Automating Complex Data Analysis: A Disruptive Business Opportunity

by John F. McGowan, Ph.D.
Web Site: http://www.mathematical-software.com/
jmcgowan79@gmail.com

ABSTRACT

Complex data analysis is a multi-billion dollar business. Major data analysis tool makers alone report revenues totaling over $4 billion per year: SAS Institute ($3.2 Billion), IBM SPSS ($0.3-1.0 Billion), MathWorks ($850 Million), Wolfram Research (at least $40 million), and a number of less well known smaller firms. Medical businesses, financial firms, and science and engineering organizations spend billions of dollars per year on these tools and the salaries of the analysts, scientists, and engineers performing the analyses.

Complex data analysis increasingly determines the approval of new drugs and medical treatments, medical treatment decisions for individual patients, investment decisions for banks, pensions, and individuals, important public policy decisions, and the design and development of products from airplanes and cars to smart watches and children's toys.

State-of-the-art complex data analysis is labor intensive, time consuming, and error prone — requiring highly skilled analysts, often Ph.D.'s or other highly educated professionals, using tools with large libraries of built-in statistical and data analytical methods and tests: Excel, MATLAB, the R statistical programming language and similar tools. Results often take months or even years to produce, are often difficult to reproduce, difficult to present convincingly to non-specialists, difficult to audit for regulatory compliance and investor due diligence, and sometimes simply wrong, especially where the data involves human subjects or human society. Many important problems in business and society remain unsolved despite modern computer-intensive data analysis methods.

A widely cited report from the McKinsey management consulting firm suggests that the United States may face a shortage of 140,000 to 190,000 such human analysts by 2018: http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation

Automating complex data analysis using Artificial Intelligence (AI) and similar technologies can substantially reduce the cost, time to completion, increase the quality, and yield results that are currently impossible. New tools that automate complex data analysis are a disruptive business opportunity.

This white paper discusses the current state-of-the-art in attempts to automate complex data analysis. It discusses widely use tools such as SAS and MATLAB and their current limitations. It discusses current products that attempt to automate complex data analysis.

The white paper presents some preliminary results from a prototype automatic data analysis system. It concludes by asking potential users of the automated data analysis system to contact us with their data analysis problems (use-cases) and representative data.

OUTLINE

1. Introduction to Complex Data Analysis
    1. Historical Examples
    2. Ball Dropped from Roof Example
2. What is meant by Automating Complex Data Analysis
3. Case Study: Vioxx
    1. Horrific failure of complex data analysis
    2. How failure might have been avoided by automation
4. Preliminary Prototype Studies
5. Current Attempts to Automate Complex Data Analysis
6. Conclusion and Call for Data

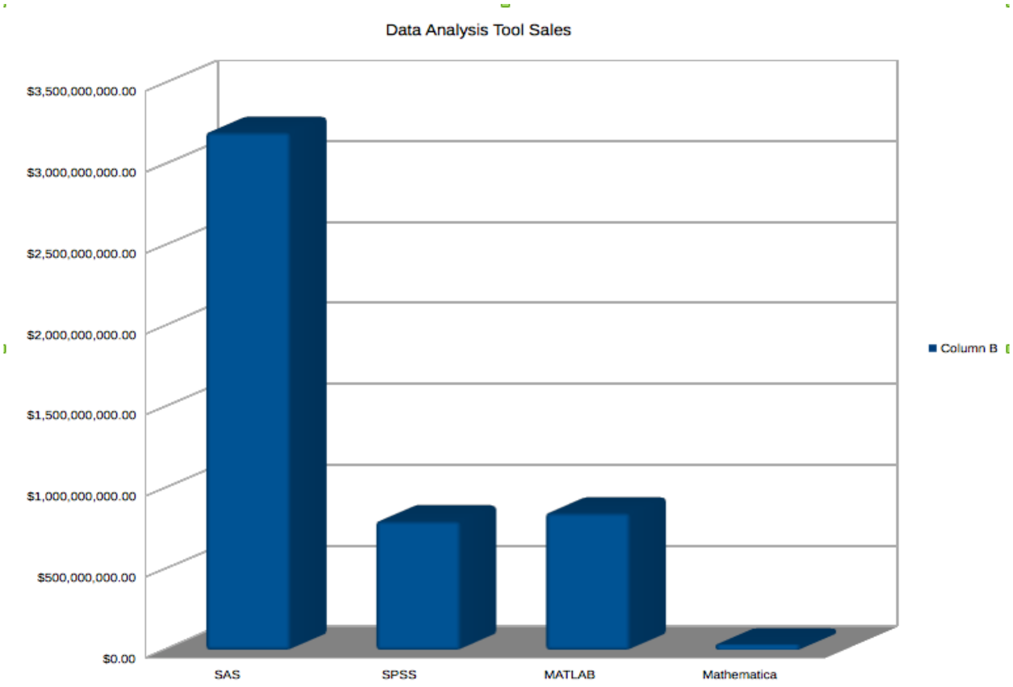
Copyright Notice and License Terms

Disclaimer

The author is *not* a qualified medical professional. Nothing in this white paper, specifically including the Vioxx case study, is intended as medical or legal advice and should not be treated as such. If you have a medical, legal, or other issue or concern that requires qualified professional advice such as a medical condition, you should get advice from a properly qualified professional such as a licensed medical doctor or attorney.

INTRODUCTION

Complex data analysis is a big business. Total annual revenues for software tools used for complex data analysis *alone* almost certainly exceeded $4 billion in 2016. The leading tools for complex data analysis are SAS from the SAS Institute, SPSS from IBM, MATLAB from MathWorks, and Mathematica from Wolfram Research. The privately-held SAS Institute claims $3.2 billion in annual revenues in 2016 on its web site. Sales for IBM's SPSS are somewhat difficult to determine, but probably fall somewhere in the range of $300 million to $1 billion. SPSS Inc. reported annual revenues of about $280 million[1] just before IBM acquired it for $1.2 billion in 2009[2]. Privately-held MathWorks reports $850 million in annual revenues for 2016 on its web site. Wolfram Research has about 700 employees[3] and probably has revenues in the range of $70 to $90 million. There are many other tools from smaller firms and organizations, notably including the free open-source *R* programming language from the R Foundation and a large collection of numerical and scientific add-ons to the Python programming language (NumPy from www.numpy.org, SciPy from www.scipy.org, and many others).



Far more money is spent annually on the salaries and overhead for analysts, scientists, and engineers using these data analysis tools. Typical salaries with overhead for analysts range from $100,000 to $200,00 per year. A single user license for market leader SAS is about $10,000, suggesting that total annual salaries and overhead for SAS analysts are in the range of $32 billion and $64 billion. IBM SPSS is difficult to estimate. A single user license for MATLAB is about $2,000, suggesting that total annual salaries and overhead for MATLAB analysts are in the range of $40 to $80 billion. A single user license for Mathematica is about $2,000, suggesting that total annual salaries and overhead for Mathematica analysts are in the range of $2 to $7 billion.

It is difficult to evaluate the amount of money spent on salaries and overhead for analysts, scientists, and engineers using the *R* programming language or *Python/NumPy/SciPy*. Both are very popular in the burgeoning "data science" field. It is common to see presentations based on *R* or *Python/NumPy/SciPy* at data science meetups and conferences. Nonetheless a search of all job advertisements in the United States posted on the popular LinkedIn Job site on May 31, 2017 turned up only 388 posts mentioning NumPy explicitly. *R* is a single letter and matches, for example, the R in "R and D," a very popular phrase.

Job Search on LinkedIn on May 31, 2017 in USA

Data Analysis Tool	**Number of Hits**
SAS	12,526
SPSS	2,857
MATLAB	7,532
Mathematica	288

Data Analysis Tool	Number of Hits
NumPy	388

The proliferation of data from cheap sensors, widespread high bandwidth wired and wireless networks, huge disk drives, and instrumentation of web sites and smartphone apps – a surveillance economy – has fueled an explosion of complex data analysis in the last few years.

What are the data analysis tools?

All of the current state-of-the-art commercial data analysis tools are quite similar. They are all interpreted scripting languages with special support for number crunching and extensive libraries of statistical tests, statistical methods, and mathematical methods. All contain a list, array, or matrix data structure for handling and processing large amounts of numerical or mathematical data. A few, notably *Mathematica*, contain a computer algebra system (CAS) for partially automating algebraic manipulations and calculus.

SAS, SPSS, and MATLAB were all developed in the 1970s. *Mathematica* in the 1980s. The *R* programming language is a free open-source implementation of the *S* statistical programming language developed at Bell Labs in the 1970s.

The tools automate lengthy and tedious numerical computations that used to be done by hand with pen and paper and sometimes with adding machines. The Manhattan Project and the space program employed armies of human “computers” to perform these calculations prior to the widespread availability of mainframe computers in the 1960s.

