Opinion Space - AMPLab Summer Retreat 2011

"Opinion Space will harness the power of connection technologies to provide a unique forum for international dialogue. This is...an opportunity to extend our engagement beyond the halls of government directly to the people of the world."
- U.S. Secretary of State Hillary Clinton (2010)

Social Media has tremendous potential for innovation and problem solving, but existing tools such as blogs, wikis, and comment lists can be quickly overwhelmed. Developed at UC Berkeley, "Opinion Space" is a new social media technology designed to help communities generate and exchange ideas about important issues and policies. The UC Berkeley AMPLab welcomed all participants of the Summer 2011 Lab Retreat to submit their opinions and perspectives on the AMPLab's research agenda.

Opinion Space is a self-organizing system that uses an intuitive graphical "map" that displays patterns, trends, and insights as they emerge and employs the wisdom of crowds to identify and highlight the most insightful ideas. The system uses a game model that incorporates techniques from deliberative polling, collaborative filtering, and multidimensional visualization. In Opinion Space, every participant chooses a "point of view" on a global opinion map. Your position is not based on geography or predetermined categories, but on similarity of opinion: those who agree on basic issues are neighbors, those who are far apart have differing opinions.

Each participant of the AMPLab Opinion Space answered the following set of 5 questions:
- I prefer Macs to PCs.
- Design by committee is deadly.
- I'm an active user of Twitter.
- Big data often trumps smart algorithms.
- With big data, privacy is dead.

Each participant also answered the following discussion question:
"What is the most compelling research question that you'd like to see AMPLab address and why?"

After entering their opinions, each participant was then projected as a point onto the Opinion Space using their answers to the first 5 statements. Participants with similar opinions were closer together on the space and those with differing opinions were further apart. Participants could then read the responses of other participants and rate them on two scales: how much they agreed with the response and how insightful they found it.

Altogether, there were 78 participants, 43 responses (32 written by UC Berkeley staff and students and 11 written by industry participants), and 760 ratings for those responses. Ideas were ranked by the ratings they received by other participants. The top rated ideas for the AMPLab's research agenda ranged from opinions on data integration, user feedback in large scale machine learning, and questions of privacy and inference surrounding the ever-increasing number of ubiquitous computing devices.

The responses of the top 20 authors of the Opinion Space are below, sorted by topic and rank. These responses can also be browsed within the Opinion Space interface at the bottom of the page.

Machine Learning
#1. How do we incorporate user corrections into large, distributed, high dimensional statistical models. By corrections, I mean direct feedback about errors: - The translation of this sentence should be X not Y - The search results for this query are missing result Z - The title of this ad is ungrammatical; add this preposition These feedback loops are tricky because no single model parameter caused the error, and per-correction retraining of large distributed models is typically not supported by our infrastructure. User corrections should be an excellent (and free) signal for data-driven applications, but they aren't today. Users don't offer corrections unless they can see the results of their changes. Note that people correct Wikipedia all the time b/c their changes go live immediately. We need that same feedback loop for large ML systems.
#20. How to strike a balance between uncertainty in real world data and response time or efficiency of learning algorithms.

Data Integration
#2. Data Integration is a huge problem that requires all three of A, M, and P. How do you take data from many different sources, all of different types, sometimes built for different purposes, and combine them in a way that can allow for meaningful queries. Without good data integration, much of the value that is hidden in our data will remain out of reach, and any bounds provided for statistical quantities on poorly matched or unmatched data will be misleading at best. Algorithms can clearly help, in terms of clustering, pattern recognition and the like. Machines can harness huge data sets and enable them to be quickly searched for similarities and relationships. Finally, people can provide judgement and common sense knowledge at a level that existing algorithmic approaches simply can't attain today.
#5. The world is awash with disparate data sources in many different formats. There is a goldmine of insight and information available with a system that allows experts in various fields to quickly synthesize and analyze that data. The AMPLab should focus on bringing robust analysis to noisy data sets, and providing a low barrier to entry for the use of the tools used.
#8. Noisy Human Data. Humans are already playing a big role in systems like recommendation systems and crowd sourcing labor markets. One problem with using humans is that when we look at a datapoint provided by a human we don't really know if they provided it properly or just randomly clicked through a couple of links. I think that is where AMP can focus. We need better statistical models to process human data and to find valuable information from the pool of noisy data. If we can somehow tell whether or not a piece of data that is recorded from human contains valuable information with no noise we can improve the

quality of our crowdsourcing tasks or recommendation systems tremendously. This problem is a good combination of psychology, statistical machine learning and algorithms. That is a big challenge and the challenge makes it an interesting research problem.
#10. Can we develop algorithms that automatically do the following: (1) Unify data from multiple data sources in a sufficiently rich way that we can perform machine learning tasks on the data (2) Make predictions using the data while trading locality (e.g. the specific situation for the prediction, such as one ad for one person/search) versus global information (say, the aggregate ad click-rates over all users given the query) (3) Automatically ask for expert (or human/crowd) advice when the algorithm does not have confidence about its predictions
#17. The real value of big-data is realized when multiple data sources are combined to allow cross-querying of these islands of data to gain new insights. How can we address the challenges associated with finding data and moving data together across boundaries with privacy and distance constraints?

