CCC

Computing Community Consortium
Catalyst

AAAI/CCC Symposium on AI for Social Good

Talk Sessions 4: Computational Sustainability
Session Chair: Dr. Fei Fang

Fei: It's a great pleasure to share the session and during this opening talk, I want to make it a little bit different. I want to make it very interactive, so that I prepared a list of questions and want to elicit answers from the audience. And some of the questions are kind of quiz questions if you listened to the invited talk by Carla yesterday, very carefully.

[laughter]

I will try to provide some example answers based on my own experience, but indeed I want to get everybody involved in the discussion so that if we run out of time, we can save the questions to the discussion session after the talks.

This session is about computational sustainability and this a very rich and interdisciplinary research area. It aims for sustainable development. We want to develop advanced computational methods that can benefit the whole sustainable development of human beings.

Then the first question is, what are the main aspects for sustainable development when we talk about it?

Any answers? You can just shout out. Okay.

Speaker 2: Number one was no poverty.

Fei: No poverty? Great. Great. Yeah. That's kind of like uh ...

Milen: Power Reduction. No hunger.

Fei: Power reduction. No hunger. So, Milen is mentioning poverty, that's kind of like the economic aspect. Also, no hunger is the social aspect of the development. anything else?

Speaker 4: [inaudible 00:01:58] [laughter]

Fei: Or we just summarize them into three major aspects. Yeah. Eugene?

Eugene: Reducing inequity.

Fei: Reducing inequity. Yeah, again that's kind of like a social aspect. Important.

Speaker 5: Environment.

Fei: Environment. Yeah, exactly. [laughter] At the very high level, this is environment, economic, and social aspects that are the main big things of sustainable development.

Then, next is why do we need advanced computational methods for sustainability? We have all other researchers working on this area. Why do computer science methods are needed? Or for what kind of problems are advanced computational methods needed?

Based on my own research, what we have seen is that in many conservation areas, the conservation agencies need to try to fight against the illegal activities. However, they often have very limited resources. Then the task for computer scientists could be to try to optimize a use of these limited resources.

In some instances, instead of asking the experienced domain experts to do the planning and optimization, we can build some formal, mathematical models for these problems and try to apply AI techniques to try to find out the solutions for the use of the limited resources.

This can happen in various domains like wildlife protection and protecting forests from illegal logging, or protecting fishery from over fishing. An example is a project that I have worked on when I was a PhD with Milen, which is a protection assistant for wildlife security named APOS. Basically, this project is trying to take past patrol and poaching information, as well as the the information about the conservation area and try to use some algorithms to try to design some patrol routes for the rangers or for the patrollers, from which, we can collect data and then further improve the whole workflow.

Previously, this PAWS core algorithm was taken by a domain expert to do all the planning, but now what we can do is we can kind of propose algorithms to learn the poachers' behavior, and then add in game theoretic reasoning, and then plan the routes in an intelligent way.

What's the benefit of this? First of all, it saves time for the experienced domain experts and also it may add into some aspects that the domain experts didn't explicitly thinking about. For example, the game theoretic aspect, like how the poachers would react or respond to the patrolling strategy. Clearly the domain expert might also think about it when they do the planning but what we can do is we provide a mathematical model that makes it explicit.

This is an example patrol route that is provided by PAWS and when it has been deployed in Southeast Asia, during which the patrollers do find sizable animals and human activities.

Here is my example answer for this question. We can, as computer scientists, we can provide tools to facilitate the decision making of the officers in the field. We can try to provide algorithms for optimization and learning and planning. So, any other thoughts on this question?

If you remember in some of the example projects yesterday during the talk, you may still recall some of the key things that we computer scientists can do.

Any thoughts?

Eugene?

Eugene: I think probably biggest marginal value of computational tools is actually the modeling and principled way to think about these problems. Framing of these computational problems and thinking about them systematically. I think that's really where a lot of the values lie. This is why we can have reasonable dialogue about what approaches are better or not.

Fei: Right, right. Yeah I completely agree. What Eugene mentioned is like we need to think about - We kind of bring in a different aspect of thinking, different way of thinking into the sustainability challenges.

Yeah?

Speaker 7: Actually, you already mentioned and your comment is very relevant and often, computer scientists have these attitudes saying, "Well you know, yeah the sustainabilities will benefit from this interaction." But what about computer scientists?

That is such a short, limited way of looking at it. Because actually often the computer scientists benefit a lot. One of the issues is that typically we are dealing with experts, who have thought about these problems. They actually have very creative solutions. But indeed, what they lack is the strengths in terms of models, in terms of computation. That's where we come in. But we actually benefit from the ideas they have.

I remember for example, Stefan who worked on fisheries. He learned quite interesting models that the mathematicians had developed for fisheries, but they not scaling up. Please don't underestimate what you can get from the sustainability field. It's actually quite exciting.

Fei: Great point.

Speaker 8: I guess after what Carla was saying, there is certain boundary, an autonomy boundary, or something like that. Human being experts will be good at certain tasks and these computational decision makers will be good at others.

We're not proposing in this method to completely displace human expertise. It's sort of a logistic relationship. Finding that right boundary A) because human beings will resent the fact that you're micromanaging their activities. In your own work, for example, you may guide people to say, "Look in this area for a sample but not this particular "x" "y" coordinate." Because they know better within that area where to look. Something like that.

Also because of this resentment, but finding that right boundary is an interesting challenge and it seems that's something that we will need to work towards to find out how to really manage that partnership, AI and human experts.

Fei: Right. I completely agree with that. Any additional comments?

Yeah.

Speaker 9: I completely agree with Milen's that it's interesting balance but I don't think it's like a fixed balance. This trying to solve for the years when people use more of this kind of tools, they will change their thinking and the balance has to change as well so we don't need just to balance but a methodology to actually recreate it every time. I think that's the real new thing also for computer science because we don't have a fixed thing. We have to have something that we can adjust easily and change the use of it.

