

Microsoft Research Faculty Summit

Rick Rashid

Redmond, Washington

July 13, 2010

RICH DEMILLO: Thanks. Good morning, everyone. As an outsider, I can say this has been a great summit. Can we just give Tom McMail and his team a round of applause? (Applause.) A really great meeting with a good program.

So welcome to the panel. Let me tell you what we're going to do this morning. I'm going to introduce the panelists. I'm going to ask them to say some brief, provocative things to get your juices flowing and to get an interactive discussion started around this topic of transformation in research. These will be short presentations with no PowerPoint just to sort of set the context. I can't think of a better person to set the context for this panel than Microsoft's head of research, Rick Rashid. So, I'm going to invite Rick up to the stage. (Applause.)

RICK RASHID: Should we bring everybody else in first?

RICH DEMILLO: Maybe we should do that. Let's bring Ed Lazowska from the University of Washington. (Applause.) Tony Hey from Microsoft Research. (Applause.) Wolfgang Gentzsch, from the DEISA project. (Applause.)

So, Rick, transformation and research.

RICK RASHID: Well, you know, one of the things that first comes to my mind when I think about this topic is the fact that, you know, I think especially over the last five, six, seven years there's been a really significant change in sort of the way computer science relates to a lot of the other sciences. Increasingly, there's an inner penetration between the field of computer science and virtually every other science, and actually many of the social sciences now as well. And that's really changed the way I think people do their work, but it's also changed the way people in other sciences think about the field and the way they do their work.

So to me, that's been probably one of the most significant things. I mean, my joke always about this is that, you know, when I first started at Microsoft Research, it would never have occurred to me that we'd be publishing papers in Science or Nature or the Journal of Medicine, I mean, those are the places that you didn't think that computer scientists would be involved with. And yet, that's really what we're doing today. So, I think that's been a significant change.

I think on the methodology side of the line, there's been this, again, and these are related, a tremendous change in the way we think about doing our work. We're much more data driven now.

I was talking with Dean Randy Bryant, Carnegie Mellon, yesterday and we were talking about how when we were both young professors at Carnegie Mellon, by comparison today, it felt like

we were sort of playing around. I mean, we didn't have really a very disciplined way of thinking about a lot of the work that we were doing.

If you were doing work in AI, you know, AI in those days felt more synonymous with heuristic. You know, people were trying things, they were experimenting. And when you got a good result, you felt happy, but you didn't always understand why you got a good result, and the results were often very fragile.

We've moved into a world where now we can collect astronomical amounts of data. We understand how to process that information and we can do it on hundreds of thousands of processors simultaneously, and that's really changed, I think, the way we think about the way we do our work and the way we do science.

On the community side, you know, I think the biggest change there has been the fact that we can now have really global organizations. We can now interact in an environment where research can be conducted and shared simultaneously among many groups around the world.

I know when I came to Microsoft Research originally, I felt like, wow, it was really important to have everybody in one place because my experience at Carnegie Mellon was that, you know, people worked really well when they were all in exactly one place and they could talk to each other. In fact, ideally, they needed to be in the same floor, and ideally, they needed to be in the offices next to each other.

That's really not true anymore. I mean, I know my organization is completely global. We have more people in Microsoft Research working outside the United States than inside, and for me, running it, it doesn't really feel like there's—that everybody's distributed. I mean, we interact, we're constantly working together, we're exchanging information, and there's really been a tremendous change there.

So that's my cut at it.

RICH DEMILLO: Ed?

ED LAZOWSKA: So I'll look at sort of a three-part breakdown too, but it's going to be a slightly different three parts, and this is the result of talking with a lot of people recently about these questions.

The first thing I try and think about is what's changing in the world, all right? What are sort of the technological and societal shifts and drivers? And there are a lot of them, and here are a few: One of them is ubiquitous connectivity and total mobility, all right? So I used to be chained to a desk, and now I'm connected everywhere.

A second is huge computational capacity available to everybody through the cloud. So, it used to be that if you had heavy duty work to do, you had to deal with the cathedral at San Diego or the cathedral at Illinois, and now it's just there for everyone.

The third that Rick touched on are these exponentially increasing data volumes from simulations, but more importantly from sensors—low-bandwidth sensors that are cheap everywhere, and high-bandwidth sensors, telescope, CCD arrays, and gene sequencers that have much greater density than they did.

A fourth that affects all of us is the end of single processor performance increases. So, the need to deal with parallelism even in sort of Office apps and sort of desktop apps.

A fifth is AI technologies entering the mainstream, right? So we've been talking for decades about natural language processing and speech and robotics and all of these things are finally entering primetime data mining and machine learning as well.

Another is social computing. And what I mean by social computing is partly Facebook-type apps, but partly crowd-sourced intelligence, crowd-sourced knowledge, the ability to get crowds to contribute to solving problems, the whole Luis Von Ahn sort of computers plus people solving problems.

And finally, the fact that all the world's transactions are online, right? And that's, again, another source of data. So, those are the changes and now what does that suggest about thinking about the field? When I think about the field going forward, I think about four main pieces: One is it's the computers and people piece by which, again, I don't mean traditional HCI, not that everybody does—anybody does traditional HCI these days, but I mean all the aspects of sort of a social and societal interactions and crowd sourcing and group problem solving and the impacts of technology on individuals.

A second is the interface of computers to the physical world. So, in my view, this goes back at least to David Tennenhouse 10 years ago at DARPA saying, "Get real, get physical." All right? So it's sensors and effectors and models.

The third is dealing with this deluge of data and the opportunity to analyze the data and move it into sort of knowledge and decisions and actions.

The fourth is the whole question of making systems scalable and trustworthy and secure. So, that's how I think about the pieces of the field going forward. And now the question is: What about research agendas? And this is going to sound I think a lot like what Rick said, so I'm worried we have a little too much commonality here, but in my view, it's more than science, it's hitching what we do, driving what we do by a set of societal challenges, which are really important these days. We've got energy, we've got transportation, we've got education, we've got healthcare and advances in computer science are necessary to tackling all of those challenges.

