2012 Edition

Big Data Now

Current Perspectives from O'Reilly Media

O'REILLY®

O'REILLY®
Strata
Making Data Work

Change the world with data. We'll show you how.

strataconf.com

O'REILLY®

Strata RX CONFERENCE

Data Makes a Difference

Sep 25 – 27, 2013
Boston, MA

O'REILLY®

Strata CONFERENCE + HADOOP WORLD

Oct 28 – 30, 2013
New York, NY

O'REILLY®

Strata CONFERENCE

Making Data Work

Nov 11 – 13, 2013
London, England

O'REILLY®

Spreading the knowledge of innovators.

Big Data Now: 2012 Edition

O'Reilly Media, Inc.

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

Big Data Now: 2012 Edition

by O'Reilly Media, Inc.

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://my.safaribooksonline.com*). For more information, contact our corporate/institutional sales department: (800) 998-9938 or *corporate@oreilly.com*.

Cover Designer: Karen Montgomery **Interior Designer:** David Futato

October 2012: First Edition

Revision History for the First Edition:

2012-10-24 First release

See *http://oreilly.com/catalog/errata.csp?isbn=9781449356712* for release details.

ISBN: 978-1-449-35671-2

Table of Contents

1. Introduction. 1

2. Getting Up to Speed with Big Data. 3
What Is Big Data? 3
What Does Big Data Look Like? 4
In Practice 8
What Is Apache Hadoop? 10
The Core of Hadoop: MapReduce 11
Hadoop's Lower Levels: HDFS and MapReduce 11
Improving Programmability: Pig and Hive 12
Improving Data Access: HBase, Sqoop, and Flume 12
Coordination and Workflow: Zookeeper and Oozie 14
Management and Deployment: Ambari and Whirr 14
Machine Learning: Mahout 14
Using Hadoop 15
Why Big Data Is Big: The Digital Nervous System 15
From Exoskeleton to Nervous System 15
Charting the Transition 16
Coming, Ready or Not 17

3. Big Data Tools, Techniques, and Strategies. 19
Designing Great Data Products 19
Objective-based Data Products 20
The Model Assembly Line: A Case Study of Optimal Decisions Group 21
Drivetrain Approach to Recommender Systems 25
Optimizing Lifetime Customer Value 28
Best Practices from Physical Data Products 31
The Future for Data Products 35

What It Takes to Build Great Machine Learning Products 35
Progress in Machine Learning 36
Interesting Problems Are Never Off the Shelf 37
Defining the Problem 39

4. The Application of Big Data. 41
Stories over Spreadsheets 41
A Thought on Dashboards 43
Full Interview 43
Mining the Astronomical Literature 43
Interview with Robert Simpson: Behind the Project and What Lies Ahead 48
Science between the Cracks 51
The Dark Side of Data 51
The Digital Publishing Landscape 52
Privacy by Design 53

5. What to Watch for in Big Data. 55
Big Data Is Our Generation's Civil Rights Issue, and We Don't Know It 55
Three Kinds of Big Data 60
Enterprise BI 2.0 60
Civil Engineering 62
Customer Relationship Optimization 63
Headlong into the Trough 64
Automated Science, Deep Data, and the Paradox of Information 64
(Semi)Automated Science 65
Deep Data 67
The Paradox of Information 69
The Chicken and Egg of Big Data Solutions 71
Walking the Tightrope of Visualization Criticism 73
The Visualization Ecosystem 74
The Irrationality of Needs: Fast Food to Fine Dining 76
Grown-up Criticism 78
Final Thoughts 80

6. Big Data and Health Care. 83
Solving the Wanamaker Problem for Health Care 83
Making Health Care More Effective 85
More Data, More Sources 89

Paying for Results 90
Enabling Data 91
Building the Health Care System We Want 94
Recommended Reading 95
Dr. Farzad Mostashari on Building the Health Information Infrastructure for the Modern ePatient 96
John Wilbanks Discusses the Risks and Rewards of a Health Data Commons 100
Esther Dyson on Health Data, "Preemptive Healthcare," and the Next Big Thing 106
A Marriage of Data and Caregivers Gives Dr. Atul Gawande Hope for Health Care 112
Five Elements of Reform that Health Providers Would Rather Not Hear About 119

CHAPTER 1
Introduction

In the first edition of *Big Data Now*, the O'Reilly team tracked the birth and early development of data tools and data science. Now, with this second edition, we're seeing what happens when big data grows up: how it's being applied, where it's playing a role, and the consequences — good and bad alike — of data's ascendance.

We've organized the 2012 edition of *Big Data Now* into five areas:

Getting Up to Speed With Big Data — Essential information on the structures and definitions of big data.

Big Data Tools, Techniques, and Strategies — Expert guidance for turning big data theories into big data products.

The Application of Big Data — Examples of big data in action, including a look at the downside of data.

What to Watch for in Big Data — Thoughts on how big data will evolve and the role it will play across industries and domains.

Big Data and Health Care — A special section exploring the possibilities that arise when data and health care come together.

In addition to *Big Data Now*, you can stay on top of the latest data developments with our ongoing analysis on O'Reilly Radar and through our Strata coverage and events series.

CHAPTER 2

Getting Up to Speed with Big Data

What Is Big Data?

By **Edd Dumbill**

Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the strictures of your database architectures. To gain value from this data, you must choose an alternative way to process it.

The hot IT buzzword of 2012, big data has become viable as cost-effective approaches have emerged to tame the volume, velocity, and variability of massive data. Within this data lie valuable patterns and information, previously hidden because of the amount of work required to extract them. To leading corporations, such as Walmart or Google, this power has been in reach for some time, but at fantastic cost. Today's commodity hardware, cloud architectures and open source software bring big data processing into the reach of the less well-resourced. Big data processing is eminently feasible for even the small garage startups, who can cheaply rent server time in the cloud.

The value of big data to an organization falls into two categories: analytical use and enabling new products. Big data analytics can reveal insights hidden previously by data too costly to process, such as peer influence among customers, revealed by analyzing shoppers' transactions and social and geographical data. Being able to process every item of data in reasonable time removes the troublesome need for sampling and promotes an investigative approach to data, in contrast to the somewhat static nature of running predetermined reports.

The past decade's successful web startups are prime examples of big data used as an enabler of new products and services. For example, by combining a large number of signals from a user's actions and those of their friends, Facebook has been able to craft a highly personalized user experience and create a new kind of advertising business. It's no coincidence that the lion's share of ideas and tools underpinning big data have emerged from Google, Yahoo, Amazon, and Facebook.

The emergence of big data into the enterprise brings with it a necessary counterpart: agility. Successfully exploiting the value in big data requires experimentation and exploration. Whether creating new products or looking for ways to gain competitive advantage, the job calls for curiosity and an entrepreneurial outlook.

What Does Big Data Look Like?

As a catch-all term, "big data" can be pretty nebulous, in the same way that the term "cloud" covers diverse technologies. Input data to big data systems could be chatter from social networks, web server logs, traffic flow sensors, satellite imagery, broadcast audio streams, banking transactions, MP3s of rock music, the content of web pages, scans of government documents, GPS trails, telemetry from automobiles, financial market data, the list goes on. Are these all really the same thing?

To clarify matters, the three Vs of *volume, velocity, and variety* are commonly used to characterize different aspects of big data. They're a helpful lens through which to view and understand the nature of the data and the software platforms available to exploit them. Most probably you will contend with each of the Vs to one degree or another.

Volume

The benefit gained from the ability to process large amounts of information is the main attraction of big data analytics. Having more data beats out having better models: simple bits of math can be unreasonably effective given large amounts of data. If you could run that forecast taking into account 300 factors rather than 6, could you predict demand better? This volume presents the most immediate challenge to conventional IT structures. It calls for scalable storage, and a distributed approach to querying. Many companies already have large amounts of archived data, perhaps in the form of logs, but not the capacity to process it.

Assuming that the volumes of data are larger than those conventional relational database infrastructures can cope with, processing options break down broadly into a choice between massively parallel processing architectures — data warehouses or databases such as Greenplum — and Apache Hadoop-based solutions. This choice is often informed by the degree to which one of the other "Vs" — variety — comes into play. Typically, data warehousing approaches involve predetermined schemas, suiting a regular and slowly evolving dataset. Apache Hadoop, on the other hand, places no conditions on the structure of the data it can process.

At its core, Hadoop is a platform for distributing computing problems across a number of servers. First developed and released as open source by Yahoo, it implements the MapReduce approach pioneered by Google in compiling its search indexes. Hadoop's MapReduce involves distributing a dataset among multiple servers and operating on the data: the "map" stage. The partial results are then recombined: the "reduce" stage.

To store data, Hadoop utilizes its own distributed filesystem, HDFS, which makes data available to multiple computing nodes. A typical Hadoop usage pattern involves three stages:

- loading data into HDFS,
- MapReduce operations, and
- retrieving results from HDFS.

This process is by nature a batch operation, suited for analytical or non-interactive computing tasks. Because of this, Hadoop is not itself a database or data warehouse solution, but can act as an analytical adjunct to one.

One of the most well-known Hadoop users is Facebook, whose model follows this pattern. A MySQL database stores the core data. This is then reflected into Hadoop, where computations occur, such as creating recommendations for you based on your friends' interests. Facebook then transfers the results back into MySQL, for use in pages served to users.

Velocity

The importance of data's velocity — the increasing rate at which data flows into an organization — has followed a similar pattern to that of

volume. Problems previously restricted to segments of industry are now presenting themselves in a much broader setting. Specialized companies such as financial traders have long turned systems that cope with fast moving data to their advantage. Now it's our turn.

Why is that so? The Internet and mobile era means that the way we deliver and consume products and services is increasingly instrumented, generating a data flow back to the provider. Online retailers are able to compile large histories of customers' every click and interaction: not just the final sales. Those who are able to quickly utilize that information, by recommending additional purchases, for instance, gain competitive advantage. The smartphone era increases again the rate of data inflow, as consumers carry with them a streaming source of geolocated imagery and audio data.

It's not just the velocity of the incoming data that's the issue: it's possible to stream fast-moving data into bulk storage for later batch processing, for example. The importance lies in the speed of the feedback loop, taking data from input through to decision. A commercial from IBM makes the point that you wouldn't cross the road if all you had was a five-minute old snapshot of traffic location. There are times when you simply won't be able to wait for a report to run or a Hadoop job to complete.

Industry terminology for such fast-moving data tends to be either "streaming data" or "complex event processing." This latter term was more established in product categories before streaming processing data gained more widespread relevance, and seems likely to diminish in favor of streaming.

There are two main reasons to consider streaming processing. The first is when the input data are too fast to store in their entirety: in order to keep storage requirements practical, some level of analysis must occur as the data streams in. At the extreme end of the scale, the Large Hadron Collider at CERN generates so much data that scientists must discard the overwhelming majority of it — hoping hard they've not thrown away anything useful. The second reason to consider streaming is where the application mandates immediate response to the data. Thanks to the rise of mobile applications and online gaming this is an increasingly common situation.

Product categories for handling streaming data divide into established proprietary products such as IBM's InfoSphere Streams and the less-polished and still emergent open source frameworks originating in the web industry: Twitter's Storm and Yahoo S4.

As mentioned above, it's not just about input data. The velocity of a system's outputs can matter too. The tighter the feedback loop, the greater the competitive advantage. The results might go directly into a product, such as Facebook's recommendations, or into dashboards used to drive decision-making. It's this need for speed, particularly on the Web, that has driven the development of key-value stores and columnar databases, optimized for the fast retrieval of precomputed information. These databases form part of an umbrella category known as NoSQL, used when relational models aren't the right fit.

Variety

Rarely does data present itself in a form perfectly ordered and ready for processing. A common theme in big data systems is that the source data is diverse, and doesn't fall into neat relational structures. It could be text from social networks, image data, a raw feed directly from a sensor source. None of these things come ready for integration into an application.

Even on the Web, where computer-to-computer communication ought to bring some guarantees, the reality of data is messy. Different browsers send different data, users withhold information, they may be using differing software versions or vendors to communicate with you. And you can bet that if part of the process involves a human, there will be error and inconsistency.

A common use of big data processing is to take unstructured data and extract ordered meaning, for consumption either by humans or as a structured input to an application. One such example is entity resolution, the process of determining exactly what a name refers to. Is this city London, England, or London, Texas? By the time your business logic gets to it, you don't want to be guessing.

The process of moving from source data to processed application data involves the loss of information. When you tidy up, you end up throwing stuff away. This underlines a principle of big data: *when you can, keep everything*. There may well be useful signals in the bits you throw away. If you lose the source data, there's no going back.

Despite the popularity and well understood nature of relational databases, it is not the case that they should always be the destination for data, even when tidied up. Certain data types suit certain classes of database better. For instance, documents encoded as XML are most versatile when stored in a dedicated XML store such as MarkLogic. Social network relations are graphs by nature, and graph databases such as Neo4J make operations on them simpler and more efficient.

Even where there's not a radical data type mismatch, a disadvantage of the relational database is the static nature of its schemas. In an agile, exploratory environment, the results of computations will evolve with the detection and extraction of more signals. Semi-structured NoSQL databases meet this need for flexibility: they provide enough structure to organize data, but do not require the exact schema of the data before storing it.

In Practice

We have explored the nature of big data and surveyed the landscape of big data from a high level. As usual, when it comes to deployment there are dimensions to consider over and above tool selection.

Cloud or in-house?

The majority of big data solutions are now provided in three forms: software-only, as an appliance or cloud-based. Decisions between which route to take will depend, among other things, on issues of data locality, privacy and regulation, human resources and project requirements. Many organizations opt for a hybrid solution: using on-demand cloud resources to supplement in-house deployments.

Big data is big

It is a fundamental fact that data that is too big to process conventionally is also too big to transport anywhere. IT is undergoing an inversion of priorities: it's the program that needs to move, not the data. If you want to analyze data from the U.S. Census, it's a lot easier to run your code on Amazon's web services platform, which hosts such data locally, and won't cost you time or money to transfer it.

Even if the data isn't too big to move, locality can still be an issue, especially with rapidly updating data. Financial trading systems crowd into data centers to get the fastest connection to source data, because that millisecond difference in processing time equates to competitive advantage.

Big data is messy

It's not all about infrastructure. Big data practitioners consistently report that 80% of the effort involved in dealing with data is cleaning it up in the first place, as Pete Warden observes in his Big Data Glossary: "I probably spend more time turning messy source data into something usable than I do on the rest of the data analysis process combined."

Because of the high cost of data acquisition and cleaning, it's worth considering what you actually need to source yourself. Data marketplaces are a means of obtaining common data, and you are often able to contribute improvements back. Quality can of course be variable, but will increasingly be a benchmark on which data marketplaces compete.

Culture

The phenomenon of big data is closely tied to the emergence of data science, a discipline that combines math, programming, and scientific instinct. Benefiting from big data means investing in teams with this skillset, and surrounding them with an organizational willingness to understand and use data for advantage.

In his report, "Building Data Science Teams," D.J. Patil characterizes data scientists as having the following qualities:

- Technical expertise: the best data scientists typically have deep expertise in some scientific discipline.
- Curiosity: a desire to go beneath the surface and discover and distill a problem down into a very clear set of hypotheses that can be tested.
- Storytelling: the ability to use data to tell a story and to be able to communicate it effectively.
- Cleverness: the ability to look at a problem in different, creative ways.

The far-reaching nature of big data analytics projects can have uncomfortable aspects: data must be broken out of silos in order to be mined, and the organization must learn how to communicate and interpet the results of analysis.

Those skills of storytelling and cleverness are the gateway factors that ultimately dictate whether the benefits of analytical labors are absorbed by an organization. The art and practice of visualizing data is becoming ever more important in bridging the human-computer gap to mediate analytical insight in a meaningful way.

Know where you want to go

Finally, remember that big data is no panacea. You can find patterns and clues in your data, but then what? Christer Johnson, IBM's leader for advanced analytics in North America, gives this advice to businesses starting out with big data: first, decide what problem you want to solve.

If you pick a real business problem, such as how you can change your advertising strategy to increase spend per customer, it will guide your implementation. While big data work benefits from an enterprising spirit, it also benefits strongly from a concrete goal.

What Is Apache Hadoop?

By Edd Dumbill

Apache Hadoop has been the driving force behind the growth of the big data industry. You'll hear it mentioned often, along with associated technologies such as Hive and Pig. But what does it do, and why do you need all its strangely named friends, such as Oozie, Zookeeper, and Flume?

Hadoop brings the ability to cheaply process large amounts of data, regardless of its structure. By large, we mean from 10-100 gigabytes and above. How is this different from what went before?

Existing enterprise data warehouses and relational databases excel at processing structured data and can store massive amounts of data, though at a cost: This requirement for structure restricts the kinds of data that can be processed, and it imposes an inertia that makes data warehouses unsuited for agile exploration of massive heterogenous data. The amount of effort required to warehouse data often means that valuable data sources in organizations are never mined. This is where Hadoop can make a big difference.

This article examines the components of the Hadoop ecosystem and explains the functions of each.

The Core of Hadoop: MapReduce

Created at Google in response to the problem of creating web search indexes, the MapReduce framework is the powerhouse behind most of today's big data processing. In addition to Hadoop, you'll find MapReduce inside MPP and NoSQL databases, such as Vertica or MongoDB.

The important innovation of MapReduce is the ability to take a query over a dataset, divide it, and run it in parallel over multiple nodes. Distributing the computation solves the issue of data too large to fit onto a single machine. Combine this technique with commodity Linux servers and you have a cost-effective alternative to massive computing arrays.

At its core, Hadoop is an open source MapReduce implementation. Funded by Yahoo, it emerged in 2006 and, according to its creator Doug Cutting, reached "web scale" capability in early 2008.

As the Hadoop project matured, it acquired further components to enhance its usability and functionality. The name "Hadoop" has come to represent this entire ecosystem. There are parallels with the emergence of Linux: The name refers strictly to the Linux kernel, but it has gained acceptance as referring to a complete operating system.

Hadoop's Lower Levels: HDFS and MapReduce

Above, we discussed the ability of MapReduce to distribute computation over multiple servers. For that computation to take place, each server must have access to the data. This is the role of HDFS, the Hadoop Distributed File System.

HDFS and MapReduce are robust. Servers in a Hadoop cluster can fail and not abort the computation process. HDFS ensures data is replicated with redundancy across the cluster. On completion of a calculation, a node will write its results back into HDFS.

There are no restrictions on the data that HDFS stores. Data may be unstructured and schemaless. By contrast, relational databases require that data be structured and schemas be defined before storing the data. With HDFS, making sense of the data is the responsibility of the developer's code.

Programming Hadoop at the MapReduce level is a case of working with the Java APIs, and manually loading data files into HDFS.

Improving Programmability: Pig and Hive

Working directly with Java APIs can be tedious and error prone. It also restricts usage of Hadoop to Java programmers. Hadoop offers two solutions for making Hadoop programming easier.

- Pig is a programming language that simplifies the common tasks of working with Hadoop: loading data, expressing transformations on the data, and storing the final results. Pig's built-in operations can make sense of semi-structured data, such as log files, and the language is extensible using Java to add support for custom data types and transformations.
- Hive enables Hadoop to operate as a data warehouse. It superimposes structure on data in HDFS and then permits queries over the data using a familiar SQL-like syntax. As with Pig, Hive's core capabilities are extensible.

Choosing between Hive and Pig can be confusing. Hive is more suitable for data warehousing tasks, with predominantly static structure and the need for frequent analysis. Hive's closeness to SQL makes it an ideal point of integration between Hadoop and other business intelligence tools.

Pig gives the developer more agility for the exploration of large datasets, allowing the development of succinct scripts for transforming data flows for incorporation into larger applications. Pig is a thinner layer over Hadoop than Hive, and its main advantage is to drastically cut the amount of code needed compared to direct use of Hadoop's Java APIs. As such, Pig's intended audience remains primarily the software developer.

Improving Data Access: HBase, Sqoop, and Flume

At its heart, Hadoop is a batch-oriented system. Data are loaded into HDFS, processed, and then retrieved. This is somewhat of a computing throwback, and often, interactive and random access to data is required.

Enter HBase, a column-oriented database that runs on top of HDFS. Modeled after Google's BigTable, the project's goal is to host billions of rows of data for rapid access. MapReduce can use HBase as both a source and a destination for its computations, and Hive and Pig can be used in combination with HBase.

In order to grant random access to the data, HBase does impose a few restrictions: Hive performance with HBase is 4-5 times slower than with plain HDFS, and the maximum amount of data you can store in HBase is approximately a petabyte, versus HDFS' limit of over 30PB.

HBase is ill-suited to ad-hoc analytics and more appropriate for integrating big data as part of a larger application. Use cases include logging, counting, and storing time-series data.

The Hadoop Bestiary

Ambari	Deployment, configuration and monitoring
Flume	Collection and import of log and event data
HBase	Column-oriented database scaling to billions of rows
HCatalog	Schema and data type sharing over Pig, Hive and MapReduce
HDFS	Distributed redundant file system for Hadoop
Hive	Data warehouse with SQL-like access
Mahout	Library of machine learning and data mining algorithms
MapReduce	Parallel computation on server clusters
Pig	High-level programming language for Hadoop computations
Oozie	Orchestration and workflow management
Sqoop	Imports data from relational databases
Whirr	Cloud-agnostic deployment of clusters
Zookeeper	Configuration management and coordination

Getting data in and out

Improved interoperability with the rest of the data world is provided by Sqoop and Flume. Sqoop is a tool designed to import data from relational databases into Hadoop, either directly into HDFS or into Hive. Flume is designed to import streaming flows of log data directly into HDFS.

Hive's SQL friendliness means that it can be used as a point of integration with the vast universe of database tools capable of making connections via JBDC or ODBC database drivers.

Coordination and Workflow: Zookeeper and Oozie

With a growing family of services running as part of a Hadoop cluster, there's a need for coordination and naming services. As computing nodes can come and go, members of the cluster need to synchronize with each other, know where to access services, and know how they should be configured. This is the purpose of Zookeeper.

Production systems utilizing Hadoop can often contain complex pipelines of transformations, each with dependencies on each other. For example, the arrival of a new batch of data will trigger an import, which must then trigger recalculations in dependent datasets. The Oozie component provides features to manage the workflow and dependencies, removing the need for developers to code custom solutions.

Management and Deployment: Ambari and Whirr

One of the commonly added features incorporated into Hadoop by distributors such as IBM and Microsoft is monitoring and administration. Though in an early stage, Ambari aims to add these features to the core Hadoop project. Ambari is intended to help system administrators deploy and configure Hadoop, upgrade clusters, and monitor services. Through an API, it may be integrated with other system management tools.

Though not strictly part of Hadoop, Whirr is a highly complementary component. It offers a way of running services, including Hadoop, on cloud platforms. Whirr is cloud neutral and currently supports the Amazon EC2 and Rackspace services.

Machine Learning: Mahout

Every organization's data are diverse and particular to their needs. However, there is much less diversity in the kinds of analyses performed on that data. The Mahout project is a library of Hadoop implementations of common analytical computations. Use cases include user collaborative filtering, user recommendations, clustering, and classification.

Using Hadoop

Normally, you will use Hadoop in the form of a distribution. Much as with Linux before it, vendors integrate and test the components of the Apache Hadoop ecosystem and add in tools and administrative features of their own.

Though not *per se* a distribution, a managed cloud installation of Hadoop's MapReduce is also available through Amazon's Elastic MapReduce service.

Why Big Data Is Big: The Digital Nervous System

By Edd Dumbill

Where does all the data in "big data" come from? And why isn't big data just a concern for companies such as Facebook and Google? The answer is that the web companies are the forerunners. Driven by social, mobile, and cloud technology, there is an important transition taking place, leading us all to the data-enabled world that those companies inhabit today.

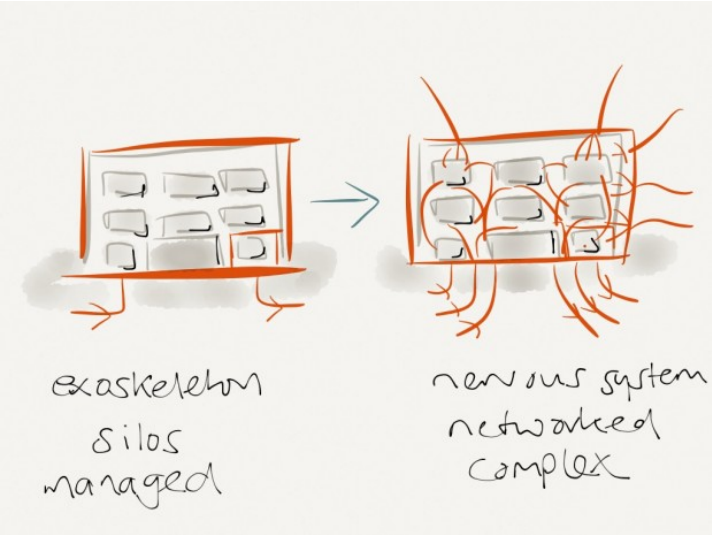
From Exoskeleton to Nervous System

Until a few years ago, the main function of computer systems in society, and business in particular, was as a digital support system. Applications digitized existing real-world processes, such as word-processing, payroll, and inventory. These systems had interfaces back out to the real world through stores, people, telephone, shipping, and so on. The now-quaint phrase "paperless office" alludes to this transfer of pre-existing paper processes into the computer. These computer systems formed a *digital exoskeleton*, supporting a business in the real world.

