UC Berkeley School of Information

INFO 290A-03, Fall 2013
Social Data Revolution
Prof. Andreas Weigend

This article was posted on 20 May 2013 to http://weigend.com/blog/2013/05/datarefineries

Beyond big data: How personal data refineries change big decisions

by Andreas S. Weigend

A hundred years ago, the only data a shopkeeper had to work with was the inventory on the shelf, and the money in the till at the end of the day. That data was recorded with a fountain pen. The consumer based her purchases on pretty pictures on the box or on anecdotes from her friends.

Fifty years ago, mail order companies knew where you lived and what you ordered. In addition, they could buy some basic demographic information about you. That was it for personal data pre-web.

With the advent of e-commerce, retailers could track every click and purchase, and capture every abandoned shopping cart.

In the 1990s, Amazon pioneered the use of data to help its customers make better decisions. First, implicit data: Clicks and purchases of all users are aggregated to suggest items to a shopper in response to their most recent click. Second, explicit data: Customers have the opportunity to publish reviews that potentially influence the purchasing decisions of other customers. User-generated content turned marketing–previously viewed as carefully controlled and released information–on its head.

I think of Amazon as a data refinery: Amazon takes the data that people create, refines the data, and returns results, allowing people to make better decisions. Amazon now influences how a billion people shop.

This article looks at three common questions that many people ask every day: (1) Who should I work with? (2) Which route should I take? (3) Where should I stay on my next trip? The answers to these

questions, their decisions, are now influenced by the personal data of a billion people.

(1) Who should I work with?

A startup I am advising recently hired a star engineer. How did they find him? Not through referrals or a headhunter, but through a post of his on Quora, a question-and-answer site. Like the shopkeeper, employers now have vastly more data resources. And like Amazon, job and professional sites now refine data that people create to help both individuals and companies make better decisions.

For example, LinkedIn provides tools for individuals to both refine their own personal data, creating a work identity that transcends a specific job, and to find others by acting as a refinery for other people's data. Similar to e-commerce, the asymmetry between buyer and seller is fading away.

This does not only apply to full-time jobs. The number of marketplaces with different mechanisms to match talent and tasks is exploding. Underlying the future of work is identity that persists across tasks and jobs where reputation is a key output of the data refinery.

Within firms, data refineries are used to create teams and track interactions. A hedge fund with more than 100 billion dollars under management captures video and audio of its meetings and other data sources and correlates them to the outcomes of trading decisions. And Google's "People Analytics" has reinvented HR.

In the future, what kinds of jobs will still require full-time employment, and what outputs of personal data refineries will be needed to power the human cloud?

(2) Which route should I take?

In the 1990s, at Xerox PARC, we used a Thinking Machines supercomputer to analyze automobile traffic patterns in order to predict when the flow would change from laminar to turbulent. Little data, and many assumptions, went into those models.

Twenty years later, a complicated prediction problem has turned into simple observations, in real time, of how cars are moving, or not. Microsoft spin-off Inrix refines geo-location data from more than 100 million individuals a day. In turn, it provides them with crowd-sourced traffic information. You may be sharing your location data without even knowing it.

The company, which sells to Garmin, MapQuest, Ford, BMW and others, collects data from mobile carriers about when a phone switches between cell towers, in addition to GPS and other data. Besides helping drivers make better decisions on which route to take, Inrix also helps cities with their planning decisions, from how to time traffic lights to where to build bridges.

As a byproduct, Inrix provides hedge funds with shopping mall traffic data to help them place bets. For example, data collected on Black Friday 2012 correctly predicted a major bump in sales for the entire

holiday season.

We are what we eat, we are what we search for, we are where we were, and we are who we were with. Location history is amongst the most sensitive data about a person. Or, as Yogi Berra said, “No matter where you go, there you are.”

(3) Where should I stay on my next trip?

In 2005, Marriott announced a breakthrough in customer service: Guests would now be able to specify their pillow preference when making their reservation! This pillow personalization represented a shift in what had become gold standard in hospitality: personality-free lodging.

