way that most of these practical problems are attacked with bigger and bigger and bigger networks. And it's reaching the point now where you know the brain has like 12 layers in the cortex, in the visual cortex. So now people are dealing with networks that have 200 layers or more. And what we didn’t know back then, and this is the key to success, is that these learning algorithms scale very well.

TERRY: 21:12 What do I mean by that? So a typical algorithm in AI is able to solve small problems where you have just a few variables that you're trying to find an optimal solution for. Traveling salesman is a good example. If you've given a bunch of cities and say what's the fastest route between the cities? So you visit them all once, right? Well that's called NP complete and what that means is that as the number of cities goes up, it becomes exponentially more difficult. It's just, you know, at some point, doesn't matter how fast your computer is, he's just going to saturate it. And that's the problem with many of the algorithms that have been used that are used in a digital computer with a single processor, which is von Neumann architecture, where you have the memory separated from the processor. So you have this bottleneck between the two.

TERRY: 21:57 Now fast forward, here we are, the beauty of these neural networks that we pioneered in the 80s was that they're massively parallel. That means that there are simple processors the memory is located on the processor. They're together so you don't have to ship it back and forth and the brain, we have a hundred billion neurons that are working together in parallels, so, so it means that you can just do much, much more computing in real time and you don't have to worry about buffers or anything. And as you add more and more neurons to your network and more and more layers, the performance gets better and better and better. And that means it scales beautifully. In fact, and this is, this is absolutely amazing, right? If you have parallel hardware, that is to say if you're simulating each unit at the same time and you're passing the information through the connection weights at the same time, then it's called order of one scaling.

TERRY: 22:50 That means it's independent. The amount of time it takes is independent of the number of units you've got. It's fixed and that's how the brain works. The brain is working order one. In other words, as a the cortex evolved and the more and more neurons in primate brains, especially in human brains, it still works in real time. It's still works with the same amount of time in order to come to a conclusion just to recognize an object about a hundred milliseconds and that that's really, you can't get better than that. So nature has found a way to scale up computation in a way that they were way ahead of us, that she was way ahead of us. And now we’re finding that out. We're actually beginning and now hardware has become a really big part of machine learning. And the reason is that up until recently, right, there were memory chips, there are CPU chips and maybe some digital signal processing chips.

TERRY: 23:40 But now these machine learning algorithms are now being put into silicon, right? Google already has a tensor processing unit, TPU, which is, which does deep learning. But you know, there's, there's a ton of other machine learning algorithms that could be put into silicon and it's going to vastly improve the amount of computing that you can do because these are like super computers now. These chips, in fact, there's one Cerebras, they have a chip that is 20 centimeters across, 400 million processing units. Right? So that's getting up to real scale and you know it's of course it's a kilowatt so you know you have to have a power generator there, but it is scaling up. It's all being scaled up and it's really, it's a, it's a completely new type of chip that people are just beginning to appreciate.

TERRY: 24:34 And some of the advantages are really, first of all, it's asynchronous and it means you don't need a clock on a chip, right? You can just let the whole thing go. Number two, you could do with low accuracy, you don't need 64 bit accuracy. You can get by with eight bits, right? So that means a vast savings on memory and then there's a high degree of connectivity locally. So that means that the processors that are near to each other have a lot of information that they're exchanging all the time. But that's okay. That's how the brain works too. And now you have all the, you, you load all the data as it's coming in. Just the way it is through your senses, it flows through. It's like a pipeline, right? Information is circulating and decisions are being made. It's a dynamical system. It's an incredibly complex dynamical system ultimately.

TERRY: 25:21 And we're faced now with an interesting problem, which is, we can see how the problem was solved by, you know, looking at the input and the output. But we really want to know is what's going on inside, what does it learn? And now it's this, this is really the hottest thing right now is, is, is probing the artificial neural network with the same experiments that neuroscience is doing on the brain. Cause we, you know, how do you figure how, what's going on in the brain where you put your electrode onto one of the units and then you see what it responds to, when it responds. Is it firing before the decision or after? And that gives you hints and it tells you a little bit about how the information is flowing through the network. And so we're doing that now with these artificial networks and it's really, it's, it's really exciting.

