Data Science by AnalyticBridge

Vincent Granville, Ph.D.

Founder, Data Wizard, Managing Partner

www.AnalyticBridge.com - www.DataScienceCentral.com

Download the most recent version at http://bit.ly/oB0zxn

This version was released on 01/03/2013 (123 pages – Gartner's contribution updated)

Previous release was on 06/05/2012 (123 pages)

Published by AnalyticBridge.

Connect with the Author:

- AnalyticBridge – www.analyticbridge.com/profile/VincentGranville
- LinkedIn – www.linkedin.com/in/vincentg
- Facebook – www.facebook.com/analyticbridge
- Twitter – www.twitter.com/analyticbridge
- GooglePlus – https://plus.google.com/116460988472927384512
- Quora – www.quora.com/Vincent-Granville

Content

Introduction

Our Data Science e-Book provides recipes, intriguing discussions and resources for data scientists and executives or decision makers. You don't need an advanced degree to understand the concepts. Most of the material is written in simple English, however it offers simple, better and patentable solutions to many modern business problems, especially about how to leverage big data.

Emphasis is on providing high-level information that executives can easily understand, while being detailed enough so that data scientists can easily implement our proposed solutions. Unlike most other data science books, we do not favor any specific analytic method nor any particular programming language: we stay one level above practical implementations. But we do provide recommendations about which methods to use when necessary.

Most of the material is original, and can be used to develop better systems, derive patents or write scientific articles. We also provide several rules of the thumbs and details about craftsmanship used to avoid traditional pitfalls when working with data sets. The book also contains interviews with analytic leaders, and material about what should be included in a business analytics curriculum, or about how to efficiently optimize a search to fill an analytic position.

Among the more technical contributions, you will find notes on

- How to determine the number of clusters
- How to implement a system to detect plagiarism
- How to build an ad relevancy algorithm
- What is a data dictionary, and how to use it
- Tutorial on how to design successful stock trading strategies
- New fast, safe and efficient random number generator
- How to detect patterns vs. randomness

The book has three parts:

- Part I: Data science recipes
- Part II: Data science discussions
- Part III: Data science resources

Part I and II mostly consist of the best Analyticbridge posts by Dr. Vincent Granville, founder of Analyticbridge. Part III consists of sponsored vendor contributions as well contributions by organizations (affiliates offering software, conferences, training, books, etc.) who make our free e-book available for download on their web site. To become a sponsor or affiliate, please contact us at vincentg@datashaping.com.

To read updates about our book and download a draft version (to be published by December), visit http://www.analyticbridge.com/group/data-science.

About the Author

Dr. Vincent Granville has successfully solved problems for 15 years in data mining, text mining, predictive modeling, business intelligence, technical analysis, web crawling, keyword intelligence, big data, unstructured data and web analytics. Vincent is widely recognized as the leading expert in click scoring and web traffic optimization. Over the last ten years, he has worked in real-time credit card fraud detection with **Visa**, advertising mix optimization with **CNET** and **NBCi**, A/B testing with **LowerMyBills**, online user experience with **Wells Fargo**, search intelligence with **InfoSpace**, click fraud and Botnet detection with major search engines and large advertising clients, statistical litigation, Google and Bing

API with **Looksmart** to predict keyword yield and commercial value, automated keyword bidding with **eBay** and **Cars.com**, as well as change point detection with Bing.com (**Microsoft**). Vincent started his career in US as statistician working with **NISS** (National Institute of statistical Sciences).

Vincent was formerly Chief Science Officer at Authenticlick, where he developed patent pending technology – the startup he co-founded raised $6 million from private investors and ITU Ventures. Most recently, Vincent launched AnalyticBridge, the leading social network for analytic professionals, with 45,000 subscribers. Vincent is a former post-doctorate of Cambridge University and the National Institute of Statistical Sciences. He was among the finalists at the Wharton School Business Plan Competition and at the Belgian Mathematical Olympiads. Vincent has published 40 papers in statistical journals and is an invited speaker at international conferences. He also developed a new data mining technology known as hidden decision trees, and is currently working on an AaaS (Analytics as a Service) project to score transactions in real time, on demand, using proprietary technology.

About Analyticbridge

Analyticbridge is the leading social network for analytic professionals. The community is focused on data science, big data, small data, visualization, business analytics, predictive models, text mining, web analytics, quant, biostatistics, computer science, econometrics, risk management, six sigma, operations research, statistics and related analytic domains.

Part I: Data Science Recipes

A.1. New random number generator: simple, strong and fast

One of the best random sets of digits (extensively tested for randomness by hundreds of scientists using thousands of tests both in small and high dimensions) is the decimals of Pi. Despite its random character being superior to most algorithms currently implemented (current algorithms typically use recursive congruence relations or compositions of random permutations, and exhibit periodicity), decimals of Pi have two big challenges, making it useless as a random number generator:

1. If everybody knows that decimals of Pi are used in many high-security encryption algorithms (to generate undecipherable randomness) then guess what... it loses this very great "undecipherable-ness" property
2. Computing millions of decimals of Pi is very difficult, it takes a lot of time, much more time than traditional random number generation

Here is my answer to these two challenges, and as a result, a proposal for a new random number generator, which overcome these two difficulties:

- Regarding speed, we have now extremely fast algorithms to compute decimals of Pi, see for instance [1] and [2] below
- Regarding using Pi, we should switch to much less popular numbers that can be computed via a very similar formula, in order to preserve speed and making reverse engineering impossible in encryption algorithms. An example of a fantastic random digit generator would be to use the digits of a number defined by formula [1], with a change like this: replace the numerators 4, -2, -1, -1 by 3, 1, -2, -2. You get the idea how trillions of random generators could be developed, using variations of formula [1].

Fast formulas to compute Pi:

[1] The formula,

$$\pi = \sum_{k=0}^{\infty} \frac{1}{16^k} \left(\frac{4}{8k+1} - \frac{2}{8k+4} - \frac{1}{8k+5} - \frac{1}{8k+6} \right),$$

is remarkable because it allows extracting any individual hexadecimal or binary digit of π without calculating all the preceding ones.

[2] The Chudnovsky brothers found in 1987 that

$$\frac{426880\sqrt{10005}}{\pi} = \sum_{k=0}^{\infty} \frac{(6k)!(13591409 + 545140134k)}{(3k)!(k!)^3(-640320)^{3k}}$$

which delivers 14 digits per term.

PS: this is a reply to the following article: Classical random number generators result in identity theft, according to IEEE, see http://www.analyticbridge.com/profiles/blogs/classical-random-numbe...

Featured Comments:

[Amy] Not sure how random these digits are, but it has an advantage over all other known random generators: it's period is infinite, because Pi is an irrational number. And it looks like you could derive a recursive formula to compute the k-th digit (in base 16) based on the (k-1)-th digit.

To test for randomness, I would consider the digits of Pi as a time series and compute the auto-correlation function (or correlogram) and see if it is statistically different from 0. To perform this test, I would do as follows:

- Compute auto-correlations c(k) of lag k = 1, 2, ...,100, on the first billion digits
- Compute m = max|c(k)|
- Test if the observed m is small enough, by comparing its value (computed on Pi), with a theoretical value computed on truly random time series of digits (the theoretical value can be computed using Monte-Carlo simulations)

If there is non-randomness, I would imagine it is more likely to occur in the first few thousand digits, rather then later on. So by choosing a large number for the seed of your random number generator, you should be OK. Note that in this case, the seed is the position you use for your 1st digit, that is, if you ignore all the first 5000 digits, your seed is 5001.

[Vincent] It would be interesting to study the function F(a,b,c,d) defined by formula [1], by replacing the numbers 4, -2, -1, -1 (numerator) by a, b, c, d. Of course, F(4,-2,-1,-1) = Pi.

Do interesting numbers share the two properties listed below? Could the number e=2.71..(or some other beautiful numbers) be a special case of F?

- a + b + c + d = 0
- |a| + |b| + |c| + |d| = 8

[Vincent] Another idea: use continued fractions to approximate Pi. Let us denote by p(n) the n-th convergent: p(n) = A(n) / B(n), where A(n) and B(n) are integers defined by simple recurrence relations. Modify these recurrence relations very slightly, et voila: we created a new interesting number, with (hopefully) random digits.

Or consider a simple continued fraction such as

F = 1 / (1 + 1 / (1 + 2 / (1 + 3 / (1 + 4 / (1 + 5 / (1 + 6 / 1 + ..))))))

Despite the very strong pattern in the construction of F, I am sure that its digits show very good randomness properties.

Read and contribute to discussion, at:
http://www.analyticbridge.com/profiles/blogs/new-state-of-the-art-random-number-generator-simple-strong-and-fa

A.2. Lifetime value of an e-mail blast: much longer than you think

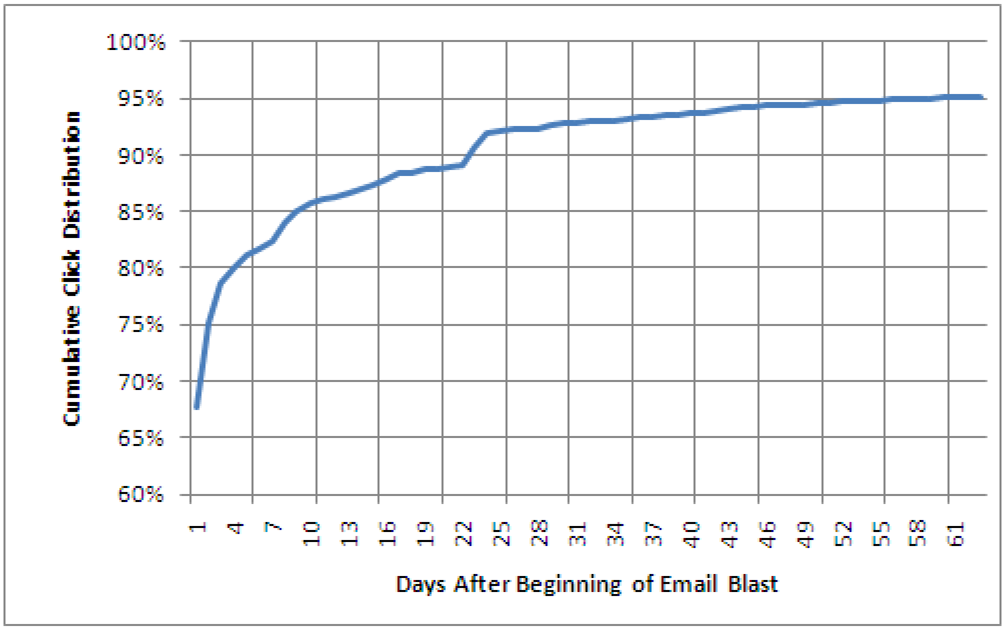
See below an example of an Analyticbridge email campaign that was monitored over a period of about **600 days**. It clearly shows that 20% of all clicks originate after day #5. Yet most advertisers and publishers ignore clicks occurring after day #3. Not only 20% of all clicks occurred after day #3, but **the best clicks** (in terms of conversions) occurred several weeks after the email blast. Also, note an organic spike occurring on day #23 in the chart below - this could be due to our vendor (iContact) featuring the newsletter in question on their website, without charging us an extra fee.

This brings interesting points:

- If you advertise a webinar taking place in 5 days, also add in your email blast a short note about webinars happening in 20, 40 or 80 days.
- If you don't renew a campaign because ROI is < 0 after 3 days, you are going to lose the "long tail" distribution that is possibly making your ROI > 0, and you might erroneously kill a profitable campaign.
- You can't just look at days #1 to #3 to estimate the lifetime value of an email blast. You are dealing with rigt-censored data, and you have to use statistical attribution and survival analysis models to measure the true impact.
- Even if you monitor clicks over a 60-day time period, you'll still miss 5% of clicks, and much more than 5% in revenue.

Thus, there's a systemic methodology flaw and bias when measuring half-life of your campaign (unless you use ad-hoc statistical methodology): the data is right-censored. How can you be sure that 97% of all clicks occur in the first 3 days? Maybe as many clicks will arrive between day 11 and day 300. But since your window of observation is only 3 days (or at best 14 days), you just can't answer the question. You can compute good estimates for half-life though, if you use a statistical model based on (say) exponential distributions, together with statistical inference, and integrate the fact that the data is right-censored, in your statistical model.

Below is a chart showing that even 60 days worth of historical data covers only 95% of your campaign in terms of clicks - and indeed much less in terms of revenue:



Here's an example where the data and conclusions are biased and wrong due to ignorance of the "right censorship" principle that applies to time series: http://blog.bitly.com/post/9887686919/you-just-shared-a-link-how-lo...

Conclusion: You don't need to track your email campaign for 600 days to measure ROI, you can monitor your campaign for 28 days, and then make interpolation using statistical models. Of course, if you started your campaign just before a great event (like Christmas shopping), then you need to take into account seasonality. That's where a statistician can help you. The above chart represents a campaign generating about 6,000 clicks.

Read and contribute to discussion, at:
http://www.analyticbridge.com/profiles/blogs/lifetime-value-of-an-e-mail-blast-much-longer-than-you-think

A.3. Two great ideas to create a much better search engine

When you do a search for "career objectives" on Google India (www.google.in), the first result showing up is from a US-based job board specializing in data mining and analytical jobs. The Google link in question redirects to a page that does not even contain the string "career objective". In short, Google is pushing a US web site that has nothing to do with "career objectives" as the #1 web site for "career objectives" in India. In addition, Google totally fail to recognize that the web site in question is about analytics and data mining.

So here's an idea to improve search engine indexes, and to develop better search engine technology:

- Allow webmasters to block specific websites (e.g. google.in) from crawling specific pages
- Allow webmasters to block specific keywords (e.g. "career objectives") from being indexed by search engines during crawling

This feature could be implemented by having webmasters using special blocking meta tags in web pages, recognized by the search engines willing to implement them.

Featured Comments:

[Vincent] Regarding the idea to build a website that provides search result pages not just for *keywords*, but also for *related links*, I've found one that provides high quality search results when someone is searching for related links. Its name is similarsites.com, and you can check the results, if you search for websites similar to Analyticbridge, by clicking on www.similarsites.com/site/analyticbridge.com.

Clearly its strengths is to show related websites (which link to the target domain, in this case Analyticbridge), by ordering the results (related domains) using a combination of outgoing links and website traffic.

You can create a search engine like Similarsites by building a table with the top 1 million websites (available for download at www.quantcast.com/top-sites-1), and for each of these 1 million websites, have up to 100 related websites (also from the same list of 1 million domains). So you could have a great search engine with a database containing less than 100 x 1 million pair of (related) domains: that's a data set fairly easy to manage (not too big).

[Jozo] To protect your webpage from unwanted traffic you may just disable Alexa, Quantcast, etc. code for bad visits.

So visitor can see his content and measurement tools aren't affected (display measure code only for good visits).

If you block a crawler you may loose you pagerank and many good visitors with it. And GoogleBot is probably the same in India and in US too.

[Vincent] Good point Jozo. Not sure where you would block the traffic, I've been thinking to block google.in via robots.txt, as this would

1. result in google.in to stop crawling the website in question
2. thus provide a better keyword and geo-data distribution on Alexa, Quantcast, Compete, etc.

3. thus make the website in question more attractive to potential advertisers who rely on Alexa, Quantcast, Compete etc. to assess the value of a website

Blocking can also be made via .htaccess. Here's an example of .htaccess file which blocks lots of undesirable traffic: http://multifinanceit.com/htaccess.txt.

If I add "career objective" in the block list, users visiting the website, following a search query with this keyword, would be redirected to an "access denied" page

---

[Jozo] Vincent, can't you write set of rules what would handle a traffic from unwanted sources?

e.g. IF HTTP_REFERRER like "%google.in%q=%career%' THEN dont_count_a_visit

---

[Vincent] See also http://www.analyticbridge.com/group/webanalytics/forum/topics/new-s... for other ideas on how to improve search.

[Vincent] Another nice feature would be to provide, for each link showing up in a search result page, the possibility (via a button or one-click action) to visit related links. This assumes the search engines uses 2 indexes: one for keywords, one for URLs (or at least, one for domain names).

[Vincent] Roberto: the answer to your question is because these unrelated terms drive CPA way up for the advertisers, as they result in no conversion. It essentially kills eCPM for the webmaster, depending on the model used to charge advertisers. In my case, I charge a flat rate, so at first glance it doesn't matter if 10% of my traffic comes from India from unrelated keywords. Yet I try to eliminate these bad sources of "free traffic" as they can negatively impact my (publicly available) web traffic statistics, and potentially scare advertisers away. Better have less, good quality traffic than more, low quality traffic - at least for my niche websites.

[Roberto] If the site makes money from new visitors, then why would they ever want not be indexed for even obscure unrelated terms. If nothing else, there is always the branding opportunity which will let a user recognize the name of a site they saw in a previous search.

[Amy] Larry: you can already define, in your meta tags, keywords that you want to be indexed, but search engines ignore these meta tags because they've been widely abused. However, a system based on keywords you want NOT to be indexed, cannot be abused.

Indeed, I'm wondering if we should use meta tags or robots.txt as the place where you specify keywords to block.

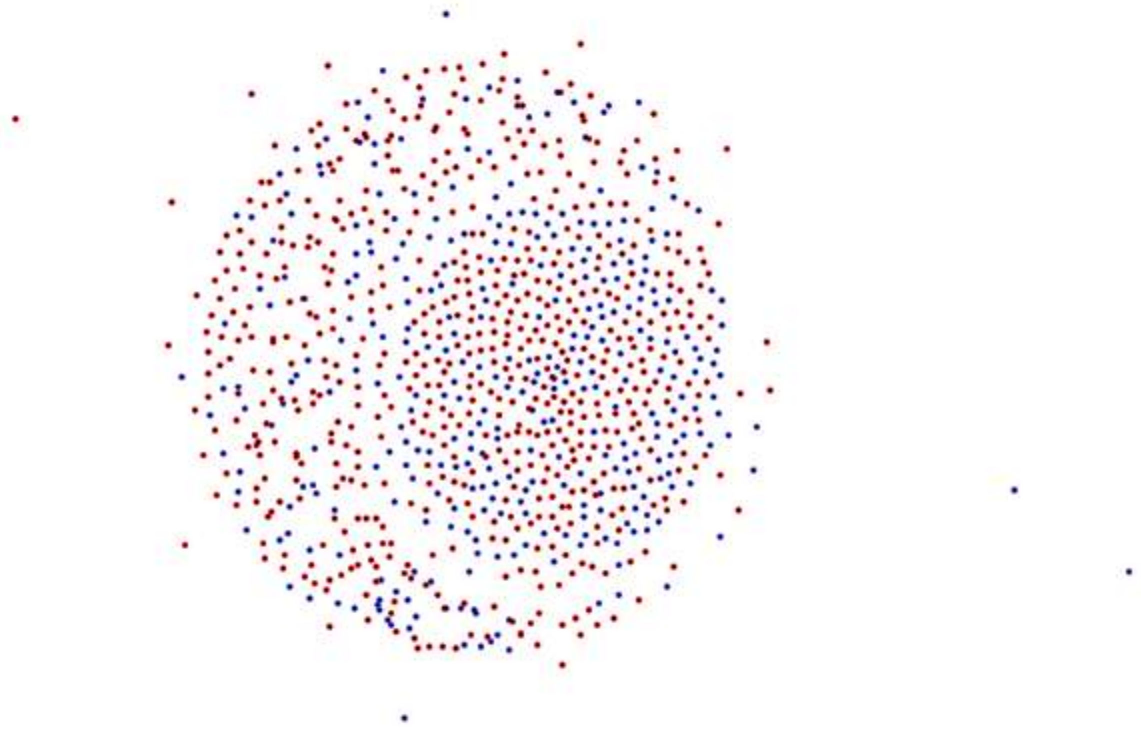
[Larry] That is an interesting idea and its got me to thinking. Wouldn't it be great if the webmasters had control of how indexing is done on their website? Perhaps a well thought out search engine could provide a api (javascript or such) to webmasters that allows them to define the specific indexing they desire. For instance if they want specific keywords, links to follow, headers, tags. The search engine will need just to look up the associated api.

Read and contribute to discussion, at:
http://www.analyticbridge.com/profiles/blogs/two-great-ideas-to-create-a-much-better-search-engine

A.4. Identifying the number of clusters: finally a solution

Here I propose a solution that can be automated and does not require visual inspection by a human being. The solution can thus be applied to billions of clustering problems, automated and processed in batch mode.

Note that the concept of cluster is a fuzzy one. How do you define a cluster? How many clusters in the chart below?

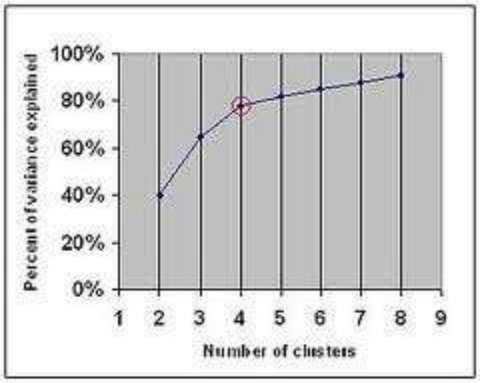


Nevertheless, in many applications, there's a clear optimum number of clusters. The methodology described here will solve all easy and less easy cases, and will provide a "don't know" answer to cases that are ambiguous.

Methodology:

- create a 2-dim table with the following rows: number of clusters in row #1, and percentage of variance explained by clusters in row #2.
- compute 3rd differences
- maximum for 3rd differences (if much higher than other values) determine number of clusters

This is based on the fact that the piece-wise linear plot of *number of cluster* versus *percentage of variance explained by clusters* is a convex function with an elbow point, see chart below. The elbow point determines the optimum number of clusters. If the piece-wise linear function is approximated by a smooth curve, the optimum would be the point vanishing the 4-th derivative of the approximating smooth curve. This methodology is simply an application of this "elbow detection" technique in a discrete framework (the number of clusters being a discrete number).



Example:

```
 1    2    3    4    5    6    7    8    9    ==> number of clusters

   40   65   80   85   88   90   91   91     ==> % variance explained by clusters

      25   15    5    3    2    1    0       ==> 1st difference

       -10   -10   -2   -1   -1   -1         ==> 2nd difference

            0    8    1    0    0            ==> 3rd difference
```

The optimum number of cluster in this example is 4, corresponding to maximum = 8 in the 3rd differences.

Note:

If you have already a strong minimum in the 2nd difference (not the case here), you don't need to go to 3rd difference: stop at level 2.

Featured Comments:

[Vincent] @Cristian: agree with you, "percentage of variance explained by clusters" might not be a good criterion depending on the configuration / shapes of the expected clusters. The purpose of this post was to illustrate the elbow detection technique and how it can be automated. In many contexts, you will need to use a different curve (not the "percentage of variance explained by clusters"), but you would still use the same automated technique for elbow detection.

[Vincent] @Sandeep: I think 3rd or 4th derivative is usually not necessary, except in a few rare cases where elbow is barely apparent (and thus clusters not well defined).

I believe that Capri's solution, based on a discrete version of curvature, is even better. And curvature only involves 1st derivative, and the angle (or its sinus) discussed by Amy is also very easy to compute. What do you think?

[Capri] The solution to finding the number of clusters is as follows:

- Let's f(k) be the the percentage of variance explained by k clusters, as in the above chart
- Compute g(k) = arctan[f(k+1)-f(k)] + arctan[f(k)-f(k-1)]
- The number k that minimizes g(k) is the optimum number of clusters

This is the solution to the min-angle rule proposed by Amy. SAS should implement it.

[Amy] Another idea, rather than looking at 2nd and 3rd differences, is to look at angles between successive line segments in the second chart. The point (joining two line segments) with the smallest angle (that is, closest to 90 degrees) determines the number of clusters.

If the curve was smooth, 2x differentiable rather than piece-wise linear, the equivalent to minimizing angle would consist in maximizing curvature. The curvature is a well defined geometrical quantity, check on Google for more details.

[Amy] Depending on the shape of the clusters, the percentage of variance explained by clusters might not be the best criterion: it works with well separated, convex clusters, and I'm not sure how efficient it is in high dimensions. I like better approaches based on fitting data with a mixture model, and estimating the number of modes.

Read and contribute to discussion, at:

http://www.analyticbridge.com/profiles/blogs/identifying-the-number-of-clusters-finally-a-solution

A.5. Online advertising: a solution to optimize ad relevancy

When you see google ads on Google search result pages or elsewhere, the ads that are displayed in front of you eyes (should) have been highly selected in order to maximize the chance that you convert and generate ad revenue for Google. Same on Facebook, Yahoo, Bing, LinkedIn and on all ad networks.

If you think that you see irrelevant ads, either they are priced very cheaply, or Google's ad relevancy algorithm is not working well.

Ad scoring algorithms used to be very simple, the score being a function of the max bid paid by the advertiser, and the conversion rate (referred to as CTR). This led to abuses: an advertiser could generate bogus impressions to dilute competitor CTR, or clicks on its own ads to boost its own CTR, or a combination of both, typically using proxies or botnets to hide its scheme, and thus gaining unfair competitive advantage on Google.

Recently, in addition to CTR and max bid, ad networks have added ad relevancy in their ad scoring mix (that is, in the algorithm used to determine which ads you will see, and in which order). In short, ad networks don't want to display ads that will make the user frustrated - it's all about improving user experience and reducing churn to boost long term profits.

How does ad relevancy scoring work?

Here's our solution. There are three components in the mix:

- The user visiting a web page hosting the ad in question
- The web page where the ad is hosted
- The ad itself, that is, its text or visual content, or other metrics such as size etc.
- The fourth important component - the landing page - is not considered in this short discussion (good publishers scrape advertiser landing pages to check the match between a text ad and its landing page, and eliminate bad adware, but that's the subject for another article)

The solution is as follows.

First create three taxonomies:

- Taxonomy A to categorize returning users based on their interests, member profile, or web searching history
- Taxonomy B to categorize web pages that are hosting ads, based on content or publisher-provided keyword tags
- Taxonomy C to categorize ads, based on text ad content, or advertiser provided keyword to describe the ad (such as bid keyword in PPC campaigns, or ad title)

The two important taxonomies are B and C, unless the ad is displayed on a very generic web page, in which case A is more important than B. So let's ignore taxonomy A for now. **The goal is to match a category from Taxonomy B with one from Taxonomy C**. Taxonomies might or might not have the same categories, so in general it will be a fuzzy match, where for instance, the page hosting the ad is attached to categories *Finance / Stock Market* in Taxonomy B, while the ad is attached to categories *Investing / Finance* in Taxonomy C. So you need to have a system in place, to measure distances between categories belonging to two different taxonomies.

How do I build a taxonomy?

There are a lot of vendors and open source solutions available on the market, but if you really want to build your own taxonomies from scratch, here's one way to do it:

- Scrape the web (DMOZ directory with millions of pre-categorized webpages, that you can download freely, is a good starting point), extract pieces of text that are found on a same web page, and create a distance to measure proximity between pieces of text
- Clean, stem your keyword data
- Leverage your search query log data: two queries found in a same session are closer to each other (with respect to the above distance) than arbitrary queries
- Create a table with all pairs of "pieces of text" that you have extracted and that are associated (e.g. found on a same web page or same user session). You will be OK if your table has 100 million such pairs.

