AAAI/CCC Symposium on Ai for Social Good

Invited Talk 1- AI for Sustainability and Public Health
Speakers: Carla Gomes (Cornell University) & Henry Kautz (University of Rochester)

Fei: Let me start by thank you all for coming, for attending this AAAI/CCC Spring Symposium; AI for social good. It's a really great opportunity to see all the people here and who care about social good and trying join the effort for contributing to the area and also framing the directions in the future of AI for social good. I'm Fei Fang from Harvard University. I'm currently a post-op at Harvard and will join [inaudible 00:47] later this year as an assistant professor. This is Amulya. Do you want to introduce yourself?

Amulya: [inaudible 00:55] in the computer science department at USC. I work with professor Milind Tambe.

Fei: It's a good honor to coach this symposium. I would like to start by thank all the organizing community members. Everyone on the committee has spent a lot of effort trying to make this symposium well organized. Also, thank you all for coming. I would also like to thank our sponsor, CCC, who kindly provides free lunch during the symposium and also supported several students for their travel and also sponsored the streaming of the whole symposium. Because the whole symposium will be streamed online, although there's no speaker here. Try to use a microphone even during the question, answering session. We will move the microphones around to make sure that people voices can be recorded and streamed. Thank you all. Next Amulya would like to introduce a goal of our symposium.

Amulya: All right. What is the goal of AI Talk 17? In recent times, conversations about future, negative consequences of AI have drowned our discussions about the potential of AI to do good for society. How can it be used to solve complex, wicked problems that society faces today? That is why we've organized this symposium to focus on the good that AI can do for society right now. We will do that by showcasing a multitude of applications of AI for social good. For example, how can AI be used to help under presented communities? How can it be used in emerging markets? How can it be used in applications like health care, social welfare, urban planning, et cetera? When we talk about AI for social good there are many questions that come up. We want this symposium to be able to answer many of the questions. Some of the questions that we want to answer are the following. This field of AI for social good is a relatively new field. We are the ones who are going to define this field. How do you define the space of AI for social good completely?

What is the kind of research that falls under its scope? What is the kind of research that we want to exclude from its scope? How do you select good problems in this space? Another important question that comes up is this mainly important application area. Or do we encounter fundamental research challenges when we solve problems unassuming. Yet another question is, are there common characteristics, which link together problems, many of the problems that we face in AI for social good? For example, is it the case that in most of the problems that we face, the data that we get is noisy. Or high performance computing is usually infeasible in problems in AI for social good. Because of these challenges, it may be the case that different kinds of solutions are needed to tackle these problems. We want the symposium to be able to find answers to these and other questions that may show up during the symposium. Now Fei will go over the agenda.

Fei: Our symposium is a two and half day program. We will start with invited talks and the first one will be by Carla Gomes and Harry Kautz. The second one will be a joint invited talk by Eric Rice and Sharad Goel. After the invited talks, we will have five technical sessions. Each including 25 paper presentations and the talk will be only eight minutes to make it short and also intense. We will have two minutes of question and answering session after each talk. Today evening, we will have the opening reception. This is for the all the symposium here. Also, tomorrow evening we will have a planner session summarizing what we have discussed here and highlight our discussions. We will also have two poster sessions during the day. We will have free lunch and during which we can continue our discussion, which is really great and sponsored by CCC. The live stream of the event can be found at the website of the symposium. I want to emphasize here that this symposium is not a regular conference or a workshop that you have maybe attended before.

It is supposed to be very interactive. Although we are expecting a lot of discussions, and we're hoping that everyone can contribute to the discussion and trying to have a lively exchange of ideas to try to answer the questions that Amulya has mentioned earlier. Please make sure that we need your voice. We need your input to these questions and to trying to forming the answer of the questions and even proposing the right questions. Let's get started. This morning we will have two joint invited talks. During which we'll have a coffee break and after that we will have a poster session and followed by a lunch discussion. The technical sessions we will have five of them. Today afternoon, we will have the health care session and social welfare session. Tomorrow we will have urban planning, computational sustainability, as well as other topics in the symposium. Let's start with the first invited talk. It's my great honor to introduce Carla and Henry for this joint invited talk on AI for sustainability and public health.

Carla is a professor of computer science at Cornell. She's the director of the Cornell Institute for Computational Sustainability. She's a lead PI for the NSF expedition on computing on the topic of computational sustainability. Carla is a fellow of AAS and Triple AI. Henry is a director of the Georgian Institute for Data Science and professor in the department of computer science at the University of Rochester. He has served as department head at AT&T Bell Lab and also as a full professor at U dub. He's also a fellow of Triple AS and ACM and Triple AI. Thank you for coming for the invited talk and let's welcome them for the talk.

Amulya: While you're setting up, would it make sense for everybody to introduce themselves among other things so that we can through the room a little better. I'll pass around the microphone.

Speaker 3: Hi. My name is Sarah McCarthy and I'm a third year PhD student working with Milind Tambe at USC.

Speaker 4: My name is Benjamin Ford. I'm a fourth year PhD candidate at University of California working with Milind Tambe.

Speaker 5: I'm Bryan Wilder and I'm a second year PhD student at USC advised by Milind Tambe.

Speaker 6: I'm Eric Rice. I'm a professor in the school of social work at USC.

Speaker 7: Hi, I'm Sharad Goel. I'm here at Stanford, assistant professor.

Speaker 8: Hi, Milind Tambe, University of Southern California. I'm a professor in computer science.

Speaker 9: Karen J. Smith. I work the alliance of NGOs on crime prevention and criminal justice. I'm here to encourage engagement with the United Nations.

Speaker 10: Hi, I'm Mia Yung and I'm a PhD candidate at USC school social work.

Speaker 11: Hi. I'm Robin Petering. I'm a PhD candidate at the school of social work at USC as well.

Speaker 12: Hi. I'm Jennifer Weeks. I'm an undergraduate student at Adam State University.

Speaker 13: Hi. I'm Oliver Bender and I'm professor for business information systems.

Speaker 14: I'm Frank Dicknum. [inaudible 09:42] University in computer science. I don't think this is working.

Fei: It is. There's no feed up here.

Speaker 15: Virginia Dicknum. I'm a professor at Delft University of Technology in the Netherlands. I'm leading the Design for Values Institute there.

Speaker 16: Hello everybody. I'm Mary [Sackarafeeli 10:06]. I'm a research assistant at Imperial College London. I'm working there at the department of computing at the Recalescence Information Security Group.