TERRY: 26:08 I heard a talk yesterday, for example, from a language researcher who is using these deep learning networks in order to create what's called word embeddings. And this is a way for language, for example, a string of words coming in as a sentence to be represented in the semantic space. And once you've done that, you can use it for answering questions from articles, news article, and it's amazing. And it will answer the question, you know, it will figure out, you know, the semantics and what's going on and it'll answer the question. But now, they went in and they said, well how was the sentence represented? And so they, what they did is they looked at the pattern of activity. They, it's in a million dimensional space, huge. And then that collapses into a much lower sub space for a single sentence.

TERRY: 26:57 And then they look at the graph of how the, the, the activity for the different words that are represented. And it basically, it parses the sentence. It knows what the noun is, it knows what the phrases, in other words, it has learned the structure of, of how syntax is organized in sentences and it only, it did that on its own, you know, by seeing a lot of sentences. So, so that, so what we're discovering, this is like a little lab for language - a new theory of linguistics is going to come out of this, right? Because in the very same network we have both knowledge of syntax and semantics just like we have in the brain. The brain doesn't have a semantics box and a syntax box, right? It's integrating that information because it's all giving you hints about meaning, which is ultimately what you need as the output.

TERRY: 27:44 You need to be able to answer a question, right? You need to understand what's going on. You can't just, you know, look at, you know, the word order, which is what linguists were doing for the last century. You, you really need to know what the words mean. Right.

INTERLUDE: 27:56 CHORD.

CRAIG: 27:56 On modeling the algorithms of the brain, Boltzmann machines as you describe, seem to come close. The field shifted to backpropagation because it was so much more efficient, but it's pretty clear that the brain doesn't do back propagation.

TERRY: Well, you know, it's doing something that may be equivalent and now, you see, this gives us a real strong hypothesis saying, okay, how could the brain do it? Right? It's not going to – You’re right, It's not going to use the same algorithm. But maybe there is, there is information, there are feedback, connections, there, more feedback, connections and feed forward connections in this hierarchy.

TERRY: 28:36 What does that, nobody knows what information is being carried, you know, it's a mystery. And so that now gives us a hypothesis. Let's go in. Maybe that information is giving the earlier layer information about error, how to, how to change the weights. But it may not be the backdrop way of doing it, but there may be an equivalent way of doing it.

CRAIG: But isn't that happening in the Boltzmann machine as well? You were saying that the information during the sleep period is...

TERRY: Right. Well, that's an example. Okay. That's an example. Now the Boltzmann machine has another assumption that we make, which is that the every pair of units has reciprocal connections with equal strength, which is a pretty strong assumption. I mean, it's approximately true within the cortex, but it's not exactly true because it is doing the equivalent of backpropagation.

TERRY: 29:23 Right? But it's doing it locally. It doesn't need to have the information flowing down from here. It's all being done at the same time over the whole network. And so it may be that the brain is somewhere between a Boltzmann machine in the back prop net. Right. And this actually leads to a really exciting new, I think area of research which is of all possible computing systems, right? That are parallel, that have this ability to learn and the ability to take in lots of data and be able to classify or predict. We're just scratching the surface here. Yeah, I mean this, this is like, you know, the beginning of a whole new mathematical exploration of this, of this space and I've written an article that was recently accepted in the Proceedings of the National Academy of Sciences. The title of it is The   
Unreasonable Effectiveness of Deep Learning because it is able to do things that are unaccountable.

TERRY: 30:21 We don't understand how it does so well. It's just like this language example I was giving you. It has done something that nobody would have been able to predict, even back in the eighties you know, if you had asked me, I would have said, well that that's unlikely. That's it's too difficult. The language is too difficult. No it wasn't. And now we have to figure out why and now the mathematicians are getting interested because they have all the tools to go in and understand something about this class of functions and try to analyze the representations and to geometry. I liked that. My analogy here is what happened in math back 250 years ago when Joseph Fourier came up with a series expansion. It basically was a series of terms that when you add them up is the solution to an equation. Now when he came up with this, you know, he was just trying to solve a practical problem.

