Harnessing the Scientific Method for Business Success

Roger Longbotham

Data & Decision Sciences Group, MSIT

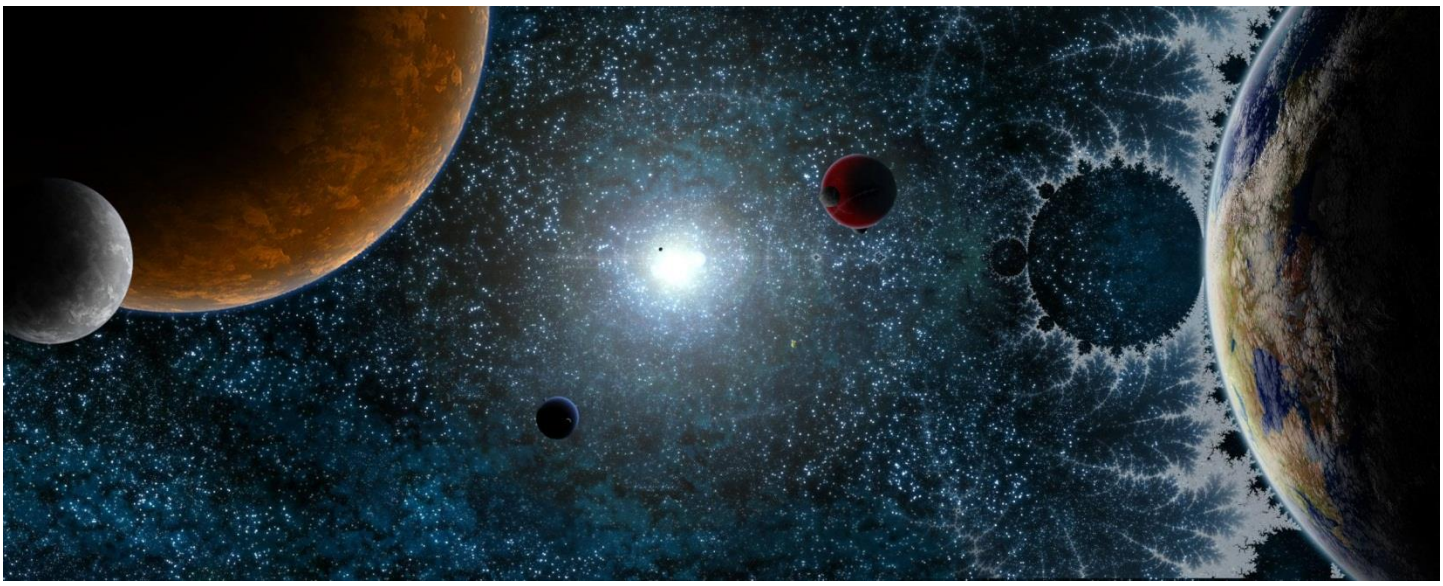


Figure: Fractal Universe[1]

Fractal geometry and statistics bring order from chaos.

[1] From http://www.wallpaperpimper.com/wallpaper/Dual_Screen/Fractal-Universe-2560x1024.jpg

Harnessing the Scientific Method for Business Success

The scientific method is an iterative process of stating a theory (or hypothesis or business question) and testing whether that theory is correct with a well-designed experiment. After the results of the test are known new theories are proposed and new experiments are conducted. This gives an upward trajectory of improvement – scientific advancement when applied in that realm, or business improvement when applied to business decision-making.

We are all familiar with the progress that has been made in science over the centuries after the scientific method began being used. Most of us think of a scientist in a laboratory with strange liquids bubbling in beakers when we hear the word "experiment". But, in the 20th century experimentation moved out of the laboratory and into the field in the disciplines of agriculture, manufacturing, pharmaceutical testing, medicine (clinical trials) and marketing. These and many more areas have benefited from the application of the scientific method and experimentation. For example, recent field experiments have been carried out in developing countries to test the effectiveness of poverty and health program options to determine how to use available resources to best serve the population (ref1, ref2). Some of these experiments are on a large scale and can span large geographical areas and involve thousands of individuals.