Let's say that (X, Y) is such a pair. Compute n1 = # of occurrences of X in your table; n2 = # of occurrences of Y in your table, and n12 = # of occurrences where X and Y are associated (e.g. found on a same web page). A metric that tells you how close X and Y are to each other would be R = n12 / SQRT(n1 * n2). With this dissimilarity metric (used e.g. athttp://www.frenchlane.com/kw8.html) you can cluster keywords via hierarchical clustering and eventually build a taxonomy - which is nothing else than an unsupervised clustering of the keyword universe, with labels manually assigned to (say) top 20 clusters - each representing a category.

Featured Comments:

[Capri] This is a good solution for companies such as LinkedIn or Facebook, where traffic is not driven by search. For search engines, you can match a user query to a bid keyword, using an algorithm that extracts potential bid keywords out of user queries (these algorithms are performing very poorly in my opinion).

Companies such as Facebook actually found the following solution to improve ad relevancy: just allow the user to flag an ad as *useful* or *annoying* or *irrelevant*. By blending this user feedback with some mathematical ad relevancy system, Facebook could achieve better results, in terms of ROI / advertiser churn.

[Amy] If you have 20 categories in taxonomy B, and 25 in taxonomy C, then you only have 20 x 25 = 500 pairs of categories. The category A to category C matching is then trivial: you just need to use a lookup table with 500 entries.

Read and contribute to discussion, at:

http://www.analyticbridge.com/profiles/blogs/online-advertising-a-solution-to-optimize-ad-relevancy

A.6. Example of architecture for AaaS (Analytics as a Service)

There is an increasing number of individuals and companies that are now delivering analytics solutions using modern web-based platforms: Data-Applied, DataSpora and AnalyticBridge to name a few.

The concept is at least 10 years old, but because inexpensive web servers can now handle a large bandwidth, and can process megabytes of data in a few seconds (even without cloud), and because Internet users have much faster (broadband) connections, it is possible to develop analytics applications capable of processing millions of observations online, on demand, in real time, and deliver results via an API, or "on the fly". In some cases, results consist of processed data sets, sometimes fairly large, where one column has been added to the input file: for instance, the new column (the output) is a score attached to each observation. This is a solution that we are working on, using ad-hoc statistical techniques to process data very efficiently with hidden decision trees, using very little memory and efficient data structures, and thus allowing users to process online, "on the fly", large data sets that R or other statistical packages would not be able to process even on a desktop. In fact, these traditional packages (R, Splus, Salford Systems) require that all your data be stored in memory, and will typically crash if your input file has more than 500,000 observations. Web 3.0 analytics can easily handle much larger data sets -- online!

Interestingly, this new type of analytics service can rely on popular statistical packages (SAS, etc.) or can use ad-hoc algorithms written in Perl (including production of charts with the GD library), Python, C, C# or Java. A version based on SAS would be called a *SAS web server* (extranet or intranet) and work as follows:

- An API call is made to an external web site where SAS is installed; parameters in the API call describe the type of analysis requested (logistic regression, training data and machine learning step, actual processing of new data, etc.)
- A Perl script processes the HTTP request, extracts the parameters and automatically writes a SAS program corresponding to the user's request.
- The SAS code is run from the Perl script in command-line mode, and produces an output file such as a chart or XML or data file.
- The Perl script reads the chart and display it in the browser (if the user is a human being using a web browser), and provides a URL where the user can fetch the chart (in case the user is a web robot executing an API call).

Once our application (analytics 3.0) will be live, we will make a public announcement, probably in January. Stay tuned!

Featured Comments:

[Vincent] Additional references:

- Analytics 3.0. - Designing an all-purpose analytics web server, off...
- Source code for web robot (for the HTTP request)
- Scoring technology that AnalyticBridge is offering as open source

---

[Vincent] One idea is that you must purchase a number of transactions before using the paid service, and add dollars regularly. A transaction is a call to the API.

The service is accessed via an HTTP call that looks like

http://www.datashaping.com/AnalyticsAPI?clientID=xxx&dataSource=yyy&service=zzz&parameters=abc

When the request is executed,

- First the script checks if client has enough credits (dollars)
- If yes it fetches the data on the client web server: the URL for the source data is yyy
- Then the script checks if source data is OK or invalid, or client server unreachable
- Then it executes the service zzz, typically, a predictive scoring algorithm
- The parameter field tells whether you train your predictor (data = training set) or whether you use it for actual predictive scoring (data outside the training set)
- Then it processes data very fast (a few secs for 1MM observations for the training step)
- Then it sends an email to client when done, with the location (on the datashaping server) of the results (the location can be specified in the API call, as an additional field, with a mechanism in place to prevent file collisions from happening)
- Then it updates client budget

Note all of this can be performed without any human interaction. Retrieving the scored data can be done with a web robot, and then integrated into the client's database (again, automatically). Training the scores would be charged much more than scoring one observation outside the training set. Scoring one observation is a transaction, and could be charged as little as $0.0025.

This architecture is for daily or hourly processing, but could be used for real time if parameter is not set to "training". However, when designing the architecture, my idea was to process large batches of transactions, maybe 1MM at a time.

Read and contribute to discussion, at:
http://www.analyticbridge.com/group/aaasanalyticsasaservice/forum/topics/example-of-architecture-for

A.7. Why and how to build a data dictionary for big data sets

One of the most valuable tools that I've used, when performing exploratory analysis, is building a data dictionary. It offers the following advantages:

- Identify areas of sparsity and areas of concentration in high-dimensional data sets
- Identify outliers and data glitches
- Get a good sense of what the data contains, and where to spend time (or not) in further data mining

What is a data dictionary?

A data dictionary is a table with 3 or 4 columns. The first column represents a label: that is, the name of a variable, or a combination of multiple (up to 3) variables. The second column is the value attached to the label: the first and second columns actually constitute a name-value pair. The third column is a frequency count: it measures how many times the value (attached to the label in question) is found in the data set. You can add a 4-th column, that tells the dimension of the label (1 if it represents one variable, 2 if it represents a pair of two variables etc.)

Typically, you include all labels of dimension 1 and 2 with count > threshold (e.g. threshold = 5), but no or only very few values (the ones with high count) for labels of dimension 3. Labels of dimension 3 should be explored after having built the dictionary for dim 1 and 2, by drilling down on label/value of dim 2, that have a high count.

Example of dictionary entry

category~keyword travel~Tokyo 756 2

In this example, the entry corresponds to a label of dimension 2 (as indicated in column 4), and the simultaneous combination of the two values (travel, Tokyo) is found 756 times in the data set.

The first thing you want to do with a dictionary is to sort it using the following 3-dim index: column 4, then column 1, then column 3. Then look at the data and find patterns.

How do you build a dictionary?

Browse your data set sequentially. For each observation, store all label/value of dim 1 and dim 2 as hash table keys, and increment count by 1 for each of these label/value. In Perl, it can be performed with code such as $hash{"$label\t$value"}++.

If the hash table grows very large, stop, save the hash table on file then delete it in memory, and resume where you paused, with a new hash table. At the end, merge hash tables after ignoring hash entries where count is too small.

Featured Comments:

[Jozo] If you got binary target variable {0,1} you add 5th column with sum(target). this allows you to calculate variable predictive power vs. target (Weight of Evidence-Information Value or ChiSquare) for all categorical variables. and when there are N binary targets, just add N more columns - get it all in the single pass through your data.

Read and contribute to discussion, at:

http://www.analyticbridge.com/profiles/blogs/why-and-how-you-should-build-a-data-dictionary-for-big-data-sets

A.8. Hidden Decision Trees: A Modern Scoring Methodology

Hidden Decision Trees is a statistical and data mining methodology (just like logistic regression, SVM, neural networks or decision trees) to handle problems with large amounts of data, non-linearities and strongly correlated dependent variables.

The technique is easy to implement in any programming language. It is more robust than decision trees or logistic regression, and help detect natural final nodes. Implementations typically rely heavily on large, granular hash tables.

No decision tree is actually built (thus the name hidden decision trees), but the final output of an hidden decision tree procedure consists of a few hundred nodes from multiple non-overlapping small decision trees. Each of these parent (invisible) decision trees corresponds e.g. to a particular type of fraud, in fraud detection models. Interpretation is straightforward, in contrast with traditional decision trees.

The methodology was first invented in the context of credit card fraud detection, back in 2003. It is not implemented in any statistical package at this time. Frequently, hidden decision trees are combined with logistic regression in an hybrid scoring algorithm, where 80% of the transactions are scored via hidden decision trees, while the remaining 20% are scored using a compatible logistic regression type of scoring.

Hidden decision trees take advantage of the structure of large multivariate features typically observed when scoring a large number of transactions, e.g. for fraud detection. The technique is not connected with hidden Markov fields.

History of HDT (Hidden Decision Trees):

- 2003: First version applied to credit card fraud detection
- 2006: Application to click scoring and click fraud detection
- 2008: More advanced versions to handle granular and very large data sets
    - Hidden Forests: multiple HDT's, each one applied to a cluster of correlated rules
    - Hierarchical HDT's: the top structure, not just rule clusters, is modeled using HDT's
    - Non binary rules (naïve Bayes blended with HDT)

Power point presentation: http://www.analyticbridge.com/group/whitepapers/forum/topics/hidden.

Featured Comments:

[Yi-Chun] Does this apply to large data set? I am currently using logistic regression to build response model on about half million customers with over 300 variable.

[Vincent] @ Yi-Chun: Yes, it was initially designed to handle data sets with 60,000,000 observations. It took 2 days for SAS EM to analyze the lift from one rule set, using *decision trees*, while *hidden decision trees* could process hundreds of rules in less than 3 hours (if written in Perl) and in less than one hour if written in C.

[Vincent] It is not available in SAS nor in other statistical packages. In SAS, you would have to call a few procedures from SAS Base and possibly write some macros to get it implemented. It's a new methodology.

[Matt] Vincent. The general idea sounds quite similar to Random Forests. Could you briefly explain how this differs?

[Vincent] @ Matt: It differs in the following ways:

- It does not involve comparing / averaging / computing a mode across multiple decision trees with (potentially) overlapping nodes
- No decision tree is actually built, so there's no splitting and no splitting criterion ever used (no pruning either)
- Final nodes are not necessarily "deepest nodes", they usually are not very deep
- Emphasis is not on producing maximum predictive power, but instead on maximum robustness to avoid over-fitting
- *Hidden decision trees* is an hybrid method. In the case I am working on, 75% of the transactions are scored via hidden decision trees nodes, and 25% are scored with another methodology. The reason being that only 75% of the transactions belong to statistically significant nodes. And the remaining 25% cannot be handled by neighboring parent nodes because of bias: in a fraud detection system, these 25% transactions tend to be more fraudulent than average.
- Eventually, all methods are equivalent. A logistic regression with dummy variables (logic logistic regression) with 2nd, 3rd and 4th order interactions, with an observations matrix with a very large number of variables (mostly cross products of initial variables), but an extremely sparse matrix at the same time, with sophisticated numerical analysis techniques to handle sparsity, is equivalent to decision trees.
- "Random forests" are to "decision trees" what "hidden forests" are to "hidden decision trees".

[Vincent] I am working on a solution where there's no need to use an hybrid strategy anymore. Observations that do not belong to a "statistically significant" node will be assigned a metric computed on the *k*-nearest nodes, rather than processed through constrained logistic regression. A correction for bias (for these observations) will be introduced. An example of a successful application will be provided: predicting the commercial value and/or volume of a keyword in Google advertising campaigns.

Read and contribute to discussion, at:

http://www.analyticbridge.com/forum/topics/hidden-decision-trees-vs

A.9. Scorecards: Logistic, Ridge and Logic Regression

In the context of credit scoring, one tries to develop a predictive model using a regression formula such as $Y = \sum w_i R_i$, where $Y$is the logarithm of odds ratio (fraud vs. non fraud). In a different but related framework, we are dealing with a logistic regression where Y is binary, e.g. $Y = 1$ means fraudulent transaction, $Y = 0$ means non fraudulent. The variables R_i, also referred to as fraud rules, are binary flags, e.g.

- high dollar amount transaction
- high risk country
- high risk merchant category

This is the first order model. The second order model involves cross products R_i x R_j to correct for rule interactions. The purpose of this question is to how best compute the regression coefficients w_i, also referred to as rule weights. The issue is that rules substantially overlap, making the regression approach highly unstable. One approach consists of constraining the weights, forcing them to be binary (0/1) or to be of the same sign as the correlation between the associated rule and the dependent variable Y. This approach is related to ridge regression. We are wondering what are the best solutions and software to handle this problem, given the fact that the variables are binary.

Note that when the weights are binary, this is a typical combinatorial optimization problem. When the weights are constrained to be linearly independent over the set of integer numbers, then each $\sum w_i R_i$ (sometimes called unscaled score) corresponds to one unique combination of rules. It also uniquely represents a final node of the underlying decision tree defined by the rules.
Contributions:

- From Mark Hansen: When the rules are binary, the problem is known as logic regression.

Read and contribute to discussion, at:
http://www.analyticbridge.com/profiles/blogs/2004291:BlogPost:17382

A.10. Iterative Algorithm for Linear Regression – Approximate vs. Exact Solution

I am trying to solve the regression $Y=AX$ where Y is the response, X the input, and A the regression coefficients. I came up with the following iterative algorithm:

$$A_{k+1} = cYU + A_k\,(I-cXU),$$

where:

- c is an arbitrary constant
- U is an arbitrary matrix such that YU has same dimension as A. For instance U = transposed(X) works.
- A_0 is the initial estimate for A. For instance A_0 is the correlation vector between the independent variables and the response.

Questions:

- What are the conditions for convergence? Do I have convergence if and only if the largest eigenvalue (in absolute value) of the matrix $I-cXU$ is strictly less than 1?
- In case of convergence, will it converge to the solution of the regression problem? For instance, if $c=0$, the algorithm converges, but not to the solution. In that case, it converges to A_0.

Parameters:

- n: number of independent variables
- m: number of observations

Matrix dimensions:

- A: $(1,n)$ (one row, n columns)
- I: (n,n)
- X: (n,m)
- U: (m,n)
- Y: $(1,m)$

Why using an iterative algorithm instead of the traditional solution?

- We are dealing with an ill-conditioned problem; most independent variables are highly correlated.
- Many solutions (as long as the regression coefficients are positive) provide a very good fit, and the global optimum is not that much better than a solution where all regression coefficients are equal to 1.
- The plan is to use an iterative algorithm to start at iteration #1 with an approximate solution that has interesting properties, then move to iteration #2 to improve a bit, then stop.

Note: this question is not related to the ridge regression algorithm described here.

Contributions:

- From Ray Koopman

  No need to apologize for not using "proper" weights. See
  Dawes, Robyn M. (1979). *The robust beauty of improper linear models in decision making*. American Psychologist, 34, 571-582.

Read and contribute to discussion, at:

http://www.analyticbridge.com/profiles/blogs/2004291:BlogPost:19370

A.11. Approximate Solutions to Linear Regression Problems

Here we assume that we have a first order solution to a regression problem, in the form

$$Y = \Sigma\, w_i\, R_i,$$

where Y is the response, w_i are the regression coefficients, and R_i are the independent variables. The number of variables is very high, and the independent variables are highly correlated.
We want to improve the model by considering a second order regression of the form

$$Y = \Sigma\, w_i\, R_i + \Sigma\, w_{ij}\, c_{ij}\, m_{ij}\, R_i\, R_j\ ,$$

where:

- c_{ij} = correlation between R_i and R_j
- $w_{ij} = |\, w_i w_j\, |^{0.5}$ x sgn($w_i w_j$)
- m_{ij} are arbitrary constants

In practice, some of the R_is are highly correlated and grouped into clusters. These clusters can be identified by using a clustering algorithm on the c_{ij}s. For example, one could think of a model with two clusters A and B such as

$$Y = \Sigma\, w_i\, R_i + m_A\, \Sigma_A\, w_{ij}\, c_{ij}\, R_i\, R_j$$

$$+\, m_B\, \Sigma_B\, w_{ij}\, c_{ij}\, R_i\, R_j$$

where

- Σ_A (resp. Σ_B) are taken over all $i < j$ belonging to A (resp. B)
- $m_{ij} = m_A$ (constant) if i, j belong to cluster A
- $m_{ij} = m_B$ (constant) if i, j belong to cluster B

An interesting case occurs when the cluster structure is so strong that

- $|\, c_{i\,j}\, | = 1$ if i and j belong to the same cluster (either A or B)
- $c_{ij} = 0$ otherwise

This particular case results in

$$m_A = 4\, /\, [1 + (1{+}8k_A)^{0.5}]$$

$$m_B = 4\, /\, [1 + (1{+}8k_B)^{0.5}]$$

where $k_A = \Sigma_A\, |\, c_{ij}\, |$ and $k_B = \Sigma_B\, |\, c_{ij}\, |$.

Question
If the cluster structure is moderately strong, with the correlations c_{ij} close to 1, -1 or 0, how accurate is the above formula involving k_A and k_B? Here we assume that the w_is are known or approximated.
Typically, w_i is a constant or w_i is a simple function of the correlation between Y and R_i.

Alternate Approach
Let us consider a simplified model involving one cluster, with m_{ij} = constant = m. For instance, the unique cluster could consist of all variables i, j with $|\, c_{ij}\, | > 0.70$. The model can be written as

$$Y = \Sigma\, w_i\, R_i + m\, \Sigma\, w_{ij}\, c_{ij}\, R_i\, R_j.$$

We want to find m that provides the best improvement over the first order model, in terms of residual error. The first order model corresponds to $m = 0$.

Let us introduce the following notations:

- $W = \Sigma\, w_{ij}\, c_{ij}\, R_i\, R_j$,
- $V = W - u$, where u = average(W) (Thus V is the centered W, with mean 0),
- $S= \Sigma\, w_i\, R_i$. (average(S) = average(Y) by construction)

Without loss of generality, let us consider the slightly modified (centered) model

$$Y = S + m\, V.$$

Then

$$m = [\text{ Transposed}(V) \text{ x } (Y\text{-}S)\ ] \,/\, [\text{ Transposed}(V) \text{ x } V\ ],$$

where

- Y, S, and V are vectors with n rows,
- n is the number of observations.

Further Improvements

The alternate approach could be incorporated in an iterative algorithm, where at each step a new cluster is added. So at each step we would have the same computation for m, optimizing the residual error on

$$Y = S + m\, V.$$

However this time, S would contain all the clusters detected during the previous step, and V would contain the new cluster being added to the model.

Read and contribute to discussion, at:

http://www.analyticbridge.com/profiles/blogs/approximate-solutions-to

A.12. Theorems for Stock Traders

The following theorems are related to the stock market and trading strategies. They have roots in the martingale theory, random walk processes, gaming theory or neural networks. We present some of the most amazing and deep mathematical results, of practical interest to the curious trader.

Lorenz Curve

Let's say that one makes 90% of his trading gains with 5% of his successful trades. We write$h(0.05) = 0.90$. The function h is known as the Lorenz curve. If the gains are the realizations of a random variable X with cdf F and expectation $E[X]$, then

$$h(x) = \left(\int_{[0,x]} F^{-1}(v)\, dv \right) / E[X], \qquad 0 \le x \le 1.$$

To avoid concentrating too much gain on just a few trades, one should avoid strategies that have a sharp Lorenz curve. The same concept applies to losses. Related keywords: inventory management, Six Sigma, Gini index, Pareto distribution, extreme value theory.

Black-Scholes Option Pricing Theory

The Black-Scholes formula relates the price of an option to five inputs: time to expiration, strike price, value of the underlier, implied volatility of the underlier, the risk-free interest rate. For technical details, check out **www.hoadley.net**. You may also look at the book *A Probability Path* by Sidney Resnik (Ed. Birkhauser, 1998).

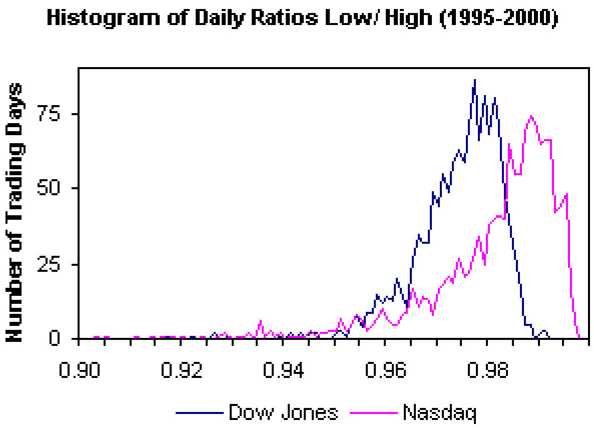
The formula can be derived from the theory of Brownian motions. It relies on the fact that stock market prices can be modeled by a geometric Brownian process. The model assumes that the variance of the process does not change over time, and that the drift is a linear function of the time. However these two assumptions can be invalidated in practice.

Discriminant Analysis

Stock picking strategies can be optimized using discriminant analysis. Based on many years of historical stock prices, it is possible to classify all stocks in three categories - bad, neutral or good - at any given time. The classification rule must be associated with a specific trading strategy, such as buying at close today and selling at close seven days later. Data Shaping Solutions is currently **investigating** this approach.

Generalized Extreme Value Theory

What is the parametric distribution of the daily ratio high/low? Or the 52-week high/low? And how would you estimate the parameter for a particular stock? Interdependencies in the time series of stock prices make it difficult to compute an exact theoretical distribution.



The distribution is characterized by two parameters: mode and interquartile. The Nasdaq has a much heavier lefthand tail, making it more attractive to day traders. As a rule of thumb, stocks with an heavy lefthand tail are good picks for Data Shaping strategies.

Random Walks and Wald's identity

> *Let us consider a random walk in Z, with transition probabilities P(k to k+1)=p, P(k to k)=q, P(k to k-1)=r, with p+q+r=1. The expected number of steps for moving above any given starting point is infinite if p is smaller than r. It is equal to 1/(p-r) otherwise.*

This result, applied to the stock market, means that under stationarity conditions (p=r), investing in a stock using the buy and hold strategy may never pay off, even after an extremely long period of time.

Arcsine Law

This result explains why 50% of the people consistently lose money, while 50% consistently win. Let's compare stock trading to coin flipping (tails = loss, heads = gain). Then

- The probability that the number of heads exceeds the number of tails in a sequence of coin-flips by some amount can be estimated with the Central Limit Theorem and the probability gets close to 1 as the number of tosses grows large.
- The law of long leads, more properly known as the arcsine law, says that in a coin-tossing games, a surprisingly large fraction of sample paths leave one player in the lead almost all the time, and in very few cases will the lead change sides and fluctuate in the manner that is naively expected of a well-behaved coin.
- Interpreted geometrically in terms of random walks, the path crosses the x-axis rarely, and with increasing duration of the walk, the frequency of crossings decreases, and the lengths of the “waves” on one side of the axis increase in length.

Read and contribute to discussion, at:

http://www.analyticbridge.com/group/computationalfinance/forum/topics/theorems-for-traders

A.13. Preserving metrics and scores consistency over time and across clients, when data sets change

Changes can come from multiple sources: definition of a visit or web site visitor is changed, resulting in visitor counts suddenly dropping or exploding. Internet log files change, for instance the full user agent string is no longer recorded, impacting traffic quality scores. Or one client has data fields that are not the same or only partially overlap with those from other clients.

How do you handle this issue?

The answer is simple: when a change in scores is detected (whether your scoring algorithm or your data has changed), apply the new scores backward to at least 2-week before the change, compare the old and new score for these 2 weeks of overlapping scores, then re-calibrate the new scores using these 2 week worth of data, to make them consistent (e.g. same median, same variance).
If the issue is not temporal but rather the fact that different clients have different data sets, then use a subset of the two data sets, where data fields are compatible, and compute scores for both clients on these reduced data sets (and compare with scores computed on full data sets). These 4 scores (2 clients, reduced data and full data) will be used for re-calibration.

Notes

- Use change-point, trend-reversal or slope-change detection algorithms to detect changes. However, the changes I am taking here are usually brutal and definitely visible with the naked eye even by a non-statistician (and in many cases unfortunately, by one of your clients).
- When you improve a scoring algorithm, if it improves scores on A but makes them worse on B, then create an hybrid, blended score consisting of old score for B and new score for A.

Read and contribute to discussion, at:

http://www.analyticbridge.com/profiles/blogs/preserving-metrics-and-scores-consistency-over-time-and-across

A.14. Advertising: mathematical formulas for reach and frequency

How many ad impressions should one purchase to achieve a specified reach on a website? We have investigated the problem and found a simple mathematical solution, for run-of-site advertising:

First formula:

Reach = Unique Users – $\boldsymbol{\Sigma}\ U_k * (1 - P)^k$,

where

- The sum is over all positive integers k = 1, 2, etc.
- U_k is the number of unique users turning *exactly* k pages, for the time period in question (e.g. 28 days). We assume that we have a tiny lookup table, mapping k to U_k. Typically, you don't need to go beyond k = 30 to take into account 99% of the web traffic. If exact values are not known, use interpolation techniques to build this look-up table.
- P is the ratio of purchased impressions by total page views, for the site and time period in question (P is always < 1 as you can't buy more impressions than the actual number of page views)

Second formula:

Number of unique users who see the ad n times = $\boldsymbol{\{\Sigma}\ U_k * C(k, n) * P^n * (1 - P)^k \boldsymbol{\}} - n$,

where

- The sum is over all positive integers k greater or equal to n
- $C(k, n) = k! / [\ n!\ (k-n)!\ ]$ are the binomial coefficients. The formula relies on the distribution of pages per unique user during the targeted period of time. It helps determine the optimum number of ad impressions to purchase.

Read and contribute to discussion, at:

http://www.analyticbridge.com/forum/topics/2004291:Topic:18137

A.15. Real Life Example of Text Mining to Detect Fraudulent Buyers

The credit card transaction described here in details is a real example of a fraudulent transaction performed by organized criminals, undetected by all financial institutions involved, and very easy to detect with simple text mining techniques.

It was not caught by any of the financial institutions involved in processing (declining or accepting) the transaction in question: merchant gateway, bank associated with the credit card holder, bank associated with the merchant, e-store, Visa. It was manually declined by the manager of the e-store, who investigated the transaction.

In short, scoring algorithms used by financial institutions, to check whether a transaction should be accepted or not, could be significantly improved using findings from the case described below. The pattern associated with this specific online purchase is very typical of traditional online fraudulent transactions.

Patterns:

- We are dealing with a B2B merchant with very good rating, located in US. Financial institutions should have a field in their transaction databases, to identify B2B from B2C or something else.
- The purchase took place Friday night. This is very unusual for the merchant in question, and it is unusual for B2B merchants in general (US merchants).
- Cardholder address is somewhere in Chicago, IL.
- Phone number (716-775-8339) is listed in Grand Island, NY, although it was reported before as an Indian cell phone number.
- Product being purchased is a product with a higher fraud risk. Historical data should show that risk associated with this product is higher than from other products from the same merchant.
- IP address from purchaser is 173.193.216.110, corresponding to gtalkproxy.com, a domain name with server in Dallas, TX, and owned by Arunava Bhowmick, a guy located in India. In addition, the domain name contains the term "proxy", a red-flag by itself (unless it's a corporate proxy, but this is not the case here).
- A Google search on the phone number points to a fraud report about "christan kingdom shipping company Renee Darrin, Terrysa Leteff free car scam Pasadena, Texas". See http://www.ripoffreport.com/auto-shipping-companies/christan-kingdo....
- Email address of purchaser is recruits@integrity-holdings.com: integrity-holdings.com is a non-existent website (the domain is hosted by Intuit.com), and quite likely, the purchaser provided a fake email address.