TERRY: 31:09 And in fact his paper was rejected by the mathematics journal. Why? Because, you know, he hadn't proved that the series converged. And furthermore, they didn't think these were functions. But now 250 years later, this area of mathematics has been extraordinarily productive in terms of understanding a whole new area of math. It's called functional analysis. It's, it's like a jewel in the crown of mathematics. Okay. It went from some anomaly as somebody who is trying to solve a radical problem, came up with something that seemed to work. No one knew why. And now it's used routinely by scientists and engineers to solve these practical problems. And you know and this is happening now. We have these networks, they're functions mathematically well-defined functions, and now there'll be a tremendous amount of effort. You know, many, many people now are beginning to use new tools and techniques and what will emerge from it, I think, is a whole new branch of mathematics that will give us new insights.

TERRY: 32:07 And not just not just the analysis. We're talking here about statistics. So statistical analysis up until now have been done on relatively small datasets with a few variables and very constrained. Well, we've gone from a few parameters up to millions of parameters, right? So that in terms of dimensionality, it's like going into hyperspace. And it turns out that the geometry of hyperspace is completely different. And I'll give you one example. Early in the, in the 80s when we were developing these networks, they were tiny by today's - they had like, you know, thousands of weights. By today's standards, you know, there are billions of, you have up to a billion weights now in these networks, so I mean it went up by a million. But already a thousand was so many more parameters. Then statisticians, they just took their hands, said you're going overfit, you know, you'll never generalize.

TERRY: 32:58 And it did. We didn't need a, a lot of data. It turns out, you only need as much data as you have parameters roughly. And nobody knew why. And then we were told that, you know, non convex optimization, that means there were a lot of local minima and of course in three dimensions, you know, we know we can see mountains and we can see valleys and we can see little lakes and so forth. And yeah, they're, you know, if you're in 3D, local minima, if you're doing grading descent, you're going to end up, you know, in one of these holes. Well we never got trapped in the holes. It is just something different when you go up there and with, what's different is that if there are a thousand parameters, then the chance that all thousand directions are upward is close to zero. There's always some parameter where you can go down a ravine.

TERRY: 33:46 There's always an escape in high dimensional spaces. It's a, they're called saddle points, right? Two directions go up, two go down, that's a saddle. And now if you're in a million dimensional space, it really becomes extremely diverse in terms of directions you can go. So you never get trapped until you get to the very bottom. So it's really a whole new mathematics of doing statistics in high dimensional spaces, which is a completely new territory. And that's what we pioneered back in the 80s. We were the first to go into this, like Lewis and Clark into this, into this jungle and find out what was there. And now, now we're at the point where we can really explore it really, really fast with modern computers.

INTERLUDE: 34:26 CHORD

CRAIG: 34:26 Yeah. A couple of questions. One historic. Marvin Minsky and Seymour Papert kind of killed the, the field with their book Perceptron. Papert seems to have come around. What was Minsky's view later on?

TERRY: 34:36 I mean, well, I can tell you because I confronted him once. This was a meeting in 2006 on the fiftieth anniversary of the famous ‘56 AI meeting at Dartmouth in the summer of ‘56 which brought together about a dozen computer scientists, including Minsky and McCarthy and, and, and, you know, these are the pioneers of new Newell, Simon, who, you know, it was computers were new and they said, well, gee, you know, we can program them to do things like play games and prove mathematical theorems and maybe, you know, we can write a program that can duplicate a human intelligence. That that's what was very exciting at the time. So, 50 years later, the goal of the meeting was to first of all, look back and see where they are. I mean, where are we and look forward, and this was a very interesting transitional time because it was, neural networks were just reaching a point where they are getting big enough to be interesting, right?