The arrival of the Internet and the Web has added a new dimension, bringing in an era of entirely digital business. Customer interaction, payments, and often product delivery can exist entirely within computer systems. Data doesn't just stay inside the exoskeleton any more, but is a key element in the operation. We're in an era where business and society are acquiring a *digital nervous system*.

As my sketch below shows, an organization with a digital nervous system is characterized by a large number of inflows and outflows of data, a high level of networking, both internally and externally, increased data flow, and consequent complexity.

This transition is why big data is important. Techniques developed to deal with interlinked, heterogenous data acquired by massive web companies will be our main tools as the rest of us transition to digital-native operation. We see early examples of this, from catching fraud in financial transactions to debugging and improving the hiring process in HR: and almost everybody already pays attention to the massive flow of social network information concerning them.



Charting the Transition

As technology has progressed within business, each step taken has resulted in a leap in data volume. To people looking at big data now, a reasonable question is to ask why, when their business isn't Google or Facebook, does big data apply to them?

The answer lies in the ability of web businesses to conduct 100% of their activities online. Their digital nervous system easily stretches from the beginning to the end of their operations. If you have factories, shops, and other parts of the real world within your business, you've further to go in incorporating them into the digital nervous system.

But "further to go" doesn't mean it won't happen. The drive of the Web, social media, mobile, and the cloud is bringing more of each business

into a data-driven world. In the UK, the Government Digital Service is unifying the delivery of services to citizens. The results are a radical improvement of citizen experience, and for the first time many departments are able to get a real picture of how they're doing. For any retailer, companies such as Square, American Express, and Foursquare are bringing payments into a social, responsive data ecosystem, liberating that information from the silos of corporate accounting.

What does it mean to have a digital nervous system? The key trait is to make an organization's feedback loop entirely digital. That is, a direct connection from sensing and monitoring inputs through to product outputs. That's straightforward on the Web. It's getting increasingly easier in retail. Perhaps the biggest shifts in our world will come as sensors and robotics bring the advantages web companies have now to domains such as industry, transport, and the military.

The reach of the digital nervous system has grown steadily over the past 30 years, and each step brings gains in agility and flexibility, along with an order of magnitude more data. First, from specific application programs to general business use with the PC. Then, direct interaction over the Web. Mobile adds awareness of time and place, along with instant notification. The next step, to cloud, breaks down data silos and adds storage and compute elasticity through cloud computing. Now, we're integrating smart agents, able to act on our behalf, and connections to the real world through sensors and automation.

Coming, Ready or Not

If you're not contemplating the advantages of taking more of your operation digital, you can bet your competitors are. As Marc Andreessen wrote last year, "software is eating the world." Everything is becoming programmable.

It's this growth of the digital nervous system that makes the techniques and tools of big data relevant to us today. The challenges of massive data flows, and the erosion of hierarchy and boundaries, will lead us to the statistical approaches, systems thinking, and machine learning we need to cope with the future we're inventing.

CHAPTER 3

Big Data Tools, Techniques, and Strategies

Designing Great Data Products

By Jeremy Howard, Margit Zwemer, and Mike Loukides

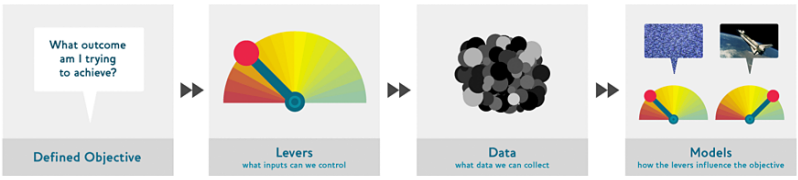
In the past few years, we've seen many data products based on predictive modeling. These products range from weather forecasting to recommendation engines to services that predict airline flight times more accurately than the airlines themselves. But these products are still just making predictions, rather than asking what action they want someone to take as a result of a prediction. Prediction technology can be interesting and mathematically elegant, but we need to take the next step. The technology exists to build data products that can revolutionize entire industries. So, why aren't we building them?

To jump-start this process, we suggest a four-step approach that has already transformed the insurance industry. We call it the *Drivetrain Approach*, inspired by the emerging field of self-driving vehicles. Engineers start by defining a clear *objective*: They want a car to drive safely from point A to point B without human intervention. Great predictive modeling is an important part of the solution, but it no longer stands on its own; as products become more sophisticated, it disappears into the plumbing. Someone using Google's self-driving car is completely unaware of the hundreds (if not thousands) of models and the petabytes of data that make it work. But as data scientists build increasingly

sophisticated products, they need a systematic design approach. We don't claim that the Drivetrain Approach is the best or only method; our goal is to start a dialog within the data science and business communities to advance our collective vision.

Objective-based Data Products

We are entering the era of data as drivetrain, where we use data not just to generate more data (in the form of predictions), but use data to produce actionable outcomes. That is the goal of the Drivetrain Approach. The best way to illustrate this process is with a familiar data product: search engines. Back in 1997, AltaVista was king of the algorithmic search world. While their models were good at finding relevant websites, the answer the user was most interested in was often buried on page 100 of the search results. Then, Google came along and transformed online search by beginning with a simple question: What is the user's main objective in typing in a search query?



The four steps in the Drivetrain Approach.

Google realized that the objective was to show the most relevant search result; for other companies, it might be increasing profit, improving the customer experience, finding the best path for a robot, or balancing the load in a data center. Once we have specified the goal, the second step is to specify what inputs of the system we can control, the *levers* we can pull to influence the final outcome. In Google's case, they could control the ranking of the search results. The third step was to consider what new *data* they would need to produce such a ranking; they realized that the implicit information regarding which pages linked to which other pages could be used for this purpose. Only after these first three steps do we begin thinking about building the predictive *models*. Our objective and available levers, what data we already have and what additional data we will need to collect, determine the models we can build. The models will take both the levers and any uncontrollable variables as their inputs; the outputs from the models can be combined to predict the final state for our objective.

Step 4 of the Drivetrain Approach for Google is now part of tech history: Larry Page and Sergey Brin invented the graph traversal algorithm PageRank and built an engine on top of it that revolutionized search. But you don't have to invent the next PageRank to build a great data product. We will show a systematic approach to step 4 that doesn't require a PhD in computer science.

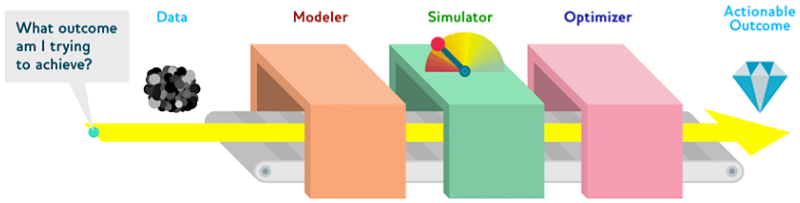
The Model Assembly Line: A Case Study of Optimal Decisions Group

Optimizing for an actionable outcome over the right predictive models can be a company's most important strategic decision. For an insurance company, policy price is the product, so an optimal pricing model is to them what the assembly line is to automobile manufacturing. Insurers have centuries of experience in prediction, but as recently as 10 years ago, the insurance companies often failed to make optimal business decisions about what price to charge each new customer. Their actuaries could build models to predict a customer's likelihood of being in an accident and the expected value of claims. But those models did not solve the pricing problem, so the insurance companies would set a price based on a combination of guesswork and market studies.

This situation changed in 1999 with a company called Optimal Decisions Group (ODG). ODG approached this problem with an early use of the Drivetrain Approach and a practical take on step 4 that can be applied to a wide range of problems. They began by defining the *objective* that the insurance company was trying to achieve: setting a price that maximizes the net-present value of the profit from a new customer over a multi-year time horizon, subject to certain constraints such as maintaining market share. From there, they developed an optimized pricing process that added hundreds of millions of dollars to the insurers' bottom lines. [*Note: Co-author Jeremy Howard founded ODG.*]

ODG identified which *levers* the insurance company could control: what price to charge each customer, what types of accidents to cover, how much to spend on marketing and customer service, and how to react to their competitors' pricing decisions. They also considered inputs outside of their control, like competitors' strategies, macroeconomic conditions, natural disasters, and customer "stickiness." They considered what additional *data* they would need to predict a customer's reaction to changes in price. It was necessary to build this da-

taset by randomly changing the prices of hundreds of thousands of policies over many months. While the insurers were reluctant to conduct these experiments on real customers, as they'd certainly lose some customers as a result, they were swayed by the huge gains that optimized policy pricing might deliver. Finally, ODG started to design the *models* that could be used to optimize the insurer's profit.



Drivetrain Step 4: The Model Assembly Line. Picture a Model Assembly Line for data products that transforms the raw data into an actionable outcome. The Modeler takes the raw data and converts it into slightly more refined predicted data.

The first component of ODG's Modeler was a model of price elasticity (the probability that a customer will accept a given price) for new policies and for renewals. The price elasticity model is a curve of price versus the probability of the customer accepting the policy conditional on that price. This curve moves from almost certain acceptance at very low prices to almost never at high prices.

The second component of ODG's Modeler related price to the insurance company's profit, conditional on the customer accepting this price. The profit for a very low price will be in the red by the value of expected claims in the first year, plus any overhead for acquiring and servicing the new customer. Multiplying these two curves creates a final curve that shows price versus expected profit (see Expected Profit figure, below). The final curve has a clearly identifiable local maximum that represents the best price to charge a customer for the first year.

Expected profit.

ODG also built models for customer retention. These models predicted whether customers would renew their policies in one year, allowing for changes in price and willingness to jump to a competitor. These additional models allow the annual models to be combined to predict profit from a new customer over the next five years.

This new suite of models is not a final answer because it only identifies the outcome for a given set of inputs. The next machine on the assembly line is a *Simulator*, which lets ODG ask the "what if" questions to see how the levers affect the distribution of the final outcome. The expected profit curve is just a slice of the surface of possible outcomes. To build that entire surface, the Simulator runs the models over a wide range of inputs. The operator can adjust the input levers to answer specific questions like, "What will happen if our company offers the customer a low teaser price in year one but then raises the premiums in year two?" They can also explore how the distribution of profit is shaped by the inputs outside of the insurer's control: "What if the economy crashes and the customer loses his job? What if a 100-year flood hits his home? If a new competitor enters the market and our

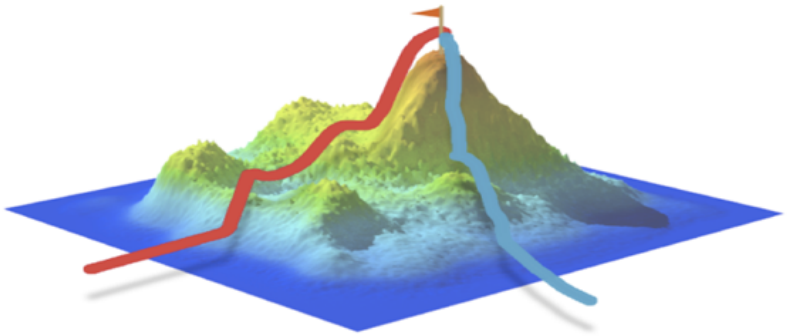
company does not react, what will be the impact on our bottom line?" Because the simulation is at a per-policy level, the insurer can view the impact of a given set of price changes on revenue, market share, and other metrics over time.

The Simulator's result is fed to an *Optimizer*, which takes the surface of possible outcomes and identifies the highest point. The Optimizer not only finds the best outcomes, it can also identify catastrophic outcomes and show how to avoid them. There are many different optimization techniques to choose from (see "Optimization in the Real World" (page 24)), but it is a well-understood field with robust and accessible solutions. ODG's competitors use different techniques to find an optimal price, but they are shipping the same over-all data product. What matters is that using a Drivetrain Approach combined with a Model Assembly Line bridges the gap between predictive models and actionable outcomes. Irfan Ahmed of CloudPhysics provides a good taxonomy of predictive modeling that describes this entire assembly line process:

> When dealing with hundreds or thousands of individual components models to understand the behavior of the full-system, a *search* has to be done. I think of this as a complicated machine (full-system) where the curtain is withdrawn and you get to model each significant part of the machine under controlled experiments and then simulate the interactions. Note here the different levels: models of individual components, tied together in a simulation given a set of inputs, iterated through over different input sets in a search optimizer.

Optimization in the Real World

Optimization is a classic problem that has been studied by Newton and Gauss all the way up to mathematicians and engineers in the present day. Many optimization procedures are iterative; they can be thought of as taking a small step, checking our elevation and then taking another small uphill step until we reach a point from which there is no direction in which we can climb any higher. The danger in this hill-climbing approach is that if the steps are too small, we may get stuck at one of the many local maxima in the foothills, which will not tell us the best set of controllable inputs. There are many techniques to avoid this problem, some based on statistics and spreading our bets widely, and others based on systems seen in nature, like biological evolution or the cooling of atoms in glass.



Optimization is a process we are all familiar with in our daily lives, even if we have never used algorithms like gradient descent or simulated annealing. A great image for optimization in the real world comes up in a recent TechZing podcast with the co-founders of data-mining competition platform Kaggle. One of the authors of this paper was explaining an iterative optimization technique, and the host says, "So, in a sense Jeremy, your approach was like that of doing a startup, which is just get something out there and iterate and iterate and iterate." The takeaway, whether you are a tiny startup or a giant insurance company, is that we unconsciously use optimization whenever we decide how to get to where we want to go.

Drivetrain Approach to Recommender Systems

Let's look at how we could apply this process to another industry: marketing. We begin by applying the Drivetrain Approach to a familiar example, recommendation engines, and then building this up into an entire optimized marketing strategy.

Recommendation engines are a familiar example of a data product based on well-built predictive models that do not achieve an optimal objective. The current algorithms predict what products a customer will *like*, based on purchase history and the histories of similar customers. A company like Amazon represents every purchase that has ever been made as a giant sparse matrix, with customers as the rows and products as the columns. Once they have the data in this format, data scientists apply some form of collaborative filtering to "fill in the matrix." For example, if customer A buys products 1 and 10, and customer B buys products 1, 2, 4, and 10, the engine will recommend that A buy 2 and 4. These models are good at predicting whether a customer will like a given product, but they often suggest products that the cus-

tomer already knows about or has already decided not to buy. Amazon's recommendation engine is probably the best one out there, but it's easy to get it to show its warts. Here is a screenshot of the "Customers Who Bought This Item Also Bought" feed on Amazon from a search for the latest book in Terry Pratchett's "Discworld series:"

All of the recommendations are for other books in the same series, but it's a good assumption that a customer who searched for "Terry Pratchett" is already aware of these books. There may be some unexpected recommendations on pages 2 through 14 of the feed, but how many customers are going to bother clicking through?

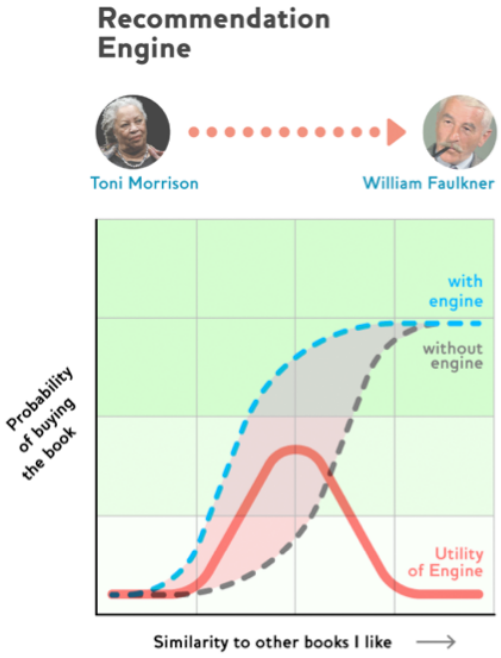
Instead, let's design an improved recommendation engine using the Drivetrain Approach, starting by reconsidering our *objective*. The objective of a recommendation engine is to drive additional sales by surprising and delighting the customer with books he or she would not have purchased *without the recommendation*. What we would really like to do is emulate the experience of Mark Johnson, CEO of Zite, who gave a perfect example of what a customer's recommendation experience should be like in a recent TOC talk. He went into Strand bookstore in New York City and asked for a book similar to Toni Morrison's *Beloved*. The girl behind the counter recommended William Faulkner's *Absolom Absolom*. On Amazon, the top results for a similar query leads to another book by Toni Morrison and several books by well-known female authors of color. The Strand bookseller made a brilliant but far-fetched recommendation probably based more on the character of Morrison's writing than superficial similarities between Morrison and other authors. She cut through the chaff of the obvious to make a recommendation that will send the customer home with a new book, and returning to Strand again and again in the future.

This is not to say that Amazon's recommendation engine could not have made the same connection; the problem is that this helpful recommendation will be buried far down in the recommendation feed, beneath books that have more obvious similarities to *Beloved*. The

objective is to escape a recommendation filter bubble, a term which was originally coined by Eli Pariser to describe the tendency of personalized news feeds to only display articles that are blandly popular or further confirm the readers' existing biases.

As with the AltaVista-Google example, the *lever* a bookseller can control is the ranking of the recommendations. New *data* must also be collected to generate recommendations that will *cause* new sales. This will require conducting many randomized experiments in order to collect data about a wide range of recommendations for a wide range of customers.

The final step in the drivetrain process is to build the *Model Assembly Line*. One way to escape the recommendation bubble would be to build a *Modeler* containing two models for purchase probabilities, conditional on seeing or not seeing a recommendation. The difference between these two probabilities is a utility function for a given recommendation to a customer (see Recommendation Engine figure, below). It will be low in cases where the algorithm recommends a familiar book that the customer has already rejected (both components are small) or a book that he or she would have bought even without the recommendation (both components are large and cancel each other out). We can build a *Simulator* to test the utility of each of the many possible books we have in stock, or perhaps just over all the outputs of a collaborative filtering model of similar customer purchases, and then build a simple *Optimizer* that ranks and displays the recommended books based on their simulated utility. In general, when choosing an objective function to optimize, we need less emphasis on the "function" and more on the "objective." What is the objective of the person using our data product? What choice are we actually helping him or her make?



Recommendation Engine.

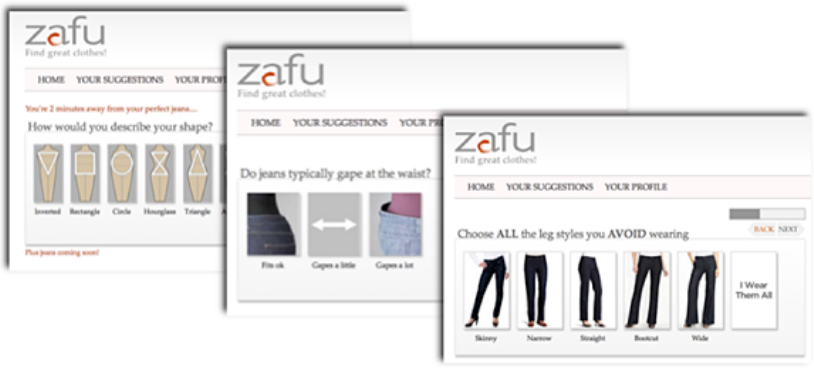
Optimizing Lifetime Customer Value

This same systematic approach can be used to optimize the entire marketing strategy. This encompasses all the interactions that a retailer has with its customers outside of the actual buy-sell transaction, whether making a product recommendation, encouraging the customer to check out a new feature of the online store, or sending sales promotions. Making the wrong choices comes at a cost to the retailer in the form of reduced margins (discounts that do not drive extra sales), opportunity costs for the scarce real-estate on their homepage (taking up space in the recommendation feed with products the customer doesn't like or would have bought without a recommendation) or the customer tuning out (sending so many unhelpful email promotions that the customer filters all future communications as spam). We will show how to go about building an optimized marketing strategy that mitigates these effects.

As in each of the previous examples, we begin by asking: "What *objective* is the marketing strategy trying to achieve?" Simple: we want to optimize the lifetime value from each customer. Second question: "What *levers* do we have at our disposal to achieve this objective?" Quite a few. For example:

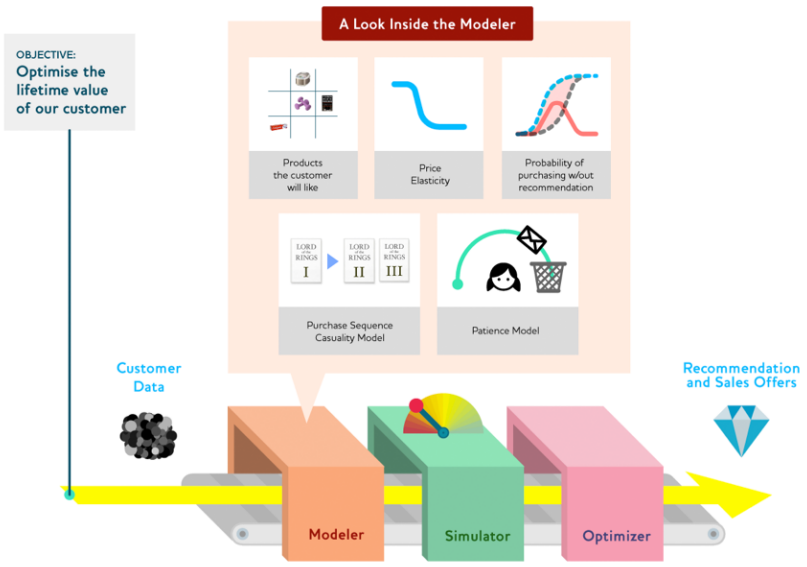
1. We can make product recommendations that surprise and delight (using the optimized recommendation outlined in the previous section).
2. We could offer tailored discounts or special offers on products the customer was not quite ready to buy or would have bought elsewhere.
3. We can even make customer-care calls just to see how the user is enjoying our site and make them feel that their feedback is valued.

What new *data* do we need to collect? This can vary case by case, but a few online retailers are taking creative approaches to this step. Online fashion retailer Zafu shows how to encourage the customer to participate in this collection process. Plenty of websites sell designer denim, but for many women, high-end jeans are the one item of clothing they never buy online because it's hard to find the right pair without trying them on. Zafu's approach is not to send their customers directly to the clothes, but to begin by asking a series of simple questions about the customers' body type, how well their other jeans fit, and their fashion preferences. Only then does the customer get to browse a recommended selection of Zafu's inventory. The data collection and recommendation steps are not an add-on; they are Zafu's entire business model — women's jeans are now a data product. Zafu can tailor their recommendations to fit as well as their jeans because their system is asking the right questions.



Starting with the objective forces data scientists to consider what additional models they need to build for the *Modeler*. We can keep the "like" model that we have already built as well as the causality model for purchases with and without recommendations, and then take a staged approach to adding additional models that we think will improve the marketing effectiveness. We could add a price elasticity model to test how offering a discount might change the probability that the customer will buy the item. We could construct a patience model for the customers' tolerance for poorly targeted communications: When do they tune them out and filter our messages straight to spam? ("If Hulu shows me that same dog food ad one more time, I'm gonna stop watching!") A purchase sequence causality model can be used to identify key "entry products." For example, a pair of jeans that is often paired with a particular top, or the first part of a series of novels that often leads to a sale of the whole set.

Once we have these models, we construct a *Simulator* and an *Optimizer* and run them over the combined models to find out what recommendations will achieve our objectives: driving sales and improving the customer experience.



A look inside the Modeler.