Speaker 17: Hi. Eugene [Verabeetrick 10:18]. I'm assistant professor at Vanderbilt University. Professor of computer science and bio medical informatics.

Speaker 18: I'm Daniel [Pristi 10:27]. I'm a professor of computer science and engineering at Lehigh University. I'm also a member of the CRACCC but you don't have to thank me for lunch.

Speaker 19: How can I follow that? Hello. I'm Danielle [Rosho 10:41]. I'm [foreign language 10:41] with the center for social services engineering at University of Toronto.

Speaker 20: I'm [foreign language 10:47].

Speaker 21: Hi. I'm Evan Patterson. I'm a PhD student in statistics here at Stanford.

Speaker 22: Good morning everyone. My name is Jason Parham. I'm a fifth year PhD student at Rennselaer Polytech Institute.

Speaker 23: Hi I'm [foreign language 11:04] and I'm a post-op here at Stanford working on applying natural language processing to social applications.

Speaker 24: Hello. My name is Helen Wright. I'm a senior program associate with the computing community consortium, which is a committee of the computing research association. We do events like this to catalyze the computing research association, the computing community, and enable to pursuit of new research. Happy to be here.

Speaker 25: My name is Hao Chen. I'm currently a post audit at Trinity University. I'll be joining University of Nebraska-Lincoln as an assistant professor in full. Or, yeah, in the full.

Speaker 26: Hi. I'm Jennifer Sleeman. I'm a PhD candidate at the University of Maryland Baltimore county in the computer science department.

Speaker 27: Hey. I'm Peter Carcos. I'm a PhD student from [inaudible 11:58] of Singapore.

Speaker 29: Hi, I'm Andrew. I'm a PhD student in computer science from Vanderbilt University. I'm working with professor [Bellermanen 12:08] and professor Eugene [Rodajick 12:11]

Speaker 28: I'm George Sellman I'm a computer science professor at Adam State University.

Speaker 30: Hi, I'm [foreign language 12:21] I'm a third year PhD student at Vanderbilt advised by Eugene [Verabeetrick 12:26]

Speaker 31: I'm [foreign language 12:28] from India. I'm a PhD fourth year student in department of computer sciences [inaudible 12:33] thank you.

Speaker 32: Hi. My name is [foreign language 12:38]. I'm a PhD candidate in political science at UC Burkeley.

Speaker 33: Hi. I'm [foreign language 12:44] I'm from University of Oxford doing a PhD in machine learning.

Speaker 34: Hi. I'm [foreign language 12:50] from [inaudible 12:52] University. I'm an associate professor in design.

Speaker 35: Sure. Brad [Mallen 13:08]. I'm professor of computer science, biomedical informatics, and bio statistics from Vanderbilt University.

Speaker 36: Hi. My name is [foreign language 13:22]. I'm student at [inaudible 13:23].

Fei: Great. Thank you again for attending and let's welcome Carla for the invited talk.

Carla: Well, thank you so much. It's a great pleasure for me to be here and these initiatives are so important. It's a great pleasure for me to be here. These initiatives are really important. I want to thank Fei and Amulya for inviting me. I do want today to give you a bit of big picture. I will talk about computation sustainability and Henry Kautz will talk about AI for public health. I always start now thanks to NSF generosity, thanking NSF for supporting this area. We received an NSF expedition in 2008 that allowed us to nucleate the computation sustainability field. I will be talking about it. More recently, we received a new expedition that really allows us to pursue research in this area. It's sometimes challenging to actually get funds to you know. For example, poverty. That is really tough to find sponsors. Thanks. I really appreciate NSF funding us.

Of course, this expedition is a large research network. Milind is a big player in this research network. Fei and there are several universities involved in addition to Cornell, Golden Caltech, CMU, Georgia Tech, Howard, OSU, Princeton, Stanford, UMASS, UFC, and Vanderbilt. This is really the goal for us is to disseminate this kind of work. This workshop is just wonderful. Let me briefly tell you what computational assistant ability is about. You may think it's about sustaining computers. That's not really the goal of computational sustainability. It is a new field that has the goal of developing computational masses to address sustainability questions. In particular, for sustainable development. The notion of sustainable development was introduced in 87 by the United Commissions. It was quite a revolutionary notion at the time. Developing that meet the needs of the present without compromising the ability of future generations to meet their needs. Basically, they wanted to frame the question and basically they actually wanted to say, sustainability is not just about the environment.

I think often computer science thinks sustainability is just about the environment. Not at all. In fact, they stress that sustainable development encompasses environmental, economic, and societal needs to address those issues. In fact, the ultimate goal of sustainable development is human well being for current and future generations. Just for you to see that. Here are the sustainable development goals. You see that frankly, the environment comes at the end. The primary goal, they are all important. But there are many more important goals. Even the perspective of human well being. No poverty. Zero hunger, et cetera. Of course, for us to make sure that we have a sustainable future, we also need to take care of our natural resources. What I want to stress is how all these goals are interlinked. And indeed above all, sustainable development is about human well being. Our CompSustNet actually has a variety of projects along different themes.

Touching for example balancing socio economic needs. Conservation and biodiversity, but also renewable energy, accelerating the discovery of sustainable materials, smart cities, et cetera. We have a variety of projects in these different areas. In fact, over 20. What is interesting and I was quite impressed how beautiful you framed some of the questions in the context of computer science. These projects and when you go to NSF obviously, NSF wants broader impact. But when you apply for an expedition in computer science, you absolutely need to show that you are also advancing computer science. That is exactly the point that we wanted to emphasize. This project, for most people, may look like just a bunch of different projects. What is exciting about it these portfolio of projects is they allow us to really address cross cutting, core computational questions. That is exactly what I'm going to try to illustrate in this doc. In fact, this slide I present the projects from a perspective of ... sorry. Is there a watch? Sorry, a clock here? No. I have my cell phone to keep check of time.

This slide shows different projects from a perspective of sustainability themes. I'm going to now give you a different perspective; the core computational questions that we are addressing. Our network is looking at issues concerning for example, big data, machine learning, constraint optimization, dynamical models, and simulation. Also, different issues concerning multi agency, some citizen science, and crowd sourcing because indeed, we are looking at a high interdisciplinary projects because we need to factor in social, economic, environmental factors. That's why we need to look at different aspects. What is exciting about this area and now you are going to tell me what the word here that gives away that this was an NSF proposal. Can you guess what word here? Transformative. What I actually truly believe in this, that when we are tackling these projects, computer science can not really stay in their little area. We are actually forced to address challenge across areas. I'm going to give examples. Often we now have to really work at the intersection of optimization and machine learning. We are optimizing models but we need to get the data. These are really very interesting challenge.

