O'REILLY®

Data: Emerging Trends and Technologies

How sensors, fast networks, AI, and distributed computing are affecting the data landscape

Alistair Croll

SAN JOSE

LONDON

NEW YORK

SINGAPORE

Strata+ Hadoop WORLD

Make Data Work

strataconf.com

Presented by O'Reilly and Cloudera, Strata + Hadoop World is where cutting-edge data science and new business fundamentals intersect—and merge.

- Learn business applications of data technologies
- Develop new skills through trainings and in-depth tutorials
- Connect with an international community of thousands who work with data

Job # 15420

Data: Emerging Trends and Technologies

How sensors, fast networks, AI, and distributed computing are affecting the data landscape

Alistair Croll

Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo O'REILLY®

Data: Emerging Trends and Technologies
by Alistair Croll

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://safaribooksonline.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or corporate@oreilly.com .

Editor: Tim McGovern

Interior Designer: David Futato
Cover Designer: Karen Montgomery

December 2014: First Edition

Revision History for the First Edition
2014-12-12: First Release

978-1-491-92073-2

[LSI]

Table of Contents

Introduction

Now in its fifth year, the Strata + Hadoop World conference has grown substantially from its early days. It's expanded to cover not only how we handle the flood of data our modern lives create, but also how that data is collected, governed, and acted upon.

Strata now deals with sensors that gather, clean, and aggregate information in real time, as well as machine learning and specialized data tools that make sense of such data. And it tackles the issue of interfaces by which that sense is conveyed, whether they're informing a human or directing a machine.

In this ebook, Strata + Hadoop World co-chair Alistair Croll discusses the emerging trends and technologies that will transform the data landscape in the months to come. These ideas relate to our investigation into the forces shaping the big data space, from cognitive augmentation to artificial intelligence.

Cheap Sensors, Fast Networks, and Distributed Computing

The trifecta of cheap sensors, fast networks, and distributing computing are changing how we work with data. But making sense of all that data takes help, which is arriving in the form of machine learning. Here's one view of how that might play out.

Clouds, edges, fog, and the pendulum of distributed computing

The history of computing has been a constant pendulum, swinging between centralization and distribution.

The first computers filled rooms, and operators were physically within them, switching toggles and turning wheels. Then came mainframes, which were centralized, with dumb terminals.

As the cost of computing dropped and the applications became more democratized, user interfaces mattered more. The smarter clients at the edge became the first personal computers; many broke free of the network entirely. The client got the glory; the server merely handled queries.

Once the web arrived, we centralized again. LAMP (Linux, Apache, MySQL, PHP) buried deep inside data centers, with the computer at the other end of the connection relegated to little more than a smart terminal rendering HTML. Load-balancers sprayed traffic across thousands of cheap machines. Eventually, the web turned from static sites to complex software as a service (SaaS) applications.

Then the pendulum swung back to the edge, and the clients got smart again. First with AJAX, Java, and Flash; then in the form of mobile apps where the smartphone or tablet did most of the hard work and the back-end was a communications channel for reporting the results of local action.

Now we're seeing the first iteration of the Internet of Things (IoT), in which small devices, sipping from their batteries, chatting carefully over Bluetooth LE, are little more than sensors. The preponderance of the work, from data cleaning to aggregation to analysis, has once again moved to the core: the first versions of the Jawbone Up band doesn't do much until they send their data to the cloud.

But already we can see how the pendulum will swing back. There's a renewed interest in computing at the edges—Cisco calls it "fog computing": small, local clouds that combine tiny sensors with more powerful local computing—and this may move much of the work out to the device or the local network again. Companies like realm.io are building databases that can run on smartphones or even wearables. Foghorn Systems is building platforms on which developers can deploy such multi-tiered architectures. Resin.io calls this "strong devices, weakly connected."

Systems architects understand well the tension between putting everything at the core, and making the edges more important. Centralization gives us power, makes managing changes consistent and easy, and cuts on costly latency and networking; distribution gives us more compelling user experiences, better protection against central outages or catastrophic failures, and a tiered hierarchy of processing that can scale better. Ultimately, each swing of the pendulum gives us new architectures and new bottlenecks; each rung we climb up the stack brings both abstraction and efficiency.

Machine learning

Transcendence aside, machine learning has come a long way. Deep learning approaches have significantly improved the accuracy of speech recognition, and many of the advances in the field have come from better tools and parallel computing.

Critics charge that deep learning can't account for changes over time, and as a result its categories are too brittle to use in many applications: just because something hurt yesterday doesn't mean

you should never try it again. But investment in deep learning approaches continues to pay off. And not all of the payoff comes from the fringes of science fiction.