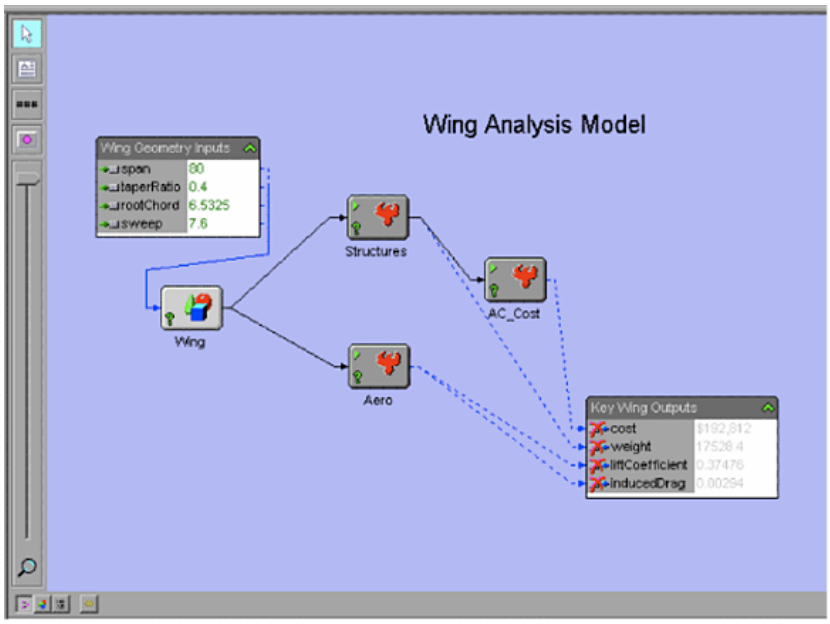
Best Practices from Physical Data Products

It is easy to stumble into the trap of thinking that since data exists somewhere abstract, on a spreadsheet or in the cloud, that data products are just abstract algorithms. So, we would like to conclude by showing you how objective-based data products are already a part of the tangible world. What is most important about these examples is that the engineers who designed these data products didn't start by building a neato robot and then looking for something to do with it. They started with an objective like, "I want my car to drive me places," and then designed a covert data product to accomplish that task. Engineers are often quietly on the leading edge of algorithmic applications because they have long been thinking about their own modeling challenges in an objective-based way. Industrial engineers were among the first to begin using neural networks, applying them to problems like the optimal design of assembly lines and quality control. Brian Ripley's seminal book on pattern recognition gives credit for many ideas and techniques to largely forgotten engineering papers from the 1970s.

When designing a product or manufacturing process, a drivetrain-like process followed by model integration, simulation and optimization is a familiar part of the toolkit of systems engineers. In engineering, it

is often necessary to link many component models together so that they can be simulated and optimized in tandem. These firms have plenty of experience building models of each of the components and systems in their final product, whether they're building a server farm or a fighter jet. There may be one detailed model for mechanical systems, a separate model for thermal systems, and yet another for electrical systems, etc. All of these systems have critical interactions. For example, resistance in the electrical system produces heat, which needs to be included as an input for the thermal diffusion and cooling model. That excess heat could cause mechanical components to warp, producing stresses that should be inputs to the mechanical models.

The screenshot below is taken from a model integration tool designed by Phoenix Integration. Although it's from a completely different engineering discipline, this diagram is very similar to the Drivetrain Approach we've recommended for data products. The *objective* is clearly defined: build an airplane wing. The wing box includes the design *levers* like span, taper ratio, and sweep. The *data* is in the wing materials' physical properties; costs are listed in another tab of the application. There is a *Modeler* for aerodynamics and mechanical structure that can then be fed to a *Simulator* to produce the Key Wing Outputs of cost, weight, lift coefficient, and induced drag. These outcomes can be fed to an *Optimizer* to build a functioning and cost-effective airplane wing.



Screenshot from a model integration tool designed by Phoenix Integration.

As predictive modeling and optimization become more vital to a wide variety of activities, look out for the engineers to disrupt industries that wouldn't immediately appear to be in the data business. The inspiration for the phrase "Drivetrain Approach," for example, is already on the streets of Mountain View. Instead of being data driven, we can now let the data drive us.

Suppose we wanted to get from San Francisco to the Strata 2012 Conference in Santa Clara. We could just build a simple model of distance / speed-limit to predict arrival time with little more than a ruler and a road map. If we want a more sophisticated system, we can build another model for traffic congestion and yet another model to forecast weather conditions and their effect on the safest maximum speed. There are plenty of cool challenges in building these models, but by themselves, they do not take us to our destination. These days, it is trivial to use some type of heuristic search algorithm to predict the drive times along various routes (a *Simulator*) and then pick the shortest one (an *Optimizer*) subject to constraints like avoiding bridge tolls or maximizing gas mileage. But why not think bigger? Instead of the femme-bot voice of the GPS unit telling us which route to take and where to turn, what would it take to build a car that would make those decisions by itself? Why not bundle simulation and optimization engines with a physical engine, all inside the black box of a car?

Let's consider how this is an application of the Drivetrain Approach. We have already defined our *objective*: building a car that drives itself. The levers are the vehicle controls we are all familiar with: steering wheel, accelerator, brakes, etc. Next, we consider what *data* the car needs to collect; it needs sensors that gather data about the road as well as cameras that can detect road signs, red or green lights, and unexpected obstacles (including pedestrians). We need to define the *models* we will need, such as physics models to predict the effects of steering, braking and acceleration, and pattern recognition algorithms to interpret data from the road signs.

As one engineer on the Google self-driving car project put it in a recent *Wired* article, "We're analyzing and predicting the world 20 times a second." What gets lost in the quote is what happens as a result of that prediction. The vehicle needs to use a simulator to examine the results of the possible actions it could take. If it turns left now, will it hit that

pedestrian? If it makes a right turn at 55 mph in these weather conditions, will it skid off the road? Merely predicting what will happen isn't good enough. The self-driving car needs to take the next step: after *simulating* all the possibilities, it must *optimize* the results of the simulation to pick the best combination of acceleration and braking, steering and signaling, to get us safely to Santa Clara. Prediction only tells us that there is going to be an accident. An optimizer tells us how to avoid accidents.

Improving the data collection and predictive models is very important, but we want to emphasize the importance of beginning by defining a clear objective with levers that produce actionable outcomes. Data science is beginning to pervade even the most bricks-and-mortar elements of our lives. As scientists and engineers become more adept at applying prediction and optimization to everyday problems, they are expanding the art of the possible, optimizing everything from our personal health to the houses and cities we live in. Models developed to simulate fluid dynamics and turbulence have been applied to improving traffic and pedestrian flows by using the placement of exits and crowd control barriers as levers. This has improved emergency evacuation procedures for subway stations and reduced the danger of crowd stampedes and trampling during sporting events. Nest is designing smart thermostats that learn the home-owner's temperature preferences and then optimizes their energy consumption. For motor vehicle traffic, IBM performed a project with the city of Stockholm to optimize traffic flows that reduced congestion by nearly a quarter, and increased the air quality in the inner city by 25%. What is particularly interesting is that there was no need to build an elaborate new data collection system. Any city with metered stoplights already has all the necessary information; they just haven't found a way to suck the meaning out of it.

In another area where objective-based data products have the power to change lives, the CMU extension in Silicon Valley has an active project for building data products to help first responders after natural or man-made disasters. Jeannie Stamberger of Carnegie Mellon University Silicon Valley explained to us many of the possible applications of predictive algorithms to disaster response, from text-mining and sentiment analysis of tweets to determine the extent of the damage, to swarms of autonomous robots for reconnaissance and rescue, to logistic optimization tools that help multiple jurisdictions coordinate their responses. These disaster applications are a particularly good

example of why data products need simple, well-designed interfaces that produce concrete recommendations. In an emergency, a data product that just produces more data is of little use. Data scientists now have the predictive tools to build products that increase the common good, but they need to be aware that building the models is not enough if they do not also produce optimized, implementable outcomes.

The Future for Data Products

We introduced the Drivetrain Approach to provide a framework for designing the next generation of great data products and described how it relies at its heart on optimization. In the future, we hope to see optimization taught in business schools as well as in statistics departments. We hope to see data scientists ship products that are designed to produce desirable business outcomes. This is still the dawn of data science. We don't know what design approaches will be developed in the future, but right now, there is a need for the data science community to coalesce around a shared vocabulary and product design process that can be used to educate others on how to derive value from their predictive models. If we do not do this, we will find that our models only use data to create more data, rather than using data to create actions, disrupt industries, and transform lives.

Do we want products that deliver data, or do we want products that deliver results based on data? Jeremy Howard examined these questions in his Strata California 12 session, "From Predictive Modelling to Optimization: The Next Frontier." Full video from that session is available here.

What It Takes to Build Great Machine Learning Products

By Aria Haghighi

Machine learning (ML) is all the rage, riding tight on the coattails of the "big data" wave. Like most technology hype, the enthusiasm far exceeds the realization of actual products. Arguably, not since Google's tremendous innovations in the late '90s/early 2000s has algorithmic technology led to a product that has permeated the popular culture. That's not to say there haven't been great ML wins since, but none have as been as impactful or had computational algorithms at their core.

Netflix may use recommendation technology, but Netflix is still Netflix without it. There would be no Google if Page, Brin, et al., hadn't exploited the graph structure of the Web and anchor text to improve search.

So why is this? It's not for lack of trying. How many startups have aimed to bring natural language processing (NLP) technology to the masses, only to fade into oblivion after people actually *try* their products? The challenge in building great products with ML lies not in just understanding basic ML theory, but in understanding the domain and problem sufficiently to operationalize intuitions into model design. Interesting problems don't have simple off-the-shelf ML solutions. Progress in important ML application areas, like NLP, come from insights specific to these problems, rather than generic ML machinery. Often, specific insights into a problem and careful model design make the difference between a system that doesn't work at all and one that people will actually use.

The goal of this essay is not to discourage people from building amazing products with ML at their cores, but to be clear about where I think the difficulty lies.

Progress in Machine Learning

Machine learning has come a long way over the last decade. Before I started grad school, training a large-margin classifier (e.g., SVM) was done via John Platt's batch SMO algorithm. In that case, training time scaled poorly with the amount of training data. Writing the algorithm itself required understanding quadratic programming and was riddled with heuristics for selecting active constraints and black-art parameter tuning. Now, we know how to train a nearly performance-equivalent large-margin classifier in linear time using a (relatively) simple online algorithm (PDF). Similar strides have been made in (probabilistic) graphical models: Markov-chain Monte Carlo (MCMC) and variational methods have facilitated inference for arbitrarily complex graphical models.[1] Anecdotally, take at look at papers over the last

1. Although MCMC is a much older statistical technique, its broad use in large-scale machine learning applications is relatively recent.

eight years in the proceedings of the Association for Computational Linguistics (ACL), the premiere natural language processing publication. A top paper from 2011 has orders of magnitude more technical ML sophistication than one from 2003.

On the education front, we've come a long way as well. As an undergrad at Stanford in the early-to-mid 2000s, I took Andrew Ng's ML course and Daphne Koller's probabilistic graphical model course. Both of these classes were among the best I took at Stanford and were only available to about 100 students a year. Koller's course in particular was not only the best course I took at Stanford, but the one that taught me the most about teaching. Now, anyone can take these courses online.

As an applied ML person — specifically, natural language processing — much of this progress has made aspects of research significantly easier. However, the core decisions I make are not which abstract ML algorithm, loss-function, or objective to use, but what features and structure are relevant to solving my problem. This skill only comes with practice. So, while it's great that a much wider audience will have an understanding of basic ML, it's not the most difficult part of building intelligent systems.

Interesting Problems Are Never Off the Shelf

The interesting problems that you'd actually want to solve are far messier than the abstractions used to describe standard ML problems. Take machine translation (MT), for example. Naively, MT looks like a statistical classification problem: You get an input foreign sentence and have to predict a target English sentence. Unfortunately, because the space of possible English is combinatorially large, you can't treat MT as a black-box classification problem. Instead, like most interesting ML applications, MT problems have a lot of structure and part of the job of a good researcher is decomposing the problem into smaller pieces that can be learned or encoded deterministically. My claim is that progress in complex problems like MT comes mostly from how we decompose and structure the solution space, rather than ML techniques used to learn within this space.

Machine translation has improved by leaps and bounds throughout the last decade. I think this progress has largely, but not entirely, come from keen insights into the specific problem, rather than generic ML improvements. Modern statistical MT originates from an amazing paper, "The mathematics of statistical machine translation" (PDF),

which introduced the noisy-channel architecture on which future MT systems would be based. At a very simplistic level, this is how the model works:[2] For each foreign word, there are potential English translations (including the null word for foreign words that have no English equivalent). Think of this as a probabilistic dictionary. These candidate translation words are then re-ordered to create a plausible English translation. There are many intricacies being glossed over: how to efficiently consider candidate English sentences and their permutations, what model is used to learn the systematic ways in which reordering occurs between languages, and the details about how to score the plausibility of the English candidate (the language model).

The core improvement in MT came from changing this model. So, rather than learning translation probabilities of individual words, to instead learn models of how to translate foreign *phrases* to English phrases. For instance, the German word "abends" translates roughly to the English prepositional phrase "in the evening." Before phrase-based translation (PDF), a word-based model would only get to translate to a single English word, making it unlikely to arrive at the correct English translation.[3] Phrase-based translation generally results in more accurate translations with fluid, idiomatic English output. Of course, adding phrased-based emissions introduces several additional complexities, including how to how to estimate phrase-emissions given that we never observe phrase segmentation; no one tells us that "in the evening" is a phrase that should match up to some foreign phrase. What's surprising here is that there aren't general ML improvements that are making this difference, but problem-specific model design. People can and have implemented more sophisticated ML techniques for various pieces of an MT system. And these do yield improvements, but typically far smaller than good problem-specific research insights.

Franz Och, one of the authors of the original Phrase-based papers, went on to Google and became the principle person behind the search company's translation efforts. While the intellectual underpinnings of Google's system go back to Och's days as a research scientist at the Information Sciences Institute (and earlier as a graduate student),

2. The model is generative, so what's being described here is from the point-of-view of inference; the model's generative story works in reverse.

3. IBM model 3 introduced the concept of fertility to allow a given word to generate multiple independent target translation words. While this could generate the required translation, the probability of the model doing so is relatively low.

much of the gains beyond the insights underlying phrase-based translation (and minimum-error rate training, another of Och's innovations) came from a massive software engineering effort to scale these ideas to the Web. That effort itself yielded impressive research into large-scale language models and other areas of NLP. It's important to note that Och, in addition to being a world-class researcher, is also, by all accounts, an incredibly impressive hacker and builder. It's this rare combination of skill that can bring ideas all the way from a research project to where Google Translate is today.

Defining the Problem

But I think there's an even bigger barrier beyond ingenious model design and engineering skills. In the case of machine translation and speech recognition, the problem being solved is straightforward to understand and well-specified. Many of the NLP technologies that I think will revolutionize consumer products over the next decade are much more vague. How, exactly, can we take the excellent research in structured topic models, discourse processing, or sentiment analysis and make a mass-appeal consumer product?

Consider summarization. We all know that in some way, we'll want products that summarize and structure content. However, for computational and research reasons, you need to restrict the scope of this problem to something for which you can build a model, an algorithm, and ultimately evaluate. For instance, in the summarization literature, the problem of multi-document summarization is typically formulated as selecting a subset of sentences from the document collection and ordering them. Is this the right problem to be solving? Is the best way to summarize a piece of text a handful of full-length sentences? Even if a summarization is accurate, does the Franken-sentence structure yield summaries that feel inorganic to users?

Or, consider sentiment analysis. Do people really just want a coarse-grained thumbs-up or thumbs-down on a product or event? Or do they want a richer picture of sentiments toward individual aspects of an item (e.g., loved the food, hated the decor)? Do people care about determining sentiment attitudes of individual reviewers/utterances, or producing an accurate assessment of aggregate sentiment?

Typically, these decisions are made by a product person and are passed off to researchers and engineers to implement. The problem with this approach is that ML-core products are intimately constrained by what

is technically and algorithmically feasible. In my experience, having a technical understanding of the range of related ML problems can inspire product ideas that might not occur to someone without this understanding. To draw a loose analogy, it's like architecture. So much of the construction of a bridge is constrained by material resources and physics that it doesn't make sense to have people without that technical background design a bridge.

The goal of all this is to say that if you want to build a rich ML product, you need to have a rich product/design/research/engineering team. All the way from the nitty gritty of how ML theory works to building systems to domain knowledge to higher-level product thinking to technical interaction and graphic design; preferably people who are world-class in one of these areas but also good in several. Small talented teams with all of these skills are better equipped to navigate the joint uncertainty with respect to product vision as well as model design. Large companies that have research and product people in entirely different buildings are ill-equipped to tackle these kinds of problems. The ML products of the future will come from startups with small founding teams that have this full context and can all fit in the proverbial garage.

CHAPTER 4

The Application of Big Data

Stories over Spreadsheets

By Mac Slocum

I didn't realize how much I dislike spreadsheets until I was presented with a vision of the future where their dominance isn't guaranteed.

That eye-opening was offered by Narrative Science CTO Kris Hammond (@whisperspace) during a recent interview. Hammond's company turns data into stories: They provide sentences and paragraphs instead of rows and columns. To date, much of the attention Narrative Science has received has focused on the media applications. That's a natural starting point. Heck, I asked him about those very same things when I first met Hammond at Strata in New York last fall. But during our most recent chat, Hammond explored the other applications of narrative-driven data analysis.

"Companies, God bless them, had a great insight: They wanted to make decisions based upon the data that's out there and the evidence in front of them," Hammond said. "So they started gathering that data up. It quickly exploded. And they ended up with huge data repositories they had to manage. A lot of their effort ended up being focused on gathering that data, managing that data, doing analytics across that data, and then the question was: What do we do with it?"

Hammond sees an opportunity to extract and communicate the insights locked within company data. "We'll be the bridge between the data you have, the insights that are in there, or insights we can gather,

and communicating that information to your clients, to your management, and to your different product teams. We'll turn it into something that's intelligible instead of a list of numbers, a spreadsheet, or a graph or two. You get a real narrative; a real story in that data."

My takeaway: The journalism applications of this are intriguing, but these other use cases are *empowering*.

Why? Because most people don't speak fluent "spreadsheet." They see all those neat rows and columns and charts, and they know something important is tucked in there, but what that something is and how to extract it aren't immediately clear. Spreadsheets require effort. That's doubly true if you don't know what you're looking for. And if data analysis is an adjacent part of a person's job, more effort means those spreadsheets will always be pushed to the side. "I'll get to those next week when I've got more time…"

We all know how that plays out.

But what if the spreadsheet wasn't our default output anymore? What if we could take things most of us are hard-wired to understand — stories, sentences, clear guidance — and layer it over all that vital data? Hammond touched on that:

> For some people, a spreadsheet is a great device. For most people, not so much so. The story. The paragraph. The report. The prediction. The advisory. Those are much more powerful objects in our world, and they're what we're used to.

He's right. Spreadsheets push us (well, most of us) into a cognitive corner. Open a spreadsheet and you're forced to recalibrate your focus to *see* the data. Then you have to work even harder to extract meaning. This is the best we can do?

With that in mind, I asked Hammond if the spreadsheet's days are numbered.

"There will always be someone who uses a spreadsheet," Hammond said. "But, I think what we're finding is that the story is really going to be the endpoint. If you think about it, the spreadsheet is for somebody who really embraces the data. And usually what that person does is they reduce that data down to something that they're going to use to communicate with someone else."

A Thought on Dashboards

I used to view dashboards as the logical step beyond raw data and spreadsheets. I'm not so sure about that anymore, at least in terms of broad adoption. Dashboards are good tools, and I anticipate we'll have them from now until the end of time, but they're still weighed down by a complexity that makes them inaccessible.

It's not that people can't master the buttons and custom reports in dashboards; they simply *don't have time*. These people — and I include myself among them — need something faster and knob-free. Simplicity is the thing that will ultimately democratize data reporting and data insights. That's why the expansion of data analysis requires a refinement beyond our current dashboards. There's a next step that hasn't been addressed.

Does the answer lie in narrative? Will visualizations lead the way? Will a hybrid format take root? I don't know what the final outputs will look like, but the importance of data reporting means someone will eventually crack the problem.

Full Interview

You can see the entire discussion with Hammond in this interview.

Mining the Astronomical Literature

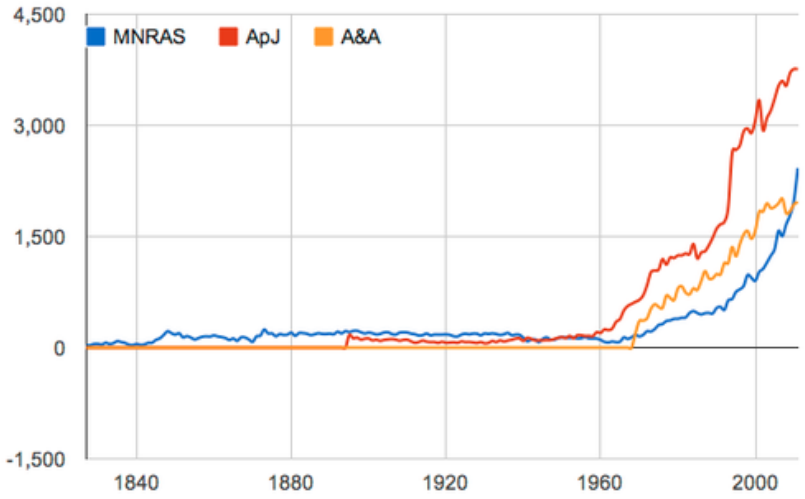
By Alasdair Allan

There is a huge debate right now about making academic literature freely accessible and moving toward open access. But what would be possible if people stopped talking about it and just dug in and got on with it?

NASA's Astrophysics Data System (ADS), hosted by the Smithsonian Astrophysical Observatory (SAO), has quietly been working away since the mid-'90s. Without much, if any, fanfare amongst the other disciplines, it has moved astronomers into a world where access to the literature is just a given. It's something they don't have to think about all that much.

The ADS service provides access to abstracts for virtually all of the astronomical literature. But it also provides access to the full text of

more than half a million papers, going right back to the start of peer-reviewed journals in the 1800s. The service has links to online data archives, along with reference and citation information for each of the papers, and it's all searchable and downloadable.



Number of papers published in the three main astronomy journals each year. Credit: Robert Simpson

The existence of the ADS, along with the arXiv pre-print server, has meant that most astronomers haven't seen the inside of a brick-built library since the late 1990s.

It also makes astronomy almost uniquely well placed for interesting data mining experiments, experiments that hint at what the rest of academia could do if they followed astronomy's lead. The fact that the discipline's literature has been scanned, archived, indexed and catalogued, and placed behind a RESTful API makes it a treasure trove, both for hypothesis generation and sociological research.

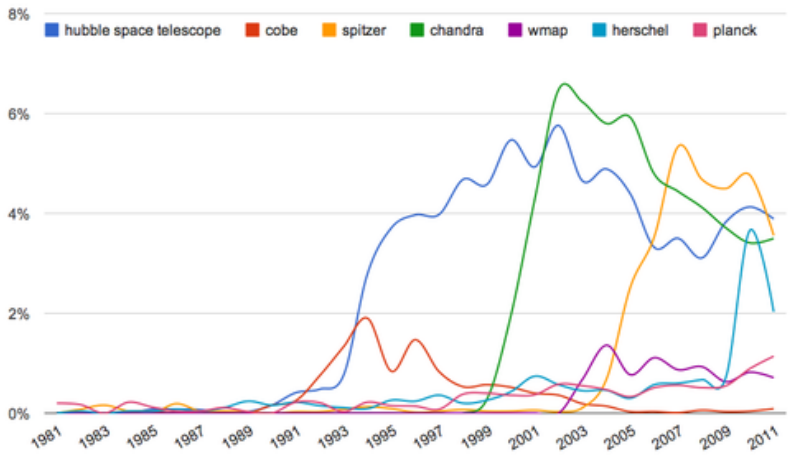
For example, the .Astronomy series of conferences is a small workshop that brings together the best and brightest of the technical community: researchers, developers, educators, and communicators. Billed as "*20% time for astronomers*," it gives these people space to think about how the new technologies affect both how research and communicating research to their peers and to the public is done.

[Disclosure: I'm a member of the advisory board to the .Astronomy conference, and I previously served as a member of the programme organising committee for the conference series.]

It should perhaps come as little surprise that one of the more interesting projects to come out of a hack day held as part of this year's .Astronomy meeting in Heidelberg was work by Robert Simpson, Karen Masters and Sarah Kendrew that focused on data mining the astronomical literature.

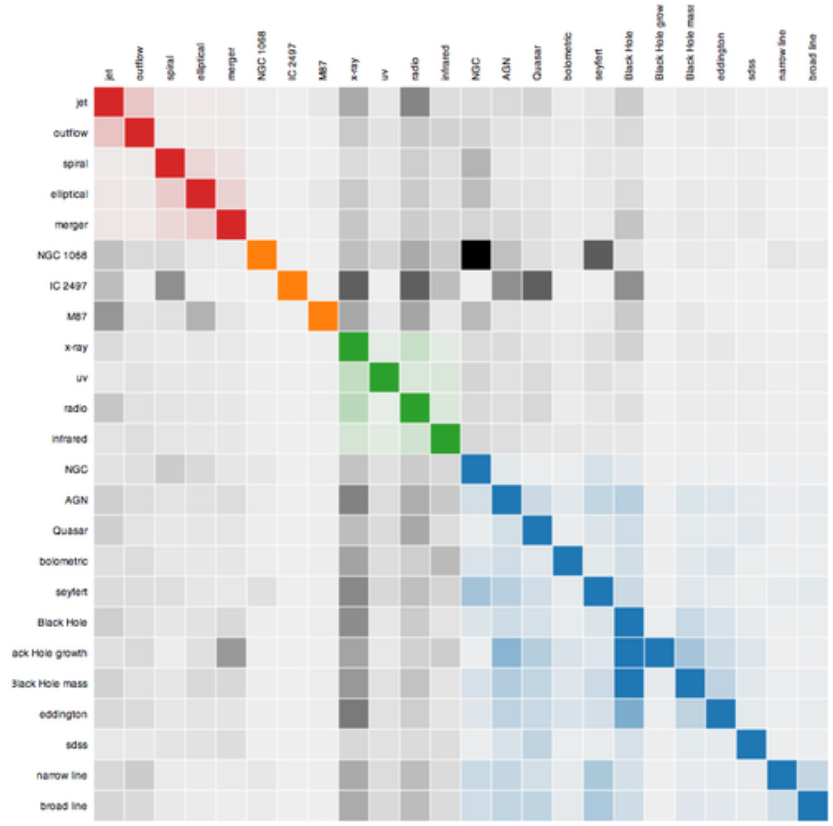
The team grabbed and processed the titles and abstracts of all the papers from the Astrophysical Journal (ApJ), Astronomy & Astrophysics (A&A), and the Monthly Notices of the Royal Astronomical Society (MNRAS) since each of those journals started publication — and that's 1827 in the case of MNRAS.