Marketing Factors to Test
- TV frequency
- TV message
- Radio timing
- Radio message
- Newspaper insert size
- Newspaper insert color
- Newspaper days
- Direct mail card vs letter
- Direct mail message
- Direct mail personalization

Professor Peter Weill in his presentation to MSIT, Case studies on Transforming the IT organization in May, 2011 calls for strategic experiments. He says that innovating via strategic experiments to test and learn is one of the four areas where we can work smarter to have the most impact. The other three are creating/revising business rules based on analytics, automating repetitive processes, and providing business partners with timely information.

This paper will highlight areas of Microsoft business that can benefit from the application of the scientific method through strategic experimentation using relevant case studies from other companies. I have personally been involved in many of these case studies and the details presented below from other companies have all been made publicly available.

The general procedure is to first conduct a screening experiment with many factors or ideas. In most situations combining many factors in a single test improves the efficiency of the test and allows you to test more ideas in a short period of time, speeding up innovation. There could be as many as 15 to 30 ideas tested in this initial experiment. This is usually followed up by a refining experiment where only the most promising ideas are tested. The business improvements that are "proven" with the refining experiment are then implemented. The appendix gives more detail on the requirements and steps needed to carry out experiments to improve business processes.

Call Center Management

Call centers are a natural place to run experiments because all the elements are present to give good experimental results. (For a list of these elements, see Appendix 2.) In a typical call center there are potentially hundreds of agents handling many calls per day. Each call can have several quantifiable outcomes related to the objective of interest. For example, the length of each call can be measured if the objective is to improve efficiency. If the objective is to increase revenue, sales per call is measured. Other metrics may include customer satisfaction, return call rate, etc.

In some cases several call centers will be part of the same experiment which will ensure the results are applicable to all call centers. Industries that use experimentation in call centers include credit card companies, banks, service and retail. Experimentation works equally well whether the calls are in-bound or out-bound.

Case Study – Financial Services Call Center

A Financial Services firm wanted to improve net sales in their call centers. They had seven call centers where they received calls regarding their credit cards and chose to use three in this improvement effort. In addition to the primary objective of increasing sales they wanted to decrease time on call and improve employee satisfaction since they were experiencing high employee turnover. After conducting a number of brainstorming sessions with customer service representatives (CSRs), team leads and managers the list of ideas was narrowed down to those that were actually tested. The experiment was in three call centers, included hundreds of CSRs, 24 team leads and tested 10 ideas.

The ideas they tested were:

1. Sales coach availability (coach was ready to coach after any sales call or not)
2. Unit manager monitoring calls (or not)
3. Use of lead associates as coaches (instead of dedicated sales coaches)
4. Operations manager available on the floor (or not)
5. Use of unit managers as coaches
6. Increase the time off the phone for call center associates
7. Unit manager availability on the floor (or not)
8. New hire coaching (or not)
9. Self-paced training for call center associates (via taped calls, or not)
10. Self-paced training for call center associates (via Web, or not)

Five of these factors were identified as improving at least one of the key metrics (increasing sales, decreasing call time, improving employee satisfaction). The increase in net sales was approximately four times what management had hoped the experiment would achieve and resulted in an additional millions of dollars in sales per year!

In addition to improvement in the business objectives, a senior vice president commented on the effect of the process on employee morale and engagement:

"One of the nicest surprises was the impact of MVT (i.e. the experiment) on the environment and the culture of the organization…. We learned a lot from our associates across all the sites that were participating in the MVT process."

Marketing Optimization

A saying attributed to John Wanamaker, an early pioneer in marketing is, "Half the money I spend on advertising is wasted; the trouble is I don't know which half" (link). The idea behind marketing experimentation is to determine which marketing efforts are working and which are not. Concentrating marketing dollars on those efforts that have the highest ROI is obviously a winning strategy.

Case Study – Effect of Online Display Advertising on Online as well as In-Store Sales

Yahoo! Research partnered with a nationwide retailer (with online and physical store presence) to study the effectiveness of online display advertising on online and in-store sales with more than three million shared customers. They measured the impact of higher online ad impression frequency using a simple design that varies the ads that users see on the Yahoo! network within identically targeted campaigns. Users in the treatment group see the retailer's ads; users in the control group see unrelated `control' ads; and users in the half treatment group see an equal probability mixture of the retailer's and control ads. They found a statistically significant increase for in-store sales as a result of the ads. Doubling the number of impressions per person--from 17 to 34 in a two-week period--approximately doubled the treatment effect. They also found evidence that the ads most strongly affected customers who live closest to the retailer's brick-and-mortar locations: those who live within two miles of a store experience six times the incremental sales lift due to ads as those living farther away. Sales to the online store had a slight increase but was not statistically significant.

