A BUSINESS REPORT ON

Big Data Gets Personal

Big data and personal data are converging to shape the Internet's most surprising consumer products. They'll predict your needs and store your memories—if you let them.

CONTENTS

The Data Made Me Do It

Has Big Data Made Anonymity Impossible?

Predictive Apps Mine Your Life

Data Won the Election. Can It Save the World?

Logging Life with a Lapel Camera

Stephen Wolfram on Personal Analytics

Marketing to the Big Data Inside Us

Intel Fuels a Rebellion Around Your Data

The Dictatorship of Data

Big Data from Cheap Phones

... and more

LAURIE ROLLITT

The Big Question

The Data Made Me Do It

The next frontier for big data is the individual.

● Would you trade your personal data for a peek into the future? Andreas Weigend did.

Weigend, the former chief scientist of Amazon.com who now directs Stanford University's Social Data Lab, told me a story about awakening at dawn to catch a flight from Shanghai. That's when an app he'd begun using, Google Now, told him his flight was delayed.

The software scours a person's Gmail and calendar, as well as databases like maps and flight schedules. It had spotted the glitch in his travel plans and sent the warning that he shouldn't rush. When Weigend finally boarded, everyone else on the plane had been waiting for hours for a spare part to arrive.

For Weigend, a consultant and lecturer on consumer behavior, such episodes demonstrate "the power of a society based on 10 times as much data."

⟶

If the last century was marked by the ability to observe the interactions of physical matter—think of technologies like x-ray and radar—this century, he says, is going to be defined by the ability to observe people through the data they share.

So-called anticipatory systems such as Google Now represent one example of what could result. We're already seeing the transformations that big data is causing in advertising and other situations where millions of people's activity can be measured at a time. Now data science is looking at how it can help individuals. Timely updates on a United Airways flight may be among the tamer applications. Think instead of statistical models that tell you what job to take, or alert you even before you feel ill that you may have the flu.

Driving this trend is a swelling amount of personal data available to computers. The amount of digital data being created globally is doubling every two years, and the majority of it is generated by consumers, in the form of movie downloads, VoIP calls, e-mails, cell-phone location readings, and so on, according to the consultancy IDC. Yet only about 0.5 percent of that data is ever analyzed.

"There is so much more data out there that you can afford to tailor it to the individual," says Patrick Wolfe, a statistician who studies social networks at University College, London. "Statistically, strength comes from pooling people together, but then the icing on the cake is when you individualize the findings."

99.5%

Percentage of newly created digital data that's never analyzed

For the data refineries of Silicon Valley, like Google, Facebook, and LinkedIn, the merger of big data and personal data has been a goal for some time. It creates tools advertisers can use, and it makes products that are particularly "sticky," too. After all, what's more interesting than yourself? Facebook suggests who your friends might be. Google Now gets better the more data you give it.

Exposing more personal data seems inevitable. With the huge jump in sales of smartphones packed with accelerometers, cameras, and GPS, "people have become instrumented to collect and transmit personal data," says Weigend. And that may be just the start. Already a fringe community of technophiles in what's known as the quantified-self movement have been equipping their bodies with sensors, pedometers, even implanted glucose monitors.

One technophile we feature in this report is Stephen Wolfram, the creator of the search engine Wolfram Alpha, who has for years engaged in a massive self-tracking project, cataloguing e-mails, keystrokes, even his physical movements. Wolfram is interested in predictive apps but also in the insights that large data sets can have on personal behavior, something he calls "personal analytics." Wolfram's idea is that just as his search engine tries to organize all facts about the world, "what you have to do in personal analytics is try to accumulate the knowledge of a person's life."

The holdup, says Wolfram, is that some of the most useful data isn't being captured, at least not in a way that's easily accessible. Part of the problem is technical, a lack of integration. But much data is warehoused by private companies like Facebook, Apple, and Fitbit, maker of a popular pedometer. Now, as the value of personal data becomes more apparent, fights are brewing. California legislators this year introduced a "Right to Know" bill that would require companies to reveal to individuals the "personal information" they store—in other words, a digital copy of every location trace and sighting of their IP address.

The bill is a part of a social movement that is demanding privacy and accountability but also a different economic arrangement between the people who supply the data and those who apply it. People want more of the direct benefits of big data, and this month's *MIT Technology Review* Business Report tracks the technology, apps, and business ideas with which industry is responding.

—Antonio Regalado

Emerged Technologies

Has Big Data Made Anonymity Impossible?

As the amount of data expands exponentially, nearly all of it carries someone's digital fingerprints.

● In 1995, the European Union introduced privacy legislation that defined "personal data" as any information that could identify a person, directly or indirectly. The legislators were apparently thinking of things like documents with an identification number, and they wanted them protected just as if they carried your name.

Today, that definition encompasses far more information than those European legislators could have imagined—more than all the bits and bytes in the entire world when they wrote their law 18 years ago.

Here's what happened. First, the amount of data created each year has grown exponentially: it reached 2.8 zettabytes in 2012, a number that's as gigantic as it sounds, and will double again by 2015, according to the consultancy IDC. Of that, about three-quarters is generated by individuals as they create and move digital files. A typical American office worker produces 1.8 million megabytes of data each year. That is about 5,000 megabytes a day, including downloaded movies, Word files, e-mail, and the bits generated by computers as that information is moved across mobile networks or the Internet.

Much of this data is invisible to people and seems impersonal. But it's not. What modern data science is finding is that nearly any type of data can be used, much like a fingerprint, to identify the person who created it: your choice of movies on Netflix, the location signals emitted by your cell phone, even your pattern of walking as recorded by a surveillance camera. In effect, the more data there is, the less any of it can be said to be private, since the richness of that data makes pinpointing people "algorithmically possible," says Princeton University computer scientist Arvind Narayanan.

We're well down this path already. The information we thought of as personal data in the past—our name, address, or credit card records—is already bought and sold by data brokers like Acxiom, a company that holds an average of 1,500 pieces of information on more than 500 million consumers. This is data that people put into the public domain on a survey form or when they signed up for services such as TiVo.

Acxiom uses information about the make and year of your car, your income and investments, and your age, education, and zip code to place you in one of 70 different "PersonicX" clusters, which are "summarized indicators of lifestyle, interests, and activities." Did you just finalize a divorce or become an empty nester? Such "life events," which move people from one consumer class to another, are of key interest to Acxiom and its advertising clients. The company says it can analyze its data to predict 3,000 different propensities, such as how a person may respond to one brand over another.

Yet these data brokers today are considered somewhat old-fashioned compared with Internet companies like Facebook, which have automated the collection of personal information so it can be done in real time. According to its financial filings at the time of its IPO, Facebook stores around 111 megabytes of photos and videos for each of its users, who now number more than a billion. That's 100 petabytes of personal information right there. In some European legal cases, plaintiffs have learned that Facebook's records of their interactions with the site—including text messages, things they "liked," and addresses of computers they used—run to 800 printed pages, adding up to another few megabytes per user.

In a step that's worrisome to digital-privacy advocates, offline and online data sets are now being connected to help marketers target advertisements more precisely. In February, Facebook announced a deal with Acxiom and other data brokers to merge their data, linking real-world activities to those on the Web. At a March investor meeting, Acxiom's chief science officer claimed that its data could now be linked to 90 percent of U.S. social profiles.

Such data sets are often portrayed as having been "anonymized" in some way, but the more data they involve, the less likely that is to be actually true. Mobile-phone companies, for instance, record users' loca-

65 billion

Location-tagged payments made in the U.S. annually

154 billion

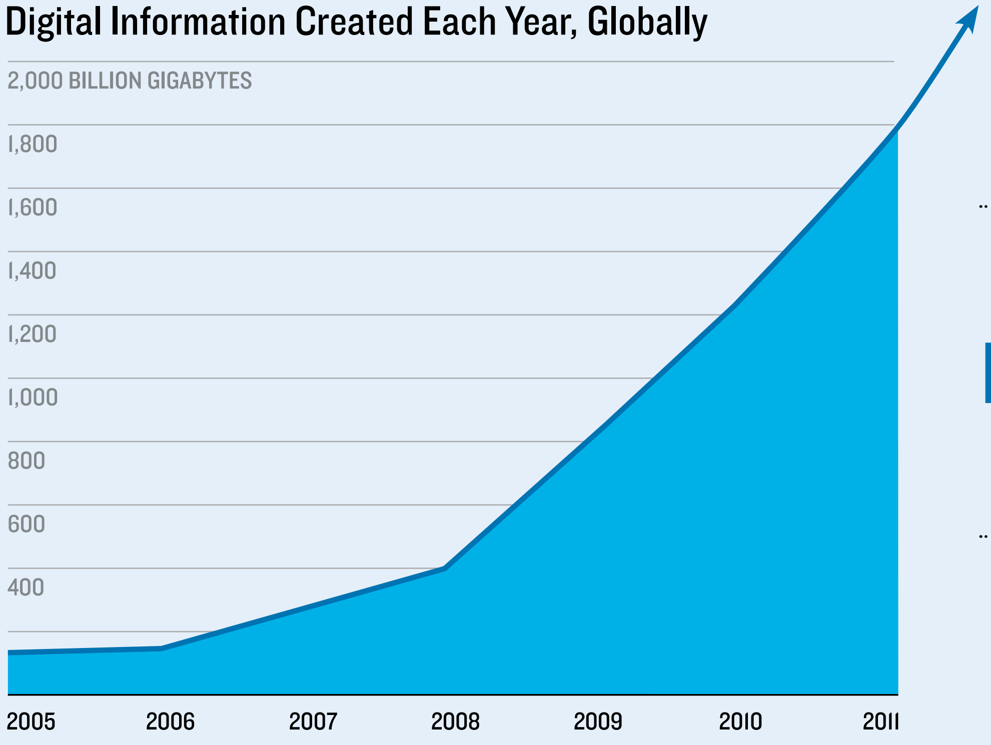
✉ ✉ ✉ ✉

E-mails sent per day

87%

U.S. adults whose location is known via their mobile phone

Digital Information Created Each Year, Globally



2,000%

Expected increase in global data by 2020

111 Megabytes

Video and photos stored by Facebook, per user

75%

Percentage of all digital data created by consumers

SOURCES: IDC, RADICATI GROUP, FACEBOOK, TR RESEARCH, PEW INTERNET tions, strip out the phone numbers, and sell aggregate data sets to merchants or others interested in people's movements. MIT researchers Yves-Alexandre de Montjoye and César A. Hidalgo have shown that even when such location data is anonymous, just four different data points about a phone's position can usually link the phone to a unique person.

The greater the amount of personal data that becomes available, the more informative the data gets. In fact, with enough data, it's even possible to discover information about a person's future. Last year Adam Sadilek, a University of Rochester researcher, and John Krumm, an engineer at Microsoft's research lab, showed they could predict a person's approximate location up to 80 weeks into the future, at an accuracy of above 80 percent. To get there, the pair mined what they described as a "massive data set" collecting 32,000 days of GPS readings taken from 307 people and 396 vehicles.

Then they imagined the commercial applications, like ads that say "Need a haircut? In four days, you will be within 100 meters of a salon that will have a $5 special at that time." Sadilek and Krumm called their system "Far Out." That's a pretty good description of where personal data is taking us. —*Patrick Tucker*

Emerged Technologies

Predictive Apps Mine Your Life

In a break from traditional software, new apps offer information proactively.

● A new type of mobile app is departing from a long-standing practice in computing. Typically, computers have just dumbly waited for their human operators to ask for help. But now applications based on machine learning software can speak up with timely information even without being directly asked for it. They might automatically pull up a boarding pass for your flight just as you arrive at the airport, or tell you that current traffic conditions require you to leave for your next meeting within 10 minutes.

The highest-profile of these apps is Google Now, which is a feature of the latest version of the Android mobile operating system and was recently added to the Google search app for the iPhone. Google Now is trained to predict when a person is about to take certain actions and offer help accordingly. It can also learn about an individual to fine-tune the assistance it offers.

Google Now's algorithms use the data in a person's Google e-mail and calendar accounts and Web searches. The app learns where you live and work and when you commute so that it can offer a virtual index card showing traffic or transit information. Other cards offer boarding passes and other handy information at appropriate times.