While hotels can capture personal data ranging from real-time minibar and video consumption, to card key accesses to room and gym, their goal still seems to be the sanitized experience.

In the meantime, their market has been threatened from a completely different direction. Airbnb offers a rich set of data to both guest and hosts enabling them to make their decision: Love pets? Want to share a hot tub? We’ve got the match for you.

However, staying in a stranger’s guest room requires a much deeper level of trust than staying in a hotel. To address this need, Airbnb verifies online identity on Facebook or LinkedIn by matching it with offline identity via Jumio.

Travel and tourism is ten percent of the world’s GDP. Beyond accommodation, matching and trust based on refining personal data now also extend to other areas from ridesharing to renting out your car.

Conclusion

A hundred years ago, data got recorded with a fountain pen. The data deteriorated over time, whereas the fountain pen got better with consistent use. In the information age, the central question for companies is: Will their product or service get better over time, or worse? Data refineries such as Amazon, Google and Facebook get better.

Like the story of the genie in the bottle, the personal data servant can wield its power for good or evil. What it cannot do, however, is go back into the bottle. The new opportunities in this abundant data ecosystem will come from new ideas about how to refine this data.

The sign has flipped, like that of the shopkeeper in the morning, from CLOSED to OPEN.

INFO 290A-03, Fall 2013
Social Data Revolution
Prof. Andreas Weigend

The future of e-commerce

by Andreas S. Weigend

Imagine walking into a mobile phone store to pick up the latest model device. As soon as you walk through the door, the sales clerk hits you with a barrage of questions: Who were the last ten people you called on your phone, and how long did you talk to each of them? How many hours a day do you spend on your phone playing games, watching videos, or browsing the internet? What websites do you visit on your phone, and what search terms do you use there? How many photos did you snap in the past week, and where was each one taken? Who are all the people you've visited today, and what are their addresses? Where do you go on the weekend? What time do you leave for work in the morning, and what route do you take to get there?

Most of us would be shocked and dismayed by a phone store employee asking us hundreds of personal questions. And yet, every day we share that information and much more with our phone providers. In order for your mobile phone to send and receive calls, your phone protocol needs to let the network know where it is. Any time your phone is turned on, you are sharing your location with your phone service provider. Sharing your location in this way sacrifices some of your privacy, but most of us find that the benefits far outweigh the risks: in exchange for telling T-Mobile or Verizon where we are, we can contact anyone at any time, anywhere, throughout the world. When we expand our phone use to include things like browsing the internet or downloading apps, we share still more data with our phone providers as well as with other parties.

Of course, it is not just our phones that work on this principle. In order to use Google Maps, we must allow Google to track our movements. In order to find a product on Amazon, we must tell Amazon what we are looking for. In order to purchase the product, we must give Amazon our credit card information, and before we can receive it, we must give them our mailing address. Certainly one could make the choice to not use a mobile phone, mapping software, or even the internet. But consider the cost of such a choice. In order to maintain total privacy, we would lose the opportunity to send email, shop online, and much more. We would live a relatively isolated, inconvenient, and information-poor life. Instead, most of us share our information willingly in exchange for what we get in return, from turn-by-turn

driving directions to real-time price comparisons. We give data to get data.

As we move through our day, making choices about everything from which freeway to take to which coffee to buy, we generate and share a tremendous amount of data about who we are and what we want. In return, we get back information that allows us to make better purchasing decisions. Fundamentally, we cannot receive information without first sharing our personal data, and the more personal and specific the information we share, the higher quality information we get back. The data revolution relies on this fundamental principle: give data to get data.

The media often presents "big data" in a predatory light. In this scenario, the customer is a powerless individual who cedes their personal data to unseen entities for unknown purposes. In reality, customers and businesses are equal partners in an information exchange that benefits both parties. In this chapter, we will investigate what it means to "give data to get data." We'll trace the evolution from consumers unintentionally generating "data exhaust" to actively creating and sharing content. We will consider why customers are motivated to share their information, weighing the risks and rewards. And we'll explore the potential for social sharing to promote openness, transparency, and creativity.

