COLUMBIA BUSINESS SCHOOL

CS@CU

Data Science and Technology Entrepreneurship

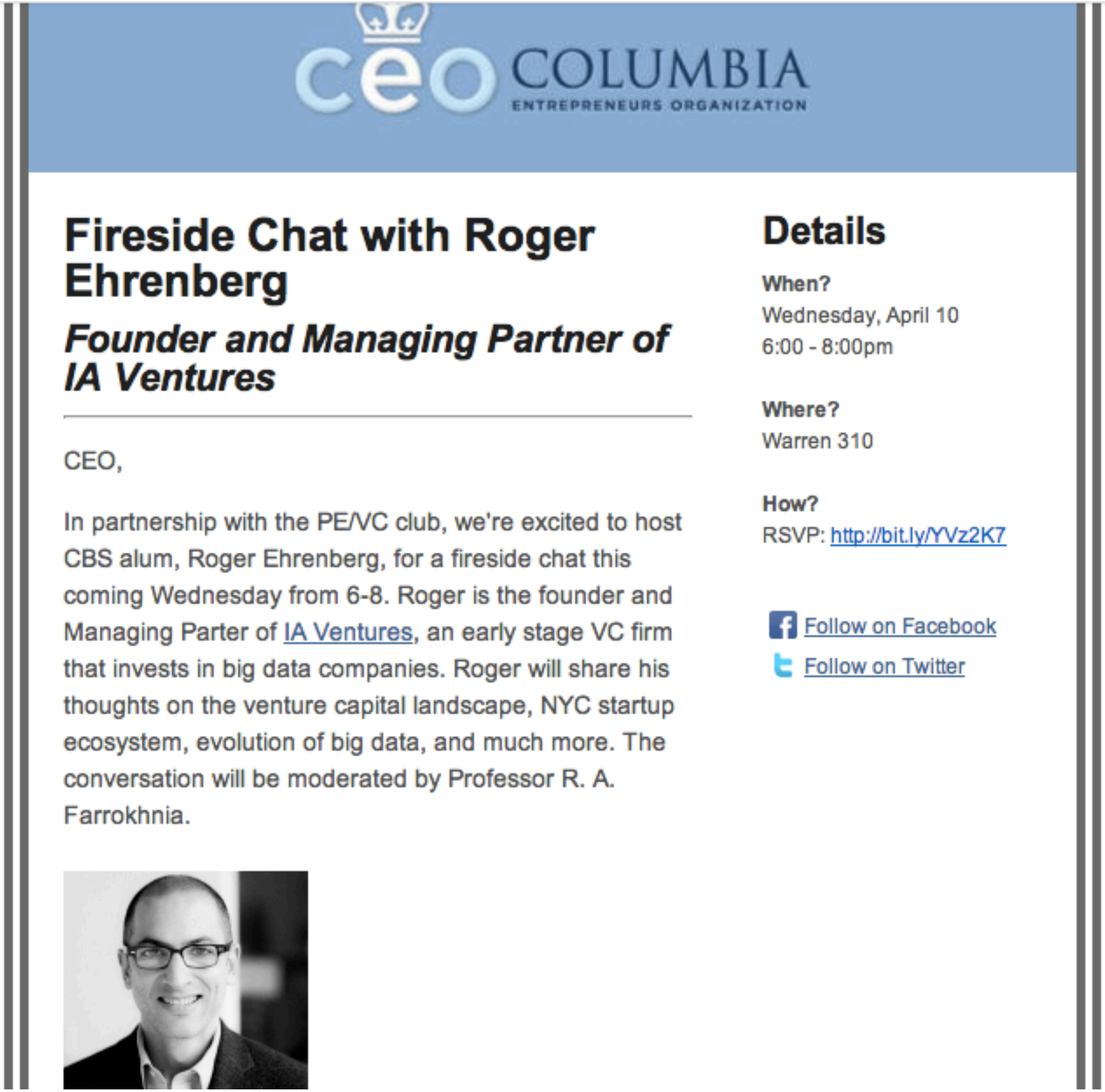
Paper Discussions
Startup Technology Stack

Sameer Maskey
Week 11

Announcements

- Friday Whole Day Open Office Hours
  - Idea is for you to come and work together as a group
  - April 26 - Whole day
  - May 3 - Whole day
- Please come to my office hours to get the feedback for the pitch

Announcement



ceo COLUMBIA
ENTREPRENEURS ORGANIZATION

Fireside Chat with Roger Ehrenberg

Founder and Managing Partner of IA Ventures

CEO,

In partnership with the PE/VC club, we're excited to host CBS alum, Roger Ehrenberg, for a fireside chat this coming Wednesday from 6-8. Roger is the founder and Managing Parter of IA Ventures, an early stage VC firm that invests in big data companies. Roger will share his thoughts on the venture capital landscape, NYC startup ecosystem, evolution of big data, and much more. The conversation will be moderated by Professor R. A. Farrokhnia.

Details

When?
Wednesday, April 10
6:00 - 8:00pm

Where?
Warren 310

How?
RSVP: http://bit.ly/YVz2K7

Follow on Facebook
Follow on Twitter

Today’s Topic

- Paper Discussion
- Data to Clusters
- Guest Lecture
  - Startup Technology Stack
  - Luis Sanz, Co-Founder, Olapic

Papers

- The Pathologies of Big Data - Adam Jacobs
- Big Data: The Management Revolution - Andrew McAfee and Erik Brynjolfsson
- 10 ways big data changes everything - The future of Foursquare is data-fueled recommendations - Editorial Staff Gigaom

Big Data : The Management Revolution

- Authors conducted interviews with 330 public North American Companies
- Gathered performance data
- Data driven decision making company
  - 6% more profitable
  - 5% more productive

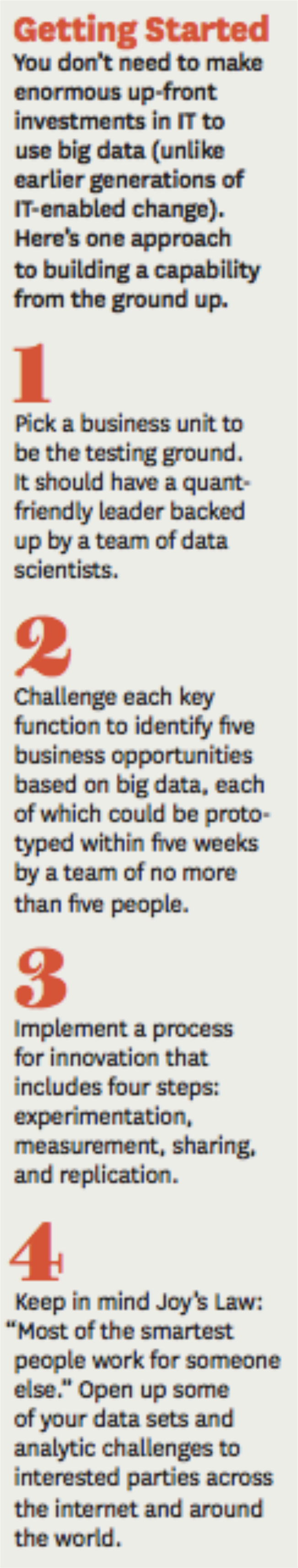
Source [2]

Big Data : The Management Revolution

- Airline Industry - ETA
  - Problem - 10% of flight arrivals had 10 min gap
  - ETA given by pilots
  - Gathered data - weather, flight schedule, feeds from passive radar stations, wide range of information on each plane
  - Eliminated the gap
- Sears : Personalized Promotion
  - 8 weeks to build promotions
  - Hadoop cluster
  - Time reduced to 1 week

Source [2]

Getting Started For a Company - Data Science



Getting Started

You don't need to make enormous up-front investments in IT to use big data (unlike earlier generations of IT-enabled change). Here's one approach to building a capability from the ground up.

