2.4.1: Business Intelligence and Data Science

First from the Economist, *A different game* (2010, http://www.economist.com/node/15557465)

> "In 1879 James Ritty, a saloon-keeper in Dayton, Ohio, received a patent for a wooden contraption that he dubbed the "incorruptible cashier". With a set of buttons and a loud bell, the device, sold by National Cash Register (NCR), was little more than a simple adding machine. Yet as an early form of managing information flows in American business the cash register had a huge impact. It not only reduced pilferage by alerting the shopkeeper when the till was opened; by recording every transaction, it also provided an instant overview of what was happening in the business.
>
> Sales data remain one of a company's most important assets. In 2004 Wal-Mart peered into its mammoth databases and noticed that before a hurricane struck, there was a run on flashlights and batteries, as might be expected; but also on Pop-Tarts, a sugary American breakfast snack. On reflection it is clear that the snack would be a handy thing to eat in a blackout, but the retailer would not have thought to stock up on it before a storm. The company whose system crunched Wal-Mart's numbers was none other than NCR and its data-warehousing unit, Teradata, now an independent firm.
>
> A few years ago such technologies, called "business intelligence", were available only to the world's biggest companies. But as the price of computing and storage has fallen and the software systems have got better and cheaper, the technology has moved into the mainstream. Companies are collecting more data than ever before. In the past they were kept in different systems that were unable to talk to each other, such as finance, human resources or customer management. Now the systems are being linked, and companies are using data-mining techniques to get a complete picture of their operations—"a single version of the truth", as the industry likes to call it. That allows firms to operate more efficiently, pick out trends and improve their forecasting.
>
> Consider Cablecom, a Swiss telecoms operator. It has reduced customer defections from one-fifth of subscribers a year to under 5% by crunching its numbers. Its software spotted that although customer defections peaked in the 13th month, the decision to leave was made much earlier, around the ninth month (as indicated by things like the number of calls to customer support services). So Cablecom offered certain customers special deals seven months into their subscription and reaped the rewards.
>
> Such data-mining has a dubious reputation. "Torture the data long enough and they will confess to anything," statisticians quip. But it has become far more effective as more companies have started to use the technology. Best Buy, a retailer, found that 7% of its customers accounted for 43% of its sales, so it reorganised its stores to concentrate on those customers' needs. Airline yield management improved because analytical techniques uncovered the best predictor that a passenger would actually catch a flight he had booked: that he had ordered a vegetarian meal.
>
> The IT industry is piling into business intelligence, seeing it as a natural successor of services such as accountancy and computing in the first and second half of the 20th

century respectively. Accenture, PricewaterhouseCoopers, IBM and SAP are investing heavily in their consulting practices. Technology vendors such as Oracle, Informatica, TIBCO, SAS and EMC have benefited. IBM believes business intelligence will be a pillar of its growth as sensors are used to manage things from a city's traffic flow to a patient's blood flow. It has invested $12 billion in the past four years and is opening six analytics centres with 4,000 employees worldwide.

Analytics—performing statistical operations for forecasting or uncovering correlations such as between Pop-Tarts and hurricanes—can have a big pay-off. In Britain the Royal Shakespeare Company (RSC) sifted through seven years of sales data for a marketing campaign that increased regular visitors by 70%. By examining more than 2m transaction records, the RSC discovered a lot more about its best customers: not just income, but things like occupation and family status, which allowed it to target its marketing more precisely. That was of crucial importance, says the RSC's Mary Butlin, because it substantially boosted membership as well as fund-raising revenue.

Yet making the most of data is not easy. The first step is to improve the accuracy of the information. Nestlé, for example, sells more than 100,000 products in 200 countries, using 550,000 suppliers, but it was not using its huge buying power effectively because its databases were a mess. On examination, it found that of its 9m records of vendors, customers and materials around half were obsolete or duplicated, and of the remainder about one-third were inaccurate or incomplete. The name of a vendor might be abbreviated in one record but spelled out in another, leading to double-counting.