In the days before the digital revolution, businesses relied primarily on manual merchandizing to sell their products. In merchandizing, salespeople and marketers try to make guesses about what people will buy based on things like market research, focus groups, and surveys; past experience; and their own intuition.

With the digital revolution came the potential for retailers to replace guesswork with a new method for assessing customer desire: learning what items customers view, purchase, or pass by while shopping online. Amazon was one of the first retailers to see the potential for this type of data-driven product promotion. Each of Amazon's millions of items is set into a matrix in which every product appears in relation to every other product. Imagine you purchase both a camera and a camera case. When the next customer comes along and adds the same camera to his shopping cart, Amazon can recommend the same case. And it doesn't only work with purchases – Amazon can make recommendations based on what you have clicked on, suggesting that many customers who viewed that camera also viewed its case. They can even take some of the guesswork out of shopping by suggesting that, for example, 85% of customers who clicked on the camera you are considering wound up buying another camera instead – a good tip that maybe you should do the same. When we aggregate the clicking and purchasing behavior of every Amazon customer in regard to every item, the results can be astonishing. Instead of relying on merchandizing experts, Amazon harvested the collective intelligence of billions of purchases to help their customers make better decisions.

At the same time that businesses like Amazon are counting clicks online, physical sensors are bringing the data mindset into brick-and-mortar stores, analyzing things like customer movements and in-store sound levels and using them to determine what does and doesn't work in the retail environment. Judicious use of physical sensors could allow stores to move employees to areas where they are most needed, minimizing wait time, or to prevent shoplifting. But physical-world data isn't limited to what goes on inside retail stores. As long as we're carrying our phone, we can send and receive information

about our location. When we enable Google Latitude or check in to Foursquare, local businesses can send us promotional messages and special deals, or simply keep track of our movements to discover, for example, that customers who have just exited a movie theater are more likely to head for dessert than those who have just exited the mall. Palo Alto-based Bay Sensors uses a combination of audio and visual sensors and data analytics software to create a complex picture of who is walking down the street and past local businesses. Mounted in the widows of retailers, and facing the public area where there is no expectation of privacy, these sensors let retailers know when you are walking by their shop and how long you are pausing in front of a compelling window display; mounted inside, the same sensors monitor everything from the store's temperature and audio volume to the male-to-female ratio of the customers inside.

In the future, we can expect that brick-and-mortar retailers will take this one step further, perhaps by allowing you to swipe a credit card or other identifier as you enter the store. Then you could browse the store armed with personalized information – for example, your phone might display which items from the store's inventory are in your size, or which items are commonly purchased with items you've bought in the past. Or skip the card swipe, and allow your phone to automatically alert your favorite businesses when you are in the neighborhood and they can woo you with specialized offers and discounts.

Most shoppers are at least dimly aware that retailers are tracking their online clicks and credit card swipes, and we implicitly agree to the arrangement every time we enter a search term or enable geolocation. But fewer people have considered the “give to get” relationship created by this click-counting. In the old days of manual merchandizing, we gave nothing and got nothing back: that is, we didn't share our personal information, and so we didn't receive personalized recommendations. With click-tracking, we share our data, aggregated with that of millions of other people, and receive data in aggregate back – a substantial improvement over the past, but still a somewhat impersonal exchange. This type of shared data can be termed “data exhaust”: like our cars leaving a trail of exhaust behind them as we speed down the highway, we leave trail of clicks and sensor pings in our wake as we move through our day. In order to get more powerful, more personalized data, we can't rely on mindlessly creating data exhaust – we need to actively contribute quality information instead.

Customer-contributed reviews everywhere from Amazon to Yelp are sources of actively created data, where consumers choose to contribute their opinions on products and services. Consider the purchases you have made in the past few months: how many were influenced by online reviews? For most of us, a powerful review for or against a product can be a key factor in our buying decision. Unlike with the exhaust we haphazardly leave behind, we spend time crafting quality data, and in return, we can get much more valuable data back.