Faced with a torrent of messy data , machine-driven approaches to data transformation and cleansing can provide a good "first pass," de-duplicating and clarifying information and replacing manual methods.

What's more, with many of these tools now available as hosted, pay-as-you-go services, it's far easier for organizations to experiment cheaply with machine-aided data processing. These are the same economics that took public cloud computing from a fringe tool for early-stage startups to a fundamental building block of enterprise IT. (More on this in "Data as a service", below.) We're keenly watching other areas where such technology is taking root in otherwise traditional organizations.

Computational Power and Cognitive Augmentation

Here's a look at a few of the ways that humans—still the ultimate data processors—mesh with the rest of our data systems: how computational power can best produce true cognitive augmentation.

Deciding better

Over the past decade, we fitted roughly a quarter of our species with sensors. We instrumented our businesses, from the smallest market to the biggest factory. We began to consume that data, slowly at first. Then, as we were able to connect data sets to one another, the applications snowballed. Now that both the front-office and the back-office are plugged into everything, business cares. A lot.

While early adopters focused on sales, marketing, and online activity, today, data gathering and analysis is ubiquitous. Governments, activists, mining giants, local businesses, transportation, and virtually every other industry lives by data. If an organization isn't harnessing the data exhaust it produces, it'll soon be eclipsed by more analytical, introspective competitors that learn and adapt faster.

Whether we're talking about a single human made more productive by a smartphone turned prosthetic brain; or a global organization gaining the ability to make more informed decisions more quickly, ultimately, Strata + Hadoop World has become about deciding better.

What does it take to make better decisions? How will we balance machine optimization with human inspiration, sometimes making

the best of the current game and other times changing the rules? Will machines that make recommendations about the future based on the past reduce risk, raise barriers to innovation, or make us vulnerable to improbable Black Swans because they mistakenly conclude that tomorrow is like yesterday, only more so?

Designing for interruption

Tomorrow's interfaces won't be about mobility, or haptics, or augmented reality (AR), or HUDs, or voice activation. I mean, they will be, but that's just the icing. They'll be about *interruption*.

In his book *Consilience*, E. O. Wilson said: "We are drowning in information…the world henceforth will be run by synthesizers, people able to put together the right information at the right time, think critically about it, and make important choices wisely." Only it won't be people doing that synthesis, it'll be a hybrid of humans and machines. Because after all, the right information at the right time changes your life.

That interruption will take many forms—a voice on a phone; a buzz on a bike handlebar; a heads-up display over actual heads. But behind it is a tremendous amount of context that helps us to decide better.

Right now, there are three companies on the planet that could do this. Microsoft's Cortana; Google's Now; and Apple's Siri are all starting down the path to prosthetic brains. A few others—Samsung, Facebook, Amazon—might try to make it happen, too. When it finally does happen, it'll be the fundamental shift of the twenty-first century, the way machines were in the nineteenth and computers were in the twentieth, because it will create a new species. Call it *Homo Conexus*.

Add iBeacons and health data to things like GPS, your calendar, crowdsourced map congestion, movement, and temperature data, etc., and machines will be more intimate, and more diplomatic, than even the most polished personal assistants.

These agents will empathize better and far more quickly than humans can. Consider two users, Mike and Tammy. Mike hates being interrupted: when his device interrupts, and it senses his racing pulse and the stress tones in his voice, it will stop. When Tammy's device interrupts, and her pupils dilate in technological

lust, it will interrupt more often. Factor in heart rate, galvanic response, and multiply by a million users with a thousand data points a day, and it's a simple baby-step toward the human-machine hybrid.

We've seen examples of contextual push models in the past. Doc Searls' suggestion of Vendor Relationship Management (VRM), in which consumers control what they receive by opting in to that in which they're interested, was a good idea. Those plans came before their time; today, however, a huge and still-increasing percentage of the world population has some kind of push-ready mobile device and a data plan.

The rise of design-for-interruption might also lead to an interruption "arms race" of personal agents trying to filter out all but the most important content, and third-party engines competing to be the most important thing in your notification center.

In discussing this with Jon Bruner, he pointed out that some of these changes will happen over time, as we make peace with our second brains:

"There's a process of social refinement that takes place when new things become widespread enough to get annoying. Everything from cars—for which traffic rules had to be invented after a couple years of gridlock—to cell phones ('guy talking loudly in a public place' is, I think, a less common nuisance than it used to be) have threatened to overload social convention when they became universal. There's a strong reaction, and then a reengineering of both convention and behavior results in a moderate outcome."