All these findings, which make this transaction highly suspicious, would have been extremely easy to detect in real-time, automatically, with a tiny bit of web crawling and text analytics, when the transaction was being reviewed by the merchant. Or even better, before it made its way to the e-store.

Methodology to detect this type of fraud:

- Capture a number of metrics on the online purchase form: phone number, e-mail address (keep in mind that the purchaser can fake these fields)
- Record IP address of purchaser
- Do a search on the e-mail address and the domain attached to it. Is the domain name empty? Is it a free email account (gmx, hotmail, yahoo, gmail)? From which country? Can you successfully ping the e-mail address?

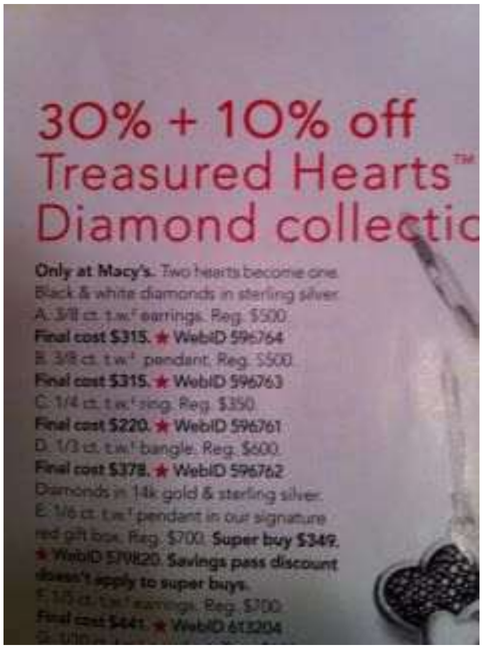
- Do a reverse lookup on the IP address to retrieve domain name. Is domain name a non-corporate proxy? From which country?
- Are IP address, phone number, email address and cardholder address all from different states or countries?
- Do a search on the phone number: does the search return results containing one of the following strings: abuse, scam, spam etc.
- Create a credit card transaction score that integrate the above rules.

Read and contribute to discussion, at:

http://www.analyticbridge.com/profiles/blogs/real-life-example-of-text-mining-to-detect-fraudulent-buyers

A.16. Discount optimization problem in retail analytics

Macy's thinks that their customers are innumerate Or maybe they use faulty calculator: 30% + 10% off of $500 results in a new price of $300, not $315 as advertised by Macy's (see ad below in their catalog).



30% + 10% off
Treasured Hearts™
Diamond collectic

Only at Macy's. Two hearts become one. Black & white diamonds in sterling silver. A: 3/8 ct. t.w.† earrings. Reg. $500. **Final cost $315.** ★ **WebID 596764** B. 3/8 ct. t.w.† pendant. Reg. $500. **Final cost $315.** ★ **WebID 596763** C: 1/4 ct. t.w.† ring. Reg. $350. **Final cost $220.** ★ **WebID 596761** D. 1/3 ct. t.w.† bangle. Reg. $600. **Final cost $378.** ★ **WebID 596762** Diamonds in 14k gold & sterling silver. E: 1/6 ct. t.w.† pendant in our signature red gift box. Reg. $700. **Super buy $349.** ★ **WebID 579820. Savings pass discount doesn't apply to super buys.** F. 1/3 ct. t.w.† earrings. Reg. $700. **Final cost $441.** ★ **WebID 613204**

Did their marketing statisticians find that lying to clients increase sales? Maybe it does, in countries where people are afraid by mathematics. But in one of these analytic-poor countries (USA), we have law against false advertising. In China (an analytic-rich country), there's no such law but then everybody would immediately notice the lie.

Featured Comments:

[Sharon] Actually, their math is correct. 30% + 10% works like this....You take 30% off the original price which is $350 in this case. Then you take 10% which makes it $315. Unfortunately, 30% + 10% is listed this way for that reason and not as 40%.
Does this apply to large data set? I am currently using logistic regression to build response model on about half million customers with over 300 variable.

[Amy] 30% + 10% = 40%. Their discount is 37%. Why do they say 30% + 10%, when they could just simply say 37% and avoid being perceived as dishonest?

[Amy] At the end of the day, this is all about optimizing revenue through marketing analytics. If they could say that 30% + 10% off on $500 is $325, and if it works well, they would do it - as long as they get the green light from their legal department. Sure 1% of prospects won't buy, but the vast majority won't check the computation and will pay more. And Macy could always argue that the $325 discounted price (when most would expect $315 and some expect $300) is the result of "discount fees" that must be added back into the final price.

Even easier, they could claim that the original price is $600, and the discounted price $325. This is what we call pricing optimization.

[Vincent] Even better: they should advertise 15% + 15% + 10%, now the discounted price climbs from $315 to $325, even though it still looks like a 40% discount. Or 1% + 1% + 1% + ... + 1% (40 times), which corresponds to a discount from $500 to $334.48. Indeed, the best they could ever get with this scheme (out of all mathematical combinations), by tricking people into believing that it's still a 40% discount, is a discount from $500 to $335.16 = exp(-40%) * $500.

Read and contribute to discussion, at:

http://www.analyticbridge.com/forum/topics/macy-s-thinks-that-their-customers-are-innumerate

A.17. Sales forecasts: how to improve accuracy while simplifying models?

The solution is simple: leverage external data, and simplify your predictive model.

Back in 2000 I was working with GE's analytic team to improve sales forecasts for NBC Internet, a web portal owned by NBC. The sales / finance people were using a very basic formula to predict next month sales, based mostly on sales from previous month. With GE, we started to develop more sophisticated models that included time series techniques (ARMA - auto regressive models) and seasonality, but still was entirely based on internal sales data.

Today, many companies still fail to use the correct methodology to get accurate sales forecasts. This is especially true for companies in growing or declining industries, or when volatility is high due to macro-economic, structural factors. Indeed, the GE move toward using more complex models was the wrong move: it did provide a small lift, but failed to generate the huge lift that could be expected switching to the right methodology.

So what is the right methodology?
Most companies with more than 200 employees use independent silos to store and exploit data: data from the finance, advertising / marketing, and operation / inventory management / product departments are more or less independent and rarely merged together to increase ROI. Worse, external data sources are totally ignored.

Even each department has its own silos: within the BI department, data regarding paid search, organic search and other avertising (print, TV, etc.), is treated separately by data analysts that don't talk to each other. While lift metrics from paid search (managed by SEM people), organic search (managed by SEO people) and other advertising are clearly highly dependent, from a business point of view, interaction is ignored and the different chanels are independently - rather than jointly - optimized.

If a sales results from a TV ad seen 3 months ago, together with a Google ad seen twice last month, and also thanks to good SEO and landing page optimization, it will be impossible to accuratly attribute the dollar amount to the various managers involved in making the sale happens. Worse, sales forecasts suffer from not using external data and econometric models.

For a startup (or an old company launching a new product), it is also important to accurately assess sales growth, using auto-regressive time series models that take into account advertising spend and a decay function of time. In the NBC Internet example, we've found that TV ads have an impact for about six months, and a simple but good model would be

$$\text{Sales}(t) = g\{\, f(\text{sales}(t-1, t-2, \ldots , t-6), \text{a1*SQRT}[\text{AdSpend}(t-1)] + \ldots + \text{a6*SQRT}[\text{AdSpend}(t-6)] \,\}$$

where the time unit is one month (28 days is indeed better), and both *g* and *f* are functions that need to be identified via cross-validation and model fitting techniques (the *f* function corresponding to the ARMA model previously mentioned).

Pricing optimization (including an elementary price elasticity component in the sales forecasting model), client feedback, new product launch and churn should be part of any basic sales forecasts. In addition, sales forecasts should integrate external data, in particular:

- Market share trends: is your company losing or gaining market share?
- Industry forecasts (growth, decline in your industry)
- Total sales volume for your industry, based on competitor data (the data in question can easily be purchased from companies selling competitive intelligence).
- Prices from your top competitors for same products, in particular, price ratios (yours vs. competition)
- Economic forecasts: some companies sell this type of data, but their statistical models have flaws, and they tend to use outdated data or move very slowly about updating their models with fresh data.
- Analyticbridge plans to provide econometric forecasts, and to create an econometric index about the future of the economy. We've identified metrics connected to some of our internal sales, which are very good 30-day predictors of the stock market indexes and of the economy in general. More on this later when we have completed all our calibration tests. We will probably make a version of our economic index (30-day forecasts) available for free.

A very simple model

Identify the top four metrics that drive sales among the metrics that I have suggested in this article (by all means, please do not ignore external data sources - including a sentiment analysis index by product, that is, what your customers write about your products on Twitter), and create a simple regression model. You could get it done with Excel (use the data analysis plug-in or the linest functions) and get better forecasts than using a much more sophisticated model based only on internal data coming from just one of your many silos. Get confidence intervals for your sales forecasts: more about this in a few days; I will provide a very simple, model-free, data-driven solution to compute confidence intervals.

How to hire a good sales forecaster?

You need to hire some sort of a management consultant with analytic acumen, who will interact with all departments in your organization, gather, merge and analyze all data sources from most of your silos, integrate other external data sources (such as our forthcoming economic index), and be able to communicate both with executives and everybody in your organization who owns / is responsible for a data silo. He / She will recommend a solution. Conversations should include data quality issues, which metrics you should track moving forward, and how to create a useful dashboard for executives.

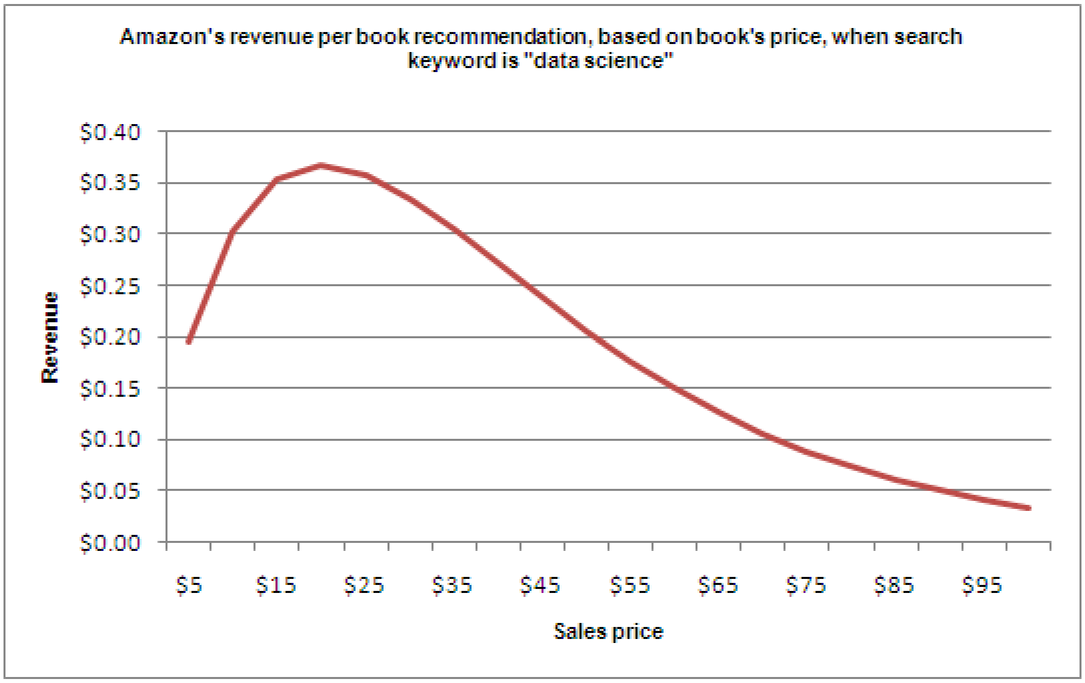
Are these data gurus expensive? Yes, they usually cost more than $150K/year in base salary, in United States. If your budget is limited, feel free to contact me at vincentg@datashaping.com: I work for free, and yes, there's a catch: I only work for projects that I am very interested in, and my solutions are eventually published in the Data Science book by Analyticbridge (although your company name will not be mentioned).

Read and contribute to discussion, at:

http://www.analyticbridge.com/profiles/blogs/sales-forecasts-how-to-improve-accuracy-while-simplifying-models

A.18. How could Amazon increase sales by redefining relevancy?

By improving its search and relevancy engines, to include item price as a main factor. The type of optimization and ROI boosting described below applies to all digital catalogs. Here we focus on books.



Search engine:
When you perform a keyword search on Amazon in the book section, Amazon will return a search result page with (say) 10 suggested books matching your search keyword. This task is performed by the search engine. The search engine will display the books in some order. The order is based either on price or keyword proximity.

Relevancy engine:
If you search for a specific book title, Amazon will also display other books that you might be interested in based on historical sales from other users. This task is performed by the relevancy engine, and it works as follows:

If m(A,B) users both purchased book A (the book you want to purchase) and another book B over the last 30 days, and if k(A) users purchased A, and k(B) users purchased B, then the association between A and B (that is, how closely these books are related from a cross-selling point of view) is defined as

$$R(A,B) = m(A,B) / SQRT\{k(A) * k(B)\}.$$

The order in which suggested books are displayed is entirely determined by the function R(A,*).

Better sorting criteria:

Very expensive books generate very few sales, but each sale generates huge profit. Cheap books generate little money, but the sales volume more than compensates for the little profit per book. In short, if you show books that all have exactly the same relevancy score to the user, the book that you should show up in the #1 position is the book with optimum price, with regard to total expected revenue. In the above chart, the optimum is attained by a booking selling for $21.

This chart is based on simulated numbers, assuming that the chance for a sale is an exponentially decreasing function of the book price. That is,

$$P(\text{sale} \mid \text{price}) = a * \exp(-b*\text{price})$$

A more general model would be:

$$P(\text{sale} \mid \text{price, relevancy score}) = a * \exp(-b*\text{price}) * f(\text{relevancy score})$$

Another way to further increase revenue is by including user data in the formula. A wealthy user has no problems purchasing an expensive book. Users who traditionally buy more expensive books should be shown more expensive books, on average.

Question:

When a sale takes place, how do you know if it is because of showing rightly priced books at the top, or because of perfect relevancy? For instance, relevancy between 'data science' and 'big data' is very good, but relevancy between 'data science' and 'cloud computing' is not as good. Does it make sense to suggest an expensive 'cloud computing' book to a wealthy user interested in a 'data science' book, or is it better to suggest a less expensive book related to 'big data', if your goal is to maximize profit? Separating the influence of relevancy from the price factor is not easy.

Note: the price factor is particularly useful when keyword or category relevancy is based on "small data".

Featured Comments:

[Vincent] Interestingly, my e-book entitled **Data Science by Analyticbridge** is now on Amazon, but when you search for "data science" on Amazon, it does not show up. Instead, other books not related to "data science" show up. Is it because I just uploaded the book a few days ago? If you search for "Analyticbridge" or "Vincent Granville" though, then my e-book does show up.

Read and contribute to discussion, at:

http://www.analyticbridge.com/profiles/blogs/how-could-amazon-increase-sales-by-redifining-relevancy

A.19. How to build simple, accurate, data-driven, confidence intervals

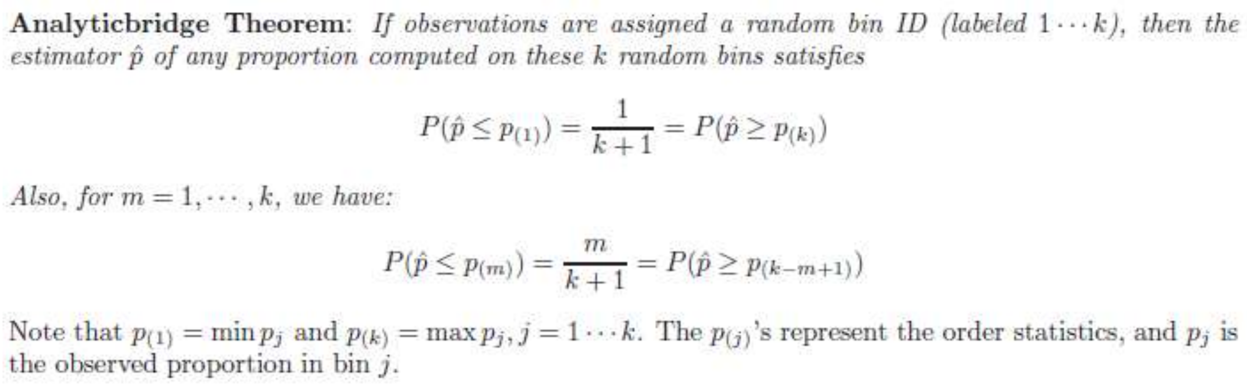
If observations from a specific experiment (for instance, scores computed on 10 million credit card transactions) are assigned a random bin ID (labeled 1, ⋯ ,k), then you can easily build a confidence interval for any proportion or score computed on these k random bins, using the Analyticridge theorem (see below).
The proof of this theorem relies on complicated combinatorial arguments and the use of the Beta function. Note that the final result does not depend on the distribution associated with your data - in short, your data does not have to follow a Gaussian (a.k.a normal) or any prespecified statistical distribution, to make the confidence intervals valid. You can find more details regarding the proof of the theorem in the book Statistics of Extremes by E.J. Gumbel, pages 58-59 (Dover edition, 2004)

Parameters in the Analyticbridge theorem can be chosen to achieve the desired level of precision - e.g a 95%, 99% or 99.5% confidence interval. The theorem will also tell you what your sample size should be to achieve a pre-specified accuracy level. This theorem is a fundamental result to compute simple, per-segment, data-driven, model-free confidence intervals in many contexts, in particular when generating predictive scores produced via logistic / ridge regression or decision trees / hidden decision trees (e.g. for fraud detection, consumer or credit scoring).

Application:
A scoring system designed to detect customers likely to fail on a loan, is based on a rule set. On average, for an individual customer, the probability to fail is 5%. In a data set with 1 million observations (customers) and several metrics such as credit score, amount of debt, salary, etc. if we randomly select 99 bins each containing 1,000 customers, the 98% confidence interval (per bin of 1,000 customers) for the failure rate is (say) [4.41%, 5.53%], based on the Analyticridge theorem, with k = 99 and m = 1 (read the theorem to understand what k and m mean - it's actual;ly versy easy to understand the signification of these parameters).

Now, looking at a non-random bin with 1,000 observations, consisting of customers with credit score < 650 and less than 26 years old, we see that the failure rate is 6.73%. We can thus conclude that the rule ***credit score < 650 and less than 26 years older*** is actually a good rule to detect failure rate, because 6.73% is well above the upper bound of the [4.41%, 5.53%] confidence interval.

Indeed, we could test hundreds of rules, and easily identify rules with high predictive power, by systematically and automatically looking at how far the observed failure rate (for a given rule) is from a standard confidence interval. This allows us to rule out effect of noise, and process and rank numerous rules (based on their predictive power - that is, how much their failure rate is above the confidence interval upper bound) at once.



Analyticbridge Theorem: *If observations are assigned a random bin ID (labeled $1 \cdots k$), then the estimator $\hat{p}$ of any proportion computed on these k random bins satisfies*

$$P(\hat{p} \leq p_{(1)}) = \frac{1}{k+1} = P(\hat{p} \geq p_{(k)})$$

Also, for $m = 1, \cdots, k$, we have:

$$P(\hat{p} \leq p_{(m)}) = \frac{m}{k+1} = P(\hat{p} \geq p_{(k-m+1)})$$

Note that $p_{(1)} = \min p_j$ and $p_{(k)} = \max p_j, j = 1 \cdots k$. The $p_{(j)}$'s represent the order statistics, and p_j is the observed proportion in bin j.

Read and contribute to discussion, at:

http://www.analyticbridge.com/profiles/blogs/how-to-build-simple-accurate-data-driven-model-free-confidence-in

A.20. Comprehensive list of Excel errors, inaccuracies and use of wrong / non-standard statistical definitions and formulas

The June 2008 issue of Computational Statistics and Data Analysis covered an analysis of Excel 2007. These faults and errors are being reviewed for inclusion.
Section 20 has been rewritten with new material based on errors pointed out by Paulo Tanimoto.
The section on RAND is not yet finished. B.D. McCullough has reported some uncertainties about RAND which are to be investigation.
Section 2 was expanded to cover the faults in Excel's "Order of Precedence"

Source: http://www.daheiser.info/excel/frontpage.html
If you have any comments, or have noted some errors or faults with Excel or my findings, please send it by email to d_heiser@att.net:

NOTES (Revised 6/25/08)

Note J: Ordinal, Nominal and Likert Scale Variables
Note K: New Display Modified Probability Distributions
Note L: Autocorrelation
Note M: An Actual Problem Requiring Unbiased Standard Deviations
Note N: Ranking, Quartiles, Medians and Percentiles
Note O: Averages, Standard Deviations and Pre-centering
Note P: Alternate Algorithms
Note Q: Data Entry For ANOVAs In The Data Analysis Tool Pac
Note R: Linear Regression
Note S: Linear Regression Throught The Origin, Excel 2000
Note T: Singularity, Multi-colinearity, Accuracy And Other Matrix Problems
Note U: Polynomial Regression
Note V: Regression Normal Probability Plot
Note W: Standardized Residuals
Note X: Support In Excel For Tests Of Significance
Note Y: Constructing A Hypothesis
Note Z: Setting Up The Excel Sheet For Calculating P Values
Note AA: Generate Diehard Test Input Files
Note AB: Diehard-II Output For RAND, Excel 2000
Note AC: Marsaglia's MWC256 RNG
Note AD: Diehard-II And -III Output For RAND, Excel 2003
Note AE: Random Number Generator VBA Routines
Note AF: The Wilkinson-Sawitski Series Of Tests On Excel 2007
Note XN: XNUMBERS, A Multi-precision Floating Point Calculus For Excel

Featured Comments:

[Vincent] Also see Raymond Panko's "What We Know About Spreadsheet Errors"
http://panko.shidler.hawaii.edu/SSR/Mypapers/whatknow.htm

Read and contribute to discussion, at:

http://www.analyticbridge.com/profiles/blogs/comprehensive-list-of-excel-errors-inaccuracies-and-use-of-wrong-

A.21. 10+ Great Metrics and Strategies for Email Campaign Optimization

This is our first article in a series about good actionable KPI's to optimize various ROI. Future articles will focus on metrics for fraud detection, user engagement etc. This one focuses on newsletter optimization. If you run an online newsletter, here are a number of metrics you need to track:

1. Open rate: proportion of uniques opening your newsletter. Anything below 10% is poor, unless your e-CPM is low.
2. Number of opens: some users will open your message multiple times.
3. Users opening more than 2 times: these people are potential leads or competitors. If very few users open more than once, your content is not interesting, or maybe there is only one clickable link in your newsletter.
4. Click rate: average number of clicks per open. If less than 0.5, your subject line might be great, but the content (body) irrelevant.
5. Open rate broken down per client (Yahoo mail, Gmail, Hotmail etc.) If your open rate for Hotmail users is very low, you should consider eliminating Hotmail users from your mailing list as it can corrupt your entire list.
6. Open rate and click rate broken down per user segment.
7. Trends: does open rate, click rate etc. per segment go up or down over time? Identify best performing segments. Send different newsletter to different segments.
8. Unsubscribe and churn rate. What subject line / content increase unsubscribe or complaint rate?
9. Spam complaints - should be kept to less than one per thousand messages sent. Identify segments and clients (e.g. Hotmail) generating high complaint rates, and remove them.
10. Geography: are you growing in India but shrinking in US? Is your open rate better in Nigeria? That's not a good sign, even if your overall trend looks good.
11. Language - do you have many Hispanic subscribers? If yes can you send a newsletter in Spanish to these people? Can you identify Spanish speakers (you can if you ask a question about language on sign-up)
12. Track open rate by day of week and time. Identify best times to send your newsletter.
13. User segmentation: ask many questions to new subscribers e.g. about their interests - make these questions optional. This will allow you to better target your subscribers.
14. Growth acceleration. Are you reaching a saturation point? If yes you need to find new sources of subscribers or reduce your frequency of email blasts (possibly finding fewer but better or more relevant advertisers to optimize e-CPM).
15. Are images causing low open rates? Are redirects (used to track clicks) causing low open rates? Some URL shorteners such as bit.ly, while very useful, can result in low open rate or people not clicking on links due to risk of computer infection.
16. Have you tracked keywords that work well or poorly, in the subject line, to drive your open rate up?
17. Have you tried changing your "from" field to see what works best? A/B testing could help you answer this question.
18. Size of message: if too large, could cause performance issue.
19. Format: text or HTML? Do some A/B testing to find optimum.

Note: e-CPM is the revenue generated per thousand impressions. It is your most important metric, together with churn.

Read and contribute to discussion, at:

http://www.analyticbridge.com/profiles/blogs/10-great-metrics-and-strategies-for-email-campaign-optimization

A.22. 10+ Great Metrics and Strategies for Fraud Detection

Emphasis here is on web log data. More than one rule must be triggered to fire an alarm. You may use a system such ashidden decision trees to assign a specific weight to each rule.

1. Monte Carlo simulations to detect extreme events. Example: large cluster of non-proxy IP addresses that have exactly 8 clicks per day, day after day. What is the chance of this happening *naturally*?
2. IP address or referral domain belongs to a particular type of blacklist, or whitelist. Classify the space of IP addresses into major clusters: static IP, anonymous proxy, corporate proxy (white-listed), edu proxy (high risk), highly recycled IP (higher risk), etc.
3. Referral domain statistics: time to load with variance (based on 3 measurements), page size with variance (based on 3 measurements), text strings found on web page (either in HTML or Javascript code). Create list of suspicious terms (viagra, online casino etc.) Create list of suspicious Javascript tags or codes but use white list of referral domains (e.g. top publishers) to eliminate false positives.
4. Analyse domain name patterns, example: a cluster of domain names, with exactly identical fraud scores, are all of the form xxx-and-yyy.com, and their web page all have the same size (1 char).
5. Association analysis: buckets of traffic with a huge proportion (>30%) of very short (< 15 seconds) sessions that have two or more unknown referrals (that is, referrals other than Facebook, Google, Yahoo or a top 500 domain). Aggregate all these mysterious referrals across these sessions - chances are that they are all part of a same Botnet scheme (used e.g. for click fraud).
6. Mismatch in credit card fields: phone number in one country, email or IP adress from a proxy domain owned by someone located in another country, physical address yet in another state, name (e.g. Amy) and email address (e.g. joy431232@hotmail.com) look very different, and a Google search on the email address reveals previous scams operated from same account, or nothing at all
7. Referral web page or search keyword attached to a paid click contains gibberish or text strings made of letters that are very close on the keyboard, such as fgdfrffrft.
8. Email address contains digits other than area code, year (e.g. 73) or zip-code (except if from someone in India or China)
9. Time to 1st transaction after sign-up is very short
10. Abnormal purchase pattern (Sunday at 2am, buy most expensive product on your e-store, from an IP outside US, on a B2B e-store targeted to US clients)
11. Same small popular dollar amount (e.g. $9.99) across multiple merchants with same merchant category, with one or two transactions per cardholder

Related articles:

- 10+ Great Metrics and Strategies for Email Campaign Optimization
- Real Life Example of Text Mining to Detect Fraudulent Buyers
- What do you think of this new data science startup idea?
- Spam detection for social networks: best practices (Part 1)
- How do you estimate the proportion of bogus accounts on Facebook?

