Welcome to MIT'S computer science and artificial intelligence labs Alliance's podcast series. My name is Steve Lewis. I'm the Assistant Director of Global Strategic alliances for CSAIL at MIT. In this podcast series I will interview principal researchers at CSAIL, to discover what they're working on, and how it will impact society.

Samuel Madden, is a distinguished professor of computing at MIT Schwarzman College of Computing. His research is in the area of database systems, focusing on database analytics and query processing from clouds, to sensors, to modern high performance server architectures. He co-directs the data systems for AI Lab Initiative in the data systems group at MIT.

Madden was named one of MIT Technology reviews 35 innovators, under 35 in 2005. And he received an NSF Career ward in 2004 in a Sloan Foundation Fellowship award in 2007.

Sam has also received several best paper awards, including a test of time award in 2019, for his work on V-Track System incenses. He was the co-founder of Vertica systems acquired by HP in 2012 and a founder of Cambridge mobile telematics, a leading vendor of smartphone based technologies for making roads safer, by making drivers better. Sam, thanks for your time today.

Let's start off by telling our listeners the focus area of your research and maybe some of your bold aspirations.

Sure Steve, it's great to be here. So my research is in the area of software systems for data management. So we build large scale pieces of software that help people access their data, faster or more easily or to get access to new types of data.

So that's everything from next generation versions of systems like relational databases like Oracle that people may have heard of, to far out new systems that lets you do things like write the equivalent of an SQL query over a large archive of video or that maybe allow you to get access to a really large archive of high rate sensor data. For example and we build these kind of high level interfaces that let people get to their data efficiently

I see, we'll talk a little bit about databases, later on in the podcast because you've done some interesting work there. Can you give us your opinion about how far can technology be pushed into other application domains? First of all, I think there's a lot of hype around AI as you're alluding to and, we're in my research anyway certainly have spent some time looking at how AI or really when we say AI we mean machine learning technology, and I would even say go a step further and say, we really think of this as incorporating components like machine learning that can make predictions about things into software systems.

And so, there's been a lot of interest in these sort of machine learning components, using machine learning components as a part of a software system because, software systems are making decisions about things all the time. So for example, your database system or your operating system, might have some queries that it has to execute, a pool of queries that it has to execute and has to allocate those queries to some hardware that it has available to, and it wants to allocate those queries in the most efficient way to execute them as quickly as possible.

That's a classical scheduling problem and there are lots of classical solutions to this. But it turns out that if you look at the machine learning literature , there are these techniques based on things like reinforcement learning, the kinds of things that get used when you're trying to train an AI to play a video game for example, where it has a set of decisions that it has to make and those decisions yield some kind of outcome or reward or you can think of scheduling just like that to. The computer playing a game or trying to schedule these little things that it has to do as quickly as possible, in order to get the reward which is to use the fewest resources possible in order to do this.

So, that's a really simple way in which you can take some off the shelf machine learning component and repurpose it to help solve a hard systems problem, even though, when you go look at the machine learning literature, they probably weren't thinking about scheduling problems for example, as the problem they're trying to solve.

And I think from my perspective and I think for a lot of us it's easy. What's cool about this way of thinking about it is that not only do you solve the system's problem, but often when you get into the details of it, you realize that there are things about the way the original machine learning problem was formulated that aren't quite right, like playing video games is like scheduling tasks, but it's not exactly like scheduling tasks. And so you actually discover ways in which you can go back and modify the underlying algorithms as well.

So, let's talk a little bit about databases. Databases have been around for a long time. Can you talk about some of the challenges of building software to process and manage data. You mentioned about a scheduling problem and I assume that's part of it.

Yeah, well so.

Sure, I think yes, it's true on one hand, the product that many of us think of as a relational database system has been out there for a long time and with the 1980s, people figured out this sort of abstraction for a database system and actually for a long time in the 90s and 2,000 I don't think there was that much that was happening. But what we've seen in recent years, partly because of the real increase in how people use computers, the number of applications for computers and the amount of data that computers produce. A real interest in a diversity of different systems for working with different types of data.