1

Pick a business unit to be the testing ground. It should have a quant-friendly leader backed up by a team of data scientists.

2

Challenge each key function to identify five business opportunities based on big data, each of which could be prototyped within five weeks by a team of no more than five people.

3

Implement a process for innovation that includes four steps: experimentation, measurement, sharing, and replication.

4

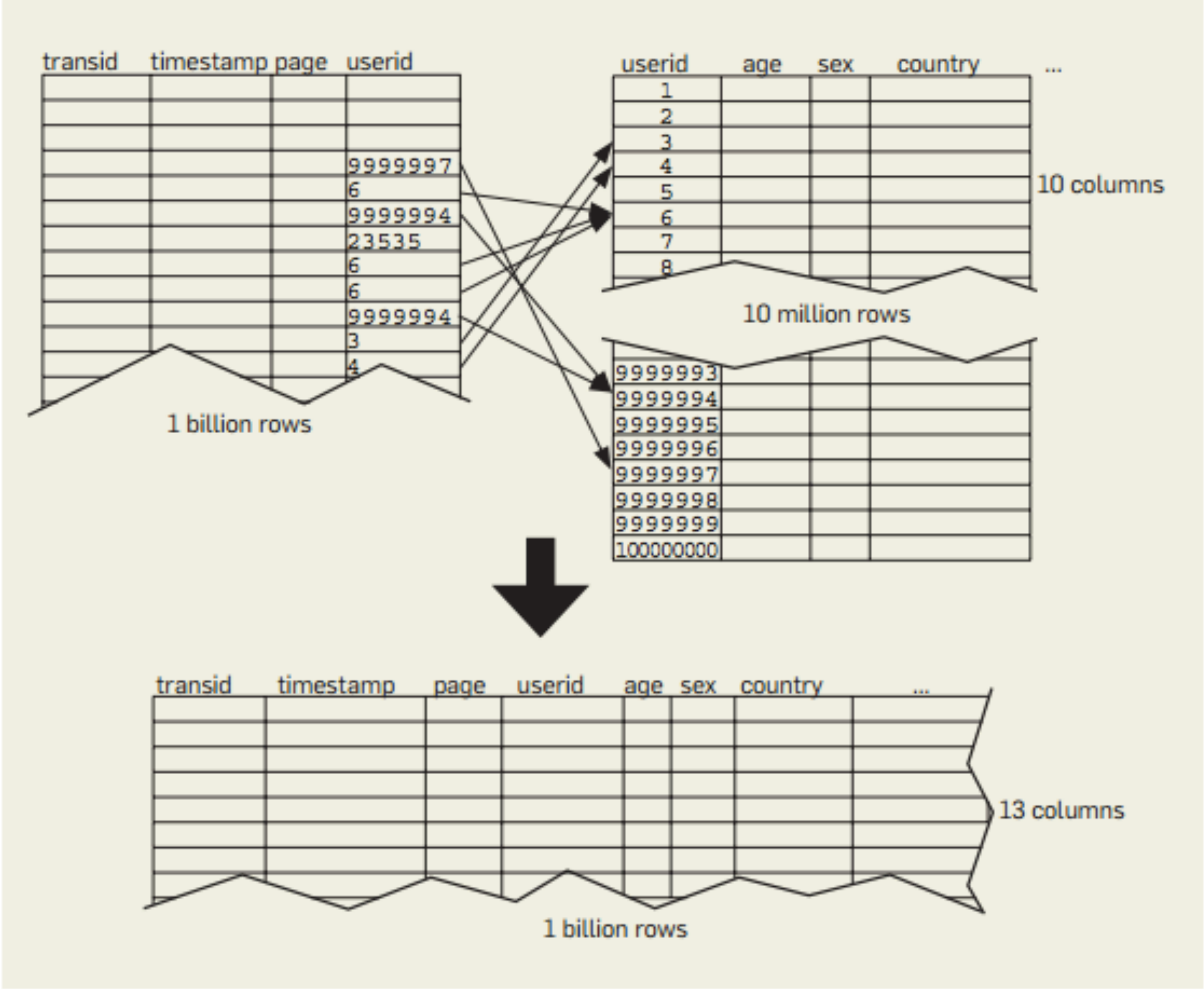
Keep in mind Joy's Law: "Most of the smartest people work for someone else." Open up some of your data sets and analytic challenges to interested parties across the internet and around the world.

- Pick data friendly unit
- Identify 5 opportunities
- experiment, measure, share, replication
- share

Source [2]

Denormalizing User Information Table

Figure 4. Denormalizing a user information table.



Source [1]

Example

- 10 years of observation
- 15 second intervals
- 1000 sensor sites
  - 20 million observations per site

Source [1]

Store by site?

Source [1]

Store by site?

- Store data by site?
  - 1 node gets 100 sites
  - 2 billion data points

Source [1]

Store by site?

- Store data by site?
  - 1 node gets 100 sites
  - 2 billion data points
- Interested in only few sites, most other nodes will be idle

Source [1]

Store by time

Source [1]

Store by time

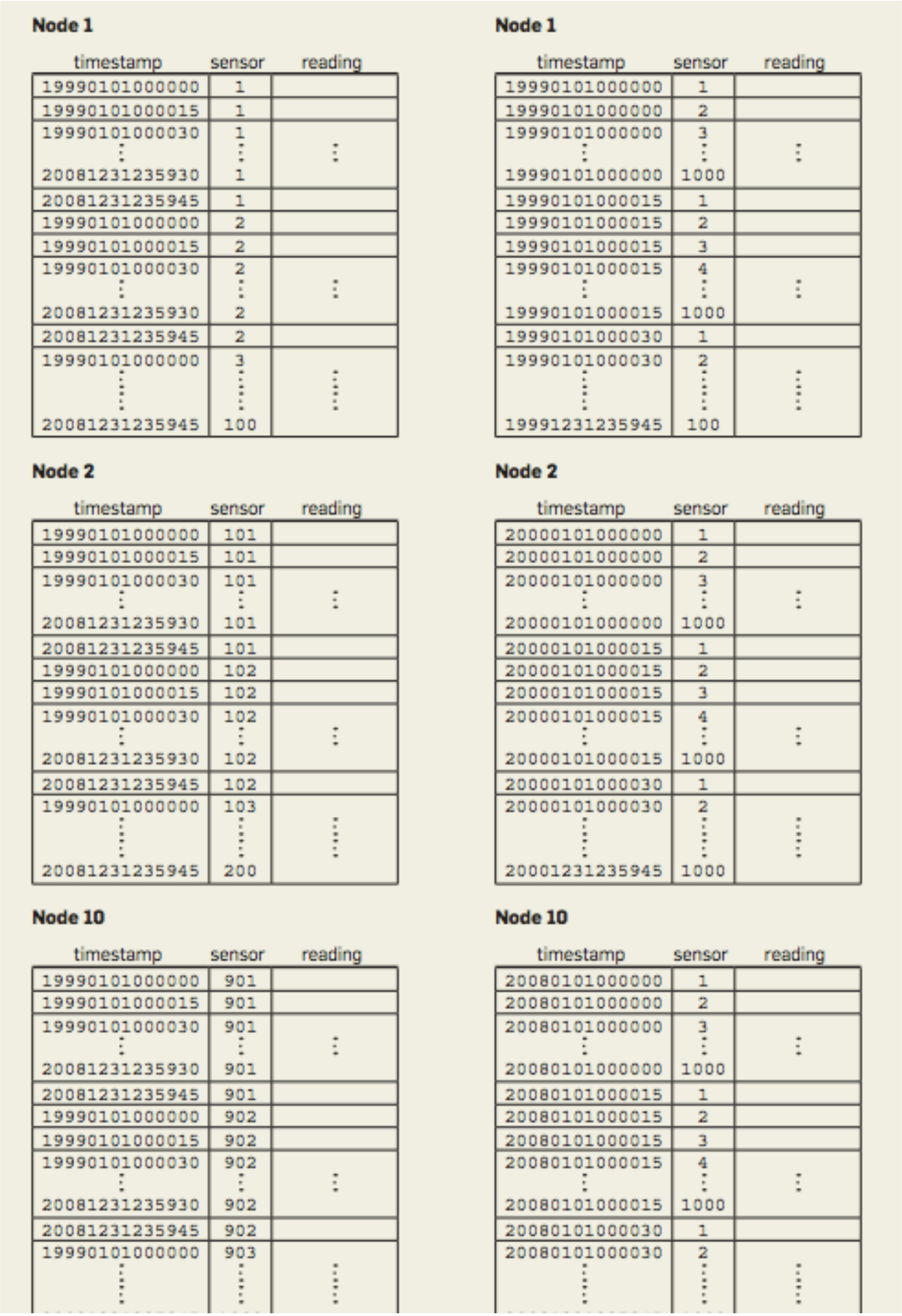
- Store data by time
    - store 1 year per node for each site