The move from clicks to reviews requires businesses to be willing to give their customers open access to their own carefully crafted content. At Amazon, Jeff Bezos championed a philosophy of removing all barriers to getting customer feedback, including allowing customers to write reviews without first signing in to the site. Furthermore, by offering users the chance to rate how useful a certain review is, Amazon captures yet more data about not only their products, but how consumers interact with those

products and with one another.

Too many companies force their customers through hoops before they allow them to communicate their ideas. This comes from an era when companies paid for customer lists and so wanted to milk them for everything from their title and responsibilities to the age of their children before allowing them to interact with the site. But little have those companies realized that the information customers want (about, say, which of two watch batteries is superior) can be found elsewhere, and those who make it easy to find are the ones making the sale. Often this inflexibility means consumers limit themselves to incomplete and unhelpful feedback, or simply give up on the whole thing and walk away. Multi-part drop-down menus mapping a world view different from the consumer (trying to hide the "other" option) and lengthy sign-in schemes (do I really need a high security password to check a "like" button?) don't allow for the kind of meaningful comments that would benefit businesses and consumers alike.

Of course, one good thing about drop-down menu boxes is, it keeps things neat. When instead you give customers a wide-open comment box, you have to be prepared to get back – well, anything, from a four-page rant to a haiku. Along with openness comes messiness, but the power of big data is that it can turn mess into meaning.

Tagging is now a regular feature on everything from Facebook and Pinterest to Flickr and Instagram, placing the responsibility for categorizing and sorting content in the hands of users as well as content creators. Naturally, this leads to everything from accidental typos to intentional trolling, but once again, the power of large numbers prevails, and enough good tags quickly push down the bad ones. As long as data is transparent and corrections are easy to make, a little "mess" can be a good thing.

In this spirit, Amazon launched a new phone app that allows users to snap and share a photo of products they find. Browsing in the grocery store and come across an interesting bottle of wine? You can grab, tag, and share a photo that may help a future user identify that same bottle when it appears on his store shelf. As with all open tagging, there will be user errors and even intentional mislabeling of items, but by harnessing hundreds of thousands of customers out in the world with their phones at the ready, Amazon is substantially enriching its image database, making search more accurate and intuitive for us all.

Being messy creates means making continuous improvement possible. At Amazon, that means building a system that continually improves the relevance of its searches: if more customers are getting where they want to go by looking for "kid's witch Halloween costume" than "children's witch Halloween costume," Amazon's product descriptions can come to reflect that preference, leading to a more "natural" search experience that mirrors how we really talk and think about products.

But where there are reviews, some of us worry whether they are provided by real customers or written by dishonest retailers to inflate their own reputation or damage a rival's. In order to make these reviews as helpful as possible – and to limit the potential for fake reviews – companies that rely on customer-contributed reviews must carefully construct their sites. Some sites require that the reviewer purchase a product before he can review it. The social ride-sharing company Lyft, ask for reviews from both passenger and driver, providing a check on inaccurate or unfair reviews. Of course, a business with

a more restrictive review policy will have fewer reviewers, but they may also be of higher quality. On less restrictive sites, like Amazon's, shoppers can rate how helpful they found a certain review, voting up insightful commentary and voting down fake reviews and other spam comments. Moreover, the very effort required to maintain a fake identity with a portfolio of fake reviews contributed over time means that it makes little economic sense for unscrupulous companies to pursue this policy for long. Whether a business chooses a more or less restrictive review policy, the goal is the same: a well-constructed ecosystem where good information can thrive.

When we create and share content like reviews, we give more and get more back. But as powerful as reviews can be, they remain a static measure of how you were feeling at one moment in time. As companies collect more and richer data, they are gradually replacing one-size-fits-all recommendations with ones that are context-dependent, refined for your present situation.