Fei: Yes basically the balancing point can change over time. As we build more trust with the agencies, they may be more acceptable to newer techniques.

Speaker 9: [inaudible 00:10:22]

Fei: Yeah [laughter]

Speaker 9: [inaudible 00:10:37] it evolves over time.

Fei: We probably only have time for one more discussion. We all know it's interdisciplinary area and we need collaborative efforts from different research communities. The question is which communities do we want to reach out or have some kind of combined efforts. Based on our experience when we walk on the passing, we definitely need to know how the animals are distributed in the area and we also want to learn how the poachers are responding.

To do that we are designing some online virtual games and ask people from [Animal Mechanical Turk 00:11:19] to play the role or poachers and try to decide where they want to place the snails and from that we learn what to do.

Clearly in this case, we are combining the knowledge of computer scientists and psychologists and ecologists. Need the input from all these experts from these areas. Of course, there are other research communities, that we want to reach out to for collaborative efforts. Anything that you can think of here?

Speaker 9: [inaudible 00:11:50]

Fei: Criminology. Yeah.

Speaker 10: I would like to think it's any area, any discipline where you have a need for automation of any kind. Any thing that's very manually labor intensive discipline, ecology comes to mind, because there's a lot of [crosstalk 00:12:14].

Fei: Ecology.

Speaker 10: Many others in prime statistics. Anything that can be automated away to eliminate that workload can better leverage that expertise in the field to go after problems that we actually care about.

Fei: Basically many fields with the need for automation. That would be involved in this collaborative efforts.

Yeah.

Speaker 11: One of the big problems is in inherent bio seeds. When you have communities that believe they know what's best, because they've worked in it for years. When you come in with new types of techniques to try to show them that this isn't necessarily the best way of doing things. It can be disruptive and it can also potentially kill your career because they come after you.

The question is what are those communities? Number one, I'll come back because I work in it as health care.

Fei: Health care.

Speaker 11: It's a very dangerous community to work in. Number two, is the banking industry. Financial services, in particular, especially when you're trying to do redistribution of wealth. Especially and particularly into from first world countries into third world countries or keeping things in third world countries.

I can go on and on but I'll stop.

Fei: Great. Thank you for the answer.

Speaker 12: I guess I mean especially because we are talking about all the dimensions, social, economic, environment. You pretty much have to involve everybody. Social sciences, after all, we are trying to improve human well being. Climate, public health. That's actually part of the challenge is how to build this interdisciplinary research project.

Fei: Very good.

Speaker 13: I think it also comes down to a powerful position. It's our obligation to explain what we do and what we bring to the table. Yes, every technology has a - it could be disruptive. I think disruptive is not necessarily a bad thing but we have to explain to them why they should want to work with us. They're not the enemy.

Fei: Exactly. Thank you all for the answers. I guess we don't have time for the rest of the questions but I just want to show the questions and we may come back to these questions later after the talks.

For example, how to bridge research and practice. What are the other research things that we can pursue and how to make the research more impactful and how to make the community sustainable itself. Funding wise and problem wise, so what are the best avenues.

With that I would like to introduce our first talk by Jennifer. She is learning the temporary evolution of climate change. Assessment research using dynamic topic models and cross domain divergence maps.

Do you need [inaudible 00:15:54]

Jennifer: No I don't thank you. [inaudible 00:16:03]

Okay hello my name is Jennifer Sleeman and today I'm going to present the work that is a collaboration between UMBC and Columbia University. The topic is "Modeling the Evolution of Climate Change, Assessment Research Using Dynamic Topic Models and Cross Domain Divergence Maps."

I'm on certain scientific disciplines. There's direct impact on society given the research. Typically those scientific disciplines have panels. Those panels are formed to assess the research and make recommendations. Often that research involves quite a bit of literature that's evolving and also that is often interdisciplinary. Hence those panels tend to involve scientists from different disciplines.

Understanding the evolution of research and how that research influences the assessment or recommendations can improve the process for future recommendations and assessments.

In particular, the IPCC, is an international body which assesses and evaluates the most recent scientific research related to climate change. There's over 20,000, I'm sorry. There's a number of researches that actually contribute to writing and reviewing these reports. They're created every five years and I think that's all I wanted to say about that.

This is the structure of these reports. The reports have basically, there's been five assessment reports thus far. It's made up of four different books. Each book has a number of chapters ranging from 11 to 18 chapters. Each chapter tends to have anywhere between 800 and 1200 citations.

The physical science basis book describes the latest research, the experimentation and numerical results. The impact book describes how the physical science research impacts our world. The mitigation book describes mitigation steps that could be taken to reduce the impacts and then there's also a summary book.

Our key contributions are well this is the first time that the IPCC reports 30 years of reports and the full text citations have been modeled using a semantic language model. Typically climate change research tends to be more of a numeric analysis.

Also this is the first time that we're combining the topic of evolution and evaluating cross domain divergences for this type of model. Understanding how the research is changing over time and understanding how those sub disciplines are interacting with each other over time.

Our work is - there's two papers in particular that are foundational to our work. The first one is "Dynamic Topic Modeling." By David Blei and group out of Princeton. This paper outlines how to essentially take a topic model for each time slice and chain them together conditioning the parameters for time slice T on the parameters of time slice T minus one.

The second paper that's foundational to our work is the "Topic Correlation Analysis for Cross Domain Text Classification." They're also using a cross domain approach. They're using it to solve a different problem. They're using it essentially perform text classification when you have unlabeled data. They're using their source data that's labeled and performing cross domain analysis to do prediction on the unlabeled data.

Roughly this is our methodology. Given that we have two domains, we could have more than two domains, but given that there are two domains, we go through this pre-processing step. The pre-processing step is the natural language processing step. This step is where we essentially extract the word engrams out of the text. This is governed by a climate change glossary that we've created.