Source [1]

Store by time

- Store data by time
  - store 1 year per node for each site
- Time series calculation means communication across nodes

Source [1]



Node 1

timestamp	sensor	reading
19990101000000	1	
19990101000015	1	
19990101000030 ⋮ 20081231235930	1 ⋮ 1	⋮
20081231235945	1	
19990101000000	2	
19990101000015	2	
19990101000030 ⋮ 20081231235930	2 ⋮ 2	⋮
20081231235945	2	
19990101000000 ⋮ 20081231235945	3 ⋮ 100	⋮

Node 2

timestamp	sensor	reading
19990101000000	101	
19990101000015	101	
19990101000030 ⋮ 20081231235930	101 ⋮ 101	⋮
20081231235945	101	
19990101000000	102	
19990101000015	102	
19990101000030 ⋮ 20081231235930	102 ⋮ 102	⋮
20081231235945	102	
19990101000000 ⋮ 20081231235945	103 ⋮ 200	⋮

Node 10

timestamp	sensor	reading
19990101000000	901	
19990101000015	901	
19990101000030 ⋮ 20081231235930	901 ⋮ 901	⋮
20081231235945	901	
19990101000000	902	
19990101000015	902	
19990101000030 ⋮ 20081231235930	902 ⋮ 902	⋮
20081231235945	902	
19990101000000 ⋮	903 ⋮	⋮

Node 1

timestamp	sensor	reading
19990101000000	1	
19990101000000	2	
19990101000000 ⋮ 19990101000000	3 ⋮ 1000	⋮
19990101000015	1	
19990101000015	2	
19990101000015	3	
19990101000015 ⋮ 19990101000015	4 ⋮ 1000	⋮
19990101000030	1	
19990101000030 ⋮ 19991231235945	2 ⋮ 100	⋮

Node 2

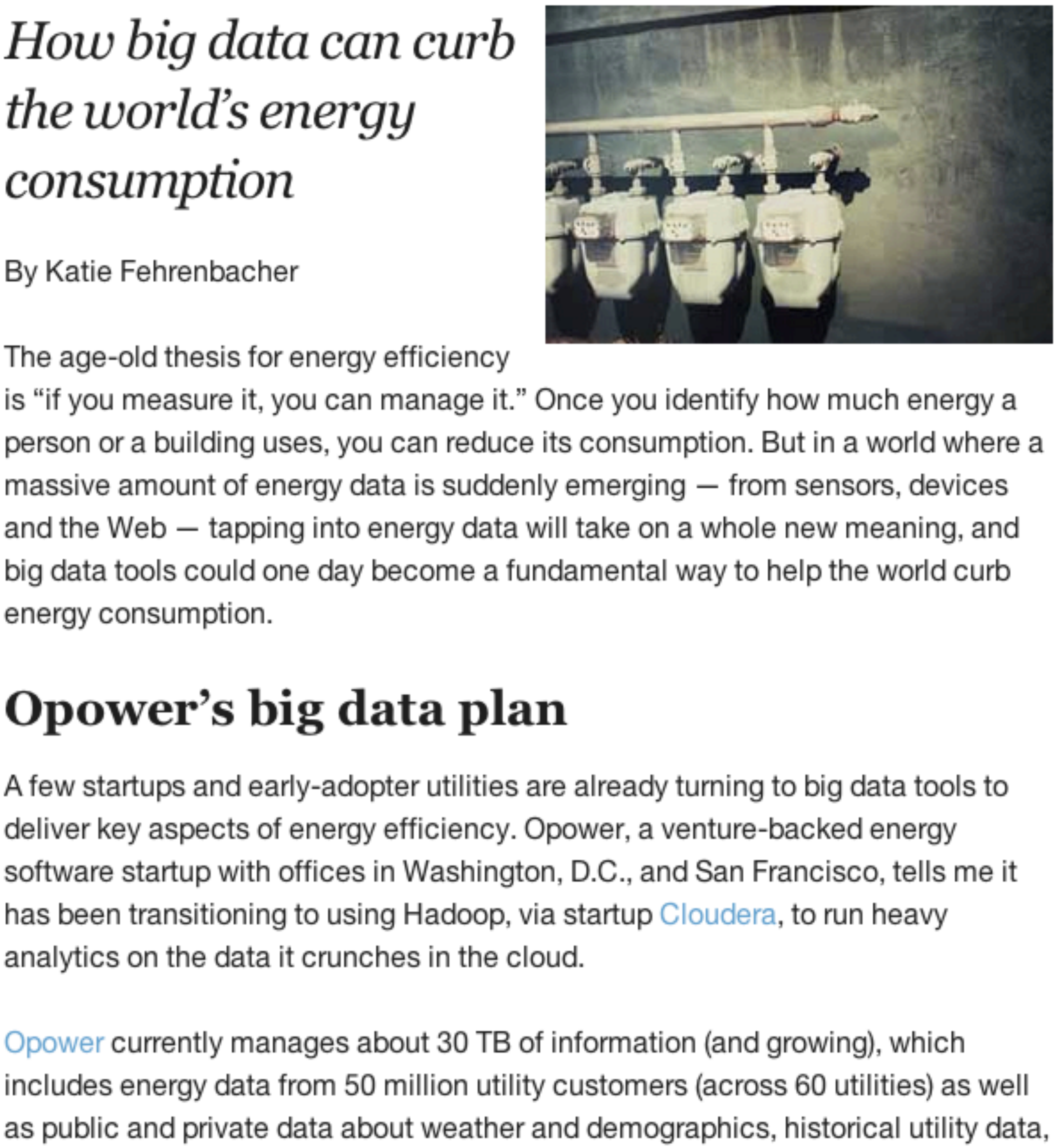
timestamp	sensor	reading
20000101000000	1	
20000101000000	2	
20000101000000 ⋮ 20000101000000	3 ⋮ 1000	⋮
20000101000015	1	
20000101000015	2	
20000101000015	3	
20000101000015 ⋮ 20000101000015	4 ⋮ 1000	⋮
20000101000030	1	
20000101000030 ⋮ 20001231235945	2 ⋮ 1000	⋮

Node 10

timestamp	sensor	reading
20080101000000	1	
20080101000000	2	
20080101000000 ⋮ 20080101000000	3 ⋮ 1000	⋮
20080101000015	1	
20080101000015	2	
20080101000015	3	
20080101000015 ⋮ 20080101000015	4 ⋮ 1000	⋮
20080101000030	1	
20080101000030 ⋮	2 ⋮	⋮