Not just pushing the frontiers of computer science. Each area of computer science but really the intersection of different varies of computer science. Here is yet another perspective. This perspective I now turn things around. What you see is each line is a computational theme. The stations here are the different projects because that's exactly what we are trying to emphasize. We are trying to emphasize sentences across computational themes and advance AI techniques. This becomes quite an exciting opportunity for the students. In fact, in general I really encourage a student not to work on a single application. They work on some computational methodology and then they apply it to different problems. For example, I will talk about how some students are working materials discovery and elephant code detection. What do they have in common? I hope that you will understand better how this subway or these themes are put together. Obviously there's no way I can talk about all these projects. But what I decide to do today is I'm going to give some examples of projects. Actually I'm going to give a bit of a start perspective how ... some projects that got us into the field.

Again, addressing your questions. I'm going to tell a story of how by studying and coming up with the ideas for design, experiments for fertilizers, that led us to introducing very powerful new techniques for counting, searching simply, et cetera. Initially these techniques again, we were addressing issues concerning sustainable agriculture but now we have generalized and developed methodology that is way beyond fertilizer. I think this is a good example. I like telling this example to my students. The importance of think at a higher level and how these problems can really lead to new fundamental foundation techniques in AI. Then I'll talk about, if there's time, about problems concerning big data, citizen science, et cetera. Let me start with ... could you give me a sense because I'm a little ...

Amulya: It's been 11 minutes.

Carla: That's fine. Let me tell you this story that I think it illustrates how we can really by tackling these problems. One thing that I think is important is when we hear, when we have a problem, we can't really say let's make it complex. Let's make it all kinds of ... we actually need to solve the problem. Actually I see Henry there. I'm going to make an analogy with Bell Labs. Bell Labs revolutionized the world because they were solving a real problem. They were trying to solve a real problem communications, phones, et cetera. I think we can make a big difference by solving real problems in sustainability, sustainable development. Now I'm going wild here but I honestly believe that, that is important. As you know, fertilizers have revolutionized food production. Really. We changed dramatically productivity thanks to fertilizers. However, fertilizers are also a major source of greenhouse gas emissions because of the nitrogen. In fact, it's estimated that more 90 percent of the agricultural nitric oxide losses come from fertilizers. They have greater global warming potential than all of US aviation.

What happens in the US particularly, farmers use ten times more the amount of fertilizer they need. We need better fertilizer management tool. At Cornell, by the way, Cornell has ... in addition, you may not know this. In addition to being a state universe ... sorry. A private university, it is also a state university. We have the Land Grand Mission. There's this major project to try to recommend the type and amount of fertilizers to use. I'm not really going to talk bout this project. I'm going to talk of another project we worked on. We are still also involved in this but here the idea is we want to design experiments to study fertilizers. If you know about experiment design, actually a key structure is this combinatorial structure called a Latin square where for example, here we are testing different drugs. Advil, aspirin, et cetera. We want the subject to the same treatment. That's why we set up as a Latin square, which is a very simple structure where here we have a four by four. Basically you want use for caller so that you don't repeat the caller in column and in a row.

Just [inaudible 28:21] I actually have worked on this problem of nausea with my students. Latin squares have many applications. In scheduling like fiber routing, Sudoku is a Latin square. We had studied this thing at nausea and in fact, when Harold, my colleague who is a professor in crops and soil science study stock, he got all excited. He said, "Carla, you are going to solve my problem. My problem is that Latin squares are good but not good enough for designing agronomic experiments." We need to do better than that because we need to factor in the geometry of the field. Basically what we need is called spatially balanced Latin squares and basically that means that all pairs of treatment have the same total distance. For example here you see this has distance one. Despair distance one, distance four, et cetera. The total distance for the rows is 14. All pairs are going to have the same constraint. I had studied this problem of nausea. I said, that's easy. He said, "you just generate all Latin squares and we select the spatially balanced Latin squares." I said, not exactly like that but I thought it would be still easy.

Straight forward, given that I had been able to solve Latin squares up to 60, 100. It was not easy at all. We tried all kinds of strategies for those familiar with constraint, with programming, and stats, ESP, integer programming, et cetera. We could only solve up to only six. I have to say I love this problem because actually it kept my sanity. I had a major accident. I was in the hospital. This was, it kept me thinking about this problem. In fact, I had wrote several problems and then I could see looking at the solutions you can see patterns here. For example, a lot of the solutions for the six are symmetric. I thought, why don't we impose additional constraints. We will have spatially balanced Latin squares and in addition they will be symmetric but they are still spatially. By pushing this reasoning, we then for example, I remember conjecturing ... by the way, in case you think I don't know how to spell Sellman. This is actually Meinolf Sellman. Then we conjectured, well maybe if we combine two small ones et cetera. By doing this we are actually able to produce larger squares.

When it came time to write a paper, Meinolf, he's German. He's very good at getting things right when they open the submission for, which the Portuguese are the opposite. They wait for the deadline. He was writing a paper. He said, "I get this going about fertilizers." In that time I said, Meinolf maybe we should try to think about what we did here. Then we thought about it. Basically we came up with this notion of streamline reasoning. The idea is you have a search space and you impost streamliners, which in the end are additional constraints. That's what we did. We impost constraints such as symmetry, such as this particular duo thing, et cetera. To narrow the search space and help find solutions. Are you following this? This is what we call streamline reasoning. Here this is actually very important. For the students, always think at a more abstract level so that things will generalize. We did this. Actually we were very happy because we thought of different ways of streamlining. Then we were able to actually generate squares for over 35.

They were thinking using about 30 treatments. Now we could generate solutions of the sizes they wanted. As a computer science, okay. Now we want to think. In fact, when we wrote the paper and this is also a recommendation that I suggest to the students and that's my rule, we generalize and we have to find three domains. Let's think about are the domains where we could apply this? We did. They were still ... the other ones were type like magic squares. I don't even know the third domain we came up with something. We showed that by adding these constraints we could actually solve, find solutions, and for larger problems, et cetera. We were happy but because we had thought about this as a technique, now the next question was can we generalize this technique for a domain or better. A technique that now it's domain independent. That was our next question. The answer is, yes. We can actually do that. That's how we came up with the notion of XOR-Streamlining based on random parity constraints.