Bill Ferrell, founder and CEO of Osito, a company with an iPhone app that offers similar functions, calls this idea "predictive intelligence." Osito's system foretells a person's actions and needs from location, e-mail, and calendar data and uses those predictions to go beyond offering just advice. It also presents ways for a person to take action. A flight reminder will include a button to summon a cab, for example.

Now that the first generation of this type of app has been well received, engineers at Google, Osito, and elsewhere seek to wring more insights from the data they collect about their users. Osito's engineers are working to learn more from a person's past location traces to refine predictions of future activity, says Ferrell. Google Now recently began showing the weather in places it believes you're headed to soon. It can also notify you of nearby properties for

Anticipatory Systems Guess Your Next Move

Smartphone apps that mine personal data in order to anticipate a person's needs

NAME	Cue	Google Now	Osito	Tempo AI	Dark Sky
RAISED	$4.7 million	N/a	$1.1 million	Incubated at SRI International	$39,376
FOUNDED BY	Y Combinator graduates Daniel Gross and Robby Walker	An internal Google team	Bill Ferrell, a former Google developer	Raj Singh, Corey Hulen, and Thierry Donneau-Golencer	Jack Turner and Adam Grossman
PREDICTIONS	Summarizes a person's day based on information scavenged from calendar, e-mail, and documents	Directions, traffic, and weather based on a person's location and calendar	Handles transactions like checking in for a flight or calling a cab after you land at the airport	Directions to appointments. Also sends messages if you're running late	Provides minute-by-minute weather forecasts for user's exact location

SOURCES: CUE, GOOGLE NOW, OSITO, TEMPO AI, DARK SKY sale if you have recently done a Web search suggesting you're looking for a new home.

Machine learning experts at Grokr, a predictive app for the iPhone, have found they can divine the ethnicity, gender, and age of their users to a high degree of accuracy, says CEO Srivats Sampath. "That can help us predict places you might like to go better," he says. The information will be used to fine-tune the recommendations Grokr offers for restaurants and music events.

These apps benefit from improved data-mining techniques, but they're also succeeding partly because of how they are presented to users. They are not cast as artificial butlers, a staple of science fiction that Apple tried to mimic with the voice-operated app Siri in 2010. Instead, apps like Google Now are intentionally made without personality and don't pretend to be people.

That plays to the strengths of today's artificial-intelligence technology, says Mike Volpi, a partner with venture capital firm Index Ventures, which invested in a predictive iPhone app called Donna. "An assistant probably is one of the most tough use cases, because you set up the expectation it will be human-level," says Volpi.

Apple's assistant has not become a core part of many iPhone users' lives, he says, because software cannot recognize speech accurately enough. Apple may have exacerbated this problem by giving its app the capacity for witty repartee and running TV ads in which Siri appears to act with almost humanlike intelligence.

Hilary Mason, chief data scientist at the Web company Bit.ly, has mixed reviews for Google Now. She finds that it's often serving up unnecessary information: for example, she says, she doesn't need to be told that she is near a Staples office supply store, which is true in many parts of Manhattan, or be given a bus schedule every time she passes a bus stop. "It's not quite tuned to what matters to me," she says.

But still, it represents a milestone in computing, she adds: "Google Now is kind of a sucky product, but I use it anyway. It's important because it's the first time Google has taken all they know about us to make a product that makes our lives better."

—Tom Simonite

Case Studies

Data Won the Election. Can It Save the World?

Data scientist Rayid Ghani helped persuade voters to reëlect President Obama. Now he's using big data to create a groundswell of social good.

● As chief scientist for President Obama's reëlection effort, Rayid Ghani helped revolutionize the use of data in politics. During the final 18 months of the campaign, a sprawling team of data and software experts sifted, collated, and combined dozens of pieces of information on each registered U.S. voter to discover patterns that let them target fund-raising appeals and ads to those most likely to respond.

Now, with Obama again ensconced in the Oval Office, some veterans of the data squad are applying lessons from the campaign to tackle social issues such as education or health care. Edgeflip, a startup Ghani founded in January with two other campaign members, plans to turn the ad hoc data analysis tools developed for Obama for America into software that can make nonprofits more effective at raising money and recruiting volunteers.

Ghani isn't the only one thinking along these lines. In Chicago, Ghani's hometown and the site of Obama for America headquarters, some campaign members are helping the city make records of utility usage and crime statistics available so that developers can build apps that attempt to improve life there.

It's all part of a bigger idea to "engineer social systems" by scanning the numerical exhaust from mundane activities—phone calls, online searches—for patterns that might bear on everything from traffic snarls to human trafficking. Among those pursuing such humanitarian goals are startups like DataKind as well as large companies like IBM, which is redrawing bus routes in Ivory Coast, and Google, with its flu-tracking software.

Ghani, 35, always had an interest in social causes, like tutoring disadvantaged kids. But he developed his data-mining savvy during 10 years as director of analytics at Accenture, helping retail chains forecast sales, creating models of consumer behavior, and writing papers with titles like "Data Mining for Business Applications."

Before joining the Obama campaign in July 2011, Ghani wasn't even sure his expertise in machine learning or predicting online prices could have an impact on a social cause. But the campaign's success in applying such methods on the fly to sway voters is now seen as having been potentially decisive in the election's outcome. "I realized two things," says Ghani. "It's doable at the massive scale of the campaign, and that means it's doable in the context of other problems."

At Obama for America, Ghani built statistical models that assessed each voter along five axes: support for the president, susceptibility to being persuaded to support the president, and the likelihood of donating money, of volunteering, and of actually casting a vote. These models allowed the campaign to target door knocks, phone calls, TV spots, and online ads to where they were most likely to benefit Obama.

One of the most important ideas Ghani developed during the campaign, dubbed "targeted sharing," now forms the

⟶

"I can imagine policies being designed a lot more collaboratively. I don't know if the politicians are ready to deal with it."

—Rayid Ghani

basis of Edgeflip's first product. It's a Facebook app that prompts people to share information from a nonprofit, but only with those friends predicted to respond favorably. That's a big change from the usual scattershot approach of posting pleas for money or help and hoping they'll reach the right people.

As Obama's Facebook app did, Ghani says, Edgeflip will ask people who share a post to provide access to their list of friends. This will pull in not only friends' names but personal details, like ages, that can feed models of who is most likely to help.

Say a hurricane strikes the southeastern United States and the Red Cross needs clean-up workers. The app would ask Facebook users to share the Red Cross message, but only with friends who live in the storm zone, are young and likely to do manual labor, and have previously shown interest in content shared by that user. But if the same person shared a donation appeal, he or she would be prompted to pass it along to friends who are older, live farther away, and have donated money in the past.

Michael Slaby, who led Obama's technology team and who hired Ghani, sees great promise in the technique. "It's one of the most compelling innovations to come out of the campaign," he says. "It has the potential to make online activism much more efficient and effective."

For instance, Ghani has been working with Fidel Vargas, CEO of the Hispanic Scholarship Fund, to increase that organization's analytical savvy. Vargas thinks social data could predict which scholarship recipients are most likely to contribute to the fund after they graduate. "Then you'd be able to give away scholarships to qualified students who would have a higher probability of giving back," he says. "Everyone would be much better off."

Ghani sees a far bigger role for technology in the social sphere. He imagines online petitions that act like open-source software, getting passed around and improved. "I can imagine policies being designed a lot more collaboratively," he says. "I don't know if the politicians are ready to deal with it." He also thinks there's a huge amount of untapped information out there about childhood obesity, gang membership, and infant mortality, all ready for big data's touch.

But one thing stands in the way of this vision: a lack of data scientists interested in applying their skills to social problems. This summer, Ghani will be teaching at a fellowship program he developed for the University of Chicago, called Data Science for Social Good, which will put roughly 40 students to work on problems facing nonprofits and governments.

"A lot of the people who have the skills to do this kind of work end up working for Facebook, Google, or the latest online ad network," he says. "I want to show them that the same kind of data is available here, and the impact is bigger." *—Ted Greenwald*

Case Studies

Logging Life with a Lapel Camera

A startup believes people will want a photographic record of their lives, taken at 30-second intervals.

● "We want to provide people with a perfect photographic memory," says Martin Källström, CEO of Memoto. His startup, based in Linköping, Sweden, is creating a tiny clip-on camera that takes a picture every 30 seconds, capturing whatever you are looking at, and then applies algorithms to the resulting mountain of images to find the most interesting ones.

Just 36 by 36 by 9 millimeters, the inconspicuous plastic camera has a lot crammed inside. The most important component is a five-megapixel image sensor originally designed for mobile phones. An ARM 9 processor running Linux powers a program that wakes the device twice a minute; takes a picture and a reading from the GPS sensor, accelerometer, and magnetometer; and promptly puts the device back to sleep.

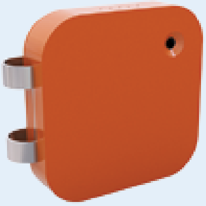
Later, a user can transfer the pictures to a computer or upload them to Memoto's cloud storage service. The pictures are then fed through an image-processing algorithm that starts to sort out the events in your day. The images are clustered by their predominant colors, and then "we get a diagram of how varied the colors are over the day," says Källström.

That processing turns your photos into "moments"—between 30 and 35 things that have happened during your day, displayed as stacks of photos in a smartphone app or on the Web. Hours in front of a computer add up to one moment, a quick coffee break to another. Each is represented by a single sharp, colorful frame—if possible, one with people in it. "It allows you, in the app, to see the good parts of your day with the boring parts hidden," says Källström.

It's this clever filtering system, Källström believes, that makes the Memoto more than just a camera. He calls it a "life logging" device that will help people remember what they've seen and experienced, or even keep a record for their descendants. "I'd like to be able to put in my will what parts of my life log are going to be available for people that come after me," says Källström. "I've always been fascinated by ways to effortlessly document life."

Källström, a 37-year-old software developer, came up with the Memoto concept in 2011 and began working on it full time the next year with partner Oskar Kalmaru and product designer Björn Wesén. Last fall, the team raised $550,189 from the public on the crowd-funding site Kickstarter, where they promised a camera to anyone who paid $279 up front.

That was far more than the $50,000 they had expected to raise. "We realized that we were going to have to build more cameras," says Källström, with typical Swedish understatement. His company now employs 17 people and he says that despite some unexpected delays in developing and producing the cameras, he expects



Memoto's clip-on camera has eight gigabytes of memory, enough to store four days' worth of photos.

COURTESY OF MEMOTO

Life Loggers

Devices and apps people are using to record aspects of their lives

NAME	AliveCor ECG	Nike+ FuelBand	Foursquare	RunKeeper	Google Glass
COST	$199	$149	Free	Free	$1,500
YEAR INTRODUCED	2013	2012	2009	2008	2013
USERS	not announced	500,000	30 million	18 million	not announced
SPECIALITY	A snap-on electrode that converts an iPhone into an electrocardiogram monitor	A digital bracelet that records physical activity, such as walking and running	An app that tracks a user's location, offering prizes and points for checking in at stores	An app that uses a smartphone's GPS to record a runner's speed and distance covered	A wearable computer that captures photos and videos

SOURCES: ALIVECOR, NIKE, FOURSQUARE, FITNESSKEEPER, GOOGLE

the first 5,000 to arrive later this year from Taiwan, where they are being assembled.

Life logging is quickly becoming a significant business as consumers embrace wearable self-tracking devices such as Nike's FuelBand, a bracelet that measures a person's movements and estimates calories burned. Sharing photos on services like Instagram or Facebook can also be considered a kind of life logging. "It's already mainstream," Källström says.

Many large technology companies are considering how to use wearable devices to collect even more personal data. Google, for instance, is now testing a head-mounted computer that can shoot video.

Recording devices that take photos, like Memoto, are going to challenge social norms and raise new privacy questions. When does recording your own life intrude on someone else's?

Stephen Wolfram, creator of the software Mathematica, has been testing an early Memoto prototype since March and says, "It's still a little bit socially weird." He adds, "I'm not completely sure what to do with the data, and it does generate a lot of data." During one day, the camera can snap 2,000 photos and generate about two gigabytes' worth of files (its flash memory has room for eight gigabytes). Wolfram says that while flipping back through the pictures, he's been able to read name tags of people at conferences whose names he'd forgotten. "I can see all kinds of stuff I didn't notice when I was actually there," he says.

Memoto is designed to automatically stop snapping pictures if it's taken off and placed on a flat surface, or put in a dark place like a pocket. Källström admits there are certain times when it's probably best to leave it at home. "Technology forces us to make new kinds of ethical judgments," he says.