Other marketing studies I have been involved with have tested market mix (TV, radio, newspaper ads) for retail store sales as well as direct mail and email. It is not easy to conduct a valid experiment with some of these channels, but given the large investment, it is worth the effort.

Retail Improvement

I have personally been involved with experiments to improve retail operations with Saks and several of its subsidiaries, Toys-R-Us and AutoNation. Other retailers that have made improvements through experimentation include Pilot, Quick Chek, Lowe's, and others. The most common goal in these experiments is revenue improvement with cost reduction or profit margin often a secondary objective. The business ideas that were tested included media factors (e.g. TV, radio, newspaper), in-store signage, in-store layout, sales force appearance, staffing levels, sales process, loyalty program features and more.

Case Study – Retail Store Chain

This case study involved a large regional retailer. This retailer had over 150 stores in a six state region. After an analysis of the most recent two years of weekly sales and discussions with operations managers 32 stores were ruled out as being ineligible. Since we were using comp sales (i.e. ratio of this year's sales to last year's sales) any store that was not open for at least a year or had an unusual disruption in sales

last year was disqualified. Also, any store where major changes had taken place since the previous year (e.g. major remodeling or increase in square footage) was eliminated as was any store that was expected to have a disruption during the time of the experiment.

After many brainstorming sessions and filtering of ideas, we were left with 23 ideas to be tested. A sampling of some of the ideas tested:

1. Change in sales associate attire,
2. Changes to the newspaper circular,
3. Store signage (prices, directions),
4. Product demos,
5. Checkout procedure,
6. Changes to sales training,
7. Sales incentives,
8. Management incentives, etc.

Since there were 23 ideas to be tested, we needed a statistical design that would be able to estimate the effect of each of these on sales independently with maximum sensitivity. (The technical term for the statistical design we used is a 24 run Placket-Burman design (ref).)

We needed 24 groups of stores for this test where each group of stores received the same set of factors. Therefore, a subset of the 118 stores was randomly assigned to the 24 groups with an almost equal number of stores in each group. Since we had some media factors (newspaper and radio) in the mix we used a restricted randomization scheme (stores in the same group had to get the same radio and newspaper factor). The screening experiment ran for six weeks.

Seven of the 23 factors in the screening test were statistically significant with five having a positive effect. These five were carried into the refining experiment. Since fewer factors were being tested the second experiment was logistically much simpler but the size of the experiment (number of stores and number of weeks) needed to be about the same to get the same sensitivity as the screening experiment. We chose a statistical design that would allow us to estimate interactions among the factors. We used a full factorial in the five factors, which required 32 groups of stores.

The final analysis showed all five of the factors in the refining design would help sales and gave a prediction of a 10% increase in sales if all five were implemented. Follow-up analysis confirmed a sales increase of approximately 10% for the chain after implementation of the five ideas.

Problem Resolution

Although we would hope it weren't the case, sometimes our customers have a problem. It may be with the software itself, or it may be with the licensing process or other issue. This often leads to what is known as the "moment of truth" when the response to the customer determines whether that customer will abandon Microsoft, remain as an unhappy customer or become a loyal advocate. These interactions should be anticipated and a planned response ready for the occasion. Some of the

responses are quick and inexpensive and others may be prolonged and expensive. Are the expensive solutions worth it? Can we implement inexpensive solutions that have higher ROI than the more expensive ones? An experiment can help determine how we should respond.

Case Study –Insurance Company

This insurance company had done several process improvement projects, some with designed experiments, to improve cost structure and customer satisfaction in the call center, insurance application and claims processing processes but still had a significant problem. There was an alarming trend toward more claimants hiring attorneys where there was bodily injury (BI) after an auto accident. When they looked at the data for a large number of claims they found that for claims with attorney involvement 1) the cost to the company was significantly higher, 2) the claims process took significantly longer and 3) the claimant received less than $100 on average. Therefore, it seemed it would be in everyone's best interest to reduce attorney involvement in the claims process (except for the attorneys ☺).