By the end of the day, they'd found some interesting results showing how various terms have trended over time. The results were similar to what's found in Google Books' Ngram Viewer.



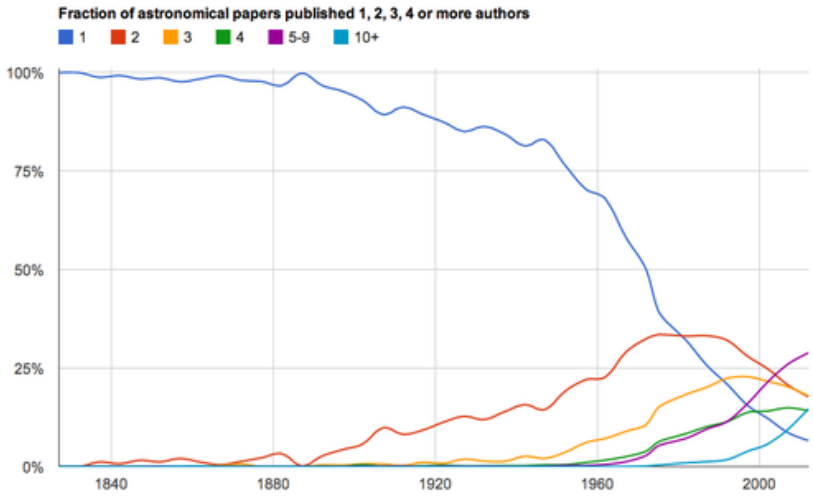
The relative popularity of the names of telescopes in the literature. Hubble, Chandra, and Spitzer seem to have taken turns in hogging the limelight, much as COBE, WMAP, and Planck have each contributed to our knowledge of the cosmic microwave background in successive decades. References to Planck are still on the rise. Credit: Robert Simpson.

After the meeting, however, Robert took his initial results and explored the astronomical literature and his new corpus of data on the literature. He has explored various visualisations of the data, including word matrixes for related terms and for various astro-chemistry.



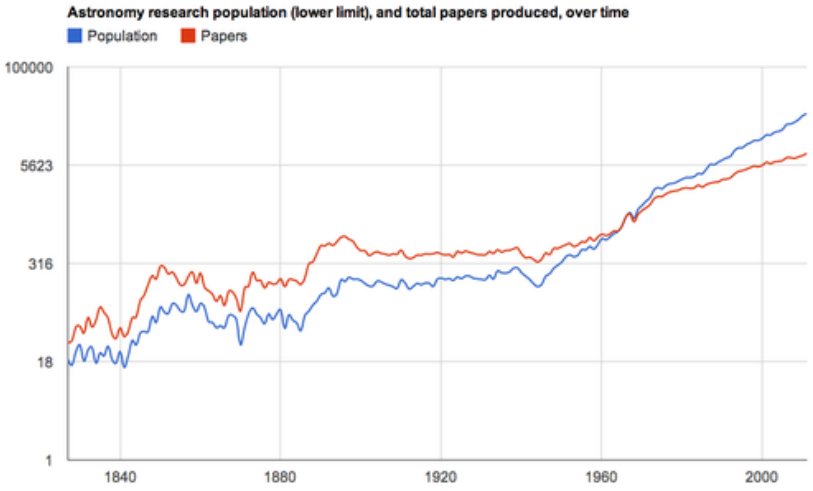
Correlation between terms related to Active Galactic Nuclei (AGN). The opacity of each square represents the strength of the correlation between the terms. Credit: Robert Simpson.

He has also taken a look at authorship in astronomy and is starting to find some interesting trends.



Fraction of astronomical papers published with one, two, three, four, or more authors. Credit: Robert Simpson

You can see that single-author papers dominated for most of the 20th century. Around 1960, we see the decline begin, as two- and three-author papers begin to become a significant chunk of the whole. In 1978, author papers become more prevalent than single-author papers.



Compare the number of "active" research astronomers to the number of papers published each year (across all the major journals). Credit: Robert Simpson.

Here we see that people begin to outpace papers in the 1960s. This may reflect the fact that as we get more technical as a field, and more specialised, it takes more people to write the same number of papers, which is a sort of interesting result all by itself.

Interview with Robert Simpson: Behind the Project and What Lies Ahead

I recently talked with Rob about the work he, Karen Masters, and Sarah Kendrew did at the meeting, and the work he has been doing since with the newly gathered data.

What made you think about data mining the ADS?

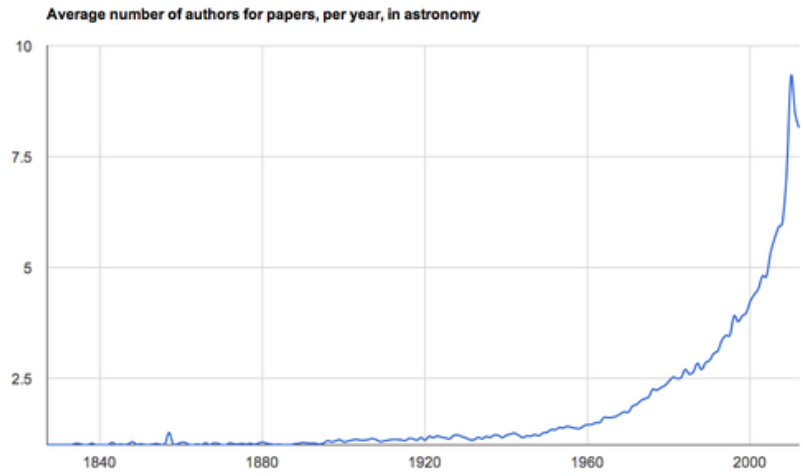
Robert Simpson: At the .Astronomy 4 Hack Day in July, Sarah Kendrew had the idea to try to do an astronomy version of BrainSCANr, a project that generates new hypotheses in the neuroscience literature. I've had a go at mining ADS and arXiv before, so it seemed like a great excuse to dive back in.

Do you think there might be actual science that could be done here?

Robert Simpson: Yes, in the form of finding questions that were unexpected. With such large volumes of peer-reviewed papers being produced daily in astronomy, there is a lot being said. Most researchers can only try to keep up with it all — my daily RSS feed from arXiv is next to useless, it's so bloated. In amongst all that text, there must be connections and relationships that are being missed by the community at large, hidden in the chatter. Maybe we can develop simple techniques to highlight potential missed links, i.e., generate new hypotheses from the mass of words and data.

Are the results coming out of the work useful for auditing academics?

Robert Simpson: Well, perhaps, but that would be tricky territory in my opinion. I've only just begun to explore the data around authorship in astronomy. One thing that is clear is that we can see a big trend toward collaborative work. In 2012, only 6% of papers were single-author efforts, compared with 70+% in the 1950s.

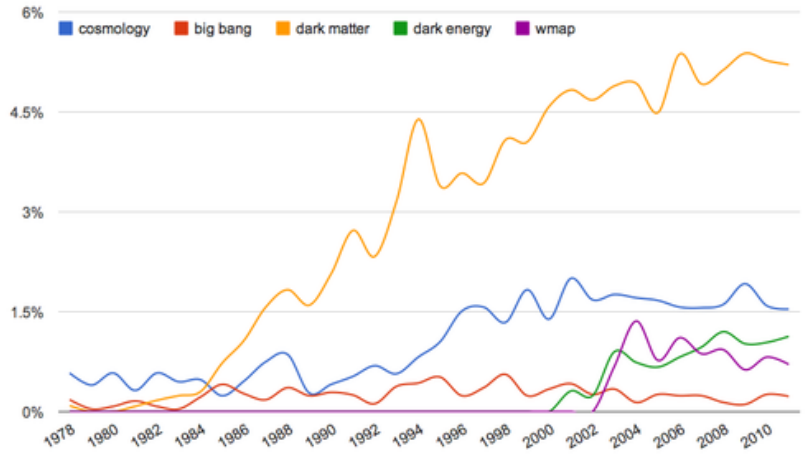


The above plot shows the average number of authors, per paper since 1827. Credit: Robert Simpson.

We can measure how large groups are becoming, and who is part of which groups. In that sense, we can audit research groups, and maybe individual people. The big issue is keeping track of people through variations in their names and affiliations. Identifying authors is probably a solved problem if we look at ORCID.

What about citations? Can you draw any comparisons with h-index data?

Robert Simpson: I haven't looked at h-index stuff specifically, at least not yet, but citations are fun. I looked at the trends surrounding the term *dark matter* and saw something interesting. Mentions of dark matter rise steadily after it first appears in the late '70s.



Compare the term "dark matter" with a few other related terms: "cosmology," "big bang," "dark energy," and "wmap." You can see cosmology has been getting more popular since the 1990s, and dark energy is a recent addition. Credit: Robert Simpson.

In the data, astronomy becomes more and more obsessed with dark matter — the term appears in 1% of all papers by the end of the '80s and 6% today.

Looking at citations changes the picture. The community is writing papers about dark matter more and more each year, but they are getting fewer citations than they used to (the peak for this was in the late '90s). These trends are normalised, so the only regency effect I can think of is that dark matter papers take more than 10 years to become citable. Either that or dark matter studies are currently in a trough for impact.

Can you see where work is dropped by parts of the community and picked up again?

Robert Simpson: Not yet, but I see what you mean. I need to build a better picture of the community and its components.

Can you build a social graph of astronomers out of this data? What about (academic) family trees?

Robert Simpson: Identifying unique authors is my next step, followed by creating fingerprints of individuals at a given point in time. When do people create their first-author papers, when do they have the most impact in their careers, stuff like that.

What tools did you use? In hindsight, would you do it differently?

I'm using Ruby and Perl to grab the data, MySQL to store and query it, JavaScript to display it (Google Charts and D3.js). I may still move the database part to MongoDB because it was designed to store documents. Similarly, I may switch from ADS to arXiv as the data source. Using arXiv would allow me to grab the full text in many cases, even if it does introduce a peer-review issue.

What's next?

Robert Simpson: My aim is still to attempt real hypothesis generation. I've begun the process by investigating correlations between terms in the literature, but I think the power will be in being able to compare all terms with all terms and looking for the unexpected. Terms may correlate indirectly (via a third term, for example), so the entire corpus needs to be processed and optimised to make it work comprehensively.

Science between the Cracks

I'm really looking forward to seeing more results coming out of Robert's work. This sort of analysis hasn't really been possible before. It's showing a lot of promise both from a sociological angle, with the ability to do research into how science is done and how that has changed, but also ultimately as a hypothesis engine — something that can generate new science in and of itself. This is just a hack day experiment. Imagine what could be done if the literature were more open and this sort of analysis could be done across fields?

Right now, a lot of the most interesting science is being done in the cracks between disciplines, but the hardest part of that sort of work is often trying to understand the literature of the discipline that isn't your own. Robert's project offers a lot of hope that this may soon become easier.

The Dark Side of Data

By Mike Loukides

Tom Slee's "Seeing Like a Geek" is a thoughtful article on the dark side of open data. He starts with the story of a Dalit community in India, whose land was transferred to a group of higher cast Mudaliars through bureaucratic manipulation under the guise of standardizing and digitizing property records. While this sounds like a good idea, it gave a wealthier, more powerful group a chance to erase older, tradi-

tional records that hadn't been properly codified. One effect of passing laws requiring standardized, digital data is to marginalize all data that can't be standardized or digitized, and to marginalize the people who don't control the process of standardization.

That's a serious problem. It's sad to see oppression and property theft riding in under the guise of transparency and openness. But the issue isn't open data, but how data is used.

Jesus said "the poor are with you always" not because the poor aren't a legitimate area of concern (only an American fundamentalist would say that), but because they're an intractable problem that won't go away. The poor are going to be the victims of any changes in technology; it isn't surprisingly that the wealthy in India used data to marginalize the land holdings of the poor. In a similar vein, when Europeans came to North America, I imagine they asked the natives "So, you got a deed to all this land?," a narrative that's still being played out with indigenous people around the world.

The issue is how data is used. If the wealthy can manipulate legislators to wipe out generations of records and folk knowledge as "inaccurate," then there's a problem. A group like DataKind could go in and figure out a way to codify that older generation of knowledge. Then at least, if that isn't acceptable to the government, it would be clear that the problem lies in political manipulation, not in the data itself. And note that a government could wipe out generations of "inaccurate records" without any requirement that the new records be open. In years past the monied classes would have just taken what they wanted, with the government's support. The availability of open data gives a plausible pretext, but it's certainly not a prerequisite (nor should it be blamed) for manipulation by the 0.1%.

One can see the opposite happening, too: the recent legislation in North Carolina that you can't use data that shows sea level rise. Open data may be the *only* possible resource against forces that are interested in suppressing science. What we're seeing here is a full-scale retreat from data and what it can teach us: an attempt to push the furniture against the door to prevent the data from getting in and changing the way we act.

The Digital Publishing Landscape

Slee is on shakier ground when he claims that the digitization of books has allowed Amazon to undermine publishers and booksellers. Yes,

there's technological upheaval, and that necessarily drives changes in business models. Business models change; if they didn't, we'd still have the Pony Express and stagecoaches. O'Reilly Media is thriving, in part because we have a viable digital publishing strategy; publishers without a viable digital strategy are failing.

But what about booksellers? The demise of the local bookstore has, in my observation, as much to do with Barnes & Noble superstores (and the now-defunct Borders), as with Amazon, and it played out long before the rise of ebooks.

I live in a town in southern Connecticut, roughly a half-hour's drive from the two nearest B&N outlets. Guilford and Madison, the town immediately to the east, both have thriving independent bookstores. One has a coffeeshop, stages many, many author events (roughly one a day), and runs many other innovative programs (birthday parties, book-of-the-month services, even ebook sales). The other is just a small local bookstore with a good collection and knowledgeable staff. The town to the west lost its bookstore several years ago, possibly before Amazon even existed. Long before the Internet became a factor, it had reduced itself to cheap gift items and soft porn magazines. So: data may threaten middlemen, though it's not at all clear to me that middlemen can't respond competitively. Or that they are really threatened by "data," as opposed to large centralized competitors.

There are also countervailing benefits. With ebooks, access is democratized. Anyone, anywhere has access to what used to be available only in limited, mostly privileged locations. At O'Reilly, we now sell ebooks in countries we were never able to reach in print. Our print sales overseas never exceeded 30% of our sales; for ebooks, overseas represents more than half the total, with customers as far away as Azerbaijan.

Slee also points to the music labels as an industry that has been marginalized by open data. I really refuse to listen to whining about all the money that the music labels are losing. We've had too many years of crap product generated by marketing people who only care about finding the next Justin Bieber to take the "creative industry" and its sycophants seriously.

Privacy by Design

Data inevitably brings privacy issues into play. As Slee points out (and as Jeff Jonas has before him), apparently insignificant pieces of data

can be put together to form a surprisingly accurate picture of who you are, a picture that can be sold. It's useless to pretend that there won't be increased surveillance in any forseeable future, or that there won't be an increase in targeted advertising (which is, technically, much the same thing).

We can bemoan that shift, celebrate it, or try to subvert it, but we can't pretend that it isn't happening. We shouldn't even pretend that it's new, or that it has anything to do with openness. What is a credit bureau if not an organization that buys and sells data about your financial history, with no pretense of openness?

Jonas's concept of "privacy by design" is an important attempt to address privacy issues in big data. Jonas envisions a day when "I have more privacy features than you" is a marketing advantage. It's certainly a claim I'd like to see Facebook make.

Absent a solution like Jonas's, data is going to be collected, bought, sold, and used for marketing and other purposes, whether it is "open" or not. I do not think we can get to Jonas's world, where privacy is something consumers demand, without going through a stage where data is open and public. It's too easy to live with the illusion of privacy that thrives in a closed world.

I agree that the notion that "open data" is an unalloyed public good is mistaken, and Tom Slee has done a good job of pointing that out. It underscores the importance of a still-nascent ethical consensus about how to use data, along with the importance of data watchdogs, DataKind, and other organizations devoted to the public good. (I don't understand why he argues that Apple and Amazon "undermine community activism"; that seems wrong, particularly in the light of Apple's re-joining the EPEAT green certification system for their products after a net-driven consumer protest.) Data collection is going to happen whether we like it or not, and whether it's open or not. I am convinced that private data is a public bad, and I'm less afraid of data that's open. That doesn't make it necessarily a good; that depends on how the data is used, and the people who are using it.

CHAPTER 5

What to Watch for in Big Data

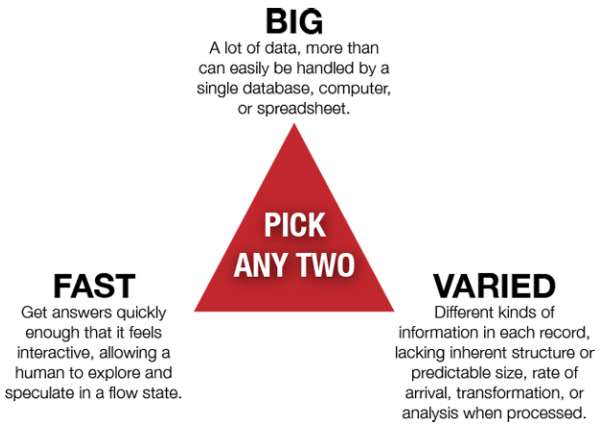
Big Data Is Our Generation's Civil Rights Issue, and We Don't Know It

By Alistair Croll

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.



I 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.

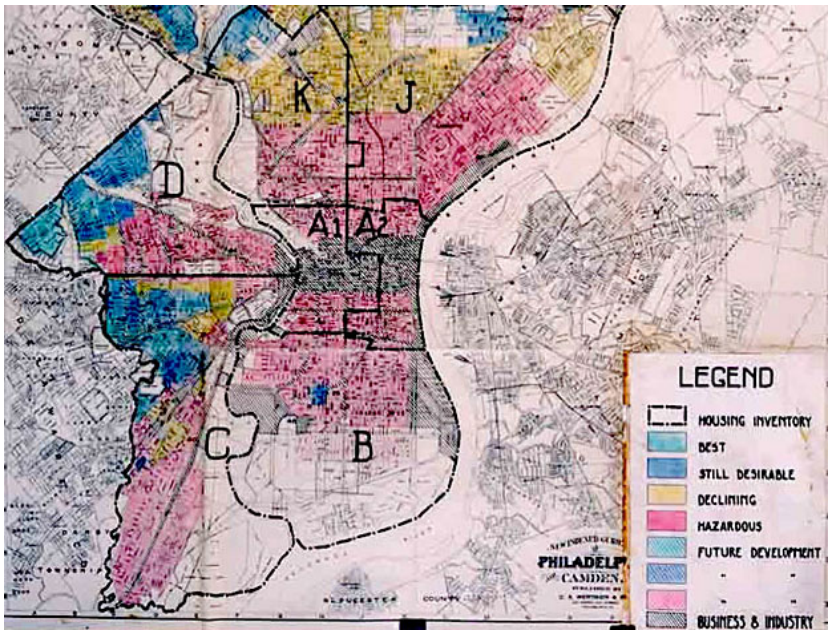
That needs repeating:

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 has 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. Credit: Wikipedia

"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

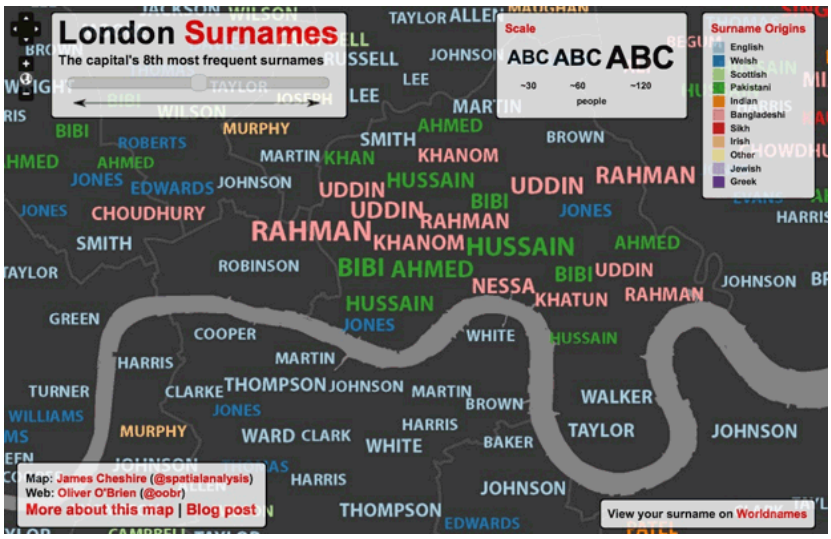
(and the guys who wrote it didn't include everything they learned — some of it was a bit too controversial). They simply looked at the words one group used which others didn't often use. The result was a list of "trigger" words for a particular race or gender.

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.

Three Kinds of Big Data

By Alistair Croll

In the past couple of years, marketers and pundits have spent a lot of time labeling everything "big data." The reasoning goes something like this:

- Everything is on the Internet.
- The Internet has a lot of data.
- Therefore, everything is big data.

When you have a hammer, everything looks like a nail. When you have a Hadoop deployment, everything looks like big data. And if you're trying to cloak your company in the mantle of a burgeoning industry, big data will do just fine. But seeing big data everywhere is a sure way to hasten the inevitable fall from the peak of high expectations to the trough of disillusionment.

We saw this with cloud computing. From early idealists saying everything would live in a magical, limitless, free data center to today's pragmatism about virtualization and infrastructure, we soon took off our rose-colored glasses and put on welding goggles so we could actually build stuff.

So where will big data go to grow up?

Once we get over ourselves and start rolling up our sleeves, I think big data will fall into three major buckets: Enterprise BI, Civil Engineering, and Customer Relationship Optimization. This is where we'll see most IT spending, most government oversight, and most early adoption in the next few years.

Enterprise BI 2.0

For decades, analysts have relied on business intelligence (BI) products like Hyperion, Microstrategy and Cognos to crunch large amounts of information and generate reports. Data warehouses and BI tools are great at answering the same question — such as "what were Mary's sales this quarter?" — over and over again. But they've been less good

at the exploratory, what-if, unpredictable questions that matter for planning and decision-making because that kind of fast exploration of unstructured data is traditionally hard to do and therefore expensive.

Most "legacy" BI tools are constrained in two ways:

- First, they've been schema-then-capture tools in which the analyst decides what to collect, then later captures that data for analysis.
- Second, they've typically focused on reporting what Avinash Kaushik (channeling Donald Rumsfeld) refers to as "known unknowns" — things we know we don't know, and generate reports for.

These tools are used for reporting and operational purposes, and are usually focused on controlling costs, executing against an existing plan, and reporting on how things are going.

As my Strata co-chair Edd Dumbill pointed out when I asked for thoughts on this piece:

> The predominant functional application of big data technologies today is in ETL (Extract, Transform, and Load). I've heard the figure that it's about 80% of Hadoop applications. Just the real grunt work of log file or sensor processing before loading into an analytic database like Vertica.

The availability of cheap, fast computers and storage, as well as open source tools, have made it okay to capture first and ask questions later. That changes how we use data because it lets analysts speculate beyond the initial question that triggered the collection of data.

What's more, the speed with which we can get results — sometimes as fast as a human can ask them — makes data easier to explore interactively. This combination of interactivity and speculation takes BI into the realm of "unknown unknowns," the insights that can produce a competitive advantage or an out-of-the-box differentiator.

Cloud computing underwent a transition from promise to compromise. First, big, public clouds wooed green-field startups. Then, a few years later, incumbent IT vendors introduced private cloud offerings. These private clouds included only a fraction of the benefits of their public cousins — but were nevertheless a sufficient blend of smoke,

mirrors, and features to delay the inevitable move to public resources by a few years and appease the business. For better or worse, that's where most IT cloud budgets are being spent today according to IDC, Gartner, and others. Big data adoption will undergo a similar cycle.