The company's business model is to sell the devices and charge about $8 per month for online storage of people's photos. "It's a lot like a mirror in your bathroom, perhaps," says Källström. "You look into it in the morning, and you know a little bit more about yourself." —*Duncan Geere*

Leaders

Q&A: Stephen Wolfram on Personal Analytics

The creator of the Wolfram Alpha search engine explains why he thinks your life should be measured and analyzed.

● Don't be surprised if Stephen Wolfram, the renowned complexity theorist, software company CEO, and night owl, wants to schedule a work call with you at 9 P.M. After a decade of logging every phone call he makes and keystroke he types, Wolfram knows exactly the probability he'll be on the phone with someone at that time: 39 percent.

→

Wolfram, a British-born physicist, is the creator of the software Mathematica and of Wolfram Alpha, the nerdy "computational knowledge engine" that can tell you the distance to the moon right now, in units including light-seconds.

Now Wolfram wants to apply similar techniques to people's personal data, an idea he calls "personal analytics." He started with himself. In a blog post last year, Wolfram analyzed a detailed record of his life stretching back three decades, including hundreds of thousands of e-mails and 10 years of computer keystrokes.

Last year, his company released its first consumer product in this vein, Personal Analytics for Facebook. In under a minute, the software generates a detailed study of a person's relationships and behavior on the site. It looks like a dashboard for your life, which Wolfram says is exactly the point. In a phone call that was recorded and whose start and stop times were entered into Wolfram's life log, he discussed why personal analytics will make people more efficient at work and in their personal lives.

"You're seeing the story of people's lives played out on the level of data."

—Stephen Wolfram

What do you record about yourself?
E-mails, documents, and normally, if I was in front of my computer, it would be recording keystrokes. I have a motion sensor for the room that records when I pace up and down. Also a pedometer, and I am trying to get an eye-tracking system set up. Oh, and I've been wearing a sensor to measure my posture.

Do you think that you're the most quantified person on the planet?
I couldn't imagine that that was the case until maybe a year ago, when I collected a bunch of this data and wrote a blog post on it. I was expecting that there would be people who would come forward and say, "Gosh, I've got way more than you." But nobody's come forward. I think by default that may mean I'm it, so to speak.

You coined this term "personal analytics." What does it mean?
There's organizational analytics, which is looking at an organization and trying to understand what the data says about its operation. Personal analytics is what you can figure out applying analytics to the person, to understand the operation of the person.

Why are you analyzing Facebook data?
We are trying to feel out the market for personal analytics. Most people are not recording all their keystrokes like I am. But the one thing they are doing is leaving lots of digital trails, including on Facebook, and that is one of the pieces we've been experimenting with.

We've accumulated a lot of Facebook data—you're seeing the story of people's lives played out on the level of data. You can see relationship status as a function of age, or the evolution of the clustering of friends at different ages. It's fascinating to see how all this stuff is just right there in the data.

What's the connection to the search engine you built?
Right now Wolfram Alpha is strong on public knowledge: accumulating and searching the knowledge of the civilization. But what you have to do in personal analytics is try to accumulate the knowledge of a person's life. Then the two can actually be integrated, and I'll give a kind of silly example. You might ask: "Who do I know that can go out into their backyard and go and look at the night sky right now?" For that you have to be able to compute who is in nighttime, who doesn't have cloudy weather, and things like this. And we can compute all that stuff.

What do you see as the big applications in personal analytics?
Augmented memory is going to be very important. I've been spoiled because for years I've had the ability to search my e-mail and all my other records. I've been the CEO of the same company for 25 years and so I never changed jobs and lost my data. That's something that I think people will just come to expect. Pure memory augmentation is probably the first step.

The next is preëmptive information delivery. That means knowing enough about people's history to know what they're going to care about. Imagine someone is reading a newspaper article, and we know there is a person mentioned in it that they went to high school with, and so we can flag it. I think that's the sort of thing it's possible to dramatically automate and make more efficient.

Then there will be a certain segment of the population that will be into the self-improvement side of things, using analytics to learn about ourselves. Because when a pattern is explicit, we can decide, "Do we like that behavior, do we not?" Very early on, back in the 1990s, when I first analyzed my e-mail archive, I learned that a lot of e-mail threads at my company would, by a certain time of day, just resolve themselves. That was useful to know, because if I jumped in too early I was just wasting my time.

Are you commercializing these ideas?
The personal analytics of Facebook for Wolfram Alpha is a deployed project, and there will be more of those in the personal-analytics space. We think we can do terrific things, but you have to be able to get to the data. That has been the holdup. The data isn't readily available. Recently we've been working with different companies to try and make sure we can connect their sensors to kind of a generic analytics platform, to take people's data, move it to the cloud, and do analytics on it.

How much better can people become with some data feedback?
I think it will be fairly dramatic. It's like asking how much more money can you make if you track your portfolio rather than just vaguely remembering what investments you made.

—Antonio Regalado

COURTESY OF WOLFRAM ALPHA

Case Studies

Marketing to the Big Data Inside Us

In your genome are clues to your health, your ancestry, even your purchasing preferences.

● Companies market to you according to your shopping habits, your age, your salary, and your social-media activities. In the future, they may be able to advertise to you on the basis of your DNA.

Do you carry the genetic variants associated with lactose intolerance? Here, Lactaid has a coupon for you. The genes for male-pattern baldness? That's accelerated by stress, so maybe you should come in for a discounted massage Jimmy's Spa & Bath.

A Minneapolis-based startup called Miinome plans to build what it calls the first "member-controlled human genetic marketplace." The company, which has just three full-time employees and is hunting for financing, is notable mostly for its bold idea: to sell DNA information to marketers.

250,000

Number of people who will get their genomes analyzed in 2013

The amount of digital information about people is exploding, and DNA data is no exception. This year, as many as 250,000 people could get their genomes entirely or partly sequenced, according to some estimates. That figure, while still modest, has been climbing exponentially and will soon reach into the millions.

Still missing, though, are widely accepted ways to store, share, and mine the data for medical and commercial applications. Miinome's idea is a membership club that electronically stores the DNA data of people who have had their genomes sequenced, who then get to decide what information is made available and for what purpose. "[It] could be anything from donating their information for philanthropic purposes, such as an American Diabetes Association study, to savings on certain products," says Paul Saarinen, the advertising executive who is CEO of Miinome.

Widspread gene-based marketing isn't feasible yet, but it could be in the future. One reason is that the cost of decoding just the most important parts of a person's DNA—only those stretches that code for proteins, known as the "exome"—is only about $700, and that's expected to decline. Miinome may eventually offer to sequence the genomes of people for free, so long as they participate in its database and perhaps agree to receive a few ads.

That could take human genetics toward an ad-supported model, just like social networks and search engines. James Ostheimer, the data scientist who is Miinome's cofounder, says consumer Internet companies like Amazon and Twitter have developed the key technology that's needed—big server farms, software to handle massive amounts of personal data, and algorithms to mine it.

One technical problem is that while DNA sequencing is cheap, analyzing the data isn't. Some of the work is still done manually, by PhD scientists who sift through tens of thousands of DNA letters looking for those that might be the cause of a disease. Dietrich Stephan, CEO of SVBio, a company in Foster City, California, that has developed software to analyze genomes automatically, says the volume of data produced by modern DNA-sequencing instruments has overwhelmed hospitals and diagnostic labs.

Furthermore, although medical researchers typically want to link genetic differences to specific diseases, common diseases like diabetes are influenced by many factors, including the environment. As a result, Stephan says, making meaningful health predictions from someone's DNA will require "rigging up all data pipes," including medical records, the pollutants and infectious diseases a person has been exposed to, maybe even grocery receipts. "Until we get that data all in one place, it's going to be difficult," he says.

Sean George, a senior executive at InVitae, a San Francisco company now offering a DNA test for 270 genetic conditions to a handful of hospitals, says too few people's genomes have been sequenced—and shared—to power the kind of analysis researchers really want to do. "At a low number, that data is not super useful," he says. "We do believe there is a tipping point [coming]."

The entrepreneurs behind Miinome think consumers will start to share their DNA data when there are applications beyond medicine. "In general, we'd like our members to hook up as many feeds as they want to," says Ostheimer. He thinks if people link their DNA to their Twitter or Facebook feeds it could be possible to "look for new associations"—say, between certain genes and a taste for spicy foods.

Scientifically, that sounds like a stretch. But commercially it might not be. "That's an interesting marketing opportunity [for] people selling groceries or trying to get you to eat in their restaurant," says Ostheimer.

—Susan Young

Case Studies

Intel Fuels a Rebellion Around Your Data

The world's largest chip maker wants to see a new kind of economy bloom around personal data.

● Intel is a $53-billion-a-year company that enjoys a near monopoly on the computer chips that go into PCs. But when it comes to the data that underlies big companies like Facebook and Google, it says it wants to "return power to the people."

Intel Labs, the company's R&D arm, is launching an initiative around what it calls the "data economy"—or how consumers might capture more of the value of their personal information like digital records of their location or work history. To make this possible, Intel is funding hackathons to urge developers to explore novel uses of personal data. It has also paid for a rebellious-sounding website called We the Data (wethedata.org) featuring raised fists and stories comparing Facebook to Exxon Mobil.

Intel's effort to stir a debate around "your data" is just one example of how some companies—and society more broadly—are grappling with a basic economic asymmetry of the big data age: they've got the data, and we don't.

"As consumers, we have no right to know what companies know about us. As companies, we have few restrictions on what we can do with this data." —Hilary Mason, chief data scientist, Bit.ly

Valuable data about consumers is being concentrated by Internet firms, like Google and Amazon.com, at an unprecedented scale as people click around the Web. But regulations and social standards haven't kept up with the technical and economic shift, creating a widening gap between data haves and have-nots.

"As consumers, we have no right to know what companies know about us. As companies, we have few restrictions on what we can do with this data," says Hilary Mason, chief data scientist at Bit.ly, a social media company in New York. "Even though people derive value, and companies derive value, it's totally chaotic who has rights to what, and it's making people uncomfortable."

In February, for instance, legislators in California introduced the first U.S. law to give individuals a complete view into their online personas. The "Right to Know" bill would let citizens of the state demand a detailed copy of all the information stored on them—and whom it had been shared with—by companies like LinkedIn and Google.

That bill quickly got shelved under pressure from lobbyists for technology companies who called it "unworkable" and financially damaging to Internet firms, and said lawmakers don't understand "how the Internet works." Some of the data covered in the bill, like a computer's IP address, or location, is so basic to communication between machines on the Internet that companies admitted they don't even know where it ends up.

And that's the wider dilemma: our personal data is inextricably tied to "big data"—those far larger data sets that now power many of the online services we use. If you don't tell a navigation app where you are, it can't tell you where to turn, or tell others there's traffic ahead. One doesn't work without the other. What's more, the economic importance of products fueled with personal data is growing rapidly. According to the Boston Consulting Group, as methods for basing transactions on a person's digital records have spread from banks to retailers and other sectors, the financial value that companies derived from personal data in Europe was $72 billion in 2011. The consultants concluded that "personal data has become a new form of currency."

Yet that doesn't mean it's a currency that's easily understood or traded on by individuals. Although a few startups have attempted to help individuals monetize their personal facts, the truth is that information about people's identity and habits has financial value mostly in the aggregate. A single user's value to Facebook, for instance, is only about $5 a year. Mason, the Bit.ly executive, says trying to put a value on one person's data is like calculating the value of one unmatched shoe. "And here we are talking about sets of millions or billions of shoes," she says. "I just don't think that data plays by the economics of any goods we are familiar with."

Some believe the market may have already found the right economic balance. "It seems like we have a working model where companies own our data and we're okay with that because of the free stuff, personalization, and convenience we get in return," says Gam Dias, CEO of First Retail, an e-commerce consulting company. "There's not a lot I'm going to do with my extra data anyway. I already know who I am and what I want."

Intel this year judged the questions swirling around personal data important enough to launch a "Data Economy Initiative," a multiyear study whose goal is to explore new uses of technology that might let people benefit more directly, and in new ways, from their own data, says Ken Anderson, a cultural anthropologist who is in charge of the project.