And if I think about energy and transportation, for example, it's a great example because it draws on all four of those cores I've identified. You've got to worry about the interface to humans. You've got to worry about the interface to sensors and the physical world. You've got to worry about exploiting enormous amounts of data, whether it's managing the grid of managing the modeling of the behavior of your home and its resonance. And you've got to deal with security and privacy.

So it seems to me that to drive us forward, we can hitch ourselves to what everybody acknowledges as really, really important national, societal, global challenges in the same way that maybe 20 or 30 years ago we honestly hitched ourselves or were hitched to high-end computing. Not that that was the only thing going on, but that that was something that a group of people understood is important and that motivated an entire research agenda for the field.

The final point I'd make is that letting your research be motivated by the problems that actual people have does not mean your research has a short time horizon. It just means that you're trying to solve problems and trying to relate to your work to things that a broader base than those of us in this room care about. (Applause.)

RICH DEMILLO: Tony Hey is going to be controversial and disagree with everything.

TONY HEY: Well, I'm going to ignore the sub bullet on the first one, research content. So, for me, research content—we're in the midst not only of a scientific evolution driven by data, but we're also in the midst of a scholarly communication evolution because, increasingly, besides text, you need data, you need software, you need images, they all need to be linked. You need the ability to do sort of scientific mash-ups where you take this data set, that data set, and you combine it with your data set to create new knowledge.

And in addition to that, you'll have work flows which you'll save and exchange and use. You'll have blogs, you'll have wikis, you'll have all this sort of Web 2.0 tagging and social networking available to science literature. And it really is a game-changer when you can make period of time digital copies for nothing and you have the Web to disseminate, you don't need a printing press. And that's why there's a problem for the publishers and that's why there's currently an issue between Nature and the University of California libraries about the budget increases and the California libraries have not had a large budget increase.

So I think there's a really, really interesting transformation in scholarly publishing and how it's being led, for example, by NIH with the National Library of Medicine, open access is a part, what's the future of research libraries? Each university should have an institutional repository to display its research. If you claim you're the best university, well, people should be reading or downloading your papers. And if they aren't, why not?

So these are things that I think are inevitable and going to come and they're really going to be transformative. And so it's one of the sessions that is really interesting for me is the sessions on the data challenge, which is actually for the information sciences, library publishing community. And it's a very interesting theme going through this conference. So, that's how I interpret the search content.

Research methodology, well, we wrote a book on it, all right? So the fourth paradigm was Jim Gray's way of expressing the fact that we really are seeing a sea change in the way the scientists have to do their research. The examples I like to give are similar to Ed's, the fact that the human genome, the gene sequencing devices now, huge amounts of data. David Heckerman is rubbing his hands in glee because he sees biology becoming the province of machine learning experts

like him, and he's really doing some really exciting stuff with HIV/AIDS and other major diseases threatening the world.

The other example I would pick up is one, again, that's dear to me. It seems a good idea since we live in Seattle to instrument the earthquake plate, the Juan de Fuca plate just off Puget Sound, and instead of a boat steaming across the surface taking occasional data points when it's out there, the transformation of ocean science will be that this data will be streaming in 24 hours a day, seven days a week, 365 days a year. And suddenly from being data poor, they're going to data rich overnight, as John Delaney says, and that's really a transformation, and I think they will need help in understanding how to manipulate, how to move, how to analyze, how to process, how to visualize all these sorts of data and data mine it and machine learning techniques, I think I agree, will be coming to the mainstream.

So I think there really is, in addition to the traditional experiment theory and computation, I think data-intensive science involves a new set of skills, and I think it's a unique opportunity for the computer science community to make a real difference in solving problems people care about.

Research Community. I sit on the advisory committee for cyber infrastructure for NSF and I think they've understood the key point. In the U.K. when we did e-science, you know, putting together computer scientists with scientists and of course they didn't understand each other for the first six months, scientists had this strange idea that computer scientists wrote programs and computer scientists thought that they had this cool tool which solved all the scientists' problems, and neither was correct, and it took them at least six months to understand that and I was able to bribe them to be together by they only got the money if they talked to each other.

But actually, what I think is needed is a third community coming together. So, multi-disciplinary is fine. Science and computer science have really things to do, but I think you also need people to take the tools and research prototypes which demonstrate how these cool techniques from computer science can really help to tools that are robust enough for scientists to use in their everyday life. So, I think there are three communities involved: The scientists, the computer scientists, and the engineers, software engineers who can actually build systems that are reliable and people can really use.

The last thing I would like to pick up, Ed's technology, I absolutely agree with him, the multi-core revolution is the end of Moore's Law meaning that things get faster automatically. We have no the problems of multi-core chips and so on. We have the cloud, which is going to transform everything, and we have sensors becoming cheap and ubiquitous, RFIDs, and everything.

So I particularly liked Butler Lampson's categorization of the phases of computing. The first 25 years it was all about computation is simulation doing calculations and so on, the second phase, 25 years from mid '80s onwards was all about communication and the Internet and the Web and so on. And the third phase, which we're just now entering, he calls "embodiment." It's a way computers are really integral to our life and they act on our behalf, they take care of our cars. He had a scenario where you could prevent cars crashing—I forget what he called it, crashless driving, but Ed had a much better way, called it reckless driving, which I thought was a very good one.

So I think the embodiment, and I think this is really a challenge for the computer science community and it's a great challenge and I think we can also help the scientists solve some of the problems facing the world today. So, I'm full of optimism for the future.

RICH DEMILLO: Great, thank you. Wolfgang?

WOLFGANG GENTZSCH: Yeah, thank you very much. And to your disappointment, there's nothing really controversial I can say.

RICH DEMILLO: Uh-oh, we'll change that.

WOLFGANG GENTZSCH: Especially when you—like I think many of us—have lived through the last 35 to 40 years, starting with linear partial differential equations in the late '60s. My thesis was about the mathematics of bending of a plate. Fourth order, partial differential equation system, two equations. And we learned how to make these things ready then, finally, for the computer, using languages like ALGOL and Fortran, and then C. So, we all—I think—we lived through this evolution.