2 ways to distribute data

Source [1]



How big data can curb the world's energy consumption

By Katie Fehrenbacher

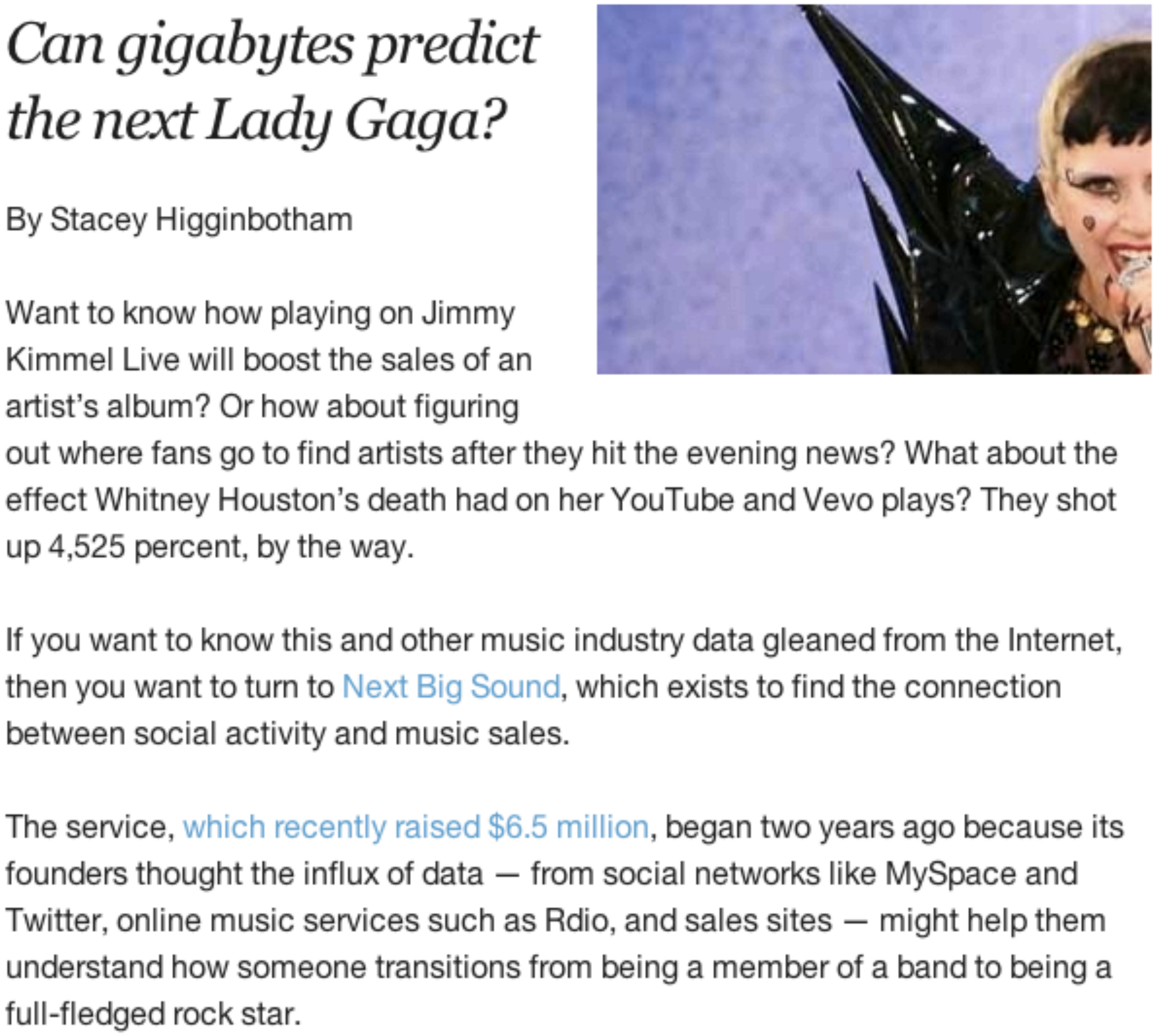
The age-old thesis for energy efficiency is "if you measure it, you can manage it." Once you identify how much energy a person or a building uses, you can reduce its consumption. But in a world where a massive amount of energy data is suddenly emerging — from sensors, devices and the Web — tapping into energy data will take on a whole new meaning, and big data tools could one day become a fundamental way to help the world curb energy consumption.

Opower's big data plan

A few startups and early-adopter utilities are already turning to big data tools to deliver key aspects of energy efficiency. Opower, a venture-backed energy software startup with offices in Washington, D.C., and San Francisco, tells me it has been transitioning to using Hadoop, via startup Cloudera, to run heavy analytics on the data it crunches in the cloud.

Opower currently manages about 30 TB of information (and growing), which includes energy data from 50 million utility customers (across 60 utilities) as well as public and private data about weather and demographics, historical utility data,

Source [3]



Can gigabytes predict the next Lady Gaga?