Anderson, who once helped Apple develop the sliding application bar that appears on Mac computers (after studying how people organized their desks and stacked items on shelves), says Intel believes technology based on personal data may end up in the control of individuals, in much the same way that mainframe computers gave way to PCs. "It doesn't matter what you look at in terms of technology. Usually, there is this move toward individualization," he says.

Intel, which has started surveying consumer opinions, has also been supporting efforts like a competition in New York last fall in which developers wrote apps for single mothers and the elderly. It's also underwriting the National Day of Civic Hacking, an event focused on new uses of municipal data being released by city governments, such as records of health inspections.

$72 billion

Commercial value of personal data in Europe, 2011

It's too early to say just what kind of products might result for Intel, Anderson says.

"When you talk about the data economy, it's really something that doesn't yet exist," he says. "There are people who [are] trying to control a lot of your personal data. But that's not an economy, that's just profit for one company."

—Antonio Regalado and Jessica Leber

IMAGINE
HEALTHCARE LIKE NEVER BEFORE
SAP®

Emerged Technologies

Augmenting Social Reality in the Workplace

A new line of research examines what happens in an office where the positions of the cubicles and walls—even the coffee pot—are all determined by data.

● Can we use data about people to alter physical reality, even in real time, and improve their performance at work or in life? That is the question being asked by a developing field called augmented social reality.

Here's a simple example. A few years ago, with Sandy Pentland's human dynamics research group at MIT's Media Lab, I created what I termed an "augmented cubicle." It had two desks separated by a wall of plexiglass with an actuator-controlled window blind in the middle. Depending on whether we wanted different people to be talking to each other, the blinds would change position at night every few days or weeks.

The augmented cubicle was an experiment in how to influence the social dynamics of a workplace. If a company wanted engineers to talk more with designers, for example, it wouldn't set up new reporting relationships or schedule endless meetings. Instead, the blinds in the cubicles between the groups would go down. Now as engineers passed the designers, it would be easier to have a quick chat about last night's game or a project they were working on.

Human social interaction is rapidly becoming more measurable at a large scale, thanks to always-on sensors like cell phones. The next challenge is to use what we learn from this behavioral data to influence or enhance how people work with each other. The Media Lab spinoff company I run uses ID badges packed with sensors to measure employees' movements, their tone of voice, where they are in an office, and whom they are talking to. We use data we collect in offices to advise companies on how to change their organizations, often through physical changes to the work environment. For instance, after we found that people who ate in larger lunch groups were more productive, Google and other technology companies that depend on serendipitous interaction to spur innovation installed larger cafeteria tables.

In the future, some of these changes could be made in real time. At the Media Lab, Pentland's group has shown how tone of voice, fluctuation in speaking volume, and speed of speech can predict things like how persuasive a person will be in, say, pitching a startup idea to a venture capitalist. As part of that work, we showed that it's possible to digitally alter your voice so that you sound more interested and more engaged, making you more persuasive.

Another way we can imagine using behavioral data to augment social reality is a system that suggests who should meet whom in an organization. Traditionally that's an ad hoc process that occurs during meetings or with the help of mentors. But we might be able to draw on sensor and digital communication data to compare actual communication patterns in the workplace with an organizational ideal, then prompt people to make introductions to bridge the gaps. This isn't the LinkedIn model, where people ask to connect to you, but one where an analytical engine would determine which of your colleagues or friends to introduce to someone else. Such a system could be used to stitch together entire organizations.

Unlike augmented reality, which layers information on top of video or your field of view to provide extra information about the world, augmented social reality is about systems that change reality to meet the social needs of a group.

For instance, what if office coffee machines moved around according to the social context? When a coffee-pouring robot appeared as a gag in TV commercial two years ago, I thought seriously about the uses of a coffee machine with wheels. By positioning the coffee robot in between two groups, for example, we could increase the likelihood that certain coworkers would bump into each other. Once we detected—using smart badges or some other sensor—that the right conversations were occurring between the right people, the robot could move on to another location. Vending machines, bowls of snacks—all could migrate their way around the office on the basis of social data. One demonstration of these ideas came from a team at Plymouth University in the United Kingdom. In their "Slothbots" project, slow-moving robotic walls subtly change their position over time to alter the flow of people in a public space, constantly tuning their movement in response to people's behavior.

The large amount of behavioral data that we can collect by digital means is starting to converge with technologies for shaping the world in response. Will we notify people when their environment is being subtly transformed? Is it even ethical to use data-driven techniques to persuade and influence people this way? These questions remain unanswered as technology leads us toward this augmented world.

—Ben Waber, cofounder and CEO of Sociometric Solutions

Leaders

The Dictatorship of Data

Robert McNamara epitomized the hyper-rational executive led astray by numbers.

● Big data is poised to transform society, from how we diagnose illness to how we educate children, even making it possible for a car to drive itself. Information is emerging as a new economic input, a vital resource. Companies, governments, and even individuals will be measuring and optimizing everything possible.

But there is a dark side. Big data erodes privacy. And when it is used to make predictions about what we are likely to do but haven't yet done, it threatens freedom as well. Big data also exacerbates a very old problem: relying on the numbers when they are far more fallible than we think. Nothing underscores the consequences of data analysis gone awry more than the story of Robert McNamara.

McNamara was a numbers guy. Appointed the U.S. secretary of defense when tensions in Vietnam rose in the early 1960s, he insisted on getting data on everything he could. Only by applying statistical rigor, he believed, could decision makers understand a complex situation and make the right choices. The world in his view was a mass of unruly information that—if delineated, denoted, demarcated, and quantified—could be tamed by human hand and fall under human will. McNamara sought Truth, and believed that Truth could be found in data. Among the numbers that came back to him was the "body count."

McNamara developed his love of numbers as a student at Harvard Business School and then as its youngest assistant professor at age 24. He applied this rigor during the Second World War as part of an elite Pentagon team called Statistical Control, which brought data-driven decision making to one of the world's largest bureaucracies. Before this, the military was blind. It didn't know, for instance, the type, quantity, or location of spare airplane parts. Data came to the rescue. Just making armament procurement more efficient saved $3.6 billion in 1943. Modern war demanded the efficient allocation of resources; the team's work was a stunning success.

At war's end, the members of this group offered their skills to corporate America. Ford Motor Company was floundering, and a desperate Henry Ford II handed them the reins. Just as they knew nothing about the military when they helped win the war, so too were they clueless about cars. Still, the so-called "Whiz Kids" turned the company around.

McNamara rose swiftly up the ranks, trotting out a data point for every situation. Harried factory managers produced the figures he demanded—whether they were correct or not. When an edict came down that all inventory from one car model must be used before a new model could begin production, exasperated line managers simply dumped excess parts into a nearby river. The joke at the factory was that a fellow could walk on water—atop rusted pieces of 1950 and 1951 cars.

McNamara epitomized the hyper-rational executive who relied on numbers rather than sentiments, and who could apply his quantitative skills to any industry he turned them to. In 1960 he was named president of Ford, a position he held for only a few weeks before being tapped to join President Kennedy's cabinet as secretary of defense.

As the Vietnam conflict escalated and the United States sent more troops, it became clear that this was a war of wills, not of territory. America's strategy was to pound the Viet Cong to the negotiation table. The way to measure progress, therefore, was by the number of enemy killed. The body count was published daily in the newspapers. To the war's supporters, it was proof of progress; to critics, evidence of its immorality. The body count was the data point that defined an era.

McNamara was a numbers guy who saw the world as a mass of unruly information. The "body count" was the data point that defined his era.

McNamara relied on the figures, fetishized them. With his perfectly combed-back hair and his flawlessly knotted tie, McNamara felt he could comprehend what was happening on the ground only by staring at a spreadsheet—at all those orderly rows and columns, calculations and charts, whose mastery seemed to bring him one standard deviation closer to God.

In 1977, two years after the last helicopter lifted off the rooftop of the U.S. embassy in Saigon, a retired Army general, Douglas Kinnard, published a landmark survey called *The War Managers* that revealed the quagmire of quantification. A mere 2 percent of America's generals considered the body count a valid way to measure progress. "A fake—totally worthless," wrote one general in his comments. "Often blatant lies," wrote another. "They were grossly exaggerated by many units primarily because of the incredible interest shown by people like McNamara," said a third.

The use, abuse, and misuse of data by the U.S. military during the Vietnam war is a troubling lesson about the limitations of information as the world hurls toward the big-data era. The underlying data can be of poor quality. It can be biased. It can be misanalyzed or used misleadingly. And even more damningly, data can fail to capture what it purports to quantify.

We are more susceptible than we may think to the "dictatorship of data"—that is, to letting the data govern us in ways that may do as much harm as good. The threat is that we will let ourselves be mindlessly bound by the output of our analyses even when we have reasonable grounds for suspecting that something is amiss. Education seems on the skids? Push standardized tests to measure performance and penalize teachers or schools. Want to prevent terrorism? Create layers of watch lists and no-fly lists in order to police the skies. Want to lose weight? Buy an app to count every calorie but eschew actual exercise.

The dictatorship of data ensnares even the best of them. Google runs everything according to data. That strategy has led to much of its success. But it also trips up the company from time to time. Its cofounders, Larry Page and Sergey Brin, long insisted on knowing all job candidates' SAT scores and their grade point averages when they graduated from college. In their thinking, the first number measured potential and the second measured achievement. Accomplished managers in their 40s were hounded for the scores, to their outright bafflement. The company even continued to demand the numbers long after its internal studies showed no correlation between the scores and job performance.

41

Shades of blue tested by Google to see which ones people used most.

Google ought to know better, to resist being seduced by data's false charms. The measure leaves little room for change in a person's life. It counts book smarts at the expense of knowledge. Google's obsession with such data for HR purposes is especially queer considering that the company's founders are products of Montessori schools, which emphasize learning, not grades. By Google's standards, neither Bill Gates nor Mark Zuckerberg nor Steve Jobs would have been hired, since they lack college degrees.

Google's deference to data has been taken to extremes. To determine the best color for a toolbar on the website, Marissa Mayer, when she was one of Google's top executives before going to Yahoo, once ordered staff to test 41 gradations of blue to see which ones people used more. In 2009, Google's top designer, Douglas Bowman, quit in a huff because he couldn't stand the constant quantification of everything. "I had a recent debate over whether a border should be 3, 4 or 5 pixels wide, and was asked to prove my case. I can't operate in an environment like that," he wrote on a blog announcing his resignation. "When a company is filled with engineers, it turns to engineering to solve problems. Reduce each decision to a simple logic problem. That data eventually becomes a crutch for every decision, paralyzing the company."

This is the dictatorship of data. And it recalls the thinking that led the United States to escalate the Vietnam war partly on the basis of body counts, rather than basing decisions on more meaningful metrics.

"It is true enough that not every conceivable complex human situation can be fully reduced to the lines on a graph, or to percentage points on a chart, or to figures on a balance sheet," said McNamara in a speech in 1967, as domestic protests were growing. "But all reality can be reasoned about. And not to quantify what can be quantified is only to be content with something less than the full range of reason." If only the right data were used in the right way, not respected for data's sake.

Robert Strange McNamara went on to run the World Bank throughout the 1970s, then painted himself as a dove in the 1980s. He became an outspoken critic of nuclear weapons and a proponent of environmental protection. Later in life he produced a memoir, *In Retrospect*, that criticized the thinking behind the war and his own decisions as secretary of defense. "We were wrong, terribly wrong," he famously wrote.

But McNamara, who died in 2009 at age 93, was referring to the war's broad strategy. On the question of data, and of body counts in particular, he remained unrepentant. He admitted that many of the statistics were "misleading or erroneous." "But things you can count, you ought to count. Loss of life is one."

Big data will be a foundation for improving the drugs we take, the way we learn, and the actions of individuals. However, the risk is that its extraordinary powers may lure us to commit the sin of McNamara: to become so fixated on the data, and so obsessed with the power and promise it offers, that we fail to appreciate its inherent ability to mislead.

—Kenneth Cukier and Viktor Mayer-Schönberger.