We create dynamic topic models for each of our domains. We filter those topics. You can think of it as a feature reduction. We reduce the number of features that is the number of terms. We take the union of those two vocabularies that are formed from those future reductions and we form this combined vocabulary. We re-normalize and from there we are able to do topic divergences using Jensen-Shannon which is just looking at how two probability distributions diverge.

This becomes the basis for us to cluster documents from the two domains into one space. We set a second threshold at the document level for the topic probabilities and we play with that threshold to tell us how many documents that will include given that couple of topics from the two domains.

We ran a number of experiments thus far. We've run physical science citations and impact report. Physical science and mitigation report. Impact citation and mitigation report. These are some of the statistics on the document counts. Our documents in this case are chapters. We play with that a little bit. Sometimes with our topic modeling, we experiment with subsections as documents and then full documents.

In the case of citations, we use the full document. This is just an example of one of the cluster results that we have. The top right corner here shows a couple of the two topics across the domains, the terms that are in common. Then at the bottom right here, we have the results of one of the clusters so what we show here is our model was able to pick out all of the chapters across of all of our assessment periods that were related to coastal and oceanic issues. In this case, we picked out a citation. It's the same citation that was cited in two different assessment periods.

We have a lot of work. The biggest effort for us was getting all of the documents in a representation where we could actually model it. We're at the point now were we're benchmarking our method against other methods.

We're playing with two concepts here. One is relatedness which is more tangible and easier to show. The other is influence which is actually a lot harder to measure. The approach that we're taking at this point is something that's known as a dynamic data assimilation where we basically take a report, we add in the citations from that point to the point of the next report and see if we can predict what would be in the next report.

We're also experimenting with neural base variational inference. We've done some preliminary work with this where we've actually been able to compare the output from our topic model with the output from the neural base variational model. This just allows us a lot more flexibility. There are some weaknesses in the dynamic topic modeling.

This is a big effort from an IPCC perspective. What we're able to do is take 30 years of research and model it and allow someone to be able to use it for search and discovery. Also, another nice side effect from our work is that all of the authors from our citations are modeled so we can do social network analysis and see how those networks are changing over time.

Okay. That's it. Thank you.

[applause]

Speaker 15: Very interesting talk. Thank you. I was just wondering why use controlled vocabularies and not anthologies? That has ramification for everything including relatedness which controlled vocabularies would only allow you to calculate using some form of correlation. You cannot actually qualify how things relate to each other in what way. Also you may have articles that talk about the exact same thing. They just use disjoint terminology from a strictly lexical graphical perspective.

Jennifer: Right. From the topic modeling standpoint. We are able to overcome some of those barriers because it's looking at co-occurrences of terms so if - it's true. A controlled vocabulary does have some limitations and an ontological representation would also be very interesting. I haven't thought about that. We do a lot of work with ontology so that might be something we consider.

Fei: [inaudible 00:28:10]

Jennifer: Okay thank you.

[applause]

Fei: Our second talk is by [coughs] sorry, Jason.

Jason: That's me. All right.

Good morning. I would say afternoon but it's not just quite there yet. My name is Jason Parham. I'm a PhD student with Rensselaer Polytech Institute. I'm going to be presenting a paper today on animal population censusing at scale with citizen science and photographic identification.

Our work is a multifaceted, multi-university, multi-organization, collaboration between Rensselaer Polytech Institute, the University of Illinois, Chicago, Colleges at Princeton University, and then actually a nonprofit in Portland called Wild Me. This is all funded by the Kenya Wildlife Service in Kenya and also our own national science foundation.

Our whole problem is to try to produce an automatic census of animal populations. If we compare from the previous kind of traditional way of doing this, it's generally invasive. It requires doing some sort of ear notching or ear tagging or radio collars to track animals and track populations, which makes it expensive. Generally this requires a lot of time and money, special equipment, things like that.

It's also error prone. It requires either accounting blocks and methods to breaking up a particular area that you want to survey to try to capture all of the animals that you've seen over time. Generally because it's so demanding, it's generally one off.

How can we make this process which infeasible for large populations, so you're trying to track 2,000 zebras walking around a part as oppose to ten rhinoceroses. How can you use computer vision and computer science to make this better?

What we propose is a passive appearance based model, where we actually use the physical appearance of the animal to be able to distinguish them. Because of this, it's inexpensive, we basically only require cameras. It's very easy to train people to just go out and just take pictures of animals, right? Tourists, school children, even a static thing like a camera trap is a good place to get information.

Therefore it's evidence based. We have an actual picture we are able to correlate between a sighting and an individual. It's incremental. We're able to add to the analysis over time by having new sightings over time which is great because it's ideal for large populations, makes a very distributed, decentralized, that's what we would like.

To put into context. All of our work is in Kenya which is on the eastern shore of Africa. I'm going to be focusing on two censusing events that we actually did in 2015/2016. The first was the GZGC which the Great Zebra and Giraffe Count. It's just south of Nairobi which is the capital of Kenya. Then we came back a year later and we actually refined the process and we scaled up to actually doing the GGR which is the Great Grevy's Rally.

Let's start with that first one. This is the map of the sightings that we asked people to go out into the park and actual take pictures of zebras and giraffes. We brought them back and actually did analysis on these. These are actual places in the park where images were taken. We kind of painstakingly did this with little GPS dongles and we gave people cameras and we asked them to go out and come back. All the red dots are animals that were seen on day one. This is a two day event. All the purple dots, are days that are, excuse me, all the blue dots were animals were sighted on day two only and then the purple dots were animals that we saw both days.

Looking at the map for the GGR, we can see that, the area that we're actually surveying is much, much larger. If we go back to this map here, we went from a very small area around the capital to a very larger area around the Laikipia region of central Kenya.

By scaling up and covering roughly about 100,000 square kilometers, we're able to evaluate our process at scale which is what we also like to do. How good is the power of partnerships? Right? We asked random volunteers to go out into these parks, got out to these areas and just take pictures. Very little training. We asked them to take a specific view point of an animal. That was basically all the required training that was needed.

