ASSUREDSCIENCE

Discover how and why Assured Science came to be

6: Data problems in the traditional system

The majority of data are too idiosyncratic, incomplete, un-validated, un-integrated, incoherent, and unwieldy to foster efficient and rapid progress.

I started thinking about my research subject, a gene called *Notch*, about 25 years ago, in 1989-1990. *Notch* had been extensively analyzed since the 1930s in the fruit fly *Drosophila*, which is an excellent model system for studying the genetic regulation of animal development. I was looking for a gene to serve as my 'informant' on how a group of similar cells differentiate into different cell types or tissues. *Notch* fit the purpose very well, as it mediated communications between cells through cell-cell contact and appeared to be required for the development of all tissues in the animal. Now we know that *Notch* and its relatives are most likely required for the differentiation of all tissues in all animals.

The next few years, I read almost all the publications on the subject (which was possible then) and performed exploratory experiments. While data from others in that period suggested that *Notch* functions involved more than one kind of activity, my data suggested that as cells differentiated, the protein synthesized from the *Notch* gene became different (via trimming) in the different emerging cell types. As the different parts of the *Notch* protein had different purposes, it occurred to me that this protein becoming different as cells differentiated could be an important part of development regulation. So it was sometime in 1993-94 that I embarked on the journey to study *Notch* functions from that perspective. My general strategy was to keep my model in mind while designing experiments but follow all data that are interesting, insightful, or informative whether or not they fit my model.

After more than twenty years of additional experiments and data, all I know is that my model might very well be right but it is much more complicated than I imagined. And there is no way that I would be able to rigorously test the model in my lifetime. Sadly, it appears that neither would anyone else in my lifetime. You see, for a while I interpreted my discovering new data as meaning that I was doing well and staying ahead of the competition. Pretty soon others would join the path and together we would generate more data

that would establish and expand a strong line of research. Such was my belief until about 2006-2007. This belief was shaken when one colleague said to me "You always manage to find new things," while another said, "You might be right, but you are the only person working on it." I began to wonder, am I so far off on my own path that there is no one around me? Am I generating new data in which no one is interested? Are my data useless to the field?

I had to find out, if only to reaffirm myself in my own mind. So I reviewed data in my papers and grant applications, anonymous reviews of these data, personal discussions with colleagues, and progress in various aspects of the field since my entry. I was relieved: my data are published in good peer-reviewed journals, cited by others in their publications, and others have validated some of my results in their publications. I now think that my finding so many new data was simply because I was first on the path. My working alone is because many of my colleagues are interested in different issues while others prefer the simpler popular model in the field. As has happened so many times in my own research, I would flip on a dime if I see data that better reconciles all the data in the field. Until then I stick to Einstein's advice: "Explanations should be as simple as possible but no simpler."

While I take comfort in my review of my published data, I find only extreme discomfort when I consider my unpublished data. I am talking only about data sets that I think are well-established. For some of them I understand their significance very well, for some others I have good guesses, and for the remaining I have no clue. But I am convinced that most of them would provide important insights into how the *Notch* gene works in different tissues at different stages of development, and what its ultimate biological function could be. The thought that some day, pretty soon, these data will be discarded and the knowledge I gathered from them will die with me is sad, to say the least.

Actually it is a shame considering how much time, effort, and money I spent on generating the data that are now buried in the piles of folders and boxes around my bench and desk. I bet that any laboratory bench occupied by the same researcher for five years or more contains a significant amount of data that are reliable but unpublished for a variety of reasons. The value of these data is obvious to those that produced them and not apparent to others, even if they ever take the time and effort to go through them.

I do not view my unpublished data any differently than the way I view my published data, in terms of their value for understanding my subject. To me they are only different in the degree to which they have been developed to

significantly advance our knowledge of a problem or an issue, which is commonly referred to as *telling a story*. When I pondered how the appeal of a good story influences our perception of what the data could actually mean, it changed the way I looked at existing data (including my own). The closest analogy I can give of the before and after of that question is this: clocks before and after Salvador Dali's melting clocks. Even the hardest data could soften (melt) in the wrong story. The contrary, although rarer is also true: even the softest data could appear hard in a great story.