TERRY: 35:35 But not yet solving any real-world problem, but getting, getting there. And so one speaker after the next got up and said basically the same story, like a Takeo Kanade who was a computer vision guy at Carnegie Mellon said, you know, when I did my thesis, memory could only hold one image. And so I was able to detect a tank in one image, but it wouldn't work with the second image. He said, now my students have access to millions of images on the internet. And so we've, we now can solve it. We can detect tanks in any image and everybody basically said the same thing. And same thing with parsing sentences, right? Gene Charniak, Brown University, says, you know, with the generative grammars that were available at the time, you know, from linguists like Chomsky, you know, we, we tried to write a computer program that would use that structure to be able to take a sentence and parse it.

TERRY: 36:29 It never worked. So we started collecting data from students who would parse Wall Street Journal articles. I don't know why Wall Street Journal, but they, they, there's, you know, thousands and thousands of articles and, and what they did was they collected statistics on trigrams three words in a row and what, what their parts of speech were. And with that they were able to like, like a template to go through and match new sentences coming in. And lo and behold, he said, we can now, we've made a lot of progress to make and parse a lot of sentences pretty well. Okay. Everybody actually said the same thing, which was that without data you can't solve these difficult problems. But as soon as you have enough data and know how to use it, then you can make progress. And so at the end of the, the scientific session, Minsky gets up and here's what he said: shame on you. You have given up the goal of general intelligence. You're just working on applications. That's not intelligence. And I felt really embarrassed because a lot of these were students, former students and close colleagues. And this is like being present when a father is braiding his son. You know, for not living up to us.

TERRY: 37:42 Later after the banquet, after the banquet, I asked a question to Minsky, I said, look, there are people in the neural network field that think that you're the devil because of your book on perceptrons, right? That you, you said in your view that no one will ever come up with a learning algorithm. Well, we've done that now. Are you the devil? And it was, it was like I just pressed this button and he just launched into this huge, and he's very smart guy. He's is, he's a good mathematician. And he's saying, you guys don't really understand what you're talking about. We're dealing here with a mathematical problem that is so difficult and you know, you see this is never going to scale and here's really what’s going to make progress. And, and I said, Dr. Minsky, I asked you a yes or no question: are you or are you not the devil?

TERRY: 38:26 And sputters and said, yes, I am the devil.

CRAIG: 38:30 That's a great story. Yeah.

INTERLUDE: 38:30 CHORD

CRAIG: I want to ask just on the analogies of the algorithms in the brain. Everyone says, and I've read over and over and I had a conversation with Rich Sutton, that temporal difference learning is recognized as the algorithm that the brain is using in reinforcement learning.

TERRY: And my lab was the place where that came together. Oh, is that right? Yes. So Richard, Richard engineered reinforcement learning, control theory. I had two postdocs in my lab, Peter Dayan and Read Montague. And we came up with the hypothesis that the dopamine neurons in the basal ganglia were computing what's called reward prediction error. That's the key thing to temporal difference. And that has subsequently been tested in monkeys and it gave rise to a huge, you know, with humans with functional imaging, but huge, you know, research projects.

TERRY: 39:25 And now it actually created a field called neuro economics, which is how humans make decisions based on the reward prediction error coming out of the dopamine neurons. So, so that was, he was absolutely right. This is really had a, again, feedback from, from AI directly to neuroscience. And this was back in the 90s when we did that. Yeah, so bio inspired. I mean this is what's happening now over and over again. This is as, as we understand a little bit more about what works in machine learning. We can we take that and we go back to the brain and look for it, look for something, not the details necessarily like you know the back prop. We look for the principles. The principle is that you have to have information about error somewhere in your system and you have to get it to the right place at the right time.

TERRY: 40:13 That's the principle, now, that we're working with in neuroscience.

CRAIG: I'll ask probably an ignorant question, but are these algorithms restricted to certain classes of neurons or do algorithms in the brain function across groups of neurons and could those same groups of neurons be executing other algorithms? Or is there one, you know, master algorithm for the brain that is flexible enough, that is doing everything?