And that's what excites me, is not just about a relational database system is like a system that stores tables of facts about employees or something. It's not just about that, now it's about video and sensor data and audio and all this other stuff that has to be stored and processed and accessed, and that's really inspired me through my whole research and there's just a ton of new problems there when you start thinking about a new data type.

I think the other set of new problems often are, when some new way of computing arises. So for example, 10, 15 years ago, cloud computing became the next big thing and there as a result of cloud computing has been a decade or more of research on how you take this classic extraction of a database system and you make it work in this cloud computing infrastructure. And so a lot of the really hot cool startup companies these days like you probably have heard of Snowflake.

Snowflake is an example of a company that essentially, took this research that was having not my research but research in the community that I work in over the last decade, on how to build database systems for the cloud and they built a really compelling product offering that's super competitive against the existing players in the space, because they had this purpose built system for the cloud and now everybody's moving all of their computing infrastructure to the cloud.

So, one of your more recent research projects is called SAJBD. Can you tell our listeners a little bit more about that project?

Yeah, so SAJBD, coming back to this theme of integrating machine learning, AI technologies into database systems, that the idea and SAJDB that we set out to solve is really to build from the ground up, a new database system that used machine learning and all of its components, from the storage system to the scheduling system I talked about to database systems have these complicated systems that are responsible for translating high level SQL query like select average salary of employees into an execution plan and that those things are called query optimizers.

So revisiting each of these components of a database system, where we think about how to integrate machine learning into that. And this has been really, I would say this project has been pushed by my colleague Tim Carrusca, he's done a bunch of amazing work in this space and I'm along for the ride. But it's been very fun to imagine this so, I gave the example of scheduling, that's one area where we've done some work, but another area that Tim has worked on a lot is in this area that we call LERN storage systems.

And so the observation here is that, classic old data structures for storing data, don't assume anything about or typically don't assume anything about the structure or layout of the data. So a big tree for example, is a data structure that a database system uses to store data. So that it can look up a particular value efficiently. So you can ask a Battery what's Sam's salary? And it can instead of looking at all the records to find Sam's salary, can directly offset to the record that is in Sam salary right.

But this bitrate structure doesn't assume anything about the way that the data is internally organized or stored, and the sort of observation that we and Really Tim calls this. He calls it instance optimized but the sort of observation that we had is that, if you know something about the way that the data is distributed, if you use machine learning to learn something about the way the data is distributed. You can exploit that knowledge to make the data structures more compact or more efficient.

So for example, if the data is mostly in order, or mostly sorted, or the values that you're storing occupy a very dense, short range, you might instead of using a bitrate, you might be able to represent the data using some other data structure like an array. And you could build a system that could dynamically make a decision about a battery versus an array. But you can actually do even, if it turns out there are a bunch of techniques you can do that are better than this that use, kind of simple machine learning but in clever ways to build these really efficient new data structures.

So it's been really fun. We've looked at all these different components of the data processing stack and been able to achieve state of the art results in a number of different sub problems in the database field, so it's been cool.

We had Tim on our podcast last week, and we were talking about his work in those optimized systems, and well it's very interesting that there's still more work to be done in databases. I think was the net of it is they could be faster, more optimized, more intelligent.

So, let's switch subjects a little bit, and talk about a subject that's near and dear to my heart video. You've recently working on some systems for processing video, can you talk a little bit about your work in the space?

Yeah, so I think this is a really good example of a place where machine learning, AI meets systems, and you can build systems that have very new capability and interesting new capabilities. So we've been inspired by, if you go out and you look at what machine learning can do for example, object recognition or scene recognition, you can give a machine learning algorithm a photograph, and it can identify all kinds of interesting things in it. So we said OK, this is a pretty cool new capability. What if we tried to build a database system, that used this capability as its core thing to answer questions.