Featured Comments:

[Amy] Gregory: possibly a bigger issue is how frequently lookup tables are updated. For instance, how often do you update your blacklist of IP addresses? Better blacklists should have several fields:

- IP address or range
- date when flagged
- date when flag should automatically clear
- reason code for flagging
- severity or score

[Amy] If you analyze referral data, scrape all referral domains (good and bad), and create a data dictionary containing all the terms found across all domain webpages, with a fraud score attached to each term.

Read and contribute to discussion, at:

http://www.analyticbridge.com/profiles/blogs/10-great-metrics-and-strategies-for-fraud-detection

A.23. Case Study: Four different ways to solve a data science problem

Here we discuss four approaches to solve the following marketing problem: identify, each day, the most popular Google groups, within a large list of target groups. You want to post in these groups only. The only information that is quickly available for each group is the time when the last posting occurred. Intuitively, the newer the last posting, the most active the group. There are some caveats such as groups where all postings come from one single user - a robot - for instance groups that focus on posting job ads exclusively. They should be in your black list.

So how do you estimate the volume of activity based on time-to-last-posting, for a particular group? This metric is actually what we want to guess, and rank groups according to estimated traffic volumes.

Four approaches can be used:

1. Intuitive (business analyst with great intuition)

The number of posts per time unit is roughly 2x the time since last posting. If you have a good sense of numbers, you just know that, even if you don't have an analytic degree. There's actually a simple empirical explanation to this. Probably very few people have this level of (consistently correct) intuition. Maybe none of your employees. If this is the case, this option (the cheapest of the four) must be ruled out.

2. Monte Carlo simulations (software engineer)

Any good engineer with no or almost no statistical training can perform simulations of random postings in a group (without *actually* posting anything), testing various posting frequencies, and for each test, pick up a random (simulated) time and compute time-to-last-posting. Then based on (say) 20,000 group posting simulations, you can compute (in fact reconstruct) a table that maps time-to-last-transaction to posting volume. Caveats: the engineer must use a good random generator and be able to assess the accuracy of his table, maybe building confidence intervals using the Analyticbridge theorem - a great and simple technique to use for non-statisticians.

3. Statistical modeling (statistician)

Based on the theory of stochastic processes (Poisson processes) and the Erlang distribution, the estimated number of postings per time unit is indeed 2x the time since last posting. The theory will also give you the variance for this estimator (infinite) and will tell you that it's much more robust to use time to 2nd or 3rd or 4th previous posting, which have finite and known variances. Now if the group is inactive, the time to previous posting itself can be infinite. In practice this is not an issue. Note that the Poisson assumption would be violated in this case. The theory will also suggest how to combine time to 2nd, time to 3rd and time to 4th previous posting to get a better estimator, read my paper Estimation of the Intensity of a Poisson process by means of neares... for details. You can even get a better estimator if instead of doing just one time measurement per day per group, you do multiple measurements per day per group and average them.

4. Big data (computer scientist)

You crawl all the groups every day and count all the postings for all the groups, rather than simply crawling the summary statistics. Emphasis is on using a distributed architecture for fast crawling and data processing, rather than a good sampling mechanism on small data.

Featured Comments:

[Capri] The best groups to target might not be the ones with highest volumes. Groups where posting occur every second are less valuable (for your marketing campaign) than groups where posts occur every 10 minutes. Another way to select groups is by checking your response rate to your own posts, and increase posting frequency in groups where response rate is higher... until you reach saturation, then you must pause postings for a while and resume later with a low frequency posting, steadily increasing as long as response rate (leads per posting or total leads) is good.

Read and contribute to discussion, at:
http://www.analyticbridge.com/profiles/blogs/four-ways-to-solve-a-data-science-problem-case-study

A.24. Email marketing: analytic tips to boost performance by 300%

This post is part of our blog post series on data science case studies and success stories. Analyticbridge improved open rates by 300%, and dramatically improved total clicks and click-through rates using the following strategies:

1. Remove subscribers who did not open the newsletter during the last 8 deployments
This produced a spectacular increase in open rate, and also significantly improved our "spam score", as our newsletter chances of ending up in a spam box or a spam trap is reduced to almost 0.

2. Segmentation of subscriber base to better target members, for instance to send a UK conference announcement to members located in Europe, but not to members in Asia or America.

3. Detect and stop sending messages that produce low open rates

4. Capturing member information (profession, experience, industry, location etc.) on sign up to create better segments.

5. Grow the list by

- offering great products for free to subscribers only: we are going to implement this idea with our data science eBook, requesting visitors to subscribe to our newsletter if they want to download our book.
- requesting our LinkedIn new members (> 600 per week) to become a member of Analyticbridge (the IEEE organization is using the same strategy)

6. A/B testing in real time when deploying a email blast: try 3 different subject lines with 3,000 subscribers, then use the subject line with highest click through rate for the remaining 40,000 subscribers.

7. Identify patterns in subject lines that work well

- research reports from well-respected companies
- case studies, success stories
- announcement about 10 great articles recently published in top news outlets
- salary surveys, job ads from great companies, data science programs from top universities
- not using the same great keywords over and over as it kills efficiency (poor subject titles that change each time work better than great subject titles that are over-used)

Example of successful articles recently posted:

- Four different ways to solve a data science problem - case study
- More resources for data scientists and analytic professionals
- Seven questions about real time analytics
- Data visualization: example of a great, interactive chart
- Three blog posts with tons of valuable information
- Free courses from top universities | Coursera.com

Read our previous case study: Quickly start and optimize keyword advertising campaigns on Google ...

Read and contribute to discussion, at:

http://www.analyticbridge.com/profiles/blogs/email-marketing-analytic-tips-to-boost-performance-by-300-case-st

A.25. Optimize keyword campaigns on Google in 7 days: an 11-step procedure

This is what I did, and it worked quite well.

1. Identify 10 top, high volume, well targeted keywords for your business. These are your seed keywords
2. Create large lists of bid keywords: use Google tools to identify "related keywords" that are related to your seed keywords
3. Grow further by finding keywords that are related to the keywords obtained in previous step
4. Use Google adWords tools to find keywords that are on your webpage, and most importantly, on your competitor web pages, choosing the website selection under "find keywords" under "add keywords" when you manage your ad group. Only pick keywords that have a moderate to high volume, and skip keywords with high competition (competition and volume statistics are provided by Google, on the same page).
5. Create a campaign, then add one ad group for all these 2,000 or so keywords. Select your target market (e.g. North America + Europe, Google Search + Partners Search) in your campaign settings. Set a max bid that is 3-4 times higher than what you want.
6. Create 4 or 5 different ads, so that at least one of them will have a great CTR and will boost click volume
7. After 4 hours, check your traffic report and pause all keywords with a CTR < 0.20% and impression count > 400. This will boost your CTR and will further improve your campaigns, and eventually your ROI. Then, reduce your max bid by 10%, as you are probably burning your budget too quickly
8. On day #2: Check again the stats: remove low CTR keywords, reduce max bid by 10%
9. On day #3: Check again the stats: remove low CTR keywords, reduce max bid by 10%.
   Then isolate keywords with high volume and good CTR, put them in a separate campaign or ad group. If you are still burning your daily budget too fast, improve targeting, by excluding some country, excluding display advertising or other sources of lower traffic quality
10. On day #4 to #7: proceed as for day #3, but be less aggressive in your efforts to reduce your max bid: if your CPC goes down too much too quickly, your campaign will collapse
11. By day #8, you should now have several ad groups, each maybe with 50 keywords, that are stable and great for your website, plus a huge ad group (the leftover of day #1-#7) that will continue to work as well, requiring occasional monitoring

Featured Comments:

[Vincent] Also, this is a quick, simple solution that works quite well. Ideally, if you have more time to work on your campaigns, each keyword (at least each ad group) should have its own landing page and max bid updated each day, and each ad group should have its own set of 4-5 text ads. Managing such a system, with (say) 10,000,000 keywords, can be done automatically. It's indeed one of the great applications of data science. The interesting part is to automatically assign a meaningful bid to new keywords with no history - and to permanently harvest (from your search logs?) and add many new keywords. Also, the top

100 "brand keywords" (with extremely large volume) should be processed manually each day, the remaining 9,999,980 being processed automatically. That's how it works at eBay.

Read and contribute to discussion, at:

http://www.analyticbridge.com/profiles/blogs/quickly-start-and-optimize-keyword-advertising-campaigns-on-googl

A.26. How do you estimate the proportion of bogus accounts on Facebook?

Facebook has 800MM users. Out of these 800MM "users", how many are duplicate (or triplicate), fake, dummy, inactive, decoy, stolen IDs, non-users (e.g. a book) and other artificial accounts?

How do you go about estimating this proportion? My guess is that less than 50% are unique, real, "non-dead" users (by "non-dead", I mean users with at least one activity over the last 6 months - such as logon, posting a message, inviting a friend, updating profile).

Featured Comments:

[Vincent] It does not matter how many accounts are bogus. The only thing that matters is impressions to clicks ratio - far below Google: it is very low on FB, but I'm not sure if it's due to the large number or artificial accounts. It has to do with sub-optimal ad targeting. Read Online advertising: a solution to optimize ad relevancy to find out how to optimize ad targeting.

Note: If you remove these artificial users, the value of a FB member increases from $4/year to $8/year

---

[Mirko] I did a quick test, creating five random names from scratch (vincent75, robert64, amy15, amy6, didierf) and checked their most recent activity on FB. Based on time since last action, 75% of the 4 existing profiles are active. Here are the results:

- vincent75 - redirects to vincent.75 - last posting Friday at 9:28am, seehttp://www.facebook.com/Vincent.75?sk=wall
- robert64 - redirects to robert.64 - updated his cover photo 16 hours ago
- amy15 - updated her cover photo on May 16
- amy6 - last activity on December 3 and thus technically inactive, see http://www.facebook.com/amy64?sk=wall (note that the profile is semi-private)
- didierf - profile does not exist

[Vincent] Very interesting Mirko! This could be a great project for a data science candidate (someone who wants to become a data scientist). Create 100,000 bogus names (with the help of an online dictionary and by adding combinations of digits at the end), see how many exists as FB profiles, and how many are active. Use a web crawler to complete the task, it should not take more than a day of work, including for the crawling activity (if it's organized using a rudimentary distributed architecture).

[Amy] Low impression-to-click ratio is good for the advertiser (it's like free branding) assuming CPC is the same as on Google, but it's bad for the publisher (Facebook) because it means that FB is doing a poor job at ad targeting. Either FB does not have enough rich ad inventory and thus targeting is difficult, or they have plenty of ads, and in this latter case, they'll make 10 times more money when they hire the right data scientist to help them with ad targeting optimization.

[Vincent] Here's another way Facebook generate revenue: when you post a Wall Street article on your FB timeline, any click that is generated results in a commission paid by the Wall Street Journal, to Facebook.

I checked a link to a WSJ article that I posted on my Facebook account, and magically, the following tags were added to the query string:
fb_rev=wsj_share_FB and *fb_source=timeline*.
The full link, on my FB page, is:
http://online.wsj.com/article/SB10001424052702303360504577408431211...
This brings an interesting issue: link fraud, by posting the same URL on various places, but substituting the tags by fake ones to claim the revenue: you need to be an approved WSJ publisher or sub-publisher or sub-sub-publisher to get the fraudulent credits, but you get the idea about how this fraud scheme would work.

This type of fraud could be motivated by different reasons, not necessarily with direct financial incentives. For instance, one might generate fake traffic (fake monitoring tags and/or fake referral)

- to kill a competing sub-publisher (in this case the fraudster hopes that the fraud scheme will be caught and attributed to the competitor),
- or for political reasons (e.g. someone who does not like a company's advertising campaigns and burns their advertising budget on fake traffic)
- or a smart kid who thinks that generating fake clicks is a challenging, fun and interesting project in itself

---

Read and contribute to discussion, at:

http://www.analyticbridge.com/forum/topics/how-do-you-estimate-the-proportion-of-bogus-accounts-on-facebook

Part II: Data Science Discussions

B.1. Statisticians Have Large Role to Play in Web Analytics

Read my full interview for AMSTATat http://magazine.amstat.org/blog/2011/09/01/webanalytics/. You will also find my list of recommended books. Here is a copy of the interview, in case the original article (posted on AMSTAT News) disappears.

(Dr. Granville's Interview for The American Statistical Association)

Vincent Granville is chief scientist at a publicly traded company and the founder of AnalyticBridge. He has consulted on projects involving fraud detection, user experience, core KPIs, metric selection, change point detection, multivariate testing, competitive intelligence, keyword bidding optimization, taxonomy creation, scoring technology, and web crawling.

Web and business analytics are two areas that are becoming increasingly popular. While these areas have benefited from significant computer science advances such as cloud computing, programmable APIs, SaaS, and modern programming languages (Python) and architectures (Map/Reduce), the true revolution has yet to come.

We will reach limits in terms of hardware and architecture scalability. Also, cloud can only be implemented for problems that can be partitioned easily, such as search (web crawling). Soon, a new type of statistician will be critical to optimize “big data” business applications. They might be called data mining statisticians, statistical engineers, business analytics statisticians, data or modeling scientists, but, essentially, they will have a strong background in the following:

- Design of experiments; multivariate testing is critical in web analytics
- Fast, efficient, unsupervised clustering and algorithmic to solve taxonomy and text clustering problems involving billions of search queries
- Advanced scoring technology for fraud detection and credit or transaction scoring, or to assess whether a click or Internet traffic conversion is real or Botnet generated; models could involve sophisticated versions of constrained or penalized logistic regression and unusual, robust decision trees (e.g., hidden decision trees) in addition to providing confidence intervals for individual scores
- Robust cross-validation, model selection, and fitting without over-fitting, as opposed to traditional back-testing
- Integration of time series cross correlations with time lags, spatial data, and events categorization and weighting (e.g., to better predict stock prices)
- Monte Carlo; bootstrap; and data-driven, model-free, robust statistical techniques used in high-dimensional spaces
- Fuzzy merging to integrate corporate data with data gathered on social networks and other external data sources
- Six Sigma concepts, Pareto analyses to accelerate software development lifecycle
- Models that detect causes, rather than correlations
- Statistical metrics to measure lift, yield, and other critical key performance indicators
- Visualization skills, even putting data summaries in videos in addition to charts

An example of a web analytics application that will benefit from statistical technology is estimating the value (CPC, or cost-per-click) and volume of a search keyword depending on market, position, and match type — a critical problem for Google and Bing advertisers, as well as publishers. Currently, if you use the Google API to get CPC estimates, Google will return no value more than 50% of the time. This is a classic example of a problem that was addressed by smart engineers and computer scientists, but truly lacks a statistical component—even as simple as naïve Bayes—to provide a CPC estimate for any keyword, even those that are brand new. Statisticians with experience in imputation methods should solve this problem easily and help their companies sell CPC and volume estimates (with confidence intervals, which Google does not offer) for all keywords.

Another example is spam detection in social networks. The most profitable networks will be those in which content—be it messages posted by users or commercial ads—will be highly relevant to users, without invading privacy. Those familiar with Facebook know how much progress still needs to be made. Improvements will rely on better statistical models.
Spam detection is still largely addressed using naïve Bayes techniques, which are notoriously flawed due to their inability to take into account rule interactions. It is like running a regression model in which all independent variables are highly dependent on each other.

Finally, in the context of online advertising ROI optimization, one big challenge is assigning attribution. If you buy a product two months after seeing a television ad twice, one month after checking organic search results on Google for the product in question, one week after clicking on a Google paid ad, and three days after clicking on a Bing paid ad, how do you determine the cause of your purchase?

It could be 25% due to the television ad, 20% due to the Bing ad, etc. This is a rather complicated advertising mix optimization problem, and being able to accurately track users over several months helps solve the statistical challenge. Yet, with more user tracking regulations preventing usage of IP addresses in databases for targeting purposes, the problem will become more complicated and more advanced statistics will be required. Companies working with the best statisticians will be able to provide great targeting and high ROI without “stalking” users in corporate databases and data warehouses.

Read and contribute to discussion, at:

http://www.analyticbridge.com/profiles/blogs/statisticians-have-large-role-to-play-in-web-analytics-american-s

B.2. Future of Web Analytics: Interview with Dr. Vincent Granville

Dr. Granville is the founder of Analyticbridge, the leading social network for analytic professionals, with more than 30,000 members. He has created several patents related to web traffic quality scoring, and he is an invited speaker at leading international data mining conferences. Vincent has consulted with Visa, eBay, Wells Fargo, Microsoft, CNET, LowerMyBills, InfoSpace and a number of startups on projects such as fraud detection, user experience, core KPIs, metric selection, change point detection, multivariate testing, competitive intelligence, keyword bidding optimization, taxonomy creation, scoring technology and web crawling.

Q: What is web analytics, vs. Advanced web analytics?

Web analytics is about extracting data, creating database schemas, defining core metrics that have potential to be exploited to increase profits or reduce losses, and reporting / visualization of summary statistics that could lead to actions, for instance detecting terrorists based on analyzing Twitter posts. This task is typically handled by business analysts and very senior software engineers, in collaboration with executive management.

Advanced web analytics is about designing pattern recognition algorithms and machine learning strategies that will actually catch terrorists and spammers, better target customers, create better advertising campaigns, make web sites easier to navigate, reduce churn, root cause analysis, etc. This task is handled by statisticians and scientists. Metrics used to measure success or improvement are called lift measures.

Q: How do you see the future of web analytics?

Integration of external data (harvested on the web, social networks and other sources) with internal corporate data, via fuzzy merging. Increased concern about scoring users, page views, keywords, referrals: not all page views are created equal. Text mining and taxonomy improvement. On-demand, web-based AaaS (Analytics as a Service) provided by programmable APIs that use scoring algorithms, able to process more than one million rows in real time. Also, blending technologies from fields as varied as biometrics, military intelligence, statistics, operations research, quant, econometrics, psychometrics, computer science, six sigma etc.

Q: What is your plan regarding Analyticbridge, with respect to the web analytics community?

We are growing fast and we want to reach as many web analytic professionals as possible, and provide them with valuable resources: jobs, courses, articles, news, think tank, software reviews, success stories, etc. We will continue to post more and more seminal articles and offer state-of-the-art technology to the community, such as HDT (hidden decision trees, designed in our research laboratory) as open source or co-branded with our partners.

Q: Which books, conferences, certifications and software do you recommend?

They are too numerous to mention. Visit our website to check new books and new journals, webinars, certifications, awards, etc. (www.analyticbridge.com). In terms of conferences, I recommend eMetrics, AdTech, SES, Predictive Analytics world, Text Analytics News and the SAS data mining conferences. Note that we offer a free web analytics certification based on your experience (minimum requirement is a master degree from a respected University). The Web Analytics Association also offers a certification.

Q: How did you started you career in web analytics?

I've been interested in mathematics for as long as I can remember, started a doctorate in computational statistics in Belgium in 1993, earning a postdoctoral degree at the statlabs at Cambridge University

(England), then moved to the states and got my first job with CNET then NBCI. Initially working in market research, then fraud detection, user behavior, traffic scoring and keyword intelligence. By 2007, I created Analyticbridge, one of the few profitable social networks.

Q: How do you participate in creating standards for our community?

I've patented a few scoring technologies and I continue to work on HDT and AaaS. I plan to deliver these technologies as open source. I've also designed countless metrics that can be used to assess lift in keyword campaigns: coverage or yield, keyword commercial value etc. Most importantly, I publish and present at conferences and discuss the correct methodology to use when dealing with sampling, Monte Carlo and model fitting. In particular, I've discussed at lengths about how to do correct cross-validation, how to compute meaningful confidence intervals for scores and why you need to provide them in the first place, and the importance of assessing the quality of your data sources through proper QA - and what to do when data is poor, wrong, or missing.

Q: Any success story you want to mention?

Detection of multiple Botnets generating more than $10 million yearly in fraud, resulting in developing sophisticated new rules involving association / collusion detection. Creation of a list of about 100,000 keywords representing 85% of the keyword pay-per-click commercial universe, in terms of Google advertising revenue. Currently working on a Google keyword price and volume forecaster. Developing scoring algorithms that are 200 times faster than algorithms available in the marketplace (without using cloud).

Q: 10 mistakes web analytics consultant should avoid?

- not listening well when discussing client requests
- trying to impress client with obscure terminology, rather than with past success stories expressed in simple English
- not understanding the big picture
- be limited to just one or two analytical techniques
- not using external data which could help detect flaws in client's internal data
- not understanding where the bias might be, not understanding the metrics well enough
- your model, no matter how good, can't be better than your data
- lack of cross-validation or improper cross validation
- failure to correctly deal with significant cross-correlations
- no plan for maintenance, or not updating data / model at the right frequency
- believing in the fact that R square is the perfect criterion for model validation
- ignoring or not properly detecting outliers
- using standard, black box techniques when robust, ad-hoc methodology should be preferred, or the other way around
- lack of good judgment / gut feelings, too much faith in data or model, or the other way around
- Ignore the 80/20 rule

Q: What do you suggest to new graduates?

Check certifications and training - visit our website, The Data Mining Blog, KDNuggets, Statistics.com, The Predictive Modeling Agency and Association websites: INFORMS, AMSTAT, ACM, WAA, SEMPO. Also get familiar with the Google Analytics andBing Intelligence blogs. Get an internship with a company that is good with web analytics. Download free data from the web (write your own web robot) and analyze it. Create your own web site or social network (check ning.com) and campaigns to have a feel about metrics and concepts such as user engagement, collaborative filtering, semantic web, page view value, churn etc. Indeed, one of the largest low-frequency click fraud Botnets ever detected was found by analyzing traffic from the small website I created in 1999. Download and try open source data mining software, e.g. Rapid Miner.

Read and contribute to discussion, at:

http://www.analyticbridge.com/profiles/blogs/future-of-web-analytics

B.3. Connecting with the Social Analytics Experts

Social Media Tips for Analytics Professionals

From Text and Data Mining to Market Research and Social Media Consulting, few are more influential than today's guests. In advance of the West Coast Text Analytics Summit (Nov. 10-11, San Jose), Text Analytics News caught up with four analytics leaders who are helping connect and educate text analytics professionals on the Web.

Tom H. C. Anderson
Managing Partner Anderson Analytics (OdinText)

The first marketing research firm to leverage modern text analytics, and currently in development of patent pending OdinText, Anderson Analytics has been a long time supporter of the Text Analytics Summit. CEO, Tom Anderson is the thought leader and proponent of social media analytics as well as advanced techniques in the field of marketing research. He founded and manages the largest and most engaged group of marketing researchers on the web, Next Gen Market Research (NGMR), as well as one of the first text mining groups on LinkedIn, Data & Text Analytics Professionals (DTAP).

- Blog (http://www.tomhcanderson.com/)
- NGMR on LinkedIn (http://www.linkedin.com/e/gis/31804)
- NGMR on Ning (http://www.nextgenmr.com/)
- DTAP on LinkedIn (http://www.linkedin.com/e/gis/22313)

Cliff Figallo
Senior Site Curator at **Social Media Today**

Editor and Community Manager **of Social Media Today**

Cliff Figallo has a long history of helping groups start online communities that will be both useful and lasting, and provides marketing analysis for the use of social media

Social Media Today is an independent, online community for professionals in PR, marketing, advertising, or any other discipline where a thorough understanding of social media is mission-critical. The site provides insight and hosts lively debate about the tools, platforms, companies and personalities that are revolutionizing the way we consume information. Content is contributed by members and curated by their editorial staff.

- Personal Blog http://www.cfigallo.com/

- Social Media Today http://socialmediatoday.com/

Vincent Granville

Chief Scientist at **LookSmart**

Chief Architect, Executive Director of **Analytic Bridge**

Dr. Vincent Granville has successfully solved problems for 15 years in data mining, text mining, predictive modeling, business intelligence, technical analysis, keyword and web analytics.
Most recently, he successfully launched DataShaping andAnalyticBridge, the largest social network for analytic professionals. Thanks to their network of talented Statistical Consultants, Data Shaping Solutions also offers a wide array of expertise in design of experiments, time series, predictive modeling, survey analysis, customer profiling, pattern recognition, statistical testing, and data mining across several industries.

- Data Shaping http://www.datashaping.com/index.shtml
- Analytics Bridge on NING http://www.AnalyticBridge.com
- Analytic Bridge on LinkedIn http://www.linkedin.com/groups?about=&gid=35222&trk=anet_ug...

Gregory Piatetsky-Shapiro

Founder **KDD/SIGKDD**

Editor, **KDnuggets** http://www.kdnuggets.com

Data Mining and Dr. Gregory Piatetsky-Shapiro are inextricably linked. Before staring KDnuggets he led data mining and consulting groups at GTE Laboratories, Knowledge Stream Partners, and Xchange.

He serves as the current Chair of ACM SIGKDD, the leading professional organization for Knowledge Discovery and Data Mining. He is also the founder of Knowledge Discovery in Database (KDD) conferences, having organized and chaired the first three Knowledge Discovery in Databases workshops. KDNuggets remains one of the Must Go To sites for the Data Mining Community.

Q. Why did you decide to start your social media group?

Gregory: I started publishing KDnuggets email Newsletter back in 1993, before the term social media existed, as a way to connect people who attended KDD-93, Knowledge Discovery in Data in workshop. From the beginning it was designed to have social content - people would contribute and my role would be as a moderator - select most relevant content and keep the spam away.
I added a website in 1994 and moved to current website www.KDnuggets.com in 1997.

In the last couple of years KDnuggets also added other social media channels (twitter, FB, LinkedIn), because this is where a lot of conversation in analytics space is happening. I find twitter.com/kdnuggets especially useful for broadcasting real-time or "smaller" items.

Tom: For much the same reason that I started Anderson Analytics in 2005. Coming from the Marketing Research/Consumer Insights industry I was frustrated by how slow my industry was in adopting new techniques especially in the area of data and text mining.

I founded Next Gen Market Research (NGMR) in 2007 for like minded marketing researchers, though the membership of about 13,000 professionals now include those in several other fields from Competitive and Business Intelligence to CRM and Web Analytics. Analytics is the common ground.

Vincent: The idea started after checking large social networks set up by recruiters on Ning.com, back in 2007. I had a fairly large network already at that time, I decided that it would be useful to create one big network for all analytic professionals, rather than multiple independent smaller communities (data miners, operations research, statisticians, quant, econometrics, biostatisticians etc.)