I see three basic problems with data in the traditional system. Please bear in mind that I am not saying that these problems affect all data or any specific data set but the majority of data in aggregate. Please also bear in mind that it is this majority data that consumes the most resources, including time, effort, and funds.

a. Data are Extremely Idiosyncratic

One major problem with data in the traditional system is the extreme idiosyncratic nature of most of them. I am not talking about the small fraction of data in every field that is related to very basic aspects or widely useful empirical methods. These data become part of the research of a good number of scientists in the field, sooner or later, directly or indirectly. But most other kinds of data have a different 'life.' They are generated in independent laboratories, each following it's own questions, approaches, empirical procedures, adaptations of methods, or using unique materials. Data from my own laboratory is a part of this majority. Instead of becoming part of other people's research, these data have a 'following' to different degrees, in the sense that some are cited more often than others in publications by other independent laboratories in the field.

I say following because citations are generally used to provide relevance or significance, in other words, justifications or support for the study or the results obtained. To grasp what I mean, consider the simple fact that most citations in publications appear when authors present the background to their study or discuss the significance of their results. The Results section, in other words the Data section, which is most often the largest section in a publication, has very few, if any, citations. This widespread practice has a purpose: to separate new data from old or to segregate data of the authors from those of others in the field. Unfortunately, given the insular way most laboratories function, data within a field exist more or less as independent lineages with different degrees of popularity.

Let me explain. In my view the relationship between citations and data in a publication is intimate in terms of context but distant in terms of

dependence, except when the citation refers to data from the same laboratory or that of a collaborator. As only a relatively small number of laboratories actively collaborate, and rarely for long periods of time, data from one laboratory by and large do not actually determine or cause the data from another laboratory. They might initiate, influence, or explain, but they do not determine or cause. In other words, there is no dependence in the sense that one piece of data depends on another in the results section of a publication. This is true even when different, independent laboratories use the same materials or methods.

To better understand that situation, imagine the different laboratories in a field as houses in a neighborhood. Each family in a house goes about its living more or less independently of other families in the neighborhood. There might be casual interactions, socializations, joint actions, exchange of materials (gifts and borrowing), and shared interests but what happens in one house rarely determines what happens in others. When offspring leave home, more often than not they continue a different line of work and are generally averse to competing with their parents. When scientists work in such a manner, data suffers. The adverse effect is so pervasive and widely acknowledged that it has been written about in highly regarded scientific journals and newspapers (Figure 1). The titles of such articles are obviously provocative and simplistic, possibly to draw attention. But they capture very well, much better than I ever could present, the essence of the data problem in the traditional system: they are too idiosyncratic.


1. Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

PLoS Medicine | www.plosmedicine.org August 2005 | Volume 2 | Issue 8 | e124

2. The New York Times JAN. 20, 2014

SCIENCE

New Truths That Only One Can See

Raw Data

By GEORGE JOHNSON

Figure 1

In my view, such a high amount of idiosyncratic data disabuses the notion that competition within today's traditional system is promoting progress. Before going any further, let me stress that I am not talking about the minority of truly revolutionary or game-changing data, but about the majority of data that are expected to produce incremental progress. Scientists involved here compete for funding, publication, or position but their data are effectively insulated from competition. In other words, whether data lineages flourish, perish, or are ignored is determined by the success of scientists. There is an argument that scientists with better data succeed. But that is more of an assumption, even wishful thinking, because for most data it is not possible to know which is better without having data

compete directly in terms of explaining a phenomenon or directing efficient progress over long periods of time.

Furthermore, when competition is intense, as it is now, scientists survive and flourish by working secretively and produce even more idiosyncratic data in order to quell competition. And it gets worse. Scientists working on the same or closely related problem are the ones who could be the most effective collaborators, but they are forced into becoming intense competitors for not only success (funds, publication in high-profile journals, and prestige) but also survival. Consequently, collaboration is suppressed, which renders the data even more idiosyncratic.

b. Data are Very Selected or Incomplete

Another basic problem with data in the traditional system is that they are either selected or incomplete because only data that are consistent with a hypothesis or a model, or can be otherwise explained, are generally published. Inconsistent or unexpected data are ignored because the results cannot be convincingly explained given what is known about the subject. Even data that can be explained but are inconsistent with popular models are often abandoned due to difficulties in obtaining funds or publishing them. What may be worse than abandonment is the pressure to find explanations that are consistent with popular models however contrived they might be. This compulsion would take scientists on the wrong path, resulting in even more unexpected or inexplicable data and even more contrived hypotheses. In other words, good data becomes housed in the wrong story.