In the next few years, then, look for acquisitions and product introductions — and not a little vaporware — as BI vendors that enterprises trust bring them "big data lite": enough innovation and disruption to satisfy the CEO's golf buddies, but not so much that enterprise IT's jobs are threatened. This, after all, is how change comes to big organizations.

Ultimately, we'll see traditional "known unknowns" BI reporting living alongside big-data-powered data import and cleanup, and fast, exploratory data "unknown unknown" interactivity.

Civil Engineering

The second use of big data is in society and government. Already, data mining can be used to predict disease outbreaks, understand traffic patterns, and improve education.

Cities are facing budget crunches, infrastructure problems, and crowding from rural citizens. Solving these problems is urgent, and cities are perfect labs for big data initiatives. Take a metropolis like New York: hackathons; open feeds of public data; and a population that generates a flood of information as it shops, commutes, gets sick, eats, and just goes about its daily life.

I think municipal data is one of the big three for several reasons: it's a **good tie breaker for partisanship**, we have **new interfaces everyone can understand**, and we finally have a **mostly-connected citizenry**.

In an era of partisan bickering, hard numbers can settle the debate. So, they're not just good government; they're good politics. Expect to see big data applied to social issues, helping us to make funding more effective and scarce government resources more efficient (perhaps to the chagrin of some public servants and lobbyists). As this works in the world's biggest cities, it'll spread to smaller ones, to states, and to municipalities.

Making data accessible to citizens is possible, too: Siri and Google Now show the potential for personalized agents; Narrative Science takes complex data and turns it into words the masses can consume easily; Watson and Wolfram Alpha can give smart answers, either through curated reasoning or making smart guesses.

For the first time, we have a connected citizenry armed (for the most part) with smartphones. Nielsen estimated that smartphones would overtake feature phones in 2011, and that concentration is high in urban cores. The App Store is full of apps for bus schedules, commuters, local events, and other tools that can quickly become how governments connect with their citizens and manage their bureaucracies.

The consequence of all this, of course, is more data. Once governments go digital, their interactions with citizens can be easily instrumented and analyzed for waste or efficiency. That's sure to provoke resistance from those who don't like the scrutiny or accountability, but it's a side effect of digitization: every industry that goes digital gets analyzed and optimized, whether it likes it or not.

Customer Relationship Optimization

The final home of applied big data is marketing. More specifically, it's improving the relationship with consumers so companies can, as Sergio Zyman once said, sell them more stuff, more often, for more money, more efficiently.

The biggest data systems today are focused on web analytics, ad optimization, and the like. Many of today's most popular architectures were weaned on ads and marketing, and have their ancestry in direct marketing plans. They're just more focused than the comparatively blunt instruments with which direct marketers used to work.

The number of contact points in a company has multiplied significantly. Where once there was a phone number and a mailing address, today there are web pages, social media accounts, and more. Tracking users across all these channels — and turning every click, like, share, friend, or retweet into the start of a long funnel that leads, inexorably, to revenue is a big challenge. It's also one that companies like Salesforce understand, with its investments in chat, social media monitoring, cobrowsing, and more.

This is what's lately been referred to as the "360-degree customer view" (though it's not clear that companies will actually act on customer data if they have it, or whether doing so will become a compliance minefield). Big data is already intricately linked to online marketing, but it will branch out in two ways.

First, it'll go from online to offline. Near-field-equipped smartphones with ambient check-in are a marketer's wet dream, and they're coming to pockets everywhere. It'll be possible to track queue lengths, store traffic, and more, giving retailers fresh insights into their brick-and-mortar sales. Ultimately, companies will bring the optimization that online retail has enjoyed to an offline world as consumers become trackable.

Second, it'll go from Wall Street (or maybe that's Madison Avenue and Middlefield Road) to Main Street. Tools will get easier to use, and while small businesses might not have a BI platform, they'll have a tablet or a smartphone that they can bring to their places of business. Mobile payment players like Square are already making them reconsider the checkout process. Adding portable customer intelligence to the tool suite of local companies will broaden how we use marketing tools.

Headlong into the Trough

That's my bet for the next three years, given the molasses of market confusion, vendor promises, and unrealistic expectations we're about to contend with. Will big data change the world? Absolutely. Will it be able to defy the usual cycle of earnest adoption, crushing disappointment, and eventual rebirth all technologies must travel? Certainly not.

Automated Science, Deep Data, and the Paradox of Information

By Bradley Voytek

A lot of great pieces have been written about the relatively recent surge in interest in big data and data science, but in this piece I want to address the importance of deep data analysis: what we can learn from the statistical outliers by drilling down and asking, "What's different here? What's special about these outliers, and what do they tell us about our models and assumptions?"

The reason that big data proponents are so excited about the burgeoning data revolution isn't just because of the math. Don't get me wrong, the math is fun, but we're *excited* because we can begin to distill patterns that were *previously invisible* to us due to a lack of information.

That's big data.

Of course, data are just a collection of facts; bits of information that are only given context — assigned meaning and importance — by human minds. It's not until we *do something* with the data that any of it matters. You can have the best machine learning algorithms, the tightest statistics, and the smartest people working on them, but none of that means anything until someone makes a story out of the results.

And therein lies the rub.

Do all these data tell us a story about ourselves and the universe in which we live, or are we simply hallucinating patterns that we want to see?

(Semi)Automated Science

In 2010, Cornell researchers Michael Schmidt and Hod Lipson published a groundbreaking paper in *Science* titled "Distilling Free-Form Natural Laws from Experimental Data." The premise was simple, and it essentially boiled down to the question, "can we algorithmically extract models to fit our data?"

So they hooked up a double pendulum — a seemingly chaotic system whose movements are governed by classical mechanics — and trained a machine learning algorithm on the motion data.

Their results were astounding. (See them here.)

In a matter of minutes the algorithm converged on Newton's second law of motion: $f = ma$. What took humanity tens of thousands of years to accomplish was completed on 32-cores in essentially no time at all.

In 2011, some neuroscience colleagues of mine, lead by Tal Yarkoni, published a paper in *Nature Methods* titled "Large-scale automated synthesis of human functional neuroimaging data." In this paper, the authors sought to extract patterns from the overwhelming flood of brain imaging research.

To do this, they algorithmically extracted the 3D coordinates of significant brain activations from thousands of neuroimaging studies,

along with words that frequently appeared in each study. Using these two pieces of data along with some simple (but clever) mathematical tools, they were able to create probabilistic maps of brain activation for any given term.

In other words, you type in a word such as "learning" on their website search and visualization tool, NeuroSynth, and they give you back a pattern of brain activity that you should *expect* to see during a learning task.

But that's not all. Given a pattern of brain activation, the system can perform a reverse inference, asking, "given the data that I'm observing, what is the most probable behavioral state that this brain is in?"

Similarly, in late 2010, my wife (Jessica Voytek) and I undertook a project to algorithmically discover associations between concepts in the peer-reviewed neuroscience literature. As a neuroscientist, the goal of my research is to understand relationships between the human brain, behavior, physiology, and disease. Unfortunately, the facts that tie all that information together are locked away in more than 21 million static peer-reviewed scientific publications.

How many undergrads would I need to hire to read through that many papers? Any volunteers?

Even more mind-boggling, each year more than 30,000 neuroscientists attend the annual Society for Neuroscience conference. If we assume that only two-thirds of those people actually *do* research, and if we assume that they only work a meager (for the sciences) 40 hours a week, that's around 40 *million* person-hours dedicated to but one branch of the sciences.

Annually.

This means that in the 10 years I've been attending that conference, more than 400 million person-hours have gone toward the pursuit of understanding the brain. Humanity built the pyramids in 30 years. The Apollo Project got us to the moon in about eight.

So my wife and I said to ourselves, "there has to be a better way."

Which lead us to create brainSCANr, a simple (simplistic?) tool (currently itself under peer review) that makes the assumption that the more often two concepts appear together in the titles or abstracts of published papers, the more likely they are to be associated with one another.

For example, if 10,000 papers mention "Alzheimer's disease" that *also* mention "dementia," then Alzheimer's disease is probably related to dementia. In fact, there are 17,087 papers that mention Alzheimer's *and* dementia, whereas there are only 14 papers that mention Alzheimer's and, for example, creativity.

From this, we built what we're calling the "cognome," a mapping between brain structure, function, and disease.

Big data, data mining, and machine learning are becoming critical tools in the modern scientific arsenal. Examples abound: text mining recipes to find cultural food taste preferences, analyzing cultural trends via word use in books ("culturomics"), identifying seasonality of mood from tweets, and so on.

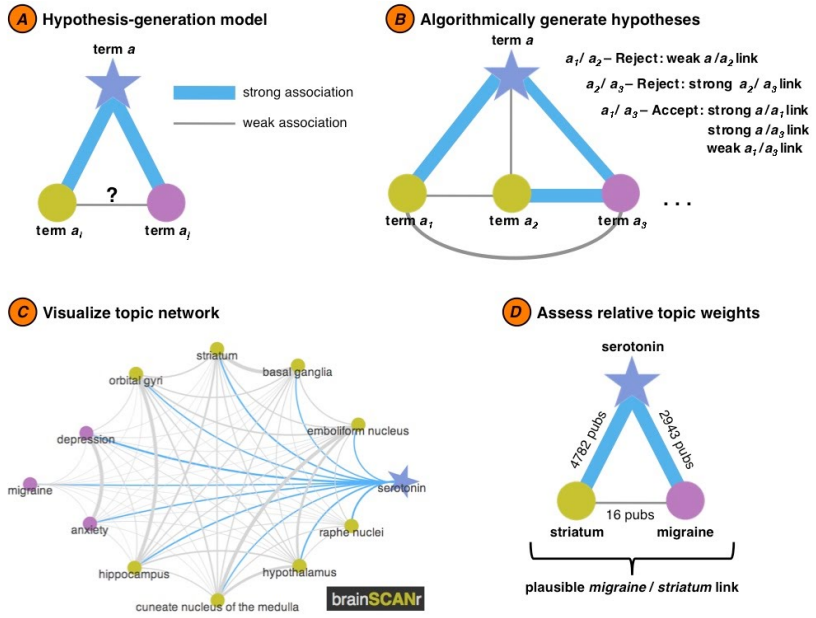
But so what?

Deep Data

What those three studies show us is that it's possible to automate, or at least semi-automate, critical aspects of the scientific method itself. Schmidt and Lipson show that it is possible to extract equations that perfectly model even seemingly chaotic systems. Yarkoni and colleagues show that it is possible to infer a complex behavioral state given input brian data.

My wife and I wanted to show that brainSCANr could be put to work for something more useful than just quantifying relationships between terms. So we created a simple algorithm to perform what we're calling "semi-automated hypothesis generation," which is predicated on a basic "the friend of a friend should be a friend" concept.

In the example below, the neurotransmitter "serotonin" has thousands of shared publications with "migraine," as well as with the brain region "striatum." However, migraine and striatum only share 16 publications.



That's very odd. Because in medicine there is a serotonin hypothesis for the root cause of migraines. And we (neuroscientists) know that serotonin is released in the striatum to modulate brain activity in that region. Given that those two things are true, why is there so little research regarding the role of the striatum in migraines?

Perhaps there's a missing connection?

Such missing links and other outliers in our models are the essence of deep data analytics. Sure, any data scientist worth their salt can take a mountain of data and reduce it down to a few simple plots. And such plots are important because they tell a story. But those aren't the only stories that our data can tell us.

For example, in my geoanalytics work as the data evangelist for Uber, I put some of my (definitely rudimentary) neuroscience network analytic skills to work to figure out how people move from neighborhood to neighborhood in San Francisco.

At one point, I checked to see if men and women moved around the city differently. A very simple regression model showed that the number of men who go to any given neighborhood significantly predicts the number of women who go to that same neighborhood.

No big deal.

But what's cool was seeing where the outliers were. When I looked at the models' residuals, *that's* where I found the far more interesting story. While it's good to have a model that fits your data, knowing where the model *breaks down* is not only important for internal metrics, but it also makes for a more interesting story: What's happening in the Marina district that so many more women want to go there? And why are there so many more men in SoMa?

The Paradox of Information

The interpretation of big data analytics can be a messy game. Maybe there are more men in SoMa because that's where AT&T Park is. But maybe there are just five guys who live in SoMa who happen to take Uber 100 times more often than average.

While data-driven posts make for fun reading (and writing), in the sciences we need to be more careful that we don't fall prey to *ad hoc*, just-so stories that sound perfectly reasonable and plausible, but which we cannot conclusively prove.

In 2008, psychologists David McCabe and Alan Castel published a paper in the journal *Cognition*, titled "Seeing is believing: The effect of brain images on judgments of scientific reasoning." In that paper, they showed that summaries of cognitive neuroscience findings that are accompanied by an image of a brain scan were rated as more credible by the readers.

This should cause any data scientist serious concern. In fact, I've formulated three laws of statistical analyses:

1. The more advanced the statistical methods used, the fewer critics are available to be properly skeptical.
2. The more advanced the statistical methods used, the more likely the data analyst will be to use math as a shield.
3. Any sufficiently advanced statistics can trick people into believing the results reflect truth.

The first law is closely related to the "bike shed effect" (also known as Parkinson's Law of Triviality) which states that, "the time spent on any item of the agenda will be in inverse proportion to the sum involved."

In other words, if you try to build a simple thing such as a public bike shed, there will be endless town hall discussions wherein people argue over trivial details such as the color of the door. But if you want to build a nuclear power plant — a project so vast and complicated that most people can't understand it — people will defer to expert opinion.

Such is the case with statistics.

If you make the mistake of going into the comments section of any news piece discussing a scientific finding, invariably someone will leave the comment, "correlation does not equal causation."

We'll go ahead and call that truism Voytek's fourth law.

But people rarely have the capacity to argue against the methods and models used by, say, neuroscientists or cosmologists.

But sometimes we get perfect models without any understanding of the underlying processes. What do we learn from that?

The always fantastic Radiolab did a follow-up story on the Schmidt and Lipson "automated science" research in an episode titled "Limits of Science." It turns out, a biologist contacted Schmidt and Lipson and gave them data to run their algorithm on. They wanted to figure out the principles governing the dynamics of a single-celled bacterium. Their result?

Well sometimes the stories we tell with data…they just don't make sense to us.

They found "two equations that describe the data."

But they didn't know what the equations *meant*. They had no context. Their variables had no meaning. Or, as Radiolab co-host Jad Abumrad put it, "the more we turn to computers with these big questions, the more they'll give us answers that we just don't understand."

So while big data projects are creating ridiculously exciting new vistas for scientific exploration and collaboration, we have to take care to avoid the Paradox of Information wherein we can know too many *things* without knowing what those "things" are.

Because at some point, we'll have so much data that we'll stop being able to discern the map from the territory. Our goal as (data) scientists should be to distill the essence of the data into something that tells as

true a story as possible while being as simple as possible to understand. Or, to operationalize that sentence better, we should aim to find balance between minimizing the residuals of our models and maximizing our ability to make sense of those models.

Recently, Stephen Wolfram released the results of a 20-year long experiment in personal data collection, including every keystroke he's typed and every email he's sent. In response, Robert Krulwich, the other co-host of Radiolab, concludes by saying "I'm looking at your data [Dr. Wolfram], and you know what's amazing to me? How much of you is missing."

Personally, I disagree; I believe that there's a humanity in those numbers and that Mr. Krulwich is falling prey to the idea that science somehow ruins the magic of the universe. Quoth Dr. Sagan:

> It is sometimes said that scientists are unromantic, that their passion to figure out robs the world of beauty and mystery. But is it not stirring to understand how the world actually works — that white light is made of colors, that color is the way we perceive the wavelengths of light, that transparent air reflects light, that in so doing it discriminates among the waves, and that the sky is blue for the same reason that the sunset is red? It does no harm to the romance of the sunset to know a little bit about it.

So go forth and create beautiful stories, my statistical friends. See you after peer-review.

The Chicken and Egg of Big Data Solutions

By Jim Stogdill

Before I came to O'Reilly I was building the "big data and disruptive analytics practice" at a major systems integrator. It was a blast to spend every week talking to customers in different industries who were waking up to the possibilities that technologies like Hadoop offered their businesses. Many of these businesses are going to fundamentally change as they embrace this stuff (or be replaced by those that do). But there's a catch.

Twenty years or so ago large integrators made big business building applications on the then-new relational paradigm. They put in Oracle

databases with custom code, wrote PowerBuilder apps on Sybase, and of course lots of businesses rolled their own with VB and SQL Server. It was an era of custom coding where Oracle, Sybase, SQL Server, Informix and etc. were thought of as platforms to build stuff on.

Then the market matured and shifted to package solution implementation. ERP, CRM,…, etc. The big guys focused on integrating again and told their clients there was no ROI in building custom stuff. ROI would come from integrating best-of-breed solutions. Databases became commodity back ends to the applications that were always the real focus.

Now along comes big data, NoSQL, data science, and all that stuff and it seems like we're starting the cycle over again. But this time clients, having been well trained over the last decade or so, aren't having any of that "build it from scratch" mentality. They know that Hadoop and other new technologies can be transformative to their business, but they want it packaged up and solution'ified like they are used to. I heard a lot of "let us know when you have a solution already built or available to buy that does X" in the last year.

Also, lots of the shops that do this stuff at scale are built and staffed around the package implementation model and have shed many of the skills they used to have for custom work. Everything from staffing models to methodologies are oriented toward package installation.

So, here we are with all of this disruptive technology, but we seem to have lost the institutional wherewithal to do anything with it in a lot of large companies. Of course that fact was hard on my numbers. I had a great pipeline of companies with pain to solve, and great technologies to solve it, but too much of the time it was hard to close it without readymade solutions.

Every week I talked to the companies building these new platforms to share leads and talk about their direction. After a while I started cutting them off when they wanted to talk about the features of their next release. I just got to the point where I didn't really care, it just wasn't all that relevant to my customers. I mean, it's important that they are making the platforms more manageable and building bridges to traditional BI, ETL, RDBMS, and the like. But the focus was too much on platforms and tools.

I wanted to know "What are you doing to encourage solution development? Are you staffing a support system for ISVs? What startups

and/or established players are you aware of that are building solutions on this platform?" So when I saw this tweet I let out a little yelp. Awesome! The lack of ready-to-install solutions was getting attention, and from Mike Olsen.

You can watch the rest of what Mike Olson said here and you'll find he tells a similar story about the RDBMS historical parallel.

I talked to Mike a few weeks ago to find out what was behind his comment and explore what else they are doing to support solution development. It boils down to what he said — he will help connect you with money — plus a newly launched partner program designed to provide better support to ISVs among others. Also, the continued attention to APIs and tools like Pig and Hive should make it easier for the solution ecosystem to develop. It can only be good for his business to have lots of other companies directly solving business problems, and simply pulling in his platform.

Hortonworks also started a partner program in the fall, and I think we'll see a lot more emphasis on this across the space this year. However, at the moment wherever I look (Hortonworks Partners, Cloudera Partners, Accel big data portfolio) the focus today remains firmly on platform and tools or partnering with integrators. Tresata, a startup focused on financial risk management, pops up in a lot of lists as the obvious odd one out — an actual domain-specific solution.

What about other people that could be building solutions? Is it the maturity level of the technology, the lack of penetration of Hadoop, etc., into your customer's data centers, or some combination of other factors that is slowing things down?

Of course, during the RDBMS adoption it took a lot of years before the custom era was over and thoroughly replaced by the era of package implementation. The question I'm pondering is whether customer expectations and the pace of technology will make it happen faster this time? Or is the disruptive value of big data going to continue to accrue only to risk-taking early adopters for the foreseeable future?

Walking the Tightrope of Visualization Criticism

By **Andy Kirk**

In a talk at the SEE conference in Germany, data illustrator Stefanie Posavec opened her talk with a sobering observation about how she had found the data visualization field really intimidating.

Her experience was that many visualization bloggers and active participants seem to believe in one right way and lots of wrong ways to create a visualization. To those entering the field, these types of views will create a fair amount of confusion, inconsistency, and contradiction. It demonstrates our current glass-is-never-full tendency toward critical evaluation.

This should act as an important wake-up call to all of us who care about maintaining an accessible and supportive community around data visualization and infographic design, particularly as these disciplines continue to penetrate the mainstream consciousness.

The fear is that Posavec's expressing of this view is just the tip of an iceberg. Who knows how many designers outside of the spotlight hold a similar perception and are reluctant to share their work and engage with the field?

In this article, I seek to take a detached view of the visualization field and weave in my experiences from delivering introductory data visualization training courses over the past year. I want to take a look at the constituency of this discipline and the role of critique to see how Posavec's experiences could have materialised and contrast them with the people I meet in my classes.

The Visualization Ecosystem

One of the most rewarding personal experiences from my training courses has been getting the chance to mix with a variety of delegates from different countries, cultures, occupations, and industries. Spending time with essentially everyday people, learning as much from them as they do from me, has been hugely refreshing. The term "everyday people" could be perceived as condescending, but far from it. Allow me to elaborate.

When you are an active participant in a field like data visualization you spend most of your time consuming and digesting information from your peers. This can create a bubble of exposure to just those hardcore connoisseurs — the academics, the authors, the designers, and the bloggers — who have spent years refining their knowledge and perfecting their craft.

This is the sharp-end of the field where the intensity of debate, knowledge exchange, and opinion expression is high. The observations emerging at this level represent the most perceptive, creative, and comprehensive insights into the design techniques on show today. The attention to detail, the care for quality, and the commitment to evaluation and feedback is significant. However, it can inadvertently create a certain suffocating or perhaps inhibiting barrier for many looking to learn and develop their capabilities.

A key observation from my training courses has been the sense that we in the field could be accused, at times, of a certain amount of design snobbery. We criticise and lambast many of the popular but "trashy" infographics, and believe them to be an inferior practice. However, during training sessions, I invite delegates to assess a variety of different types of visualization design, including such infographic pieces. I often hear comments that express and reason a preference or even a "like" for pieces that I would not. This has proven to be a highly illuminating experience.

A consequence of associating with or belonging to this top-tier "bubble" is that you can become somewhat detached and even oblivious to the opinions of those who might be considered to exist in the real world. These are the casual enthusiasts, the everyday people I mentioned early. They are likely to be beginning their journey into the field or have been nibbling around the edges for a while, but probably never too seriously until now. In contrast to the hardcore connoisseurs, this lower-expertise but more highly populated tier of the field's pyramid of participants makes up a totally different demographic and psychographic.

These people provide a great tapestry of different opinions, backgrounds, and capabilities and, generally, they offer a more sympathetic, fresh, and open-minded view on visualization design. Without the burden of knowledge, theories, and principles that the rest of us carry around with us all the time, and by not living and breathing the subject across every waking hour, their appreciation of visualization is more rooted in taste and instinct and fueled by a fresh enthusiasm to consume information in visual form.

Beyond and beneath this middle-tier sits, well, everybody else. These are the purely occasional consumers and nothing more. Their daily roles may not have anything to do with data, they possibly don't even know or probably care what visualization is. Yet, they belong to the

almost silent but abundant cohort of people who are occasionally curious enough to look at an attractive visualization or light-weight infographic. They don't want or need to learn about the field, they just find enough interest in having a look at some of its output.

This is the true make-up of the visualization and infographic field, and we need to appreciate its relevance.

The Irrationality of Needs: Fast Food to Fine Dining

There is a prominent, long-established film critic in the U.K. who is generally considered a fair and sound judge of movies. He has deep subject expertise and is capable of fully reasoning all his reviews with thorough analysis. Despite this, he does occasionally resort to the riposte "other opinions are available, but they're the wrong ones" when challenged by readers or viewers.

As with any subject's "expert" tier, we in data visualization can find ourselves being a little too closed off, perhaps believing the merit of our views hold greater weight than other, contrary opinions from outside. But this is largely because we don't always entirely appreciate the variety of intentions and needs behind visualization designs. Furthermore, there are so many different contexts, target audiences, and formats through which visual communication of data can exist.

Sometimes we're looking to impart a data-driven communication where the absolute accuracy of interpretation is vital. On other occasions it might be about creating a visual representation of data to impact more on an emotional level, trying to change behaviour and connect with people through non-standard methods. Sometimes we are working on subjects that are important, complex, and deep, and require a more engaging and prolonged interactive exploratory experience. By contrast, we might just be presenting some rather lightweight facts or stats that relate to a harmless, maybe even "fun," subject matter.

This is where a comparison with other creative territories like music, TV, movies, and food is appropriate to help illustrate how fundamentally impulsive, inconsistent, and irrational our tastes can be. Of course, the intention is very different with these channels of expression, but still we can relate to experiences when we sometimes prefer a fast food meal or to feast on junk food snacks as opposed to sitting down to a wholesome, home-cooked meal. We know it's probably bad for us, we'll probably spend more money on it, and we know we're likely to be hungry again in an hour, but we still do it.

