DSD 2017: A Personal Perspective

Kevin R. Vixie

Disclaimer

The viewpoints here, while almost uniformly shared by other faculty in the Analysis+Data Group, are my opinions. Because I led the organization of the data science day and linked events – see https://analysisplusdata.community/, these opinions have heavily influenced those events.

Comments: Philosophy and Perspective

Data Science

Like anything with real power, Big Data can be good and bad, liberating and enslaving. But no matter what we might wish for, it is never truly neutral.[1] From the invasive snooping of the intelligence agencies, to the manipulative technologies that enter our most private lives through mobile technology, to the ability to understand much more subtle phenomena and see deeper and further into the ultrasmall and ultralarge universe, the growing power of data science is enabling the very good and the very bad. As Lord Acton famously stated, "Power corrupts; absolute power corrupts absolutely."

The questions a good future depends on are not only those that are primarily scientifically interesting, but also those of wisdom and forethought:

* How do we learn from massive data streams while preserving freedom and privacy in this new age of high-performance data science?
* How do we shift the focus from massive data to thick data, from clever manipulation to wise use, from profit-driven data exploitation to data insights that actual liberate us and make the world a better place?
* How do we inspire a shift back to taking the time to think, to see, to feel, towards a world that realizes the insanity of always bigger, faster, richer?

But these are questions that humans operating in a connected, empathetically viable communitarian environment would naturally come up with. And it is only by returning to such environments that will we avoid the race to the moral bottom that the power of technology enables. (And in fact, the misguided idea that anything, like data science, can be neutral, follows directly from not understanding the connected nature of reality.)

[1] The comments in this section are from my perspective and while many in the Analysis + Data Group share the same or similar views, I can only claim them as my thoughts. KRV

As might be expected, what data science is varies depending on who you talk to. But the broadest, most organic views admit that it is far more than just some part of computer science and/or statistics.

> A reasonable definition of data science is that it has three components: domain expertise, mathematical sciences, and computational horsepower.

The *domain expertise* is those pieces of chemistry, physics, engineering, biology, sociology, psychology, medicine, etc that deal with large, and often complicated, data streams. The *mathematical sciences* includes big pieces of mathematics, statistics, computer science, and electrical engineering as well as physics and economics. And by *computational horsepower*, I mean high performance computing along with other rather low level computational tasks that are critically important to any solution involving large data streams.

Another critical issue, far to often ignored is the fact that there is a difference between statistics and understanding, between finding a needle in a haystack and knowing why the needle was there in the first place. Finding correlations and understanding are two very different things – seeing something, even if that something is surprising, is not the same as understanding that observation. And when understanding is the focus, we move from mere manipulation of data to the patient, thoughtful science that lies at the triple point joining the mathematical sciences, the domain expertise and the computing power that make up data science.

(Yet another extremely common, and ultimately wasteful idea, is that the solution to every problem involving computing, is a bigger, faster computer. This shallow approach to problems wastes enormous amounts of resources and distracts us from the truth that real solutions require truly new ideas and a great deal of reflection and wisdom guided cleverness and insight. Mere computing power by itself, falls prey to the ubiquitous exponential growth in complexity; simply throwing a bigger computer at the problem will never, ever be enough.)

Of course, the reason scientists work on data science is that it is truly fascinating and useful. Here are examples of important threads with many fascinating problems to work on:

Graphs and Networks From social networks to networks of neurons, from links between webpages to dendritic structures in nature, graphs and networks are everywhere. And there are lots of problems to inspire anyone with an inclination for mathematics and computing.

Learning in High Dimensions: Images and graphs and text documents and audio recordings and high bandwidth measurements of anything generate data that is very productively thought of as points in very high dimensional spaces. This brings with it a whole host of problems and conundrums, the unfolding of which has begun to show us that high dimensions are truly surprising and rich in opportunity for study.

Modeling Singularities: While most of the mathematical models you first encounter in mathematics are very nice and smooth, reality is full of things that are best modeled by rough sets and crazy functions and wild measures. Understanding how to calibrate the wildness, to choose just enough that what you need to model can be modeled, but not so wild that you cannot work with it, is an art form that is endlessly fascinating.

Images and Video Streams: Though already mentioned about in the "Learning from high dimensions" above, image and video data is such a big piece of the data flood overwhelming us, that it deserves its own bullet. The work inspired by this data is truly gargantuan, yet in no way have all the problems associated with that data been resolved, partly because new and interesting ways to generate image and video data is being dreamed up every day.

Learning from Text: Text mining is still, in may ways, in its infancy, if for no other reason, that the aims that you might have in this thread so naturally include artificial intelligence which is yet another enormous can of worms.

First Do No Harm: At least some of us also argue that these fascinating scientific questions, are no less interesting when you insist on the the principle, "first do no harm". (Even though this may limit the funded problems you choose to work on.)