Reprinted with permission from Big Data: A Revolution That Will Transform How We Live, Work, and Think, *Houghton Mifflin Harcourt, 2013.*

What We Are Reading

REPORTS

Rethinking Personal Data
World Economic Forum, 2013
Prepared by the Boston Consulting Group

The Value of Our Digital Identity
Boston Consulting Group, 2012
John Rose, Olaf Rehse, Bjorn Rober

NetMob 2013 Program and Abstracts
MIT Media Lab, 2013

The New Data Democracy: How Big Data Will Revolutionize the Lives of Small Business Owners and Consumers
Intuit Emergent Research, 2012

Big Data for Development: Challenges & Opportunities
United Nations, 2012
Emmanuel Letouzé

BOOKS

Big Data: A Revolution That Will Transform How We Live, Work, and Think
Houghton Mifflin Harcourt, 2013
Viktor Mayer-Schonberger and Kenneth Cukier

The Victory Lab: The Secret Science of Winning Campaigns
Crown, 2012
Sasha Issenberg

The Human Face of Big Data
Against All Odds Productions, 2012
Rick Smolan and Jennifer Erwitt

Who We Are Following

Jure Leskovek, professor of computer science, Stanford University, @**jure**

Rayid Ghani, chief scientist, Obama for America, @**rayidghani**

Michael Cavaretta, head of predictive analytics, Ford Motor Company, @**mjcavaretta**

Pierre Levy, professor of collective intelligence, University of Ottawa, @**plevy**

Ben Lorica, chief data scientist, O'Reilly Media, @**bigdata**

Doug Laney, vice president of research, Gartner, @**Doug_Laney**

Special Focus

Mobile Phones: Big Data on People

Cell phones have become powerful and nearly universal sensors of human behavior: where people are, what they're doing, and what they're planning to do next. Fast-developing methods for collecting and analyzing these types of "big data" are creating new questions around privacy, but also promise valuable insights to retailers, urban planners, and for individuals.

Emerged Technologies

Smartphone Tracker Gives Doctors Remote Viewing Powers

Not taking your pills? Here's a smartphone technology that alerts a doctor when patients are headed for trouble.

● At the Forsyth Medical Center in Winston-Salem, North Carolina, nurses can see into the lives of some diabetes patients even when they're not at the clinic. If a person starts acting lethargic, or making lengthy calls to their mom, a green box representing them on an online dashboard becomes yellow, then red. Soon, a nurse will call to see if the patient is still taking his medication.

This novel way of keeping tabs on patients is one of several studies of an app called Ginger.io taking place at hospitals in the United States. Once installed on a patient's smartphone, the app silently logs data about what they do and where they go. It's looking for signs that something in their life has changed.

The company, also called Ginger.io, that makes the app was spun out of MIT's Media Lab in 2011 from a group that applies computer algorithms to mobile phone data to learn about the health of individuals and entire populations. That work, sometimes called "reality mining," has shown that shifts in how people use their phones or where they go can reflect the onset of a common cold, anxiety, or stress.

Anmol Madan, cofounder and CEO of Ginger.io, says that research suggested a new, inexpensive way to automate monitoring of people with conditions like diabetes or mental illnesses. They generally care for themselves, taking drugs at home, but often stop taking medication if they get depressed, then run up medical bills when they have to see a doctor.

The Ginger.io app doesn't diagnose patients directly. But it does warn that a person's behavior has changed in ways that are linked to what doctors call "noncompliance" with the drugs they're supposed to take. With the app silently logging those changes, says Madan, "now the doctor or nurse can get a sense of the patient's life and help as needed."

Ginger.io is available in both the Android and Apple app stores but can only be activated by a hospital or health care company. Once installed, Ginger.io takes about a week to record the normal patterns of a person's life. It collects motion data from a phone's accelerometers; notes the places a person visits; and logs the timing, duration, and recipients of phone calls and patterns in text messaging. After that, algorithms watch for any significant deviations and notify hospital staff if they occur.

One of the hospitals and health-care companies testing Ginger.io is Novant Health, which operates the Forsyth Medical Center and 13 other medical centers across the Carolinas and Virginia. Matthew Gymer, Novant's director of innovation, says his group approached Ginger.io because it wanted an "early warning system" that could cut the number of times patients came to its clinics. Gymer declined to say how many patients are involved in Novant's trial, but Madan says tests of Ginger.io typically start with hundreds of patients.

Eight months after installing Ginger.io on the smartphones of some diabetes patients, it's still too soon to say what the financial and health effects have been. But Gymer says patients "love the app because they have quicker access to caregivers" and says Novant is considering expanding the test to patients with heart problems or chronic back pain.

In the Novant trial, nurses respond to alerts generated by the app, but Madan says his company is now working on how to automate interaction with patients, too. For people with Crohn's disease, or inflammation of the gastrointestinal tract, an automated message might ask if a person is experiencing stomach pain, which may be an effective way to catch people on the verge of a flare-up.

Other automatic responses to Ginger.io warnings being tested by the company are a bit further from the methods of traditional medicine. "We've also explored the idea of sending a funny picture," says Madan, "or messaging a patient's friend suggesting they call." —*Tom Simonite*

Case Studies

African Bus Routes Redrawn Using Cell-Phone Data

The largest-ever release of mobile-phone data yields a model for fixing bus routes.

● Researchers at IBM, using movement data collected from millions of cell-phone users in Ivory Coast in West Africa, have developed a new model for optimizing an urban transportation system.

The IBM model prescribed changes in bus routes around Abidjan, the nation's largest city. These changes—based on people's movements as discerned from cell-phone records—could, in theory, reduce travel times by 10 percent.

While the results were preliminary, they point to the new ways that urban planners can use cell-phone data to design infrastructure, says Francesco Calabrese, a researcher at IBM's research lab in Dublin, and a coauthor of a paper on the work. "This represents a new front with a potentially large impact on improving urban transportation systems," he says. "People with cell phones can serve as sensors and be the building blocks of development efforts."

"This represents a new front. People with cell phones can serve as sensors and be the building blocks of development efforts."—Francesco Calabrese, IBM research lab, Dublin

The IBM work was done as part of a research challenge dubbed Data for Development, in which the telecom giant Orange released 2.5 billion call records from five million cell-phone users in Ivory Coast. The records were gathered between December 2011 and April 2012. The data release is the largest of its kind ever done. The records were cleaned to prevent anyone identifying the users, but they still include useful information about these users' movements. The IBM paper is one of scores being aired later this week at a conference at MIT.

The IBM work centered on Abidjan, where 539 large buses are supplemented by 5,000 mini-buses and 11,000 shared taxis. The IBM researchers studied call records from about 500,000 phones with data relevant to the commuting question.

Mobility data is created when someone uses a phone for a call or text message. That action is registered on a cell-phone tower and serves as a report on the user's general location somewhere within the tower's radius. The person's movement is then ascertained as the call is transferred to a new tower or when a new call is made that connects to a different tower.

While the data is rough—and of course not everyone on a bus has a phone or is using it—routes can be gleaned by noting the sequence of connections. And IBM and other groups have found that these mobile phone "traces" are accurate enough to serve as a guide to larger population movements for applications such as epidemiology and transportation (see "Big Data from Cheap Phones," below).

Cell-phone data promises to be a boon for many industries. Other research groups are using similar data sets to develop credit histories based on a person's movements and phone-based transactions, to detect emerging ethnic conflicts, and to predict where people will go after a natural disaster to better serve them when one strikes.

To do such tasks in the developing world, there may be little or no other data to work with. Owners of smartphones that have GPS can allow apps like Google Maps to use their location data for traffic sensing information shared with others. But location information on the simple phones that are far more prevalent in the developing world is known only to the mobile carriers. And that data is available only by special arrangement with the carriers.

In the case of transportation, improving roads and public transit systems often depends on labor-intensive work such as the traveler surveys done commonly in the rich world. "The cost of traditional surveys is very high for developing world applications, but cell-phone use is high, so cell traces are a terrific data opportunity. This is a valuable investigative effort," says Kara Kockelman, a transportation researcher at the University of Texas, Austin.

Indeed, while in a number of past studies, mobile phone data was used to infer travel routes and demand, IBM says this was the first time such data was used in an effort to actually optimize a city transit network.

IBM calls its model AllAboard. For Abidjan, the model selected among 65 possible improvements to conclude that adding two routes and extending an existing one would do the most to optimize the system, with a 10 percent time savings for commuters.

Of course, unclogging one transportation route can have unanticipated problems, like attracting more people to use that route, perpetuating the problem. "If travel times noticeably fall on many roadways, many travelers may shift back to peak times" and popular roadways, Kockelman says.

Still, if the data about people's movements were available in real-time—rather than months after it was created—the results could be even more powerful. "This would provide snapshots of people moving around in a city, allowing the optimal shifting of routes, and reducing travel and wait times," Calabrese says.

—David Talbot

Emerged Technologies

Big Data from Cheap Phones

Collecting and analyzing information from simple cell phones can provide surprising insights into how people move about and behave—and even help us understand the spread of diseases.

● At a computer in her office at the Harvard School of Public Health in Boston, epidemiologist Caroline Buckee points to a dot on a map of Kenya's western highlands, representing one of the nation's thousands of cell-phone towers. In the fight against malaria, Buckee explains, the data transmitted from this tower near the town of Kericho has been epidemiological gold.

SAP Makes Big Data Real – And Real-Time

By Steve Lucas
Executive Vice President, Database and Technology
SAP®

The theme of this MIT Technology Review special report—"Big Data Gets Personal"—fits perfectly with SAP's view of this important new direction in technology. We regard Big Data as nothing less than enabling people to reimagine the world by tracking signals that have previously been impossible to track. We're already using Big Data at SAP to change how we work, how we play, and how we literally stay alive. We're helping major-league sports teams tap Big Data to reshape the game experience, letting fans use their mobile phones to follow their favorite players and plays, whether they're in their homes or at the ballpark. At Burberry's, when an established customer with a loyalty card walks in, a sales associate will know their ordering history and be able to escort them on a shopping tour personalized to their tastes. And soon, cancer patients will routinely be able to have their DNA tested to help determine the specific treatment that will be most effective for their condition, with the fewest side effects.

Many people, when they hear "Big Data," think "big business." Not SAP. Via our Startup Focus program, we are actively working with young companies, helping them take advantage of our revolutionary HANA in-memory Big Data platform. We're discovering that some of the most adroit users of Big Data can be five-person startups.

And of course many large companies are also launching their own Big Data projects. But when we ask Fortune 1000 CIOs about Big Data, we often hear the same thing: "Well, we've got a lot of data, but we haven't figured out a way to translate it into real results."

The reason? Many people have an extremely skewed view of Big Data. They associate it exclusively with Hadoop, and they assume that "doing Big Data" is a simple matter of setting up a Hadoop cluster.

Of course, we at SAP are big believers in Hadoop, but it addresses only one portion of a Big Data strategy: storing data in a low-cost archive, often one that can only be accessed in batch jobs that might take all night to run. The problem is that Big Data isn't just about storing information—it's also about extracting useful insights from it in real time. And for that, you need much more than Hadoop.

In fact, at the big Internet companies most associated with Big Data—Google, Facebook, and the like—IT budgets and innovation are mostly dedicated not to Hadoop, but to a more comprehensive platform that acquires, accelerates and analyzes insights to meet their business needs. They don't just use Big Data to generate insights —they also connect these insights directly into their business processes in real time.

SAP HANA was the brainchild of Dr. Hasso Plattner, co-founder of SAP AG and the Chairman of the Supervisory Board, and Dr. Vishal Sikka, member of Executive Board, Technology and Innovation. In his many public appearances, Dr. Sikka has often stressed that SAP needs to go outside its own walls and harness the collective imagination of all of us to realize the full potential of real time and big data.

SAP's commitment to taking Big Data beyond the simple commodity of Hadoop is exemplified by our HANA platform, which takes advantage of a new generation of columnar databases running on multicore processors. The entire system—both application and data—is in RAM memory, and the resulting speed increases are startling. HANA users routinely tell us that data queries that used to take days can now be executed in a few seconds.

HANA is changing the way companies do business. Our enterprise customers don't have to separate data anymore into two silos, with servers for real-time OLTP and data warehouses for longer-term OLAP. With HANA, accounting can store receivable information while business units seek out customer patterns, all on the same platform.

Having instant, actionable information is crucial for companies competing in a global marketplace, and SAP's enterprise customers have enthusiastically embraced HANA. But the platform is sufficiently adaptable that even small startups not normally associated with SAP are finding that HANA can easily handle data-processing tasks that would make traditional systems choke and sputter.

One example is Feedzai, a Silicon Valley company using HANA to perform real-time analytics on credit card transactions, to spot and block fraudulent transactions as they are occurring. That's a big improvement over most current analytic products, which only flag problematic transactions for follow-up by a human help desk.