TERRY: I can assure you that's not the case. There is no master algorithm. We know an enormous amount about plasticity. Plasticity is the change in the strength of connections between the neuron at synapses. It's the change in the excitability of neurons. It's the change in threshold. So all of those variables in neurons are under constant shifts in order to maximize information flow, in order to be able to keep track of information, content coming in that you want to hold onto.

TERRY: 41:16 And so we've already, I would say we know at least 20 algorithms for synaptic plasticity that are used in different parts of the brain for different purposes. So the, the answer is, short answer is that nature has taken advantage of many, many mechanisms. And I'll give one example just to bring home the point. We've been talking about deep learning. Well that's a model for the cerebral cortex, which is the top of the brain. Right under that is the basal ganglia where I was telling you about the dopamine neurons, but that's where they live and that's where temporal difference learning works. So there, here's two different learning algorithms and they're both in the brain. There's a deep learning algorithm of some sort in the cortex and then there is a temporal difference learning in the basal ganglia and the basal ganglia talks back to the cortex. It is a big loop there and we know the basal ganglia is there for being able to learn sequences of actions that take you to a goal or reward and that could be far in the future. And that's what temporal difference does for you. And you put these two together and what do you get? AlphaGo.

TERRY: 42:14 AlphaGo depended on having a really rich representation of the board - that's deep learning - at the same time that it was making decisions about what moves to make - and that's temporal difference. And it's the talk between those across, talk back and forth that produces magic. By the way, I mean this is another sidelight, which is Noam Chomsky - I brought, brought him up earlier in the case of linguistics. He speculated again like Minsky and Papert, he speculated that there would be never any way that a learning system could learn language. It was just too complicated. It was a New York Review of Books essay in ‘77 and if you analyze the logic, okay, of that, here's the logic of both Minsky, Papert and Chomsky: I'm the smartest person in this field and I don't think it'll ever happen. It's inconceivable to me that this would work.

TERRY: 43:02 Therefore it's impossible. It's called argument from ignorance. And they were both wrong. I mean, you know, and so if you're a young, kids getting started and if some expert tells you that's not possible, don't believe them. Right. The chances are that just because they know all the ways it won't work doesn't mean that they know a way that it can work and that will require somebody new and fresh to come into the field. Like we were back in the 80s. We were very, in some ways, taking a big risk because the all the experts telling us that this is a dead end. Forget it. We couldn't get, you know, you couldn't get funding, you couldn't get resources. They were the ones that really had all of the power in AI. And that's all changed, hasn't it?

INTERLUDE: 43:42 MUSIC

CRAIG: 43:42 Yeah. Yeah, I do want to talk quickly about temporal dynamics of learning centers and the science of learning and when can we expect that to be productized for students in the market, whatever advances are being made.

TERRY: 43:59 Yeah. Temporal difference of learning center was one of six science of learning centers funded by the NSF and that was 10 years ago and it was a $30 million 10 year project that we had, which involved 12 institutions and 50 investigators and 50 fellows, students, who were working together collaboratively and had a wide range of projects. Our center focused on machine learning and neuroscience, putting those two together. And I'll just give you a couple of examples of what we did. So there was a robot that Javier Movellan put together called Rubi that he brought into a classroom, preschool, 18 month old toddlers. And the purpose of the project was to try to see if you can get the toddlers to interact with the robot. And the robot wasn't very sophisticated, just sat there, but it had very expressive eyes that moved around and it had hands that could pick things up and had a Teletubby.

TERRY: 44:54 So that it can have things that the kids can press. And it was able to play music. So the idea was to interact with little kids. So here's what happened. They put this robot in so that know the kids are running around, you know, they have very short attention spans and so they run over to the robot, what is this? And so the boys run to the robot and grab the arm and pull it off because they're testing things. And so you go back to the shop and you've got to repair it. And rather than put in an industrial strength arm, which they could have, but that would be very dangerous, what they did was they put a pressure sensor. And so when that arm was yanked, the next time Ruby would cry; boys would back off, girls would hug it.

