O'REILLY radar Insight, analysis, and research about emerging technologies

Big data is our generation's civil rights issue, and we don't know it

What the data is must be linked to how it can be used.

by Alistair Croll | @acroll | +Alistair Croll | Comments: 11 | August 2, 2012

Data doesn't invade people's lives. *Lack of control over how it's used does.*

What's really driving so-called big data isn't the volume of information. It turns out big data doesn't have to be all that big. Rather, it's about a reconsideration of the fundamental economics of analyzing data.

For decades, there's been a fundamental tension between three attributes of databases. You can have the data fast; you can have it big; or you can have it varied. The catch is, you can't have all three at once.

analysis when processed.

I'd first heard this as the "three V's of data": Volume, Variety, and Velocity. Traditionally, getting two was easy but getting three was very, very, very expensive.

The advent of clouds, platforms like Hadoop, and the inexorable march of Moore's Law means that now, analyzing data is trivially inexpensive. And when things become so cheap that they're practically free, big changes happen — just look at the advent of steam power, or the copying of digital music, or the rise of home printing. Abundance replaces scarcity, and we invent new business models.

In the old, data-is-scarce model, companies had to decide what to collect first, and then collect it. A traditional enterprise data warehouse might have tracked sales of widgets by

color, region, and size. This act of deciding what to store and how to store it is called designing the schema, and in many ways, it's the moment where someone decides what the data is about. It's the instant of context.

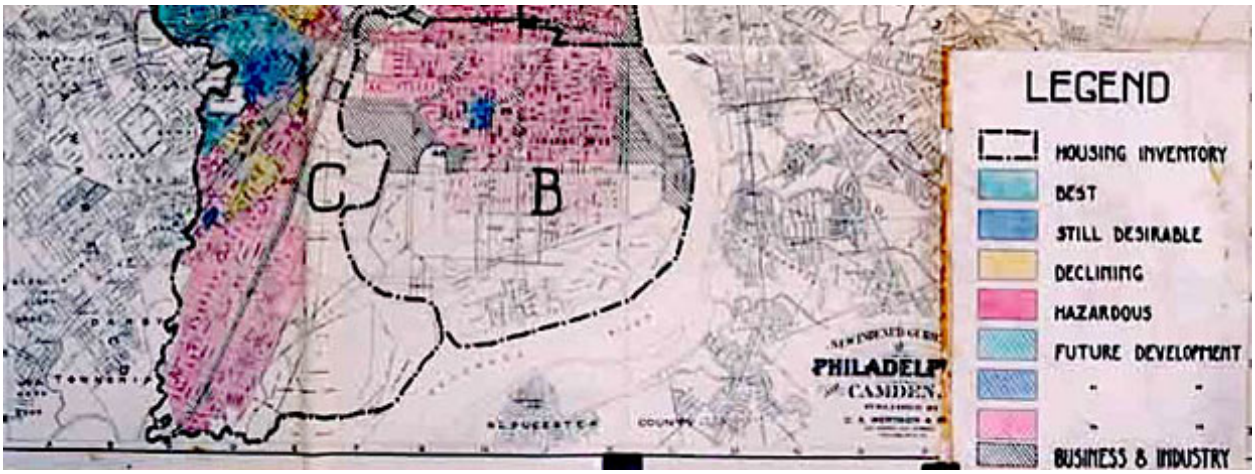
That needs repeating:

You decide what data is *about* the moment you define its schema.

With the new, data-is-abundant model, we collect first and ask questions later. The schema comes *after* the collection. Indeed, big data success stories like Splunk, Palantir, and others are prized because of their ability to make sense of content well after it's been collected — sometimes called a schema-less query. This means we collect information long before we decide what it's for.

And this is a dangerous thing.

When bank managers tried to restrict loans to residents of certain areas (known as redlining) Congress stepped in to stop it (with the Fair Housing Act of 1968). They were able to legislate against discrimination, making it illegal to change loan policy based on someone's race.



Home Owners' Loan Corporation map showing redlining of "hazardous" districts in 1936.

"Personalization" is another word for discrimination. We're not discriminating if we tailor things to you based on what we know about you — right? That's just better service.

In one case, American Express used purchase history to adjust credit limits based on where a customer shopped, despite his excellent credit limit:

> Johnson says his jaw dropped when he read one of the reasons American Express gave for lowering his credit limit: "Other customers who have used their card at establishments where you recently shopped have a poor repayment history with American Express."

We're seeing the start of this slippery slope everywhere from tailored credit-card limits like this one to car insurance based on driver profiles. In this regard, big data is a civil rights issue, but it's one that society in general is ill-equipped to deal with.