In most cases, the tools require substantial custom programming to produce useful results, hence the large number of jobs for analysts. In addition to the SAS, SPSS, MATLAB, *Mathematica*, or other programming language, the analyst typically must have a good knowledge of several areas of mathematics and statistics to use the tools successfully. Generally, at least familiarity with the linear algebra and statistics usually taught in second year mathematics courses at a good university or college is a minimum requirement.

What do analysts do?

Analysts perform a large number of critical tasks that have either not been automated or have proven difficult to automate. These include:

1. Selection and validation of the data
2. Identification of candidate mathematical models for the data
3. Customization of the model fitting process for the data and candidate model or models
4. Evaluation of the “goodness of fit” of the model to the data
5. Finding a better mathematical model if the fit is judged a failure
6. Inventing new mathematics if no known mathematics matches the data
7. Writing a final report
    1. Persuading specialists such as colleagues

    2. Persuading non-specialists such as policy makers, opinion leaders, funding agencies, senior executives, venture capitalists, etc.
8. Integrating the results of the analysis into hardware or software for practical use.

Frequently, the practical goal of the analysis is to predict the future or to engineer a system to do something better or perform a task that has not been possible so far. As will be discussed further below, Galileo's successful mathematical model of gravity had the practical purpose of predicting the trajectory of cannon balls, a matter of no small importance in the war-torn Renaissance, more accurately than the largely non-quantitative Aristotelian theory of motion.

In practice, data analysis, even with modern tools, is slow, expensive, error-prone, and often unconvincing. The case study on Vioxx, the deadly pain-killer, below demonstrates many of these problems.

Historical Examples

Complex data analysis is not new. In fact, it dates back thousands of years to astronomical observations in ancient Sumeria (modern day Iraq) and some methods still in use are very old. Ancient records are fragmentary, sometimes contradictory, and of uncertain reliability, surviving manuscripts being purported copies of copies of copies of... often dating from around 1000 A.D. Nonetheless, there is good reason to think Pythagoras (c. 570 – c. 495 B.C.) brought then very advanced mathematical methods used in astronomy, astrology, and mysticism back to Greece from Babylonia/Sumeria and Egypt in the sixth century B.C. Western astronomy, astrology, physics, and specifically mathematical modeling and statistics traces its roots to these ancient methods, preserved and extended by Plato (c. 428-348 B.C.), his students and followers, and others.

Data selection and validation

In 1600, the Danish astronomer, astrologer, and nobleman Tycho Brahe (1546-1601) had accumulated over a lifetime by far the most accurate measurements of the positions of the planets over time, especially the planet Mars thought by astrologers and kings to influence the occurrence and outcomes of wars and conflict. After years of lavish royal patronage in Denmark, Tycho had a falling out with the new king and fled to the mostly German-speaking Holy Roman Empire of Rudolf II (1552-1612). Here with funding from Rudolf II he hoped to analyze his data and confirm his own novel theory of the solar system, the known universe at the time, in which the Earth was the center with the Sun and Moon orbiting the Earth and all the other planets orbiting the Sun. He hired the brilliant young up-and-coming astronomer and mathematician Johannes Kepler (1571-1630) to analyze his data. Kepler hoped to use Tycho's data to confirm *his* own Theory of Everything based on the hot new Sun-centered theory of Nicolaus Copernicus (1473-1543). Tycho and Kepler had a stormy working relationship until Tycho's untimely death in 1601 which left Kepler with the access to Tycho's data that he desired.

Tycho Brahe and Johannes Kepler

One of the first tasks that consumed Kepler was to assess Tycho's data and determine which numbers to trust and use. Not a trivial undertaking. Tycho's data had been collected over many years in Denmark and Prague by Tycho and various assistants using the naked eye and giant, extremely precise sextants that enabled measurement of the position of Mars and the other planets to within 1-2 arc minutes (an arc minute is 1/60th of a degree). This was a phenomenal improvement in accuracy over existing measurements.

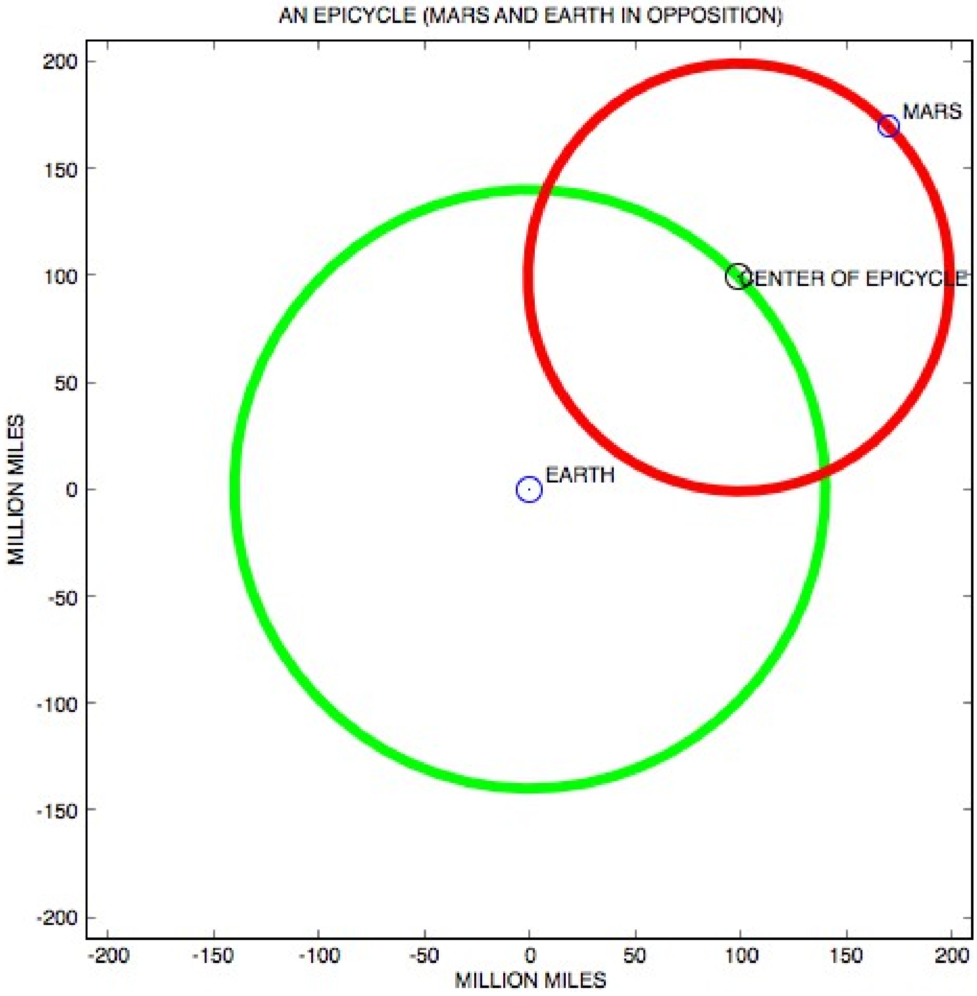
However, human beings make mistakes, transpose digits when writing down numbers, and other errors. In addition, Kepler realized that he had to correct the numbers for the refraction of light in the atmosphere, forcing him to invent the foundations of the modern theory of optics! Eventually, Kepler settled on a set of twelve measurements of the position of Mars, one in each sign of the Zodiac over many years – *tiny data.*

Finding a mathematical model

Kepler had three mathematical models to choose from when he started his analysis. All three were extremely complex and all three turned out to be wrong. The first was the Earth-centered solar system of Klaudius Ptolemy (c. 100 – c. 170 A.D.) as elaborated by subsequent astronomers and astrologers. The second was the hot new Sun-centered solar system of Nicolaus Copernicus (1473-1543), which Kepler hoped to prove as his own theory was built on top of the Copernican model. The third was the hybrid-model of Tycho in which the Sun and Moon orbited a stationary Earth, but all the other planets orbited the Sun.

All three of the models assumed that the motion of the planets was constructed of uniform circular motion and incorporated *epicycles* to patch the otherwise obvious problems with the models. Ordinary observation of Mars with the naked eye – no expensive sextants or telescopes required – had long shown that Mars advances through the Zodiac but about every two years *backs up* for a month or two and then resumes its forward motion. This same retrograde motion is also observed with Jupiter and Saturn. This is grossly inconsistent with uniform circular motion or even non-uniform circular motion in one direction around the Earth.

The ancients, probably in Sumeria, solved the mystery of retrograde motion by proposing that Mars moves in uniform circular motion around a point which then in turn moves in uniform circular motion around the Earth. This *epicycle* reproduces the correct general behavior of Mars. It is however wrong in detail, predicting the position of Mars incorrectly by several degrees. Thus, over the millennia astronomers added more and more epicycles, epicycles on top of epicycles, to the model. Nonetheless even with hundreds of epicycles, the Ptolemaic model was only accurate to about one percent (3.6 degrees).