This trend leads to fascinating moral and ethical questions:

- Will a connected, augmented species quickly leave the disconnected in its digital dust, the way humans outstripped Neanderthals?
- What are the ethical implications of this?
- Will such brains make us more vulnerable?
- Will we rely on them too much?
- Is there a digital equivalent of eminent domain? Or simply the equivalent of an Amber Alert?
- What kind of damage might a powerful and politically motivated attacker wreak on a targeted nation, and how would this affect productivity or even cost lives?

- How will such machines "dream" and work on sense-making and garbage collection in the background the way humans do as they sleep?
- What interfaces are best for human-machine collaboration?
- And what protections of privacy, unreasonable search and seizure, and legislative control should these prosthetic brains enjoy?

There are also fascinating architectural changes. From a systems perspective, designing for interruption implies fundamental rethinking of many of our networks and applications, too. Systems architecture shifts from waiting and responding to pushing out "smart" interruptions based on data and context.

The Maturing Marketplace

Here's a look at some options in the evolving, maturing marketplace of big data components that are making the new applications and interactions that we've been looking at possible.

Graph theory

First used in social network analysis, graph theory is finding more and more homes in research and business. Machine learning systems can scale up fast with tools like Parameter Server, and the TitanDB project means developers have a robust set of tools to use.

Are graphs poised to take their place alongside relational database management systems (RDBMS), object storage, and other fundamental data building blocks? What are the new applications for such tools?

Inside the black box of algorithms: whither regulation?

It's possible for a machine to create an algorithm no human can understand. Evolutionary approaches to algorithmic optimization can result in inscrutable—yet demonstrably better—computational solutions.

If you're a regulated bank, you need to share your algorithms with regulators. But if you're a private trader, you're under no such constraints. And having to explain your algorithms limits how you can generate them.

As more and more of our lives are governed by code that decides what's best for us, replacing laws, actuarial tables, personal trainers and personal shoppers, oversight means opening up the black box of algorithms so they can be regulated.

Years ago, Orbitz was shown to be charging web visitors who owned Apple devices more money than those visiting via other platforms, such as the PC. Only that's not the whole story: Orbitz's machine learning algorithms, which optimized revenue per customer, learned that the visitor's browser was a predictor of their willingness to pay more.

Is this digital goldlining an upselling equivalent of redlining? Is a black-box algorithm inherently dangerous, brittle, vulnerable to runaway trading and ignorant of unpredictable, impending catastrophes? How should we balance the need to optimize quickly with the requirement for oversight?

Automation

Marc Andreesen's famous line that "software eats everything" is pretty true. It's already finished its first course. Zeynep Tufecki says that first, machines came for physical labor like the digging of trenches; then for mental labor (like Logarithm tables); and now for mental skills (which require more thinking) and possibly robotics.

Is this where automation is headed? For better or for worse, modern automation isn't simply repetition. It involves adaptation, dealing with ambiguity and changing circumstance. It's about causal feedback loops, with a system edging ever closer to an ideal state.

Past Strata speaker Avinash Kaushik chides marketers for wanting real-time data, observing that we humans can't react fast enough for it to be useful. But machines can, and do, adjust in real time, turning every action into an experiment. Real-time data is the basis for a perfect learning loop.

Advances in fast, in-memory data processing deliver on the promise of cybernetics—mechanical, physical, biological, cognitive, and social systems in which an action that changes the environment in turn changes the system itself.

Data as a service

The programmable web was a great idea, here far too early. But if the old model of development was the LAMP stack, the modern equivalent is cloud, containers, and GitHub.

- Cloud services make it easy for developers to prototype quickly and test a market or an idea — building atop Paypal, Google Maps, Facebook authentication, and so on.
- Containers, moving virtual machines from data center to data center, are the fundamental building blocks of the parts we make ourselves.
- And social coding platforms like GitHub offer fecundity, encouraging re-use and letting a thousand forks of good code bloom.

Even these three legs of the modern application are getting simpler. Consumer-friendly tools like Zapier and IFTTT let anyone stitch together simple pieces of programming to perform simple, repetitive tasks across myriad web platforms. Moving up the levels of complexity, there's now Stamplay for building web apps as well.

When it comes to big data, developers no longer need to roll their own data and machine learning tools, either. Consider Google's prediction API and BigQuery, Amazon Redshift and Kinesis. Or look at the dozens of start-ups offering specialized on-demand functions for processing data streams or big data applications.

What are the trade-offs between standing on the shoulders of giants and rolling your own? When is it best to build things from scratch in the hopes of some proprietary advantage, and when does it make sense to rely on others' economies of scale? The answer isn't clear yet, but in the coming years the industry is going to find out where that balance lies, and it will the decide the fate of hundreds of new companies and technology stacks.