For me, there is basically no surprise, and I don't expect any surprise at all in the future, it's all evolutionary. I mean, at one point in time, we had so many new tools that then there was a certain breakthrough, but we have seen this already in the past, several times, so we expect a couple of similar great breakthroughs in the near future as well.

Our research content is basically still the same. We want to solve the grand challenge, big science problems in many different areas like astrophysics, climate, and earth systems or geophysics, particle physics, nanosciences, material sciences, you name it. This, for me at least, hasn't changed. What has changed is certainly the accuracy of the results and the power of prediction we now have in many of these areas, based on the tools we are now able to use.

But, you know, this went on for the past 40 years, at least. I mean, I can only talk about the last 40 years. The challenges, the applications we were tackling are always progressing, they went hand in hand with the tools, therefore improvement of one side kind of stimulated improvement on the other side, and vice versa. So, the problems got bigger and the challenges got bigger, but also the tools got more powerful.

Now the methodology. As I mentioned before, I have worked on science applications like elasticity theory, plasma physics, and CFD. And when you do that, like many of us in the room here, you figure out that you always need more and better tools.

Over the last years, we have built great cyberinfrastructures, e-infrastructures, with all the different components we have mentioned. Like the peta-scale machines, and in about eight years, the exa-scale machines, being compute nodes in these cyberinfrastructures, plus the data nodes, as well as the fourth paradigm Tony mentioned.

In addition, we have the grids and the clouds, so we have a very broad spectrum of tools. When we started in the early days, we often had just one tool, which was a mainframe. Over time, then, in the

late '70s, early '80s, we had the choice of several powerful supercomputers, monolithic ones, the vector and the parallel machines.

And these days, when you enter cyberinfrastructure, we see a large variety of very different tools now which allow us to map our applications, or our workflow, which consists of several applications, because in many multi-physics problems, the components interact very strongly and tightly, but you can map these different components (or nodes of a workflow) now onto different systems in the cyberinfrastructure.

So that's what has changed, which however means that the complexity has increased, too. Now, we are looking forward for simplifications in that space. And one of those is making IT, making computing and data processing a utility. This won't work for every application because some applications really are in need for hands-on fine tuning, when you want to map a complex workflow to the underlying available architectures of heterogeneous compute nodes.

It is not that clouds will replace grids. In my humble opinion, this is not correct. For the time being, with the complexity of our applications and workflows, we still need the grid for the researcher who needs to access the individual nodes to customize and optimize them according to the specific needs of the application.

On the other hand, there are many applications, you know, like the parameter-space type of applications: you have just one and the same application, but you have many parameters, like in automotive crash simulation where you are trying different materials, different geometries, different physics, et cetera, but it's always the same application and 500 different parameters resulting in 500 jobs which you can submit into the cloud, in parallel.

So the cloud is now a very nice additional tool for researchers which might be somewhat disruptive, Rich, if I may, which is taking us to a new paradigm, which is Science as a Service. Thinking outside the box, and using as much external services as possible to solve your problem. At the end of the day, it is not just using a utility, but you combine your grid-based workflow wherever you can with a cloud node. We are already seeing this hybrid approach these days. And nobody really knows how far we can go with the cloud. Currently it is an interesting research area, how HPC fits into the cloud, because of potential bottlenecks coming from the virtualization layer and the interconnects between the cloud servers (in the form of high latency and low bandwidth). Also, clouds hide the complexity nicely, but they also hide the system details underneath which you might want to have access to for performance optimization, for example.

One more thing is coming to our community: the big problems are tackled these days by virtual organizations, VOs. In the early 2000s, for example, the BIRN project started, which is a biomedical informatics virtual organization. Also NEESgrid, a virtual earthquake engineering organization, and many more virtual organizations. And then, we have the more generic grids, like Teragrid, or DEISA, the Distributed European Infrastructure for Supercomputing Applications, combining 15 of the largest supercomputer centers in Europe into one virtual supercomputer center, donating hundreds of millions of CPU core hours to research simulations every year within the DEISA Extreme Computing Initiative.

Let me close with one remark which I think is extremely important. We take it for a given, that all these young people here are working on real challenging problems. But, we often forget, especially

governments forget, that we need more of these young and well-educated people, educated in sciences and engineering.

In Europe, a famous report from a commission lead by Professor Gago in 2005 predicted that in 2010 Europe will need 600,000 scientists and engineers more than we had then. Unfortunately, nothing happened so far. For me, it's too late to start in the universities to improve the educational system, especially in the sciences. We have to start in the sixth to eighth grade already to excite our kids for sciences and engineering.

And that's what I see as another trend coming, more and more with the cloud. I mean, you will have an education science cloud that our kids access from their classroom or from home. They will access simplified and nice scientific and engineering problems, do simulations interactively, embedded in a Web 2.0 Collaboratory, and sharing their own science examples, experiments, et cetera, with their peers.

This will help to move the more frontal education of today then towards a more creative, interactive and highly motivating science education, which then creates more students studying sciences and engineering, leading to more scientists and researchers and engineers jointly tackling all our grand challenges.

PARTICIPANT: Okay, I'm prepared to be controversial.

RICH DEMILLO: Let me do my—

PARTICIPANT: That would never happen, Ed.

ED LAZOWSKA: No, because Rich is stopping me.

RICH DEMILLO: Let me do my controversial remarks first, and then we'll flip it over. I had the advantage of seeing the comments from the panel yesterday and to think for a day or so about what I was going to add to these points, which I agree with. And a couple things occur to me. We're used to, in technology, thinking about technology curves. And technology curves are all about change and we can precisely characterize the rate of change and what it means. It's rarely that change by itself leads to a transformation.

I went back to Bob Metcalf's original 1980-'81 talk where he presented what's called the Metcalf's Law, the value of a network is proportional to the square of the elements of the network, or exponential, depending on what religion you belong to. But that actually was not Metcalf's Law. Metcalf's Law was a critical mass crossover principle. Metcalf's Law was as soon as the value of the interaction has exceeded the value of the components, something transformational happens.