I am not implying that people who review papers and grants are averse to data contradicting popular models or new approaches. In fact many of my publications contained contrarian data and I still treasure some of those long and well-constructed criticisms that made my publications better. Thus, it is not reviewers *per se* that cause selected and incomplete data sets to be published but the natural biases in the way the traditional system works.

Given that data have to be explained for publication, data and explanations that are consistent with popular models receive less scrutiny simply because they are familiar or reassuring. On the other hand, data or explanations that are contrary to popular models, or are entirely new, require a lot of additional data to convince. Often, not enough is known to even design experiments that yield convincing data. If one is stressed for time and resources, the practical choices are either to abandon the data or find an agreeable explanation in order to pursue it. It might seem like a lose-win choice but in actuality it is a lose-lose choice. I have been there and know it well. In one the scientist loses, in the other science loses.

Data in the traditional system are also biased in favor of trends. Whenever a research area is trending, generally following an important discovery, it is easier to publish papers and obtain funds for the 'hot' topic. For a brief period the interests of scientists, journals, and funding agencies match.

Why Are So Few Blockbuster Drugs Invented Today?
By DAN HURLEY NOV. 13, 2014, New York Times

'The claims made for genomics in the 1990s sound a bit like predictions made in the 1950s for flying cars and anti-gravity devices.' So far it has produced fewer returns on greater investments.

The spirit that animates the trial-and-error chemists is an enthusiasm for constructing new drugs piece by molecular piece, like children playing with building blocks.

Figure 2

However good that can be, it comes with costs. One, the interest is fleeting in comparison to the time required to understand any phenomenon well, which creates cycles of boom and bust that result in uneven progress across a field. Two, since funding is limited, other fields suffer. Three, as there is no mechanism for continuously assessing emerging data, the lack of commensurate progress in knowledge becomes apparent only after a huge amount of time and resources are already spent. Therefore, it is not surprising to see articles like the one shown in Figure 2. This article focuses on the belated realization that the new area of genomics is not delivering on its promise despite considerable investment in time and money for almost two decades now.

Similar conclusions can be made about a number of trends or fashions that periodically sweep different areas of research. To me these are simply expected consequences of selected and incomplete data. In the example shown in Figure 2, data are selected because genomics (studies involving the whole complement of genes or their products in cells, tissues, or organisms) is but one level at which genes function and not all genetic regulation can be studied effectively at that level. For example, the RNA product of my subject, the *Notch* gene, is highly regulated, which when perturbed has serious consequences for development. However, most genomic studies fail to pick up changes in *Notch* RNA levels. Genomics data are also incomplete because they are by and large based on correlations or associations and there could be a big gap between those and causation. Filing such holes in data would require tedious, slow, and difficult experiments that go by the non-trending name of Genetics.

c. Data are Too Disparate, Un-integrated, and Unwieldy

The third major data problem within the traditional system is that the data are disparate, un-integrated, and unwieldy. These three issues are linked and arise from the way data are published. There are thousands of scientific

journals of diverse focus, varying from a specific area in a field to covering all scientific fields. Furthermore, each journal has its own standards and criteria for publishing papers. As a consequence, journals differ in quality of publications, number of publications, depth of coverage of topics, scope of interest, and the format of data. The main goals of journals also differ. Some like to publish papers with the potential to make a high impact on their respective fields, some just publish useful or interesting papers, and some others publish for profit. In reality, most journals mind all three goals albeit to different degrees and care most about performance of publications in their journals.

Many measures are used to assess performance of publications. A popular one is impact factor, which is the frequency of citations of a typical publication in the journal. The higher the number of citations, the higher the assessment of performance in the field (i.e., impact). For example, the impact factor for *New England Journal of Medicine* is ~54 (currently the highest value on the scale), which means that a typical publication in that journal was cited 54 times in the last two years. High impact factor journals are also very selective, publishing only a small fraction of manuscripts they receive. There are variations and elaborations on impact factor, but the basic one is sufficient to make my point.