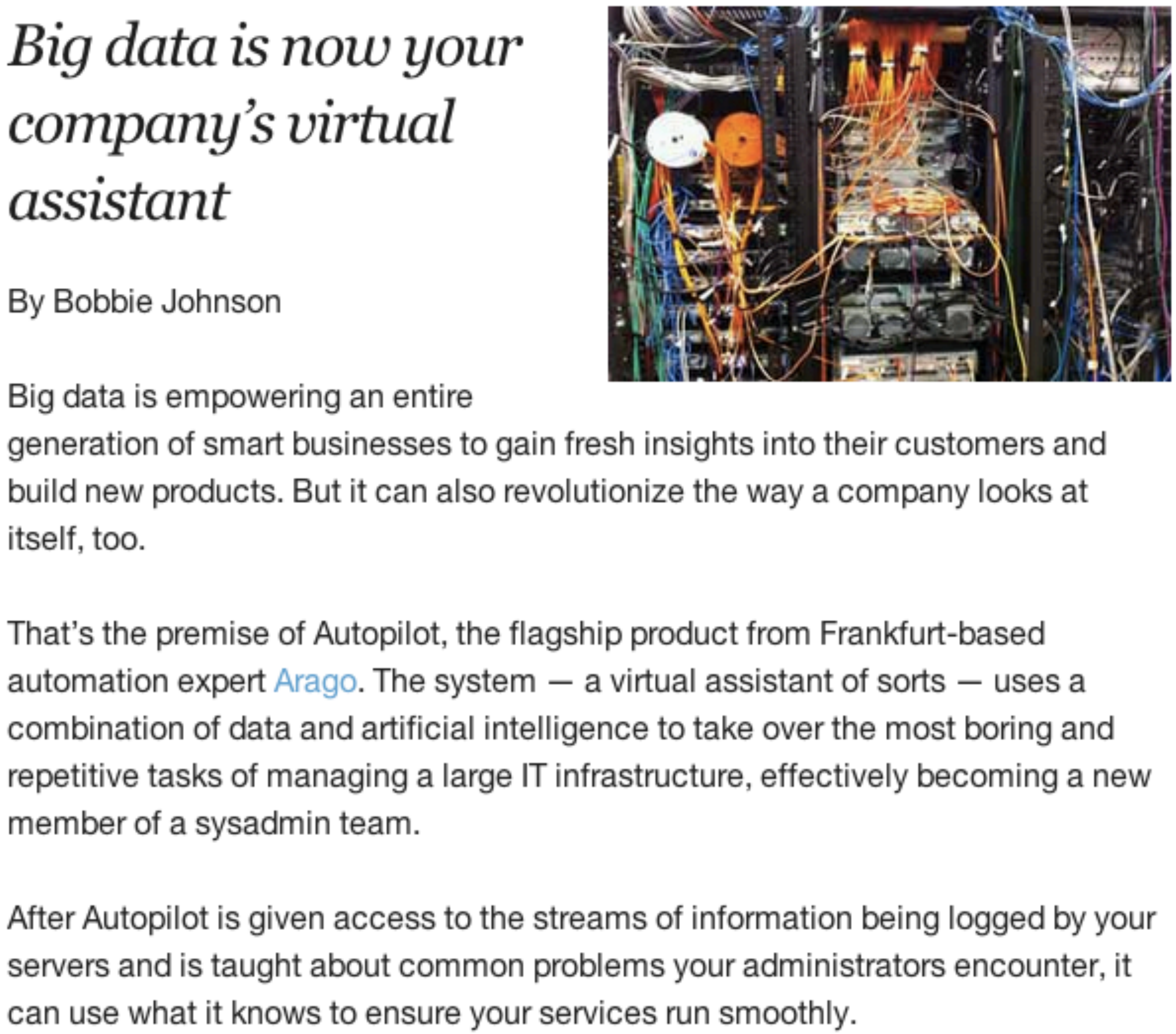
By Stacey Higginbotham

Want to know how playing on Jimmy Kimmel Live will boost the sales of an artist's album? Or how about figuring out where fans go to find artists after they hit the evening news? What about the effect Whitney Houston's death had on her YouTube and Vevo plays? They shot up 4,525 percent, by the way.

If you want to know this and other music industry data gleaned from the Internet, then you want to turn to Next Big Sound, which exists to find the connection between social activity and music sales.

The service, which recently raised $6.5 million, began two years ago because its founders thought the influx of data — from social networks like MySpace and Twitter, online music services such as Rdio, and sales sites — might help them understand how someone transitions from being a member of a band to being a full-fledged rock star.

Source [3]



Big data is now your company's virtual assistant

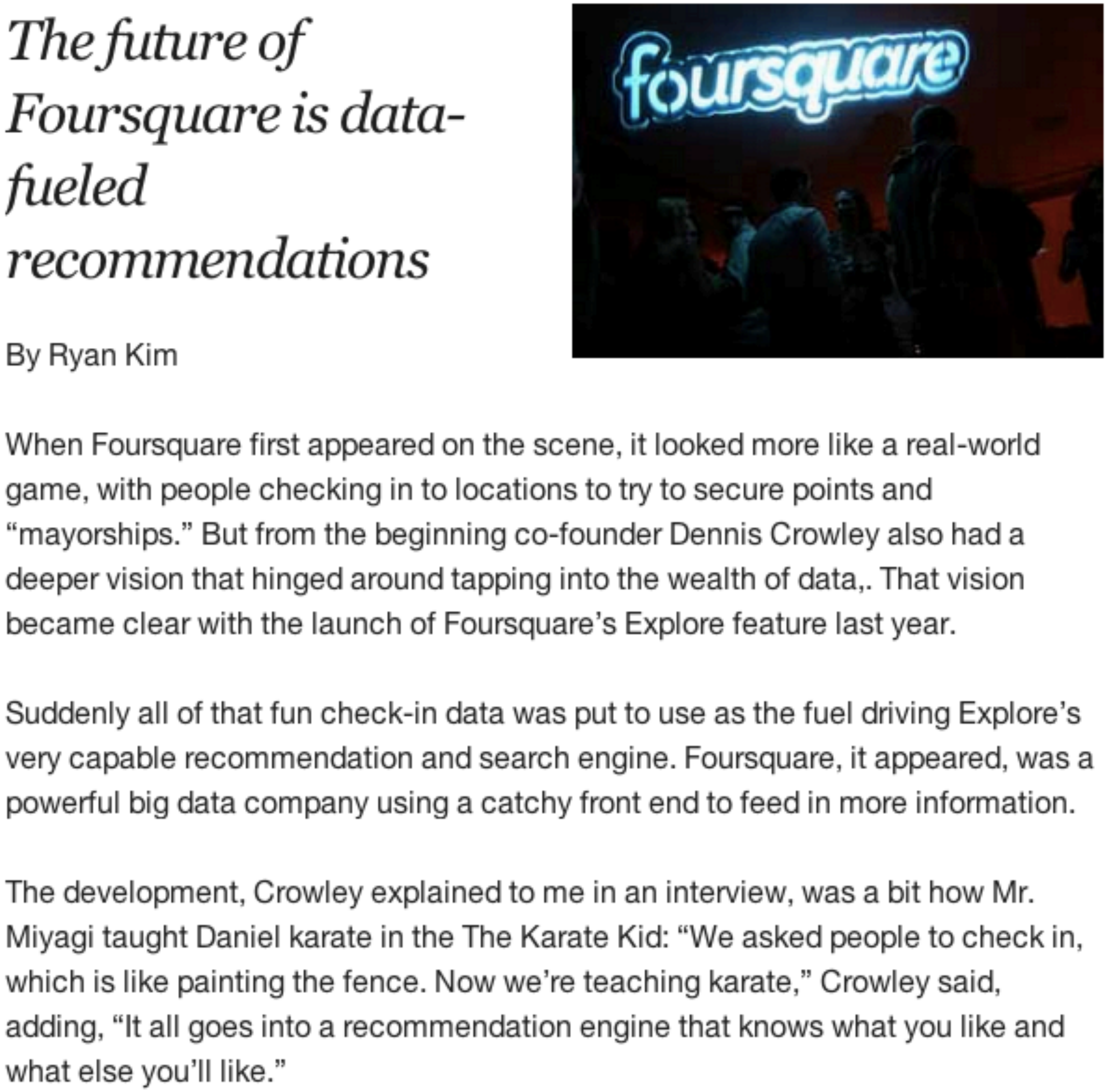
By Bobbie Johnson

Big data is empowering an entire generation of smart businesses to gain fresh insights into their customers and build new products. But it can also revolutionize the way a company looks at itself, too.

That's the premise of Autopilot, the flagship product from Frankfurt-based automation expert Arago. The system — a virtual assistant of sorts — uses a combination of data and artificial intelligence to take over the most boring and repetitive tasks of managing a large IT infrastructure, effectively becoming a new member of a sysadmin team.

After Autopilot is given access to the streams of information being logged by your servers and is taught about common problems your administrators encounter, it can use what it knows to ensure your services run smoothly.

Source [3]