You can actually see the number of images that had been collected on an individual basis is fairly high. For the GZGC, the highest person gave us roughly 1,200 images, thereabouts. For the GGR, that got significantly higher, very over 3,000. We get a good amount of data per sighting and these are actual images that were taken.

We actual images of zebras that we care about. We get random images of buses, rhinoceroses. These are tourists. They are interested in going to the park and they seen an animal they never seen before. You get alligators and people never seen alligators before.

How do we take all of this information and try to deal with it in a systematic way where we can actually produce a population estimate?

The first thing we do is we take an image like this. I love this image because it perfectly represents the problem that at least I'm trying to solve in my PhD which is how many animals are in this image?

[laughter]

It's complicated. You have this little guy here who is just off the frame of the image so you just see a sliver of a neck. You see the little head poking out in the left which basically you see only that and maybe a little bit of a leg. You can see the complexity of things that we're trying to do here. Via machine learning and deep learning, we're able to put bounty boxes around these animals to do a classical detection problem.

We're able to classify them into species and then actually do some sort of background subtraction to get rid or the information that we don't really care about like grass.

How well do our classifiers work? We do into two stages. First we try to figure out the species of the animal. Is it zebra verus giraffe? Then we try to figure out the actual view point that we saw this animal. The orange boxes are species. Anything outside is a species classification that was pour. Anything inside the boxes is a viewpoint classification.

Given a particular annotation. What we ask the algorithm to do is find all of places on this animal that are distinctive, that have a lot of variation, a lot of contrast. We ask the algorithm to take these key points, we normalize them to a unit circle and then we use sift to actually give it a numerical value that we can build into a big database across all these different sightings and be able to search for things where we actually take a particular animal, search in the database for all of these descriptors and then actually start doing matching.

In this way, we can start to identify individuals in the population, instead of just sightings. This is a very different procedure than like eBird where instead of tracking just - we saw a zebra, this is that specific zebra on that day at this location. It's much more powerful.

How does this actually work in terms of the matching. We get roughly 80 to 90 percent depending on what level of review you're willing to do. How many databases matches you're allowed to return. The actual collection data statistics are also interesting because we're asking these volunteers with very little training to go out. How good is the data that we end up getting?

We gave them the very simply restriction of just take a picture of a viewpoint, one specific viewpoint. Half of our images that we got were of the correct viewpoint of the correct species that we wanted. That's great. Anybody that's done any kind of citizen science, decentralized data collection, that's a phenomenal number.

You actually see the number re-sighting across the two different events. This actual gets fairly robust as you scale up which means that the data collection holds or even gets better at scale which is always encouraging. It doesn't start to crumble.

The actual number sightings. Oops excuse me, I went the wrong way. The actual number of sightings per individual indicates how well the actual algorithm is performing for the GZGC, the first census rally we did, we saw individuals only once about 900. Not terrible great. We would prefer to get repeated sightings that we can make the estimates more robust. We actually ended up getting better when we scaled up, got more images, got more participants involved. The population estimate actually ends up to converging nicely.

For the GZGC as over time as new sightings are seen, you would hope that that would start to asymptote along since you've actually captured all of your animal population. I would like to say, that the GZGC did a very good job at this but it's still sloping up a little bit. We're not capturing everybody. We came back, we scaled up, we were fine, and we actually get a much better asymptote.

The power of big data, the power of using more collection, gives you better confidence balance. What are the ultimate numbers. This is the golden nugget. These are what the scientists, the ecologists care about. How many animals are in a park or in an area? We can actually estimate how many animals are seen in a particular area by knowing how many we actually seen, what's the convergence rates. You can see for the GZGC, we estimated roughly 2,300 zebras in the Nairobi National Park with very high confidence balance because we didn't see everybody.

When we came back we estimated roughly the same number of Grevy's with a much better confidence bound which is great. You can see the number of participants, the cameras almost tripled and the number of photographs also quadrupled.

What does this mean as from a goal orientated? We want to maximize speed and we want to maximize accuracy of these predictions because we want to get the ecologists that we actually care about.

In conclusion, we want to achieve speed and accuracy for large scale animal population censusing. We want to eliminate the process bottleneck for ecologists. We want to use computer vision algorithms that can enable detection identification algorithms to actually perform these things automatically so you don't have to rely on experts to do this population estimates. The last that citizen scientists can be effective and not only that, they can be contributors of high-quality data. We can rely on them to give us the data that we need, completely decentralized for free essentially with very little training.

What does this actually mean? Participating volunteers actually become engaged as community advocates for conservation which is the best part of this whole thing is that we're getting the community involved in this whole idea.

Thank you. I think I'm out of time.

[applause]

Speaker 17: Is there any sort of understanding of how well do these populations censusing is which rely on crowd sourcing work as opposed to I mean, how close are the data to the ground truth?

Jason: It's new. The ground truth being the existing method that ecologists have been doing which they block base count and they're traditionally fairly err on the very confidence balance. Our estimates are following along with those numbers. They're roughly in those bounds but those population estimates seem to be anywhere from 1,000 to 3,000 animals, roughly give or take in this area. Our confidence balance would be there's 1,800 animals and we think there's plus or minus 100. We have much more accurate numbers that you're able to make decisions. Your turning a college into a data driven science where you have conservationists and managers of these parks to be able to make decisions on where these animal are going. What do they do? How long to they live? If they don't have the data to be able to make those decisions, it's hard for them to be effective managers of this area.

Speaker 17: Won't the balance depend on the number of crowd workers that you have and so I guess ...

Jason: Not necessarily no. It's more of how fast you can get through the processing because a lot of the workers that we end up using for review ends ups being for actually getting through the things that are uncertain. What the algorithm is uncertain about.