And as I think about the technology curves in computing, the curves by themselves are very incremental. They tell us what's going to happen over the next year, over the next 18 months. It's not until you get curves intersecting, you get a critical mass crossover principle working that you really have the opportunity for transformation.

I'm thinking here about some examples like biodiversity. So, we heard in some of the sessions yesterday about the great work being done in Brazil on biodiversity and the worldwide information facility. That's not about Moore's Law, that's not even about data size increase. It's about the crossover that you get when you get enough computing power, enough genomic data populating genomic databases, and enough observations available at the same time. So, that's a transformational principle.

Robotics has gone through a transformation. Robotics is a crossover of computing power, programming capability, vision, and low-cost, off-the-shelf assemblies. You get transformations that way.

So I think there's a tendency, particularly in computer science, to take a look at how great we're doing as we march along these curves and to not step back and say, well, you know what? All the action is what happens when these curves cross each other at the right time and provide things that are just destructive of the way we used to do things in the past. And that strikes me as transformational.

The second thing I wanted to mention is this idea of community, and a couple people pointed out the fact that we have sort of a new interaction paradigm in social networking that I don't think has been really exploited in a transformational way. And, again, this is not across the board. It's easy to imagine research areas, research problems where transformations don't occur by getting enough people together. On the other hand, the classification of simple groups took place over 40 years, involved tens of thousands of mathematicians. In retrospect, had the data been mined from group theory very quickly, the classification program could have been completed very, very quickly.

And you wonder about problems like $P=MP$. If we had a massively open project to resolve P and P , in which we could mine data that goes back 200 years, mathematical data that goes back 200 years, reject crackpot participation, embrace participation by not 100 mathematicians, but 10,000; a million mathematicians, is that a new paradigm for conducting research? That strikes me as being transformational. And without the technologies, it's hard to imagine how that would be done.

Let me just finish with a word that I've been torturing these guys with all morning: ephemeralization. Who knows what ephemeralization is? Nobody knows. Who knows what dematerialization is?

PARTICIPANT: Perfect.

RICH DEMILLO: So it's an economic principle which basically means doing more with less. Dematerialization is the way that economists talk about—Buckminster Fuller had this wonderful term called "ephemeralization." And ephemeralization has been on my mind because I've been going back over some software tools that I wrote 20 years ago in trying to update them for virtual machines. And it turns out that we spent an enormous amount of time—our graphics display was a blit. So, if you remember the dual-ported, green-screen monster from Bell Labs that was

impossible to program, we spent an enormous amount of time laying bits on this green screen. And we don't have to do that today.

And if you think about what Wolfgang was talking about, sort of taking this infrastructure, this tool layer and using it to wrap up conceptually a lot of stuff that we used to spend energy on so that energy can be spent somewhere else, that's a form of ephemeralization.

You know, I work with world-class mathematicians, they use Maple. You know, a lot of symbolic computation takes place on the machine that used to take place on scraps of paper and napkins. So, where does that time go? Does it go towards problem-solving? Does it go towards rearranging your office, as Rick would suggest, or doing other non-productive things that people need to think about when they solve problems?

So I don't know how to break those three ideas apart into content, methodology, and community, but they all strike me as things that are suggestive of transformations, not simply incremental improvements along already-defined technology curves.

So with that, I'm going to throw the floor open to, I guess, Ed and the audience to have at it.

ED LAZOWSKA: So Wolfgang and Tony will have to forgive me, and you'll have to realize that I don't necessarily believe what I'm saying, but here's what I'm going to say: (Laughter.) I think we have to—

(Crosstalk.)

PARTICIPANT: We know this about you, Ed.

ED LAZOWSKA: We have to stop focusing on science and we have to stop focusing on peta-scale and exo-scale and whatever the next obscene word is, okay? We focus on those things because we can measure them, it's like the drunk looking for his keys under the lamppost because there's a light there, okay?

What's going to get more people into our field is the fact that we can change individuals' lives, all right? And when you focus on science (applause) what you do is change the lives of a few thousand scientists, and indirectly of many more people, right? But it seems to me that that has been the horse we've ridden for 40 or 50 years in this field, and it's not like it's a dead horse, it's rather than it's no longer the only horse, all right?

So what I want is for people's bodies to be instrumented as well as their cars so that they understand their health situation. I want people to be able to live at home longer. I want automobiles not only to drive wreck-lessly, but I want their utilization to increase. And I want public transit to be better able to be used by people because you know where the vehicles are and because they route around congestion properly. I want your home to be instrumented in a way that you understand your water and your gas and your power consumption, and your home is intelligent and it's managing those things for you and it has a model of how its systems behave and how its occupants behave and on and on and on.

And I think these are the new horses for us to ride, and they are things that change people's lives and that folks can relate to. And we have to give it up on the exo-scale and the peta-scale, and we have to give it up on being solely focused on helping the chemists and physicists and astronomers, and we have to start helping people. And that's the power of our field and that's what will attract people to our field. So, let's ride these other horses and let the other one just kind of keep moving along with it.

WOLFGANG GENTZSCH: Just a little one which is, I agree—

ED LAZOWSKA: And let's admit that the grid was a fiasco and a failure.

WOLFGANG GENTZSCH: Not true. Oh, no, no, no. (Laughter.)

ED LAZOWSKA: Waste of money.

WOLFGANG GENTZSCH: There are several hundred of focus groups in action which are the basis for—

ED LAZOWSKA: Remember, I don't believe what I'm saying.

WOLFGANG GENTZSCH: —virtual organizations—

ED LAZOWSKA: Thought I'd be clear on that.

WOLFGANG GENTZSCH: I forgot that one, you warned us before. But I want to say peta-scale and exo-scale is really a small part of the big picture.

ED LAZOWSKA: Right. Right.

WOLFGANG GENTZSCH: And—

ED LAZOWSKA: And there's an obsession because you can count it and because you can say, oh my God, the Chinese exo-scale is bigger than our exo-scale.