Ubiquitous Computing and Privacy
#3. This isn't exactly an answer to the question that was asked, but think the AMP is not focused enough on mobile applications. There are already 5 *billion* phones in use in the world, and they are in many ways revolutionizing society in even more dramatically than the PC did. Though many of the issues with mobile data are similar to those from other forms of "big data", there are unique challenges too, including: - A huge variety of unique privacy challegnes - Complex inference and de-noising problmes that arise from location and other sensor data - The ability to to ask and infer things about the real world, including who interacts with whom, what someone is doing, whether someone is healthy or sick, etc.
#4. Right now a lot of the crowdsourcing work in the AMP Lab seems to be focusing on explicit mechanical turk style integration of people. In addition to that, I think there is a lot of promise in utilizing people as implicit sensors (a la GPS data from phones being used for traffic prediction) or data labelers (gmail priority inbox). Interesting research problems include: * protecting privacy so people will be willing to contribute more behavioral data * the ability to quickly join disparate data sources and draw conclusions from dirty data * active learning, allowing the system to know when to bother users for more information and when passive collection is adequate * context aware, pervasive collection of user activity
#6. Making analyzing big data accessible to small users (e.g. not Google, FB, Microsoft sized companies). This means better infrastructure, programming models, and analysis packages that make it easy for unsophisticated users to crunch tera to petabytes of data.
#11. I would love to see more discussions how the vast amount of sensor and devices entering the market may impact the way we today look at big data. What additional challenges does that bring, in terms of data collection, algorithms, privacy etc. What are the oppertunities and how much data do we foresee? I guess this is very compelling since this will create almost a new paradigm shift. We need to not only address a distributed system, multitude of users and complex analysis which to different levels been addressed for years. The future will also bring diverse set of devices that has different possibilities to handle data, but should in the same time be involved in the same applications that ultimately will generate the desired knowledge.
#19. Data is important, but perhaps we could start a discussion about metadata management as well. Without it, especially in sensor deployments, interpretation of the data is impossible. For Amplab, this may becme more relevant if the question of mobility comes into play. Mobile user context is necessary for interpretting the data coming from the phone for the user in that context. Is this metadata management problem fundamentally different from data management? How is it similar/different? Does it matter in other contexts?

Crowdsourcing
#7. Is there a way to effectively allow interaction and communication amoung crowd-sourced employees to answer non-trivial tasks? Humans tend to work well when they are able to communicate, and ideally all workers would work towards a task. Allowing them to communicate ought to improve the quality of their answers. Ideally, we would not need to directly movtivate them to communicate, such as via an incentive to post drivel on a forum. Another issue is avoiding the effects of group think and personalities dominating discussion.
#9. What are the limits of computer systems for enabling new approaches (abstractions) to human powered computing? E.g. Can Mesos manage people in addition to CPUs, ram, etc.? I would like to see a deep exploration of the interplay between _Machines_ and _People_, in particular how can computer systems support human computation in fundamentally new ways. I believe that crowd-sourcing, as it is realized today, is the rough equivalent of punch cards on time-shared computers with human operators used in the early days of computing.
#12. Cloud computing and crowdsourcing are both great pay-as-you-go resources. The AMP lab should investigate how to most effectively use the cloud and the crowd to answer data analysis queries. This effort includes determining which types of queries are better suited for crowd, cloud, or crowd+cloud, as well as deciding how a user's expectations of answer quality and response time translate to a particular execution plan using these pay-as-you-go resources. The crowd+cloud angle really captures what makes the AMP lab different than other distributed systems or crowdsourcing projects.
#13. What is the killer app that will emerge from the trend towards big, unstructured data and cloud computing, and why is the combination of machine learning and crowd sourcing necessary to realize the killer app? Relying on big, noisy data and the nature of crowds makes it hard to build a predictable system, so the benefits must outweigh the challenges of unpredictability.
#14. How can you guarantee, in crowd sourcing experiments, that your user sampling is sufficiently random? Other sciences spend a lot of tiime and energy on this in order to minimize biased results. Does having your entire user population computer literate and willing to perform microtasks for micropayments, and possibly in developing regions, bias a study?
#15. Leverage "wisdom of crowds" to make predictions. Wikipedia collects expert knowledge but it is not directed towards predictions or decision making. The challenge is to move beyond simple rating systems towards methods that leverage crowd knowledge for broader predicitons or decision making. For example, individual ecomonists are often asked to predict things like the price of oil next year. Each has a different domain of expertise. One might be an expert in exploration, one in new processes (like with shale oil), one in global conflict, on in alternative energies. A system that gathered their knowledge and made predictions would be better than one that asked each to extrapolate and averaged their answers.
#16. There have been many remarkable successes using the crowd for what could be called matters of common sense (eg, labeling images, mining address info from business cards) that are still difficult for algorithms. There are also many areas that

are difficult for algorithms and thus attractive as an application for crowdsourcing; however, they are in the realm of expert knowledge (eg, medical diagnosis, debugging). The AMP Lab should address the following: Can you approach expert performance using a crowd of semi-informed people? How do you know (1) how much knowledge these semi-informed people need, and (2) how many of them you need?
#18. For what kinds of data and applications is it useful to have crowds of people analyzing the data in a symbiotic manner with algorithms? When is it helpful and when does it hurt or slow progress? When is it more useful just to have a small group of experts or even just machines work on a problem? Can we identify various features of applicatons and data that would allow us to automatically figure this out? As big data and crowdsourcing platforms become more common, I think it will be very important to know what the strengths and limitations of the various approaches will be and identifying varius paradigms will be extremely useful.

For more information:
Visit opinion.berkeley.edu or contact:
Ken Goldberg
IEOR and EECS, College of Engineering and School of Information
425 Sutardja Dai Hall, Berkeley, CA 94720-1758
(510) 643-9565 http://goldberg.berkeley.edu
goldberg@berkeley.edu