EasyAsk, a Massachusetts startup, is using HANA to allow enterprise customers to ask questions about ERP data verbally, in plain language, and get an immediate answer—in a cheerful voice.

NextPrinciples, a next-generation Silicon Valley social media startup, uses HANA to help companies monitor the effectiveness and reach of their social media campaigns. Companies can learn in real time what is working and what isn't, and shift strategies accordingly.

And we at SAP are especially proud of our work with MIBS, an Indian company working with point-of-sale data from thousands of pharmacies to provide early warnings of disease outbreaks in India—and thus help public health officials better prepare for them.

Improving the health of an entire country: Is there any better use of Big Data? It's one of countless examples of how SAP is making Big Data a reality for companies both small and large. Real data, with real insights, in real time for your real customers. That's as personal as it gets.

When she and her colleagues studied the data, she found that people making calls or sending text messages originating at the Kericho tower were making 16 times more trips away from the area than the regional average. What's more, they were three times more likely to visit a region northeast of Lake Victoria that records from the health ministry identified as a malaria hot spot. The tower's signal radius thus covered a significant waypoint for transmission of malaria, which can jump from human to human via mosquitoes. Satellite images revealed the likely culprit: a busy tea plantation that was probably full of migrant workers. The implication was clear, Buckee says. "There will be a ton of infected [people] there."

"The exponential adoption of mobile phones in low-income settings will lead to new technological tools that could change everything."—Vincent Blondel, University of Louvain, Belgium

This work is now feeding into a new set of predictive models she is building. They show that even though malaria cases were seen at the tea plantation, taking steps to control malaria there would have less effect on the disease's spread than concentrating those efforts at the source: Lake Victoria. That region has long been understood as a major center of malaria, but what hasn't been available before is detailed information about the patterns of human travel there: how many people are coming and going, when they're arriving and departing, which specific places they're coming to, and which of those destinations attract the most people traveling on to new places.

Existing efforts to gather that kind of travel data are spotty at best; sometimes public-health workers literally count people at transportation hubs, Buckee says, or nurses in far-flung clinics ask newly diagnosed malaria victims where they've been recently. "At many border crossings in Africa, they keep little slips of paper—but the slips get lost, and nobody keeps track," she says. "We have abstractions and general models on travel patterns but haven't been able to do this properly—ever."

The data mining will help inform the design of new measures that are likely to include cheap, targeted campaigns of text messages—for example, warning visitors entering the Kericho tower's signal zone to use bed netting. And it will help officials choose where to focus mosquito control efforts in the malarial areas. "You don't want to be spraying every puddle for mosquito larvae all the time. But if you know there is a ton of importation from a certain spot, you want to increase your control program at that spot," Buckee says. "And now I can pinpoint where the importation of a disease is especially important."

Buckee's most recent study, published last year in *Science* and based on records from 15 million Kenyan phones, is a result of a collaboration with her husband, Nathan Eagle, who has been working to make sense of cell-phone data for more than a decade. In the mid-2000s, after getting attention for his work mining data from the phones of volunteers at MIT, Eagle started to get calls from mobile carriers asking for insight into questions like why customers canceled their phone plans. Eagle began working with them. And when the couple spent 18 months in Africa starting in 2006—Buckee was doing work on the genetics of the malaria parasite—he studied call data for various purposes, trying to understand phenomena like ethnic divisions in Nairobi slums and the spread of cholera in Rwanda. Buckee's results show what might be possible when the technology is turned on public-health problems. "This demonstrated 'Yeah, we can really provide not just insight, but actually something that is actionable,'" says Eagle, now CEO of Jana, which runs mobile-phone surveys in the developing world. "This really does work."

That demonstration suggests how such data might be harnessed to build tools that health-care workers, governments, and others can use to detect and monitor epidemics, manage disasters, and optimize transportation systems. Already, similar efforts are being directed toward goals as varied as understanding commuting patterns around Paris and managing festival crowds in Belgium. But mining phone records could be particularly useful in poor regions, where there's often little or no other data-gathering infrastructure. "We are just at the start of using this data for these purposes," says Vincent Blondel, a professor of applied mathematics at the University of Louvain in Belgium and a leading researcher on data gleaned from cell phones. "The exponential adoption of mobile phones in low-income settings—and the new willingness of some carriers to release data—will lead to new technological tools that could change everything."

BLANK SLATE

The world's six billion mobile phones generate huge amounts of data—including location tracking and information on commercial activity, search history, and links in social networks. Innumerable efforts to mine the data in different ways are under way in research and business organizations around the world. And of those six billion phones, five billion are in developing countries. Many of them are cheap phones that can do little besides make calls and send text messages. But all such activity can be tracked back to cell-phone towers, providing a rough way to trace a person's movements. Throw in the spread of mobile payment technology for simple commerce and you have the raw material for insights not only into epidemiology but into employment trends, social tensions, poverty, transportation, and economic activity.

The prospect of mining data from phones is especially tantalizing in poor countries, where detailed, up-to-date information on these matters has been scarce. "In the developing world, there isn't a functioning census, you don't know where traffic is, you don't always have the data-gathering infrastructure of government," says Alex "Sandy" Pentland, director of the Human Dynamics Lab at MIT, who has long been interested in insights from data created by mobile-phone use. "But all of a sudden, the one thing you do have—cell phones everywhere, especially in the past few years—can give you the equivalent of all that infrastructure already built in the developed world."

When a call connects to a given base station, that station logs the ID number of the phone and the duration of the call; over time, this information can be used to get a sense of people's regional movements and the shape of their social networks. Purchasing history on phones is also invaluable: records of agricultural purchases could be used to predict food supplies or shortages. And financial data collected by mobile payment systems can build credit histories and help millions of people without access to banking qualify for conventional loans. "The database analysis methods and the computers are very standard," Pentland says. "It's a matter of doing science and finding the right patterns." Certain mobility patterns might relate to the spread of a disease; purchasing patterns could signify that a person has had a change in employment; behavioral changes or movement patterns might relate to the onset of an illness.

> **"There are risks and benefits of having a data-driven society. There is a question of who owns the data and who controls it. "**—Alex "Sandy" Pentland, director of the Human Dynamics Lab at MIT

A powerful demonstration of how useful data from cheap phones can be came after the January 2010 earthquake in Haiti, which killed more than 200,000 people. Researchers at Sweden's Karolinska Institute obtained data from Digicel, Haiti's largest mobile carrier. They mined the daily movement data from two million phones—from 42 days before the earthquake to 158 days after—and concluded that 630,000 people who had been in Port-au-Prince on the day of the earthquake had left the city within three weeks. They also demonstrated that they could do such calculations in close to real time. They showed—within 12 hours of receiving the data—how many people had fled an area affected by a cholera outbreak, and where they went.

Most important, their work led to a model that could guide responses to future disasters. After analyzing data on pre-earthquake travel habits, the Swedish group found that Haitians generally fled the city for the same places where they'd spent Christmas and New Year's Day. Such findings make it possible to predict where people will go when disaster hits.

SCALING UP

Until recently, these studies were done by researchers who made some special arrangement with carriers to get the data (Eagle obtained it through his academic connections). But last year Orange, the France-based global telecom giant, released to the world's research community—subject to certain conditions and restrictions—data based on 2.5 billion anonymized records from five months' worth of calls made by five million people in Ivory Coast. The first phase of this grand experiment involves seeing just what it's possible to do with the data.

Nearly a hundred research groups worldwide leaped at the opportunity to analyze the records. The resulting papers were scheduled to be presented in May at a conference at MIT under the name Data for Development, part of a larger conference of data-mining projects in both the poor and rich worlds. "It's the first time a large-scale mobile-phone data set has been released at that scale," says Blondel, who is chairing the conference. The papers had not been formally released at the time of this writing. But one charts social and travel interactions across a traditional north-south ethnic divide, providing insights into how conflict might be averted; another proposes tools for mapping the spread of malaria and detecting disease outbreaks. One corporate lab built a transportation model using cell-phone data to track ridership on 539 buses, 5,000 mini-buses, and 11,000 shared taxis.

Even if the Ivory Coast experiment succeeds, replicating it in other countries may not be easy. Last year the World Economic Forum—the group of leading industry, academic, and political figures who converge annually at Davos, Switzerland—issued a call for governments, development organizations, and companies to develop data analysis tools to improve the lives of people in the poor world. "I shouldn't have to go to operators and say 'I'll do free consulting for you—and in exchange I want to use your data to improve lives,' " Eagle says. "The operators should want to be affiliated with this. Right now many of them don't see the upside, but if we can get world leaders knocking on their doors saying 'Let's do this!' maybe we can get a lot more done."

This will take some careful work to protect privacy and prevent the data from being used in the service of oppression. Orange says it took pains to anonymize its data, but the field needs clear and widely agreed-upon ways to bring the information to market. "There are risks and benefits of having a data-driven society," Pentland says. "There is a question of who owns the data and who controls it. You can imagine what Muammar Qaddafi would have done with this sort of data. Orange is taking the steps to figure out how to create a data commons that induces greater transparency, accountability, and efficiency—to tell where there are unusual events, extreme events, to tell us where the infrastructure is breaking down. There are all sorts of things we can do with it—but it has to be available."

As these larger questions play out, Buckee and Eagle are working on refining and augmenting the data-mining tools in Kenya. Eagle aims to use surveys to sharpen and confirm the picture created by mining cell-phone data on a large scale. Call records alone are often not enough, he says; surveying even a few people could allow researchers to weed out erroneous assumptions about what those records show. Once, while analyzing phone data in Rwanda, Eagle noted that people had not moved around very much after a flood. At first, he theorized that many of them were bedridden with cholera. But it turned out that the flood had washed out the roads.

Buckee hopes to mine phone data to target drug-resistant strains of the malaria parasite. These strains, emerging in Cambodia and elsewhere, could reverse progress against the disease if allowed to proliferate, she warns. So she wants to begin merging data on the parasites' spread into mobility models to help produce targeted disease-fighting strategies.

"This is the future of epidemiology," she says. "If we are to eradicate malaria, this is how we will do it."

—David Talbot

Case Studies

How Wireless Carriers Are Monetizing Your Movements

Data that shows where people live, work, and play is being sold to businesses and city planners, as mobile operators seek new sources of revenue.

● Wireless operators have access to an unprecedented volume of information about users' real-world activities, but for years these massive data troves were put to little use other than for internal planning and marketing.

This data is under lock and key no more. Under pressure to seek new revenue streams, a growing number of mobile carriers are now carefully mining, packaging, and repurposing their subscriber data to create powerful statistics about how people are moving about in the real world.

More comprehensive than the data collected by any app, this is the kind of information that, experts believe, could help cities plan smarter road networks, businesses reach more potential customers, and health officials track diseases. But even if shared with the utmost of care to protect anonymity, it could also present new privacy risks for customers.

Verizon Wireless, the largest U.S. carrier with more than 98 million retail customers, shows how such a program could come together. In late 2011, the company changed its privacy policy so that it could share anonymous and aggregated subscriber data with outside parties. That made possible the launch of its Precision Market Insights division last October.

The program, still in its early days, is creating a natural extension of what already happens online, with websites tracking clicks and getting a detailed breakdown of where visitors come from and what they are interested in.

Similarly, Verizon is working to sell demographics about the people who, for example, attend an event, how they got there or the kinds of apps they use once they arrive. In a recent case study, says program spokeswoman Debra Lewis, Verizon showed that fans from Baltimore outnumbered fans from San Francisco by three to one inside the Super Bowl stadium. That information might have been expensive or difficult to obtain in other ways, such as through surveys, because not all the people in the stadium purchased their own tickets and had credit card information on file, nor had they all downloaded the Super Bowl's app.

Other telecommunications companies are exploring similar ideas. In Europe, for example, Telefonica launched a similar program last October, and the head of this new business unit gave the keynote address at new industry conference on "big data monetization in telecoms" in January.

"It doesn't look to me like it's a big part of their [telcos'] business yet, though at the same time it could be," says Vincent Blondel, an applied mathematician who is now working on a research challenge from the operator Orange to analyze two billion anonymous records of communications between five million customers in Africa.

The concerns about making such data available, Blondel says, are not that individual data points will leak out or contain compromising information but that they might be cross-referenced with other data sources to reveal unintended details about individuals or specific groups.