Over the past ten years Nestlé has been overhauling its IT system, using SAP software, and improving the quality of its data. This enabled the firm to become more efficient, says Chris Johnson, who led the initiative. For just one ingredient, vanilla, its American operation was able to reduce the number of specifications and use fewer suppliers, saving $30m a year. Overall, such operational improvements save more than $1 billion annually.

Nestlé is not alone in having problems with its database. Most CIOs admit that their data are of poor quality. In a study by IBM half the managers quizzed did not trust the information on which they had to base decisions. Many say that the technology meant to make sense of it often just produces more data. Instead of finding a needle in the haystack, they are making more hay.

Still, as analytical techniques become more widespread, business decisions will increasingly be made, or at least corroborated, on the basis of computer algorithms rather than individual hunches. This creates a need for managers who are comfortable with data, but statistics courses in business schools are not popular.

Many new business insights come from “dead data”: stored information about past transactions that are examined to reveal hidden correlations. But now companies are increasingly moving to analysing real-time information flows.

Wal-Mart is a good example. The retailer operates 8,400 stores worldwide, has more than 2m employees and handles over 200m customer transactions each week. Its

revenue last year, around $400 billion, is more than the GDP of many entire countries. The sheer scale of the data is a challenge, admits Rollin Ford, the CIO at Wal-Mart's headquarters in Bentonville, Arkansas. “We keep a healthy paranoia.”

Wal-Mart's inventory-management system, called Retail Link, enables suppliers to see the exact number of their products on every shelf of every store at that precise moment. The system shows the rate of sales by the hour, by the day, over the past year and more. Begun in the 1990s, Retail Link gives suppliers a complete overview of when and how their products are selling, and with what other products in the shopping cart. This lets suppliers manage their stocks better.

The technology enabled Wal-Mart to change the business model of retailing. In some cases it leaves stock management in the hands of its suppliers and does not take ownership of the products until the moment they are sold. This allows it to shed inventory risk and reduce its costs. In essence, the shelves in its shops are a highly efficiently managed depot.

…

Two technology trends are helping to fuel these new uses of data: cloud computing and open-source software. Cloud computing—in which the internet is used as a platform to collect, store and process data—allows businesses to lease computing power as and when they need it, rather than having to buy expensive equipment. Amazon, Google and Microsoft are the most prominent firms to make their massive computing infrastructure available to clients. As more corporate functions, such as human resources or sales, are managed over a network, companies can see patterns across the whole of the business and share their information more easily.

A free programming language called R lets companies examine and present big data sets, and free software called Hadoop now allows ordinary PCs to analyse huge quantities of data that previously required a supercomputer. It does this by parcelling out the tasks to numerous computers at once. This saves time and money. For example, the New York Times a few years ago used cloud computing and Hadoop to convert over 400,000 scanned images from its archives, from 1851 to 1922. By harnessing the power of hundreds of computers, it was able to do the job in 36 hours.

Visa, a credit-card company, in a recent trial with Hadoop crunched two years of test records, or 73 billion transactions, amounting to 36 terabytes of data. The processing time fell from one month with traditional methods to a mere 13 minutes. It is a striking successor of Ritty's incorruptible cashier for a data-driven age.

“Business intelligence” is about companies using data and analytics to make smarter decisions and improve processes. What’s the difference then with business analytics? Let’s look to excerpts from a piece in CIO: *Business Intelligence Versus Business Analytics--What's the Difference?* http://www.cio.com/article/print/18095

“The marketing and analyst airwaves are flooding with speculation about what is next for

business intelligence (BI). What will comprise BI 2.0?

Historically, this market has been served by vendors such as Business Objects and Cognos. But the competitive landscape is changing. Microsoft has now shrewdly entered the market by driving the placement of SQL servers into the space in order to broadly deploy and deliver its BI suite and reporting services in volume. Oracle has seen the effect of companies moving data out of the database to stage it for analysis. The resulting data warehouses have provided a degree of utility in housing, manipulating and delivering "strategic" information across the organization.