CRAIG: That's so interesting.

TERRY: 45:37 So this is, this is social engineering you're talking about. In other words, what is it that humans respond to? How do you get humans to behave? How to interact with them to get their attention and hold their attention? That's because if you have a teaching robot, that's what you've got to do. You've got to be very interactive. And so they discovered many things like common - this is also very well known among people who study child development, which is when a mother and a child are together, they have a common attention. When the mother points to an object, the baby will look at it, right? Most species don't do that. They don't have this common attention. And so Rubi was programmed to have a common attention. So if a little kid pointed to the clock, Ruby would look at the clock and the kids love this.

TERRY: 46:16 They could do this for hours because they had some control over this creature. And then you start incorporating learning like you know, new language, new words and language learning and so forth. And they also, this was for us a great experiment. We've learned so much from it and we were afraid that the teacher was going to get threatened. Why? I mean you hear about AIs going to get, you know, jobs, are going to lose jobs. Our teachers are going to lose jobs. No. The teacher loved it. Why? Because for the teacher it helped her with crowd control. One of the biggest problems you have when you're teaching is just keeping everybody in line. And Rubi was a way of doing that, helping her as an assistant. It was assisting her be a better teacher. And that's been the experience we've had all along. Whenever we do something that looks like it's gonna threaten jobs or not, it basically helps people do their job better.

TERRY: 47:02 Okay. Now the second project, this is one that I undertook with Barbara Oakley who's an engineer at Oakland University in Michigan. We put together a MOOC, massive open online course, which became widely popular, Learning How to Learn. And what was our goal? Our goal was to try to help students become better learners, taking advantage of what we know about how the brain learns. And so Barbara would give a lecture on some practical advice about how to solve a problem, like an exam, anxiety, you know, mental block, procrastination. And then I would give what's going on in your brain and why this is happening, why this works, why this is going to work and you know, and so I'm giving them brain lessons, at the same time, Barbara is giving them practical lessos and we use green screen so that we could have things in the background, pictures of neurons, things flying around, just like the weatherman.

TERRY: 47:54 We use humor to try to get people's emotions involved and not just, you know, have a dry talking head. In any case it became the most popular MOOC. Now over 3 million people have taken it. I get fan mail every day, people in 200 countries, ages, you know, 10 to 90 and it really is having an impact. And this is I think where we are headed, which is that we can use the assistants, right? Like for example, Alexa, you can imagine Alexa being programmed so that it interacts and help educate, you know, individuals - have a model of that child, not just all one way. Although Coursera is a really great platform because it allows you to have a forum where students can ask questions and we have TAs that go and help. And there were meetups in cities at Starbucks meet there on Sunday at noon. And so there's a lot of machinery here to help, you know, get the students interacting with each other in education. It's been well established that by far the best education is when you have one on one with an adult who is a really good educator and a child and the adult understands the child very well over many years and can help the child get through mental blocks and so forth. But that's very expensive, very labor intensive.

INTERLUDE: 49:07 CHORD

CRAIG: 49:08 That's right. But, but presumably as natural language processing improves and all of these other facets - I've been talking to companies that are already putting in the market, AI tutors that track your performance and kind of understand the patterns of your learning and when to intervene.

TERRY: 49:26 So they're actually, there have been programs that have provided that now for, you know, 2030 years. So that's not new. What's new is the interface. And if you actually look at what deep learning has been successful at, it's been the machine-human interface, speech, vision, language, but that's what humans communicate with each other through, right through these channels. And that's where we've made the most progress. And that can be put into use now immediately in order to be able to bring this to the student, instead of sitting in front of a computer and pressing the buttons, now the student can talk to Alexa or the tutor or the system or whatever. And now that can be a dialogue. It could, you know, it's going to be much more efficient. It'll be much more natural and it'll make it more fun to have someone that you can talk to, right.