In terms of the robust estimate, how many people you can participate in the data collection plays a huge role of how robust it is. Generally the order of scale is much different than previous older techniques, it generally is a team of five or six rangers that would go out into the park and they say, we just saw 18 animals here and 12 animals here. You hope that they don't cross boundary lines because you don't want to double count.

For this you are able to use an order of magnitude more, a number of volunteers, many more sightings, evidence based to be able to make these refine predictions.

Speaker 18: What we hear from our collaborators from Wildlife conservation society is they'll fly helicopters with people that over Queen Elizabeth National Park or something like that and they'll be looking over and counting how many zebras or giraffes they saw and based on that they'll give a count. I guess the question is, this method could compliment that I would assume.

Jason: Yes. That it's not necessarily one or the other. It really comes down to resource constraints. If you don't own a fixed wing or a helicopter to be able to do these population and things multiple times, maybe you can rely on both methods to sample more robustly in time instead of doing a full blown census every three months. That's cost, not cost effective, you'd require volunteer to come continuously over time so maybe it can be augmented with a ariel based survey to help maintain that population estimate.

Speaker 19: Typically tourists in those parks, they go in big groups like several jeeps together or a bus in a cougar park or whatever. You will have typically like 50 tourists making a picture of the same zebra. How do you account for that?

Jason: You can actually see that reflected here. The number of cars that we have for the GZGC was 27. The number of cameras were 55. There's more, roughly two, two and a half cameras per car. You get a lot of these repeats sightings. What that means is that from a data analytic standpoint, a data management standpoint. How do you filter these multiple sightings down to the processing that you actually are about? What we're able to do is actually refine the identification process over time to take the best sightings of these animals, the best viewpoint, the best sighting. To be able to match against our database more effectively and efficiently.

Speaker 19: Are you determined that it's the best picture of one animal or two bad pictures of two different animals.

Jason: That's essentially what we do. That's the goal at least. That work is actually done by another PhD student. His name is Johnathan [Crawl 00:42:41]. It has a lot of work of what's the photographic quality of these images? Are they blurry? Do they have a lot of shadowing? How well does a particular match in the database go against how many come back that are true positives and picking the best ones that are the best exemplars of that individual at a certain viewpoint?

Speaker 20: So how confident are you about the confidence bounds and particular what distribution. I didn't get quite get what distribution assumptions you have to make. Does it account for the biases in terms of what people select to take photographs of and so forth.

Jason: That gets into a level of detail that is hard to do in even a talk but I'm willing to go into it now. Ecologists have been doing these kind of the mark recapture studies for 30,40 years now. The whole idea is to take a capture of a certain population on day one, mark them, take another sampling that is hopefully to be unbiased. You hope that there's certain restraints that are maintained in that time window. No births, no deaths, no immigrations, things like that. If you can maintain those constraints, you can actually get very confident bounds. All of these numbers up here are reported with a 95 percent confidence bound. What we end up doing is transforming a mark recapture study into a sight re-sight study which is slightly different in terms of what it's actually tracking but the mathematics and the statistics underneath still hold. They're still valid. We assume these animals aren't migrating out of the park between days. We assume that there's no significant change in births and deaths of these megafauna over days. A lot of those statistical balances and assertions can stay held on something like this.

Fei: Great. Let's thank the speaker again.

[applause from crowd]

Next talk will by Zhiyu

Zhiyu: Hi everyone i'm Zhiyu Wan, a PhD student from Vanderbilt University. It is a great honor to stand here to present our work on Game Survey and Data Sharing and Privacy.

Okay now lets get started. I forgot to mention, a paper based on this work has already been published on American Journal of Human Genetics.

Now let's get started. With some background and motivations of our work. One reason we need to share genomic data is because it is beneficial to the whole of society. First of all, the genomic data is extraordinary valuable. For example, genetic testing can help the doctor and patients with diagnose of diseases using the associations between genes and the disease.

This genetic testing was brought to the public spotlight by Angelina Jolie about four years ago because she has some genes that's associated with breast cancer. Genomic data is also influenced the drug effects and treatment.

That's why the genomic data sharing will accelerates the discovery of new associations and especially for the rare diseases that needs a lot of data. The NIH incentivizes investigators to share genomic data by make some policy for funding and building some data repositories for sharing genomic data. Also trying to protect the privacy of the data subjects.

We can see when our big genomic data era. In the past two decades, because of the sequencing genomic data, jobs from 100,000,000s to about 1,000 dollars. In the beginning of the 21st century. The International HapMap Project has only about 100 subjects. One decade ago 1000 Genomes Project has about 1000 subjects. Recently the Precision Medicine Initiative aims to collect data from one million patients. These transfer will keep growing.

However there is privacy risk ensuring these data. If it is shared in individual-level with sensitive attribute, it is very risky. For example there is table that show the shared genomic records. If an attacker can collect and sequence DNA samples from identified targets he can conduct linkage attack.

If there is a match between two tables than we can adjust the target has a particular disease which is a privacy breech. However sharing the summary statistics is still useful but also risky.

In 2008, Homer introduced an attack. The attacker knows the genome of the target denoted at "Y." The allele frequencies of the Mixture he's attacking, denoted at "M." The population allele frequencies denoted as "Pop." There's basically three cases what happened.

In first case, the genomes target small close to the mixture so there is a distant measure that is positive. The attacker cursing, it is most likely the target is in the mixture.

In the second case, the genomes attacker is equally close to the mixture and the population so that the attacker will think it's equally likely to be in the mixture and the references population.

In the last case, the genome of the target is more close to the population. The distant measure is inactive. The attacker will believe that the target is more likely to be in the reference population. Because they are millions of snips adds them up, this type of attacks shows to be very powerful.

Because of these attacks, I need to stop sharing any summary testing from the dbGaP, public website and more powerful attacks came out after this like once attack, several more attack.