Cliff: I've been working in social media for 25 years as the technical environments have evolved. That's my profession, but the companies I've worked for have had various reasons for starting social media groups. In the current case, with Social Media Today, the founders recognized that there was value in providing central sites where writers on a range of specialties could share their ideas and talents with their natural professional communities.

I started a second group, Data & Text Analytics Professionals (DTAP) just a few days later for those of us who were more involved in the specifics of text analytics, that group now has well over 4,000 members.

Q. What kind of professionals tend to frequent your site?

Vincent: We see analytic professionals from government and from all industries (especially Finance, Health Care), as well as a good share of University students. Proportionally, consultants and startup executives are over-represented, while data miners from big corporations such as Google, Microsoft or IBM are under-represented. Job titles include web analyst, data analyst, analytic recruiter, database marketer, statistical programmer, military officer, head of analytics, scientist, VP of analytics, software architect, risk analyst, University professor, SEO or SEM expert, etc. According to Quantcast, our US demographics is as follows: 5 times more Asian than an average web site, 1.4 more in the 35-49 age range, 1.4 more with income above $100K, and of course 2.3 more with a graduate degree.

Tom: NGMR is still heavy marketing research. In our last survey we had almost an 20/80 Client to Supplier ratio which is far higher than the other groups. We were also the heaviest US based research group initially, but we have much more global representation now.

Our visitors come for the engaging discussion. There’s no other place like it, where you can ask a question on almost any analytics topic and expect to get several great answers in a few minutes. Many members also share Information via their blogs (http://www.tomhcanderson.com/next-gen-market-research-top-blogs/) or on Twitter, and the group now runs various competitions and is giving out our second annual innovation in research awards this fall.

Gregory: I have done a number of surveys of KDnuggets visitors and about 2/3 of them are technical people who analyze data, and about 10% analytics managers/directors. The rest are academic researchers and students.

Cliff: In the case of the Social Media Today site we attract practitioners in social media marketing, online community management, enterprise-level directors in marketing and PR, social applications development and business leaders looking for best practices in use of social media channels.

Q. What part does Text Analytics specifically play on your site?

Cliff: We realize the need for more sophisticated text analytics to better understand what attracts readers to our republished content. Our audience is looking for answers and out of hundreds of articles addressing "best practices for Facebook" (for example), we need to be able to drill down deeper than categories and tags can take us.

Gregory: I use web analytics to understand the behaviour of visitors to KDnuggets.
I have experimented with text analytics and word clouds many times, but found that the results were rather predictable with most important words being Data, Mining, Analytics, Jobs, Technology, etc . So, I am still looking for an effective way to use text analytics.

Vincent: We have a special group dedicated just to text mining, see http://www.analyticbridge.com/group/textmining. It features references, conferences, books and posting from members, including from myself. But many other text mining discussions are spread throughout the network, including in forums and groups such as Collective Intelligence and Semantic Web, Social Network Analytics, Web Analytics. Google analyticbridge+text+mining to find more" to find more. Also, many Analyticbridge members have included text mining or NLP in their domains of expertise, on their profile.

Tom: Text Analytics is often discussed more generally in NGMR where market researchers are trying to make sense of what social media monitoring tools to use/not use, and understand what role if any text analytics should play in their specific research area.

The DTAP group tends to get a lot more technical, though there are also a lot more text analytics suppliers who are competitors (including my own firm) in that group, so the conversation there tends to be a bit more academic relating to text analytics.

Q, In your opinion, what role does or should text analytics play in relation to social media?

Gregory: Text analytics is a key component of understanding social media, but it should also be integrated with social network analysis and data analytics.

Vincent: Better detection of spam, commercial or irrelevant posts. Also by clustering members or groups using text mining techniques, one could create segments which can then be used for highly targeting advertising.

Other areas of interests: crime and infringement detection based on analyzing and classifying posts, analyzing corporate image (what people think about your product or company), and leverage postings from social networks by blending this data with internal company data to create richer data sets. This means creating a strong structure on otherwise loosely structured data, using text mining technologies such as NLP, text clustering, and taxonomy creation..

Cliff: Text analysis can help organizations better understand their communities of customers, fans, advocates and colleagues by surfacing commonly-used phrases and memes. Revealing the juxtaposition of key terms across hundreds or thousands of posts and conversations would reveal deeper levels of shared experience and sentiment. It would also bring more understanding of disagreement and conflict within communities, telling organizations how to better address and serve people with varied attitudes toward an organizations products and services.

Tom: You really can't separate data and text mining, and both have a critical role in helping to leverage social media for insights. We're going to see more real time use of text analytics based models in the near future.

My problem is rarely convincing people that text analytics is critical for social media, but more often getting them to take a step back to realize where else they should be using it.

Q. What three pieces of advice would you give analytics professionals who are interested in participating more in social media?

Vincent: Start with a LinkedIn profile, join analytic groups on LinkedIn, and see what other members are postings before contributing to the discussions.

You can search LinkedIn groups by keywords: some of these groups have more than 20,000 members, some but not all are vendor-neutral, some are very specialized, and some are very good at filtering out spam. Then visit or sign-up with popular analytic networks such as KDNuggets, AnalyticBridge, SmartDataCollective, Quora. Check what your professional association offers in terms of networking.

Cliff: Participate regularly and deeply on social media platforms - immerse yourself in them so that you understand them. Put yourself in the role of a marketing or public relations person and ask the questions that they would have about mining conversational content.

Try to understand the difference between text as "content" and text as "conversation."

Gregory: Contribute - where you know the material and topics.
Learn from others - see what they do right. It is a constantly shifting landscape.
Have a sense of humor

Tom: Just do it!

Don't be afraid to ask questions.
Try to contribute, people really appreciate it.
Realize just like traditional networking it's a give and take, you need to be ready to return favors as well.

Read and contribute to discussion, at:

http://www.analyticbridge.com/profiles/blogs/connecting-with-the-social-analytics-experts

B.4. Interesting note and questions on mathematical patents

I was reading about the "*Automated Reduced Error Predictive Analytics*" patent secured by Rice Analytics (see below) and my first question is:

How can you successfully sue competitors about using a mathematical technology? After all, most vendors offer error and variance reduction as well as dimension reduction and automated model selection (based on optimizing goodness-of-fit) in their software. All statistical and data mining consultants, including myself, also use similar techniques to help solve business problems from their clients. For instance, I have developed methodology that achieves the same goal, and my methodology (hidden forests, see http://www.analyticbridge.com/forum/topics/hidden-decision-trees-vs) is public domain, non-patented, and everybody can use it freely.

Any claim about patent violation would most likely fail, the defendant's argument being "my algorithm is different, the only thing that our technology shares with the defendant's system is a methodology - well known and used by analytic professionals for decades - to reduce dimensionality, automate model selection and reduce error".

What about the newly recently published algorithm for random number generation based on the decimals of numbers similar to Pi (see http://www.analyticbridge.com/profiles/blogs/new-state-of-the-art-r...). This is public domain and non-patented. Could such a methodology be patented (assuming it would never have been published)? I don't think so, but would like to have your opinion on this.

The Rice Analytics Patent

Rice Analytics Issued Fundamental Patent on RELR Method

This Patent Covers RELR Error Modeling and Related Dimension Reduction

St. Louis, MO (USA), October 4, 2011 – Rice Analytics, the pioneers in automated reduced error regression, announced today the issuance to it by the US Patent Office for a patent for fundamental aspects of its Reduced Error Logistic Regression (RELR) technology. This patent covers important error modeling and dimension reduction aspects of RELR. Dan Rice, the inventor of RELR and President of Rice Analytics, stated the significance of this RELR patent as follows:

“While large numbers of patents are important in many technology applications, it is also clear that just one fundamental patent can lead to the breakthrough commercialization of an entire industry. The MRI patent in the early 1970’s had such an effect and by the 1990’s had resulted in billions of dollars in licensing fees and enormous practical applications in medicine. We believe that this RELR patent could have a similar effect in the field of Big Data analytics because RELR completely avoids the problematic and risky issues related to error and arbitrary model building choices that plague all other Big Data high dimensional regression algorithms. RELR finally allows Big Data machine learning to be completely automated and interpretable. Just as the MRI allowed the physician to work at a much higher level and avoid arbitrary diagnostic choices where two physicians would come to completely different and inaccurate diagnoses, RELR allows analytic professionals to work at a much higher level and completely avoid arbitrary guesses in model building. Thus, different modelers will no longer either build completely different models with the very same data or have to rely upon pre-filled parameters that are the arbitrary choices of others. Most modelers would spend significant time testing arbitrary parameters because they are worried about the large risk associated with such parameters, but then it is very hard for them to find the time to be creative. The complete automation that is the basis of RELR frees analytic professionals to work at a much higher and creative level, so they can pose better modeling problems and develop insightful model interpretations. Most importantly, unlike parsimonious variable selection in all other algorithms, RELR’s Parsed variable selection models actually can be interpreted because these models are not built with arbitrary choices and because they are consistent with maximum probability statistical theory.”

Read more about this patent at http://www.riceanalytics.com/_wsn/page9.html

Featured Comments:

[Vincent] @Daniel: While I have developed a few patents back in 2006 (related to Internet traffic scoring and decision trees), I moved from being a corporate scientist to becoming a publisher. As a result, I don't want to patent new techniques anymore, but instead my interest is to make my new analytic inventions freely available to a large audience of analytic professionals, in order to attract new members and thus advertisers. This could indeed create problems, as I might publish patented material without even knowing it. Since I am not paid by a University or any organization to do my own research (you could call me an *independent data scientist*), I need to do my research at the speed of the light. It took me 10 minutes to produce my new random number generator based on decimals of interesting numbers (similar to Pi), while it would take 3 years to a PhD student to develop the same ideas. In 10 minutes, there is no way I can check whether the idea is new or not, or whether it is already patented. In my quest to provide high-quality content to my readers, I might inadvertently, one day, reinvent the wheel and publish techniques very similar to what other people have already patented. As a publisher with little money, what would be the outcome, should this issue arise? It would be very easy to prove that what I published is not plagiarism.

Read and contribute to discussion, at:

http://www.analyticbridge.com/profiles/blogs/interesting-note-and-questions-on-mathematical-patents

B.5. Big data versus smart data: who will win?

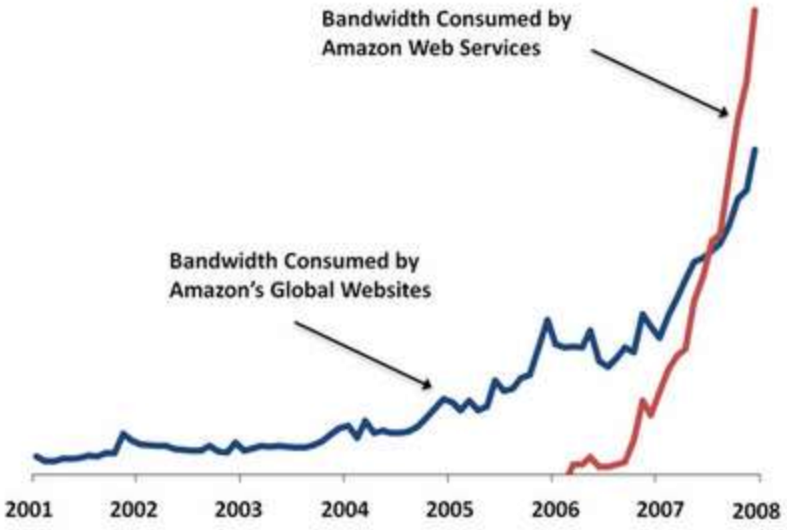
Much of the explosion of big data has been driven by increased efficiency in sever performance, memory cost, distributed architecture improvements (cloud, and truly parallel databases, e.g. noSQL) and essentially, by how much it costs to process a terabyte of data, both in terms of memory and bandwidth resources.

However, most of the very big data is very sparse, from an information point of view : big data is essentially made of noise or redundant information (think about videos or tweet data where information redundancy is huge) and can be compacted by 90-95% without any significant information loss. Storing and processing the entire data is a very inefficient process. I believe we can do much better by smartly sampling and smartly summarizing very big data (particularly stuff that is more than 4 week old) - a process known as data reduction or signal processing - rather than storing everything. The sampling / summarizing process is a task that should be left to expert, very senior statisticians, not to computer scientists.

At the end of the day, you should answer the following questions:

- How much lift or increased ROI / reduced risk do you get by storing everything, rather than storing the 5% "core" of your data (even if this means that you still store 100%, but only for the most recent 60 minutes, and less than 1% for data 5-week old and older). My guess is that you gain very little. But have you ever tested this?
- How much does it cost to store and keep everything, versus storing 5% of very carefully, smartly selected / sampled / summarized "core" data?
- What about keeping 5% core of your data, but in addition add 3 external big data sources for which you also only keep the core? Now you have potentially 4 times as much predictive power as before for 20% (20% = 4 x 5%) of the cost of storing all your internal big data, with very minimum information loss.

Think about this: to extrapolate how many users visit your very large website on a particular month, you don't need to store all user cookies for 28 days in a row. You can extrapolate by sampling 10% of your users, and sample 7 days (1 Monday, 1 Tuesday, 1 Wednesday, etc.) out of 28, and use a bit of statistical modeling and Monte Carlo simulations. So you can very accurately answer your question by using 40 times less data than you think.



Read and contribute to discussion, at:

http://www.analyticbridge.com/profiles/blogs/big-data-versus-smart-data-what-is-your-choice-do-you-think-smart

B.6. Creativity vs. Analytics: Are These Two Skills Incompatible?

In a nutshell, are great analytic people lacking creative skills? And are great creators lacking analytic skills? How to fix this gap?

Here are a few interesting questions:

- Should analytic people focus on measuring, and nothing else?
- Do analytic people lack business skills because they were odd kids in high school, because the way the school system is working in US?
- Is it impossible to hire great analytic people combining both soft and hard skills, because these very people are CEO's competing with your business and trying to kill you (business-wise) rather than work with you?
- Do you think analytic people should not be involved in providing ideas to improve business?
- Here are 5 ideas that were brought by analytic people:
    - Idea suggested to **Microsoft**: add an "advertise with us" link on your Bing.com front page
    - Idea suggested to **Colgate**: produce tooth paste with original flavors, for kids and for people who do not like fake mint
    - Idea suggested to **Apple**: turn off spell checker on web search, but turn it on when writing an email
    - Idea suggested to **Google**: index "related web pages", not just keywords, so that people can easily find "related links" as opposed to doing a pure keyword search - with all the limitations associated with keyword search
    - Idea suggested to the **FBI**: use decoy bank accounts to catch Nigerian and other fraudsters

Do you think analytic people should not be involved in providing this type of insights to corporate executives, and if not, who should?

Featured Comments:

[Vincent] You can deploy creativity not just to solve business problems, but to analytics itself. For instance, I believe that my new random number generator
(see http://www.analyticbridge.com/profiles/blogs/new-state-of-the-art-r...) is the result of thinking creatively, but NOT analytically.

Finally, too much creativity will cause you problems, be it in the corporate world or academic research. Highly creative and analytic people are better off being entrepreneur (although you will need additional skills - social skills, sales - unless your business involves no human interaction / no client, such as day trading with your own money)

The most successful analytic professionals have developed great craftsmanship: that's something in-between science and art, something that you can't possibly learn in a university curriculum.
But something not unlike what it takes to be a great cook or a great carpenter.

[Vincent] Great analytic professionals are not just data scientists; they are essentially both data and business architects, at the same time.

[Lisa] It is my opinion that what may ***appear*** to be creativity is **actually** deep subject matter expertise.

Formal education (supplemented with training/ certification) is how an analyst acquires technical skills. Content familiarity is acquired in a less structured way, over the passage of time. It takes awhile to acquire **both** knowledge sets! What may seem like "lack of creativity" in some analysts can be due to this: Analytic skills are versatile, applicable in many fields, industries. So minimal content knowledge is necessary to do an adequate job. That can cause the impression that those analysts are not creative.

It is too simplistic to think that quantitative analysts are unable to provide new product ideas, conceptual insights etc. A quantitative analyst with content knowledge **IS** capable of providing creative insights to corporate executives.

Read and contribute to discussion, at:

http://www.analyticbridge.com/profiles/blogs/creativity-vs-analytics-are-these-two-skills-incompatible

B.7. Barriers to hiring analytic people

You hear all the time that it is very hard to find and hire a great data scientist. Yet these scientists can't find a job, and are typically unemployed for many months after graduating or after being laid off - even those with 15 years of experience and stellar degree and accomplishments.

So what's the problem? I'm suggesting a few possibilities:

- Candidates lack business acumen
- Candidates have poor communication skills, are asking too much money
- Candidates have outdated skills - or University is teaching outdated material
- Recruiters are very slow in the recruiting process - eventually candidates evaporate
- Recruiters are very concerned about hiring because of the economy
- Candidates should not apply for a job, but instead create their own career or enterprise
- Recruiters don't know how to measure added value provided by analytic talent
- Certifications, regulations (e.g. regarding data privacy) or US citizenship requirements is a barrier
- Candidates will not relocate because they can't sell their house
- Too many analytic people, we should discourage prospective students to pursue an analytic career

What do you think?

Featured Comments:

[Vincent] See another post on the subject: http://www.analyticbridge.com/profiles/blogs/new-strategies-to-get-...

I think candidates can feel there's no job because they are still looking in the wrong places (big job boards) or still act like 10 years ago: send a resume, and wait. This approach does not work anymore. And recruiters can have a feeling that candidates are rare because they've moved to different places (like Analyticbridge!) where candidates are of much better quality. Indeed, smart candidates post great articles and get hired without ever submitting a resume or applying for a job.

But barriers exist, e.g. for some analytic jobs you must be a US citizen. It makes it very hard, if you are recruiting analytic people with security clearance, to find available people. But in this case, the scarcity is created by artificial conditions (regulation).

[Amy] @Jozo: I think in this economy, the skill #1 to succeed is sales, regardless of your profession.
Being freelancers or consultant is the first step towards entrepreneurship.
Analytic people can also find work that require no interaction with human beings, and thus no sales. But competition for these jobs is fierce. These jobs include arbitrage (on the stock market or in pay-per-click advertising), sport betting, etc.

Read and contribute to discussion, at:

http://www.analyticbridge.com/profiles/blogs/barriers-to-hiring-analytic-people

B.8. Salary report for selected analytical job titles

Source: Indeed. Click on the links below or visit http://www.indeed.com/salary to get more granular data (by region, degree, years of experience):

Predictive Analytics Expert	$61,000
Web Analytics Specialist	$77,000
Web Analyst	$67,000
Director of Analytic	$113,000
Seo Manager	$78,000
Quantitative Analyst	$96,000
Senior Data Architect	$121,000
Marketing Analyst	$54,000

Featured Comments:

[Vincent] Here are some numbers from 2007, broken down per metro area:

Keyword	City	Salary	Nationwide	Delta
quant	Atlanta	108,000	98,000	10%
statistician	Atlanta	78,000	70,000	11%
senior statistician	Atlanta	82,000	74,000	11%
sas	Atlanta	80,000	72,000	11%
biostatistician	Atlanta	86,000	77,000	12%
data mining	Atlanta	79,000	71,000	11%
data analyst	Atlanta	64,000	58,000	10%
quant	Boston	116,000	98,000	18%
statistician	Boston	84,000	70,000	20%
senior statistician	Boston	88,000	74,000	19%
sas	Boston	86,000	72,000	19%
biostatistician	Boston	92,000	77,000	19%
data mining	Boston	85,000	71,000	20%
data analyst	Boston	69,000	58,000	19%
quant	Chicago	113,000	98,000	15%
statistician	Chicago	82,000	70,000	17%
senior statistician	Chicago	85,000	74,000	15%
sas	Chicago	83,000	72,000	15%
biostatistician	Chicago	89,000	77,000	16%
data mining	Chicago	83,000	71,000	17%
data analyst	Chicago	67,000	58,000	16%
quant	Denver	88,000	98,000	-10%
statistician	Denver	63,000	70,000	-10%
senior statistician	Denver	66,000	74,000	-11%
sas	Denver	65,000	72,000	-10%
biostatistician	Denver	69,000	77,000	-10%
data mining	Denver	64,000	71,000	-10%
data analyst	Denver	52,000	58,000	-10%
quant	Los Angeles	95,000	98,000	-3%

statistician	Los Angeles	69,000	70,000	-1%
senior statistician	Los Angeles	72,000	74,000	-3%
sas	Los Angeles	70,000	72,000	-3%
biostatistician	Los Angeles	75,000	77,000	-3%
data mining	Los Angeles	70,000	71,000	-1%
data analyst	Los Angeles	56,000	58,000	-3%
quant	New York	122,000	98,000	24%
statistician	New York	88,000	70,000	26%
senior statistician	New York	92,000	74,000	24%
sas	New York	90,000	72,000	25%
biostatistician	New York	96,000	77,000	25%
data mining	New York	89,000	71,000	25%
data analyst	New York	72,000	58,000	24%
quant	Philadelphia	106,000	98,000	8%
statistician	Philadelphia	77,000	70,000	10%
senior statistician	Philadelphia	80,000	74,000	8%
sas	Philadelphia	78,000	72,000	8%
biostatistician	Philadelphia	84,000	77,000	9%
data mining	Philadelphia	78,000	71,000	10%
data analyst	Philadelphia	63,000	58,000	9%
quant	San Diego	94,000	98,000	-4%
statistician	San Diego	67,000	70,000	-4%
senior statistician	San Diego	71,000	74,000	-4%
sas	San Diego	69,000	72,000	-4%
biostatistician	San Diego	74,000	77,000	-4%
data mining	San Diego	68,000	71,000	-4%
data analyst	San Diego	55,000	58,000	-5%
quant	San Francisco	120,000	98,000	22%
statistician	San Francisco	87,000	70,000	24%
senior statistician	San Francisco	91,000	74,000	23%
sas	San Francisco	89,000	72,000	24%
biostatistician	San Francisco	95,000	77,000	23%
data mining	San Francisco	88,000	71,000	24%
data analyst	San Francisco	71,000	58,000	22%
quant	Seattle	95,000	98,000	-3%
statistician	Seattle	68,000	70,000	-3%
senior statistician	Seattle	71,000	74,000	-4%
sas	Seattle	70,000	72,000	-3%
biostatistician	Seattle	75,000	77,000	-3%
data mining	Seattle	69,000	71,000	-3%
data analyst	Seattle	56,000	58,000	-3%
quant	Washington	90,000	98,000	-8%
statistician	Washington	65,000	70,000	-7%
senior statistician	Washington	68,000	74,000	-8%
sas	Washington	66,000	72,000	-8%
biostatistician	Washington	71,000	77,000	-8%
data mining	Washington	66,000	71,000	-7%
data analyst	Washington	53,000	58,000	-9%

Source: Internal report, September 2007 (www.datashaping.com/salaries.shtml)

[Vincent] See also:

- Salary surveys for actuaries and statisticians
- Hourly rates for statistical consultants: $89 to $189 | AMSTAT survey (2006) - Going Rates for Statistical Consulting: Results from the Statistical Consulting Section Rates Survey.
- http://www.payscale.com/ (search for *Analytics* in the career search box)

Read and contribute to discussion, at:

http://www.analyticbridge.com/group/salary-trends-and-reports/forum/topics/salary-report-for-selected-analytical-job-titles

B.9. Are we detailed-oriented or do we think "big picture", or both?

Hiring managers always assumed that I was a very detailed-oriented person. It turns out that this is not the case: I'm certainly a very analytic person, yet I always think "big picture", and everybody who knows me well would say that I am everything but detail-oriented.

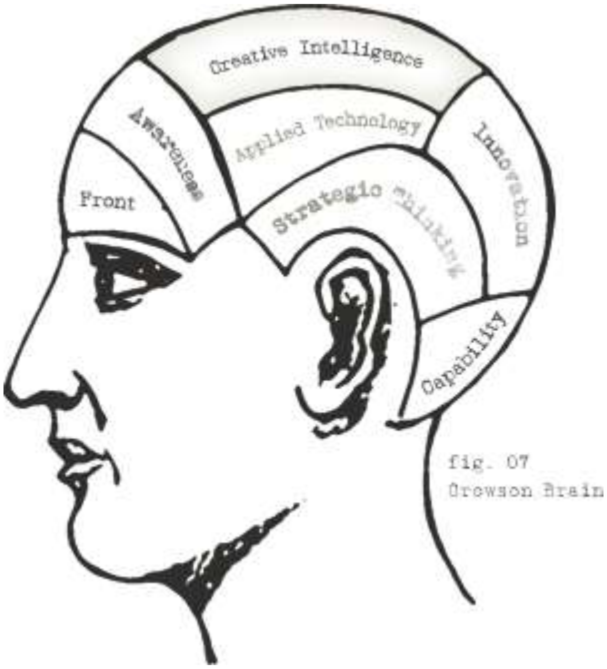


fig. 07
Crowson Brain

Since I clearly lack this skill, I surround myself with truly detail-oriented people, in particular for filing tax forms and keeping track of financial transactions. I do have a number of strengths, being detailed-oriented is just not one of them:

- I'm good at being organized (my desk looks like a pile of garbage, but it is extremely organized garbage - please don't try to clean it)
- I conceive, design (algorithms or great visual dashboards), model, investigate, solve, optimize
- I use good judgment more so than insights from statistical reports, to boost revenue and growth
- I use craftsmanship skills more so than recipes learned at university or some other training, to manage my operations / client operations
- My brain uses fuzzy logic to solve problems
- I don't like coding (except in Perl), but I have designed a few systems, even distributed systems, without writing one line of code
- I have a strong sense of intuition, and can "feel" patterns in big data sets using simple visual techniques or data dictionaries
- I sometimes "feel" what the future will be (for a specific issue), without running any predictive models.
- Typically, I combine statistical modeling with intuition and use of external data sources. For instance, I like to combine data from sales, advertising, finance, sales from competitors, social networks (to measure brand trends), industry trends, external data agencies (providing market share trends), economic data, etc. to answer a question about sales forecasts.

I believe that the above strengths are more typical of a senior analytic professional, and I want your opinion on this.

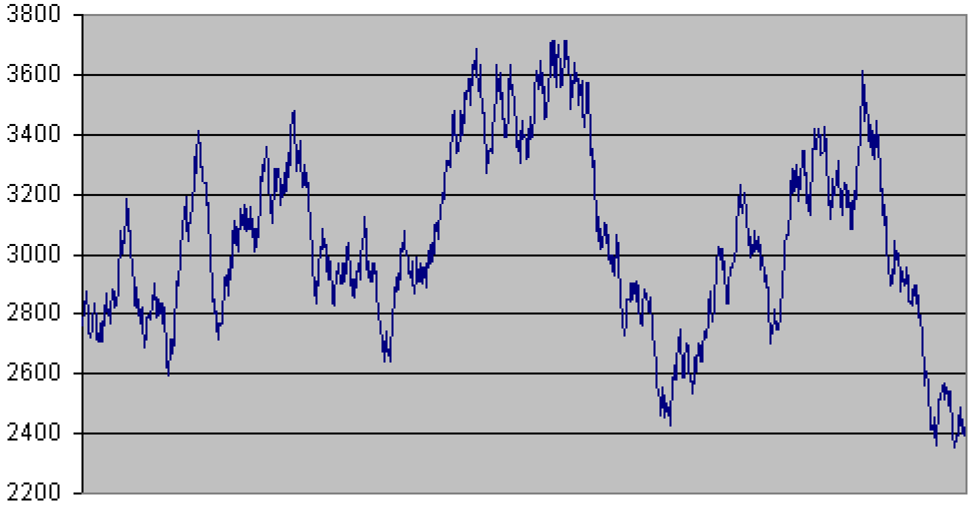
Question: Do you think that typical analytic professionals are detailed-oriented but lack the big picture and vision? Or do you think it is possible to have both (I don't think so). Or do you think there are two types of analytic professionals: detailed-oriented vs. visionaries / big picture thinkers.