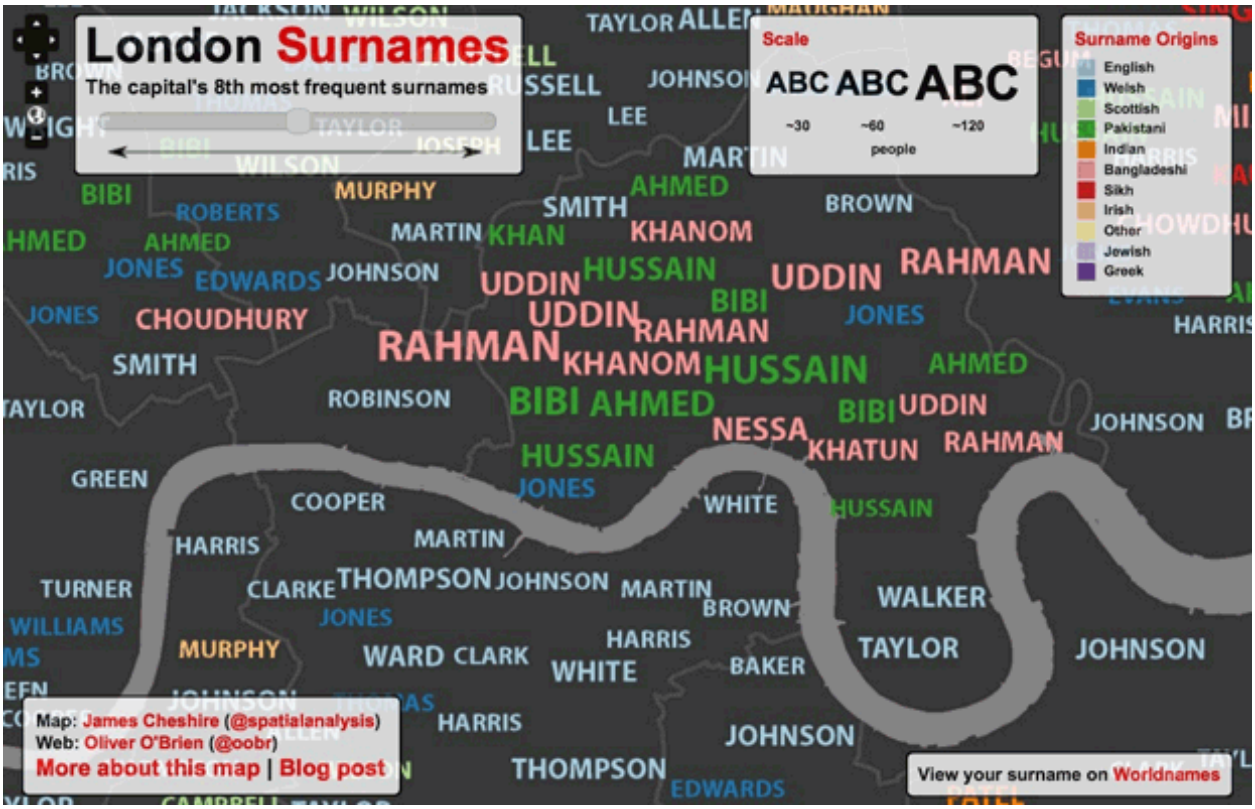
We're great at using taste to predict things about people. OKcupid's 2010 blog post "The Real Stuff White People Like" showed just how easily we can use information to guess at race. It's a real eye-opener (and the guys who wrote it didn't include everything they learned — some of it was a bit too controversial). They simply looked at the words one group used which others didn't often use. The result was a list of "trigger" words for a particular race or gender.

Now run this backwards. If I know you like these things, or see you mention them in blog posts, on Facebook, or in tweets, then there's a good chance I know your gender and your race, and maybe even your religion and your sexual orientation. And that I can personalize my marketing efforts towards you.

That makes it a civil rights issue.

If I collect information on the music you listen to, you might assume I will use that data in order to suggest new songs, or share it with your friends. But instead, I could use it to guess at your racial background. And then I could use that data to deny you a loan.

Want another example? Check out Private Data In Public Ways, something I wrote a few months ago after seeing a talk at Big Data London, which discusses how publicly available last name information can be used to generate racial boundary maps:



Screen from the Mapping London project.

This TED talk by Malte Spitz does a great job of explaining the challenges of tracking citizens today, and he speculates about whether the Berlin Wall would ever have come down if the Stasi had access to phone records in the way today's governments do.

So how do we regulate the way data is used?

The only way to deal with this properly is to somehow link *what the data is* with *how it can be*

used. I might, for example, say that my musical tastes should be used for song recommendation, but not for banking decisions.

Tying data to permissions can be done through encryption, which is slow, riddled with DRM, burdensome, hard to implement, and bad for innovation. Or it can be done through legislation, which has about as much chance of success as regulating spam: it feels great, but it's damned hard to enforce.

There are brilliant examples of how a quantified society can improve the way we live, love, work, and play. Big data helps detect disease outbreaks, improve how students learn, reveal political partisanship, and save hundreds of millions of dollars for commuters — to pick just four examples. These are benefits we simply can't ignore as we try to survive on a planet bursting with people and shaken by climate and energy crises.

But governments need to balance reliance on data with checks and balances about how this reliance erodes privacy and creates civil and moral issues we haven't thought through. It's something that most of the electorate isn't thinking about, and yet it affects every purchase they make.

This should be fun.

This post originally appeared on Solve for Interesting. This version has been lightly edited.

Related:

- Private Data In Public Ways
- Cooking the data
- There's no such thing as big data