However we think we that our previous attack imagine a worse case scenario where the attacker has unlimited means and resources. The decision the people are making is based on what is possible but not what is probable. We want to - out of our three models, our attacker is driven by economic incentives. We think a portion of the data can be shared with an acceptable risk level.

Beyond that we want to find the perfect balance between the sharing utility and privacy. We assume that the data shares is also driven by economic incentives. We use Stackelberg game to solve this problem.

Here's the genomic data sharing process. We have a sharer and a recipient. The sharer will decide which region of the genome to review and how much to penalize the recipient in the event of a privacy breech. For each target, the recipient have to decided whether or not to run attack. There's a cause associated with the attack. When privacy is cracked, the recipient receives a payment and the share loss it's money due to the privacy breech.

In the game for insurance strategy for the sharer, the recipient will find the best attacking strategy that maximizes his path which is the difference between his battery and the cost. After simulating the attackers offer no choice. The data sharer can choose the best sharing strategy that maximizes his payoff.

Here is some points about our experimental set up. We use data from the SPHINX project. This is his website. This is the website. Once from Genome. These are valuations settings from the experiment. This is the result. Privacy is a proportion of the successful detected individuals in the pool and the utility is the proportion of released SNPs. This is Ideal policy which is not realistic.

The payoff is a functional to the privacy. This is the result when the summary data is released without any protection. These are results of different policy for SPHINX pressure policy. These is our printing point of SPHINX pressure based on Michael Jordan's paper.

This is a result if we change the evaluation of data by penalizing the attacker for violating the contract of data use agreement. This is where our work arrives and policies end.

We asked you to trace some usually for privacy compared to do it policy. Most importantly it has higher payoff than all other policies. We can also ask have some constraints on the desire part privacy level.

For example can you high privacy level and still obtain a pair that larger than existing SNPs pressure policy. We also do a serious sensitivity analysis on some key parameters in the model. For example this sensitivity analysis is out on the penalty. This shows no matter how the payoff changes, out games are policy. Out performs or other policies.

The take home message is that blending economical, legal and technical approaches can help us balance between data utility and privacy risk. There is some limitations in our work. First our model may be too simple and in the real world, there may be multiple advisories that not driven by economic incentives.

That's it. Thank you.

[applause]

Fei: Are you assuming that it's all monetary inceptive for both sides of the game.

Zhiyu: Yes.

Fei: Is is the case when the recipient is willing to pay for the data, he has already got a plan. If he don't have plan to get leads to a positive utility, who would just not buy or not participate in the game.

Zhiyu: In our game, we assume the recipient has already has the data.

Fei: Okay

Zhiyu: This payment here is called attack.

Fei: Oh that's payment from the attack and they have the option of no action.

Zhiyu: Yeah.

Fei: Let's thank the speaker again.

[Applause from audience]

Next to talk will by Sara

Sara: Hi everybody. Today I'm going to talk to you about our application PAWS-LITE. This is done in collaboration with our partners at Penthera.

Environmental sustainability is a serious issue where natural resources all around the world are being threatened by different human activities such as encroachment, deforestation, poaching and overfishing.

As such, many different government and non governmental organizations have started initiatives and taken up different measures such as the creation of national parks and different wildlife and conservation areas in order to protect these natural resources.

However these organizations are then faced with the daunting task of having these extremely massive conservation areas to protect which could be thousands of square miles in area with extremely limited budget and limited set of resources for which to protect these areas.

The challenge is how do you efficiently use these limited resources in order to best protect these extremely large conservation areas.

Game theory has been extremely successful in addressing these challenges in the past. It is currently being used to model different security challenges all around the world such as in the forests of Malaysia, in Queen Elizabeth National Park in Uganda and the forests of Madagascar.

In all of these areas, rangers are tasked with conducting patrols throughout the conservation area in order to protect the area. Where game theory comes in, it can be used to compute intelligently randomized patrols. This prevents the patrols from being predictable to any attackers. It kind of makes the patrollers appear to be everywhere in the park.

Additionally the patrols can be computed intelligently so higher risk areas or higher valued areas are patrol more frequently than lower valued one.

A recently successful deployment of such game theoretic model has been the PAWS project of the Protection Assistant for Wildlife Security. It was first deployed in Uganda and is currently being used in Malaysia in order to optimize the patrol schedules of the rangers.

PAWS integrates different machine learning techniques within the game theoretic model in order to predict where poaching activity will be and generate optimal patrol schedules for the rangers.

It has been extremely successful and you can see here different signs of human activity and snares found during PAWS computed patrols. Although it has been extremely successful, the system is very complex as it incorporates many different domain features in order to compute these patrols.

For example, the frequent repeated attacks in these domains mean that past patrolling data and past attack data can be incorporated into these predictive models. In order to better predict where future poaching activity will be.

Additionally adversaries in these domains are bound to be rational which means essentially that they don't always do the perfect thing. The PAWS system incorporates complex behavioral models in order to better predict the activity that the adversary will perform.

Additionally, patrols in these areas are very difficult to conduct because the train is very different to navigate which means complex spatial temporal constraints need to be incorporated into the model as well. As is often the case that rangers in these areas can only patrol along ridge lines and rivers which means there needs to be route planning incorporated into the model in order to generate patrols that are actually feasible for these rangers to conduct.

However, all of this requires a significant amount of maintenance on the part of the user to maintain. Both computationally and manually collecting the data in order to input into the predictive analytic model.

In all of these, can make adoption of this type of software extremely difficult and can discourage people from using it. In response to this, we propose PAWS-LITE which is a light-weight game theoretic application. The point of this software is to encourage early adoption of these type of game theoretic models and provide immediate benefit to the users in these areas.

Additionally, as trust and use of these types of software grow, you can incorporate feedback and additional complications into the model in order to make it more realistic and more useful. We also have some preliminary field tests of this software ongoing.