You will typically never be too far away from running across intelligent, well-written movies or TV programs, but sometimes a trashy, loud, special-effects-laden blockbuster just does the trick. The critics might have told us how much we should hate them and how we should have spent our time with a more critically acclaimed work, but we don't care; we just want some mindless escapism. You can extend this to writing. Maybe we should all be sitting down in our spare time reading Shakespeare or Keats, enriching our minds. But most of us aren't. I know I'm not.

You can extend this to music, art, or really to any other creative channel. Of course, there are many other factors at play (access, time, resources, peer influences, etc.), but we still instinctively seek to mix things up on occasion and go against the grain. Being told what we should and shouldn't do can create as many followers as it does opponents.

It's the same with visualization. For many people, sometimes a harmless infographic showing some throw-away facts or stats about social media, or demonstrating how to avoid getting bitten by a shark is just what people fancy viewing at that point in time. This explains the vast success of works presented on gallery sites like visual.ly, the growth of design agencies like Column Five, and the general phenomenon of modern-day tower infographics.

Whilst more important subjects and works from leading organisations like the New York Times are arguably where we should be paying our attention to learn and respond to critical issues, occasionally we just need a release. We just want a blend of different visuals. This is the visualization ecosystem, and we need to appreciate its value. Nathan Yau recently wrote an insightful comment piece about this pattern.

Extend this discussion further and consider the appeal of fun and of aesthetic attraction to help stimulate the brain into engaging and learning with representations of information. This has been proposed as an important attribute of design for a long time but still exists as such a divisive issue within the data visualization field.

Whilst I recently remarked that there might be a sense that the traditional factions in the field were starting to better appreciate each other, I feel there is still more visible polarity than harmony. Indeed, arguably more polarity than even co-existence. This is an indication that the

field is still evolving but needs to mature, and it is through our critique where these fault-lines and opinion clashes manifest themselves. Most of it is valuable and healthy debate, but equally, we need to make sure it remains reasoned and accessible.

Grown-up Criticism

A key part of the training sessions I deliver is focused on trying to equip delegates with a more informed sense of how to evaluate a visualization piece. It urges them to attempt to understand the process, the purpose, and the parameters that have surrounded a project. Rather than drawing conclusions from a superficial "taste" reaction, they are asked to take a forensic approach to assessing the quality and effectiveness of a visualization, peeling through the layers of a visualization's anatomy and putting themselves into the mind of the designer.

This is something we should all try to do before publishing our knee-jerk conclusions to the world. To empathise with the constraints that might have existed within the project, the limitations of the data, try to imagine the brief and the influencing factors the designer had to contend with. When we view and evaluate a piece, we are looking at something that has not benefited from infinite time, endless resources, and limitless capability. Could we have done better ourselves given the same context?

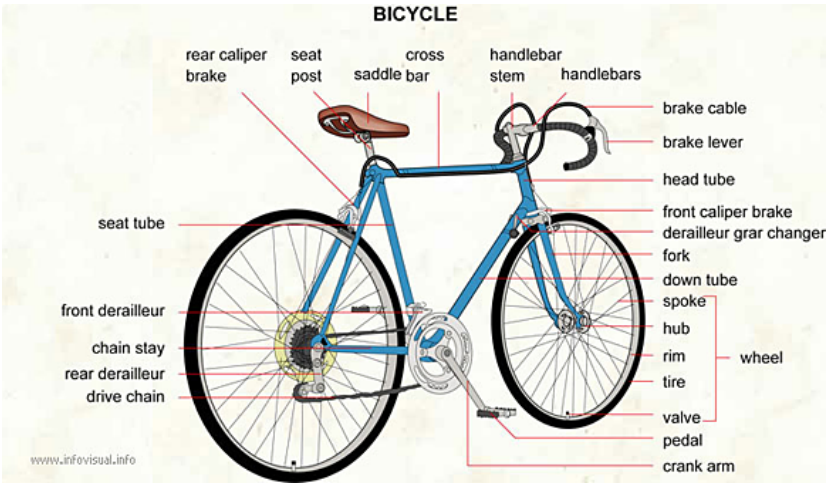
On a perverse level, I feel this part of the training risks eroding the raw innocence (without being disrespectful) that enables more casual observers to take visualizations and infographics on face value. They are not cursed by the depth of analysis and variety of lenses through which they should evaluate a piece.

However, I shouldn't worry because what always comes across from the delegates when we do this exercise is the very grounded, realistic, and practical appreciation of what works and doesn't work in different contexts. There is a mature and pragmatic acceptance and appreciation of the type of limitations, pressures, constraints, and interferences that might have shaped the resulting design.

Such experiences in my training course have made me think that those of us in the connoisseur's cohort are occasionally guilty of assessing visualization pieces too harshly, too readily, and too rapidly. This was the essence of Stefanie Posavec's observation. It's not so much looking at the glass being half empty; it's more akin to seeing the slightest shortcoming and amplifying the importance of this perceived flaw.

A recent observation on Twitter from Santiago Ortiz highlights this idea, characterising the type of critique that often exists about different visualization methods and approaches.

And here's the vehicle Ortiz was referring to:



Via the Visual Dictionary.

This observation resonates with a question I have been asked on several occasions by training course delegates. Many express a frustration in their struggle to understand and identify what makes a perfect visualization. By extension, they admit to a difficulty in establishing clarity in their own convictions about judging what is a right way and a wrong way to approach a visualization design.

Entering the field, you begin with fundamentally no informed reasoning for appreciation of quality; it is a gut instinct based on the effect it has on you. Yet, through the influence of reading key articles and exposure to social media, when you see others expressing a conviction, you feel obliged to jump off the fence and hurriedly wave a flag, any flag, of your judgment. It's not so much a case of following the crowd, rather more about feeling a need to express an opinion as quickly and as clearly as everybody else seems to.

Here's the truth: Developing clarity of your design conviction is difficult. If it were purely about taste, it would be easier. That's why you can be much more affirmative about your tastes in things like music, art, or movies. "Did it connect with you?" is a very open but fitting question that easily allows you to arrive at a Boolean type of response and the clarity of your judgment.

I recently wrote an article to discuss the visualizations I like. In this piece, I talked much less about style, approach, subjects, technique, or principles, but instead focused on those visualizations that give back more in return than you put in. That is my conviction, but it has taken a long while to arrive at that level of clarity. As many others will, I've been through a full discovery cycle of liking things that I now don't like and disliking things that I now do.

This conviction is informed by knowledge, by exposure to other disciplines and methods, and also through greater appreciation of what it takes to craft an effective visualization solution that works for the problem context it is responding to. Fundamentally, this is a hard discipline to do well.

Final Thoughts

The balance, fairness, and realism of our criticism needs to improve.

The desire of those active "experts" in the field to influence widespread effective practice needs to be matched by a greater maturity and sensitivity in the way we also evaluate the output of this creativity. Moreover, commentators and critics, myself included, need to develop a smarter appreciation of the different contexts in which these works are created.

A creative field, by its very nature, will have many different interpretations and perspectives, and the resolution and richness of this opinion is important to safeguard. Of course, promoting a more open-minded approach to evaluation doesn't mean to say there should be no critical analysis. We also need to ensure there isn't too much demonstration of the emperor's new clothes attitude, especially when a work looks cool or demonstrates impressive technical competence.

There is great importance in having the conviction and confidence to ask the question "so what?," to engage in constructive and mature critique (for example), and to exhibit a desire to understand and probe the intention behind all visualisation work. From this, we will all learn so much more and help create an environment that facilitates encouragement rather than discouragement.

This article likely contains some sweeping generalisations that manage to over-simplify things, but hopefully they help illustrate the impor-

tance of lifting our heads above the noise and seeing what's actually going on, who is active in this field, what roles they are taking on, and the value they are bringing to the whole visualization ecosystem, not just to the top table.

Fundamentally, what we need to avoid is inadvertently creating barriers to people trying to enter and develop in this field by creating the impression that a 1% missed opportunity is more important than the 99% of a design's features that were a nailed-on success.

I know I will be making a concerted effort to achieve this balance and fairness in my own analyses.

CHAPTER 6

Big Data and Health Care

Solving the Wanamaker Problem for Health Care

By Tim O'Reilly, Julie Steele, Mike Loukides and Colin Hill

> The best minds of my generation are thinking about how to make people click ads.
>
> — Jeff Hammerbacher
> *early Facebook employee*

> Work on stuff that matters.
>
> — Tim O'Reilly

In the early days of the 20th century, department store magnate John Wanamaker famously said, "I know that half of my advertising doesn't work. The problem is that I don't know which half."

The consumer Internet revolution was fueled by a search for the answer to Wanamaker's question. Google AdWords and the pay-per-click model began the transformation of a business in which advertisers paid for ad impressions into one in which they pay for results. "Cost per thousand impressions" (CPM) was outperformed by "cost per click" (CPC), and a new industry was born. It's important to understand why CPC outperformed CPM, though. Superficially, it's because Google was able to track when a user clicked on a link, and was therefore able to bill based on success. But billing based on success doesn't fundamentally change anything unless you can also change the success rate, and that's what Google was able to do. By using data to understand each user's behavior, Google was able to place advertisements that an individual was likely to click. They knew "which half" of their advertising was more likely to be effective, and didn't bother with the rest.

Since then, data and predictive analytics have driven ever deeper insight into user behavior such that companies like Google, Facebook, Twitter, and LinkedIn are fundamentally data companies. And data isn't just transforming the consumer Internet. It is transforming finance, design, and manufacturing — and perhaps most importantly, health care.

How is data science transforming health care? There are many ways in which health care is changing, and needs to change. We're focusing on one particular issue: the problem Wanamaker described when talking about his advertising. How do you make sure you're spending money effectively? Is it possible to know what will work in advance?

Too often, when doctors order a treatment, whether it's surgery or an over-the-counter medication, they are applying a "standard of care" treatment or some variation that is based on their own intuition, effectively hoping for the best. The sad truth of medicine is that we don't always understand the relationship between treatments and outcomes. We have studies to show that various treatments will work more often than placebos; but, like Wanamaker, we know that much of our medicine doesn't work for half or our patients, we just don't know which half. At least, not in advance. One of data science's many promises is that, if we can collect enough data about medical treatments and use that data effectively, we'll be able to predict more accurately which treatments will be effective for which patient, and which treatments won't.

A better understanding of the relationship between treatments, outcomes, and patients will have a huge impact on the practice of medicine in the United States. Health care is expensive. The U.S. spends over $2.6 trillion on health care every year, an amount that constitutes a serious fiscal burden for government, businesses, and our society as a whole. These costs include over $600 billion of unexplained variations in treatments: treatments that cause no differences in outcomes, or even make the patient's condition worse. We have reached a point at which our need to understand treatment effectiveness has become vital — to the health care system and to the health and sustainability of the economy overall.

Why do we believe that data science has the potential to revolutionize health care? After all, the medical industry has had data for generations: clinical studies, insurance data, hospital records. But the health care industry is now awash in data in a way that it has never been before: from biological data such as gene expression, next-generation DNA sequence data, proteomics, and metabolomics, to clinical data and health outcomes data contained in ever more prevalent electronic health records (EHRs) and longitudinal drug and medical claims. We have entered a new era in which we can work on massive datasets effectively, combining data from clinical trials and direct observation by practicing physicians (the records generated by our $2.6 trillion of medical expense). When we combine data with the resources needed to work on the data, we can start asking the important questions, the Wanamaker questions, about what treatments work and for whom.

The opportunities are huge: for entrepreneurs and data scientists looking to put their skills to work disrupting a large market, for researchers trying to make sense out of the flood of data they are now generating, and for existing companies (including health insurance companies, biotech, pharmaceutical, and medical device companies, hospitals and other care providers) that are looking to remake their businesses for the coming world of outcome-based payment models.

Making Health Care More Effective

What, specifically, does data allow us to do that we couldn't do before? For the past 60 or so years of medical history, we've treated patients as some sort of an average. A doctor would diagnose a condition and recommend a treatment based on what worked for most people, as reflected in large clinical studies. Over the years, we've become more sophisticated about what that average patient means, but that same

statistical approach didn't allow for differences between patients. A treatment was deemed effective or ineffective, safe or unsafe, based on double-blind studies that rarely took into account the differences between patients. With the data that's now available, we can go much further. The exceptions to this are relatively recent and have been dominated by cancer treatments, the first being Herceptin for breast cancer in women who over-express the Her2 receptor. With the data that's now available, we can go much further for a broad range of diseases and interventions that are not just drugs but include surgery, disease management programs, medical devices, patient adherence, and care delivery.

For a long time, we thought that Tamoxifen was roughly 80% effective for breast cancer patients. But now we know much more: we know that it's 100% effective in 70 to 80% of the patients, and ineffective in the rest. That's not word games, because we can now use genetic markers to tell whether it's likely to be effective or ineffective for any given patient, and we can tell in advance whether to treat with Tamoxifen or to try something else.

Two factors lie behind this new approach to medicine: a different way of using data, and the availability of new kinds of data. It's not just stating that the drug is effective on most patients, based on trials (indeed, 80% is an enviable success rate); it's using artificial intelligence techniques to divide the patients into groups and then determine the difference between those groups. We're not asking whether the drug is effective; we're asking a fundamentally different question: "for which patients is this drug effective?" We're asking about the patients, not just the treatments. A drug that's only effective on 1% of patients might be very valuable if we can tell who that 1% is, though it would certainly be rejected by any traditional clinical trial.

More than that, asking questions about patients is only possible because we're using data that wasn't available until recently: DNA sequencing was only invented in the mid-1970s, and is only now coming into its own as a medical tool. What we've seen with Tamoxifen is as clear a solution to the Wanamaker problem as you could ask for: we now know when that treatment will be effective. If you can do the same thing with millions of cancer patients, you will both improve outcomes and save money.

Dr. Lukas Wartman, a cancer researcher who was himself diagnosed with terminal leukemia, was successfully treated with sunitinib, a drug

that was only approved for kidney cancer. Sequencing the genes of both the patient's healthy cells and cancerous cells led to the discovery of a protein that was out of control and encouraging the spread of the cancer. The gene responsible for manufacturing this protein could potentially be inhibited by the kidney drug, although it had never been tested for this application. This unorthodox treatment was surprisingly effective: Wartman is now in remission.

While this treatment was exotic and expensive, what's important isn't the expense but the potential for new kinds of diagnosis. The price of gene sequencing has been plummeting; it will be a common doctor's office procedure in a few years. And through Amazon and Google, you can now "rent" a cloud-based supercomputing cluster that can solve huge analytic problems for a few hundred dollars per hour. What is now exotic inevitably becomes routine.

But even more important: we're looking at a completely different approach to treatment. Rather than a treatment that works 80% of the time, or even 100% of the time for 80% of the patients, a treatment might be effective for a small group. It might be entirely specific to the individual; the next cancer patient may have a different protein that's out of control, an entirely different genetic cause for the disease. Treatments that are specific to one patient don't exist in medicine as it's currently practiced; how could you ever do an FDA trial for a medication that's only going to be used once to treat a certain kind of cancer?

Foundation Medicine is at the forefront of this new era in cancer treatment. They use next-generation DNA sequencing to discover DNA sequence mutations and deletions that are currently used in standard of care treatments, as well as many other actionable mutations that are tied to drugs for other types of cancer. They are creating a patient-outcomes repository that will be the fuel for discovering the relation between mutations and drugs. Foundation has identified DNA mutations in 50% of cancer cases for which drugs exist (information via a private communication), but are not currently used in the standard of care for the patient's particular cancer.

The ability to do large-scale computing on genetic data gives us the ability to understand the origins of disease. If we can understand why an anti-cancer drug is effective (what specific proteins it affects), and if we can understand what genetic factors are causing the cancer to spread, then we're able to use the tools at our disposal much more effectively. Rather than using imprecise treatments organized around

symptoms, we'll be able to target the actual causes of disease, and design treatments tuned to the biology of the specific patient. Eventually, we'll be able to treat 100% of the patients 100% of the time, precisely because we realize that each patient presents a unique problem.

Personalized treatment is just one area in which we can solve the Wanamaker problem with data. Hospital admissions are extremely expensive. Data can make hospital systems more efficient, and to avoid preventable complications such as blood clots and hospital readmissions. It can also help address the challenge of health care hotspotting (a term coined by Atul Gawande): finding people who use an inordinate amount of health care resources. By looking at data from hospital visits, Dr. Jeffrey Brenner of Camden, NJ, was able to determine that "just one per cent of the hundred thousand people who made use of Camden's medical facilities accounted for thirty per cent of its costs." Furthermore, many of these people came from only two apartment buildings. Designing more effective medical care for these patients was difficult; it doesn't fit our health insurance system, the patients are often dealing with many serious medical issues (addiction and obesity are frequent complications), and have trouble trusting doctors and social workers. It's counter-intuitive, but spending more on some patients now results in spending less on them when they become really sick. While it's a work in progress, it looks like building appropriate systems to target these high-risk patients and treat them before they're hospitalized will bring significant savings.

Many poor health outcomes are attributable to patients who don't take their medications. Eliza, a Boston-based company started by Alexandra Drane, has pioneered approaches to improve compliance through interactive communication with patients. Eliza improves patient drug compliance by tracking which types of reminders work on which types of people; it's similar to the way companies like Google target advertisements to individual consumers. By using data to analyze each patient's behavior, Eliza can generate reminders that are more likely to be effective. The results aren't surprising: if patients take their medicine as prescribed, they are more likely to get better. And if they get better, they are less likely to require further, more expensive treatment. Again, we're using data to solve Wanamaker's problem in medicine: we're spending our resources on what's effective, on appropriate reminders that are mostly to get patients to take their medications.

More Data, More Sources

The examples we've looked at so far have been limited to traditional sources of medical data: hospitals, research centers, doctor's offices, insurers. The Internet has enabled the formation of patient networks aimed at sharing data. Health social networks now are some of the largest patient communities. As of November 2011, PatientsLikeMe has over 120,000 patients in 500 different condition groups; ACOR has over 100,000 patients in 127 cancer support groups; 23andMe has over 100,000 members in their genomic database; and diabetes health social network SugarStats has over 10,000 members. These are just the larger communities, thousands of small communities are created around rare diseases, or even uncommon experiences with common diseases. All of these communities are generating data that they voluntarily share with each other and the world.

Increasingly, what they share is not just anecdotal, but includes an array of clinical data. For this reason, these groups are being recruited for large-scale crowdsourced clinical outcomes research.

Thanks to ubiquitous data networking through the mobile network, we can take several steps further. In the past two or three years, there's been a flood of personal fitness devices (such as the Fitbit) for monitoring your personal activity. There are mobile apps for taking your pulse, and an iPhone attachment for measuring your glucose. There has been talk of mobile applications that would constantly listen to a patient's speech and detect changes that might be the precursor for a stroke, or would use the accelerometer to report falls. Tanzeem Choudhury has developed an app called Be Well that is intended primarily for victims of depression, though it can be used by anyone. Be Well monitors the user's sleep cycles, the amount of time they spend talking, and the amount of time they spend walking. The data is scored, and the app makes appropriate recommendations, based both on the individual patient and data collected across all the app's users.

Continuous monitoring of critical patients in hospitals has been normal for years; but we now have the tools to monitor patients constantly, in their home, at work, wherever they happen to be. And if this sounds like big brother, at this point most of the patients are willing. We don't want to transform our lives into hospital experiences; far from it! But we can collect and use the data we constantly emit, our "data exhaust,"

to maintain our health, to become conscious of our behavior, and to detect oncoming conditions before they become serious. The most effective medical care is the medical care you avoid because you don't need it.

Paying for Results

Once we're on the road toward more effective health care, we can look at other ways in which Wanamaker's problem shows up in the medical industry. It's clear that we don't want to pay for treatments that are ineffective. Wanamaker wanted to know which part of his advertising was effective, not just to make better ads, but also so that he wouldn't have to buy the advertisements that wouldn't work. He wanted to pay for results, not for ad placements. Now that we're starting to understand how to make treatment effective, now that we understand that it's more than rolling the dice and hoping that a treatment that works for a typical patient will be effective for you, we can take the next step: Can we change the underlying incentives in the medical system? Can we make the system better by paying for results, rather than paying for procedures?

It's shocking just how badly the incentives in our current medical system are aligned with outcomes. If you see an orthopedist, you're likely to get an MRI, most likely at a facility owned by the orthopedist's practice. On one hand, it's good medicine to know what you're doing before you operate. But how often does that MRI result in a different treatment? How often is the MRI required just because it's part of the protocol, when it's perfectly obvious what the doctor needs to do? Many men have had PSA tests for prostate cancer; but in most cases, aggressive treatment of prostate cancer is a bigger risk than the disease itself. Yet the test itself is a significant profit center. Think again about Tamoxifen, and about the pharmaceutical company that makes it. In our current system, what does "100% effective in 80% of the patients" mean, except for a 20% loss in sales? That's because the drug company is paid for the treatment, not for the result; it has no financial interest in whether any individual patient gets better. (Whether a statistically significant number of patients has side-effects is a different issue.) And at the same time, bringing a new drug to market is very expensive, and might not be worthwhile if it will only be used on the remaining 20% of the patients. And that's assuming that one drug, not two, or 20, or 200 will be required to treat the unlucky 20% effectively.

It doesn't have to be this way.

In the U.K., Johnson & Johnson, faced with the possibility of losing reimbursements for their multiple myeloma drug Velcade, agreed to refund the money for patients who did not respond to the drug. Several other pay-for-performance drug deals have followed since, paving the way for the ultimate transition in pharmaceutical company business models in which their product is health outcomes instead of pills. Such a transition would rely more heavily on real-world outcome data (are patients actually getting better?), rather than controlled clinical trials, and would use molecular diagnostics to create personalized "treatment algorithms." Pharmaceutical companies would also focus more on drug compliance to ensure health outcomes were being achieved. This would ultimately align the interests of drug makers with patients, their providers, and payors.

Similarly, rather than paying for treatments and procedures, can we pay hospitals and doctors for results? That's what Accountable Care Organizations (ACOs) are about. ACOs are a leap forward in business model design, where the provider shoulders any financial risk. ACOs represent a new framing of the much maligned HMO approaches from the '90s, which did not work. HMOs tried to use statistics to predict and prevent unneeded care. The ACO model, rather than controlling doctors with what the data says they "should" do, uses data to measure how each doctor performs. Doctors are paid for successes, not for the procedures they administer. The main advantage that the ACO model has over the HMO model is how good the data is, and how that data is leveraged. The ACO model aligns incentives with outcomes: a practice that owns an MRI facility isn't incentivized to order MRIs when they're not necessary. It is incentivized to use all the data at its disposal to determine the most effective treatment for the patient, and to follow through on that treatment with a minimum of unnecessary testing.

When we know which procedures are likely to be successful, we'll be in a position where we can pay only for the health care that works. When we can do that, we've solved Wanamaker's problem for health care.

Enabling Data

Data science is not optional in health care reform; it is the linchpin of the whole process. All of the examples we've seen, ranging from cancer

treatment to detecting hot spots where additional intervention will make hospital admission unnecessary, depend on using data effectively: taking advantage of new data sources and new analytics techniques, in addition to the data the medical profession has had all along.

But it's too simple just to say "we need data." We've had data all along: handwritten records in manila folders on acres and acres of shelving. Insurance company records. But it's all been locked up in silos: insurance silos, hospital silos, and many, many doctor's office silos. Data doesn't help if it can't be moved, if data sources can't be combined.

There are two big issues here. First, a surprising number of medical records are still either hand-written, or in digital formats that are scarcely better than hand-written (for example, scanned images of hand-written records). Getting medical records into a format that's computable is a prerequisite for almost any kind of progress. Second, we need to break down those silos.

Anyone who has worked with data knows that, in any problem, 90% of the work is getting the data in a form in which it can be used; the analysis itself is often simple. We need electronic health records: patient data in a more-or-less standard form that can be shared efficiently, data that can be moved from one location to another at the speed of the Internet. Not all data formats are created equal, and some are certainly better than others: but at this point, any machine-readable format, even simple text files, is better than nothing. While there are currently hundreds of different formats for electronic health records, the fact that they're electronic means that they can be converted from one form into another. Standardizing on a single format would make things much easier, but just getting the data into some electronic form, any, is the first step.

Once we have electronic health records, we can link doctor's offices, labs, hospitals, and insurers into a data network, so that all patient data is immediately stored in a data center: every prescription, every procedure, and whether that treatment was effective or not. This isn't some futuristic dream; it's technology we have now. Building this network would be substantially simpler and cheaper than building the networks and data centers now operated by Google, Facebook, Amazon, Apple, and many other large technology companies. It's not even close to pushing the limits.