In fact, as Kepler eventually discovered, Mars follows an elliptical orbit with the Sun at one focus of the ellipse. Because of this, Copernicus could not reproduce the observed motions of Mars by assuming uniform circular motion around the Sun. Accordingly, he added epicycles to his model to reproduce the observed data. Tycho Brahe, who also assumed uniform circular motion, also had to add epicycles to his hybrid model to fit the data.

Kepler tried and tried for years to get any of the models, especially his favorite the Copernican model, to match Tycho's data. He added epicycles. He tried different parameters for the epicycles. None worked. He was eventually able to show that all three models could be made to make the same incorrect predictions with the proper choice of epicycles. The models were *mathematically equivalent* although their physical meaning and interpretation differed.

Is the data bad?

At this point in an analysis, an analyst needs to ask the hard question: *is something wrong with the data?* Could there be sampling bias, measurement error, misinterpretation, fraud? Kepler however was sure from his extensive study of Tycho's data that his chosen measurements were good.

Finding new math

In 1605 Kepler was on the verge of giving up. He was stumped, deeply frustrated. Then, he took a break and while on holiday on Easter weekend in 1605, the answer hit him. What if Mars was following an elliptical orbit and not moving at a uniform speed at all? In 1605, the ellipse was known mathematics but extremely advanced, understood by few even Kepler. Kepler had to gain access to a copy of *Conics* by Apollonius of Perga (c. 240-190 B.C.) which contained the proper math expressed in the crude notation and diagrams of ancient Greek geometry.

Today we have graph paper, analytic geometry, algebraic notation, and many other intellectual tools which make the ellipse an easy subject for high school or college geometry classes. Not so in Kepler's time. For Kepler and his colleagues, working with the ellipse was as challenging as advanced number theory or super-strings or quantum field theory is today.

Kepler was lucky that as complicated as the motion of Mars and the other planets seemed to be 1605, it could be explained by the known mathematics of the time. Kepler soon worked out what was happening. Mars was following an elliptical orbit with the Sun at one focus of the ellipse at a varying speed such that Mars swept out the same area in the same time, moving faster near the Sun and slower farther away from the Sun. These conclusions are now known as Kepler's First and Second Laws.

Kepler did not express his discovery in the simple, clear, concise modern way used above. There is a small chance that he did not realize that his mathematics implied that the planet was sweeping out the same area in the same time.

Presenting the final results to a skeptical audience

Kepler wrote up his results in a book *Astronomia Nova (New Astronomy)*, around six hundred pages of Latin filled with computations and expositions in the ancient Greek geometric style published in 1609. Very few people then or now have read the entire book. The part many people read was the introduction, which summarized his results in clear language and was widely translated into other languages. Kepler was careful to address the apparent contradiction between his conclusions and the literal language of the Old Testament cautiously and diplomatically. In modern language, *Astronomia Nova's* introduction was the *Executive Summary* – key to persuading patrons like Rudolf II and other interested parties.

Inventing new math

Kepler was fortunate that known math explained the motion of Mars and the other planets. Otherwise, he probably would have failed and at best would have needed years to invent new mathematics that did explain the data. In fact, sometimes the analysts, scientists, and engineers have to invent new mathematics to explain their data. James Clerk Maxwell (1831-1879) invented the system of differential equations that bear his name to explain Michael Faraday (1791-1867)'s data on electricity and magnetism. Albert Einstein (1879-1955) and Marcel Grossman (1878-1936) invented the system of differential equations that constitute the General Theory of Relativity (Einstein's famous theory of gravitation), a difficult task as they lacked expertise in the field of differential geometry which they adapted to the theory of gravity. Inventing new mathematics often takes many years and remains largely a human activity.

A Simpler Example: a Ball Dropped from a Roof

Even today with modern tools like SAS, MATLAB or the *R* programming language, Kepler's data analysis would be quite difficult and take a long time, perhaps even as long as Kepler took. All of the models were quite complex with many epicycles and many adjustable parameters from each epicycle that would be computed by a modern curve of function fitting program. Modern tools like SAS and MATLAB are not able to automate Kepler's critical insight that the motion of Mars was elliptical. **The human analyst must still recognize the mathematics.**

To get a better sense of what a modern analysis is like, let's look at a much simpler analysis similar to one performed by Kepler's contemporary Galileo Galilei (1564-1642). The modern analysis is done with MATLAB. This is constructing a mathematical model of a ball dropping off a roof. There is an incorrect legend that Galileo dropped two balls off the Leaning Tower of Pisa, one heavy and one light, to prove his theory. This did not happen and, in fact, the heavier ball would have hit the ground slightly sooner than the light ball due to air resistance, arguably confirming the Aristotelian theory of motion. What, in fact, Galileo did was to roll balls down inclined panes indoors where air resistance and turbulence were not an issue and the fall was slowed by the incline so that he could measure the time with the crude water clocks he had.

A modern equivalent would be to take a video of a ball dropped from the roof of a building. This is governed by a simple formula where the distance of the ball below the roof is proportional to the square of the time since the ball was dropped. The proportionality constant, the Earth's gravitation acceleration constant, g is about 32 feet/second/second (United States) or 9.8 meters/second/second (everywhere else). The gravitational constant varies slightly from place to place on Earth.

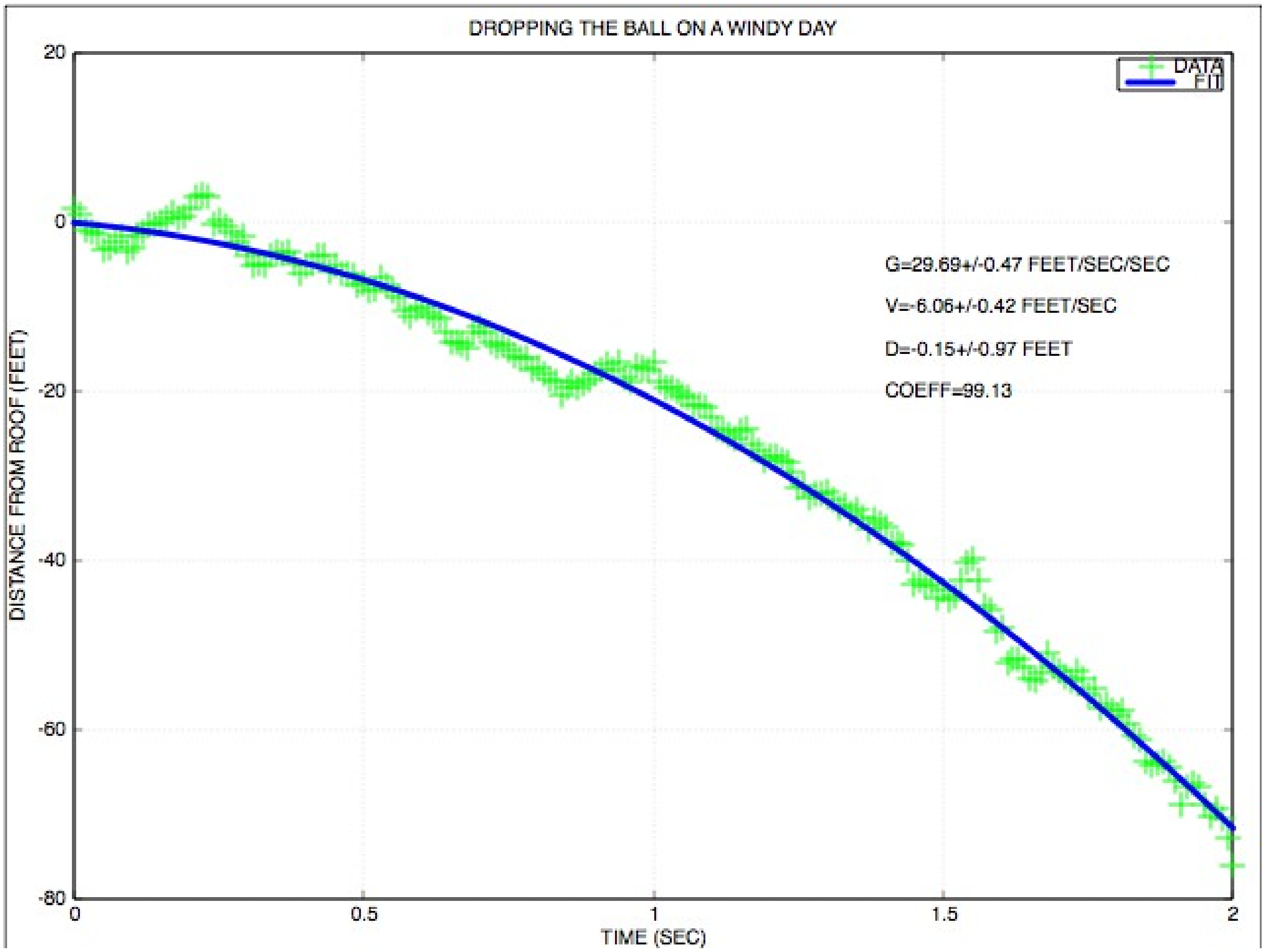
The Mathematics of Dropping a ball

$$X = -(1/2)\, g\, t^2 + E_p + E_m$$

Distance from roof	X
Elapsed time	t
Gravitational Acceleration	g *(32 feet/sec/sec)*
Process error (e.g. wind)	E_p
Measurement error (e.g. camera jitter)	E_m

Note: Errors are *often* assumed to have a Gaussian/Normal/Bell Curve distribution.