The Promise and Problems of Big Data

Finally, we'll look at both the light and the shadows of this new dawn, the social and moral implications of living in a deeply connected, analyzed, and informed world. This is both the promise and the peril of big data in an age of widespread sensors, fast networks, and distributed computing.

Solving the big problems

The planet's systems are under strain from a burgeoning population. Scientists warn of rising tides, droughts, ocean acidity, and accelerating extinction. Medication-resistant diseases, outbreaks fueled by globalization, and myriad other semi-apocalyptic Horsemen ride across the horizon.

Can data fix these problems? Can we extend agriculture with data? Find new cures? Track the spread of disease? Understand weather and marine patterns? General Electric's Bill Ruh says that while the company will continue to innovate in materials sciences, the place where it will see real gains is in analytics.

It's often been said that there's nothing new about big data. The "iron triangle" of Volume, Velocity, and Variety that Doug Laney coined in 2001 has been a constraint on all data since the first database. Basically, you can have any two you want fairly affordably. Consider:

- A coin-sorting machine sorts a large volume of coins rapidly—but assumes a small variety of coins. It wouldn't work well if there were hundreds of coin types.

- A public library, organized by the Dewey Decimal System, has a wide variety of books and topics, and a large volume of those books — but stacking and retrieving the books happens at a slow velocity.

No, what's new about big data is that the cost of getting all three Vs has become so cheap, it's almost not worth billing for. A Google search happens with great alacrity, combs the sum of online knowledge, and retrieves a huge variety of content types.

With new affordability comes new applications. Where once a small town might deploy another garbage truck to cope with growth, today it can affordably analyze routes to make the system more efficient. Ten years ago, a small town didn't rely on data scientists; today, it scarcely knows it's using them.

Gluten-free dieters aside, Norman Borlaug saved billions by carefully breeding wheat and increasing the world's food supply. Will the next billion meals come from data? Monsanto thinks so, and is making substantial investments in analytics to increase farm productivity.

While much of today's analytics is focused on squeezing the most out of marketing and advertising dollars, organizations like Datakind are finding new ways to tackle modern challenges. Governments and for-profit companies are making big bets that the answers to our most pressing problems lie within the very data they generate.

The death spiral of prediction

The city of Chicago thinks a computer can predict crime. But does profiling doom the future to look like the past? As Matt Stroud asks: is the computer racist?

When governments share data, that data changes behavior. If a city publishes a crime map, then the police know where they are most likely to catch criminals. Homeowners who can afford to leave will flee the area, businesses will shutter, and that high-crime prediction turns into a self-fulfilling prophecy.

Call this, somewhat inelegantly, algorithms that shit where they eat. As we consume data, it influences us. Microsoft's Kate Crawford points to a study that shows Google's search results can sway an election.

Such feedback loops can undermine the utility of algorithms. How should data scientists deal with them? Do they mean that every algorithm is only good for a limited amount of time? When should the algorithm or the resulting data be kept private for the public good? These are problems that will dog the data scientists in coming years.

Sensors, sensors everywhere

In a Craigslist post that circulated in mid-2014 (since taken down), a restaurant owner ranted about how clients had changed. Hoping to boost revenues, the story went, the restaurant hired consultants who reviewed security footage to detect patterns in diner behavior.

The restaurant happened to have 10-year-old footage of their dining area, and the consultants compared the older footage to the new recordings, concluding that smartphones had significantly altered diner behavior and the time spent in the restaurant.

If true, that's interesting news if you're a restaurateur. For the rest of us, it's a clear lesson of just how much knowledge is lurking in pictures, audio, and video that we don't yet know how to read but soon will.

Image recognition and interpretation—let alone video analysis—is a Very Hard Problem, and it may take decades before we can say, "Computer, review these two tapes and tell me what's different about them" and get a useful answer in plain English. But that day will come — computers have already cracked finding cats in online videos.

When that day arrives, every video we've shot and uploaded—even those from a decade ago—will be a kind of retroactive sensor. We haven't been very concerned about being caught on camera in the past because our behavior is hidden by the burden of reviewing footage. But just as yesterday's dumpster-diving and wiretaps gave way to today's effortless surveillance of whole populations, we'll realize that the sensors have always been around us.

Already obvious are the smart devices on nearly every street and in every room. Crowdfunding sites are a treasure-trove of such things, from smart bicycles to home surveillance. Indeed, littleBits makes it so easy to create a sensor, it's literally kids' play. And when Tesla pushes software updates to its cars, the company can change what it

collects and how it analyzes it long after the vehicle has left the showroom.

The evolution of how we collect data in a world where every output is also an intput—when you can't read a thing without it reading you back—poses immense technical and ethical challenges. But it's also a massive business opportunity, changing how we build, maintain, and recover almost everything in our lives.