So, instead of find the average salary of my employees in my database table, what if I wanted to, find the count the number of red cards, in some archive of video that I have right? And so you might say, well, OK that sounds pretty easy. I've got this off the shelf machine learning algorithm, I'll just run it on every frame of this video, and find all the red cards and I'll count them up. And that would be a solution to this but, you know, when you start actually trying to build systems that work this way, what you find is that, these machine learning algorithms, they're they weren't typically built with real high performance in mind.

So imagine, some of the people we've been working with are cities for example, who have hundreds or thousands of traffic cams, that are at intersections, and they want to do these traffic planning exercises, where they understand how intersections are being used or how many cars are at intersections at different times of day, the things that they would have to use manual counting now, they could replace with video. But they need to be able to answer these high level questions, about what's happening in these scenes.

And if they try to just naively apply these machine learning algorithms to every frame of the video, it turns out is just crazy, prohibitively expensive because it might take, you might be able to do this at like one third of real time, or three times real time on a GPU. But that means, if I've got 10,000 cameras, I need like 3,000 GPUs to keep up with a continuous video feed. Right, It's a crazy expense, nobody is going to do this. So we said, let's try to build software systems that don't require us to run on every frame of video, they're much more efficient by that than this.

And we developed all these techniques to do things like subsample the video, look at fewer frames, focus on the portions of frames that tend to have interesting stuff happen in them, be able to skip over periods of time, when we can guess that nothing is happening, or be able to exclude certain kinds of detections from needing to be processed. If I'm looking for red cards well maybe I don't, I look for red things first, and then I figure out whether there's a card there or optimizations like this that exploit the relative difficulty of different types of computations and run the computations in the right order to be able to get to the answer as efficiently as possible.

So we've built a couple of cool systems. My student Fabian Bustani, has this really cool paper that appeared last year in segment called Meurice MRMRS, that's all about how do you do this kind of video processing really efficiently.

So, it's been super fun I've got to learn a lot about computer vision technology and also build, what I think is a cool new system and we've got some people are starting to use it. Which is always fun to have people using your software so.

Yeah, that is great and I can definitely see how, if you were average 30 frames a second, and maybe high resolution video, how that would be so expensive computationally to try to do any real time analytics on.

You are also the co-founder and chief scientist for Cambridge Mobile Telematics, which is a startup company. Can you tell us more about the company and your work there?

Yeah, so this is a company I actually grew out of some research that I did and see scale with my colleague Hari Balakrishnan. We had a project, many years ago now called cartwheel that's car telecommunications, and where we were putting sensors on vehicles and using them to measure things of the world and we, spun out this company Cambridge mobile telematics. And at the time, we weren't exactly sure what the application was going to be, but we eventually hit on road safety and safe driving as a really critical application area.

So, the company is basically about making our roads safer by making people better drivers. We do that using a combination of smartphone apps and some embedded hardware that we put into vehicles, they look like little *I actually have one here,* little tall transponder, you can see this in the video. I'm going to turn it to Steve on the podcast. You can't see this on the podcast, I'm talking to Steve.

Does that go in the OB, the on-board computer?

So it doesn't go in the OBD. It gets suffixed to where you're told transponder.

OK, anyway, we have this smartphone apps and sensors that measure how people drive, and then give people feedback to help them become better drivers and so, we've had a lot of success, particularly in personal lines insurance business so many of you have probably seen these ads on TV for safe driving programs, don't mess with my discount et cetera, et cetera. And so those are really powered many of those programs are powered by the technology that we build that at CMT, Cambridge Mobile Telematics.

It's pretty exciting because on one hand, there's a lot of underlying interesting technology that came out of CSAIL, how do you take this data from vehicles. It looks like, what we're looking at are things like the acceleration signal from your smartphone, which you have to recognize, when there's driving happening and pick out the part of the acceleration signal that, represents the acceleration and deceleration of the vehicle and then you can build a profile as to whether somebody is a safe driver or not based on how they accelerate or decelerate.