Scientists, journal editors, and administrators rely heavily on impact factor in one way or the other for the assessment of science and scientists. Using a measure like impact factor to assess performance is laudable, but given the way the system works it is in my view *the most disruptive force today that is responsible for data dispersion, devaluation, and disuse.*

To get an idea of what I am talking about, consider the following. In the traditional system, having a publication in high impact factor journals is so richly rewarded (funds, jobs, status) that almost every scientist dreams of publishing there (I did too!). Most of these journals have very broad scope or focus, in effect covering whole fields or all of science. As a consequence, important papers across all fields are scattered among these journals. The same thing happens at every rung down the impact factor ladder. Now consider what has happened to all the data in any specific field. They are now dispersed among journals with different impact factors. This dispersion has a significant consequence for how data develop that is not good for either science or efficient progress. Let me explain.

Although it is a simple numerical value based on citation of one publication in another, the impact factor transforms into a complex measure through the way it is used by scientists, journals, and funding agencies. *It is transformed from an assessor into an influencer.* More or less objective

assessments such as rigor and validity of data or experiments are combined with more or less subjective assessments such as interest, significance, value, trends, and future prospects of data to boost the value of the impact factor. Unfortunately, these subjective assessments are very difficult for the majority of data and are essentially value judgments based on anticipation and prediction that may or may not be fully tested for years to come or even decades.

In the meantime, as the system favors and rewards data in high impact factor journals, more scientists will be attracted to this subset of data in the field and generate more data related to it. Consequently, this 'popular' line will flourish and proliferate. Other lines of research with data that are also valuable, often related to the same phenomenon, receive less attention, even disregard, because they are in journals with relatively lower impact factors that are not magnets for funds, jobs, or status. Data in these journals may be even more intriguing or interesting but are not developed enough, easily understood, or are contrary to the data in the popular lines of research. These data are effectively devalued and most of them fall into disuse for lack of support and following. However justified popularity is in the short term, it is disastrous for the field in the long term because it results in uneven development of data.

Impact factor-engendered bias perpetuates in other, less obvious ways. Consider the most common means by which people access data, through the Internet via search engines such as Google. Upon entry of search terms (which could include author name), one retrieves search engine results pages (SERPs), often pages and pages of them, that contain a list of publications (often with a brief description). Within minutes, if not seconds, one can be clicking on or touching specific publications of interest.

The first issue that confronts the searching scientist is access to data. Not all publications are freely accessible. Many require subscriptions for immediate access or a waiting period of 1-5 years when they become free. Many institutions subscribe to the journals commonly used by their scientists (generally journals higher up on the impact factor scale) and scientists with funds can subscribe to additional journals relevant to their research (journals lower down on the impact factor scale). Thus, data in higher impact factor journals enjoy more exposure, and scientists with funds have access to more data including those in low impact factor journals. The irony of the latter issue is that scientists with funds are generally focused on popular data, which are in high impact factor journals. Nowadays, there are a number of free-access journals that charge authors instead of readers. Here also, only scientists with funds can afford to pay. The positive feedback cycle favoring

data in higher impact factor journals extends to attendance at scientific conferences and meetings at which the latest data are presented.

The next issue confronting a searching scientist is the bias in the list of publications on SERPs. If one is directly using a search engine like Google, the ranking order of items presented to the searcher is essentially, even primarily, based on the popularity of the link between search terms and webpages (in this case, publications) as it exists in the web or extracted from previous users and SERPs. Therefore, the most popular publications appear near the top and the least popular ones appear somewhere near the end, close to items with very few or no links. Obviously, data from popular lines of research in high impact journals are higher up on the SERPs and those in low impact factor journals are way down.

Some websites make a fair presentation using a different criterion. For example, PubMed uses a chronological order for listing publications on SERPs. Still, bias creeps in due to human nature. As most data are not integrated, and data in one publication rarely directly depends on data in another, the items on a SERP are generally incoherent. Data in one paper might be at the level of behavior of an organism, while that in the next paper might be at the level of interaction between two molecules in a cell. Unless, of course, one is using a very specific string of search terms, in which case they would more likely be locating a specific publication instead of searching for related publications for use in planning research or experiments.