Read and contribute to discussion, at:

http://www.analyticbridge.com/profiles/blogs/analytic-professionals-are-we-detailed-oriented-or-do-we-think-bi

B.10. Why you should stay away from the stock market?

The stock market is much more random than what most investors think. Indeed, it looks almost like a random walk: even if it has been going down 7 days in a row, the probability that it will go dow tomorrow is still 50%: random walks are memory-less processes. The chart below illustrates a simulated, perfectly random / unpredictable stock market.



Explanations:

- **Why can't you consistently win?** Events that can boost or kill the stock markets are quickly exploited by professional traders. This explains the gap between yesterday close price, and today open price. Indeed many traders don't keep open position overnight because of this.
- For large indices (QQQQ etc.), the **correlation with today's price and the price from yesterday** or from k days ago (k = 1, 2, 3, etc.) is essentially zero. If it was significantly different from zero (from a statistical point of view), then this pattern would have been exploited by professional traders - many holding a MIT or Princeton PhD in mathematics - and eventually the pattern would die. Note that the time series correlogram (some call it the spectral signature) associated with stock prices uniquely identifies the nature of the underlying stochastic process. Time series with a flat correlogram represent pure random walks, that is, processes where you can't predict the future.
- **So is the stock market truly random?** No. It is true that when a large number of independent algorithms (run by hedge funds) compete in real time to extract money from Wall Street, eventually nobody wins or lose - and the stock market becomes as random as a lottery: some can win for some time, but on the long term they can't. The return becomes zero. Think of your daily commute when you are stuck on highways driving at 5 miles an hour: can you find a pattern to beat the rat race, maybe a new road that you can use to beat your fellow commuters? If you do, that road will be found by thousands of commuters fighting the same battle to minimize commute time, and your miracle road will turn into rat race again, in a matter of days. The same mathematical principles (arbitrage) apply to the stock market. However, even if you exclude insider trading and after hours trading, the market, while almost random, is not entirely random. Most of Wall Street trading algorithms have been designed by mathematicians trained in the same universities, and are not independent: it is like all search engine algorithms providing the same search results for most keyword searches. This indeed explains why the stock market can experience flash crashes or flash spikes.

- **This discussion about randomness applies to short-term trading**. But is there a long-term trend that can be leveraged? In my opinion, since year 2000, the stock market has no up or down trends: while highly volatile, it is mostly flat. I think there's a good chance for a slow downward trend given the fact that baby boomers are retiring, demographic pressures, and the young generation having lost trust in 401K plans. Yet government might impose some regulations to artificially keep the market from collapsing - otherwise long term shorting would be a great option for the smart people who are currently "all cash".
- **Does it mean that there's no money to be made for the average investor**? Well, it is becoming more like a lottery, and a few people win money, sometimes big money, in all lotteries. Look at my above chart showing a purely random (simulated) stock market: there are local trends, patterns such as short squeeze, "head and shoulder", market crash, steady growth, runs (going up or down for 7 days in a row). Every kind of pattern can be found in this simulated, random, trend-less stock market, just like in the real stock market. It is clear that you could make money in this simulated market - what is less clear is that you have no way to predict if your algorithm will or will not make money. Note that to increase your odds, you have to back test your strategies and perform sound cross-validation or "walk-forward" (http://en.wikipedia.org/wiki/Walk_forward_optimization) to predict what your return might be.
- Despite the increasingly random nature of the stock market, **it is still possible, for the average educated trader, to exploit patterns**. Typically, any pattern that Wall Street is reluctant to exploit - such as staying all cash for 5 years in a row and trading like crazy when the right opportunity presents itself. In a nutshell, Wall Street traders must think very short term or otherwise risk losing their job. This provides an opportunity for the amateur trader. As far a short-term patterns are concerned, we've found that they occur no more than three times, and can be exploited only during the third occurrence (the first two occurrences being used for pattern detection and confirmation): after 3 occurences, Wall Street gurus have exploited and killed the pattern (the patten - e.g. a short squeeze on a particular stock - might still be there, but it's parameters have shifted on the fourth occurence, due to heavy exploitation).
- **The future could be rosier**: with more traders leaving this extraordinary competitive environment (e.g. they leave because they are laid off due to poor performance) and less money circulating in the stock market due to baby boomers retiring, the nature of the stock market will become less random, whether the general market trend is up, down or neutral. This will provide new opportunities for traders who haven't been killed in the 2000-2010 "wild fire".

Featured Comments:

[Allan Miler] Lyapunov exponents in this system are probably positive, where the time constant are getting shorter and shorter the instability will get getting and greater. Chaotic behavior and mode competition is something I would expect to see more often especially with automated trading. There comes a point when all these systems oscillate between each other as competing hedge fund super computers try and out manipulate each other...

It's not really random it just looks that way in the linear world. If someone has the time use Wigner Voltera time series reconstruction and you would probably be able to recover he chaotic dynamics underneath the noise. Its all non linear, don't waste your time on linear analysis if you want to understand it, your variances will be too high to make sense of it all.

Read and contribute to discussion, at:

http://www.analyticbridge.com/forum/topics/why-you-should-stay-away-from-the-stock-market-unless-you-read-th

B.11. Gartner Executive Programs' Worldwide Survey

Gartner Executive Programs' Worldwide Survey of More Than 2,300 CIOs Shows Flat IT Budgets in 2012, but IT Organizations Must Deliver on Multiple Priorities

Survey Shows CIOs are Using Technology to "Amplify" the Enterprise

IT organizations will have to deliver on multiple priorities without an increase in their IT budget, as CIO IT budgets are expected to be flat, increasing just 0.5 percent, with declining IT budgets in North America and Europe, according to a global survey of CIOs by Gartner, Inc.'s Executive Programs.

The worldwide CIO survey was conducted in the fourth quarter of 2011, and it included 2,335 CIOs, representing more than $321 billion in CIO IT budgets and covering 37 industries in 45 countries. The Gartner Executive Programs report, "Amplifying the Enterprise: The 2012 CIO Agenda" represents the world's most comprehensive examination of business priorities and CIO strategies.

"Technology's role in the enterprise is increasing. This does not mean, however, that the role of the IT organization is increasing," said Mark McDonald, group vice president for Gartner Executive Programs and Gartner Fellow. "CIOs concentrating on IT as a force of operational automation, integration and control are losing ground to executives who see technology as a business amplifier and source of innovation. Effective leaders use technology, which includes IT, to strengthen the customer experience and eliminate costly internal distortions. They are using technology to 'amplify' the enterprise."

"In the face of continued economic uncertainty and government austerity, business strategies call for a combination of growth and operational efficiency. As reflected in the 2012 CIO Agenda survey findings, effective leaders see customers as the key factor in both of these strategic components, with the customer experience their focal point in reconciling potentially conflicting goals," Mr. McDonald said. "Present economic conditions may tempt CIOs to force IT back into cost-cutting mode, but senior executives expect technology — and this includes IT — to address the tough challenges by amplifying enterprise strategies and operations."

CIO's increasingly see technologies such as analytics/business intelligence, mobility, cloud and social in combination rather than isolation to address business priorities. Changing the customer experience requires changing the way the company interacts externally rather than operates internally.

Analytics/business intelligence was the top-ranked technology for 2012 (see Table 1) as CIOs are combining analytics with other technologies to create new capabilities. For example, analytics plus supply chain for process management and improvement, analytics plus mobility for field sales and operations, and analytics plus social for customer engagement and acquisition.

Table 1
Top 10 CIO Business and Technology Priorities in 2012

Top 10 Business Priorities	**Ranking**	**Top 10 Technology Priorities**	**Ranking**
Increasing enterprise growth	1	Analytics and business intelligence	1
Attracting and retaining new customers	2	Mobile technologies	2
Reducing enterprise costs	3	Cloud computing (SaaS, IaaS, PaaS)	3
Creating new products and services (innovation)	4	Collaboration technologies (workflow)	4
Delivering operational results	5	Virtualization	5
Improving efficiency	6	Legacy Modernization	6
Improving profitability (margins)	7	IT Management	7
Attracting and retaining the workforce	8	CRM	8
Improving marketing and sales effectiveness	9	ERP Applications	9

Expanding into new markets and geographies	10	Security	10

Source: Gartner Executive Programs (January 2012)

Sixty-one percent of enterprises responding to the survey say they will be improving their mobile capability over the next three years. The majority have a mobility strategy that calls for becoming a market leader in their industry — so there will be significant competition as everyone seeks to be "above average" in its industry.

Overall, CIOs rank growth as their top priority — despite tough economic conditions and future uncertainties. They are particularly attentive to attracting and retaining customers and to creating products and services.

Meeting business expectations for increased growth, reduced cost or a transformed customer experience normally involves a significant increase in IT resources. Forty-six percent of CIOs reported that their CIO IT budget would increase from 2011 to 2012 in terms of actual spending. The average firm in this year's survey will see a modest budget increase of between 2 and 3 percent.

On a global weighted average basis, CIO IT budgets are anticipated to be essentially flat for 2012. These investments are strongest among enterprises in Latin America (with a 12.7 percent IT budget increase) and the Asia/Pacific region (with a 3.4 percent increase), while investments are weakest among the largest enterprises in North America (decreasing 0.6 percent) and Europe (down 0.7 percent). Larger organizations, those with IT budgets more than $500 million, have continued to cut their IT expenditures, offsetting modest growth in the rest of the survey population.

"The 2012 Gartner CIO Agenda survey results show that CIOs believe that the customer experience is the greatest opportunity for IT-enabled innovation," said Dave Aron, vice president and Gartner Fellow. "As business executives see the potential of technology to transform customer channels and the customer experience, their view of technology has leapfrogged conventional ideas of IT."

Technology is playing an increasing role in enterprise growth, innovation and operational performance while technology's definition now incorporates new combinations of traditional IT systems, consumer devices and their respective services.

"Applying technology as part of amplifying the enterprise reflects both the changing nature of business strategies, and executive expectations about the role of technology in realizing those strategies. Amplifying products, services and operations requires an enterprise to strengthen the customer experience and send clearer market signals," Mr. McDonald said. "Mobility, social media, information and analytics can be used to re-imagine the customer experience, as well as sales and service channels. These technologies do more than automate or administer processes; they are the processes and the sources of value."

About Gartner Business Intelligence Summit

Business intelligence and analytics is the number one technology priority for 2012, followed by cloud and SaaS at number three, according to a new Gartner survey of more than 2,300 CIOs. To deliver the analytics and insight your business needs, join us for the upcoming **Gartner Business Intelligence Summit**, **April 2 – 4, in Los Angeles, CA**. The summit presents the latest research and explore new BI best practices, including how to:

- Develop an effective BI strategy that improves the business
- Assess the value of cloud and SaaS offerings
- Manage big data challenges and improve data quality

- Prepare for next generation analytics

The spotlight is on BI and cloud, where innovation has opened the door to new efficiencies, capabilities and business improvements. CIOs have taken notice; now it’s your job to lead the way. With trends evolving at high speed, you need information based on leading-edge research. At the upcoming summit, you’ll meet the best in the business, hear the latest case studies and gain the informed perspective essential to making today’s pressing BI decisions.

B.12. Gartner Says Big Data Creates Big Jobs: 4.4 Million IT Jobs Globally to Support Big Data By 2015

Worldwide IT spending is forecast to surpass $3.7 trillion in 2013, a 3.8 percent increase from 2012 projected spending of $3.6 trillion, but it's the outlook for big data that is creating much excitement, according to Gartner, Inc.

"By 2015, 4.4 million IT jobs globally will be created to support big data, generating 1.9 million IT jobs in the United States," said Peter Sondergaard, senior vice president at Gartner and global head of Research. "In addition, every big data-related role in the U.S. will create employment for three people outside of IT, so over the next four years a total of 6 million jobs in the U.S. will be generated by the information economy."

"But there is a challenge. There is not enough talent in the industry. Our public and private education systems are failing us. Therefore, only one-third of the IT jobs will be filled. Data experts will be a scarce, valuable commodity," Mr. Sondergaard said. "IT leaders will need immediate focus on how their organization develops and attracts the skills required. These jobs will be needed to grow your business. These jobs are the future of the new information economy." (see Mr. Sondergaard's commentary on the Gartner YouTube channel).

Mr. Sondergaard provided the latest outlook for the IT industry today to an audience of more than 8,000 CIOs and IT leaders at Gartner Symposium/ITxpo, which is taking place here through October 25. He said the IT industry is entering the Nexus of Forces, which includes a confluence and integration of cloud, social collaboration, mobile and information.

"This is a time of accelerating change, where your current IT architecture will be rendered obsolete," Mr. Sondergaard said. "You must lead through this change, selectively destroy low impact systems, and aggressively change your IT cost structure. This is the New World of the Nexus, the next age of computing."

Cloud

The cloud is the carrier for the three other Forces: mobile is personal cloud, social media is only possible via the cloud, and big data is the killer app for the cloud. Cloud will be the permanent fixture, the foundation.

"Cloud is not merely about cost-cutting, the end game is not just cheap on-demand services. In fact, 90 percent of these services are still subscription based, not pay-as-you-go," Mr. Sondergaard said. "We are just at the beginning of realizing the cost benefits of cloud, but organizations moving to the cloud are also attracted by the new capabilities they do not get today. It is bringing new approaches to designing applications, specifically for the cloud, and providing more resilience by architecturing failure as a design concept. Cloud also teaches us about services and service levels, and the contrast between what the business wants for outcomes versus IT's old methods of getting there."

Mobile

In 2016, more than 1.6 billion smart mobile devices will be purchased globally. Two-thirds of the mobile workforce will own a smartphone, and 40 percent of the workforce will be mobile. The challenge for IT leaders is determining what to do with this new channel to their customers and employees.

"Mobile is about computing at the right time, in the moment. It is the point of entry for all applications, delivering personalized, contextual experiences," Mr. Sondergaard said. "It means: marketing gets more time with the customer; employees become more productive; and process flows get dramatically cut."

In less than two years, iPads will be more common in business than Blackberries. Mr. Sondergaard said some CIOs are now placing orders for tens of thousands of iPads at a time. Productivity is the driver. Two years from now, 20 percent of sales organizations will use tablets as the primary mobile platform for their field sales force. As a result, by 2018, 70 percent of mobile workers will use a tablet or a hybrid device that has tablet-like characteristics. (see Mr. Sondergaard's comments regarding mobile on the Gartner YouTube channel).

Gartner forecasts that in 2016, half of all non-PC devices will be purchased by employees. By the end of the decade, half of all devices in business will be purchased by employees.

Social Computing

In the next three years, the dominant consumer social networks will the limits of their growth. However, social computing will become even more important. Companies are establishing social media as a discipline. Gartner predicts that in three years, 10 organizations will each spend more than $1 billion on social media.

"Social computing is moving from being just on the outside of the organization to being at the core of business operations," Mr. Sondergaard said. "It is changing the fundamentals of management: how you establish a sense of purpose and motivate people to act. Social computing will move organizations from hierarchical structures and defined teams to communities that can cross any organizational boundary."

Big Data

By tapping a continual stream of information from internal and external sources, businesses today have an endless array of new opportunities for: transforming decision-making; discovering new insights; optimizing the business; and innovating their industries.

Big data creates a new layer in the economy which is all about information, turning information, or data, into revenue. This will accelerate growth in the global economy and create jobs.

"Big data is about looking ahead, beyond what everybody else sees," Mr. Sondergaard said. "You need to understand how to deal with hybrid data, meaning the combination of structured and unstructured data, and how you shine a light on 'dark data.' Dark data is the data being collected, but going unused despite its value. Leading organizations of the future will be distinguished by the quality of their predictive algorithms. This is the CIO challenge, and opportunity."

About Gartner Business Intelligence & Analytics Summit

The **Gartner Business Intelligence & Analytics Summit 2013**, March 18 – 20, in Grapevine, TX, provides independent advice, thought leadership and insight you need to fast-forward with BI and analytics, discover how to build a strategy, architecture and team that delivers value now and in the future. Delegates will benefit from a three day learning experience where you can learn directly from Gartner analysts and expert BI practitioners, hear the latest research findings, take-away actionable insights and also share knowledge and experience with your peers. Hot topics include: Predictive analytics, mobile BI, big data analytics, logical data warehousing, and visual data discovery.

Additional information from the event will be shared at www.gartner.com/us/bi and on Twitter at http://twitter.com/Gartner_inc using #GartnerBI.

B.13. Gartner: Fewer Than 30 Percent of BI initiatives Will Align Analytic Metrics With Enterprise Business Drivers by 2014

Gartner Says Fewer Than 30 Percent of Business Intelligence initiatives Will Align Analytic Metrics Completely With Enterprise Business Drivers by 2014

Analysts Explore the Future of Business Intelligence and Analytics at Gartner Business Intelligence Summit 2012, April 2-4, in Los Angeles

By 2014, fewer than 30 percent of business intelligence (BI) initiatives will align analytics completely with enterprise business drivers, despite alignment being the foremost BI challenge, according to Gartner, Inc. Cloud offerings will account for just 3 percent of BI revenue by 2013, despite every major BI platform vendor presenting one. In addition, Gartner analysts said that by 2013, BI initiatives will be based on an organizational model that strikes a balance between centralized and decentralized delivery.

“The immediate future of the BI landscape is one of a disconnect between marketing hype about pressing challenges on the one hand and reality on the other,” said Andreas Bitterer, research vice president at Gartner. “The need for analytics does not match most organizations’’ skill requirements; vendor hype for cloud-based BI is not reflected in revenue and customer adoption, and there is a struggle between centralized and decentralized organizational models of BI delivery.”

Gartner’s three central predictions for the BI market are:

By 2013, every major BI platform vendor will present a cloud offering, but these will account for just 3 per cent of total BI revenue.

The BI market is not exempt from cloud-related hype. Current adoption of "cloud BI" by user organizations lags far behind the expectations of vendors, which are busy creating and marketing new off-premises solutions. Organizations that have already invested in on-premises BI infrastructure are hesitating to identify a segment of their BI initiative for which data can be moved into the cloud and reports and dashboards received from a cloud provider. However, companies that have subscribed to a specific cloud application, such as customer relationship management, payroll or help desk service, are more inclined to use BI functionality delivered by their cloud provider, as they see it essentially as an extension of the cloud application.

By 2013, BI initiatives will be based on an organizational model that strikes a balance between centralized and decentralized delivery.

Many BI programs have departmental roots with analytical resources embedded in the business. This model has worked well in serving departmental needs, but it lacks consistency in terms of data definitions and measures across an entire organization. Often, the IT organization has solved this inconsistency problem by establishing a central team to deliver BI. However, such an overly centralized model lacks the agility and familiarity of the decentralized model. A hybrid delivery model enables greater consistency and economies of scale, more autonomy and faster turnaround times.

By 2014, fewer than 30 percent of BI initiatives will align analytic metrics completely with enterprise business drivers.

The foremost BI challenge is to align initiatives with corporate strategy and objectives, but fewer than one-third of organizations have a documented analytics, BI or performance management strategy. Organizations often develop and deploy hindsight-oriented reports and/or query applications focusing on metrics that users may find interesting, but they don't represent the operational or strategic controls used to facilitate business performance.

With the increasing consumerisation of BI (for example, mobile BI), the growing volume and variety of available data, and the soaring speed of business, it can be challenging to establish appropriate "guard rails" for analytic implementations to ensure that the right data is presented to the right people and processes at the right time. These user/data growth factors also challenge the cohesion of metrics frameworks among lines of business, resulting in business functions that operate in conflict with one another; for example, one group may focus on profitability, while another concentrates on market share.

“Throughout 2012 and beyond, BI will remain subject to nontechnical challenges,” said Mr. Bitterer. “IT leaders should concentrate not only on the technological aspects of BI, but also on the severe lack of analytical skills. Second, they should use a ‘think global, act local’ approach in their BI programs to provide the right level of autonomy and agility to avoid the bottlenecks that overly centralized BI teams create, while simultaneously establishing enough consistency and standards for enterprise wide BI adoption.”

More information is available in the report "Predicts 2012: Business Intelligence Still Subject to Nontechnical Challenges,” available on Gartner's website at www.gartner.com/resId=1873915. This document is part of Gartner’s overall 2012 Predicts coverage, which is available at www.gartner.com/predicts. The Gartner Predicts Special Report overview includes links to more than 70 Predicts reports, categorized by topic, industry and market.

Mr. Bitterer will speak on BI market trends at the Gartner Business Intelligence Summit 2012.

About Gartner Business Intelligence Summit 2012
The Gartner Business Intelligence Summit is the premier business analytics event that provides the must-have insights, frameworks and best practices for maximizing the business impact of information management and business analytics initiatives. This year’s summit will help organizations transform their decision-making by examining new developments in BI, how analytics and BI relate, improvements in data quality, analytics in the cloud, and the linking of BI to master data management. Additional information from the event will be shared on Twitter at http://twitter.com/Gartner_inc using #GartnerBI.

The Gartner Business Intelligence Summit in Los Angeles is being held on 2-4 April at the JW Marriott hotel at L.A. Live. Additional information is available at www.gartner.com/us/bi.

B.14. Twenty Questions about Big Data and Data Sciences

1. How and when did you become interested in analytics?
2. Do companies treat data and data science differently in Europe, America and Asia?
3. What are your predictions for the next 5 years, regarding the evolution of data science?
4. Is there still interest in small data, classical statistical models, simulation and sampling?
5. Are poor models on comprehensive data better than great models on silos?
6. How to get data silos, internal and external data sources, to blend together?
7. What skills should data scientist acquire?
8. What should colleges teach?
9. I believe great data scientists are also good management consultants. Do you agree?
10. Which areas are going to benefit most from cloud technology?
11. What is the difference between computational statistics and data mining?
12. With the advent of huge data, what is the future for QA, fuzzy merging, data compression, sampling, interactive dashboards and smart visualization?
13. Is there a lot of hype surrounding real time analytics?
14. Do you think in-memory analytics will become more widespread? What would make in-memory analytics and in-database data mining more attractive?
15. Will the data become more or less structured?
16. What do you think of companies heavily relying on social media for data intelligence? Aren't social users different, possibly more liberal, than others?
17. What about security, regulation and privacy issues?
18. What about automating the process of analyzing big data?
19. Is the return on big data bigger than on small data, once you factor in infrastructure, learning curve and human resources? How to improve return on investment?
20. What are the competitive advantages offered by your company?

Read and contribute to discussion, at:

http://www.analyticbridge.com/group/data-science-q-a

B.15. Interview with Drew Rockwell, CEO of Lavastorm

1. Short Bio

I started my career in the communications industry, where I spent 20 years with a Tier 1 carrier in probably 15 different jobs across the entire organization: Marketing, Advertising, Product Management, Operations, Sales, General Management, Strategy and Business Development. I basically experienced a multi-billion business from many different functional areas, at increasingly responsible management levels. When I was in my early 40s, I decided to embark on a “second” career, to take what I had learned and to try to animate and build companies, which ultimately led me to MDS Lavastorm Analytics, where I am CEO today.

2. How and when did you become interested in analytics?

I think in all the various jobs I had, I was left a little cold by “reports”, which later in my career became visually more appealing dashboards, but in the end seemed more as ways to describe a certain situation, or a function, or a customer segment, etc. They were and are necessary, and sometimes they were cool to look at, but for me not sufficient.

Analytics to me are less focused at describing a situation and more focused on understanding “why” something is happening, so that I could do something about it, or simulating or predicting what might happen, so I could plan for it.

I was always interested in connecting things, in understanding relationships between things, and I think that was what made me gravitate to analytics as a career.

3. Do you have any predictions for the coming year or two in the field of analytics?

The field of analytics is so dynamic with technology changes increased investment. There is a great deal of change going on and a great deal of opportunity in front of us. We posed that exact question to the Lavastorm Analytics LinkedIn community, an online community we manage and we got a tremendous list of predictions for the field. Personally, I see a few themes gaining more traction in the coming 24 months:

- Analytic power will continue to become much more decentralized, moving from IT organizations to business users, moving from the exclusive domain of highly technical people to less technical users, moving from dependency on large data warehouses to a variety of data sources, tools, and methodologies to get to insight and action quickly. One data point: 40% of analytics budget spend will move to business departments in the next 3 years (Gartner)
- Analytic methodologies are becoming more discovery-driven and less dependent on the crafting of a question or a query. With the proliferation of “big data” there will be a need for more agile ways to test hypotheses, to join disparate data, both structured and unstructured together, and to more easily construct analytics. We have made huge strides in optimizing the processing of data, but we will see in the next few years huge strides in optimizing the analytic process itself, which I think will create a new wave of insight and action. The key here is to be able to gain analytic insight from within a business process itself to add context to the data you are analysing.
- At the same time as there is growth in the profession of analytics, and the continued emergence of data “scientists,” this specialized knowledge will create more powerful software assets to extend that knowledge to a much broader group of analytic “consumers” who will be focused on capturing value, on the answers not the methodology.

4. How do you see analytic models in the era of big data?
It seems to me there is a need for analytic models to become less and less “rigid” and more and more “adaptive”. For example, as you inspect data at a detailed level and wonder about new questions that you want to understand, the “cost” in terms of time necessary to pursue those new questions or models should be virtually free. This is something I think Lavastorm does very well. In addition, the nature of ‘audit analytics’ is changing and moving much closer from single source data requirements to multiple source and also with a much greater degree of focus on auditing of the business process itself. This will help finance departments turn audit from a cost center to a money making function.

5. How do we get data silos, internal and external sources, to blend together?
I view this as one of the key enablers of true analytics. In general, BI technologies have failed to make the bringing together of disparate data easy enough and they haven’t been able to create an analytic connective layer without having to put everything in a data warehouse. At Lavastorm, we have focused a lot of engineering talent on simplifying the joining of disparate data while maintaining the traceability of any data used in the analysis back to its original sources. This “traceability” builds confidence in the results and can be applied to a new generation of audit analytics.

6. What do you think of real time analytics?
To my thinking, true insight that yields action trumps all. But I do think we will see more timely analytics, whether it is real time or near-real time will depend on the analytics and the value of timely action. The important point is that the analytics are reflecting current conditions. For example, the Lavastorm Analytics Platform gives organizations the ability to push the analytic closer to the source of data creation because the data doesn’t have to go through a data warehouse and, therefore, the data is closer to the business process itself. That allows for faster detection, faster reaction, and greater control.

I have been intrigued for years by using analytics to correct mistakes as they are happening. I think there are some interesting examples of this, but I expect there is much more to come.