Electronic health records enable us to go far beyond the current mechanism of clinical trials. In the past, once a drug has been approved in

trials, that's effectively the end of the story: running more tests to determine whether it's effective in practice would be a huge expense. A physician might get a sense for whether any treatment worked, but that evidence is essentially anecdotal: it's easy to believe that something is effective because that's what you want to see. And if it's shared with other doctors, it's shared while chatting at a medical convention. But with electronic health records, it's possible (and not even terribly expensive) to collect documentation from thousands of physicians treating millions of patients. We can find out when and where a drug was prescribed, why, and whether there was a good outcome. We can ask questions that are never part of clinical trials: is the medication used in combination with anything else? What other conditions is the patient being treated for? We can use machine learning techniques to discover unexpected combinations of drugs that work well together, or to predict adverse reactions. We're no longer limited by clinical trials; every patient can be part of an ongoing evaluation of whether his treatment is effective, and under what conditions. Technically, this isn't hard. The only difficult part is getting the data to move, getting data in a form where it's easily transferred from the doctor's office to analytics centers.

To solve problems of hot-spotting (individual patients or groups of patients consuming inordinate medical resources) requires a different combination of information. You can't locate hot spots if you don't have physical addresses. Physical addresses can be geocoded (converted from addresses to longitude and latitude, which is more useful for mapping problems) easily enough, once you have them, but you need access to patient records from all the hospitals operating in the area under study. And you need access to insurance records to determine how much health care patients are requiring, and to evaluate whether special interventions for these patients are effective. Not only does this require electronic records, it requires cooperation across different organizations (breaking down silos), and assurance that the data won't be misused (patient privacy). Again, the enabling factor is our ability to combine data from different sources; once you have the data, the solutions come easily.

Breaking down silos has a lot to do with aligning incentives. Currently, hospitals are trying to optimize their income from medical treatments, while insurance companies are trying to optimize their income by minimizing payments, and doctors are just trying to keep their heads above water. There's little incentive to cooperate. But as financial pres-

sures rise, it will become critically important for everyone in the health care system, from the patient to the insurance executive, to assume that they are getting the most for their money. While there's intense cultural resistance to be overcome (through our experience in data science, we've learned that it's often difficult to break down silos within an organization, let alone between organizations), the pressure of delivering more effective health care for less money will eventually break the silos down. The old zero-sum game of winners and losers must end if we're going to have a medical system that's effective over the coming decades.

Data becomes infinitely more powerful when you can mix data from different sources: many doctor's offices, hospital admission records, address databases, and even the rapidly increasing stream of data coming from personal fitness devices. The challenge isn't employing our statistics more carefully, precisely, or guardedly. It's about letting go of an old paradigm that starts by assuming only certain variables are key and ends by correlating only these variables. This paradigm worked well when data was scarce, but if you think about, these assumptions arise precisely because data is scarce. We didn't study the relationship between leukemia and kidney cancers because that would require asking a huge set of questions that would require collecting a lot of data; and a connection between leukemia and kidney cancer is no more likely than a connection between leukemia and flu. But the existence of data is no longer a problem: we're collecting the data all the time. Electronic health records let us move the data around so that we can assemble a collection of cases that goes far beyond a particular practice, a particular hospital, a particular study. So now, we can use machine learning techniques to identify and test all possible hypotheses, rather than just the small set that intuition might suggest. And finally, with enough data, we can get beyond correlation to causation: rather than saying "A and B are correlated," we'll be able to say "A causes B," and know what to do about it.

Building the Health Care System We Want

The U.S. ranks 37th out of developed economies in life expectancy and other measures of health, while by far outspending other countries on per-capita health care costs. We spend 18% of GDP on health care, while other countries on average spend on the order of 10% of GDP. We spend a lot of money on treatments that don't work, because we have a poor understanding at best of what will and won't work.

Part of the problem is cultural. In a country where even pets can have hip replacement surgery, it's hard to imagine not spending every penny you have to prolong Grandma's life — or your own. The U.S. is a wealthy nation, and health care is something we choose to spend our money on. But wealthy or not, nobody wants ineffective treatments. Nobody wants to roll the dice and hope that their biology is similar enough to a hypothetical "average" patient. No one wants a "winner take all" payment system in which the patient is always the loser, paying for procedures whether or not they are helpful or necessary. Like Wanamaker with his advertisements, we want to know what works, and we want to pay for what works. We want a smarter system where treatments are designed to be effective on our individual biologies; where treatments are administered effectively; where our hospitals our used effectively; and where we pay for outcomes, not for procedures.

We're on the verge of that new system now. We don't have it yet, but we can see it around the corner. Ultra-cheap DNA sequencing in the doctor's office, massive inexpensive computing power, the availability of EHRs to study whether treatments are effective even after the FDA trials are over, and improved techniques for analyzing data are the tools that will bring this new system about. The tools are here now; it's up to us to put them into use.

Recommended Reading

We recommend the following articles and books regarding technology, data, and health care reform:

- Ahier, Brian. "Big data is the next big thing in health IT," O'Reilly Radar. February 27, 2012.
- Bigelow, Bruce. "Big Data, Big Biology, and the 'Tipping Point' in Quantified Health," Xconomy. April 26, 2012.
- Brawley, Otis Webb. *How We Do Harm: A Doctor Breaks Ranks About Being Sick in America*. St. Marten's Press, 2012.
- Christensen, Clayton M. et al. *The Innovator's Prescription: A Disruptive Solution for Health Care*. McGraw Hill, 2008.
- Howard, Alex. "Data for the Public Good," O'Reilly Radar. February 22, 2012.
- Manyika, James et al. "Big data: The next frontier for innovation, competition, and productivity," McKinsey Global Institute. May, 2011.

- Oram, Andy. "Five tough lessons I had to learn about health care," O'Reilly Radar. March 26, 2012.
- Shah, Nigam H and Jessica D Tenenbaum. "The coming age of data-driven medicine: translational bioinformatics' next frontier," *Journal of the American Medical Informatics Association* (JAMIA). March 26, 2012.
- Trotter, Fred and David Uhlman. *Meaningful Use and Beyond*. O'Reilly Media, 2011.
- Wilbanks, John. "Valuing Health Care: Improving Productivity and Quality" [PDF], Ewing Marion Kauffman Foundation. April, 2012.

Dr. Farzad Mostashari on Building the Health Information Infrastructure for the Modern ePatient

By Alex Howard

To learn more about what levers the government is pulling to catalyze innovation in the healthcare system, I turned to Dr. Farzad Mostashari (@Farzad_ONC). As the National Coordinator for Health IT, Mostashari is one of the most important public officials entrusted with improving the nation's healthcare system through smarter use of technology.

Mostashari, a public-health informatics specialist, was named ONC chief in April 2011, replacing Dr. David Blumenthal. Mostashari's full biography, available at HHS.gov, notes that he "was one of the lead investigators in the outbreaks of West Nile Virus and anthrax in New York City, and was among the first developers of real-time electronic disease surveillance systems nationwide."

I talked to Mostashari on the same day that he published a look back over 2011, which he hailed as a year of momentous progress in health information technology. Our interview follows.

What excites you about your work? What trends matter here?

Farzad Mostashari: Well, it's a really fun job. It feels like this is the ideal time for this health IT revolution to tie into other massive megatrends that are happening around consumer and patient empowerment, payment and delivery reform, as I talked about in my TED Med Talk with Aneesh Chopra.

These three streams [how patients are cared for, how care is paid for, and how people take care of their own health] coming together feels great. And it really feels like we're making amazing progress.

How does what's happening today grow out of the passage of the Health Information Technology for Economic and Clinical Health Act (HITECH) Act in 2009?

Farzad Mostashari: HITECH was a key part of ARRA, the American Recovery and Reinvestment Act. This is the reinvestment part. People think of roadways and runways and railways. This is the information infrastructure for healthcare.

In the past two years, we made as much progress on adoption as we had made in the past 20 years before that. We doubled the adoption of electronic health records in physician offices between the time the stimulus passed and now. What that says is that a large number of barriers have been addressed, including the financial barriers that are addressed by the health IT incentive payments.

It also, I think, points to the innovation that's happening in the health IT marketplace, with more products that people want to buy and want to use, and an explosion in the number of options people have.

The programs we put in place, like the Regional Health IT Extension Centers modeled after the Agriculture Extension program, give a helping hand. There are local nonprofits throughout the country that are working with one-third of all primary care providers in this country to help them adopt electronic health records, particularly smaller practices and maybe health centers, critical access hospitals, and so forth.

This is obviously a big lift and a big change for medicine. It moves at what Jay Walker called "med speed," not tech speed. The pace of transformation in medicine that's happening right now may be unparalleled. It's a good thing.

Healthcare providers have a number of options as they adopt electronic health records. How do you think about the choice between open source versus proprietary options?

Farzad Mostashari: We're pretty agnostic in terms of the technology and the business model. What matters are the outcomes. We've really left the decisions about what technology to use to the people who have to live with it, like the doctors and hospitals who make the purchases.

There are definitely some very successful models, not only on the EHR side, but also on the health information exchange side.

(*Note: For more on this subject, read Brian Ahier's Radar post on the Health Internet.*)

What role do open standards play in the future of healthcare?

Farzad Mostashari: We are passionate believers in open standards. We think that everybody should be using them. We've gotten really great participation by vendors of open source and proprietary software, in terms of participating in an open standards development process.

I think what we've enabled, through things like modular certification, is a lot more innovation. Different pieces of the entire ecosystem could be done through reducing the barrier to entry, enabling a variety of different innovative startups to come to the field. What we're seeing is, a lot of the time, this is migrating from installed software to web services.

If we're setting up a reference implementation of the standards, like the Connect software or popHealth, we do it through a process where the result is open source. I think the government as a platform approach at the Veterans Affairs department, DoD, and so forth is tremendously important.

How is the mobile revolution changing healthcare?

We had Jay Walker talking about big change [at a recent ONC Grantee Meeting]. I just have this indelible image of him waving in his left hand a clay cone with cuneiform on it that is from 2,000 B.C. — 4,000 years ago — and in his right hand he held his iPhone.

He was saying both of them represented the cutting edge of technology that evolved to meet consumer need. His strong assertion was that this is absolutely going to revolutionize what happens in medicine at tech speed. Again, not "med speed."

I had the experience of being at my clinic, where I get care, and the pharmacist sitting in the starched, white coat behind the counter telling me that I should take this medicine at night.

And I said, "Well, it's easier for me to take it in the morning." And he said, "Well, it works better at night."

And I asked, acting as an empowered patient, "Well, what's the half life?" And he answered, "Okay. Let me look it up."

He started clacking away at his pharmacy information system; clickity clack, clickity clack. I can't see what he's doing. And then he says, "Ah hell," and he pulls out his smartphone and Googles it.

There's now a democratization of information and information tools, where we're pushing the analytics to the cloud. Being able to put that in the hand of not just every doctor or every healthcare provider but *every patient* is absolutely going to be that third strand of the DNA, putting us on the right path for getting healthcare that results in health.

We're making sure that people know they have a right to get their own data, making sure that the policies are aligned with that. We're making sure that we make it easy for doctors to give patients their own information through things like the Direct Project, the Blue Button, meaningful use requirements, or the Consumer E-Health Pledge.

We have more than 250 organizations that collectively hold data for 100 million Americans that pledge to make it easy for people to get electronic copies of their own data.

Do you think people will take ownership of their personal health data and engage in what Susannah Fox has described as "peer-to-peer healthcare"?

Farzad Mostashari: I think that it will be not just possible, not even just okay, but actually *encouraged* for patients to be engaged in their care as partners. Let the epatient help. I think we're going to see that emerging as there's more access and more tools for people to do stuff with their data once they get it through things like the health data initiative. We're also beginning to work with stakeholder groups, like Consumer's Union, the American Nurses Association, and some of the disease groups, to change attitudes around it being okay to ask for your own records.

This interview was edited and condensed.

John Wilbanks Discusses the Risks and Rewards of a Health Data Commons

By Alex Howard

As I wrote earlier this year in an ebook on data for the public good, while the idea of data as a currency is still in its infancy, it's important to think about where the future is taking us and our personal data.

If the Obama administration's smart disclosure initiatives gather steam, more citizens will be able to do more than think about personal data: they'll be able to access their financial, health, education, or energy data. In the U.S. federal government, the Blue Button initiative, which initially enabled veterans to download personal health data, is now spreading to all federal employees, and it also earned adoption at private institutions like Aetna and Kaiser Permanente. Putting health data to work stands to benefit hundreds of millions of people. The Locker Project, which provides people with the ability to move and store personal data, is another approach to watch.

The promise of more access to personal data, however, is balanced by accompanying risks. Smartphones, tablets, and flash drives, after all, are lost or stolen every day. Given the potential of mhealth, and big data and health care information technology, researchers and policy makers alike are moving forward with their applications. As they do so, conversations and rulemaking about health care privacy will need to take into account not just data collection or retention but context and use.

Put simply, businesses must confront the ethical issues tied to massive aggregation and data analysis. Given that context, Fred Trotter's post on who owns health data is a crucial read. As Fred highlights, the real issue is not ownership, per se, but "What rights do patients have regarding health care data that refers to them?"

Would, for instance, those rights include the ability to donate personal data to a data commons, much in the same way organs are donated now for research? That question isn't exactly hypothetical, as the following interview with John Wilbanks highlights.

Wilbanks, a senior fellow at the Kauffman Foundation and director of the Consent to Research Project, has been an advocate for open data

and open access for years, including a stint at Creative Commons; a fellowship at the World Wide Web Consortium; and experience in the academic, business, and legislative worlds. Wilbanks will be speaking at the Strata Rx Conference in October.

Our interview, lightly edited for content and clarity, follows.

Where did you start your career? Where has it taken you?

John Wilbanks: I got into all of this, in many ways, because I studied philosophy 20 years ago. What I studied inside of philosophy was semantics. In the '90s, that was actually sort of pointless because there wasn't much semantic stuff happening computationally.

In the late '90s, I started playing around with biotech data, mainly because I was dating a biologist. I was sort of shocked at how the data was being represented. It wasn't being represented in a way that was very semantic, in my opinion. I started a software company and we ran that for a while, [and then] sold it during the crash.

I went to the World Wide Web Consortium, where I spent a year helping start their Semantic Web for Life Sciences project. While I was there, Creative Commons (CC) asked me to come and start their science project because I had known a lot of those guys. When I started my company, I was at the Berkman Center at Harvard Law School, and that's where Creative Commons emerged from, so I knew the people. I knew the policy and I had gone off and had this bioinformatics software adventure.

I spent most of the last eight years at CC working on trying to build different commons in science. We looked at open access to scientific literature, which is probably where we had the most success because that's copyright-centric. We looked at patents. We looked at physical laboratory materials, like stem cells in mice. We looked at different legal regimes to share those things. And we looked at data. We looked at both the technology aspects and legal aspects of sharing data and making it useful.

A couple of times over those years, we almost pivoted from science to health because science is so institutional that it's really hard for any of the individual players to create sharing systems. It's not like software, where anyone with a PC and an Internet connection can contribute to free software, or Flickr, where anybody with a digital camera can license something under CC. Most scientists are actually restricted by their institutions. They can't share, even if they want to.

Health kept being interesting because it was the individual patients who had a motivation to actually create something different than the system did. At the same time, we were watching and seeing the capacity of individuals to capture data about themselves exploding. So, at the same time that the capacity of the system to capture data about you exploded, your own capacity to capture data exploded.

That, to me, started taking on some of the interesting contours that make Creative Commons successful, which was that you didn't need a large number of people. You didn't need a very large percentage of Wikipedia users to create Wikipedia. You didn't need a large percentage of free software users to create free software. If this capacity to generate data about your health was exploding, you didn't need a very large percentage of those people to create an awesome data resource: you needed to create the legal and technical systems for the people who did choose to share to make that sharing useful.

Since Creative Commons is really a copyright-centric organization, I left because the power on which you're going to build a commons of health data is going to be privacy power, not copyright power. What I do now is work on informed consent, which is the legal system you need to work with instead of copyright licenses, as well as the technologies that then store, clean, and forward user-generated data to computational health and computational disease research.

What are the major barriers to people being able to donate their data in the same way they might donate their organs?

John Wilbanks: Right now, it looks an awful lot like getting onto the Internet before there was the Web. The big ISPs kind of dominated the early adopters of computer technologies. You had AOL. You had CompuServe. You had Prodigy. And they didn't communicate with each other. You couldn't send email from AOL to CompuServe.

What you have now depends on the kind of data. If the data that interests you is your genotype, you're probably a 23andMe customer and you've got a bunch of your data at 23andMe. If you are the kind of person who has a chronic illness and likes to share information about that illness, you're probably a customer at PatientsLikeMe. But those two systems don't interoperate. You can't send data from one to the other very effectively or really at all.

On top of that, the system has data about you. Your insurance company has your billing records. Your physician has your medical records. Your pharmacy has your pharmacy records. And if you do quantified self, you've got your own set of data streams. You've got your Fitbit, the data coming off of your smartphone, and your meal data.

Almost all of these are basically populating different silos. In some cases, you have the right to download certain pieces of the data. For the most part, you don't. It's really hard for you, as an individual, to build your own, multidimensional picture of your data, whereas it's actually fairly easy for all of those companies to sell your data to one another. There's not a lot of technology that lets you share.

What are some of the early signals we're seeing about data usage moving into actual regulatory language?

John Wilbanks: The regulatory language actually makes it fairly hard to do contextual privacy waiving, in a Creative Commons sense. It's hard to do granular permissions around privacy in the way you can do granular conditional copyright grants because you don't have intellectual property. The only legal tool you have is a contract, and the contracts don't have a lot of teeth.

It's pretty hard to do anything beyond a gift. It's more like organ donation, where you don't get to decide where the organs go. What I'm working on is basically a donation, not a conditional gift. The regulatory environment makes it quite hard to do anything besides that.

There was a public comment period that just finished. It's an announcement of proposed rulemaking on what's called the Common Rule, which is the Department of Health and Human Services privacy language. It was looking to re-examine the rules around letting de-identified data or anonymized data out for widespread use. They got a bunch of comments.

There's controversy as to how de-identified data can actually be and still be useful. There is going to be, probably, a three-to-five year process where they rewrite the Common Rule and it'll be more modern. No one knows how modern, but it will be at least more modern when that finishes.

Then there's another piece in the U.S. — HIPAA — which creates a totally separate regime. In some ways, it is the same as the Common

Rule, but not always. I don't think that's going to get opened up. The way HIPAA works is that they have 17 direct identifiers that are labeled as identifying information. If you strip those out, it's considered de-identified.

There's an 18th bucket, which is anything else that can reasonably identify people. It's really hard to hit. Right now, your genome is not considered to fall under that. I would be willing to bet within a year or two, it will be.

From a regulatory perspective, you've got these overlapping regimes that don't quite fit and both of them are moving targets. That creates a lot of uncertainty from an investment perspective or from an analytics perspective.

How are you thinking about a "health data commons," in terms of weighing potential risks against potential social good?

John Wilbanks: I think that that's a personal judgment as to the risk-benefit decision. Part of the difficulty is that the regulations are very syntactic — "This is what re-identification is" — whereas the concept of harm, benefit, or risk is actually something that's deeply personal. If you are sick, if you have cancer or a rare disease, you have a very different idea of what risk is compared to somebody who thinks of him or herself as healthy.

What we see — and this is born out in the Framingham Heart Study and all sorts of other longitudinal surveys — is that people's attitudes toward risk and benefit change depending on their circumstances. Their own context really affects what they think is risky and what they think isn't risky.

I believe that the early data donors are likely to be people for whom there isn't a lot of risk perceived because the health system already knows that they're sick. The health system is already denying them coverage, denying their requests for PET scans, denying their requests for access to care. That's based on actuarial tables, not on their personal data. It's based on their medical history.

If you're in that group of people, then the perceived risk is actually pretty low compared to the idea that your data might actually get used or to the idea that you're no longer passive. Even if it's just a donation, you're doing something outside of the system that's accelerating the odds of getting something discovered. I think that's the natural group.

If you think back to the numbers of users who are required to create free software or Wikipedia, to create a cultural commons, a very low percentage is needed to create a useful resource.

Depending on who you talk to, somewhere between 5-10% of all Americans either have a rare disease, have it in their first order family, or have a friend with a rare disease. Each individual disease might not have very many people suffering from it, but if you net them all up, it's a lot of people. Getting several hundred thousand to a few million people enrolled is not an outrageous idea.

When you look at the existing examples of where such commons have come together, what have been the most important concrete positive outcomes for society?

John Wilbanks: I don't think we have really even started to see them because most people don't have computable data about themselves. Most people, if they have any data about themselves, have scans of their medical records.

What we really know is that there's an opportunity cost to *not* trying, which is that the existing system is really inefficient, very bad at discovering drugs, and very bad at getting those drugs to market in a timely basis.

That's one of the reasons we're doing this is as an experiment. We would like to see exactly how effective big computational approaches are on health data. The problem is that there are two ways to get there.

One is through a set of monopoly companies coming together and working together. That's how semiconductors work. The other is through an open network approach. There's not a lot of evidence that things besides these two approaches work. Government intervention is probably not going to work.

Obviously, I come down on the open network side. But there's an implicit belief, I think, both in the people who are pushing the cooperating monopolies approach and the people who are pushing the open networks approach, that there's enormous power in the big-data-driven approach. We're just leaving that on the table right now by not having enough data aggregated.

The benefits to health that will come out will be the ability to increasingly, by looking at a multidimensional picture of a person, predict

with some confidence whether or not a drug will work, or whether they're going to get sick, or how sick they're going to get, or what lifestyle changes they can make to mitigate an illness. Right now, basically, we really don't know very much.

Esther Dyson on Health Data, "Preemptive Healthcare," and the Next Big Thing

By Alex Howard

If we look ahead to the next decade, it's worth wondering whether the way we think about health and health care will have shifted. Will health care technology be a panacea? Will it drive even higher costs, creating a broader divide between digital haves and have-nots? Will opening health data empower patients or empower companies?

As ever, there will be good outcomes and bad outcomes, and not just in the medical sense. There's a great deal of thought around the potential for mobile applications right now, from the FDA's potential decision to regulate them to a reported high abandonment rate. There are also significant questions about privacy, patient empowerment, and meaningful use of electronic health care records.

When I've talked to US CTO Todd Park or Dr. Farzad Mostashari they've been excited about the prospect for health data to fuel better dashboards and algorithms to give frontline caregivers access to critical information about people they're looking after, providing critical insight at the point of contact.

Kathleen Sebelius, the U.S. Secretary for Health and Human Services, said at this year's Health Datapalooza that venture capital investment in the health care IT area is up 60% since 2009.

Given that context, I was more than a little curious to hear what Esther Dyson (@edyson) is thinking about when she looks at the intersection of health care, data, and information technology.

Dyson, who started her career as a journalist, is now an angel investor and philanthropist. Dyson is a strong supporter of "preemptive health care" — and she's putting her money where her interest lies, with her investments.

Our interview, which was lightly edited for content and clarity, follows.

How do you see health care changing?

Dyson: There are multiple perspectives. The one I have does not invalidate others, nor it is intended to trump the others, but it's the one that I focus on — and that's "health" as opposed to "health care."

If you maintain good health, you can avoid health care. That's one of those great and unrealizable goals, but it's realizable in part. Any health care you can avoid because you're healthy is valuable.

What I'm mostly focused on is trying to change people's behavior. You'll get agreement from almost everybody that eating right, not smoking, getting exercise, avoiding too much stress, and sleeping a lot are good for your health.

The challenge is what makes people do those things, and that's where there's a real lack of data. So a lot of what I'm doing is investing in that space. There's evidence-based medicine. There's also evidence-based prevention, and that's even harder to validate.

Right now, a lot of people are doing a lot of different things. Many of them are collecting data, which over time, with luck, will prove that some of these things I'm going to talk about are valuable.

What does the landscape for health care products and services look like to you today?

Dyson: I see three markets.

There's the **traditional health care market**, which is what people usually talk about. It's drugs, clinics, hospitals, doctors, therapies, devices, insurance companies, data processors, or electronic health records.

Then there's the **market for bad health**, which people don't talk about a lot, at least not in those terms, but it's huge. It's the products and all of the advertising around everything from sugared soft drinks to cigarettes to recreational drugs to things that keep you from going to bed, going to sleep, keep you on the couch, and keep you immobile. Look at cigarettes and alcohol: That's a huge market. People are being encouraged to engage in unhealthy behaviors, whether it's stuff that might be healthy in moderation or stuff that just isn't healthy at all.

The new [third] market for health existed already as health clubs. What's exciting is that there's now an explicit **market for things that are designed to change your behavior**. Usually, they're information- and social-based. These are the quantified self — analytical tools, tools for sharing, tools for fostering collaboration or competition with people that behave in a healthy way. Most of those have very little data to

back them up. The business models are still not too clear, because if I'm healthy, who's going to pay for that? The chances are that if I'll pay for it, I'm already kind of a health nut and don't need it as much as someone who isn't.