In practice, there are process errors, the ball actually moves around in the wind, and measurement errors, jitter of the camera. These are often assumed to be Gaussian/Normal/Bell Curve distributed. This assumption can be wrong, sometimes with serious consequences.

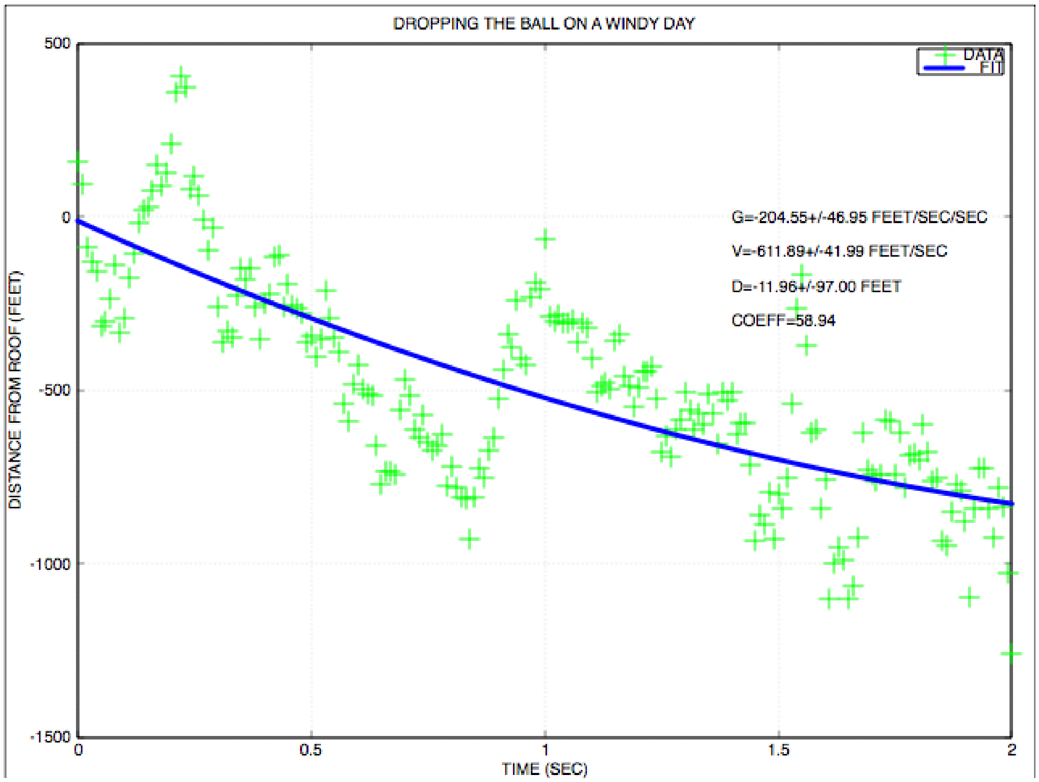


It is fairly simple to fit this model to data in MATLAB with a few lines of code (Kepler's models would take hundreds, possibly thousands of lines of code).

This fit is pretty good but off by about one percent. The *coefficient of determination*, a measure of how well the model explains the variation in the data, is 99.13 percent; 100 percent would be perfect. In many cases, an analyst would say “good enough” and accept the results and move on. Kepler and his contemporaries were dissatisfied with this level of performance in the Ptolemaic and other models of the planetary motions. It is a human value judgement how good is “good enough.”

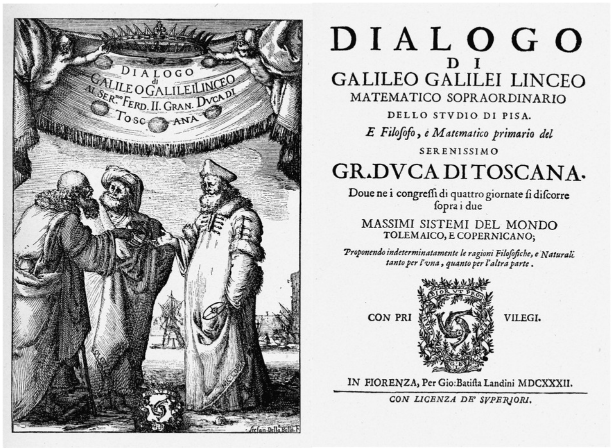
The model is imperfect because we do not have model for the wind and turbulence in the air. Turbulence is difficult to understand even today. The solution to the Navier-Stokes equation thought to describe turbulence in the air, water and other fluids remains an open problem in higher mathematics.

In some cases, the model and data clearly disagree. Below is a plot of dropping a ball in a cyclone. The coefficient of determination is only 58 percent and the disagreement between the data and the model is clearly visible in the plot. The ball goes all over the place and gravity no longer provides an adequate explanation of the motion. As discussed, then the human analyst must find a new mathematical model or even invent completely new mathematics.



Fortunately, Galileo's theory of gravitation matched his experiments with inclined planes quite well. He was not forced to hunt through the mathematical literature like Kepler or even more difficult invent entirely new mathematics.

Losing your audience


DIALOGO
di
GALILEO GALILEI LINCEO
AL SER.mo FERD. II. GRAN. DVCA DI
TOSCANA

DIALOGO
DI
GALILEO GALILEI LINCEO
MATEMATICO SOPRAORDINARIO
DELLO STVDIO DI PISA.
E Filoſofo, e Matematico primario del
SERENISSIMO
GR.DVCA DI TOSCANA.
Doue ne i congreſſi di quattro giornate ſi diſcorre
ſopra i due
MASSIMI SISTEMI DEL MONDO
TOLEMAICO, E COPERNICANO;
Proponendo indeterminatamente le ragioni Filoſofiche, e Naturali
tanto per l'vna, quanto per l'altra parte.

CON PRI VILEGI.

IN FIORENZA, Per Gio:Batiſta Landini MDCXXXII.

CON LICENZA DE' SVPERIORI.

Like Kepler before him and analysts today, Galileo had to persuade others of his analysis and theories. In this he failed spectacularly. The illustration above is the frontispiece and title page of Galileo's famous *Dialogue Concerning the Two Chief Systems* (1632) in which he insults the Pope, attributing the Pope's Aristotelian views to a character diplomatically named *Simplicio*, meaning “Simpleton,” and propounds a grossly inaccurate theory to explain the tides, attributing the tides to sloshing motion from the spinning of the Earth and discarding thousands of years of observations that the tides follow the Moon and the Sun – the tides are highest when the Moon, the Sun, and the Earth are located along a line.

Galileo, Kepler, and the other advocates of the new Copernican theory were confronted by a physical question. Their critics conceded that the mathematics worked; Kepler had demonstrated this. However, the theory seemed to require that the Earth was physically spinning at thousands of miles per hour and flying through space at thousands of miles per hour. And yet standing on the Earth, no motion was evident. They confronted Galileo with the obvious question: *can you prove the Earth is moving?* This is not easy to show. There is a way but Galileo did not know how to do it, which led him to unwisely embrace the incorrect theory of the tides.

Galileo also had a long history of putting down opponents as idiots which proved disastrous with the Pope. He probably intended to put down only his rivals and did not consider that he might offend the Pope as well.

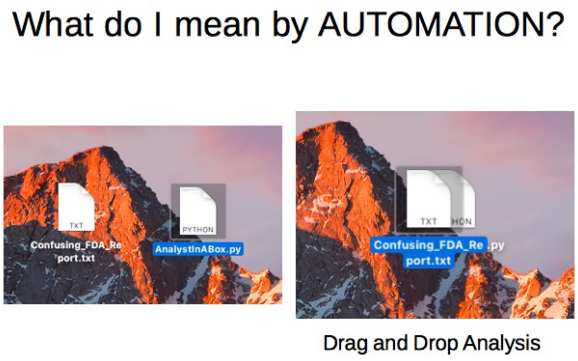
Egos, personalities and caustic often counter-productive put-downs remain a fixture of complex data analyses to the present day. We have excellent tools for automating the numerical computations but many human factors remain the same. **The state of the art in complex data analysis remains slow, expensive, error-prone, and often unconvincing.**

What is meant by Automating Complex Data Analysis?