foursquare

The future of Foursquare is data-fueled recommendations

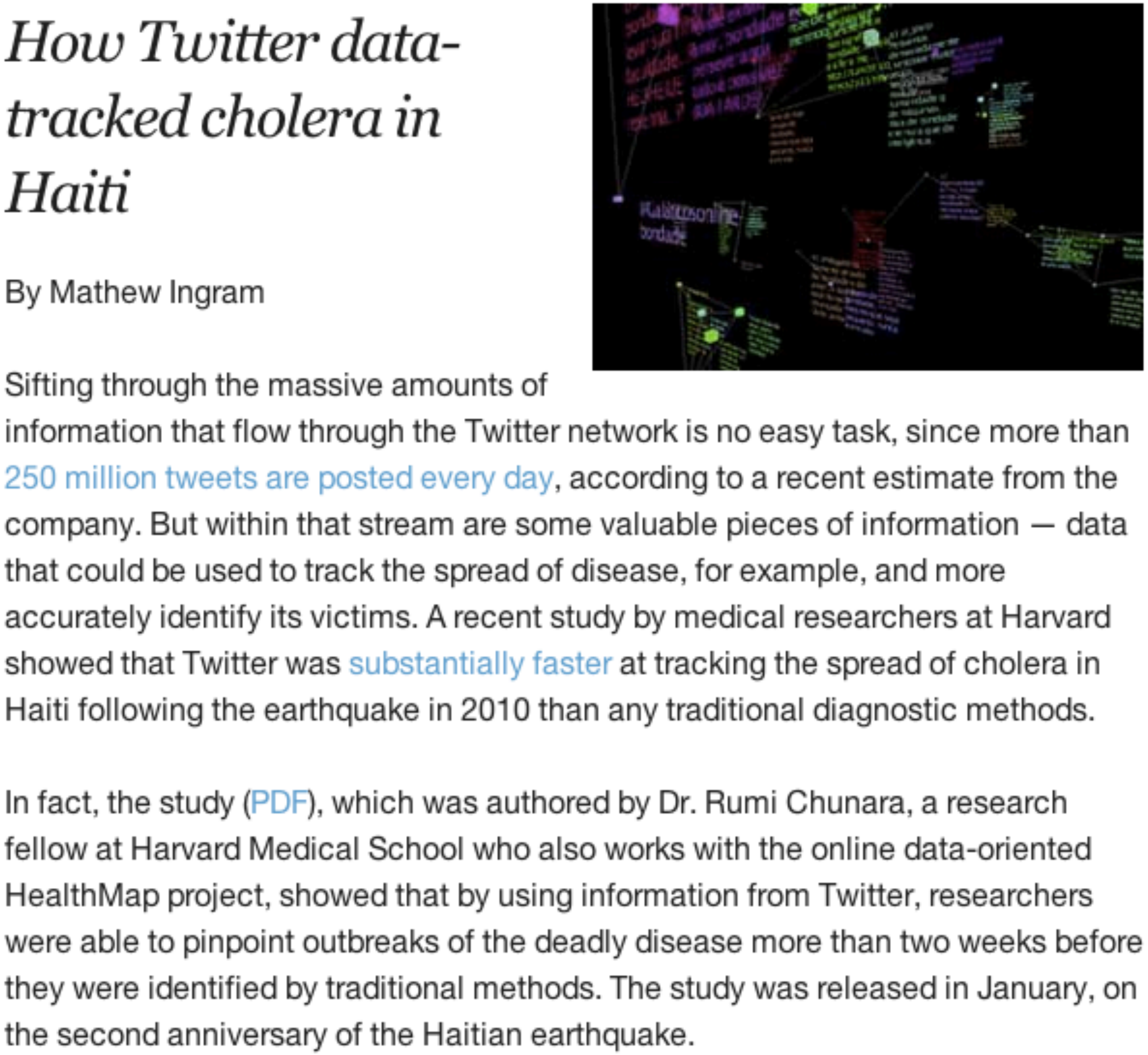
By Ryan Kim

When Foursquare first appeared on the scene, it looked more like a real-world game, with people checking in to locations to try to secure points and "mayorships." But from the beginning co-founder Dennis Crowley also had a deeper vision that hinged around tapping into the wealth of data,. That vision became clear with the launch of Foursquare's Explore feature last year.

Suddenly all of that fun check-in data was put to use as the fuel driving Explore's very capable recommendation and search engine. Foursquare, it appeared, was a powerful big data company using a catchy front end to feed in more information.

The development, Crowley explained to me in an interview, was a bit how Mr. Miyagi taught Daniel karate in the The Karate Kid: "We asked people to check in, which is like painting the fence. Now we're teaching karate," Crowley said, adding, "It all goes into a recommendation engine that knows what you like and what else you'll like."

Source [3]



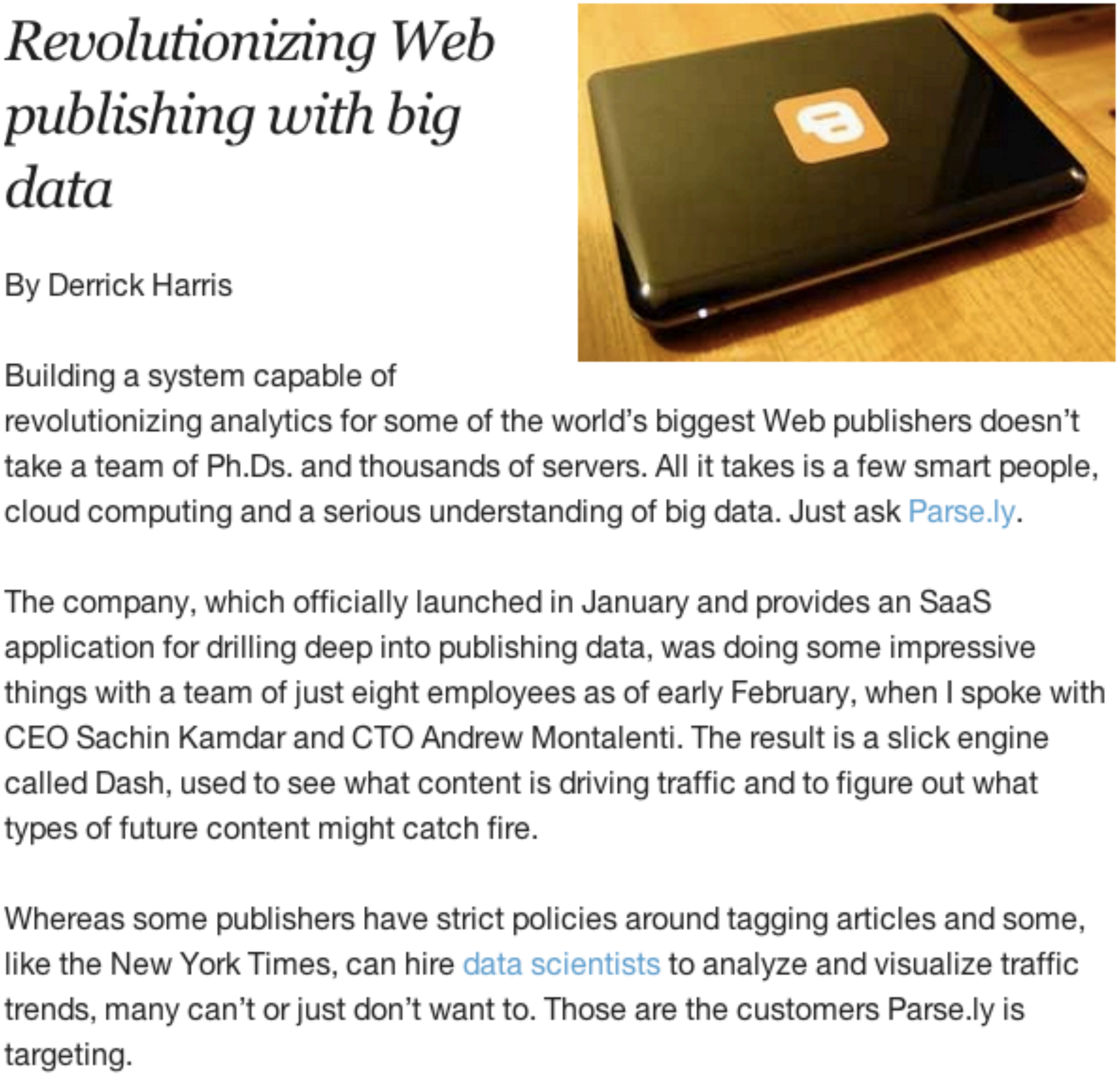
How Twitter data-tracked cholera in Haiti

By Mathew Ingram

Sifting through the massive amounts of information that flow through the Twitter network is no easy task, since more than 250 million tweets are posted every day, according to a recent estimate from the company. But within that stream are some valuable pieces of information — data that could be used to track the spread of disease, for example, and more accurately identify its victims. A recent study by medical researchers at Harvard showed that Twitter was substantially faster at tracking the spread of cholera in Haiti following the earthquake in 2010 than any traditional diagnostic methods.