Already, some startups are building businesses by aggregating this kind of data in useful ways, beyond what individual companies may offer. For example, AirSage, an Atlanta, Georgia, a company founded in 2000, has spent much of the last decade negotiating what it says are exclusive rights to put its hardware inside the firewalls of two of the top three U.S. wireless carriers and collect, anonymize, encrypt, and analyze cellular tower signaling data in real time. Since AirSage solidified the second of these major partnerships about a year ago (it won't specify which specific carriers it works with), it has been processing 15 billion locations a day and can account for movement of about a third of the U.S. population in some places to within less than 100 meters, says marketing vice president Andrea Moe.

As users' mobile devices ping cellular towers in different locations, AirSage's algorithms look for patterns in that location data—mostly to help transportation planners and traffic reports, so far. For example, the software might infer that the owners of devices that spend time in a business park from nine to five are likely at work, so a highway engineer might be able to estimate how much traffic on the local freeway exit is due to commuters.

Other companies are starting to add additional layers of information beyond cellular network data. One customer of AirSage is a relatively small San Francisco startup, Streetlight Data which recently raised $3 million in financing backed partly by the venture capital arm of Deutsche Telekom.

Streetlight buys both cellular network and GPS navigation data that can be mined for useful market research. (The cellular data covers a larger number of people, but the GPS data, collected by mapping software providers, can improve accuracy.) Today, many companies already build massive demographic and behavioral databases on top of U.S. Census information about households to help retailers choose where to build new stores and plan marketing budgets. But Streetlight's software, with interactive, color-coded maps of neighborhoods and roads, offers more practical information. It can be tied to the demographics of people who work nearby, commute through on a particular highway,

or are just there for a visit, rather than just supplying information about who lives in the area.

"If you're a retailer and you are on someone's commute path, they may pass by you 600 times a year," says founder and CEO Laura Schewel, a transportation researcher who is also finishing her PhD at the University of California, Berkeley. "If they never come in, that's a big missed opportunity."

Streetlight's work shows why such data has the potential to improve city planning. One of the company's first customers is the Oakland Business Development Corporation, which is trying to attract businesses to a city with a reputation for crime and poverty, says Schewel. Streetlight's data shows the patterns by which droves of wealthier people commute through the city from outlying suburbs on their way to San Francisco, and it also shows that young people visit Oakland for the growing nightlife scene. The group's goal is to convince national chains that don't know Oakland well and might look only at household demographics that it makes sense to open up a shop there, and then help them chose ideal locations.

Of course, all the companies involved are aware that people may have privacy concerns, and they wrestle with the challenge of conveying what "anonymous" and "aggregated" data means to people who are increasingly aware that they are carrying around a tracker in their pocket. (Verizon Wireless also does allow its customers to opt out of the program.)

The companies appear to be taking many precautions to protect privacy. In Streetlight's case, the data is obtained in batches and algorithms are used to tease out patterns that help it match locations to other sources of data, such as DMV records or U.S. Census information. The company says it can't track any individual device on a map or target ads to any one specific person with the data it buys. Schewel also emphasizes that the data is at least a day, if not a month, old and its software doesn't report statistics for groups of less than 15 devices.

A more important question, probably, is how people will balance their privacy concerns with the benefits of making their location data, or other mobile data about them, more available. Research and experience suggest that in practice most people don't mind, or don't care as much as they think they do about privacy.

There will be many uses for such data. Some will involve the race to become the platform that provides the best mobile ads and deals. Verizon, for example, now has a pilot effort within its Precision Market Insights program that allows customers to opt in to (rather than opt out of) giving Verizon permission to share more information about who and where they are, including Web browsing and app usage data. In return, they might receive deals and offers on their phone from marketers who want to reach them.

The data could also save budget-strapped governments some money. In California, the state is undertaking an unusual pilot experiment by licensing GPS data from navigation providers rather than repairing its crumbling network of sensors and detectors in the state's roadways. Alexandre Bayen, a researcher at the University of California, Berkeley, who is helping put together the request for bidders, believes markets that give governments access to mobility data will eventually become common.

"If you're a retailer and you're on someone's commute path, they may pass by you 600 times a year. If they never come in, that's a big missed opportunity."—Streetlight CEO Laura Schewel

Schewel, who came to the idea for Streetlight as she was trying to give green transportation advocates better data about who uses the roads, also views her company as a long-term way to create more efficient cities by helping businesses locate closer to their customers and reducing shopping trips.

"A green transportation startup, where the core value proposition of the company is saving transportation fuel, is not a scalable idea," she says. "But a marketing analytic startup that spits out mile reduction as its side effect? I think that is the most powerful way to be an agent for change."

—Jessica Leber

Emerged Technologies

How to Mine Cell-Phone Data Without Invading Your Privacy

Researchers use phone records to build a mobility model of the Los Angeles and New York City regions with new privacy guarantees.

● Researchers at AT&T, Rutgers University, Princeton, and Loyola University have devised a way to mine cell-phone data without revealing your identity, potentially showing a route to avoiding privacy pitfalls that have so far confined global cell-phone data-mining work to research labs.

Working with billions of location data points from AT&T mobile phone calls and text-messages around Los Angeles and New York City, they've built a "mobility model" of the two regions that aggregates the data, produces representative "synthetic call records"—then mathematically obscures any data that could tend to identify people.

The model can do things like rapidly predict how a new development or telecommuting policy would affect overall transportation, or it could be a new tool for planning at a town level where little mobility data is available, says Margaret Martonosi, a computer scientist at Princeton who is working on the model.

Right now, planners generally rely on road sensors and the limited number of people who permit their GPS position to be captured.

Vincent Blondel, a computer scientist at the Université Catholique de Louvain, Belgium, and a leader in research efforts surrounding call data records and privacy issues, says the work was impressive. "This is an excellent work that will help explore ways of making the best use of important data in a privacy protective way," he says.

Even the simplest phone leaves behind extensive digital traces—called call detail records, or CDRs—that are preserved by mobile carriers. These records—on the time a voice call or text message was placed, and the identity and location of the cell tower involved—give the approximate locations of the phone's owner. Over time, they can be used to develop an accurate trace of the user's movements.

In aggregate—but mostly in theory so far—this data can be used to guide epidemiology research, or to unsnarl traffic by giving an unprecedented view on all human movement patterns (see "How Wireless Carriers Are Monetizing Your Movements" on p. 22). It can also guide development efforts in poorer parts of the world (see "Big Data from Cheap Phones" on p. 17).

But building in guaranteed privacy protections represents the toughest hurdle to the growing number of research efforts that tap CDRs. Even if such records are stripped of names and numbers, the identity of the person can often be revealed through other means. For example, a single cell-tower ping at 4:12 a.m. could be connected to a public tweet made at 4:12 a.m. that includes the location and identity of the tweeter. Similar risks crop up for data belonging to people who live in a remote area or have unusual home-work commuting patterns.

The new approach starts by aggregating traces of real human movements, then identifying common locations that might indicate home, work, or school. Next, it creates a set of transportation models. These models generate route tracks of people that the researchers call "synthetic," because they are merely representative of the aggregate data, and not of actual people.

But the third part is the key. Even these supposedly synthetic records can closely match real ones (especially when the underlying aggregate sample is small). So an algorithm, using an emerging technique known as differential privacy, calculates exactly how high this risk is, and how to reduce it by altering the data. "Noise is injected into the model at points in order to reduce the likelihood of individuals being identifiable," says Martonosi.

Injecting noise includes deliberately altering the aggregated home and work locations to reduce the reliance on any one individual's data. Likewise, the aggregated call times are changed to mask any individual's contribution. Taken together, such tweaks to the data would throw off any efforts to align databases.

Part of this new mobility modeling work was first presented at a conference last year, but refinements and the differential privacy variant were presented last week at a conference at MIT. At the same conference, IBM researchers showed how call records could help optimize public transportation routes (see "African Bus Routes Redrawn Using Cell Phone Data" on p. 16).

Martonosi says that publicly releasing the mobility models she and her colleagues have built of New York and Los Angeles metro areas won't happen before additional publications finalize the work and prove the privacy approach, since the models indirectly draw from real user data.

In the meantime, the methods she and her colleagues used to build the model are publicly published. So other groups could build similar models for other metro areas if they have their own call data records to work with, she says. AT&T collaborated on the research, which was done at an AT&T facility on three months' worth of customer data from 300,000 of the carrier's customers each in the New York and Los Angeles areas. AT&T declined to comment.

Amid surging research interest in mobile data, the groups' approach is garnering considerable interest. William Hoffman, who heads the World Economic Forum's data-based development efforts, says the approach showed promise.

"I thought the concept was quite interesting as a means of 'de-risking' the ability of researchers to explore the data," he says. "It's one of multiple steps data holders can pursue to strike the balance of using data while protecting the individual."

One key question is whether a system of synthetic data records could get the carriers around the delicate matter of obtaining user consent. "That's one of the big issues I took away from the [recent MIT] conference," says Hoffman. The answer might depend how the data was used or sold, he says.

Nicolas Decordes, a vice president at Orange, the European carrier, says that the company's R&D team said the techniques "would be feasible and could be helpful" for transportation modeling. Because the method does not use real-time data, however, it is better for planning and cannot guide response to events.

300,000

Number of AT&T subscriber records in a recent privacy study.

The process of obtaining and using cell-phone data is already very touchy. When Orange released data from Ivory Coast to researchers last year, a process Decordes oversaw, that nation was chosen because its Information and Communications Technology (ICT) ministry hadn't signed on to a regulatory framework restricting such use, in contrast to nearby African nations. And even so, Orange required researchers to sign agreements barring them from trying to identify individuals.

Linus Bengtsson, an epidemiologist at Sweden's Karolinska Institute and a founder of Flowminder, which provides mobility data to NGOs and relief agencies, says that however advanced the privacy protections get, the research community will always need codes of conduct to protect privacy.

"Researchers in many areas analyze datasets where someone—with enough determination—could be able to identify people," he says. "I think [developing] rules for this is actually a more important point than the difficult task of creating special anonymized data sets."

Other recent research results included ones that show how call records can be used to follow soccer fans as they leave a match or even map poverty levels inside a country, if airtime purchasing habits are analyzed. *—David Talbot*

Emerged Technologies

Why Big Companies Are Investing in a Service That Listens to Phone Calls

A startup that converts conversations to text so it can offer instant information gets financing from Telefónica, Samsung, and Intel.

● Would you give your wireless carrier permission to listen in on your phone calls? Telefónica, one of the world's largest mobile carriers, is testing a technology that can understand conversations and quickly pull up relevant information. If that info turns out to be useful, customers may want to invite it to listen in.

The technology is being built by a small San Francisco startup, Expect Labs, which is announcing strategic investments from the venture capital arms of three weighty backers today: Samsung, Intel, and Telefónica Digital, the business unit the telecom company launched in 2011 to unearth new revenue opportunities. The size of the investments was not disclosed.

"Google would love to have the amount of information that these telecommunications companies have every day."

—Expect Labs CEO Timothy Tuttle

Expect Labs has attracted attention because its technology is in line with the general direction that search technology has been taking with the advent of wearable computers such as watches and glasses, and Internet-connected cars and TVs. Rather than wait for users to search for something, the new technology offers up info that it thinks the user might need. The Google Now software for mobile devices, for example, already monitors its users' locations, search history, and e-mail to call up traffic reports and other information. But Google Now doesn't mine old-fashioned voice conversations yet. In October, Google Ventures invested in Expect Labs as part of a $2.4 million financing round by the startup.

Expect Labs has spent more than two years developing artificial intelligence technology that can parse the meaning of real-time conversations (Apple's Siri, in contrast, can interpret only relatively simple spoken commands). Expect Labs's "anticipatory computing engine" extracts the most relevant terms and uses them to offer potentially helpful information." For example, if two friends are having a discussion about grabbing some Thai food, it might call up reviews of nearby restaurants. If a company's revenue comes up in a videoconference, it could display recent revenue charts. Expect Labs has built an app, called MindMeld, to demonstrate to partners how this works.

"The great thing about the platform," says Telefónica Digital vice president of investments Christian Craggs, "is that voice is translated into text, and that is translated into an element of data."

Telefónica, Craggs says, intends to test the technology internally and with some of its customers over the next six months. It could introduce it into products as soon as next year, he says. The technology could also appear inside TuGo, the company's voice app for subscribers, or as a tool for customer service agents or doctors. Telefónica—which operates mainly in Europe and South America—will also explore how it could use the technology to create targeted advertising, he says. In all cases, he emphasizes, customers would have to explicitly opt in. Their conversations would be analyzed in real time, but not stored on company servers.