So, it is up to the scientist to individually examine each document within each SERP to verify the relevance and validity of data. The publication with the most relevant data might be buried deep in the almost endless pages of SERPs retrieved and the scientist would have no *a priori* way of knowing it. Going through all the items on the SERPs is so unwieldy and daunting that even the most scrupulous and patient scientist would end up resorting to secondary sources that are also on SERPs in order to identify papers closely related to their research. Therefore, it is not surprising that many journals that publish reviews, digests, or trends are very popular (and have high impact factors). In fact, the top items of the first Google SERP on a search with terms related to my research subject, "Notch and Cell Fate," are indeed reviews.

Gone are the days when reviews were comprehensive or exhaustive; now there is too much data for that. Many 'reviews' nowadays are more status reports of the popular data in the field or a summary of knowledge that is naturally based on data that can be explained or is well understood, which in general are data in journals at the higher end of the impact factor scale.

Furthermore, authors of most popular reviews (in high impact factor journals) are generally those who have published in high impact factor journals and the review format tolerates to a higher degree personal views, if not advocacy. Sadly, all these biases within biases would be perpetuated in the next generation of scientists, as they would also rely on secondary sources because the primary data are so disparate and unwieldy.

The sadness I feel is not just an analytical response but a very personal one as well. Data from my own laboratory is published in journals with impact factors varying from ~2.5 to ~11. Data I generated in a similar manner, which are of equal validity or usefulness, are also in publications in journals with impact factors between ~30 and ~40. If I have to pick from among my published data those with the potential to resolve many perplexing issues related to my research subject, they would be from journals at the lower end of the impact factor scale. These data are effectively isolated, even segregated, almost like the unpublished data on my bench. Now, I can only wonder how much more we would have known about my subject had these data been given the chance to develop to the extent necessary for integration with the popular data.

I have talked about data in the traditional system being idiosyncratic, trend-driven, disparate, un-integrated, and unwieldy. But does any evidence exist suggesting a negative consequence to science from such data? Well, it is not possible to assess that, because until now the traditional system has been the only system for research. But it is possible to get some idea by considering results of clinical research and trials, as they are a test of the performance of biological data relevant to humans. For example, more than 75 years of data on my research subject, the *Notch* gene, obtained from nematode worms, fruit flies, mice, humans, other animals, and cultured cells are tested (directly or indirectly) in any clinical trial related to cancer or Alzheimer's disease. The titles of articles shown in Figure 3 indicate that the data underperforms, to say the least. The example of the shown trial (Figure 3.2) failed in part because of the unexpected responses of the *Notch* gene. Thus, I can well imagine how the nature of data in the traditional system could have contributed to the failure and premature termination of this Alzheimer's clinical trial.


1. Contradicted and Initially Stronger Effects in Highly Cited Clinical Research

John P. A. Ioannidis, MD

218 JAMA, July 13, 2005—Vol 294, No. 2 (Reprinted)

2. Lessons from a Failed γ-Secretase Alzheimer Trial

Bart De Strooper[1,2,3,*]

Cell *159*, November 6, 2014 ©2014 Elsevier Inc. 721

Figure 3

The overall problem is that now most data more or less simply coexist without being co-dependent, integrated, or validated. That is because in the traditional system data is effectively a product of selection applied to the scientists or publications. The situation would be quite the opposite if selection were applied directly to data, focusing only on fitness of different data sets in explaining a phenomenon. In this situation, there will not be different classes of data (popular, devalued, and unpublished), but only one class wherein all data would be good, validated, and integrated. And, all lines of research in a field would progress at their natural rate, with competition and collaboration working together instead of in opposition. Such a system is not just desirable but also possible.

The new system of research is indeed such a data-driven system that is designed to also eliminate, mitigate, or avoid all the funding, mechanism, and data problems in the traditional system that have been discussed in this essay and in the previous two essays, essay 4 and essay 5. In the new system, competition would be focused on data and the only way forward would be collaboration among scientists to seek accommodation or fitness, whether the data are concordant, discordant, or intriguing. Such 'competitive collaboration' is a better driver of efficient progress in research. In this new system for research, measures such as impact factor would be a potent synergistic force instead of a disruptive one.

In the next essay, essay 7, I will present you the reasoning and guidelines I used in designing the new system for research so that you will have a good idea of its philosophy, principles, and purpose. After reading that essay I believe that you will have sufficient information to decide whether or not you want to become involved in the effort.

Cedric