This really shows the power when you think in general terms, you can actually come up with ideas that can be applied across different domains. To give you a sense, imagine and I want to count what we mean by this. Imagine that I want to count how many people are in this room. I could do it. I could go brute force one by one. I could do all kinds of techniques like some kind of approximate branch and band, sample some areas and stipulate from there. I could also do something different. I could do a coin flipping strategy where I would give everybody a coin and everybody puts the hand up. You toss a coin, if you get heads, you keep the hand up. Otherwise, you bring the hand down. You do this until you have only one hand up. When you do this, now you can estimate there are two to the number of rounds, number of people here. This is a strategy. How do we do this in terms of a real problem? We apply this for set and the analogy here is this auditor is the search space. The seats are my possible assignments.

The occupied seats are the satisfying assignments. Bringing a hand down means I'm going to add a constraint, eliminating the set dose satisfying assignments. How do we now make this operational? We can use what is called a parity constraint, an XOR constraint. Basically this is satisfied when an odd number of variables is said to be true. What's interesting, we came up with this idea because this parity constraint were used by Valiant-Vazirani to show that unique set is hard. They used this in a negative context to show that unique Sudoku. Sudoku has a solution but it's still a hard problem. I'm not going to go into the details but we can actually now, based on this, we came up with an algorithm that tells me if I have a random and I throw k random XOR constraints. I check for satisfiability using a SAT solver. This was in the set context. I repeat this procedure, say four times. Then I can provide some guarantee of the number of solutions that my formula has.

What I want to emphasize, I don't want to go. I'll fly and I can go over the details. Just see from fertilizers what we came with strategies for counting, sampling. Stephen Erman who I think is now at Stanford, generalized our probability inference and actually [foreign language 38:03], my student also generalized this for stochastic optimization. Vardi's group has also further pushed this in different directions. They actually used this technique for verification in program synthesis and also reliability of power grids. This started by studying fertilizers. Think big and we can really go far. Here I show and I again, I'm happy to go over the details offline. We are actually using this technique for stochastic optimization. Basically we are designing corridors in Ecuador for Indian bears. Now we have stochastic optimization problem where basically we want to buy parcels to protect the land for the bears.

This is, I should have used the other version of PowerPoint. Basically I want to maximize this expectation. By protecting different parcels so I can buy or use [inaudible 39:23] or et cetera. This problem is very hard. This is actually a dynamical model in the sense that what I want to do is maximize the connectivity for the bears, which is in expectation. Now I have actually maximization of an expectation. It's dynamically in the sense that when I'm deciding what kinds of parcels I'm going to buy, the connectivity of a parcel depends heavily on whether or not I buy the neighbors. Correct? This is a challenge for standard sampling strategies.

Amulya: 20 minutes.

Carla: I have a few more minutes?

Amulya: You have ten more minutes.

Carla: Ten more, oh. Awesome. How are we solving this? Basically we need here to do the expectation and the partition function. It's very challenge. We are using this XOR-streamlining to approximate these expectation. We are actually using another trick that

I will talk later that is we actually find a way of folding. This expectation in the end it's a sum. Folding this problem into, flattening it so that we have a single optimization problem. I will not explain but I'll show you, mention the trick that was also ... we developed that approach based on another problem. This again, it's just to emphasize how a simple problem like the fertilizers led us to these techniques. Let me now give another example or talk about other projects. In particular, bird conservation. I have to say, our portfolio of problems is very much dependent on access to experts. As you know, you don't want a bunch of scientists pontificating about something they know nothing about.

Just like Milind who is teaming with all kinds of experts, we also go for problems for which we have access to experts. Cornell has the lab of O, the lab of sorry, of ornithology. Ornithology is such a long word that we just say lab of O. There's this program, a citizen science program called eBird. Are you familiar with eBird? I think you should join or at least your kids should join because this is fantastic. Actually, that's how James Wattson ... sorry. Let me not go into another story. Basically when it comes actually to sustainability issues, we need data. Nobody is going to pay for bird observations or if we were to use experts, which would be awesome. It would be too expensive. This is an example where we have to rely heavily on citizen science. But look at that. We have about over 300 thousand volunteer birders. They've submitted over 300 million observations corresponding to over 21 million hours of filed work. Three times more what was used, the person's hours for building the Empire State Building.

We use this data, combining the environmental data, and then we are developing spatial and temporal machine learning models. We can now predict how birds occur. The patterns are for current soft birds across spatial and temporal spaces at a very fine grain. This really gives a much better understanding of habitat requirements for the birds. Obviously there are many challenge in terms of this problem. We are using these development model for birds but also for plants, animals, and actually for materials. Again, this problem in the end we are developing just general techniques. We actually for lab of O produces the state of the bird in collaboration with many other agencies. The state of the bird reports. That is actually released by the secretary of interior. But because now we have such a good understanding of the movements, now this leads to fundamentally novel approaches to do bird conservation. As an example here, actually we did California. Now there's no drought.

You have had a tremendous drought. The nature conservancies has this amazing program where they do revers auctions. Using our models, they can predict exactly when the birds are flying over the Sacramento Valley. The farmers are submitting bids to the nature conservancy where they are willing for some money to keep the rice fields flooded while the birds are migrating. This is a fundamentally different way of doing conservation because typically they would buy land. Obviously they can not by the entire US. This is a way more viable strategy. They've actually added additional habitats for the birds. This is only possible because now we have this fine resolution models that give us a good understanding of the bird migrations. Again, to mention the fact that we are applying models across different areas. We have a program. The US, we have a ton of data.

Africa is very poorly sensed. One of the thing that they don't even have is good maps. In particular, we can infer maps from remote sensing. But we cannot really know about edible grass because remote sense, they tell us this is green but it doesn't mean that the camels can eat that. We're trying to get better information. We have also a crowd sourcing program where the herd of the [inaudible 43:44] submit this vegetation pictures and service. It has been very successful. I think Neal is presenting here. Stefano and his collaborators are leading this very interesting initiative where they combine satellite imagery and machine learning to predict poverty mapping, to build poverty maps. Let me talk about citizen science. It is awesome but there are many challenges associated with citizen science. One of the challenge is, the fact that because this is a volunteer program we can not be demanding. We let people submit the observations from where they are. We can not put constraints.