WOLFGANG GENTZSCH: Yeah, that worries me.

ED LAZOWSKA: Right? And who cares? It doesn't matter.

WOLFGANG GENTZSCH: I never mentioned peta-scale or exo-scale.

ED LAZOWSKA: You didn't.

PARTICIPANT: I did, I'm sorry.

TONY HEY: And I agree with most of your themes, but David Heckerman trying to solve HIV and trying to solve diabetes and things, these are problems that people will care about.

I gave the oceanography example, but that's very similar to—we have a project with John McGee looking at disaster response, how you coordinate all the different agencies and so on, that's a real computer science challenge.

ED LAZOWSKA: Yeah, Tony and I worked together on these oceanography problems.

TONY HEY: What I do disagree with is Rich, all right? I actually don't agree that mass collaboration always produces more creativity. In my 40 years in academia, I have found that the really, really creative people are rather rare. And actually just putting large numbers of extra people just increases the noise. And so I actually really think that creative people are rare and you don't get more creativity necessarily by adding thousands of extra people to the mix.

ED LAZOWSKA: Someone introduced me recently to a two-year-old comment by Stephen Colbert on this thing he called "wiki-ality" you familiar with that? Wikiality is—

TONY HEY: That's sort of like wikinomics—

ED LAZOWSKA: It's the reality of the crowd. Okay? So the idea is if you can convince enough people that something is true, then it's true. And he got the Colbert Nation to do things like hack Wikipedia so it said that the population of elephants had tripled in the past three years, and he would demonstrate on the air that this must be true because it's in Wikipedia.

RICH DEMILLO: On the other hand, I agree it's not always true, on the other hand, I mean, there have been pretty careful analyses of these open mathematical questions that were resolved in the last generation where the discovery rate of new results sort of doubled every generation. And had there been tools to mine that data or to translate that data, you can imagine that the discovery time would have been compressed.

And, you know, time is not the only dimension to compress. You can compress over the population. So, if you get enough people agreeing on a direction to go because they have evidence, that's a line of research.

I don't know how to duplicate that in the small group. I don't know how to duplicate that with pencil and paper. I can imagine how to duplicate it in a massively open project, massively open online project.

(Crosstalk.)

RICH DEMILLO: I don't think those have been tried to any useful extent.

PARTICIPANT: One of the things we haven't talked about, by the way, is transformations about research universities which will be implied by this. And I wondered whether you had—there might be a book there, perhaps.

PARTICIPANT: There might be. There might be. (Laughter.)

PARTICIPANT: For pretty much the same reasons.

RICK RASHID: But the other thing that has been historically an issue is the fact that, historically, a lot of scientists, even a lot of computer scientists, you know, have actually not been particularly open with their data. Right?

PARTICIPANT: Yes.

(Crosstalk.)

RICK RASHID: I was sitting through a physics discussion six or seven months ago, you know, people were trying to understand how a particular set of experiments was getting a particular result. And somebody was pointing out, well, depending on how you process that data, you could have mapped almost any function to the data, right? Just depended on where you started the sequence and where you decided to end it and what you did for the transform.

And the particular scientist involved wouldn't give the data to anybody else. So, there was no way for anybody to know what the answer really was. And I think one of the issues that you can get into, and this can even happen in crowd sourcing if you're not careful, is that unless you provide people to access to sort of the underlying information, right, it's really easy for lots of people to look at the wrong data or inadequate amounts of information and start running off in the completely wrong direction.

RICH DEMILLO: Which suggests a bunch of science to be done to figure out how to do that.

RICK RASHID: Well, it also suggests a change in—going back to research methodology, research methodology, which is starting to say, hey, if you really don't put your information online, you know, we don't care about your paper.

ED LAZOWSKA: So another aspect of that is a computational biologist described to me a few weeks ago the notion that it used to be that there was a lot of lab work and a relatively modest amount of data analysis, because there was a relatively modest amount of data. And now there's an enormous amount of data analysis, and in many ways, the analysis of the data is directing the lab work, right?

So in some sense, the lab folks who used to drive the discovery are no longer driving the discovery, it's the analysts who are driving the discovery, and I think that's an important notion because those analysts are going to need training in biology and training in data analysis, statistics, computer science as well.

RICK RASHID: I think it's kind of hard to imagine now someone really getting a degree in computer science without a strong statistics background.

ED LAZOWSKA: Right. Right.

RICK RASHID: I mean, I think that's something that—I know a number of universities are changing their curricula specifically to put statistics into computer science as a part of the curriculum, but I think it's a critical part now.

ED LAZOWSKA: Sure.

TONY HEY: One thing on the open data, I was at a conference in Helsinki where WorldwideScience.org, which is about—I think it's 70 nations have put their databases online, it's led by the Department of Energy's Center for Technical and Scientific Information, which puts up DOE data online. And what we were there for was to celebrate the automate translation using the world of Bill Roland and his machine translation team. And it was really interesting. So, you could actually now put queries in and you could search Chinese databases which were not translated or Russian and so on.

And they had a very interesting statistic—by making this data open, if you did a Google search—I'm sorry I said Google, not Bing, but if you did a Google search—

PARTICIPANT: But we know he meant Bing.

PARTICIPANT: Where's Harry? (Laughter.)

TONY HEY: You found 3.5 percent, and 96.5 percent of the results from doing this federated search across WorldwideScience were unique and not covered, found by Google, and that was really, for me, very interesting. So, there is a lot of things about making the data available and searchable.

RICH DEMILLO: So you mentioned universities. I think one of the trends that's worth talking about here is the way that universities are measured globally for research productivity and quality. I mean, the Shanghai ranking is the most incremental, you know, publish one more paper ranking of university research output—

TONY HEY: You get 20 points for a Nobel Prize winner.

RICH DEMILLO: Yeah, if you know a Nobel Prize winner. (Laughter.) If one has ever visited your campus, but those kinds of things I think are not aligned well with what you're talking about. So, as long as there's not a more fundamental way of saying who's doing what or agreeing it doesn't matter, universities are going to be pulled towards hanging onto IP, hanging onto data, you know, hanging onto things that would make it easy for people to collaborate because they look at it as a zero-sum game.