At the beginning of this effort 40% of the claims with BI had attorney involvement. Initial improvement efforts targeting the "low-hanging fruit" lowered that to 36%. Each reduction of 1% equated to a savings of $6,000,000 to the insurance company.

They held brainstorming sessions with many employees and did several surveys to get a long list of potential ideas they could test to reduce the percentage of claims where an attorney was retained in the first 60 days after the accident. They narrowed the list of ideas to be tested down to 13.

The screening test showed four ideas that had a beneficial effect (in order of benefit):

1. Pay more than blue book value when the value of the automobile was debatably higher. (This was a controversial finding which went against the philosophy of the CEO and the industry – to pay more than necessary for property damage claims in order to reduce the overall cost of bodily injury claims.)
2. When appropriate (and legal) give the claimant an open-ended BI release form, i.e. if the claimant came back later with additional medical expenses than originally anticipated, Progressive would pay those expenses.
3. Increase the number of in-person contacts with the claimant.
4. Increase the range and discretion by agents in settling BI claims.

These ideas were all put into a second (refining) experiment and not only were they validated, they got a more precise estimate of the cost and benefit of each idea. The final, verified improvement in percentage of claims with attorney involvement was a reduction of 8%, from 36% to 28%.

Other Opportunities

In this section I will just briefly mention some other areas where business experimentation could be applied effectively to determine the optimal business "control settings".

Sales Management

Experimentation has been used effectively to determine how sales personnel manage their sales prospects, from contact strategies to sales messaging. In a previous section, telephone sales was mentioned, but this applies equally to in-person sales when a large sales force is involved.

Manufacturing quality and efficiency

Many authors have recorded successes with experimentation in manufacturing. Perhaps the best known approach is the Taguchi method, but other approaches to experimentation have been quite successful as well. The goal of these experiments is usually to improve throughput, quality, cost, efficiency or even environmental impact. Of course, all of these aspects can be measured and quite often more than one can be optimized. For example, an experiment may find some factors that increase throughput and others that improve quality leading to improvement in both metrics.

Online optimization

I, along with a number of co-authors, have written extensively about the value of testing one idea at a time (A/B tests) and more advanced experiments in ThinkWeek papers (2008, 2009) and in professional journals (Data Mining and Knowledge Discovery 2009, IEEE Computer 2007 and 2010, SIGKKD 2009) so I will leave it to the interested reader to follow up with these references. Microsoft already has a good experimentation capability for Bing and a few other online properties, but, with the retirement of the Experimentation Platform in 2010 many other Microsoft websites and web services need to find the capability through partnering with Bing or contracting with an external vendor.

Telemetry Testing

A recent experiment was completed to test the effect on performance of adding a small amount to file headers for an operating system. His test randomly assigned over 16,000 machines to get the build with the extra bits in the header or not. By using a large number of machines outside the lab setting they were able to demonstrate whether the additional load would significantly affect performance or not.

APPENDICES

I have included two topics for those who want to understand more about how these experiments may be carried out. It is not possible to give all the details in a short paper such as this, but, the sections below can help put a framework on how these experiments are conducted (Appendix 1) and the conditions that are optimal for business experimentation (Appendix 2).

Appendix 1: Steps in Business Experimentation

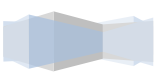
First, what type of process should you consider for experimentation? The best candidates are processes that 1) have fairly quick repetition and 2) are expected to remain in place long enough for the implementation to have an adequate payoff after the experiment is conducted. The first of these is to assure the experiment can be carried out in a reasonable time frame. If the experiment design required 24 process runs and there was only one process run per month, the experiment would nominally take two years which is outside the planning timeframe of most organizations. In some cases like this, an

innovative experimenter may fine a different approach to get the experiment conducted in a timely way, but an experiment that takes more than 6-8 weeks is probably too long for the organization to consider. The second requirement is just so the positive results of the experiment can be in place long enough to make a substantial difference to the company. If your experiment ends in May and the results are implemented in July but the process sunsets in November, you'll only get five months of value from the experiment, at most. It is better to choose a process that is expected to be in place much longer.

Conducting an experiment to improve business results is best done when part of an overall effort to improve the business process, but could be used as a special initiative. Using a baseball analogy, most process improvement efforts involve hitting many singles with continual improvement as the objective with an occasional or rare home run along the way. A well-run experiment could be one of those home runs and some would even be considered grand slams.