TERRY: 50:14 Rather than a computer. Do you know anyone that's doing that? Oh yeah. Oh yeah. Amazon actually has a competition where you know, they want people to write programs for Alexa. And so there's programs out there now for Alexa that are, are being built. Education, I think is going to be the killer app for deep learning and the, and the reason is it's so important. It's such a problem. You know, we have, we have terrible problems in the U S where our educational systems have failed us K through 12 and you know, this has long-term consequences that is, it's going to go on for generations. So this is really, if we have a new technology now that can be delivered. My MOOC is an example of the first wave of that, right. This is one way, but eventually it'll be two way. It'll be one on one and, and that's gonna make education so much richer and so much better for everybody, you know. And it's not going to get rid of the teachers.

TERRY: 51:09 Right. Cause you still need humans in the loop overseeing things and making decisions about blocks. You know, in terms of when does a student go onto the next block and so forth. Very complex decisions that have to be made. Our center TDLC, Temporal Dynamics of Learning Center …

CRAIG: It's still srunning?

TERRY: No it, the 10 year grant finished in 2018. What we are doing now is going international. So we've organized meetings with other countries and they were, by the way, we inspired science of learning centers in Australia and Brazil and and you know, many countries now are following the lead that we took 10 years ago. And now we're getting some of the big foundations interested like you know, Chan-Zuckerberg and Gates Foundation. I mean these are powerhouses, multimillion dollar powerhouses and I think that we're reaching that point now where with the right international cooperation and funding from the major foundations who are in this business, right, they’re in this sphere of trying to help education, I think this could be a real worldwide effort to create a much better learning environment that is based on the brain of how learning really works and also delivers teaching through the most powerful channel we have now for communicating, the internet.

TERRY: 52:25 All the pieces are falling into place.

CRAIG: Before that penetrates into classrooms, do you think it's five years, 10 years?

TERRY: Ah, here's what I learned. First of all, through this center we thought that we were going to be bringing science to the classroom. No, the classroom brought data to us. We learned an enormous amount, but delivering was incredibly difficult because of all the barriers. There are gatekeepers at every step. There are 12,000 school districts in the U S and if you needed to take some software into a classroom, you're going to knock on 12,000 doors. And then if you get past those doors, you have a problem with unions. You know, do teachers want to learn a completely new curriculum that you'd need to actually deliver this stuff? And so it was clear that, you know, we're academics, we don't know how to do this. So you'd have to start a business and so it would just be, and they'll take decades, many decades, right.

TERRY: 53:21 Because our educational system is so frozen and you know, and you know it's, it's huge. Multi trillion dollar industry. So this Learning How to Learn, Coursera taught us something. It taught us that the way that you get into people's homes is to bypass all the gatekeepers and go directly through the internet. Anybody with an internet connection, any place in the world now has access. It's free, right? The lesson is that you don't try to reform the system. That's not going to happen. But using the tools now that are available with machine learning and the internet, you can jump over that. And parents, of course, they want the best for their kids, and that's happening more and more. The parents are figuring out that they can get much better lessons for their kids through the internet than they're getting at school and and so that's going to happen.

TERRY: 54:10 That's going to be accelerated. And now if you have a, a personal teaching assistant, right, that's coming out of 10 million of these smart speakers that people have in their homes, they already have the technology in their homes. You’ve just got to know how to use it. You have to program it, you've got to get deep learning in the loop and that and that. So that's where I think things are headed. And how long will that take? Well, you know, in a sense the hardest part has already been done, which is building up the infrastructure. The infrastructure is there and now it's going to be companies going in with software and building up better and better interfaces. You know, it's not easy. I mean, you know, we're, we're dealing with very complex problems that kids have when they're learning. So, you know, you have to really understand something about, you know, being a really good educator at the same time that you're being a good engineer.

TERRY: 54:54 But that will happen. It's already happening.