Pharma companies will pay for some such things, especially if they think they can sell people drugs in conjunction with them. I'll sell you a cholesterol-lowering drug through a service that encourages you to exercise, for example. That's a nice market. You go to the pre-diabetics and you sell them your statin. Various vendors of sports clubs and so forth will fund this. But over time, I expect you're going to see employers realize the value of this, then finally, long-term insurance companies and perhaps government. But it's a market that operates mostly on faith at this point.

Speaking of faith, Rock Health shared data that around 80% of mobile health apps are being abandoned by consumers after two weeks. Thoughts?

Dyson: To me, that's infant mortality. The challenge is to take the 20% and then make those persist. But you're right, people try a lot of stuff and it turns out to be confusing and not well-designed, et cetera.

If you look ahead a decade, what are the big barriers for health data and mobile technology playing a beneficial role, as opposed to a more dystopian one?

Dyson: Well, the benign version is we've done a lot of experimentation. We've discovered that most apps have an 80% abandon rate, but the 20% that are persisting get better and better and better. So the 80% that are abandoned vanish and the marketplace and the vendors focus on the 20%. And we get broad adoption. You get onto the subway in New York and everybody's thin and healthy.

Yeah, that's not going to happen. But there's some impact. Employers understand the value of this. There's a lot more to do than just these [mobile] apps. The employers start serving only healthy food in the cafeteria. Actually, one big sign is going to be what they serve for breakfast at Strata RX. I was at the Kauffman Life Sciences Entrepreneur Conference and they had muffins, bagels, and cream cheese.

Carbohydrates and fat, in other words.

Dyson: And sugar-filled yogurts. That was the first day. They responded to somebody's tweet [the second day] and it was better. But it's not just the advertising. It's the selection of stuff that you get when you go to these events or when you go to a hotel or you go to school or you go to your cafeteria at your office.

Defaults are tremendously important. That's why I'm a big fan of what [Michael] Bloomberg is trying to do in New York. If you really want to buy two servings of soda, that's fine, but the default serving should be one. All of this stuff really does have an impact.

Ten years from now, evidence has shown what works. What works is working because people are doing it. A lot of this is that social norms have changed. The early adopters have adopted, the late adopters are being carried along in the wake — just like there are still people who smoke, but it's no longer the norm.

Do you have concerns or hopes for the risks and rewards of open health data releases?

Dyson: If we have a sensible health care system, the data will be helpful. Hospitals will say, "Oh my God, this guy's at-risk, let's prevent him from getting sick." Hospitals and the payers will know, "If we let this guy get sick, it's going to cost us a lot more in the long run. And we actually have a business model that operates long-term rather than simply tries to minimize cost in the short-term."

And insurance companies will say, "I'm paying for this guy. I better keep him healthy." So the most important thing is for us to have a system that works long-term like that.

What role will personal data ownership play in the health care system of the future?

Dyson: Well, first we have to define what it is. From my point-of-view, you own your own data. On the other hand, if you want care, you've got to share it.

I think people are way too paranoid about their data. There will, inevitably, be data spills. We should try to avoid them, but we should also not encourage paranoia. If you have a rational economic system, privacy will be an issue, but financial security will not. Those two have gotten mingled in people's minds.

Yes, I may just want to keep it quiet that I have a sexually transmitted disease, but it's not going to affect my ability to get treatment or to get

insurance if I've got it. On the other hand, if I have to pay a little more for my diet soda or my hamburger because it's being taxed, I don't think that's such a bad idea. Not that I want somebody recording how many hamburgers I eat, just tax them — but you don't need to tax me personally: tax the hamburger.

What about the potential for the quantified self-movement to someday reveal that hamburger consumption to insurers?

Dyson: People are paranoid about insurers, but they're too busy. They're not tracking the hamburgers you eat. They're insuring populations. I went to get insurance and I told Aetna, "You can have my genetic profile." And they said, "We wouldn't know what to do with it." I'm not saying that [tracking is] entirely impossible, but I really think people obsess too much about this kind of stuff.

How should — or could — startups in health care be differentiating themselves? What are the big problems they could be working on solving?

Dyson: There's the whole social aspect. How do you design a game, a social interaction, that encourages people to react the way you want them to react? It's like the difference between Facebook and Friendster. They both had the same potential user base. One was successful; one wasn't. It's the quality of the analytics you show individuals about their behavior. It's the narratives, the tools and the affordances that you give them for interacting with their friends.

For what it's worth, of the hundreds of companies that Rock Health or anybody else will tell you about, probably a third of them will disappear. One tenth will be highly successful and will acquire the remaining 57%.

What are the health care startup models that interest you? Why?

Dyson: I don't think there's a single one. There's bunches of them occupying different places.

One area I really like is user-generated research and experiments. Obviously, there's 23andMe.[1] **Deep analysis** of your own data and the option to share it with other people and with researchers. **User-generated data science research** is really fascinating.

1. Dyson is an investor in 23andMe.

And then **social affordance**, like HealthRally, where people interact with each other. Omada Health — which I'm an investor in — is a Rock Health company that says we can't do it all ourselves — there's a designated counselor for a group. Right now it's focused on pre-diabetics.

I love that, partly because I think it's going to be effective, and partly because I really like it as an employment model. I think our country is too focused on manufacturing and there's a way to turn more people into health counselors. I'd take all of the laid off auto workers and turn them into gym teachers, and all the laid off engineers and turn them into data scientists or people developing health apps. Or something like that.

What's the biggest myth in the health data world? What's the thing that drives you up the wall, so to speak?

Dyson: The biggest myth is that any single thing is the solution. The biggest need is for long-term thinking, which is everything from an individual thinking long-term about the impact of behavior to a financial institution thinking long-term and having the incentive to think long-term.

Individuals need to be influenced by psychology. Institutions, and the individuals in them, are employees that can be motivated or not. As an institution, they need financial incentives that are aligned with the long-term rather than the short-term.

That, again, goes back to having a vested interest in the health of people rather than in the cost of care.

Employers, to some extent, have that already. Your employer wants you to be healthy. They want you to show up for work, be cheerful, motivated and well rested. They get a benefit from you being healthy, far beyond simply avoiding the cost of your care.

Whereas the insurance companies, at this point, simply pass it through. If the insurance company is too effective, they actually have to lower their premiums, which is crazy. It's really not insurance: it's a cost-sharing and administration role that the insurance companies play. That's something a lot of people don't get. That needs to be fixed, one way or another.

A Marriage of Data and Caregivers Gives Dr. Atul Gawande Hope for Health Care

By Alex Howard

Dr. Atul Gawande (@Atul_Gawande) has been a bard in the health care world, straddling medicine, academia and the humanities as a practicing surgeon, medical school professor, best-selling author, and staff writer at the *New Yorker* magazine. His long-form narratives and books have helped illuminate complex systems and wicked problems to a broad audience.

One recent feature that continues to resonate for those who wish to apply data to the public good is Gawande's *New Yorker* piece "The Hot Spotters," where Gawande considered whether health data could help lower medical costs by giving the neediest patients better care. That story brings home the challenges of providing health care in a city, from cultural change to gathering data to applying it.

This summer, after meeting Gawande at the 2012 Health DataPalooza, I interviewed him about hot spotting, predictive analytics, networked transparency, health data, feedback loops, and the problems that technology won't solve. Our interview, lightly edited for content and clarity, follows.

Given what you've learned in Camden, N.J. — the backdrop for your piece on hot spotting — do you feel hot spotting is an effective way for cities and people involved in public health to proceed?

Gawande: The short answer, I think, is "yes."

Here we have this major problem of both cost and quality — and we have signs that some of the best places that seem to do the best jobs can be among the least expensive. How you become one of those places is a kind of mystery.

It really parallels what happened in the police world. Here is something that we thought was an impossible problem: crime. Who could possibly lower crime? One of the ways we got a handle on it was by directing policing to the places where there was the most crime. It sounds kind of obvious, but it was not apparent that crime is concentrated and that medical costs are concentrated.

The second thing I knew but hadn't put two and two together about is that the sickest people get the worst care in the system. People with complex illness just don't fit into 20-minute office visits.

The work in Camden was emblematic of work happening in pockets all around the country where you prioritize. As soon as you look at the system, you see hundreds, thousands of things that don't work properly in medicine. But when you prioritize by saying, "For the sickest people — the 5% who account for half of the spending — let's look at what their $100,000 moments are," you then understand it's strengthening primary care and it's the ability to manage chronic illness.

It's looking at a few acute high-cost, high-failure areas of care, such as how heart attacks and congestive heart failure are managed in the system; looking at how renal disease patients are cared for; or looking at a few things in the commercial population, like back pain, being a huge source of expense. And then also end-of-life care.

With a few projects, it became more apparent to me that you genuinely could transform the system. You could begin to move people from depending on the most expensive places where they get the least care to places where you actually are helping people achieve goals of care in the most humane and least wasteful ways possible.

The data analytics office in New York City is doing fascinating predictive analytics. That approach could have transformative applications in health care, but it's notable how careful city officials have been about publishing certain aspects of the data. How do you think about the relative risks and rewards here, including balancing social good with the need to protect people's personal health data?

Gawande: Privacy concerns can sometimes be a barrier, but I haven't seen it be *the* major barrier here. There are privacy concerns in the data about households as well in the police data.

The reason it works well for the police is not just because you have a bunch of data geeks who are poking at the data and finding interesting things. It's because they're paired with people who are responsible for responding to crime, and above all, reducing crime. The commanders who have the responsibility have a relationship with the people who have the data. They're looking at their population saying, "What are we doing to make the system better?"

That's what's been missing in health care. We have not married the people who have the data with people who feel responsible for ach-

ieving better results at lower costs. When you put those people together, they're usually within a system, and within a system, there is no privacy barrier to being able to look and say, "Here's what we can be doing in this health system," because it's often that particular.

The beautiful aspect of the work in New York is that it's not at a terribly abstract level. Yes, they're abstracting the data, but they're also helping the police understand: "It's this block that's the problem. It's shifted in the last month into this new sector. The pattern of the crime is that it looks more like we have a problem with domestic violence. Here are a few more patterns that might give you a clue about what you can go in and do." There's this give and take about what can be produced and achieved.

That, to me, is the gold in the health care world — the ability to peer in and say: "Here are your most expensive patients and your sickest patients. You didn't know it, but here, there's an alcohol and drug addiction issue. These folks are having car accidents and major trauma and turning up in the emergency rooms and then being admitted with $12,000 injuries."

That's a system that could be improved and, lo and behold, there's an intervention here that's worked before to slot these folks into treatment programs, which by and large, we don't do at all.

That sense of using the data to help you solve problems requires two things. It requires data geeks and it requires the people in a system who feel responsible, the way that Bill Bratton made commanders feel responsible in the New York police system for the rate of crime. We haven't had physicians who felt that they were responsible for 10,000 ICU patients and how well they do on everything from the cost to how long they spend in the ICU.

Health data is creating opportunities for more transparency into outcomes, treatments, and performance. As a practicing physician, do you welcome the additional scrutiny that such collective intelligence provides, or does it concern you?

Gawande: I think that transparency of our data is crucial. I'm not sure that I'm with the majority of my colleagues on this. The concerns are that the data can be inaccurate, that you can overestimate or underestimate the sickness of the people coming in to see you, and that my patients aren't like your patients.

That said, I have no idea who gets better results at the kinds of operations I do and who doesn't. I do know who has high reputations and who has low reputations, but it doesn't necessarily correspond to the kinds of results they get. As long as we are not willing to open up data to let people see what the results are, we will never actually learn.

The experience of what happens in fields where the data is open is that it's the practitioners themselves that use it. I'll give a couple of examples. Mortality for childbirth in hospitals has been available for a century. It's been public information, and the practitioners in that field have used that data to drive the death rates for infants and mothers down from the biggest killer in people's lives for women of childbearing age and for newborns into a rarity.

Another field that has been able to do this is cystic fibrosis. They had data for 40 years on the performance of the centers around the country that take care of kids with cystic fibrosis. They shared the data privately. They did not tell centers how the other centers were doing. They just told you where you stood relative to everybody else and they didn't make that information public. About four or five years ago, they began making that information public. It's now available on the Internet. You can see the rating of every center in the country for cystic fibrosis.

Several of the centers had said, "We're going to pull out because this isn't fair." Nobody ended up pulling out. They did not lose patients in hoards and go bankrupt unfairly. They were able to see from one another who was doing well and then go visit and learn from one and other.

I can't tell you how fundamental this is. There needs to be transparency about our costs and transparency about the kinds of results. It's murky data. It's full of lots of caveats. And yes, there will be the occasional journalist who will use it incorrectly. People will misinterpret the data. But the broad result, the net result of having it out there, is so much better for everybody involved that it far outweighs the value of closing it up.

U.S. officials are trying to apply health data to improve outcomes, reduce costs and stimulate economic activity. As you look at the successes and failures of these sorts of health data initiatives, what do you think is working and why?

Gawande: I get to watch from the sidelines, and I was lucky to participate in Datapalooza this year. I mostly see that it seems to be following a mode that's worked in many other fields, which is that there's a fundamental role for government to be able to make data available.

When you work in complex systems that involve multiple people who have to, in health care, deal with patients at different points in time, no one sees the net result. So, no one has any idea of what the actual experience is for patients. The open data initiative, I think, has innovative people grabbing the data and showing what you can do with it.

Connecting the data to the physical world is where the cool stuff starts to happen. What are the kinds of costs to run the system? How do I get people to the right place at the right time? I think we're still in primitive days, but we're only two or three years into starting to make something more than just data on bills available in the system. Even that wasn't widely available — and it usually was old data and not very relevant to this moment in time.

My concern all along is that data needs to be meaningful to both the patient and the clinician. It needs to be able to connect the abstract world of data to the physical world of what really happens, which means it has to be timely data. A six-month turnaround on data is not great. Part of what has made Wal-Mart powerful, for example, is they took retail operations from checking their inventory once a month to checking it once a week and then once a day and then in real-time, knowing exactly what's on the shelves and what's not.

That equivalent is what we'll have to arrive at if we're to make our systems work. Timeliness, I think, is one of the under-recognized but fundamentally powerful aspects because we sometimes over prioritize the comprehensiveness of data and then it's a year old, which doesn't make it all that useful. Having data that tells you something that happened this week, that's transformative.

Are you using an iPad at work?

Gawande: I do use the iPad here and there, but it's not readily part of the way I can manage the clinic. I would have to put in a lot of effort for me to make it actually useful in my clinic.

For example, I need to be able to switch between radiology scans and past records. I predominantly see cancer patients, so they'll have 40 pages of records that I need to have in front of me, from scans to lab tests to previous notes by other folks.

I haven't found a better way than paper, honestly. I can flip between screens on my iPad, but it's too slow and distracting, and it doesn't let me talk to the patient. It's fun if I can pull up a screen image of this or that and show it to the patient, but it just isn't that integrated into practice.

What problems are immune to technological innovation? What will need to be changed by behavior?

Gawande: At some level, we're trying to define what great care is. Great care means being able to provide optimally knowledgeable care in the right time and the right way for people and not wasting resources.

Some of it's crucially aided by information technology that connects information to where it needs to be so that good decision-making happens, both by patients and by the clinicians who work with them.

If you're going to be able to make health care work better, you've got to be able to make that system work better for people, more efficiently and less wastefully, less harmfully and with much better teamwork. I think that information technology is a tool in that, but fundamentally you're talking about making teams that can go from being disconnected cowboys in care to pit crews that actually work together toward solving a problem.

In a football team or a pit crew, technology is really helpful, but it's only a tiny part of what makes that team great. What makes the team great is that they know what they're aiming to do, they're very clear about their goals, and they are able to make sure they execute every basic thing that's crucial for that success.

What do you worry about in this surge of interest in more data-driven approaches to medicine?

Gawande: I worry the most about a disconnect between the people who have to use the information and technology and tools, and the people who make them. We see this in the consumer world. Fundamentally, there is not a single [health] application that is remotely like my iPod, which is instantly usable. There are a gazillion number of ways in which information would make a huge amount of difference.

That sense of being able to understand the world of the user, the task that's accomplished and the complexity of what they have to do, and

connecting that to the people making the technology — there just aren't that many lines of marriage. In many of the companies that have some of the dominant systems out there, I don't see signs that that's necessarily going to get any better.

If people gain access to better information about the consequences of various choices, will that lead to improved outcomes and quality of life?

Gawande: That's where the art comes in. There are problems because you lack information, but when you have information like "you shouldn't drink three cans of Coke a day — you're going to put on weight," then having that information is not sufficient for most people.

Understanding what is sufficient to be able to either change the care or change the behaviors that we're concerned about is the crux of what we're trying to figure out and discover.

When the information is presented in a really interesting way, people have gradually discovered — for example, having a little ball on your dashboard that tells you when you're accelerating too fast and burning off extra fuel — how that begins to change the actual behavior of the person in the car.

No amount of presenting the information that you ought to be driving in a more environmentally friendly way ends up changing anything. It turns out that change requires the psychological nuance of presenting the information in a way that provokes the desire to actually do it.

We're at the very beginning of understanding these things. There's also the same sorts of issues with clinician behavior — not just information, but how you are able to foster clinicians to actually talk to one another and coordinate when five different people are involved in the care of a patient and they need to get on the same page.

That's why I'm fascinated by the police work, because you have the data people, but they're married to commanders who have responsibility and feel responsibility for looking out on their populations and saying, "What do we do to reduce the crime here? Here's the kind of information that would really help me." And the data people come back to them and say, "Why don't you try this? I'll bet this will help you."

It's that give and take that ends up being very powerful.

Five Elements of Reform that Health Providers Would Rather Not Hear About

By Andy Oram

The quantum leap we need in patient care requires a complete overhaul of record-keeping and health IT. Leaders of the health care field know this and have been urging the changes on health care providers for years, but the providers are having trouble accepting the changes for several reasons.

What's holding them back? Change certainly costs money, but the industry is already groaning its way through enormous paradigm shifts to meet current financial and regulatory climates, so the money might as well be directed toward things that work. Training staff to handle patients differently is also difficult, but the staff on the floor of these institutions are experiencing burn-out and can be inspired by a new direction. The fundamental resistance seems to be expectations by health providers and their vendors about the control they need to conduct their business profitably.

A few months ago I wrote an article titled "Five Tough Lessons I Had to Learn About Health Care." Here I'll delineate some elements of a new health care system that are promoted by thought leaders, that echo the evolution of other industries, that will seem utterly natural in a couple decades — but that providers are loathe to consider. I feel that leaders in the field are not confronting that resistance with an equivalent sense of conviction that these changes are crucial.

1. Reform Will Not Succeed Unless Electronic Records Standardize on a Common, Robust Format

Records are not static. They must be combined, parsed, and analyzed to be useful. In the health care field, records must travel with the patient. Furthermore, we need an explosion of data analysis applications in order to drive diagnosis, public health planning, and research into new treatments.

Interoperability is a common mantra these days in talking about electronic health records, but I don't think the power and urgency of record formats can be conveyed in eight-syllable words. It can be conveyed better by a site that uses data about hospital procedures, costs, and patient satisfaction to help consumers choose a desirable hospital. Or an app that might prevent a million heart attacks and strokes.

Data-wise (or data-ignorant), doctors are stuck in the 1980s, buying proprietary record systems that don't work together even between different departments in a hospital, or between outpatient clinics and their affiliated hospitals. Now the vendors are responding to pressures from both government and the market by promising interoperability. The federal government has taken this promise as good coin, hoping that vendors will provide windows onto their data. It never really happens. Every baby step toward opening up one field or another requires additional payments to vendors or consultants.

That's why exchanging patient data (health information exchange — HIE) requires a multi-million-dollar investment, year after year, and why most HIEs go under. And that's why the HL7 committee, putatively responsible for defining standards for electronic health records (EHR), keeps on putting out new, complicated variations on a long history of formats that were not well-enough defined to ensure compatibility among vendors.

The Direct Project and perhaps the nascent RHEx RESTful exchange standard will let hospitals exchange the limited types of information that the government forces them to exchange. But it won't create a platform (as suggested in this PDF slideshow) for the hundreds of applications we need to extract useful data from records. Nor will it open the records to the masses of data we need to start collecting. It remains to be seen whether Accountable Care Organizations (ACO), which are the latest reform in U.S. health care and are described in this video, will be able to use current standards to exchange the data that each member institution needs to coordinate care. Shahid Shaw has laid out in glorious detail the elements of open data exchange in health care.

2. Reform Will Not Succeed Unless Massive Amounts of Patient Data Are Collected

We aren't giving patients the most effective treatments because we just don't know enough about what works. This extends throughout the health care system:

- We can't prescribe a drug tailored to the patient because we don't collect enough data about patients and their reactions to the drug.
- We can't be sure drugs are safe and effective because we don't collect data about how patients fare on those drugs.
- We don't see a heart attack or other crisis coming because we don't track the vital signs of at-risk populations on a daily basis.

- We don't make sure patients follow through on treatment plans because we don't track whether they take their medications and perform their exercises.
- We don't target people who need treatment because we don't keep track of their risk factors.

Some institutions have adopted a holistic approach to health, but as a society there's a huge amount more that we could do in this area.

Leaders in the field know what health care providers could accomplish with data. A recent article even advises policy makers to focus on the data instead of the electronic records. The question is whether providers are technically and organizationally prepped to accept it in such quantities and variety. When doctors and hospitals think they own the patients' records, they resist putting in anything but their own notes and observations, along with lab results they order. We've got to change the concept of ownership, which strikes deep into their culture.

3. Reform Will Not Succeed Unless Patients Are in Charge of Their Records

Doctors are currently acting in isolation, occasionally consulting with the other providers seen by their patients but rarely sharing detailed information. It falls on the patient, or a family advocate, to remember that one drug or treatment interferes with another or to remind treatment centers of follow-up plans. And any data collected by the patient remains confined to scribbled notes or (in the modern Quantified Self equivalent) a website that's disconnected from the official records.

Doctors don't trust patients. They have some good reasons for this: medical records are complicated documents in which a slight rewording or typographical error can change the meaning enough to risk a life. But walling off patients from records doesn't insulate them against errors: on the contrary, patients catch errors entered by staff all the time. So ultimately it's better to bring the patient onto the team and educate her. If a problem with records altered by patients — deliberately or through accidental misuse — turns up down the line, digital certificates can be deployed to sign doctor records and output from devices.

The amounts of data we're talking about get really big fast. Genomic information and radiological images, in particular, can occupy dozens of gigabytes of space. But hospitals are moving to the cloud anyway.

Practice Fusion just announced that they serve 150,000 medical practitioners and that "One in four doctors selecting an EHR today chooses Practice Fusion." So we can just hand over the keys to the patients and storage will grow along with need.

The movement for patient empowerment will take off, as experts in health reform told U.S. government representatives, when patients are in charge of their records. To treat people, doctors will have to ask for the records, and the patients can offer the full range of treatment histories, vital signs, and observations of daily living they've collected. Applications will arise that can search the data for patterns and relevant facts.

Once again, the U.S. government is trying to stimulate patient empowerment by requiring doctors to open their records to patients. But most institutions meet the formal requirements by providing portals that patients can log into, the way we can view flight reservations on airlines. We need the patients to become the pilots. We also need to give them the information they need to navigate.

4. Reform Will Not Succeed Unless Providers Conform to Practice Guidelines

Now that the government is forcing doctors to release information about outcomes, patients can start to choose doctors and hospitals that offer the best chances of success. The providers will have to apply more rigor to their activities, using checklists and more, to bring up the scores of the less successful providers. Medicine is both a science and an art, but many lag on the science — that is, doing what has been statistically proven to produce the best likely outcome — even at prestigious institutions.

Patient choice is restricted by arbitrary insurance rules, unfortunately. These also contribute to the utterly crazy difficulty determining what a medical procedure will cost as reported by e-Patient Dave and WBUR radio. Straightening out this problem goes way beyond the doctors and hospitals, and settling on a fair, predictable cost structure will benefit them almost as much as patients and taxpayers. Even some insurers have started to see that the system is reaching a dead-end and they are erecting new payment mechanisms.

5. Reform Will Not Succeed Unless Providers and Patients Can Form Partnerships

I'm always talking about technologies and data in my articles, but none of that constitutes health. Just as student testing is a poor model for education, data collection is a poor model for medical care. What patients want is time to talk intensively with their providers about their needs, and providers voice the same desires.

Data and good record keeping can help us use our resources more efficiently and deal with the physician shortage, partly by spreading out jobs among other clinical staff. Computer systems can't deal with complex and overlapping syndromes, or persuade patients to adopt practices that are good for them. Relationships will always have to be in the forefront. Health IT expert Fred Trotter says, "Time is the gas that makes the relationship go, but the technology should be focused on fuel efficiency."

Arien Malec, former contractor for the Office of the National Coordinator, used to give a speech about the evolution of medical care. Before the revolution in antibiotics, doctors had few tools to actually cure patients, but they live with the patients in the same community and know their needs through and through. As we've improved the science of medicine, we've lost that personal connection. Malec argued that better records could help doctors really know their patients again. But conversations are necessary too.