Typical steps in a successful business experiment are

1. **Clear business objective and an environment for success**
   Clear business objective
   This should be a well-defined goal with high value to the organization.
   Environment for success
   A process owner with sufficient clout should be champion of the improvement effort with a PM (or equivalent) running the day-to-day activities. In some cases a cross-functional team would be needed to coordinate the activities.
2. **Valid measurements**
   You need the ability to measure process results directly related to the business objective, which could be a KPI or other results metric. You may have more than one result metric of interest (e.g. profit, revenue and customer satisfaction). Every effort should be made to make sure the measurements are timely, consistent, and repeatable, with low variation. Measurement studies should always be done. Most measurement systems are found to be inadequate at the beginning of one of these tests.
3. **Some level of process stability**
   This is to ensure the effects of the ideas being tested can be seen above the noise of the process. If the process is not stable across time (as is the case with most business and online experiments) you may be able to gain stability by conducting all the experiment variations at the same time. (If that last sentence does not mean a lot to you, don't worry. It just means there may be more than one way to gain the stability needed for your experiment, you just may need to get creative or call on an experimentation expert.)
4. **Brainstorm ideas to test**
   Get ideas from many people, focusing on the experimental objective(s). All primary functions in the process should be represented, especially individuals who are closest to the day-to-day operation of the process.
5. **Determine which ideas would be tested**
   Do not choose the ideas that you think will work the best. Our ability to predict which ideas will work and which will not has been shown to be almost the same as flipping a coin. Instead use



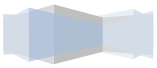
criteria such as ideas that are easy to test, inexpensive to test and can be ready to be tested quickly.

6. **Screening test to determine which ideas are best candidates for improvement**
   Since most experiments will have many ideas being tested, the first experiment is to determine which ideas are the most promising. Organizations that have conducted experiments consistently find that 25% of the ideas tested or less have the potential to improve results. So if you have 16 ideas in your screening experiment you may have 4 (more or less) ideas that make it to the next step.
7. **Refining test to validate winners and optimize results**
   Many organizations consider this step optional, but best practice is to run a second experiment with the promising ideas from the screening experiment. The objectives of the refining experiment are to validate these ideas, get an improved estimate of the actual amount of improvement you can expect from them and to get the optimal combination. (Technically, a screening experiment does not give you any information about interactions between the ideas in that experiment so the optimal combination can only be determined from the results of the refining experiment.)
8. **Implement and monitor improvement**
   Only the ideas from the screening phase that are validated should be implemented. You should be prepared to monitor a) how well the ideas are implemented and b) the impact the improvements have on the KPIs. A control chart is a good tool for the latter. In some cases, you may want to run a follow-up experiment to test ideas that couldn't be tested in the first experiment or that were conceived during the first test. Also, some ideas may have additional settings that could be tested for further improvement. For example, if you found that offering the customer a one month extension on the annual license yielded 15% improvement in sales, in your next experiment you may want to test whether offering a 2 or 3 month extension would have a sufficient ROI.

Appendix 2: Elements that are Conducive to Good Experimental Results

It is not necessary to have all of these present for a good or valid experiment, but the more that are satisfied, the more likely it is you will have a valid, impactful experimental result.

1. Clear quantifiable outcome metrics that are directly related to business objectives.
2. A sufficient number of potential/eligible "experimental units". This depends on the test, but 12 is a bare minimum for almost any test and many tests may require 30 or more. (The outcomes for the experiment can be measured independently for the different experimental units and the treatments can also be applied independently to the units.) The number of experimental units needed is related to number 5 below. The larger the variation in the primary metric, the more experimental units are required.
3. Relatively stable environment, or can be made so through appropriate controls, test design or transformation. For example,



    a. In online tests, this is achieved by requiring all treatment groups be collecting data at the same time (concurrent) and the percentage in each group does not change throughout the test.
    b. For retail experiments, this would involve choosing stores that are in a stable time period and using year-over-year ratios for outcomes.
4. Controllable test variables/ideas in order to optimize outcomes.
5. Either a large amount of data or small variation in order to separate the signal from the noise. (With online tests the variation relative to the change we want to see is usually so large that we need a large number of users in these experiments.)
6. Results of actions that can be seen in relatively short period of time (days or weeks, not months).
7. Results have sufficient business merit to justify the expense of running the experiment.