Ideally, we would like to automate all of the steps currently performed by human beings. This would be pretty close to replicating or even exceeding human intelligence. This is almost certainly many years in the future, if it is even possible. We could envision a fully "drag and drop analysis," where a non-specialist can drag a data file or report onto an *Analyst in a Box* application which would perform the appropriate analysis and generate a clear, readable, convincing put-down free report of the results. The *Analyst in a Box* would systematically apply the best practices in statistics and data analysis at every stage, generate a detailed log of every step and the reasons for each choice, and source code in a free open-source language such as *R* that would enable anyone to quickly reproduce the analysis. The *Analyst in a Box* would bundle the final report, log of analysis steps, source code, and data in a single archive (such as a ZIP file) with identifying hash codes (e.g. SHA) and time stamps.

Realistically, this is almost certainly not fully possible in the immediate future. We can automate many best practices such as automatically performing the relevant statistical power[i] calculations – a major issue in the Vioxx case study below – and reporting the results in a clear way accessible to general audiences such as doctors and patients. With modern pattern recognition, computer vision, and artificial intelligence techniques we can automate some, perhaps all, of the currently arduous task of identifying candidate mathematical models that resemble the data and are good candidates for model fitting and evaluation.

Some tasks will still require the domain expertise and conceptual understanding of human beings. Early automated data analysis systems will have Turbo Tax-like interfaces where the *Analyst in a Box* will pop-up intermediate results and ask for human input from experts. In the case of FDA approvals, questions such as: *Have I interpreted this passage as the description of a clinical trial correctly?*



Screenshot of Drag and Drop Analysis

---

[i] The power of a binary hypothesis test is the probability that the test correctly rejects the null hypothesis (H0) when the alternative hypothesis (H1) is true. It can be equivalently thought of as the probability of accepting the alternative hypothesis (H1) when it is true—that is, the ability of a test to detect an effect, if the effect actually exists. In the Vioxx case study, the null hypothesis would be that Vioxx is as safe as a placebo or comparison drug; the alternative hypothesis is that Vioxx kills more patients than the placebo or comparison drug.

Case Study: Vioxx

Vioxx (generic name *rofecoxib*) was a pain-killer marketed by the giant pharmaceutical company Merck (NYSE:MRK) between May of 1999 when it was approved by the United States Food and Drug Administration (FDA) and September of 2004 when it was withdrawn from the market. Vioxx was marketed as a "super-aspirin," allegedly safer and implicitly more effective than aspirin and much more expensive, primarily to elderly patients with arthritis or other chronic pain. Vioxx was a "blockbuster" drug with sales peaking at about $2.5 billion in 2003[4] and about 20 million users[ii]. Vioxx probably killed between 20,000 and 100,000 patients between 1999 and 2004[5].

Faulty blood clotting is thought to be the main cause of most heart attacks and strokes. Unlike aspirin, which lowers the probability of blood coagulation (clotting) and therefore heart attacks and strokes, Vioxx increased the probability of blood clotting and the probability of strokes and heart attacks by about two to five times.

FDA drug approvals require randomized clinical trials and detailed sophisticated statistical analyses of the results to demonstrate both the safety and the efficacy of the approved drug. SAS dominates the statistical analysis of clinical trials although it is not technically required by law[6]. Most analyses of clinical trials use SAS.

Vioxx was a horrific example of a complex data analysis gone awry. It is not unique. At least thirty-five prescription drugs approved by the FDA have been withdrawn due to major safety concerns, generally killing patients[7]. Serious safety problems have been discovered after approval in many drugs that have not been withdrawn. Generally, the FDA places restrictions on the use of the drug and requires a "black box warning" on the final printed label for the drug.

The FDA has instituted an FDA Adverse Events Reporting System (FDAERS) for doctors and other medical professionals to report deaths and serious health problems such as hospitalization suspected of being caused by adverse reactions to drugs. In 2014, 123,927 deaths were reported to the FDAERS and 807,270 serious health problems. Of course, suspicion is not proof and a report does not necessarily mean the reported drug was the cause of the adverse event.

Aspirin, ibuprofen (the active ingredient in *Advil* and *Motrin*), naproxen (the active ingredient in *Aleve*), and many other alternative "aspirins" are known as NSAIDs (Non Steroidal Anti-Inflammatory Drugs). The NSAIDs all inhibit an enzyme known as cyclooxygenase (COX). COX indirectly affects blood coagulation and the gastrointestinal system. Aspirin is known to act as a "blood thinner," reducing blood clotting and reducing the chance of heart attacks and strokes by about 25% at certain low doses. However, aspirin and other NSAIDs can aggravate the stomach with repeated use, sometimes causing ulcers, bleeding, and even death from uncontrolled bleeding.

In the 1990's, studies of arthritis patients at Stanford led to estimates that about 16,500 patients were dying from aspirin and other NSAID induced bleeding each year[8]. This "silent epidemic" received considerable press both in the scientific and medical literature[9] and in popular media like the *Los*

[ii] A "blockbuster" drug is pharmaceutical industry jargon for a drug with at least $1 billion in annual sales. Like Vioxx, it need not be a "wonder drug" that cures or treats a fatal or very serious disease or condition.

Angeles Times[10]. Some physicians at the time expressed skepticism that the bleeding episodes and deaths were this common[11,12].

About 42.7 million Americans had arthritis in 1999[13]. Loosely, the 16,500 deaths number meant that 3.8 in 10,000 (0.038 %) arthritis patients would die each year from aspirin and other NSAIDs. As a point of reference, in 1999 about 1.526 in 10,000 (0.01526) of Americans died in car accidents[14]. The risk of death in car accidents is somewhat lower today (2017) according to official figures. To beat aspirin and other NSAIDs, a new drug would need to beat the 3.8 deaths per 10,000 per year rate (the 16,500 deaths per year number divided by 42.7 million Americans with arthritis).

At about the same time in the 1990s, it was discovered that there were two variants of COX, COX-1 and COX-2. Supposedly, only COX-1 was connected to the gastrointestinal system. Thus, in theory, a drug that inhibited only COX-2 would have all the benefits of aspirin and little or no stomach irritation or death. These new drugs were called COX-2 inhibitors. Pfizer's Celebrex, still on the market, and Merck's Vioxx were the two leading COX-2 inhibitors.

However, from the very beginning, before Vioxx and Celebrex were approved by the FDA, there were theoretical reasons to fear that the COX-2 inhibitors, unlike aspirin, would *increase* blood coagulation function, making heart attacks and strokes more probable in patients. This was due to a complex balancing of push and pull effects on blood coagulation by COX-1 and COX-2. If both forms of COX were inhibited as in aspirin, blood coagulation dropped overall, but if only COX-2 was inhibited, the uninhibited COX-1 boosted blood coagulation function without a larger compensating reduction from COX-2.

Remarkably, Merck proposed and the FDA approved Phase III clinical trials of Vioxx with too few patients to show that Vioxx was actually safer than the putative 3.8 deaths per 10,000 patients rate (16,500 deaths per year) from aspirin and other NSAIDs.

The FDA guideline, *Guideline for Industry: The Extent of Population Exposure to Assess Clinical Safety: For Drugs Intended for Long-Term Treatment of Non-Life-Threatening Conditions* (March 1995), only required enough patients in the clinical trials to reliably detect a risk of about 0.5 percent (50 deaths per 10,000) of death in patients treated for *six months or less* (roughly equivalent to one percent death rate for *one year* assuming a constant risk level) and about 3 percent (300 deaths per 10,000) for one year (recommending about 1,500 patients for six months or less and about 100 patients for at least one year *without* supporting statistical power computations and assumptions in the guideline document).

The implicit death rate detection threshold in the FDA guideline was well *above* the risk from aspirin and other NSAIDs and at the upper end of the rate of cardiovascular “events” caused by Vioxx. FDA did not tighten these requirements for Vioxx even though the only good reason for the drug was improved safety compared to aspirin and other NSAIDs. In general, the randomized clinical trials required by the FDA for drug approval have too few patients – insufficient *statistical power* in statistics terminology – to detect these rare but deadly events[15].

The relevant section of the original final printed label from the FDA Approval of Vioxx in May 1999 reads:

ADVERSE REACTIONS

Osteoarthritis

Approximately 3600 patients with osteoarthritis were treated with VIOXX: approximately 1400 patients received VIOXX for 6 months or longer and approximately 800 patients for one year or longer. The following table of adverse experiences lists all adverse events regardless of causality, occuring in at least 2% of patients receiving VIOXX in nine controlled studies of 6 weeks to 6 months duration conducted in patients with OA at the therapeutically recommended doses of (12.5 and 25 mg), which included a placebo and/or positive control group.

As will be elaborated further below, 3600 patients is too small a sample to reliably detect the increased risk of death from Vioxx – well above the risk of death from aspirin – at the 95 percent confidence level used by the FDA. When I have given talks about the Vioxx tragedy, some audience members are incredulous that the FDA allowed this. Indeed I was incredulous going through the FDA approval documents. Nonetheless this is what the FDA did.