Recently though, established vendors such as SAP and Siebel have unveiled BI product suites under the banner of "analytics." SAS, a perennial stalwart of the statistics market, is suddenly being touted as the number-three BI vendor and frequently positions itself as an analytics vendor.

With analytics finding its place within many functions and business processes it seems clear that it will be a defining feature of next generation business intelligence. Particularly, a significant new group of business users—a group I like to call "Go-To Guys"—are in need of analytics tools to tackle daily problems and opportunities. Go-To Guys are the operating managers of company—product managers, sales managers, researchers, engineers and marketers.

So, what is analytics? Neil Raden of Hired Brains, a market research and management consulting firm, has said that, "the proper term for interacting with information at the speed of business, analyzing and discovering and following through with the appropriate action, is 'analytics'." CIOs often assume that business analytics (BA) comes along with BI. The traditional BI market has been associated with providing executive dashboards and reporting to monitor the assumptions and key performance metrics that are part of long term planning cycles.

Everybody wants a dashboard. To the extent that all of us are CEO's of our own business discipline, we want a simple measurement display of how we are doing and an alert mechanism of when something goes wrong. Additionally, dashboards address the growing urgency around Sarbanes Oxley. Monitoring planning assumptions and key performance metrics has now become mission critical from a regulatory and compliance standpoint.

But BI reporting ends with the dashboard, which is sufficient only for some business planning, and BA picks up the rest for the Go-To Guys. Simply, this group must interact with data in a much different way from what traditional BI allows.

The Go-To Guys deal daily in unanticipated outcomes and unknown results and it is their job to mitigate risk and capitalize on opportunities. **BI is not architected to iterate on new scenarios or for immediate response to unanticipated questions because it is set up to automate the distribution of standardized reports that monitor pre-determined key performance metrics and planning assumptions.** BI's answer to analytics has been to deliver the report to the business user and the business user typically takes the data in the report and dumps it into Microsoft Excel in order to do his

own analysis.

As a result, there are $8B (yes, billion) of internally developed analytic applications with Excel as their front end. The BI players treat the output to Excel as a feature. But I actually think it's a tremendous failing. It is proof that you don't get BA when you buy BI. The BI architecture cannot support the operating needs of the business users to ask and answer their own questions in response to new occurrences and events in the marketplace.

Secondly, Excel is not an answer either. **As soon as the data is dumped into Excel, the user is out of the BI system with no way back in.** Any insight that the business user gains while interpreting Excel spreadsheets tends to stay with him—all opportunity for organizational learning or process improvement is lost.
So requirements for analytics are different than the requirements for BI, but the benefits are different as well.

There's also a technical component to all of this reinforcing the claim that the technical requirements to support analytics are different from the technical requirements that enable BI. To facilitate reporting and dashboards, BI traditionally works with aggregated data. **Business users cannot rely solely on aggregated data in the operating environment. They have to be able to get to the details.** The aggregated data will many times obscure the key issue or opportunity in your information.

BI data is typically staged in an OLAP cube to support drill-down. In analytics the Go-To Guys have to be able to get directly to the source data in the database. The key facts needed to make your operating decision are often not in the cube because they haven't been anticipated by the IT department. This is not a question of the trees obscuring the forest—you have to be able to see both. The business users cannot be disconnected from the critical data needed to make a key business decision.

And lastly, **the requirement of the BI system has been to monitor the data based on pre-configured questions requiring only a thin client environment to inform the user.** In the operating world, users need to engage with the information requiring a richer client to support interactivity and the ability to ask and answer their own question without having to go back to IT.

What are the characteristics of an analytic savvy organization? First of all, even the planners want into the act. Analytics is enabling more proactive, high-frequency planning cycles. Planners are better able to refine and iterate the plan, shifting resources to higher performing areas with the goal of being first-to-market and never having a warehouse full of trendy goods once the trend is over. Secondly, the analytically savvy organization is more agile—able to adapt and respond—whether that's to a competitor that releases a new product, a change to the pricing structure in the marketplace or the success of its own marketing campaign.