Expect Labs's mix of strategic investors also illustrates how previously distinct business models are shifting in the wireless world.

For example, Expect Labs CEO Timothy Tuttle says that electronics maker Samsung is interested in his company because, in the smartphone and smart TV market, owning and controlling software and services is becoming more important—as Apple has shown. Similarly, with traditional voice service becoming commoditized by the Internet, Telefónica Digital is looking to mine new revenue from the company's existing assets. This includes using its wealth of data and access to customers to launch a market research service (see "How Wireless Carriers are Monetizing Your Movements" on p. 22).

"Google would love to have the amount of voice information that these telecommunications companies have every day," says Tuttle.

All that data will also help Expect Labs improve its technology, which is still a work in progress. By tapping into more voice conversations in its work with Telefónica, its algorithms will have more testing data to improve their predictive abilities.

Still, Tuttle says, the technology will take a few years to become mainstream. "We're a small company that's trying to solve a very hard problem," he says.

—Jessica Leber

Special Focus:

People Analytics: Big Data in Hiring

Finding the right person to hire from a pile of resumes is a huge challenge for any company that's growing. Now, several technology startups, as well as human resources departments at some large companies, believe they've found ways to automate hiring using software that grades prospective employees on their work or that compares them with statistical profiles of ideal workers. Next comes the debate: is data-driven hiring unfair or actually far more meritocratic?

Emerged Technologies

A Startup That Scores Job Seekers, Whether They Know It or Not

To help recruiters, a startup called Gild has created a database of four million software developers and rated their work. Could other fields be next?

● Winning over recruiters and potential bosses can be hard enough. Now there's something else job seekers have to woo: an algorithm.

A San Francisco startup called Gild has created a program that evaluates and scores software developers on the work they have publicly released. Recruiters for technology companies can use this "Gild score" to see through the top-tier degrees, vague descriptions of skill sets, or polished testimonials of well-connected programmers whose coding skills may be below par. Less-obvious candidates, such as a junior in college who has been building great apps since she was 16, might rise into view instead.

For now, Gild is evaluating only software developers, whose work can often be freely found in repositories for open-source software, coder Q&A forums, and other online developer hangouts. But CEO Sheeroy Desai says that Gild hopes to bring its "talent acquisition technology" beyond the realm of software programmers, especially as more work products start to appear online.

He says it's too early to detail what those possibilities could be. But one could imagine some future algorithm evaluating a teacher's online courses, a journalist's articles, or a scientist's open-access data. (A company called Klout already scores how influential people are in social media.) "This is massively useful beyond just tech recruiting," says Bryan Power, director of talent at Square, a payment technology company that has used Gild's software to help vet job candidates for the last three months. "There's so much more that will be online in the next couple of years," he says.

Since launching a beta version of its software last March, Gild has profiled four million software developers and has 70 customers, from high-profile Silicon Valley startups such as Palantir Technologies and Box to large IT providers such as Salesforce and EMC.

"This is massively useful beyond just tech recruiting. There's so much more that will be online in the next couple of years."—Bryan Power, director of talent at Square

Its technology stitches together profiles of individual coders from their activities in open-source forums and public websites. It can "scrape" information from popular developer hangouts even if those sites don't have formal APIs, or application programming interfaces, to facilitate the transfer of data. Gild also uses image recognition to match up profile pictures on different sites. It then assesses two scores—one for work quality, one for influence.

One of Gild's biggest sources of data is Github, a software developer collaboration site that hosts the most open-source code in the world. Github profiles are already replacing programmers' résumés in many cases.

Desai, the former chief operating officer of the IT company Sapient, cofounded Gild with Luca Bonmassar, who had been leading a software team at Vodafone, because they were tired of the time they had wasted in hiring software developers. "You'd bring them in, throw code at them, and realize they didn't know what they were doing," says Desai.

4 million

Number of software developers whose work has been ranked by an algorithm.

As with any grading scheme, Gild's algorithm—which scores from 1 to 100—is making judgments about what makes a good developer. The software grades the quality of someone's code by checking for basic errors and also gauging its complexity. It also looks at how extensively programmers' open-source code has been taken up by other projects. It tends to reward developers who know a small number of programming languages really well and dabble in several others, Desai says.

This approach has several limitations. Power, at Square, points out that not every company would make the same judgments about candidates that Gild's algorithm does. It can be tricky to untangle the contributions that individual programmers have made to a group project on Github or a similar social forum. And not every developer has worked on a large number of open-source projects.

But Desai says that by accumulating so much data, learning patterns, and making predictions, Gild is lowering the threshold of information it needs to score a given candidate.

—Jessica Leber

Case Studies

The Machine-Readable Workforce

Companies are analyzing more data to guide how they hire, recruit, and promote their employees.

● Xerox is screening tens of thousands of applicants for low-wage jobs in its call centers using software from a startup company called Evolv that automatically compares job seekers against a computer profile of the ideal candidate.

According to these data, culled from studying job records of many similar workers, past experience working in call centers isn't a good predictor of success. Instead, a person should be a "creative" type, though not too inquisitive. Participating in one social network like Facebook is a plus, but involvement in too many is a negative. A short commute is a must—that means a person is less likely to quit before Xerox can recoup its cost to train them.

While personality exams aren't new to business, large employers like Xerox are beginning to embrace a concept called "workforce science" that promises to make performance reviews and judging résumés far more data-driven. One of the best-known attempts to hire and fire by the numbers is at Google, whose HR department, called "People Operations," has turned hiring into a kind of engineering project, using computer models to determine how many times each candidate should be interviewed, how larges raises should be, and nearly every other personnel decision.

"More and more companies are doing preemployment testing. Whether this really indicates some job-related quality in the applicant is questionable." —Christopher Moody, lawyer

Evolv, based in San Francisco and founded in 2007, bases its advice on data gleaned from tens of thousands of employee files on hourly workers, who also make up 60 percent of the United States workforce. Applicants have to take a half-hour online test that ranks them against a profile of a successful center worker. Another startup, Gild, has begun using software to score computer programmers who place their work in public repositories, locating job candidates whose résumés might otherwise end up in a trash bin (see "A Startup That Scores Job Seekers, Whether They Know It or Not" on p. 27).

Lawyers who practice anti-discrimination law are watching these trends. While it's legal to give aptitude tests, hiring based on a computer's assessment of seemingly unconnected factors—like how many social networks you join—could raise new questions. "They're creating these big databases of people," says Christopher Moody, an employment lawyer in Los Angeles. "More and more companies are doing preemployment testing. Whether this really indicates some job-related quality in the applicant is questionable."

It's easy to see why Xerox wants to turn to automated methods. Although it still sells photocopiers, Xerox has also become one of the world's largest outsourcing companies. It provides services like running customer service centers, and processing health claims and credit-card applications that brought in $11.5 billion in revenue last year.

That business relies on a huge workforce of 54,000 customer service agents, and because of high attrition in hourly jobs (pay in the U.S. ranges from $9 to $20 an hour), Xerox will have to replace 20,000 of them this year, says Teri Morse, vice president for recruiting at Xerox Services. Morse says employees that stay less than six months cause a loss for Xerox, due to the expense of training them.

Since the company began pilot tests of Evolv's analytics software two years ago, Morse says employees are on average staying longer at Xerox and their performance is 3 to 4 percentage points better, as measured by factors like how many complaints they resolve or how long it takes to handle a call. The software has also started to influence other subtle factors, like what time of year Xerox hires people.

Morse says basing decisions on data means Xerox has been able to broaden the base of people it will consider for hourly jobs, including those who have been unemployed for long periods. But the data also rules people out. Morse says Xerox today won't even look at résumés of those who score in the "red" category of Evolv's initial behavioral assessment, a 30-minute online exam that workers fill out at home. "Individuals that test strongly perform better and survive longer," she says. Early on, while piloting the system, Morse says Xerox still hired against the advice of the data. Now, she says, "people who do poorly we no longer hire."

20,000

Customer service agents Xerox will replace this year because of attrition.

—Jessica Leber

Case Studies

In a Data Deluge, Companies Seek to Fill a New Role

A job born in Silicon Valley is going mainstream as new industries are deluged in useful data.

● The job description "data scientist" didn't exist five years ago. No one advertised for an expert in data science, and you couldn't go to school to specialize in the field. Today, companies are fighting to recruit these specialists, courses on how to become one are popping up at many universities, and the *Harvard Business Review* even proclaimed that data scientist is the "sexiest" job of the 21st century.

Data scientists take huge amounts of data and attempt to pull useful information out. The job combines statistics and programming to identify sometimes subtle factors that can have a big impact on a company's bottom line, from whether a person will click on a certain type of ad to whether a new chemical will be toxic in the human body.

While Wall Street, Madison Avenue, and Detroit have always employed data jockeys to make sense of business statistics, the rise of this specialty reflects the massive expansion in the scope and variety of data now available in some industries, like those that collect data about customers on the Web. There's more data than individual managers can wrap their minds around—too much of it, changing too fast, to be analyzed with traditional approaches.

As smartphones promise to become a new source of valuable data to retailers, for example, Walmart is competing to bring more data scientists on board and now advertises for dozens of open positions, including "Big Fast Data Engineer." Sensors in factories and on industrial equipment are also delivering mountains of new data, leading General Electric to hire data scientists to analyze these feeds.

The term "data science" was coined in Silicon Valley in 2008 by two data analysts then working at LinkedIn and Facebook. Now many startups are basing their businesses on their ability to analyze large quantities of data—often from disparate sources. ZestFinance, for example, has a predictive model that uses hundreds of variables to determine whether a lender should offer high-risk credit. The underwriting risk it achieves is 40 percent lower than that borne by traditional lenders, says ZestFinance data scientist John Candido. "All data is credit data to us," he says.

Data scientist has become a popular job title partly because it has helped pull together a growing number of haphazardly defined and overlapping job roles, says Jake Klamka, who runs a six-week fellowship to place PhDs from fields like math, astrophysics, and even neuroscience in such jobs. "We have anyone who works with a lot of data in their research," Klamka says. "They need to know how to program, but they also have to have strong communications skills and curiosity."

The best data scientists are defined as much by their creativity as by their code-writing prowess. The company Kaggle organizes contests where data scientists compete to find the best way to make sense of massive data sets. Many of the top Kagglers (there are 88,000 registered on the site) come from fields like astrophysics or electrical engineering, says CEO Anthony Goldbloom. The top-ranked participant is an actuary in Singapore.

Universities are starting to respond to the job market's needs. Stanford University plans to launch a data science master's track in its statistics department, says department chair Guenther Walther. A dozen or so other programs have already been started at schools including Columbia University and the University of California, San Francisco. Cloudera, a company that sells software to process and organize large volumes of data, announced in April that it would work with seven universities to offer undergraduates professional training on how to work with "big data" technologies.

Cloudera's education program director, Mark Morissey, says a skills shortage is looming and that "the market is not going to grow at the rate it currently wants to." That has driven salaries up. In Silicon Valley, salaries for entry-level data scientists are around $110,000 to $120,000.

Others think the trend could create a new area of outsourcing. Shashi Godbole, a data scientist in Mumbai, India, who is ranked 20th on Kaggle's scoreboard, recently completed a Kaggle-arranged hourly consulting gig, a new business the platform is getting into. He did work for a tiny health advocacy nonprofit located in Chicago and is now bidding on more jobs (he earns $200 per hour, and Kaggle collects $300 an hour). His Kaggle work is part time for now, but he says it's possible that it could be his major source of income one day.

"I'm a data janitor. That's the sexiest job of the 21st century. It's very flattering, but it's also a little baffling."

—Josh Wills, senior director of data science at Cloudera

To the data scientists themselves, the job is certainly less sexy than it's being made out to be. Josh Wills, a senior director of data science at Cloudera, says most of the time it involves cleaning up messy data—for example, by putting it in the right columns and sorting it.

"I'm a data janitor. That's the sexiest job of the 21st century," he says. "It's very flattering, but it's also a little baffling."

—Jessica Leber

MIT Technology Review published six business reports per year. For reports on Innovation Funding, Digital Education, The Future of Work, and more, please visit www.technologyreview.com/businessreports