So, there's a lot of technology to make this work but then, it gets packaged up and these applications that really about. Now one hand maybe people are getting a discount, but it's also giving you all kinds of feedback about how you could be a better or safer driver.

And so we've got, millions of people in the US who are using this thing every day. And we can show that this is reducing the accident rate by significant amounts, like we see people who are actively engaged in these programs see 10, 20, 30% lower accident rates, than people who don't use these programs.

So, you're making just this tremendous difference in road safety, by packaging up and game of buying, driving a little bit to get people to want to be better drivers and incentivize them in some cases, to do that through things like insurance discounts.

Now, is this an app that any users could download of Play Store, or is this something that they have to do in conjunction with their insurance provider?

There is an app that you can get called*safest driver,* which we build or insurance. Many insurance companies now offer versions of these apps, but the safest driver app, people can check it out, download it, install it, it doesn't do anything separate measure your driver and give you some feedback and tell you how to be a better driver so, signing up for it doesn't mean no commitment, no data being given to insurance companies or anything.

So, you mentioned CSAIL, can you tell us about your research efforts and how CSAIL alliance partners have maybe helped you with your research efforts.

So, I love to work with industry for a couple of reasons and CSAIL alliances is the primary way that I connect with industry sponsors. I think the transactional part of it, of course, is that some of these industry programs, fund the research that we need to support the graduate students, and do all the cool research we've been talking about. And that's great, but I think from my point of view, the more valuable thing almost about industry is that, it gives you an eye for what are the problems that really matter. You know it's easy to sit-in our office, or in this case, it's in my office at home not sitting in CSAIL yet.

But hopefully we'll be back there soon. But sit-in your office at CSAIL and imagine what people might want, but when you can go talk to industrial sponsors about what's a problem, you really have, that there's no better way to motivate the research that you're doing. So it's been really great. In fact, the CMT itself was, we had this technology for measuring driving and we actually discovered the connection to safe driving and insurance applications by talking to industrial sponsors that we'd met through MIT CSAIL so.

How about that?

So what excites you most about the research that you're doing?

I just like building cool software systems. I think at some level most of the software systems I work on are motivated by some data processing or certain some data capture. But I'm a computer geek at heart, a software geek at heart and I just like building, cool software that doesn't exist that hasn't existed before or figuring out how to build a better mousetrap or a faster database system or whatever it is. So, I feel like it's just I've got the best job in the world because that's what I get to do every day.

That's awesome.

Yeah.

And what would you recommend to a young researcher just starting out in computer science to work on?

Well, I think the first thing I would say is, on the topic of AI and AI will eat the world. There's a lot of people who are coming in to computer science thinking that all computer science is AI and feel like I'm the poster child for, although I do work on AI a little bit, there's a whole lot more to computer science than just AI machine learning and being a little bit application focused, focusing on building software systems, that real people want.

There is a ton of really cool work to be done there still, and I think as a young person that's tempting to jump into the thing that everybody else is doing. But one thing that I found in my research is that, oftentimes the most successful research projects are finding the place where other people aren't, and pushing on that place, as opposed to trying to do the same thing that everybody else is doing.

And so, I would encourage people to think broadly, look outside of what everybody else is doing and, when academia works, that's the great thing about it, is that we can push on new things that people haven't thought of before and make progress.

That's good advice. Well Sam thank you very much for your time today. We appreciate you participating in the CSAIL Alliance's podcast.

If you're interested in learning more about the CSAIL Alliance Program, and the latest research on CSAIL. Please visit our website @cap.csail.mit.edu. And listen to our podcast series on Spotify, Apple Music or wherever you listen to your podcasts. Tune in next month for a brand new edition of the CSAIL Alliance's podcast and stay ahead of the curve.