INTERLUDE: 54:55 MUSIC

TERRY: 54:56 But one of the big themes this year is things like bias and fairness and ethics and so forth. That's become a very major issue with AI, as you probably know. How do you get rid of bias, your data? I mean, this is a technical problem that can be solved, but here's the problem. It turns out it's very difficult to, for example, fairness. The word fair means different things in different cultures. And so there's dozens of definitions for what's fair. And so here's this poor engineer is trying to make his algorithm fair and you know, who do you listen to? What is your, what are you trying to accomplish here? And you know, and it raises these issues that before were just people debating. But now it's real. Like, you know, we have this program and it's giving advice about hiring and medical problems and you know, all, all of these things which are very important for how human beings interact with each other.

TERRY: 55:51 And now we have, we have to be explicit, we have to decide how to, how to make it fair and you tell me what I should put into my loss function. You know, if you're a company, you're trying to maximize profit, but then you also want to have fairness. You have to decide what's the balance between the two and what do you mean by fairness and what is, what is bias? And you know, there are programs now that are being used for screening applicants, job applicants and people are getting very upset and say, what do you mean AI is making these decisions and it's all biased and so forth. And I agree that the program is biased, but do you think that the human who is making decision, are they biased? You know, we have this program, we have control over it, we can actually go in and we can adjust it so it's not biased.

TERRY: 56:32 How easy do you think it would be to go into the brain of the human and change them? Right. In other words, in some sense the, the biases that were in the programs were reflecting the biases of the humans and the differences that we can actually, we have access to the machine learning in a way that we don't with humans. And so therefore if humans are not even aware of their bias, this is, you know, this is part of the way the brain works. So this is where it's, what's going to happen over the next decade is there's going to be this learning process where people are beginning to learn how to be explicit about what are the goals and who's making decisions about that. How do you balance between all the different goals you might have. And that, and that's great because we needed this debate. You know, people have been debating this for a long time, but now we can actually do something about it.

INTERLUDE: 57:16 MUSIC

CRAIG: 57:17 Where is NeurIPS going? There was a joke I think in Montreal that if it keeps on increasing at the rate it's been increasing in however many years everyone in the world will be going to NeurIPS. So what is the plan?

TERRY: 57:32 Well, we're reaching a point now where the meeting has grown and grown exponentially. It's like 40, 50% per year for the last five years. And you're right, that's the joke. We wanted to maintain the high quality. We are the premier machine learning AI conference now, but we can't continue to grow like this and we don't know how much longer. You know right now there's a great demand. A lot of the reason people come to the meeting and sponsors come to the meeting is that they're looking for jobs and sponsors are trying to recruit. But you know, having one monolithic meeting, which worked for us for a long time is probably no longer practical. And so one of the things the board is now discussing is having additional meetings during the year. Smaller meetings spread out, distributed over many locations and countries on specific themes. And the idea is that by having a bunch of the satellite meetings, we will then take the burden off the main meeting.

TERRY: 58:26 All of our lectures are streaming. And by the way, we have now four tracks, right? We went from one to two. Now we are four. All that is streaming through the internet. Anybody now can in a sense watch the lectures, but that's not as much fun as being here. So meetups are places where people can come and congregate. Maybe a movie theater, you know, Google has a big auditorium and the, they have a lot of machine learning people and so what you do is you bring all together, they listen to the lectures, you have some food for them to interact with each other, coffee and so forth. And so there are 40 of these meetups, right? And they are self organized, you know, all over the world and concurrent delivery in real time watching these lectures as they're streaming across. Again, taking advantage of modern technology to be able to distribute the information to draw in more people. It's also better for carbon. You don't have as many pounds of carbon dioxide going into the air. Right?

CRAIG: 59:22 That's it for this week's podcast. I want to thank Terry for his time. I encourage all of you to read his book, Deep Learning Revolution. If you want to learn more about what we talked about today, you can find a transcript of this episode on our website, EYE on AI, that's E-Y-E hyphen O-N dot AI. We love to hear from listeners, so feel free to contact us with comments or suggestions.

The singularity may not be near, but AI is about to change your world, so pay attention.