7. MDS Lavastorm talks about “controls” in the context of analytics? Can you shed some light on that?
Yes, from our work in Fraud Analytics and Revenue Assurance, we have come to believe that there is value in running persistent analytics over business processes, to continuously identify data that do not conform to rules.

A simple example of this is order accuracy – we run analytics for companies that inspect an order against a number of highly conditional business rules or logic, checking to be sure that things like promotional codes are correct, discounts are correct, addresses match, etc. We basically call out errors very soon after they happen, allow the business to fix them before they cause downstream issues, and to understand the root cause of the error so that the business process can be fixed quickly. This is a good example of a control – once you have captured the correct business rules (what is supposed to happen) there is enormous value in continuously monitoring a process to be sure the rules are followed. Finding the percent that is wrong, correcting it, understanding why, and correcting that, has enormous value. A key principle that we built into the Lavastorm Analytics Platform is the ability to easily create business controls, store them in a library and reuse them.

8. How has big data changed the way we use analytics?

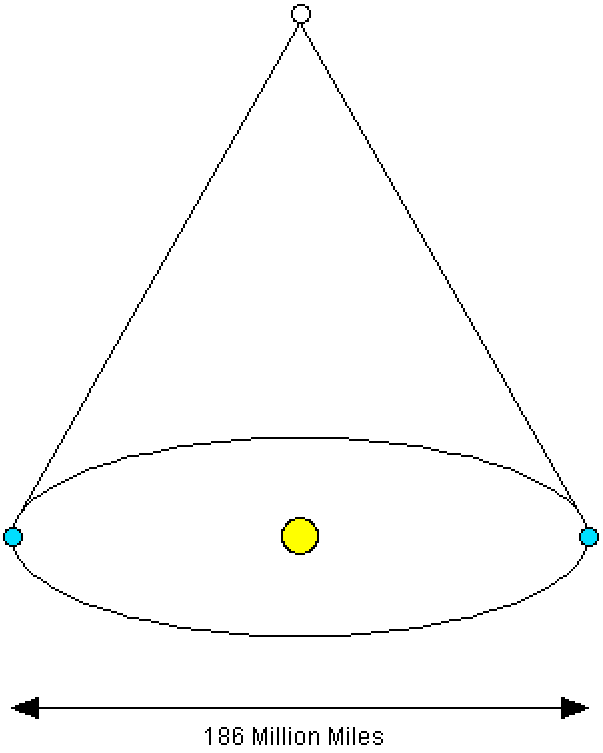
Well it has obviously created the need for more and more powerful appliances and techniques for processing huge volumes of data, as well as dealing with the complexity that this brings. It has also created the need to cost effectively and quickly join together multiple sources of data and data types. It is creating a greater need to do "discovery" based analytics rather than pure query or model-based analytics.

Read and contribute to discussion, at:

http://www.analyticbridge.com/profiles/blogs/interview-with-drew-rockwell-ceo-of-lavastorm

B.16. Can we use data science to measure distances to stars?

This is a very challenging problem: even as of today, is still very challenging. For the closest few hundred stars (and the focus on the discussion is on those close stars exclusively), the methodology to measure the distance is based on parallax measurements. You measure the angle to the star using 2 other reference points (the sun being one of them), and six months later, when you have traveled 300 million kilometers in the sky and are just opposite to the sun, you measure the angle again. Based on this triangulation, you can then estimate the distance to the star in question. See illustration below, and you will immediately understand the mechanism.



The big issue is that the 300,000,000 kilometers that separate the two observation points (blue circles 6 months apart) is at least 600,000 times smaller than the distance to the (closest) start. In other words, the angles in question are extremely acute, less than 1/4,000 of a degree. You first need very accurate instruments to measure the parallax, perfectly calibrated, unbiased, and be located in a very stable location to eliminate tiny side effects that will impact the measurement.

The first astronomer to make this measurement (Bessel) repeated his calculations 16 times, and averaged the 16 measurements. This significantly reduces measurement errors. But what if the resolution of your device is not granular enough (for a distant star)? For instance you can measure a deviation of 1/1,000 of a degree, but not one of 1/4,000 of a degree.

That's where the data science magic kicks in:
Instead of making 16 repetitions of a single measurement, do 10 measurements each day at exactly the same time, over a period of two years. For each measurement, repeat the process 16 times just like Bessel did. You now have 58,400 pairs of measurements (6-month apart) that you can analyse to identify patterns, trend and compute a much more accurate parallax (with small enough confidence interval) and thus a solution to the problem, by averaging grouping / measurements after removing outliers. The real magic in

this is that you can in fact, thanks to good *design of experiment* and *statistical inference*, measure a distance that your instrument cannot technically do due to its too low resolution. Isn't this amazing?
Note: Variations in these 58,400 measurements should be white noise: if you see patterns, something is wrong with the experiment or the device, and need to be fixed.

Featured Comments:

[Vincent] Could a similar methodology be used to detect very rare occurrences of fraud (occurring say one in 100,000 transactions), by magnifying the imperceptible signal, just as used in the distance-to-star problem?

[Jean-Paul] Looks like you use a software rather than an hardware solution to fix a problem with an instrument (to boost resolution). Interesting and new approach, much less costly than an hardware fix

---

Read and contribute to discussion, at:

http://www.analyticbridge.com/forum/topics/can-we-use-data-science-to-measure-distances-to-stars

B.17. Eighteen questions about real time analytics

1. Besides transaction scoring (credit card transactions, online ad delivery systems), what other applications need true real time?
2. What types of applications work better with near real time, rather than true real time?
3. How do you boost performance of true real time scoring? e.g. by having pre-loaded, permanent "in-memory" small look up tables (updated hourly) or other mechanisms? Please explain.
4. How do you handle server load at peak times, as it can be 10x higher than (say) at 2am? And at 2am, do you use idle servers to run hourly or daily algorithms to refine / adjust real time scores?
5. Are real time algorithms selected and integrated into production by data scientists, mostly based on their ability to be easily deployed in a distributed environment?
6. Examples of hard-to-solve problems, how is it done? Example: 3-D streaming video processing in real time from a moving observation point (to automatically fly a large plane at low elevations in crowded skies)
7. Do you think end users (e.g. decision makers) should have access to dashboards updated in true real-time, or is it better to offer 5-minute delayed statistics to end users? In which application real time is better, for end users?
8. Is real time limited only to machine generated data?
9. What is machine generated data? What about a real-time trading system that is based on recent or even extremely recent tweets or Facebook posts? Do you call this real time, big data, machine data, machine talking to machines, etc.?
10. What is the benefit of "true real time" over (say) "5-minute delayed" signals in question 9)? Does the benefit (increased accuracy) outweigh the extra costs? (On Wall Street, usually the answer is yes. But what about keyword bidding algorithms? - delayed reaction is OK?)
11. Any rules of a thumb regarding optimum latency (when not implementing true real time) depending on the type of application? For instance, for Internet traffic monitoring, 15 minutes is good because it covers most user sessions.
12. What kind of programming environments are well suited for big data in real-time (SQL is not, C++ is better, what about Hadoop? What technology do you use?)
13. What kind of applications are well suited for big data in real-time?
14. Example of metrics heavily / easily used in real time (e.g. time to last transaction)?
15. To deliver clicks evenly and catch fraud right away, do you thing that the only solution for Google is to monitor ad delivery at the advertiser level, in true real time?
16. Do you think Facebook use true real time for ad targeting? How could they improve their very low impression-to-click ratio? (10x below Google, I think) Why is this ratio so low despite the fact that they know so many things about their users? Could technology help?
17. Future of real time over the next 10 years? What will become real time? What will stay hourly of end-of-day systems?
18. Are all real-time systems actually hybrid, relying also on hourly and daily or even yearly (with seasonality) components to boost accuracy? How are real-time predictions performed for very sparse highly granular data, such as predicting the yield of any advertising keyword in real time for any advertiser? [answer: group sparse data into bigger buckets, make forecasts for the entire bucket]

Read related article: Seven questions about real time analytics.

Read and contribute to discussion, at:

http://www.analyticbridge.com/forum/topics/18-questions-about-real-time-analytics

B.18. Can any data structure be represented by one-dimensional arrays?

I believe so. For example,

- Each node of a binary tree would use 4 array cells: one for a pointer to the father node, two pointers to the two sons, and one for the value.
- Each element of a hash table would use 4 or 5 array cells: the index, possibly a random key that efficiently encode the index, the value, a pointer to the previous index, and a pointer to the next index. This would make finding, updating, deleting or inserting an element a bit slow, but performance can be boosted by storing, at the beginning of the array, a list of pointers to 1,000 indices evenly spaced in the index universe.

A similar arguments can be used for graphs (as in graph theory), non-binary trees, heaps, stacks, linked lists etc.

Indeed, before programming languages offered advanced data structures, sophisticated objects and types, recursion (and much more) -- all the data -- had to be stored in (organized) arrays. Trees, hash tables etc. were simulated by using arrays and pointers. Even recursion.

Are there exceptions? In some ways, we are getting back to the old times, with unstructured data, such as member postings on social networks. Although structuring unstructured data (by putting it into clusters and taxonomies) allow it to be manipulated much more easily.

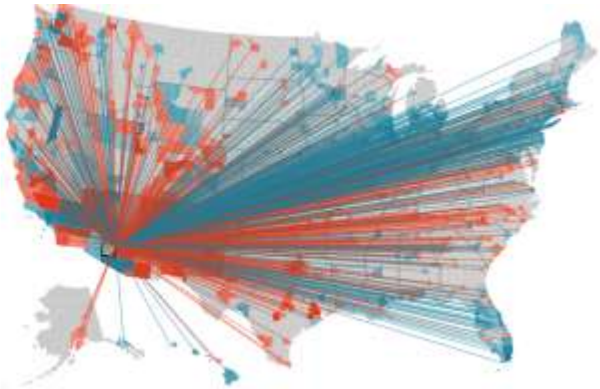
Read and contribute to discussion, at:

http://www.datasciencecentral.com/forum/topics/can-any-data-structured-be-efficiently-represented-by-one-dimensi

B.19. Data visualization: example of a great, interactive chart

This chart was published in the Forbes magazine by Jon Bruner, and the data comes from the IRS tax stats tables.

Chart:



American Migration [Interactive Map] Close to 40 million Americans move from one home to another every year. Click anywhere on the map below: blue counties send more migrants to the selected county than they take; red counties take more than they send.

View interactive chart on Forbes website

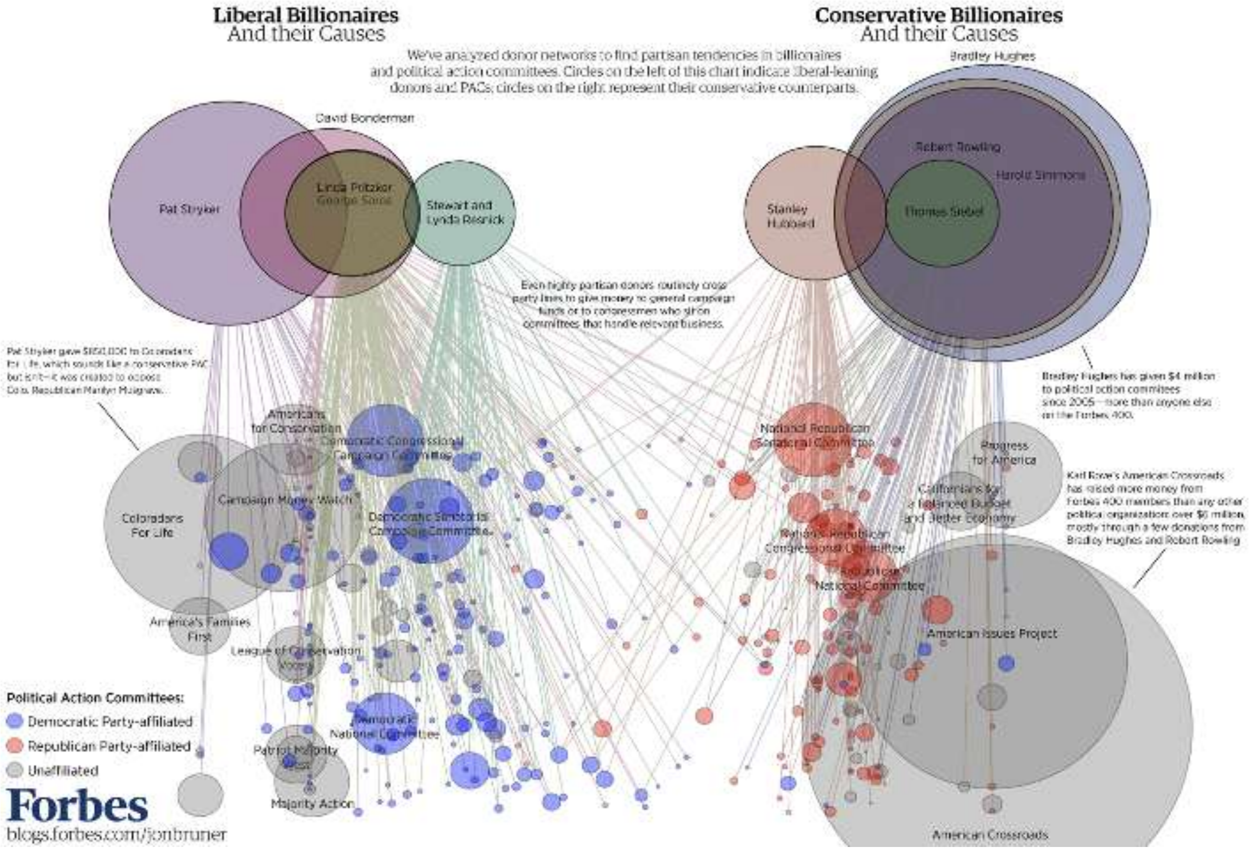
Other articles by Jon Bruner:

- Will Data Monopolies Paralyze the Internet? - http://www.forbes.com/sites/jonbruner/2012/04/12/will-data-monopoli...
- Five Steps For Making Data-Driven Decisions - http://www.forbes.com/sites/jonbruner/2012/04/20/five-steps-for-mak...
- Tim O'Reilly on the Future of Location: "The Guy with the Most Data Wins" - http://www.forbes.com/sites/jonbruner/2012/04/04/tim-oreilly-on-the...

Featured Comments:

[Vincent] I don't agree with "the guy with the most data wins". Comprehensive data with poor predictive modeling bring less value than correctly sample data combined with good statistical models. The latter is indeed the future of big data.

[Vincent] By the way, here's another great chart from the same guy:



Liberal Billionaires
And their Causes

Conservative Billionaires
And their Causes

We've analyzed donor networks to find partisan tendencies in billionaires and political action committees. Circles on the left of this chart indicate liberal-leaning donors and PACs; circles on the right represent their conservative counterparts.

Even highly partisan donors routinely cross party lines to give money to general campaign funds or to congressmen who sit on committees that handle relevant business.

Pat Stryker gave $850,000 to Coloradans for Life, which sounds like a conservative PAC but isn't—it was created to oppose Colo. Republican Marilyn Musgrave.

Bradley Hughes has given $4 million to political action commitees since 2005—more than anyone else on the Forbes 400.

Karl Rove's American Crossroads has raised more money from Forbes 400 members than any other political organization over $6 million, mostly through a few donations from Bradley Hughes and Robert Rowling.

David Bonderman
Pat Stryker
Linda Pritzker
George Soros
Stewart and Lynda Resnick
Stanley Hubbard
Bradley Hughes
Robert Rowling
Harold Simmons
Thomas Siebel
Americans for Conservation
Democratic Congressional Campaign Committee
Campaign Money Watch
Coloradans For Life
Democratic Senatorial Campaign Committee
America's Families First
League of Conservation Voters
Democratic National Committee
Patriot Majority West
Majority Action
National Republican Senatorial Committee
Progress for America
Californians for a Balanced Budget and Better Economy
National Republican Congressional Committee
Republican National Committee
American Issues Project
American Crossroads
Political Action Committees:
● Democratic Party-affiliated
● Republican Party-affiliated
● Unaffiliated
Forbes
blogs.forbes.com/jonbruner

Read and contribute to discussion, at:

http://www.analyticbridge.com/profiles/blogs/data-visualization-example-of-a-great-interactive-chart

B.20. Data science jobs not requiring human interactions

For the very smart data science geek who wants to avoid all kinds of social interactions (that is: having no boss, no employee, no colleague, no client, no customer, no contractors, no interaction with vendors etc.), there are a few options. All of them allows you to work from home. Some (when automated) allow you to not work at all. They can generate good complimentary income or (if scalable) great income.
In practice, these opportunities are not real jobs but rather entrepreneurial initiatives that leverage the most sohisticated data science techniques, and they bring a complement of revenue rather than a full income. Also, these "occupations" are usually not fully automated (when they are - and sometimes they indeed are, you are literally making money when sleeping). Only top talent with strong business acumen is able to succeed due to massive competition.

Here are a few of these jobs or activities:

1. Pay-per-click arbitrage: buy and sell clicks on ad networks (can be fully automated thanks to automated bidding and automated click fraud detection relying on carefully crafted and continuously tested algorithms)
2. Stock trading strategies: competition is so stiff that there are only two ways to succeed: (1) insider trading, e.g. you try to obtain job interviews with small publicly traded companies, then based on information gleaned during the interview, perform trades and (2) use trading strategies that professional traders will never use, e.g. stay "all cash" for several years on your trading account, and when the right event occurs, massively trade major indexes for a couple of days, then go dormant for another few years. You need sophisticated statistical models to succeed in this, with good back testing, walk-forward and robustness based on state-of-the-art cross-validation.
3. Sport bets and gaming. Requires very good statistical models (include fraud patterns in your model) and deep domain expertise, and to carefully select which brokers you are going to work with. Horse racing (Australia) is a good choice, depending on the broker.
4. Become a digital publisher of data science (or any) content, and use Google Adsense to monetize your websites. Requires great SEO / SEM skills. This can be fully automated thanks to content syndication. Not for the amateur if you want to make a living out of it without working at all - it is not easy to automatically build real, targeted and growing traffic in a purely automated way, but it is feasible. Revenue could also come from your e-newsletter, thanks to Google or other ads.
5. Write data science e-books or reports, and sell them on Kindle (Amazon).
6. Launch a job board where recruiters automatically purchase via credit card and post job ads. Once again, for full automation, the challenge is to automatically grow targeted traffic. Difficult, but not impossible thanks to automated promotion in various social networks, and via Google Adwords and careful CPC pricing strategies.

Question: Are men more likely than women to be interested in these "no human interaction" types of activities?

Featured Comments:

[Titus] I like these ideas. It is in many ways more challenging than working a regular job, and by the same token, more exciting. It's like climbing the Everest in winter, solo, with no oxygen (a few did it).

[Vincent] Interesting: the concept of extracting competitive intelligence / insider information, via bogus job interviews. Anyway, from my past experience, companies that invited me for a job interview usually had a boost in stock price in the next 30 days. So, just the fact that you are invited for a job interview is a buy signal, for the stock in question. You might also be able to create a start-up whose sole purpose is to

1. have bogus candidates applying for various positions from a list of target companies,
2. securing job interviews,
3. then producing reports based on these interviews
4. and finally sell these reports.

Interestingly, corporations sometimes do the reverse: they invite you for a job interview just to get free advice from an expert (you), or worse to try to learn about your IP and steal it.

Read and contribute to discussion, at:

http://www.analyticbridge.com/profiles/blogs/data-science-jobs-not-requiring-human-interactions

B.21. Featured Data Scientist: Vincent Granville, Analytic Entrepreneur

This Analyticbridge interview is part of a blog post series about data scientists , what they do in their daily life, how they found their job, what techniques they use, corporate culture and benefits in their work place, etc. The previous posting was featuring people working for Deloitte Analytics, click here for details.

Can you summarize your career path in a few sentences?
PhD in computational statistics / image analysis, 1993. Post doctorate (Cambridge UK and NISS, North Carolina), then 15 years of corporate experience ranging from statistician, market research, fraud detection expert, business intelligence, analytics consulting, to Chief Scientist roles. Industries: Internet, online advertising, finance.

Founded and self-funded Analyticbridge, a leading community for analytic professionals, in 2008. Raised 6MM in VC funding in 2006, worked with Wells Fargo, InfoSpace, eBay, Microsoft and a few startups. Own patents on Internet traffic quality scoring (2008-2010), author of the first ebook on data science (2012), listed as top 20 most influential big data experts by Forbes (2012).

How did you find your current job? What skills does it require?
Working exclusively for one employer is dangerous from a risk management point-of-view: it's like having a stock portfolio with just one stock. Being independent is much more fun, more financially rewarding, and much less risky if done well. Most importantly, I wanted to help people connect and leverage skills by bridging together several communities that are heavy analytic users: statistics, operations research, BI, quant, data mining, six sigma,

econometrics etc. The opportunity appeared in 2008, with the emergence and growth of social networks.

My job is similar to being COO / CEO, but I still perform statistical analyses. It requires good judgment, acting quickly and using the right business metrics (to assess long term potential, and success), change management, forecasting, management consulting type of skills, opportunity detection, being aware of new technologies and competitors, taking calculated risks, assessing vendors and partnerships, sales and communications with clients / vendors / users, accounting, a bit of law / finance, and computational marketing (we have a secret recipe for growth). These skills and craftsmanship's are more important than technical skills learned at school.

What do you enjoy most and least in your current occupation?
I am happy when great partnerships are created thanks to my company, when applicants find jobs thanks to us, when we make clients happy or valuable members successful (some got PhD scholarships in US thanks to us). I don't like mundane activities (about 40% of my workload right now), but we are recently doing a better job at outsourcing most of them (accounting, etc.)

What kind of models and data do you use?
I use data gathered on the Internet with web crawlers, or internal unstructured data combined with text mining to produce reports about the analytic community, or to generate stock price forecasts (buy and sell signals for major indices) based on internal job ads (sales) data and other metrics, or testing new patentable algorithms that we describe in our open-source, free data science eBook sponsored by advertisers.

How can someone become a data scientist?
We start to see a few interesting curricula: Northwestern University, Berkeley, Stanford -- check our course section on Analyticbridge. Attending these online classes is a good start, as well as being intern for an analytic company (including ourselves). A lot of the technical stuff and resources can be found and learned online or in in our e-book, for free: open-source languages (Python, Hadoop environment, SQL, Java, C++), cross-validation, analytics as a service, model fitting, machine learning, clustering techniques, hidden decision trees, lift metrics, design of experiments, Monte-Carlo simulation, non parametric confidence intervals, association rules and scoring technology etc. We plan to offer a course, and we already offer a free data science certification based on your bio (no exam required). Another idea is to play with Google keyword API's and start creating your search engine or your own taxonomies. Attend conferences (check out our conference section). Or download free trial copy of vendor software (Lavastorm, Data Miner, etc.) And buy books on Amazon: read Read our book and journal section on Analyticbridge for details.

What is the corporate culture in your company?
Optimizing all processes by outsourcing as much as possible to carefully selected vendors: mailing list and server management, advertising delivery, social network platform, credit card processing, etc. Making partners, clients and valuable members happy. Killing spam. Offering salaries or hourly rates above average thanks to smart business optimization and the leverage of a number of business "unfair advantages".

Read and contribute to discussion, at:

http://www.analyticbridge.com/group/faces-of-analytics-and-data-science/forum/topics/featured-data-scientist-vincent-granville-analytic-entrepreneur

B.22. Healthcare fraud detection still uses cave-man data mining techniques

The Washington Education Association (WEA, in Washington State) is partnering with Aon Hewitts (Illinois), a verification company, to eliminate a specific type of health insurance fraud: teachers reporting non-qualifying people as dependents, such as an unemployed friend with no health insurance. The fraud is used by "nice" people (teachers) to provide health insurance to people who would otherwise have none, by reporting them as spouse or kids.

Interestingly, I saw the letter sent to all WEA teachers. It requires you to fill lots of paperwork and provide multiple identity proofs (tax forms, birth certificates, marriage certificates etc.) similar to ID documents (I9 form) requested to be allowed to work for a company.

It is easy to cheat on the requested paper documentation that you have to mail to the verification company (e.g. by producing fake birth certificates or claiming you don't have one, etc.) In addition, asking people to fill so much paperwork is a waste of time and natural resources (trees used to produce paper), and results in lots of errors, privacy issues and ID theft risk, and costs lots of money to WEA.

So why don't they use modern methods to detect fraud: data mining techniques to detect suspicious SSN's, identifying SSN's reported as dependent by multiple households based on IRS tax data, SSN's not showing up in any tax forms submitted to the IRS, address mismatch detection, etc. (note that a 5-day old baby probably has no record in the IRS database, yet he is eligible as a dependent for tax or health insurance purposes).

Why not use data mining technology, instead of paper - with all the advantages that data mining offers over paper? What advantages does paper offer? I don't see any.

Featured Comments:

[Vincent] Here's how data collection and processing should have been performed:

- Ask Washington teachers to provide list of dependents, with relationship (spouse / kid), address, phone number(s), date of birth, maybe driving license with state, for each dependent - and nothing more, no paper.
- Teachers fill the form online (those without Internet access use school computers) and the online questionnaire is designed to minimize the risk of errors / typos. IP addresses are also tracked.
- The verification agency use a service such as Intelius to check whether the names (dependents) provided by a teacher live under the same roof, and that the age / date of birth matches the number reported by the respondent. The data mining algorithm will perform fuzzy matching. Also do dependents share the same last name? If not additional scrutiny required.
- Teachers providing more than 3 dependents are subject to increased scrutiny by data mining algorithms, and random manual verification (e.g. via phone calls).

This should eliminate most of the fraud, at a very low cost, and with very little burden on teachers.

Read and contribute to discussion, at:

http://www.analyticbridge.com/forum/topics/healthcare-fraud-detection-still-uses-cave-man-data-mining

B.23. Why are spam detection algorithms so terrible?

It looks like most of them still rely on Naive Bayes applied to individual keywords, to flag messages. They fail to catch 90% of the spam, yet have a terrible "false positive" rate - as high as 5%.

Are there any companies working on customized (e.g. per email account) solutions? Are there any spam detector that

- use Botnet lists of (blacklisted) IP addresses for filtering as well as white lists,
- use lists of scammy URLs (embedded in an email message) as well as white lists
- use metrics other than individual keywords or combination of two keywords (e.g. positive / negative keywords) for spam detection, such as return address different from sender address, or return address looks spammy
- use algorithms that are much more modern than Naive Bayes, such as hidden decision trees?

Featured Comments:

[Vincent] I think spam detection algorithms to block outgoing mail are especially poor. The ones to block incoming mail are better.