You see here is the map. You see a good coverage. But you see that mainly you can't even identify the big cities. Of course, the lab of O is there. This is a problem. What are we doing to address this problem? We are developing these games. If you know Pokemon Go, we came up with this game before Pokemon Go. I have to admit. It's the real version of Pokemon Go. The idea is we are trying to get people to go to under sampled areas. What we do is we assign points to the under sample areas, people see those points and they are going to collect the points. At the end there's a lottery where they get binoculars proportionate to the number of points they collected. This is actually what it's called, the principal agent framework. I have to say that we thought about this after my student went to a talk by Milind and he starts thinking because they've developed these kinds of models. I'll talk a little bit about that for the patrolling. I'm sure there will be several talks about it here.

That was really the inspiration. Can we use this model, this principal agent model to design this game? The idea is eBird, we are the principal and we are going to assign rewards or points that agents see their rewards. They are going to optimize their own utilities. The game is we want to put the points to induce a behavior that will lead to a more uniform distribution. Do you see that? This is what it's called, a bi level optimization problem. The principal is trying to maximize the uniform distribution of the observations. The agents have their own utility they are trying to maximize. We want to set up the points based on their behavior to induce this overall goal. Does it make sense? This is a bi level optimization. It is a two stage game. In principal, you would sink. It is a very hard computation problem. What's awesome about this is we are working with our colleagues and they get so excited about the fact that computer science actually are working with them.

That you come up with the idea and they say we were working on this. Next time we'll run a pilot. Wow, that's awesome. We did. Now we're running all these pilots. You see here we starting with Tompkin counties and Tompkin, our county and [Cortland 51:14] you see how in a year we were ... from one year to another we had a 20 percent shift of the observation. It's a little less concentrated. Now again, again, my rule is we need three domains. What have we done here? Of course, we developed this game but what was the key strategy that Yexiang actually come up with? Yexiang is really a smart student and he's really pushing this. Basically the problem here is we have a two level optimization problem. Because of that, this is quite hard computation. What his key insight was what if we embed the agent's optimization problem into the problem of the principal?

We flatted. Rather than have this two stage, we just ... we do this by approximating this. We are able to embed this into the principal problem. Again, can we further explore this? We actually use this idea in the stochastic optimization for the corridors. He actually had a paper at nip and we also show that we can use this same embedding technique for standard machine learning techniques. For example, it ran on several machine learning benchmarks that can be framed like this. Again, this is the idea of generalizing and to address this challenge we come up with new computation foundation. You had that question. Trust me. There are many opportunities here. Again, I already mentioned that this is mechanism design. In fact, we were heavily inspired by the work of Milind and Fei and all the students where they used the same the stackable games or the principal agent framework to place patrols to prevent poaching and illegal fishing. Computationally speaking, there are several analogies with this model.

I think I'm running out of time but I wanted to mention another example. I really believe in pushing this cross cutting computational themes. We are working on a very exciting domain that is quite different. Here, the goal is we want to develop [inaudible 54:12] techniques to analyze material data. Material science have developed these techniques where they can rapidly sensitize thousands of materials, rapidly characterize them with say x-ray diffraction patterns, thousands of materials. The bottleneck is then how then to analyze the data. In particular, just like for DNA, it's important to know the structure of DNA. They also need to know the structure of materials. A problem that we are working on is give an x-ray diffraction patterns, we want to infer the crystal structure of materials. This problem is very similar to this problem, which is the problem of monitoring elephant populations by analyzing recordings. What's the analogy? The analogy is both problems involve I have a signal.

Let's say an x-ray diffraction signal. From the x-ray, I want to decompose this signal into different components corresponding to different crystal structures. For here, I have the audio signal and I want to identify when there's an elephant call or a frog call or a plane going by or a gun shot. Both problems are separation problems where they involve the decomposition of signal into patterns. The crystal structures of are the type of sounds. If I tell you topic modeling, does it ring a bell? Are you familiar with topic modeling? Basically in topic modeling, we have a bunch of documents and we want to identify the key topics of the set of documents. Here, we can think of bases if you have a matrix factorization it's basic. Here, the key topis are the key crystal structures or the key sounds. Can you see at a high level the analogy? That's exactly my philosophy. My students working on their problem are actually working on this problem. The problem of identifying gun shots or whales or other problems that is for example, high prospector imaging where we are trying to identify disease in plants, which is a similar.

In the end, these are dimensionally reduction. What I love about this problem and this is where we advanced the state of the art, is that in general the machine learning techniques fail completely to address this problem. The reason is, when you do topic modeling or whatever. If you don't get it completely right, it doesn't matter. For us, it matters. For the scientists, they want the solutions to follow the physical loss. We have this really dimensional reduction with compact constraints. The solutions have to be physically meaningful. That makes it so exciting for computer science. In fact, there's a big initiative we actually ... Toyota, for example is now very much interested in this. This is becoming a big area. I think computer science can make a big role. As I said, we are working with Caltech on identifying new solar fuels. In fact, we just had a paper in material science journal that was actually the cover article because with this technique we discovered new solar absorbers. It's a very exciting area.

This is again, how we have the subway lines. For example, here I talked about this project and the elephant calls. We also have the size group with pheno typing. For those doing you can think of there's also the electricity usage desegregation, where from the electricity bill and electricity signal you want to decompose into did you spend most of the money on the dishwasher or actually probably the ride, et cetera. I hope you have the big picture of this. I also talked about different mechanism designs, problems that are on this line. I think I'm already going. Thank you for not really being very tough on me. I think computational sustainable, AI for social good, these are really exciting areas for computer science. We can have tremendous impact. What I try to emphasize, we can also have impact in terms of pushing the frontiers of computer science, addressing, challenge that, [inaudible 59:41] we can also advance computer science dramatically. Thank you.

Amulya: The connectivity, the problem that you're laying out in order to make sure that different areas remain connected in conservation. You see that there is some special tricks or special ideas you were going to return to that later on. You did not. Could you elaborate on that?