Remember, you don't get business analytics when you buy business intelligence. The requirements are different and the benefits are different. The return on information and expertise achieved by arming your operating managers with analytics will supercharge

your existing BI investment."

So there's a difference between business intelligence and business analytics. But what about data science? Is it more related to intelligence or analytics, or is it too early to know? Again, does it really matter? While reading this excerpt reflect on how you would describe data science to someone coming from business intelligence or analytics.

An excerpt from *Smart Data Collective*: *Statistics vs. Data Science vs. BI (2013),* *http://smartdatacollective.com/davidmsmith/124376/statistics-vs-data-science-vs-bi*)

> "As someone who trained as a statistician, I've always struggled with that title. I love the rigor and insight that Statistics brings to data analysis, but let's face it: Statistics — the name — has always had a bit of a branding problem. Telling someone I was a statistician was more likely to conjure up images of me counting runs at a baseball (or cricket) game than pursuing serious science. And the image of what Statistics ideally is about — collaborative, interactive, applied, fun — was too often subsumed by the stereotype image — isolated, actuarial, ivory tower, report driven. (And hey, even actuaries can be fun sometimes.)
>
> That's why I'm a fan of the term "data scientist" — it embodies everything that Statistics always should be, without the baggage and tradition of the term "statistician". So I enjoyed participating in yesterday's Kalido webinar "Data Scientist: Your Must-Have Business Investment Now" where I could make the following contrast between the images of Statisticians and Data Scientists:

	Statistician	Data Scientist
Image	Baseball (Cricket)	HBR Sexiest Job of 21st Century
Mode	Reactive	Consultative
Works	Solo	In a team
Inputs	Data File, Hypothesis	A Business Problem
Data	Pre-prepared, clean	Distributed, messy, unstructured
Data Size	Kilobytes	Gigabytes
Tools	SAS, Mainframe	R, Python, awk, Hadoop, Linux, …
Nouns	Tables	Data Visualizations
Focus	Inference (why)	Prediction (what)
Output	Report	Data App / Data Product
Latency	Weeks	Seconds
Stars	G.E.P Box Trevor Hastie	Hilary Mason Nate Silver

(A quick aside on the "Data Size" row above: while the unstructured or unaggregated data source data that data scientist work with can be in the terabytes range or even large, by the time it's cleaned and prepared for statistical modeling, a file in the gigabytes range is even more typical — even at "Big Data" companies like Facebook. This is a topic I cover in more detail in my recent Strata talk on real-time predictive analytics.)

So bottom line: while I am a statistician, and I love Statistics dearly, I do prefer to call myself a Data Scientist today, because it better represents to me what Statistics really is to me (if that makes sense). And that's certainly not to diminish the achievements of those who do call themselves Statistician. In particular, I want to recognize George Box: a true hero of mine, coiner of the idiom "all models are wrong, but some are useful", and one of the nicest people I ever met, who sadly passed away in March.

On the other hand, I have no qualms about making a competitive comparison between

Data Science and Business Intelligence:

	Business Intelligence	Data Science
Perspective	Looking backwards	Looking forwards
Actions	Slice and Dice	Interact
Expertise	Business User	Data Scientist
Data	Warehoused, Siloed	Distributed, real-time
Scope	Unlimited	Specific business question
Questions	What happened?	What will happen? What if?
Output	Table	Answer
Applicability	Historic, possible confounding factors	Future, correcting for influences
Tools	SAP, Cognos, Microstrategy, SAS	Revolution R Enterprise QlikView, Tableau, Jaspersoft
Hot or not?	*So* 1997	Transformational

From this excerpts, think about where you see yourself in this discussion. How is data science similar to and different from like efforts in the past? How does knowing the history of business intelligence shape how you might describe your work to others, particularly those familiar with business intelligence and analytics?