ED LAZOWSKA: Let's try some audience questions.

RICH DEMILLO: Please, if there are any.

WOLFGANG GENTZSCH: Just a few, yes.

(Break for direction.)

PARTICIPANT: In regard to the transformation of the research university, could you comment on what you think the roles of libraries, research university libraries might be in helping to archive or access these massive amounts of data?

RICH DEMILLO: Let me jump in.

TONY HEY: I'm happy to say something after, Rich.

RICH DEMILLO: The role of a university library has just changed—they're unrecognizable from what I remember a library being when I was a student. If you just look at the archive requirements that are levied on particularly public universities, all that stuff goes into libraries. And those aren't books, those aren't periodicals, those aren't things that we think of as part of the normal library function librarians do. And they have no idea what to do with it.

If you had to pick one research topic for big data, you know, I would pick libraries, I would pick university libraries because their technology curve is also growing exponentially. The number of tools that they have is sort of fixed artificially by budgets. And aside from very simple search and storage tools, there really is not an awful lot that computer scientists have said about what these things should look like.

TONY HEY: Okay, so, I mean, I feel obliged to speak because my wife is a librarian, right? My engineering students when I was dean never went into the library for, quote, library purposes, they went there because they had a warm place to work, meet their friends, have a cup of coffee, doing wi-fi. And that's really confusing a library with Starbucks, all right? And really the function of a library at a research university I think should be the guardian of the intellectual output of the university and it should be actually helping the university maintain its reputation.

And if you don't like the Shanghai criteria, and for the computer science community, I always used to have problems with my other engineers who looked down on us because we don't publish in proper journals, we do conferences and workshops. We should fix that. We should actually make sure we have something which is really—you look outside and you see Microsoft Academic Search. Why doesn't the computer science community arrange that so it actually gives the results so you can get sensible results about ranking universities or computer science departments?

So I think the libraries have a big role to play with different communities. I think they have to work with communities, they have to be less isolated, they can't expect people will necessarily come to them, but I think—I passionately believe that research universities should have a library, but it may not look very much like the libraries we have at the moment.

RICK RASHID: I'll just raise the question there because—why should universities have a library? I mean, you know, this engine in question—and I'm not anti-librarian (laughter)—my

first wife was a librarian. (Laughter.) I guess the question I've got in my mind is should universities have libraries or should there be a national, you know, repository where—because, basically, everything wants to be online now. There's really no point in how—I mean, Starbucks satisfies the warm place with wi-fi.

TONY HEY: Absolutely, yeah.

RICK RASHID: You know, people don't go into the library stacks the way they used to. I'm sure they do in some disciplines, but not so much in computer science. You know, Ed, do we really need to have the expense of having different organizations in each university maintaining, in many cases, the same information—

TONY HEY: Absolutely.

RICK RASHID: —or would it be better to have sort of a national or at least a state resource for each individual state?

ED LAZOWSKA: Or disciplinary.

RICK RASHID: Or disciplinary resource, however you want to think about it, that could be staffed adequately, you know, to really archive the information that could be staffed adequately to really help with the data processing and data analysis work where you didn't have to waste a huge amount of funding on all these sort of individual organizations.

RICH DEMILLO: That's an interesting question, but that's really a phase shift for libraries. If that happens, it takes the traditional role of a library, makes it more of a facility, super computer facility, and libraries then—

RICK RASHID: It's a cloud service.

RICH DEMILLO: And then the libraries become another kind of organization that are responsible for information knowledge management on a very local scale. Massive in terms of number of bytes, but—

ED LAZOWSKA: If you even need that, right?

RICH DEMILLO: There are probably regulatory reasons that it has to be done locally today but you're right, I mean, there's no—

ED LAZOWSKA: Okay, next question.

PARTICIPANT: In what Wolfgang said, there was a little bit of almost Whitney Houston, the children are our future. (Laughter.) In order to drive this transformation in research that you all talk about, could you say a little bit about what the fundamental transformation in education that is required in order to drive that, and particularly for those responsible for education? And I'd like you to go beyond just computer science, but also talk about the transformation in education

and those responsible for education in disciplines like engineering, bioinformatics, chemistry, physics, arts, and humanities?

WOLFGANG GENTZSCH: So may I start, I feel obliged with your introduction, and I already have seven grandkids, so I feel very much obliged. I mean, even today we see this one lady one-man show, this frontal education which suppresses creativity, enthusiasm, and so on. While on the other hand when I see my grandkids growing up now with Xbox and other nice digital native toys, this I see a real discrepancy.

But I also see a real chance to get this new stuff into the classroom, to create a sandbox, a science sandbox or whatever sandbox for the kids, you know, they can really interact. You know, they can collaborate. So, where they exactly learn when they are six, seven, eight or so to work together on a little nice challenge or so, to be better prepared for the future, for the jobs and for the big paradigms which are coming over them when they get into real—into the real workplaces. And I'm very optimistic there, I see real chances. What I don't see currently is the governments moving.

Unfortunately, like in Germany, we have 16 states, therefore, we have at least 16 different education systems. And we have hundreds of textbooks in one specific discipline spread all over Germany. So, I mean, this is horrible, this is really anti-productive, and we have to think about how to change it.

In the earlier days when I was more optimistic or when I was—right, even more optimistic.

PARTICIPANT: Or younger.

WOLFGANG GENTZSCH: So I thought you could solve the problem from top to bottom, but recently, again yesterday with this very nice session here, the design, I love that, so congratulations to all those who did that. So, what I can see more and more is that these things are, indeed, coming from bottom to top. Take all the 200,000 applications now on one of these phones, for example. So, this is all individual coming out of the crowd which—basically this is a revolution which is coming from the bottom, and that makes me, again, more optimistic, more positive.