Carla: Yeah, that problem ... oh, thank you so much. The way we scaled up that problem, one was using this XOR streamlining. Basically for us to approximate the expectation, we actually ... you have a query, an [mipsova 01:00:38] by adding XOR constraints we can estimate that expectation. What we do is, the queries to the [mipsova 01:00:52] remember we want to maximize the expectation by choosing what action to, what parcel to buy. We have this maximization and then the operator that is the expectation. We solve this queries to a mipsova. Then we would have a maximization and then a maximization problem. Do you see? We want to flatten this. We use the same trick that we used in the heavy cashing where it was a two stage game but we fold. We embed the problem of the agents into the problem of the principal. Here, this expectation, we have queries to a mipsova to approximate it. Then we also have a maximization over this. We fold this problem into the other one. Again, we have a single [MIP 01:01:56] problem. Does it make sense at the high level? I'm happy to elaborate more. Thank you.

Henry: Thank you Carla. I'm actually not sure I should thank you since anyone who follows you is at an extreme disadvantage. Carla gave a whole tool kit of ways in which AI can be used for social good. I'm going to be talking about maybe one wrench or one screwdriver that I've been applying to a lot of problems. Particularly we're looking at mining social media to create tools for public health. We have lots of very noisy data.

We can look at social media as essentially a population scale distributed sensor network that is extremely moving around randomly. Very fine grained, timely data. Making more or less random observations about themselves and their environment. Some of the questions I'd like to ask. Who's likely to contract disease? What lifestyle factors influence health? What are sources of disease?

When we began this work is when people were first starting to realize how heavily our social media is being mined for marketing purposes and to sell us things. We thought, this technology is being developed. Can we turn it around and use it for a good purpose? We started with a project we called Twitterflu, tracking influenza. You're looking at different tweets and you want to identify, which are the sick tweets of self reports of illness. From that, use that data as you would in Google flu trends or in traditional reporting to estimate the levels of illness. The initial work in this area, simply was based on finding key words about tweets. The problem is that we an never come up in advance with all the ways people might describe being sick. If you're sick of homework, of course you're not really sick. If you're under the weather, you are.

We developed a pipeline. This is essentially the pipeline we used over and over again where we used some simple heuristics to gather a large set of potentially relevant tweets. Use mechanical Turk workers then to create our training data and use that then to train our machine learning algorithm. In particular, we used what we call a cascade SVM, a support vector machine. The features are going to be every word, every word bi gram, and tri gram. So triple of words that appear in a message. We want to learn a support vector machine, the thing that could separate out those very rare but significant tweets from the others. What makes this an interesting machine learning problem is the class imbalance. The tweets we care about may be one out of 30 thousand, one out of 100 thousand, one out of 200 thousand tweets. We begin gathering some tweets. We start with some labeled by human workers.

We train essentially two classifiers by adjusting the thresholds. One with high accuracy and low precision, one with low accuracy and high precision. We run this to label data using in our classifiers. Then we feed those back to the workers to correct the errors. Then we put it all together into our final corpus. When we get to the end of this process, we can come up with very high accuracy and high recall classifiers. Over 98 percent accurate of we're finding these very rare tweets. These are the kinds of features you see so sick is good evidence of sick. Sick of is a negative evidence. When we did this, Bieber fever was popular. We eliminated that. We saw that yes, we could essentially match the CDC results but we could provide it instantly on a daily basis. We wouldn't have to have weekly reports waiting for the CDC. Furthermore, we could do interesting things like actually start doing prediction. If we look at people's whose tweets were geo tagged, we could say if other people from your geographic are tweeting, by area I mean within 100 meters are tweeting about being sick and you're in that area within an hour time window, your probability of tweeting about being sick in the next couple of days is going from a very low baseline to up to 50 percent if you're in an area where that's about 60 people had tweeted about being sick.

These different lines are for different time windows. We can also look at your friends. If your Twitter friends are tweeting about being sick, that's not particularly important because most of our friends might not be with you. If your Twitter friends who are tweeting about sick have also tweeted from a location where you have tweeted from recently, then just a small number of sick Twitter friends gives you a high probability of predicting being sick. Could then combine this information with other sources of geo tagged information such as sources of air pollution to make sure the possible association between being in places that have polluted air and local micro pollutions. If you're living near the exit to the Lincoln Tunnel, there's very high carbon monoxide. Also, measuring things like how much time, how often did you tweet from a park or a gymnasium. You were exercising. You can start to then create regression models to look at, we said this, essentially you're predicting how often are you going to be tweeting about being sick.

These are the factors that predict that. It can explain a little over half of the variants of you tweeting about being sick based on these other factors about your life that we've measured. When you think of public health globally, of course the big challenge for the future is global contagions. Some people think that we ain't seen nothing yet. That there will be a dystopian possible future where because of growing inequality and pockets in the world where there's these terrible diseases can rise, combined with global transportation, those things can spread. We need to do everything we can to ... we want to improve health globally we also be aware of how can we track the disease very quickly. What we did is we ran the same system on airports and metropolitan areas around the world of a hundred airports and 75 metropolitan areas. What we were looking at is the number. We could use this to estimate travel into a city. The number of Twitter users who flew into x on a particular day because they were in a different city on one day and then the target city on the next.

You'd say, the airlines have those statistics, yes. But it's not easy to get those statistics. Airlines don't like to publish their numbers. If you're the CDC, you can get them eventually. We could also measure how many of those Twitter users had tweeted about being sick and we could also measure even if you have not tweeted about being sick, you were near with a hundred meters of people who had tweeted about being sick. Possibly you were a carrier. We added each of these factors to a model for predicting the change in flu rates in a city. You explain a larger and larger percentage from about 56 percent to 78 percent. This number is consistent. The 56 is consistent with studies that I've done where they look at data from the airlines of how you can use passenger volume to predict about half of the change in flu rates. Here we're getting all the way up to 78 percent by including the other parts of the signal. We had this nice screwdriver. What else could we apply it to in public health?

One idea is a food born illness. Can we detect places that are making people sick through their food? I guess sick in the short term. We could identify every Kentucky Fried Chicken in the world in terms of diabetes. How do we do it now is we have restaurant inspections and inspections of food distribution. The issue for public health officials is how to target these inspections more effectively. I think this is an interesting connection when you think about you're patrolling. You want to set out patrols to catch the drug smugglers and the criminals and so forth as in some of Milind's work or the poachers. Again, there's thousands of restaurants. We can only inspect so many of them each day. How do we deploy our resources? Of course, it'd be nice if we could rely upon just looking at social media posts where people say, go here. Inspect here. We have to do something a little more subtle.