In order to address this challenge, there's three main requirements that PAWS-LITE needs to satisfy. The first is that the software needs to be light weight. As many users in these area often don't have very limited computing resources. The software needs to be able to run on vary basic computers. The model also needs to be simple so that the users understand the inputs that need to go into the model and understand the outputs that they receive so that they trust the software and will continue to use it. Additionally, it needs to be flexible so that although, initially, the model may be simple, any additional domain features or challenges can be easily incorporated into the model into the future as use grows.

In order to do this. We took a simplified version of the original PAWS model. We divide the conservation area into T sectors. The defender or rangers are then tasked with protecting these T sectors with k different resources or patrollers. A single schedule assigns one patroller to one sector in the domain. Each sector has a particular value which is determined by different domain features in that area.

These features could be something like the animal distribution or the firewood distribution or distribution of valuable trees or perhaps the frequency of human activity in these areas.

We then use game theory to compute randomization over these schedules which then gives us the probability that each sector should be patrolled based on the particular target values and adversary model that we have.

Here's an example of the PAWS-LITE interface. We implemented the software into EXCEL so that it can be easily run on any laptop on very limited systems. Here you can see that the users input the number of sectors that they need to protect, the number of teams that they have and the different values of the sectors.

Solving the model then gives coverage distribution over these sectors. We also implemented a feature to generate schedules for the patrollers so they'll input the number of days that they will need to conduct a patrol for and then for each of these days, we assign a sector to each team member in the patrol. We do this by sampling from the solution distribution here.

PAWS-LITE is currently in preliminary field tests and where it's being used is working very well. At the present, the actual target values or the values of these different sectors are being collected using camera trapping data. In the future, the plan is to move towards more sophisticated ways to measure the actual values of these different targets using smart data and using the additional data that is collected as these patrols continue to be conducted.

To summarize, these game theory models are extremely useful for projecting these conservation areas. These simplified models allow for easy adoption of this type of software. It allows for the natural discovery of unique domain challenges that can be easily incorporated into these simplified models.

Also having these simple models opens easy lines of communication between these partnering NGOs and allows you to kind of open conversation and to have something that you can give them to provide them with immediate benefit and get them to trust the type of software that you're using.

Yeah that's it. Thanks.

[Applause from audience]

Speaker 23: When you compare PAWS-LITE against PAWS in terms of the quality of the schedule, are they just as good?

Sara: We don't have any direct comparisons with the two models. It's obviously not as sophisticated in terms of generating these patrols since we're not given them actual, this is where you should exactly in the forest. We're telling them this is the sector that you should be patrolling in. Where they actually go patrol in that sector will be up to them. The quality may not be the same but I think that's also kind of an additional benefit of this simplified model in that we're not immediately coming in and saying, "This is what you need to do. This is what you need to patrol." We're letting them have some control over how these patrols are being conducted. We're just suggesting like a general area. As they get to trust, kind of, our recommendations, more and more, we can be a little bit more specific, and move more towards these types of targeted patrols.

Fei: No more questions. Let's thank Sara again.

[Applause from Audience]

Last talk

Neetu: Hello everyone, my name is Neetu Pathak, today I'm presenting paper called Understanding Social Media's Take on Climate Change Through Large Scale Analysis of Target Opinion and Emotions. In short, I'm just trying to analyze how people talk about climate change.

Basically when we started this we large project, we wanted to take up a topic where we can see how general public of people show their emotions or sentiments. How they express themselves on social media and we thought of global climate change would be a really great topic for it.

What has already been done, a lot of work has been done. People predict things. They kind of study but they usually do it only on the main topic. You know, like climate change or maybe wildlife. They don't get into the subtopics of the topic and we thought if we get into the data stuff, we can understand what impact and how people behave at different levels of these subtopics.

Basically these are the questions we wanted to answer through this research paper. Can we identify climate change relevant discourse in social media? The second one is like who are the most influential people when it comes to climate change. How differently people like personal accounts and non personal accounts express their emotions when it comes to climate change. How to use the demographic like gender, age, and income influence the opinions and emotions.

The 21st Conference of United Nations Climate Change took place on the 30th November 2015 to 12 December. We thought this is a great chance for us to actually get all the data we wanted. We collected our data from first December to 31st December 2015. We had about eight million Tweets. Out of which 4.5 million of them were in English.

The classification of data. This part took a lot of time and I don't have enough time to actually get into detail of it so I'm just showing the classification that we've done. The first thing we did is we classified all the Tweets into nine categories. Five of them were taken from UN Global which has already done research in climate change which are energy, weather, economy, agriculture and water. We added four more categories, security, climate denial, air issues and animals.

Though we don't have much time, but I do want to add that these might look really simply on top of it but they are not. For instance, like security. When people talk about security they don't really talk about saving the planet earth. They also talk about how the money of each country who are participating in Paris Agreement can be used in a better way. People who are like climate deniers. They do not believe in climate change and they think when a government is using the money or investing the money to save the planet, they're kind of faking it. They worried that it can be used in improving the defense system or controlling gun violence and other things.

Each category has so many different subtopics which could be gone into more deeply but right now these are the nine topics.

Then we classified our users into person and non-person account. What I mean by non-person accounts are usually organizations and influential people who have a say in people follow them. Personal accounts are gender public. How we did that was kind of based on offense and follow ratios and other five or six most features and we made a classified while identifying them.

We got in on 13 thousand of users as influential users and we did analysis on them. The third classification we did was for each user we wanted to find what is a gender, what age category do they fall in and what is their income bracket.

Let's come to the analysis part. This is non-personal account influence. We wanted to find out how people are influencing the general public.

On my left side, you can see, those are the Twitter users. Obviously the commons ones are UNFCC, COP21, UN, and Climate Reality. Those are the users which post about climate change daily and people retweet their tweets a lot.

You can also see Barrack Obama's, Sen Sanders, and Bernie Sanders which are political leaders. We can see that they were creating a lot of buzz about climate change too.