RICH DEMILLO: Let me be pessimistic about your question. And I'm speaking now about higher education as opposed to education broadly. The two things that should bother you if you look very carefully at higher education are how we judge quality in higher education. It's almost linearly related to how much we spend per student. So, a university is ranked more or less, first approximation, for how much they spend per student. Well, how much you spend per student has no pedagogic meaning whatsoever. It's not related to any outcome, it's not related to anything that happens in later life.

So why aren't we thinking about—just back to the ephemeralization idea—why aren't we thinking about how to do a much, much better job in let's say the first two years of college with half the money because that half the money that we save there could really be well spent on the next two years of college where, you know, one on one instruction is really important.

And along the same lines, we judge quality of universities by how exclusive they are. Why is that a measure of anything? Bill Bowen, former president of Princeton, wrote a book called *Crossing the Finish Line*, in which he just beats the crap out of every public university in the country for not graduating enough students—actually, for not graduating as many students as Princeton.

Well, you know, if I get to hand pick 1,000 of the best freshmen in the country, I can pretty much guarantee that short of death or really bad illness, these guys are going to graduate. If I'm Berkeley and I have to accept this broader group, you know, I can't make those guarantees. Does that mean that on an absolute scale Princeton is doing a better job than Berkeley? No. What it means is that Princeton has narrowed its selection to ZIP codes because the best predictor of what you're going to do—what university you're going to go to is the socio-economic group where you go to high school. So, you just look at the ZIP codes in major metropolitan areas, you can map out the distribution to top 20 universities.

Those are two problems that have a technology component, they have a public policy component that I'm very pessimistic about in the U.S. I'm more optimistic globally, but I think this is a huge deal for American universities.

(Break for direction.)

PARTICIPANT: This is a question about primary education. So, I have a lot of respect for people who teach science in primary education, but the fact is that for people who are interested in something like literature or painting, being a teacher in primary education is an attractive profession, but it's not one for those who are teaching science. This is not the case in higher education, you know, teaching science or engineering in higher education, it's a very attractive profession for many of us. So, I'm just wondering what can be done to make primary—being in professional primary education attractive for people who do science.

WOLFGANG GENTZSCH: I think this is reflected very much in the quality of—you know, there are a lot of great teachers of science in primary education. You know, a lot of teachers—I've seen a lot of teachers in areas like English and theater and arts being very creative and innovative in the ways they teach these things in primary education.

RICH DEMILLO: I don't know the answer, but I have an example. You might do what Arizona State is doing. Arizona State has a goal of making sure that a quarter of its engineering graduates walk away with teaching certificates. Not that they're going to be teachers, a quarter of the engineers that leave Arizona State are going to be certified as teachers, which means that some of those people will end up teaching in Arizona public schools, and they will actually know something about math and science.

So those are the kinds of things that you might want to think about.

PARTICIPANT: But, again, even the good teachers don't have the right tools, which are required for our digital nation—

RICH DEMILLO: A lot of evidence that there are no good teachers of science and math across—

PARTICIPANT: There are no?

RICH DEMILLO: Good teachers of science and math.

WOLFGANG GENTZSCH: Oh, I know a few, yeah. But it's all about—I mean, one of the real problems in primary schools are when you have these physics and chemistry and bio labs, it's usually just the teacher who does things and who demonstrates things. This doesn't really create creative and enthusiastic young kids. So, that's another chance with a laptop for one child—one laptop for a child—that they can do experiments themselves, hands on, and become creative.

RICK RASHID: I think we don't—in a lot of cases, I think our kids are really being held back by the fact that they don't have teachers who can teach things that they could learn. And I'll give—my favorite example is my two youngest boys, I've used this in a number of talks where my wife is a computer scientist. She taught them both to program in C# when they were eight years old, right?

And there's nothing really—if you think about programming, there's nothing particularly—there's nothing about programming that has very much in the way of prerequisites associated with it. You know, it's a basic reading and writing, some basic math, a little bit of logic, right? All this stuff that a six- or seven-year-old knows how to do, right? So there's no reason you can't teach these kids—and you know, these kids—you know, they program and they know how to use exception handlers, generics, and they serialize and de-serialize XAML and do all this—I mean, they work in the Visual Studio IDE and it just works, right?

I think part of it is they had individual instruction from a single person, right, that focused on them and who knew how to teach the material to them. And I think there are a lot of kids that could learn a lot of things if we had that capacity.

RICH DEMILLO: And there's a subject matter expertise pyramid that's been turned around in education, right?

RICK RASHID: And I think that's part of the problem is you need to be able to have—and to some extent, the kids need one-on-one education, one-on-one encouragement. Unfortunately, we can't afford that and we don't have an infrastructure that supports it. So, the alternative, in some sense, is to really build much better computing tools that can provide that level of instruction.

My favorite mental image is *The Diamond Age*, you know, Neal Stephenson's novel for those of you who know it know what I'm talking about. But it's basically a young woman's primer, it takes her and educates her through her whole life. You know, it's one interactive physical document. We may not be more than 15 to 20 years away from being able to build something like that.

ED LAZOWSKA: Take some more questions.

(Break for direction.)

BEN SNYDERMAN: Thank you, Ben Snyderman, University of Maryland. I appreciate the inspiration and visionary qualities of what you've laid out here, and several of you represent institutions or universities that have made some transformations. I appreciate especially Ed Lazowska's point about social media being such an important issue, and also a devotion to aspects that make for transformation, society.

I must say, though, there's a certain part of the reality outside this room which is more troubling, that only a few institutions have made the kind of transformations that you describe. And I'd like to suggest sort of two particular issues. One is about the metrics. While I agree that the Moore's Law metrics of petabytes and gigahertz and so on have been good for the past, but in the mature notion of what computer science is, I suggest that new metrics might become a better guiding platform.

So, for example, Wikipedia counts "contribs" and we might think of the peta-contribs that we need, or "collabs" the number of collaborations. And so the tagline I'd use before that the old computing is about what computers can do, but the new computing is about what people can do, I think is one-way of describing what I'd like to see the mature sense of computer science, that we focus not on the plumbing, but the delivery of water and the good health that comes with it.