Again, we trained our algorithm to find self reports of stomach ailments only. These of course, are much more rare than flu. They have to be talking about vomiting but not a stuffy nose essentially. Then we want to link these sick tweets to restaurants where the user tweeted from within the previous few days. Then use that information to target health inspections. Again, we learned a classifier. Stomach and stomach ache. Mention I'm getting. That's interesting. If somebody says, Sam I'm getting. That turned out to be an interesting ... who knows why. We never would've guessed that, that was a good feature. It turns out I think I'm sick is a negative feature. I just realized it was because if you have food poisoning you don't think you're sick. We ran a trial in Las Vegas with a southern Nevada health district where we collected tweets on Las Vegas and said, which of these tweets is coming from restaurants. We then tracked all the tweets from the uses for the next three days collecting them and then linking them back to those restaurants and used that to rank the food vendors.

This was a system actually used by the health inspectors. They would look at the report for the last week and they did a matched trial I think. They took the top ones recommended by us and they found similar restaurants, similar size, similar kinds of food, similar price range. Did a double blind trial so that the inspector does not know, which ones we suggested, which ones just came up because it was time for them to be inspected. To our surprise, it worked. It's not a perfect system but it was quite ... they said they were very pleased. If they found 11 bad actors versus 7 bad actors, it almost doubled their effectiveness. They were very happy with that. If we just ran the numbers, they said because of those extra four we could predict that we prevented 71 cases of food poisoning. They also found an unlicensed food vendor because our system found someone who had been eating fish and they looked in their database and said there's no such fish restaurant there. We said, there is now.

That work is actually continuing on. We are working both with Nevada and with Google to see what we can do to scale this up, take it nationwide, and also rely less on Twitter but looking at some other ways of getting the data. Just a few other applications of the same social media analytics tool box. A system called GeoDrink, understanding patters of alcohol use in communities. For example, you want to cut down on teenage drinking. Is it better to close the liquor store or to close the bars? Maybe neither because maybe if you close them both, people will jut go drink in another location. We bought classifiers that could both measure people drinking and tweeting about drinking and then tweeting about drinking at that location at that moment. These are all different things. A lot of previous were blurred these all together. If someone just said I love Bud, they might classify that as a drinking tweet but maybe they were just repeating a slogan from a TV commercial.

We could also infer where the uses were located, where they were drinking. For different communities look at how much of the drinking went on essentially at home and in the neighborhood versus in distant places. How far did the person travel to do drinking? If you're controlling alcohol outlets in the neighborhood, you're going to probably effect this part, you're not going to effect that part of the curve. Finally just to briefly mention some new work we're really excited about is using similar techniques for mental health screening and for actual working with therapists who are working with patients with depression. Typically, now a subject comes in for their weekly meeting and talks about how things are going. It's all based ... the doctor asked the patient, how are you doing this past week? Patients are really not very good at remember how they were a few days ago. They're good at knowing how they feel at the exact moment.

However, their online behavior leaves a trace of what they were doing. Were they up in the middle of the night searching and posting at 3 AM? Were they obsessed with something that might indicate a mental health issue? We basically have a new join project that hopefully starting in two weeks, we have patients who have consented to allow us to access their online behavior history, their historic search logs. From that we can extract various features about sleeping patterns and topics that they're searching on. Then see can we use this to predict how the patient's overall mood is into the future? That's a little ... my little tool kit. Or not my tool kit it's my hammer or screwdriver there about social media for public health.

Amulya: We have ten minutes for questions.

Henry: By the way, you can also feel free if you have a question for Carla you can ask her too. I think you had up your hand there?

Speaker 4: Do you think in the most recent experiment or the most recent model you'll run into a problem where people will change their behavior because they know they're in the trial? How would you control for that?

Henry: That's why we're starting by doing retrospective information. People consent and because they're patients, we have their medical history. They consent to let us look at their last two or three years of search data.

Speaker 4: But if their behavior does change, does that nullify the trial?

Henry: What will do is we're looking at how their behavior over the lat year compared with what the doctor was saying. Their medical history over the last year.

Speaker 4: Oh, I see. [inaudible 01:21:05]

Henry: Yeah. We'd also like to do prospective. Yes, that would be indirect. There's a chance it could affect behavior. That will be an interesting issue to look at.

Speaker 3: I have actually already a paper on food illness and after that I started working on food poisoning. What was interesting was sometimes I found tweets where people were saying I'm definitely going to get food poisoning. For some reason I found very suspicious. I went through their tweet history and I found that people were discussing ways to get out of work. Then they were tweets like get well soon or I hope you feel better. When I went back, I found that those people had food poisoning. Going to through the positive tweets, going through previous tweets in all the cases at least 24 hours back help you to drain them out. At the same point, if you take Yelp, you already have the restaurants. If you find that these restaurants have been giving cases of food poisoning, you can go to tweets and you can actually even trace down more tweets to train for the machine learning purpose.

Henry: That's great. All papers came out. First one was three years ago. The other one was a year ago. I'm looking forward to talking to you during the break. This is very exciting. Thank you.

Speaker 4: Hi. First off, thank you for the talk. I found it very interesting. I had a specific question, actually two questions about the Twitter flu project. Did you do any analysis about the rates of flu and the robustness of the model versus the CDCC estimates versus time of year? Were the estimates better when flu season was high and you're getting more reports and it's not as sparse and super unbalanced in terms of training data?

Henry: We did not do a formal power analysis but informally that does seem to be the case. When it's not flu season, it depends also the size of the geographic area. In New York City, there's enough sick tweets throughout the year. When we were working for example just in the city of Rochester, when you're not in flu season, the sick tweets are too sparse.

Speaker 4: This leads into my second question. Were you able to train a model based no data from you said New York City, get it done, do all the training, and then apply it to a different city like San Francisco where you can try to leverage that and see if the rates of infections and flu transmission is different based on a different geographic location. Do those models hold up to different locations?

Henry: We use the same language model. Actually one of the questions we have that we still need to investigate is could we have done better if we had trained a location specific language model. One thing we noticed when I mentioned the work on alcohol that some of the people would discuss this work with, they said they weren't sure that we were capturing the language of alcohol use from their community. That we go through and they would say phrases, we'd type them into our system and it did not classify them as alcohol use that [nacking 01:24:43] is drinking. We just never got that. Definitely there is that issue of the locality of the models.