In some respects, the issue is quite simple. A clinical trial with one-hundred (100) patients taking Vioxx compared to one-hundred patients taking a placebo or comparison drug (e.g. the pain-killer naproxen) is simply unable to reliably detect an increased risk of death of 0.1 percent per year (an average rate of 1 patient death per 1000 patients) in the patients taking Vioxx. It clearly takes at least one-thousand (1000) patients to detect such a small but deadly effect.

What happens if the risk of death with Vioxx is 0.2 percent per year (2 patient deaths per 1000 treated patients) and 0.1 percent per year (1 patient death per 1000 patients) on a placebo or comparison drug like naproxen? This is similar but not identical to the actual situation with Vioxx and naproxen (from the VIGOR study discussed below). One hundred patients is clearly too few to discriminate between Vioxx and the placebo.

What happens with one-thousand (1000) patients? In this case, *on average* two patients will die in the Vioxx group and one patient will die in the control group. The number of deaths is a random process *like flipping a coin.* Zero, one, two or more patients could die in both groups. It is quite possible that a clinical trial with only one-thousand (1000) patients could give zero (0) deaths on Vioxx and two (2) deaths on the placebo, incorrectly showing that Vioxx is safer than the placebo. To be reasonably confident, for example ninety-five (95) percent confident, that an excess number of deaths in the Vioxx group compared to the control group is not due to chance, more than one-thousand patients are needed.

The exact required number of patients in the clinical trial is the result of a *statistical power* computation. This number depends in detail on the size of the effect being compared in the Vioxx and control groups *and* the desired confidence level that the result of the clinical trial is correct. For an adverse effect at a level of one in *N* (some number such as 1000), the required number of patients in the clinical trial is typically several times *N*. In general, several thousand patients are required to detect an increase in the risk of death of 0.1 percent (1 in 1000 patients) with a ninety-five percent confidence level (used by the FDA). In the example, with 10,000 patients in the clinical trial, *on average* there would be 10 deaths in the control group and 20 deaths in the Vioxx group. This excess of ten (10) deaths in the Vioxx group has slightly less than a five percent chance of being due to random chance.

The original Phase III clinical trials used for the approval of Vioxx in 1999 only had eight-hundred (800) patients who took the drug for more than one year. Since the lethal side effect of Vioxx appears to have been cumulative and associated with prolonged use, eight-hundred was far too few to detect the small but deadly effect. The original clinical trials had a total of 3600 patients, most of whom took the drug for less than one year. Even if all of these had taken the drug for one year, there was about a one in four (25 percent) chance that the clinical trials would not have detected the increased risk of death with Vioxx at the 95 percent confidence level used by the FDA.

BAD EVENT	**ANNUAL PERCENT DEATH RATE**	**DEATHS PER 10,000 PER YEAR**	**ODDS (PER YEAR)**	**NUMBER AT RISK**	**DEAD (PER YEAR)**
Reference Level	0.01%	1 in 10,000	1 in 10,000		
Die in Car Crash (1999)	0.01526%	1.526 in 10,0000	1 in 6,553	272.7 Million in 1999	41,611
Die from NSAID (Singh 1998)	0.038%	3.8 in 10,000	1 in 2,631	42.7 Million with Arthritis in 1999	16,500
Reference Level (about Vioxx best case)	0.1%	10 in 10,000	1 in 1,000	20 Million taking Vioxx	20,000
Reference Level (about Vioxx worst case)	0.5%	50 in 10,000	5 in 1,000	20 Million taking Vioxx	100,000
FDA Guideline detection level	1.0 % (or more)	100 in 10,000	10 in 1,000	20 Million taking Vioxx	200,000

The final printed label required by the FDA then and now does not clearly explain the safety level detectable in the clinical trials, only stating the number of patients in the clinical trial in the fine print. ***Most doctors and patients lacked and still lack the statistics knowledge to interpret the numbers in the final printed label*** – although the FDA could easily require that the safety level detectable in the clinical trial be prominently and clearly stated in the label.

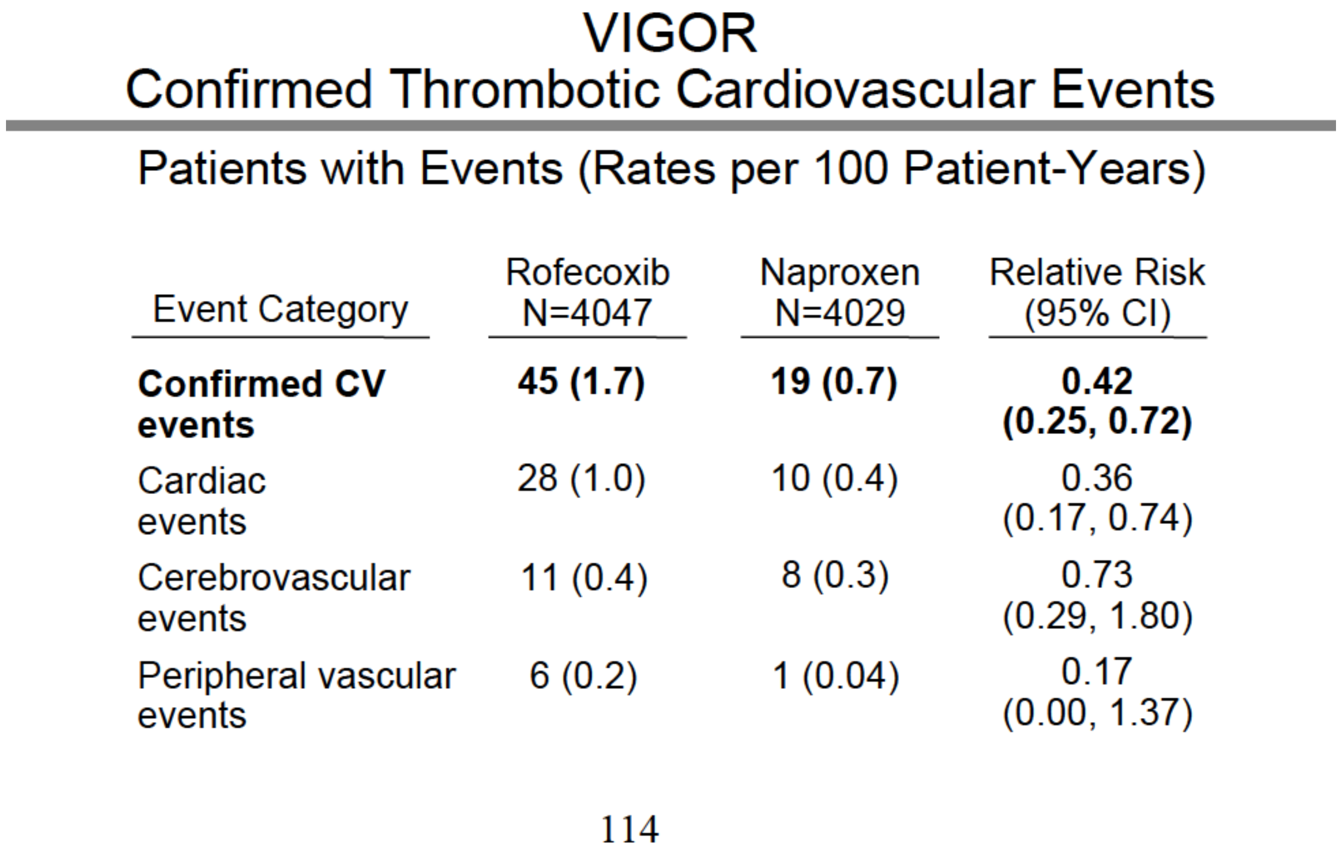
Merck was a very credible company in 1999 with many successful drugs, few safety or legal problems prior to Vioxx, a long history of high-minded rhetoric, extensive charitable activities, a reputation similar to Google at the height of its “do no evil” period in the business, scientific, and medical communities and with the general public.

Merck mounted a highly successful marketing campaign centered on the 2000 Olympics and endorsements from Dorothy Hamill, the 1976 gold-medal Olympic figure-skating champion, and Bruce Jenner, the 1976 gold-medal Olympic running champion. Dorothy Hamill claimed to have been effectively disabled by arthritis, a common problem in figure skaters, and to have been restored by taking Vioxx[16]. Her claimed experience went well beyond the benefits demonstrated in the Phase III clinical trials. Merck ran highly successful commercials for Vioxx showing Dorothy Hamill skating. Merck did internal marketing studies showing that for every dollar spent on the Dorothy Hamill commercials they saw a four dollar increase in Vioxx prescriptions[17].