This situational analysis can be viewed as the data-driven version of old-fashioned intuition and common sense: when you go into a store to buy a new suit, a good salesperson will assess the situation she sees in front of her, she doesn't ask what kind of suit you wore for your first communion twenty years earlier. Our shopping habits or interests, like so much of what we do, vary in different situations. We don't buy the same kind of wine to bring to a party as we do to drink at home, or the same kind of underwear to wear around the house alone as we might buy for a romantic night out. Where once we were reduced to a set of specific, relatively fixed criteria (male, 45 years old, married) now we are an ever-changing sum of our experiences and objectives. By knowing our geolocation, our latest updates on Twitter or Facebook, and much more, businesses can offer us much more meaningful recommendations, sometimes predicting what we want before we know it ourselves.

Google organized and made accessible the world's online content and Google Maps did the same for the world's physical structures, including highways, bridges, and buildings. Now Google Glass is set to bridge the digital and physical worlds by providing its wearers with a way to capture and categorize whatever they see in the real world. Imagine walking into a grocery store wearing Google Glass' head-mounted camera. As you stroll down the aisles, Google Glass can give you whatever information you need about product location, price comparisons, and nutritional content, even what products your friends have recommended, or what ingredients you need for a new recipe you're planning on trying that night. In exchange, you send Google a wealth of information – not just what is in the store, but how long your gaze falls on each item. It's a map of that retail environment, weighted by your attention: what in this store is important to you? What is tempting, and what is forgettable? And Google can use it to provide richer experiences to all the shoppers that come after you, the ultimate example of “give to get.”

So far, we have been discussing the ways that we, as individuals, can share our personal information in order to get back information we use. But when we share our social information, we can get back vastly richer reserves of data. The social graph is a complex and constantly changing web of relationships you build over social networks like Facebook. Every time you friend someone, like one of their posts, leave a comment on their status update, tag them in a photo, or reference them in a comment, you are strengthening a connection. This connection can also be weakened through blocking, muting, or

unfriending. Businesses can track this evolving graph to learn more about what people want and need. People recommend products to one another through formal referral programs and informal posts on places like Facebook, Pinterest, and Twitter. Unlike product reviews, which are usually written by strangers and directed toward the world at large, these recommendations are written by people you already know and trust, whose circumstances are more likely to be like your own, so they can lead you to more targeted, better-value information than what you would find somewhere like Yelp alone.

Sharing social data has fundamentally changed our society from one that valued secrecy and control to one that promotes openness and transparency. From these new values are emerging new social business models that challenge the old ways of defining companies and customers. Airbnb is a prime example of this new "social business model." Airbnb replaces hotels with house-sharing: owners can offer up their houses, apartments, or other dwellings for paying guests. These guests can peruse the Airbnb site to search through hundreds of rooms in everything from high-rise condominiums to rural cabins, and negotiate with owners about any of their preferences, from pet policies to bedding. The decision to allow a stranger to stay in your own home – or choosing to stay in the home of stranger – is predicated on trust, and it is the social graph that enables that trust. Before agreeing to a housing arrangement, you can check up on your potential tenant or landlord using the site's reviews. Still unsure? Use Facebook to check up on that person, even to see what friends you may already have in common. Ride-sharing companies Uber and Lyft use the same tools to allow potential drivers and riders to find matches they feel comfortable with. Before hopping in the car, you can use your phone to view all of the available vehicles in your area and select which car and driver you prefer, then call that driver to negotiate a price and terms – you will even be notified automatically if you have any Facebook friends in common. At the end of you drive, both you and your driver rate the experience, providing mutual feedback. Other social business models are enabling users to share, borrow, or rent everything from parking spaces to household tools, all with the increased trust and transparency that comes from a shared social network.

But openness and transparency are changing the world of commerce in ways that go beyond just sharing rides or renting rooms. The same "give to get" bargain that consumers strike with retailers and with one another also applies to business-to-business interaction. In fact, the secrecy with which traditional retailers have guarded their data may soon seem as outdated as the fax machine. Innovative businesses have found that freely sharing information – both within the organization and outside it - can improve their products and services, customer relations and their bottom line. Back in 1992, Walmart began sharing their sales data with their suppliers – once anathema to most merchants – in order to help their suppliers be proactive about ensuring availability of tomorrow's coveted items. As you might expect, this willingness to be open benefited all parties: the suppliers, the consumers, and Walmart itself.