Speaker 4: I challenge and I presume that you're running into with working with Twitter data to try to figure out whether there's something flu going on. Is that the disease itself is a latent variable. What you observe are the tweets. For example, when you look at anywhere very precise about the way you're describing that you've seen or if you have a lot of people around tweeting about having the flu, you're more likely to tweet about having the flu. The question really that people would be interested in this, their relationship between people having the flu and the likelihood of you having the flu. If the baseline, if the probability of tweeting, that presumably will vary also based the social connections and the area you're in as well. I can imagine this is an inherently challenging problem. Have you thought at all about this? How would you address it?

Henry: Definitely with Twitter you're limited to a particular segment of the population that's not representative as broadly as one would want. We had thought about using Facebook data. Facebook is very hard to work with. Unless you're a Facebook, you can't really do that. We didn't have a Facebook insider on our team. What we're looking at now is doing one more work with Google. Because as you know from things like the original Google health trends, they have access to search log data. You can do a lot of what we did but looking again instead of at tweets, looking at search histories. Again, you have to be inside the Googleplex. Fortunately my student who started this project is now inside the Googleplex.

Speaker 4: This is a question for Carla and Henry both. You talked about how there are ... we should generalize to other domains, the three domains. Henry, you talked about how you're trying to apply things for social good from the more commercial applications of this technology. Are there specific research challenges because your applications are in the kinds of areas that you have? I understand that we should generalize but is there something more specific that you think is more specialized for these types of domains that's different from the other domains? Special kinds of research challenges, computational constraints, anything else.

Carla: There are many challenges. Maybe this is not a research challenge but it is actually to get people involved in these topics. Even here you see when we talk about Twitter and whatever area. If you then go to hardcore sustainability challenge, maybe they are not as sexy. That's not what I think your question. I think your question is different if I understand this. Are there inherently recent challenges that come from addressing sustainability problems? Yes. There are many. I can say typically we number one, we are trying to truly predict and manage very complex systems. These are the smart grids or natural ecosystems such as the coral reef or poverty. Then we need to really look at factors, socio economic, very multi disciplinary. In terms of you need to gather data. Typically, we are dealing with multiple scales, multiple temporal scales. There are a phenomena for me to predict the birds. I have to factor in phenomena, characterize it minutes or seconds but also larger scales.

Months, seasons, et cetera. This multiple temporal scale is actually a huge challenge. Spacial scales, again, even some phenomena we need to characterize as small, little, or local scale. The same may also have global scales. Bringing together this. Then we are talking about really typically complex dynamics. Highly non linear where you need to ... for example, we have a project that now we are looking at that is in South America. They are planning to install, to build in the next ten years about 500 hydro plant dams. The consequence is in terms of the environment are dramatic. In terms of fishers, in terms of sediments, in terms of ... obviously these populations, they want electricity. We did. We build so many hydro dams in the US. The complexity, the highly non linear dynamics, these problems are really hard. I do think because of that for example, multi agent systems are going to play a big role because you need to look at this phenomena and understand the different agents involved.

You are trying to protect species. Above all, you are also trying to give the populations better standards. Then there are conflicting factors. Multi agencies sometimes will play a big role in this highly complex dynamic. You can not just say I'm going to ignore them. For example, in Ecuador we are trying to also build this suggest locations for or how to protect, build corridors for an Indian bear. Probably if I ask anybody do they care about an Indian bear or sending the kids to school to the university. We need to think of solutions that will also affect positively even though the livelihoods, the communities, et cetera. We can not even talk about just a corridor for the Indian bear. It has to be a social ecologic. In fact, we are talking about a social ecological corridor where we also will provide opportunities for the communities. Show them the importance of keeping the forest, the importance of keeping the water. There are lots of mining industries, et cetera, that very profitable.

Many challenge I think. As I say, I think we can't really ignore them. We need to attack them and solve these problems and factor in all these businesses. That's what I find exciting. I also gave a more restricted example. In the end, we need to develop models. We need to have data obviously. In order to collect data we also need models. Typically, and I thought that actually and you illustrated that beautifully. For you to give predictions you need to bring together very two genius sources of data. How to combine all of that. Then how to interpret the data. I think there are man challenge. As I said, for example for the material discovery where we are trying to help our collaborators with solar fuels. By the way, solar fuels are quite an exciting technology because for example when we talk about solar energy we have the panels. Then the solar energy is very intermittent. If there's sun, you can have energy. Otherwise, you don't.

Solar fuels actually you could store them. That problem would go away. I think for example, for computer science I love this problem because a lot of work has been done in terms of topic modeling, dimensional reduction et cetera. Here we do need to factor in constraints. We need to have solutions that are physically meaningful. In general more scientific domains that Henry will also talk about even though he can not give to trends or whatever. Now he has to really validate. In fact, I have a question. Is it easy for you to get the ground truth?

Henry: No. That's the easiest.

Carla: It's not. That is a key problem. For you it's for example, we throw away that problem the board is boggled. In his case I'm sure that is an even more dramatic problem.

Henry: Yes. Just to give a shorter answer. Both you and Carla have done a good job of finding problems that have real world impact but also have an interesting novel computer science side. Particularly in health care delivery there's so many places where you can basically just take what you might learn in your first course in machine learning, apply to some problem, predicting which patients get a visiting nurse after they're discharged. Immediately it has a big impact. However, to get that actually used and deployed and tested takes serious, serious money. You can't really get it from NIH, I mean NSF, because it's not novel algorithmically. For NIH, sometimes we'll fund things through a mechanism called RO1. Again, there's the challenge of getting credibility. You can have like I say, an orthopedic surgeon. You say, I'm going to predict, which orthopedic patients need to get a visiting nurse. Now I say then I want to actually then go apply this to somebody who's going in for stomach surgery.

They said that's a completely different condition. Now you have to get a stomach surgeon. I'll tell you right now. There's this notion that I think in the medical world people tend not to think as abstractly. In computer science we think things very abstractly. We feel comfortable. Medicine is this world of immense amounts of fine grain knowledge and specialists. Everything is separate. Even every little version of cancer has different groups of people working on it. You come along and you say, I have a general tool that can apply to cancer and back surgery. People just going to laugh at you.

Carla: I have a question for you Henry, another one. Have you looked into this word embedding the models?

Henry: Yeah.

Carla: Which I think this could be a very interesting area where you could try to build the embeddings. For example, we are doing that for the boards. We have these embeddings. We can talk more about-

Henry: Yeah. That's where I'm getting some more generalization in a model.