In fact, the study (PDF), which was authored by Dr. Rumi Chunara, a research fellow at Harvard Medical School who also works with the online data-oriented HealthMap project, showed that by using information from Twitter, researchers were able to pinpoint outbreaks of the deadly disease more than two weeks before they were identified by traditional methods. The study was released in January, on the second anniversary of the Haitian earthquake.

Source [3]



Revolutionizing Web publishing with big data

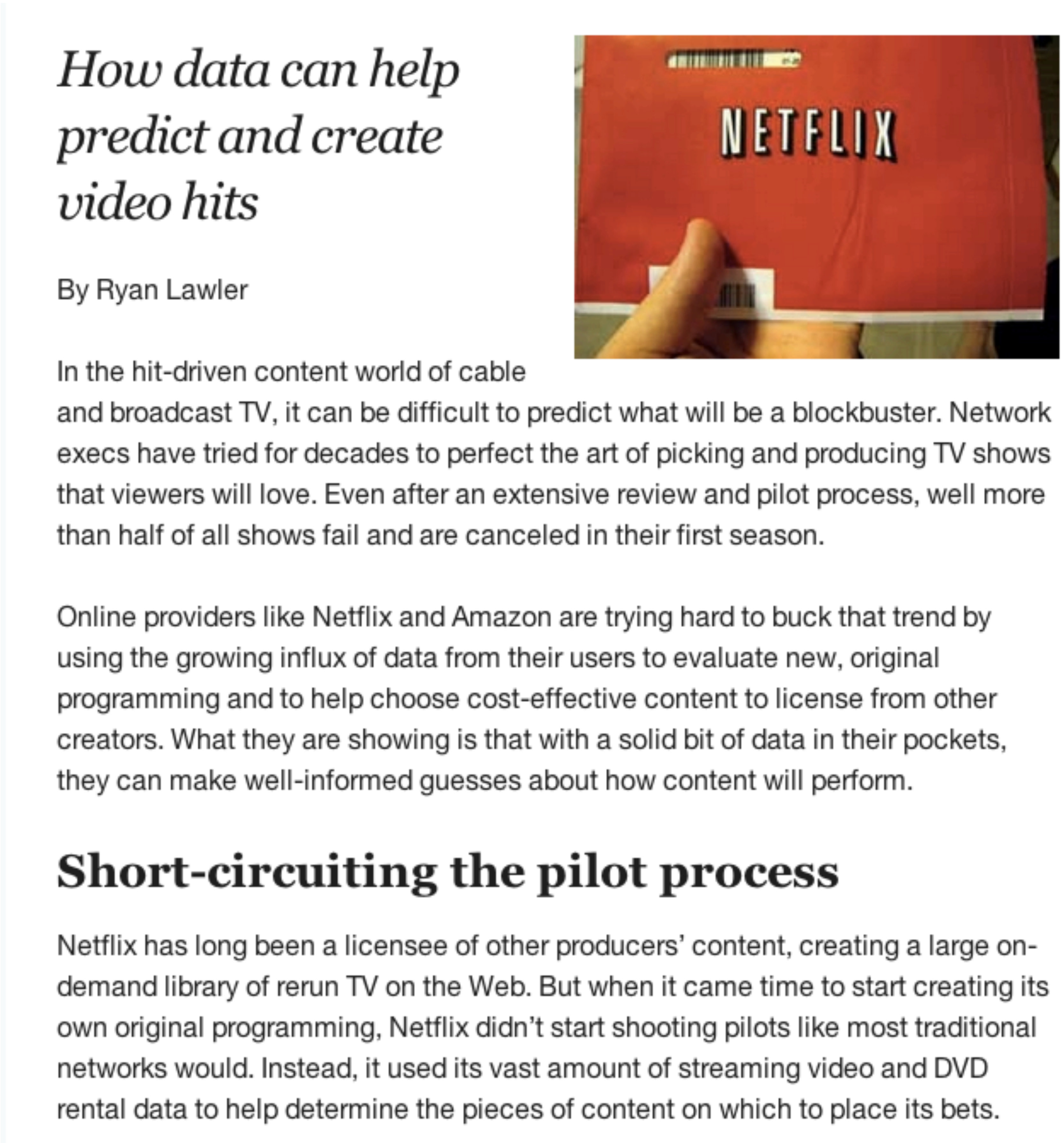
By Derrick Harris

Building a system capable of revolutionizing analytics for some of the world's biggest Web publishers doesn't take a team of Ph.Ds. and thousands of servers. All it takes is a few smart people, cloud computing and a serious understanding of big data. Just ask Parse.ly.

The company, which officially launched in January and provides an SaaS application for drilling deep into publishing data, was doing some impressive things with a team of just eight employees as of early February, when I spoke with CEO Sachin Kamdar and CTO Andrew Montalenti. The result is a slick engine called Dash, used to see what content is driving traffic and to figure out what types of future content might catch fire.

Whereas some publishers have strict policies around tagging articles and some, like the New York Times, can hire data scientists to analyze and visualize traffic trends, many can't or just don't want to. Those are the customers Parse.ly is targeting.

Source [3]



How data can help predict and create video hits

By Ryan Lawler

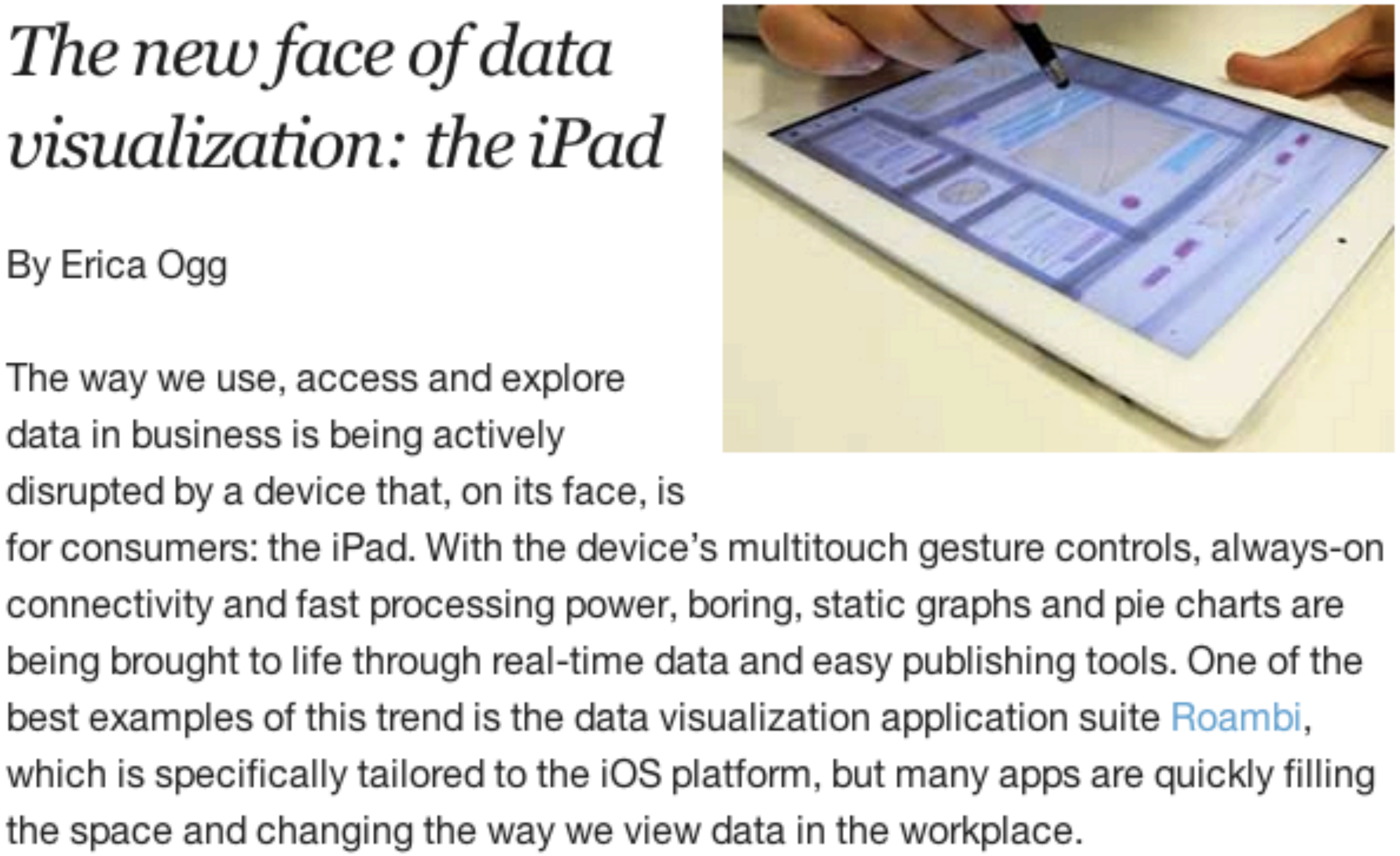
In the hit-driven content world of cable and broadcast TV, it can be difficult to predict what will be a blockbuster. Network execs have tried for decades to perfect the art of picking and producing TV shows that viewers will love. Even after an extensive review and pilot process, well more than half of all shows fail and are canceled in their first season.