Environments for Creativity and Innovation

The Data Science Day was organized by the Analysis+Data group, which works on the data science living at the intersection of data problems and various pieces of the (very broad) field of analysis in mathematics. (For more information on what we do and what our perspectives are, you can visit https://analysisplusdata.org).

Our perspective is that siloed environments are deeply unhealthy for the development of science. Empathetic environments – connected environments marked by intentional awareness that some call wisdom – are crucial for developments that are truly sustainable. As a result, we are working towards the creation of some sort of updated, improved combination of Bell Labs and the Nesin Mathematics Village. (See The idea factory by Gertner and the webpage of the Nesin Mathematics Village.) As a result, the Data Science Day is designed to be inclusive:

> It is important to know that the Data Science Day that we organize actively solicits participation from data scientists and those merely interested in data science, no matter what their particular focus is.

Part of the motivation for this is the recognition that data science is certainly wider than the data/mathematical analysis intersection, but just as important is the recognition that one of the most important ways to get new ideas in mathematics is to immerse yourself in completely new, real world ideas that do not lie close to your frequent paths.

Strong interaction across disciplines is a key component of how we believe we should operate in the scientific community. As a result:

> The design of the Data Science Day is unique. Typical conference talks are banned and instead, every talk focuses exclusively on what we do not know, on open problems, with most of the time being filled with interactive discussion of those open problems.

This year, there were 22 speakers, 18 of which talked for 5 minutes, 4 of which talked for 15 minutes. Who those speakers were and the general area from which their open problems were taken are on the schedule that can be found here: https://analysisplusdata.community/2017-data-science-day Because we believe that students and education is inseparable from research and innovation, we organized parallel events in the AMS meeting that coincided with the Data Science Day this year. In addition to the events on Friday, April 21, 2017, there were three linked events on Sunday that are also part of the American Mathematical Society Sectional meeting being held at WSU this year. These include (1) a session on Geometric Measure Theory and its applications to things like data, (2) a panel discussion aimed at students (undergraduate and graduate) who might consider internships in industry and government labs, and (3) a poster session focused in contributions from early career faculty. More information is available here: https://analysisplusdata.community/events/

At the foundation of our organizational efforts is a belief that environment is enormously important – that it actually determines the future for innovation and science. There is a mistake made by the majority of people working in STEM; as measured by their actions, they seem to think that environment does not matter that much. It is true that good things can come from people working in very difficult circumstances. But it is also true that most of us understand instinctively and intellectually, that good environments produce and sustain the best, most creative work. (Part of the explanation of how good things come from bad environments is that some are able to ignore bad environments, to take only the good from the environment and not see the rest. In other words, they are able to change their own environment in a way that is usually challenging for most people.)

Yet many mathematicians and scientists both (1) take the environment for granted and (2) are unable to understand that *their experience of the environment* is very largely responsible for the innovations they get credit for. As a result of their unrealistic view of where innovations come from, they are also curiously unable to imagine how to change their own environment, even when that is crucial for the survival of their own creative life.

The unenlightened division of academic life into research, teaching and service reflects and reinforces the idea that the environment can be separated into these pieces with no damage to our creative life. Service, which is looked down on as though it were a chore, is considered quite separate from research and teaching. Instead of understanding the essential inseparability of all these intertwined threads, we separate them and suffer the results: the decay of the rich, vibrant, living culture a wholistic, empathetic environment delivers to those willing to embrace it. (In fact, the reason service is such a chore is that it is separated from the other threads. Were it integrated it would neither be a chore, nor would it be something that was looked down on.)

The gospel "truth" that that there are normal people and geniuses and super geniuses that are the accidents of birth, is another reason for the persistence of the idea that environment is not so important. See for example *The Genius in all of us* by David Shenk.

I believe that environment, combined with one of a small number of initial predispositions explains everything, and that genius is as common as dirt, at least if one is to focus on *potential*. From this perspective, civilization deserves condemnation for squandering enormous amounts of human potential, wasted because our ill conceived notions have turned into self-fulfilling prophecies. The fact that small experiments in reform find that perturbations cannot move us to the other human potential optima that I and others believe in, is not a convincing argument that the prevailing viewpoints are correct. It just confirms that the problem is a wholistic one not easily solved.

A belief in the supreme importance of the environment is strong encouragement to put a great deal of effort into exposition and teaching, into the creation of open environments that are rich and generous, freely accessible to very wide audiences. But of course, the fundamental problems are economic – the current system literally robs the majority of people in the world of the chance to develop those amazing potentials. And of course the solution of those problems, on a large scale are out of the reach of any reasonably sized organization. *But what can be done, if we put our minds and hearts to it, is a great deal more than is currently being done.*

That, in a nutshell, is why we organize the events we organize.