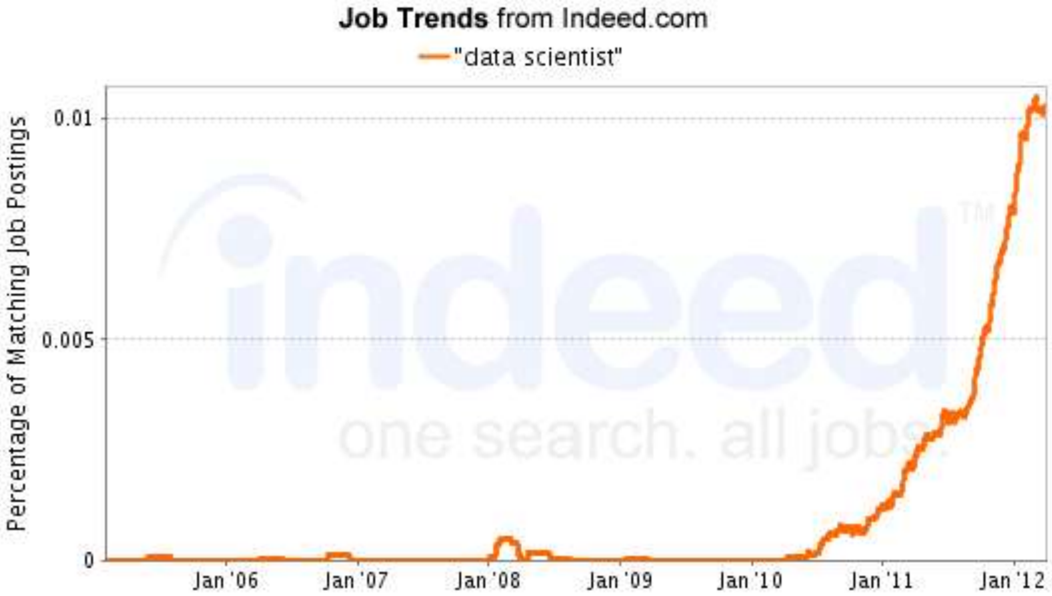
[Titus] Yahoo / Gmail / Hotmail certainly use spam detection for outgoing messages. Also, your mail client might be sending spam without your knowledge, if you've been infected by a Botnet. That's why blocking outgoing spam (which can be performed by your ISP) is as important as blocking incoming spam.

Read and contribute to discussion, at:

http://www.analyticbridge.com/forum/topics/why-are-spam-detection-algorithms-so-terrible

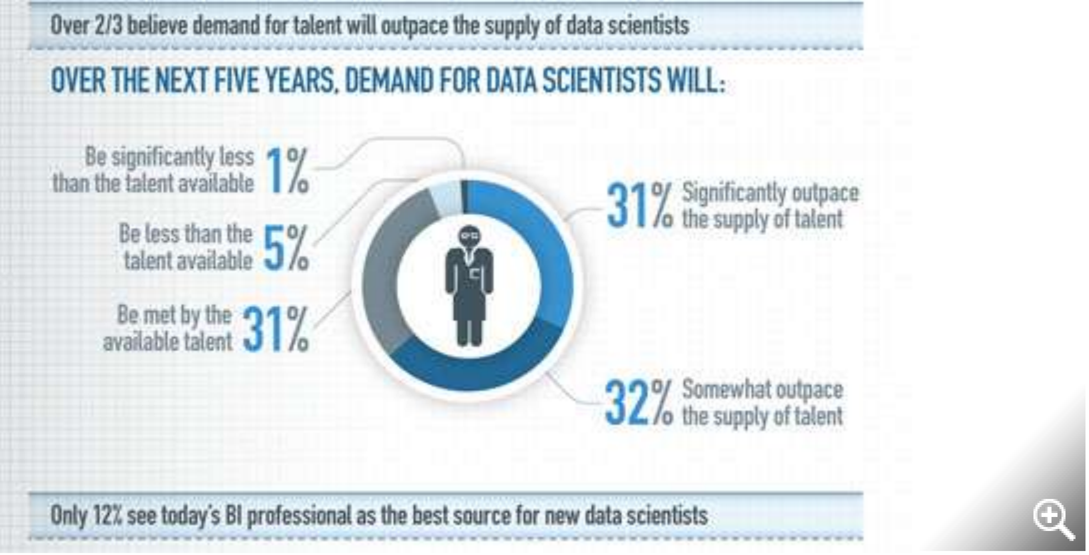
B.24. What is a Data Scientist?

Interesting article posted in *The Guardian* (see below). Here's my answer to this question:
A data scientist (I like the term data wizard better) is someone who can consistently derive money out of data, e.g. working as an employee, consultant or in an other capacity, by providing value to clients or extracting value for himself, out of data. Even a guy who design statistical models for sport bets, and use his strategies for himself alone, is a data scientist. But a journalist who writes an article about big data, is not a data scientist. In my opinion no diploma is required to succeed as a data scientist: you can learn the principles online, for free. Rather than knowlege, what makes a data scientist successful is craftsmanship, intuition and vision, to compete with peers who share the same knowledge but lack these other skills. If you work with people, in addition you need to have great communication skills and be able to guess what clients or your boss want.



What is a data scientist? (source: The Guardian)

It's the job of the moment. But what exactly is a data scientist?



What is a data scientist? EMC2 graphic representation of their survey. Click the image to see it
Everybody loves a data scientist: ever since Google's Hal Varian told the world that
the sexy job in the next ten years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s?

That, combined with the McKinsey report into big data last year is a powerful blend. The report reckoned the US alone would need 190,000 deep analytical 'data scientists' - and another 1.5m data-savvy managers to make the big decisions.
This week, in California, over 2,000 people grappled with this issue at the Strata data confe.... Data is big business, with companies like Google, Facebook and LinkedIn - and possibly every other corporate body you've heard of - creating huge profits out of the way they use data. This is the 'big data' everyone's talking about - the 2.5 quintillion bytes of information created every day from our internet searches, purchases, mobile phone calls and social networking. (If you're interested, a quintillion is 1,000 times a quadrillion, which is 1,000 times a trillion, which is 1,000 times a billion).
Read the full article at http://www.guardian.co.uk/news/datablog/2012/mar/02/data-scientist

Featured Comments:

[Vincent] See also Gartner article at http://blogs.gartner.com/doug-laney/defining-and-differentiating-th....
BI Analyst are poor statisticians. Statisticians are poor BI analysts. Data Scientists are both good BI analysts and good "big data" statisticians, and more: they are in between a traditional statistician and a computer scientist. Unlike statisticians or computer scientists, they are no geek. And most of their knowledge and art is not learned at school.

Read and contribute to discussion, at:

http://www.analyticbridge.com/profiles/blogs/what-is-a-data-scientist

B.25. Twenty seven types of data scientists: where do you fit?

Three metrics can be used to segment the population of data scientists. Each metric has three levels: high, medium, low. Hence the 27 (= 3 * 3 * 3) types of scientists.
Here are the metrics in question:

1. Soft skills: sales, business acumen, communications
2. Leardeship: vision, intuition, guessing/interpreting client needs, big picture - strategy oriented vs. tactical / detailed oriented
3. Knowledge: experience, craftsmanship, broad and deep knowledge (deep in some areas) vs. specialized only or absent.

Where do you fit? Employers tend to hire among a subset of those who rank high on all three metrics (if you are high on all three metrics but independent and making good money, you won't be hired, but you are not looking for a job anyway).
Our survey about "how hard it is to find a data scientist position" revealed that 70% of analytic people find it difficult to impossible, see http://www.analyticbridge.com/group/polls/forum/topics/poll-how-har... .
I'm interested in providing training to help candidates acquire the skills /craftsmanship required. More on this later.

Featured Comments:

[Gregory] I bet that not all 27 segments are equally populated, so there is a much smaller number of "clusters" of data scientists

[Vincent] Yes Gregory. Also this special segment made up of professionals that rank high on all three dimensions represents far less than 1/27 of the data mining community. Also, this segment is made up of people like you who:

1. have no interest in being hired and routinely turn down 98% of all employment opportunities (I can't speak for yourself, so feel free to reply if you disagree),
2. and who actually may not be a good fit in the corporate world,
3. and who make more money in their current situation than in any data scientist position in the corporate world or government.

I guess that's why employers complain that it is so hard to find talent. And why candidates complain it's hard to find a job.

Read and contribute to discussion, at:

http://www.analyticbridge.com/profiles/blogs/27-types-of-data-scientists

Part III: Data Science Resources

C.1. Vincent’s Lists

Our friend Sandro created a great list of data mining blogs. The full list contains several dozens of interesting data mining blogs: http://www.dataminingblog.com/list-of-blogs. He also maintains a list of analytic people. For a list of top 1,800 data science websites (it will be updated in the next few weeks), see http://www.analyticbridge.com/forum/topics/top-1800-analytic-websites.

We also maintain a list of resources (books, companies, training, webinars, vendors, etc.) at DataShaping.com, see http://www.datashaping.com/data_mining.shtml. One of the best, most comprehensive and up-to-date list is by Gregory at http://www.kdnuggets.com.

Vincent’s Favorite Books

- *Handbook of Natural Language Processing* by Nitin Indurkhya and Fred J. Damerau
- *Collective Intelligence* by Toby Segaran
- *Handbook of Fitting Statistical Distributions with R* by Zaven A. Karian and Edward J. Dudewicz
- *Statistics for Spatial Data* by Noel Cressie
- *Computer Science Handbook* by Allen B. Tucker
- *Data Mining and Knowledge Discovery Handbook* by Oded Maimon and Lior Rokach
- *Handbook of Computational Statistics* by James E. Gentle, Wolfgang Härdle, and Yuichi Mori
- *Handbook of Statistical Analysis and Data Mining Applications* by Robert Nisbet, John Elder, and Gary Miner
- *International Encyclopedia of Statistical Science* by Miodrag Lovric
- *The Princeton Companion to Mathematics* by Timothy Gowers
- *Encyclopedia of Machine Learning* by Claude Sammut and Geoffrey Webb
- *The Elements of Statistical Learning* by Trevor Hastie, Robert Tibshirani, and Jerome Friedman
- *Numerical Recipes: The Art of Scientific Computing* by William Press, Saul Teukolsky, William Vetterling, and Brian Flannery

Internal Analyticbridge Resources

- News | Polls | Jobs | Help
- Data Science e-Book
- Analytic Certifications
- Courses, Training
- Conferences, Webinars
- White Papers, Slides
- Text Mining Group
- Featured Profiles
- AnalyticBridge Newsletter
- Data Visualization
- Big Data
- New Books and Journals
- Competitions
- Think Tank
- Tutorials
- Salary surveys and trends

External Resources

Web Analytics

- Web Analytics Association
- Avinash's Blog

Statistics

- American Statistical Association
- Royal Statistical Society
- StatSoft Electronic Textbook
- Glossary
- Stat Links from UCLA
- ENBIS

Quant

- Portal for the Computational Finance
- Financial Intelligence Network
- Artificial Intelligence Stock Market Forum
- Wilmott
- Fisher's Comprehensive Financial Directory
- Trading Analytics, conference listing
- Futures Industry Association

Data Mining / Engineering

- KDNuggets.com
- Video Lectures, Interviews
- ACM Special Interest Group on Knowledge Discovery and Data Mining
- American Association for Artificial Intelligence
- The Advanced Computing Systems Association
- Association for Computational Linguistics
- IEEE International Conference on Data Mining
- Data Mining Digest (Blog)
- Data Mining Resources by Marcus Zillman
- Predictive Markup Modeling Language

Biostatistics

- Biospace
- MedStats Google Group
- Collection of Biostatistics Research Archive
- The International Biometric Society

BI, Market Research

- Data Warehousing and Business Intelligence Organization

Operations Research

- Six Sigma Directory
- Institute for Operations Research and Management Sciences

Text Mining

- National Center for Text Mining
- Text Mining Directory

Tutorials

- Online Statistics Handbook
- Wolfram Online Mathematical Encyclopedia
- Online statistics courses

SAS and Statistical Programming

- SASCommunity.org
- SAS-L Mailing List
- SAS Google Group

Miscellaneous

- Citation List

C.2. History of 24 analytic companies over the last 30 years

Here are Wikipedia links. Follow them to read about the history of these analytic companies and software over the last 30 years

1. Greenplum - http://en.wikipedia.org/wiki/Greenplum
2. Splunk - http://en.wikipedia.org/wiki/Splunk
3. Lavastorm - http://en.wikipedia.org/wiki/Lavastorm
4. Rapid Miner - http://en.wikipedia.org/wiki/RapidMiner
5. SAS - http://en.wikipedia.org/wiki/SAS_Institute
6. Datawatch - http://en.wikipedia.org/wiki/Monarch_(software)
7. S-Plus (Tibco) - http://en.wikipedia.org/wiki/S-PLUS
8. IMSL - http://en.wikipedia.org/wiki/IMSL_Numerical_Libraries
9. R - http://en.wikipedia.org/wiki/R_(programming_language)
10. SPSS - http://en.wikipedia.org/wiki/SPSS
11. Netteza - http://en.wikipedia.org/wiki/Netezza
12. Teradata - http://en.wikipedia.org/wiki/Teradata
13. Oracle - http://en.wikipedia.org/wiki/Oracle_Database
14. NoSQL - http://en.wikipedia.org/wiki/NoSQL
15. Map Reduce - http://en.wikipedia.org/wiki/MapReduce
16. Sybase - http://en.wikipedia.org/wiki/Sybase
17. FICO - http://en.wikipedia.org/wiki/FICO
18. PMML - http://en.wikipedia.org/wiki/Predictive_Model_Markup_Language
19. KNIME - http://en.wikipedia.org/wiki/KNIME
20. KXEN - http://en.wikipedia.org/wiki/KXEN_Inc.
21. Tableau - http://en.wikipedia.org/wiki/Tableau_Software
22. Excel - http://en.wikipedia.org/wiki/Microsoft_Excel
23. Statistica - http://en.wikipedia.org/wiki/STATISTICA
24. Mathematica - http://en.wikipedia.org/wiki/Mathematica
25. Informatica - http://en.wikipedia.org/wiki/Informatica
26. Matlab - http://en.wikipedia.org/wiki/Matlab

Featured Comments:

[Vincent] Need to add Pervasive, DataStax, Vertica, I'm sure I'm missing many others, e.g. Kaggle. Should I add Cognos, Microstrategy, SAP...

Read and contribute to discussion, at:

http://www.datasciencecentral.com/profiles/blogs/history-about-24-analytic-software-over-the-last-30-years

C.3. Fifteen great data science articles from influential news outlets

This is our third post in our series of "great articles". For each article: click on the link after the title to read the full story.

1. How Big Data Will Disrupt the $9 Billion Music Publishing Rights Business - http://blogs.wsj.com/cio/2012/05/15/how-big-data-will-disrupt-the-9...
Even successful songwriters experience their share of professional frustration, and so it was for Scott Schreer. In his view, understanding and monitoring the royalties for his compositions–which included the music for the “Have a Coke and a Smile” TV commercials, penned when he was 25–was an impossible challenge. Like most songwriters worldwide, he depended upon performing rights organizations such as ASCAP and BMI in the U.S. to negotiate and collect royalties for the performance of his work in TV productions and other public venues.

2. Data-Mining in Doctor's Office Helps Solve Medical Mysteries - http://www.businessweek.com/news/2012-05-15/data-mining-in-doctors-...
When hospitals turn to Microsoft Corp., it's no longer just for the latest office software. Some are asking the technology giant for help in diagnosing their patients.
In one instance, a hospital in Washington, D.C., asked Microsoft to examine its medical records to determine why certain patients were getting sick soon after being discharged. The company crunched the data from MedStar Washington Hospital Center and found something surprising: Patients who stayed in the same room had come down with the same infection.

3. Obama Administration Big Data Initiative Calls All Hands on Deck - http://data-informed.com/obama-administration-big-data-initiative-c...
The Obama Administration’s announcement that it is investing $200 million in a big data initiative reads like a checklist for leaders of advanced organizations who want to derive ever-more sophisticated insights from data sets growing in size and complexity by the hour.

- Advance the means for data scientists to manage, analyze, visualize and extract useful information from large and diverse data sets. Check.
- Create a cloud-based system that allows researchers to access a 200-terabyte data set. Check.
- Develop new scalable software tools for analyzing large volumes of structured and unstructured data in distributed data stores. Check.
- Create human-computer interaction tools to facilitate “rapidly customizable visual reasoning.” Check.
- Launch an online innovation marketplace for qualified bidders to contest for R&D “data to decisions” projects. Check.

A Lack of Big Data Investment
These and other projects unveiled March 29 are a response to a recent White House assessment that the federal government was underinvesting in information technologies that enable scientists, researchers and analysts to “move from data to knowledge to action,” said John P. Holdren, director of the White House Office of Science and Technology Policy.

4. Five Myths about Predictive Analytics - http://tdwi.org/articles/2012/05/01/5-predictive-analytics-myths.aspx

- Myth #1: You can't start until a data warehouse is in place
- Myth #2: Predictive analytics requires a Ph.D. or math degree
- Myth #3: There is a long time-to-value with predictive analytics
- Myth #4: Curiosity killed the cat
- Myth #5: The results are incomprehensible

5. Big data is worth nothing without big science - http://news.cnet.com/8301-1001_3-57434736-92/big-data-is-worth-noth...

As with gold or oil, data has no intrinsic value, writes Webtrends CEO Alex Yoder. Big science, which bridges the gap between knowledge and insight, is where the real value is.

6. Three Big "WHO"s to Master Big Data - http://futureofcio.blogspot.com/2012/05/three-big-whos-to-master-bi...

- Big Data Visionary
- Big Data Analytical/Technical Expert
- Big Data Solutionary

7. Imagining a Census Survey Without a Mandate - http://blogs.wsj.com/numbersguy/imagining-a-census-survey-without-a...

My print column explores a House bill that would make it voluntary for Americans who receive the Census Bureau's American Community Survey to respond to it, to protect the privacy of those who don't want to answer its dozens of questions.

Several statisticians and demographers caution that the bill would diminish the quality of ACS data by reducing response rates. "For a Congress that has shunned traditional earmarks, objective, high quality, and consistent data available for communities of all sizes are the only logical and fair way to allocate federal dollars to places and populations most in need of assistance," said Terri Ann Lowenthal, a consultant to the Census Project, a group of government agencies and advocacy groups that organized a letter signed by dozens of groups warning about the potential effects of the House bill.

8. GM Says No More Facebook Ads For Now; Sure Signs That Big Data Is Growing Up - http://www.adexchanger.com/ad-exchange-news/wednesday-05162012/

9. Fundraising Survey Results: More Organizations Considering Predictive Modeling in 2012- http://www.wealthengine.com/blog/2012/fundraising-mini-survey-more-...

10. Managed Care - Data mining technology to manage outpatient cost - http://fixushealth.com/managed care/managed-care-managed-care-data-...

11. Big Data Bubbles Up Trouble! - http://blog.pentaho.com/2012/05/16/big-data-bubbles-up-trouble/

12. Probabilistic Data Structures for Web Analytics and Data Mining- http://highlyscalable.wordpress.com/2012/05/01/probabilistic-struct...

13. Various Data Mining Techniques - http://www.europamines.net/134-various-data-mining-techniques.html

14. Art and Analysis – A Designer’s View of Data Visualization - http://blog.nielsen.com/nielsenwire/featured-insights/art-and-analy...

15. Bringing closer Data Visualization and Information design - http://www.indiegogo.com/auramavega

Read our previous "great articles from top news outlets" at:

1. http://www.analyticbridge.com/profiles/blogs/13-great-articles-abou...
2. http://www.analyticbridge.com/profiles/blogs/a-new-set-of-six-great...

Read and contribute to discussion, at:
http://www.datasciencecentral.com/profiles/blogs/15-great-data-science-articles-from-influential-news-outlets

C.4. List of publicly traded analytic companies

I've just started the list. I need your help to finish it. I want to include companies whose core business is analytics / data science / big data.

- FICO (Fico) - http://finance.yahoo.com/q?s=fico&ql=1
- SPLK (Splunk) - http://finance.yahoo.com/q?s=splk&ql=1
- DWCH (DataWatch) - http://finance.yahoo.com/q?s=dwch&ql=1
- TIBX (Tibco) - http://finance.yahoo.com/q?s=tibx&ql=1
- PVSW (Pervasive) - http://finance.yahoo.com/q?s=pvsw&ql=1

Other companies

- SAS, Asterdata, KXEN: private, not publicly traded
- Teradata, Oracle, Sybase, EMC/Greenplum: core business is databases, servers, cloud rather than analytics
- SPSS, Netezza: acquired by IBM
- Google, Facebook, Amazon, LinkedIn: business models deeply rooted into analytics, but not offering analytic solutions as core bussiness
- Deloitte, Market research companies offering consulting services
- IBM: analytics is not the core of their business

Featured Comments:

[Vincent] Informatica: stock symbol is INFA. See also this article: History about 24 analytic software over the last 30 years.

[John] Verisk Analytics might be one to include.

[Loretta] http://www.acxiom.com/ may be a firm to consider.

Read and contribute to discussion, at:

http://www.analyticbridge.com/profiles/blogs/list-of-publicly-traded-analytic-companies

C.5. Thirty unusual applications of data sciences, analytics and big data

Outside the traditional business analytics / BI fields:

1. Algorithms and Numerical Techniques
2. Application Design and Porting Techniques
3. Astronomy and Astrophysics
4. Audio, Image and Video Processing
5. Bioinformatics
6. Climate, Weather, Ocean Modeling
7. Cloud Computing
8. Cluster Management
9. Computational Photography
10. Computational Physics
11. Computational Structural Mechanics
12. Computational Fluid Dynamics
13. Computer Graphics
14. Computer Vision
15. Databases, Data Mining, Business Intelligence
16. Development Tools and Libraries
17. Digital Content Creation and Film
18. Electrical Design and Analysis
19. Energy Exploration
20. Finance
21. General Interest
22. GPU Accelerated Internet
23. Life Sciences
24. Machine Learning and AI
25. Machine Vision
26. Medical Imaging and Visualization
27. Mobile Applications and Interfaces
28. Molecular Dynamics
29. Neuroscience
30. Parallel Programming Languages and Compilers
31. Quantum Chemistry
32. Ray Tracing
33. Stereoscopic 3D
34. Supercomputing
35. Visualization

Learn more at http://www.analyticbridge.com/group/conferences/forum/topics/gpu-co...

C.6. 50 unusual ways analytics are used to make our lives better

Many more to come soon. Let's start with 10 for today.

1. Automated patient diagnostic and customized treatment. Instead of going to the doctor, you fill an online (decision tree type of) questionnaire. At the end of the online medical exam, customized drugs are prescribed, manufactured and delivered to you: for instance, male drug abusers receive a different mix than healthy females, for the same condition. In short, doctors and pharmacists are replaced by robots and algorithms.
2. Movie recommendations for family members. Allows parents to select age, gender, language, viewing history to optimize recommendations for individual family members.
3. Scoring technology. Computation of your FICO score, algorithms to determine the chance for an Internet click (in a rare bucket of clicks lacking historical data) to convert or for a patient to recover -- based on analytical scores with confidence intervals. Issues: score standardization, score blending and score consistency over time and across clients.
4. Detection of fake book reviews (Amazon) and fake restaurant reviews (Zagat).
5. Detection of plagiarism, spam, as well as people sharing paid accounts and pirated movies with friends and colleagues.
6. Automated sentencing for common crimes, using a crime score based on simple factors. This eliminates some lawyer and court costs.
7. Testing athletes for fraud: To boost performance, some athletes now keep blood samples from themselves in the freezer, and inject it back into their system when participating in a competition, to boost red cells count and performance while making detection impossible.
8. Detecting election fraud.
9. Semi-automated car driving. Software to guess user behavior, avoiding collisions by having automated pilot to bypass human driving as needed. And also providing directions when you are driving in a place where GPS does not work (due to inability to connect with navigation satellites, typically in remote areas).
10. When you obtain a mortgage, immediately sense if the calculated monthly payment makes sense. It happened to me: they mixed up 30 and 15 years (that was their explanation for the huge error in their favor): I believe it was fraud, and if you are not analytical enough to catch these "errors" right away, you will be financially abused. In this example, you need to be able to mentally compute a mortgage monthly payment given the length and interest rate. Similar arguments apply to all financial products that you purchase.

Read and contribute to discussion, at:

http://www.analyticbridge.com/profiles/blogs/50-unusual-ways-analytics-are-used-to-make-our-lives-better

C.7. Berkeley course on Data Science

Instructor is former Facebook data star. See http://datascienc.es.

Quora

- What is Data Science?
- How do I become a Data Scientist?
- How does Data Science differ from traditional statistical analysis?

Related Courses

- Concepts in Computing with Data, Berkeley
- Practical Machine Learning, Berkeley
- Artificial Intelligence, Berkeley
- Visualization, Berkeley
- Data Mining and Analytics in Intelligent Business Services, Berkeley
- Data Science and Analytics: Thought Leaders, Berkeley
- Machine Learning, Stanford
- Paradigms for Computing with Data, Stanford
- Mining Massive Data Sets, Stanford
- Data Visualization, Stanford
- Algorithms for Massive Data Set Analysis, Stanford
- Research Topics in Interactive Data Analysis, Stanford
- Data Mining, Stanford
- Machine Learning, CMU
- Statistical Computing, CMU
- Machine Learning with Large Datasets, CMU
- Machine Learning, MIT
- Data Mining, MIT
- Statistical Learning Theory and Applications, MIT
- Data Literacy, MIT
- Introduction to Data Mining, UIUC
- Learning from Data, Caltech
- Introduction to Statistics, Harvard
- Data-Intensive Information Processing Applications, University of Maryland
- Dealing with Massive Data, Columbia
- Data-Driven Modeling, Columbia
- Introduction to Data Mining and Analysis, Georgia Tech
- Computational Data Analysis: Foundations of Machine Learning and Da..., Georgia Tech
- Applied Statistical Computing, Iowa State
- Data Visualization, Rice
- Data Warehousing and Data Mining, NYU
- Data Mining in Engineering, Toronto
- Machine Learning and Data Mining, UC Irvine
- Knowledge Discovery from Data, Cal Poly
- Large Scale Learning, University of Chicago
- Data Science: Large-scale Advanced Data Analysis, University of Florida
- Strategies for Statistical Data Analysis, Universität Leipzig

Related Workshops

- Data Bootcamp, Strata 2011
- Machine Learning Summer School, Purdue 2011
- Looking at Data

Books

- Competing on Analytics
- Analytics at Work
- Super Crunchers
- The Numerati
- Data Driven
- Data Source Handbook
- Programming Collective Intelligence
- Mining the Social Web
- Data Analysis with Open Source Tools
- Visualizing Data
- The Visual Display of Quantitative Information
- Envisioning Information
- Visual Explanations: Images and Quantities, Evidence and Narrative
- Beautiful Evidence
- Think Stats
- Data Analysis Using Regression and Multilevel/Hierarchical Models
- Applied Longitudinal Data Analysis
- Design of Observational Studies
- Statistical Rules of Thumb
- All of Statistics
- A Handbook of Statistical Analyses Using R
- Mathematical Statistics and Data Analysis
- The Elements of Statistical Learning
- Counterfactuals and Causal Inference
- Mining of Massive Data Sets
- Data Analysis: What Can Be Learned From the Past 50 Years
- Bias and Causation
- Regression Modeling Strategies
- Probably Not
- Statistics as Principled Argument
- The Practice of Data Analysis

Videos

- Lies, damned lies and statistics (about TEDTalks)

- The Joy of Stats
- Journalism in the Age of Data

Source: http://datascienc.es