Then we can see of the news channels as users in down left. These are the list of the influence users which are according to the link share we found.

Speaker 25: I just wanted to ask what the access was the number of times it was shared.

Neetu: Yes, this is retweets like number of times it was shared. We also wanted to do the frequency or the reassured offered but we didn't have the correct numbers of the would be values.

This is according to the link share. We see that most of the new style of them more influential than it comes to sharing of links by the people not as a user. We do see UNFCC link here but it's way down. I mean we can say that they're most successful as a Twitter user than actually as a site.

Then you are able to see that all these influential users, how they're kind of distributed throughout the world. Where we can find the most number of accounts and this is the map for that. We can see if I have to select top ten, it's USA, France, UK, Canada, Australia, Germany, Spain, Belgium, India and Netherlands.

This is actually according to accounts by location. The other thing we did was account according to the retweets and tweets. This is according to the retweets so how each influential user retweet a lot and we wanted to find out like who's retweeted the most. This is duo graph shows that. According to this, obviously USA is at top then we have Germany, France, UK, Lebanon, Canada, Mexico, India, Kenya and Belgium which is kind of similar. The only thing is Lebanon, Mexico, and Kenya kind of replaced Australia, Spain and Netherlands in this.

This is according to the number of tweets generated. We have USA, France, UK, Canada, Germany, Belgium, Australia, India and Spain which is kind of similar. We can see this is the trend among all three geo graphs.

We have our nine categories on the top and on our "y" axis, we can see six emotions and three sentiments. We wanted to see, you know, the different kind of surge we can notice from first December to 31st December. Like for instance people showed a lot of anger when they were talking about water in the 1st ten days of December. We can see that people showed a lot of - yeah. We can see in the denial, that people showed a lot of disgust towards the end of December.

We can see a lot of trends over there so I don't know what caused this surge but there's a huge difference how people are showing their emotions in different part of the month to later to different categories.

Then obviously we wanted to find out how people express their sentiments differently and influential people, organizations when they are expressing their sentiments toward topic. How exactly they're different from general public. This is the intense red is the maximum and intense blue is the least and white is in medium so if you see already lighter color it's kind of in between that.

If I go to a security, kind of what the people show fear the most which was kind of expected. For climate and climate denial, we see that personal people show a lot of disgust. We don't see a non-personal account actually - they're kind of in between neutral. That was kind of interesting. If I have to add climate denial includes messages about climate denials and people who hate them. People who are supporters. It's not like just about people who are deniers.

This is present. Next I plotted up graph which actually shows the difference the most. If you see intense red here. That means the difference of because of the non-personal accounts. If you see intense blue there, that means the difference is actually of the personal accounts so wherever you find the intense color that means that there's the most difference. We can see in climate denial, non-personal users shows surprise the most and personal users show disgust the most or this kind of shows the difference in the opinion.

The last map is actually according to the user demographic. We didn't see too much of a difference but it was still interesting. For instance like, female tends to show more emotion towards food than male.

[Laughter from Audience]

Yeah. Questions.

[Applause from Audience]

Speaker 26: I have a couple of questions. First one, what were you expecting to see?

Neetu: There are a few things. Initially when I tell myself, when I was doing my classification, I never chose security as my topic. I never even thought that young people would be talking about gun violence, or defense or anything like that. That was kind of surprising because you don't expect to see those kinds of tweets when it comes to climate change.

Second thing that was surprising to me was obviously when people were talking about food because for some reason, people were talking about going vegetarian or going green. When it comes to climate change, so I wasn't expecting that.

When I was coming up with categories, there were many things that was very surprising. At the same time, like emotions, if I go back to this one, I thought when people will be talking about animal extinction, they'll be showing sadness more but for some reason non-personal accounts shows anger but the general product actually don't show much of sadness. That was surprising because I mean as person, I'm concerned about people, animal, animal extinction but we don't see that among general public but we do see them among influential users.

There were a lot of things which were kind of different from what we expected and that's the reason we kind of did this non-personal and personal application and we wanted to see how different they are so maybe the things that we expect are actually the views of influential user not of the general public. That was something.

Speaker 26: Real quickly, two factual questions. One, I assume the distribution of the size of these topics like how many people are actually using them is not uniform, right and so I wonder how generalizable and reliable the results are given the size of the groups.

Neetu: The heat map is not based on the number of tweets.

Speaker 26: I know.

Neetu: It's based on the probability and ratios. We have kind of normalized values and then we have plotted the seed map.

Speaker 26: Right. That will take us offline but the other question was, how did you get income?

Neetu: Okay that a different part of research on paper and that classifier was made by one of mentor, [Siplana 01:17:45] she actually boiled on top of it and it's based on the user demographics and user features which you get from the Twitter grounds.

Speaker 26: Do you know how reliable it is?

Neetu: Not much. I'm sorry.

Speaker 27: Thanks for your talk. What about social bots in this context.

Neetu: Sorry.

Speaker 27: What about social bots in this context because many accounts in social media are bots driven.

Neetu: Social buts. Yes I did find them. I mean when I was collecting my data I found that one entire file was kind of created by a bot. I don't have a proper example. I try to manual eradicate the users but obviously I wasn't that great in it. I have something to think about it in the future but I'm pretty sure that might have created a little bit of bias.

Speaker 28: Just wondering how robust is your classification to sarcasm. We did some work on political sentiment on the Canadian Twitter, let's say, related to Canadian Federal elections last year and we were very surprised by that.

Neetu: I have all the sarcasm myself. When you're working on sentiments, like they create a huge problem. I mean people say that "Damn I'm so happy about it" but they're actually not happy about it. And sentiment analysis, that's the problem with - I'm not saying the classification used that they're that accurate but this is something what we did with what we already had. I'm pretty sure they have taken - this classifieds were available to me and I've used them as it was given but I'm sure that they would have taken some features in countries creating that classified.

[Applause from Audience]