Dorothy Hamill may very well have had a good and unusually positive experience with Vioxx. As a woman and only forty-four in 2000, she was at low risk for heart attack or strokes, the primary often lethal "side effects" of Vioxx. Olympians and other elite athletes are very unrepresentative of the general population, being usually in generally excellent health but suffering from unusual sports-related problems that can appear at early ages and that may respond differently to drugs than common forms in the general population. Figure skaters have unusually high rates of hip problems including arthritis[18].

Sales of Vioxx soared even after the results of the VIGOR study, published in November 2000, with a larger sample of 4047 patients treated for eighteen months showed four (4) to five (5) times more heart attacks (20 heart attacks – the study in the *New England Journal of Medicine* initially implicitly included only 17 of the 20 heart attacks[19]) in the patients treated with Vioxx versus the control group of 4029 patients treated with naproxen (4 or 5 heart attacks)[20].

Unfortunately, the exact counts of deaths and other adverse events in the VIGOR study jump around by a few patients depending on which source or report is consulted. Bonnie Goldmann M.D., Regulatory Affairs, Merck Research Laboratories gave a detailed presentation to the FDA Advisory Committee on the VIGOR study results on February 1, 2001 with the following numbers for *cardiovascular events* (which includes heart attacks) on slide number 114 (the *relative risk* refers to the risk of *naproxen* relative to *rofecoxib/Vioxx – Merck* interpreted the alarming results as showing that *naproxen* was *less risky* than *rofecoxib)*:



VIGOR
Confirmed Thrombotic Cardiovascular Events

Patients with Events (Rates per 100 Patient-Years)

Event Category	Rofecoxib N=4047	Naproxen N=4029	Relative Risk (95% CI)
Confirmed CV events	**45 (1.7)**	**19 (0.7)**	**0.42 (0.25, 0.72)**
Cardiac events	28 (1.0)	10 (0.4)	0.36 (0.17, 0.74)
Cerebrovascular events	11 (0.4)	8 (0.3)	0.73 (0.29, 1.80)
Peripheral vascular events	6 (0.2)	1 (0.04)	0.17 (0.00, 1.37)

114

Remarkably this four to five times (4-5X) higher heart attack rate was statistically significant at the 95 percent confidence level required by the FDA. How did Merck get around this seeming red flag? Merck argued that naproxen had a hitherto undocumented cardio-protective effect similar to aspirin *but much stronger.* Hence, Vioxx was not killing more patients but rather naproxen was saving many more patients from heart attacks than aspirin.

Deaths, Hospitalizations, and Heart Attacks in the VIGOR Study (November 2000)[21]

Bad Event	**Vioxx**	**Naproxen**
Overall Deaths	22	15
Hospitalizations	338	265
Hospitalizations from heart problems	65	24
Heart attacks[22]	20	4 or 5 (depends on source)
Heart attack deaths	9	5

At this point, medical doctors, scientists, FDA officials, politicians and hedge funds began to cry foul. Major players included Dr. David Graham of the FDA Office of Drug Safety (ODS), U.S. Senator Chuck Grassley (R, Iowa), and noted cardiologist Eric Topol, who was also a paid advisor to a hedge fund (Great Point Partners LLC) that shorted Merck's stock[23,24].

Conflicts of interest are almost inevitable with complex data analyses using current (2017) state-of-the-art tools. Although some of the tools such as the *R* programming language are free, the analyses are complex, time-consuming and require a skilled analyst with substantial knowledge of mathematics and statistics. Frequently only vested interests with large amounts of money, political power, or other important factors at stake can afford to fund an analysis: pharmaceutical companies, trial lawyers suing the pharmaceutical companies, politicians and government officials with uncertain agendas, and sometimes political activists with deep pockets.

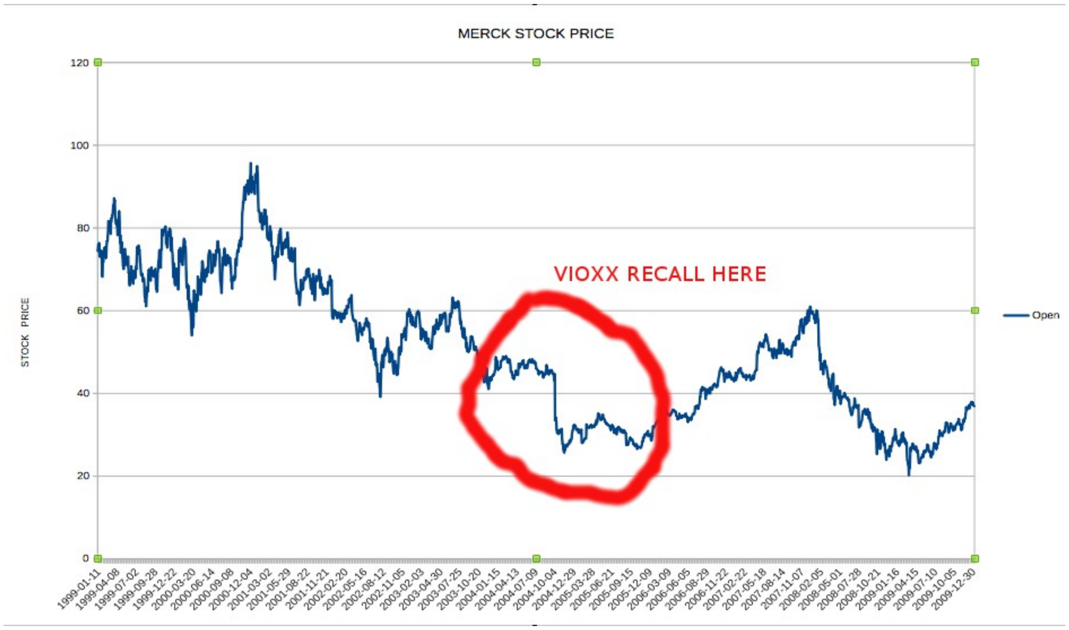
Most medical doctors and patients lack the requisite statistical skills to analyze the FDA approval documents and the scientific/medical journal articles related to a drug. Even where they do, they often lack the free time and resources to perform an independent analysis.

David Graham at the FDA was able to get access to patient data from Kaiser Permanente on patients taking Vioxx (*rofecoxib*) and patients taking Pfizer's Celebrex (*celecoxib*) and demonstrate that the data showed a statistically and practically significant higher death rate for patients on Vioxx compared to Celebrex. It is unclear how safe Celebrex, which remains on the market today, is. The imminent publication[25] of these results probably forced Merck to withdraw the drug in September 2004 although Merck cited the recent results from yet another study, the APPROVe study[26], in their recall announcement.

Bad for Merck

Merck probably lost money on Vioxx. Merck took in about $8-10 billion in sales of Vioxx between May 1999 and September 2004 (extrapolating from peak sales of $2.5 billion in 2003). Merck spent at least $160 million on advertising for Vioxx in 2000 alone[27]. Published articles[28] and media reports claim Merck spent over $100 million per year advertising Vioxx from 1999 to 2004 implying a total advertising budget of over $500 million; this does not include the costs of advertising and public relations to recover from the Vioxx recall.

Merck settled personal injury lawsuits in the United States alone for a total of $4.5 billion[29]. This was touted as a *victory* in the business press in that some analysts estimated Merck could have lost as much as $25 billion in the lawsuits[30]. Merck was sued worldwide by former patients and trial lawyers[31,32,33,34]. The Department of Justice fined Merck $950 million for deceptive marketing of Vioxx[35]. Merck settled a shareholder lawsuit in the United States for $830 million[36]. Merck paid its attorneys at least $1.5 billion[37]. Merck spent at least hundreds of millions on advertising and public relations activities to repair its reputation which remains damaged today. Merck stock dropped sharply on the recall announcement[38], rebounded, and then has dropped again. There were undoubtedly many other miscellaneous expenses largely due to the recall and its aftermath.



How the Vioxx tragedy could have been avoided by automation

Regardless of the actions of Merck or the FDA, doctors and patients using an automated tool that extracted the number of patients in the clinical trials from the final printed label and computed the statistical power of the safety evaluation implicit in the trials could have avoided the drug, warning them that the clinical trial could not reliably detect an increased risk of death from Vioxx at the few tenths of a percent per year level, well above the purported risk from aspirin and other traditional NSAIDs.

FDA approval is not unusual. Research studies in many fields, especially with human subjects, often fail to perform the standard statistical power analyses required by classical statistics or downplay the results. The probable reason is simply that many research studies, especially on human subjects, have small sample sizes and thus are not that reliable – may be wrong simply due to statistical fluctuations in the sample.

Research studies with large numbers of human subjects are generally expensive with current technologies. Phase III clinical trials with a few thousand subjects, for example, often have a total out-of-pocket cost in the range of $50 to $100 million[39]. Phase III clinical trials with the tens of thousands of subjects needed to push the detection threshold for lethal adverse effects below 0.1 percent per year (1 in 1000/10 in 10,000), would probably cost several hundred million, perhaps one billion dollars in some cases. These high costs mean research studies with human subjects often are small and lack the statistical power one would like in an ideal world.