In 2002, Amazon was faced with a quandary: should searches on Amazon.com return ads by their competitors? In the short run, there was a clean benefit, because companies like Google would pay Amazon to show their ads. But in the long run, wasn't Amazon running the risk of losing customers if they discovered a lower price on a comparable item elsewhere? The retailer decided to buck

conventional wisdom and display competitor's ads. Not only did Amazon benefit from the ad revenue, but they earned the trust of their customers. Allowing customers to see what else was out there demonstrated Amazon's confidence in its own value and convenience. Moreover, those customers no longer felt compelled to go somewhere else to look for a bargain – Amazon became their single entry point for product searches, creating a strong relationship that benefits both customer and retailer. In the long term, Amazon knew the goal of a successful company is to help their customers make better decisions.

Open inventory is another way that businesses can take the guesswork out of the shopping experience. At one time, airlines were notorious for guarding all information about their ticket prices, from average prices per flight to number of seats oversold. Keeping customers in the dark was their key to controlling prices. Airlines like Lufthansa have changed all that, publishing real-time information about how many seats are available on a given flight and at what price. This open information allows travelers to make informed decisions: if you absolutely must take a specific flight, you will be willing to pay a higher price for it, while travelers with greater flexibility can wait for a deal. In the future, perhaps airlines will go even further, allowing travelers to indicate at the time of purchase whether or not they are flexible, and then contacting the flexible travelers later with incentives for changing their tickets on overbooked flights.

Fifty years ago, there were very few ways to communicate with large corporations on a personal level. Instead, market researchers slotted you into a certain demographic – say, a white male, married, between the age of 30 and 50, upper middle class. When these demographics weren't effective enough, marketing executives responded by carving up smaller and smaller segments of the population – perhaps, a white, college-educated, married male engineer between 35 and 40 who works in technology and lives in a suburb of Los Angeles. But these smaller and smaller segments were still far more general and impersonal than direct knowledge of who you are as an individual.

Along with the movement from one-on-one knowledge to general marketing segments came a more sanitized, depersonalized consumer experience. If a company doesn't know for sure what their customers prefer, the safest route is to choose something safe, generic, and inoffensive. A hotel that doesn't know very much about its guests might choose to decorate their rooms in a bland, colorless way, and to outfit each room with the standard kit of mini fridge, coffee maker, and telephone, whether you want them or not. The digital revolution changes our relationship with businesses, from one where we are the passive recipients of marketing messages to one where we can talk back. The result is an opportunity for consumers to receive specific, personal, meaningful information and use it to make the best decisions for their own unique circumstances.

Today the cell phone in your pocket collects data on what time you leave for work in the morning and what route you take to get there. Tomorrow your phone will collect data on your mood as you drive to the office. Are you feeling anxious? Bored? Energized? The next generation of cell phones will be able to measure your attention, energy, and emotional state. Already cameras on phones like the Samsung S4 are capable of reading your facial expression as well as sensing the direction and length of your gaze, forming a good model for your level of engagement with whatever it is you are looking at. Start-ups like

Color and FaceSense are working to turn your smart phone into a virtual mood ring. In the future, more sophisticated measurements of things like your heart beat, blood oxygenation, and pupil dilation will create a still more refined map of your internal state. Each time you allow your phone to measure your pulse or snap a photo of your face, you will be giving up more data about who you are – what will you get back in return? A map of your world, made richer and more meaningful because it has been annotated not just by your emotions and experiences, but by those of your friends and family and the world at large.

Naturally, the data generated by these new technologies will revolutionize the retail landscape. Yet this "give data to get data" relationship extends far beyond learning about new books on Amazon or renting rooms on Airbnb. In the next chapters, we will see how the data we send and receive is changing the way we manage our health, our careers, our communities, and our personal relationships. As informed participants in this new economy, we can choose to use our data both creatively and judiciously, empowering us to make better decisions every day.