And so it's very hard to change our colleagues. And one of the key things about changing our colleagues and our students is the participation of women, which is a difficult issue. I think if we look around the room, we look at the panel and others, I think this discipline would be enriched by having more participation from diverse sources, especially from women. So, those are two things that I think we need to do (applause) in changing our discipline. If we change our discipline, we can have the aspirations, we can succeed in the aspirations that you bring, but we need to change our colleagues, not just the few people that are here in the room. We have a lot of work to do, it's not easy, but I appreciate your inspirational vision.

PARTICIPANT: I read an article in Scientific American that basically says we produce too many PhDs. And among the things that it talked about saying, it says that only 15 percent of PhDs are able to get into academia or research universities. And when I see the incoming students and I ask them, "Why do you want to do a PhD?" A large number of them said they want to do it because they can go into research or academia or industry research.

So the interesting thing seemed to happen at least lately in the last ten years or so. For an academic position, we have 300 to 400 applicants. So, clearly there is a huge oversupply of PhDs who want to do research in academia. I think the same thing seems to be happening in the select—when I graduated, the number of research labs were quite a few, and it was prestigious to go into a research lab. More so than academia. That has changed, but the number of real research labs with the flexibility, giving you 20 percent of time, 50 percent of time to do what you want or invent something are very few. I mean, you see Microsoft Research, IBM Research, and two or three others.

So the number of quality jobs or researcher jobs are significantly lower now than the supply there is. Any comments on that? Or do you see it from your perspectives—

RICK RASHID: Yeah, let me make a couple comments on the workforce issue. First of all, there's obviously an oscillation—the economy goes up, the economy goes down. If you look at graduate enrollments in computer science, for example, the number of PhDs per year we produce had been at 1,000 per year for more than a decade and in the past few years, it's risen to about 1700. Why is that? It's because in 2001 because of the downturn, the very best undergraduates, bachelor's graduates, weren't able to get the jobs that they wanted, and therefore they went to graduate school and low and behold, six years later they're coming out. Unfortunately, it's a time of another economic downturn. So, you have to sort of damp out these oscillations.

The fact is that by all measures in at least the United States from the Bureau of Labor and Statistics, a significant shortfall of people at all degree levels in the computing field, and 70 percent of all new jobs in all fields of science and engineering together over the next decade are projected to be in the computing fields.

The guy at the Bureau of Labor and Statistics who's in charge of these statistics has said that all the rest of science and engineering is hiding behind computer science, all right? By which means there is no workforce shortage across all of science and engineering, but there is such a big workforce shortage in computer science that it makes all of science and engineering look short.

So that's not to say that everybody gets the job that they want, it's also the case that all of us in universities, we reproduce like bunnies and we're trying to produce people who look just like us and want to do exactly what we did. And it's important to realize that getting a job at a research-intensive university is not the sine qua non for a PhD graduate.

So I am convinced that our field needs to produce far more people of far more varied flavors as well, and that there's huge opportunity going forward for the next decade. And I think the statistics bear that out, but it's not true for all of science and engineering, it's true for our field.

RICH DEMILLO: We're counting down on time, I think we have time for one short question.

PARTICIPANT: Just a quick question—well, maybe motivational statement. I'm in computer engineering and I'm in a school of engineering science and everybody wants to go into mechatronics because it's the hot, new, exciting thing, or biomedical because it's the hot, new, exciting thing. And computing engineering and computer science used to be that. And when I talk to the students, what I find is we've got lousy marketing for lack of a better way of putting it. They see Dilbert and they think they're going to be sitting in a cubicle. They see video games, and they don't see what else they're going to do.

I spend a lot of time saying, look, computing science, computer engineering, two sides of the same coin, it's a tool. You can put it with anything that you like and do something with it. And I guess what I'd like to ask on behalf of the educators would be, I think it's important that industry gives some visibility to the cool things that kids can do to get them back into it, to get them

excited about it so now we're not the sort of fallback thing, oh, I guess maybe I can do that. No, computer science, computer engineering, this is amazing stuff. I can do anything with this. And I think that some of that goes to industry to try and bring that back as shiny, new play toys.

ED LAZOWSKA: Here's a little campaign of mine that you can all join, okay? And that is write to whoever you know at Google and tell them to stop referring to their employees as engineers and start referring to them as computer scientists, which is what they are. Right? They all have degrees in our field, right? So let them know that they're doing damage to the field by that form of reference.

RICK RASHID: I'll just make a comment. I think actually within industry we really do try to—because obviously we have a big recruiting issue. So, we put a lot of energy into trying to make the jobs we have attractive and to get the word out. So, there's a lot of that that goes on. I think within the universities you have a responsibility as well to make things exciting, especially—so many students go through the first course in computer science. A lot of universities—as many as two-thirds to three-quarters of all the students go through the first course in computer science.

RICH DEMILLO: Yeah, all of them at Georgia Tech.

RICK RASHID: And all of them at Georgia Tech. This is a huge marketing opportunity, right? If you make it boring, if you make it all about programming and doing boring little things, then everybody except the nerds will leave, right? And the people who really love doing boring things will stay, right? (Laughter.) And they'll propagate that view of the world.

If you make it exciting and get people excited about what is computing and why is computing—you know, how does it make the world better? I mean, how can you solve real problems? And you put that in the very first class, you know, that I think can get people excited. I know some of the universities are doing that—

RICH DEMILLO: Well, you did that with the Personal Robotics Institute.

RICK RASHID: Right. And we see people doing this, and where they do it, we see a much better uptake among the students, we see much greater diversity among the students because it actually appeals to them. So, I think that's—you know, yes, industry can do a lot in terms of what we do and if you can convince Google to call it computer scientists, that's great. I don't think they'll listen to me. (Laughter.)

But also, think about the marketing opportunity you have to make people excited about the field and I think there's a—you know, we don't put enough energy into that first class.

RICH DEMILLO: I think we're going to wrap it up at this point. Let me thank the panel, which was a great panel, and the audience, which was a great audience. (Applause.)

(Break for direction.)

END