Online providers like Netflix and Amazon are trying hard to buck that trend by using the growing influx of data from their users to evaluate new, original programming and to help choose cost-effective content to license from other creators. What they are showing is that with a solid bit of data in their pockets, they can make well-informed guesses about how content will perform.

Short-circuiting the pilot process

Netflix has long been a licensee of other producers' content, creating a large on-demand library of rerun TV on the Web. But when it came time to start creating its own original programming, Netflix didn't start shooting pilots like most traditional networks would. Instead, it used its vast amount of streaming video and DVD rental data to help determine the pieces of content on which to place its bets.

Source [3]

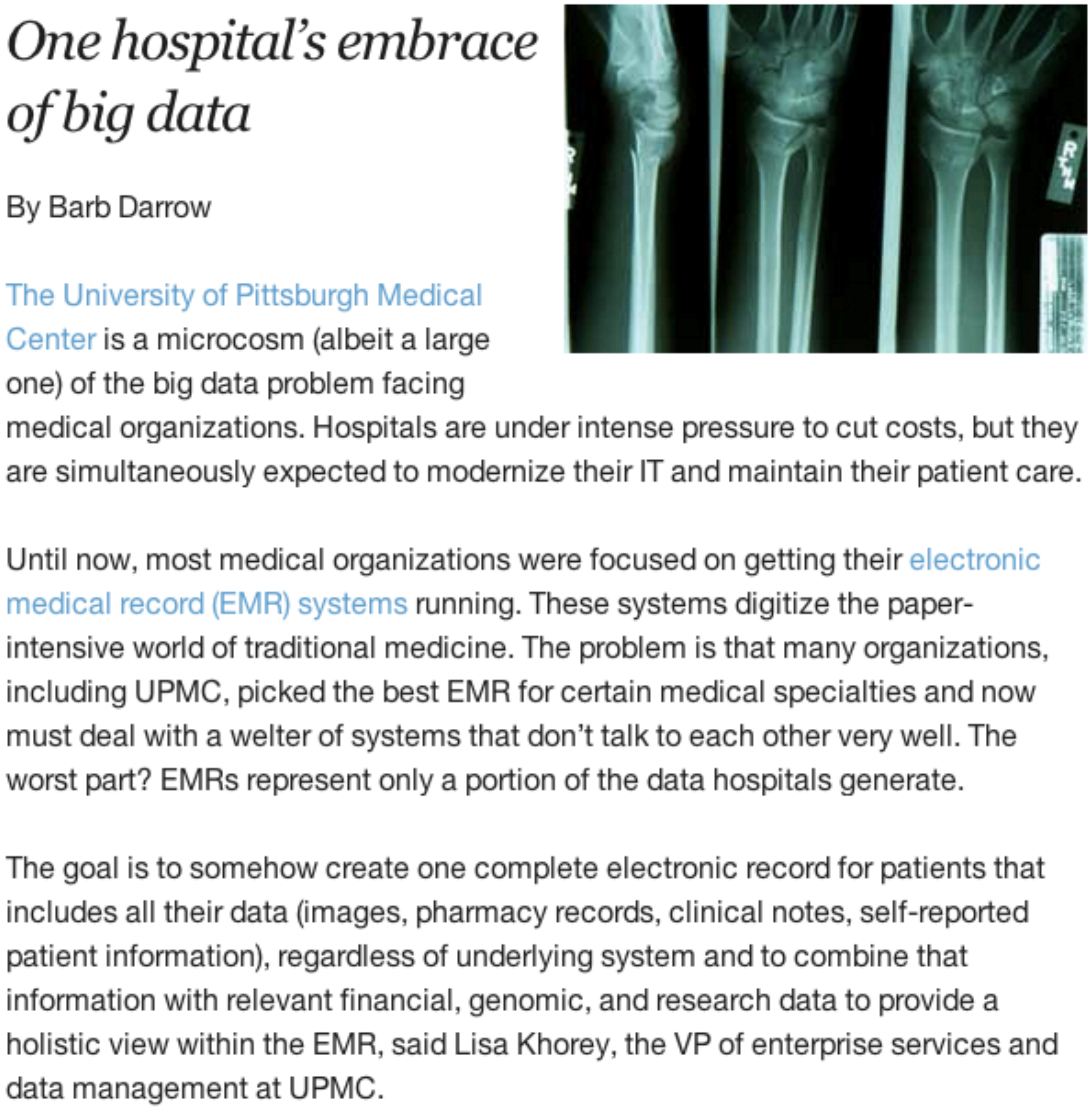


The new face of data visualization: the iPad

By Erica Ogg

The way we use, access and explore data in business is being actively disrupted by a device that, on its face, is for consumers: the iPad. With the device's multitouch gesture controls, always-on connectivity and fast processing power, boring, static graphs and pie charts are being brought to life through real-time data and easy publishing tools. One of the best examples of this trend is the data visualization application suite Roambi, which is specifically tailored to the iOS platform, but many apps are quickly filling the space and changing the way we view data in the workplace.

Source [3]



One hospital's embrace of big data

By Barb Darrow

The University of Pittsburgh Medical Center is a microcosm (albeit a large one) of the big data problem facing medical organizations. Hospitals are under intense pressure to cut costs, but they are simultaneously expected to modernize their IT and maintain their patient care.

Until now, most medical organizations were focused on getting their electronic medical record (EMR) systems running. These systems digitize the paper-intensive world of traditional medicine. The problem is that many organizations, including UPMC, picked the best EMR for certain medical specialties and now must deal with a welter of systems that don't talk to each other very well. The worst part? EMRs represent only a portion of the data hospitals generate.

The goal is to somehow create one complete electronic record for patients that includes all their data (images, pharmacy records, clinical notes, self-reported patient information), regardless of underlying system and to combine that information with relevant financial, genomic, and research data to provide a holistic view within the EMR, said Lisa Khorey, the VP of enterprise services and data management at UPMC.

Source [3]

Guest Lecture - Luis Sanz

- Guest lecture on Startup Technology Stack
  - Luis Sanz

References

- [1] The Pathologies of Big Data - Adam Jacobs
- [2] Big Data: The Management Revolution - Andrew McAfee and Erik Brynjolfsson
- [3] 10 ways big data changes everything - Editorial Staff Gigaom