The blood coagulation system is extremely complex and not fully understood. There are many quantitative tests of the blood coagulation system including measurements of how quickly the blood clots when exposed to air or other stimulants, direct measurements of the blood concentration levels of the known blood clotting factors, and many more. Merck could have measured these values for the patients in the clinical trials. This likely would have shown the probability of blood coagulation rising as the patients took Vioxx relative to a placebo or comparison drugs like Naproxen in many or all patients, not just the tiny number who had heart attacks. The effect might have been clear in the smaller, cheaper, earlier Phase II clinical trials.

A more advanced automated analysis tool with *mathematics recognition* capabilities might have been able to extract the underlying mathematics of the blood coagulation system and perhaps enable systematic engineering of the drug to avoid the heart attacks and strokes.

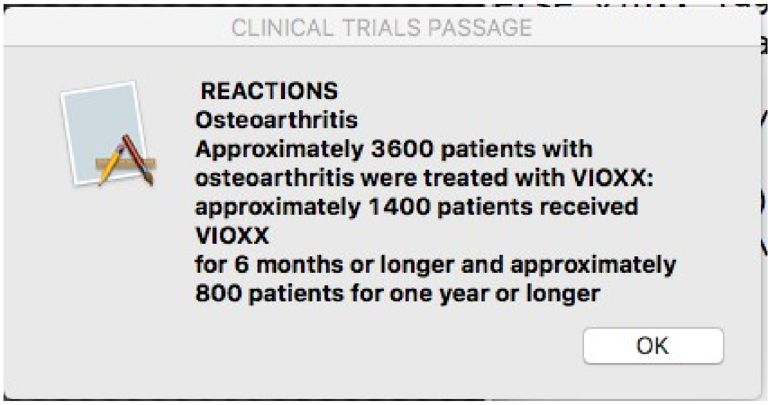
Preliminary Prototype Studies

There are two major technical obstacles to automating complex data analysis. The first involves text comprehension to extract and "understand" key information (*meta-data*) about the numerical data being analyzed. The key information includes the units used such as grams, milligrams, weeks, months, seconds. It includes what is being measured such as grams of water, grams of Vioxx (*rofecoxib*), etc. as well as how the measurements are made in some cases. For example, FDA approved final printed labels often include phrases such as "4000 patients were treated with 25 milligrams of *wonderdrugex* per day for six months." An automatic data analysis tool does not need to understand *all* of the text in a report or data file but it needs to identify phrases like this and "understand" them in order to apply the correct statistical procedures such as a statistical power analysis.

The second major technical obstacle is automatically recognizing candidate mathematical models for the data. This is generally done today by skilled analysts who may recognize that the data resembles a known function (e.g. the ellipse in Kepler's case) or a composition of known functions. This recognition by human analysts is generally a time consuming trial and error process. Once a good candidate model is recognized, this can be fed into a function or curve-fitting function in a tool such as SAS or MATLAB which automates much of the numerical calculations and determination of parameters of the model.

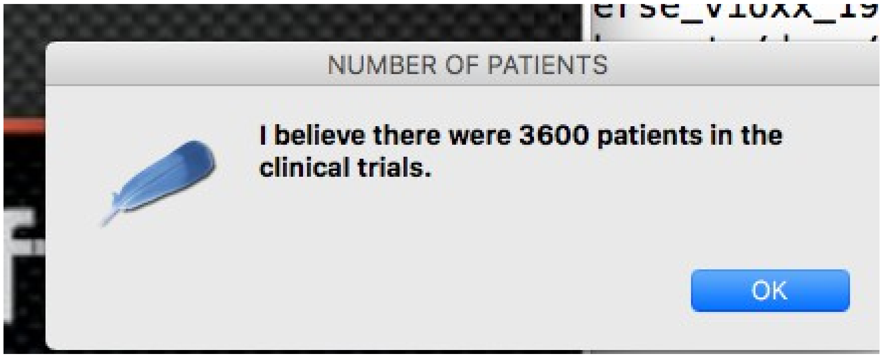
These are a few results of a prototype system:

Analyzing the original 1999 Vioxx final printed label



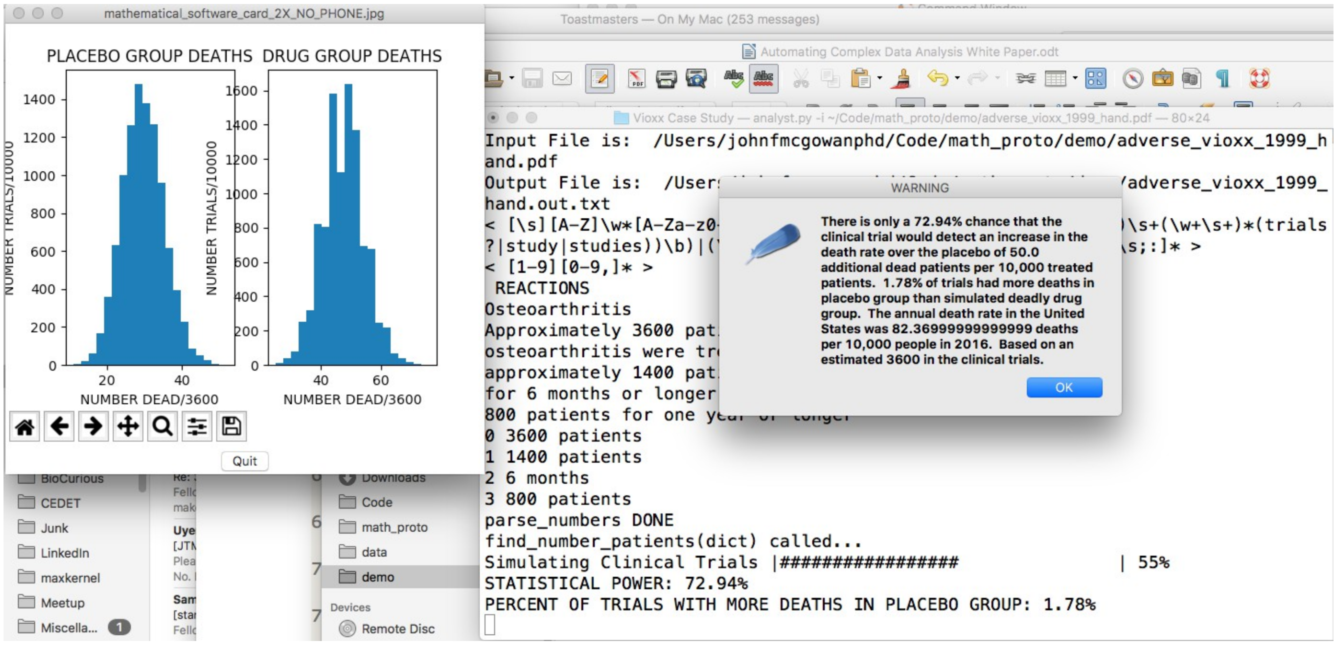
REACTIONS
Osteoarthritis
Approximately 3600 patients with osteoarthritis were treated with VIOXX: approximately 1400 patients received VIOXX
for 6 months or longer and approximately 800 patients for one year or longer

First the prototype scans through the final printed label text to identify passages that state the number of patients in the clinical trial or trials. In a commercial product, the doctor or other operator would be asked to confirm that the passage refers to a clinical trial.



I believe there were 3600 patients in the clinical trials.

Next the prototype extracts the number of patients from the passage. In this simple example, it uses the largest number of patients – what is the best case? Again, in a commercial system, the doctor or other analyst would be asked to confirm the *Analyst in a Box's* interpretation of the passage.



```
Input File is:  /Users/johnfmcgowanphd/Code/math_proto/demo/adverse_vioxx_1999_h
and.pdf
Output File is:  /User                                        /adverse_vioxx_1999_
hand.out.txt
< [\s][A-Z]\w*[A-Za-z0                                        )\s+(\w+\s+)*(trials
?|study|studies))\b)|(                                        \s;:]* >
< [1-9][0-9,]* >
 REACTIONS
Osteoarthritis
Approximately 3600 pat
osteoarthritis were tr
approximately 1400 pat
for 6 months or longer
800 patients for one ye
0 3600 patients
1 1400 patients
2 6 months
3 800 patients
parse_numbers DONE
find_number_patients(dict) called...
Simulating Clinical Trials |#################                  | 55%
STATISTICAL POWER: 72.94%
PERCENT OF TRIALS WITH MORE DEATHS IN PLACEBO GROUP: 1.78%
```

There is only a 72.94% chance that the clinical trial would detect an increase in the death rate over the placebo of 50.0 additional dead patients per 10,000 treated patients. 1.78% of trials had more deaths in placebo group than simulated deadly drug group. The annual death rate in the United States was 82.36999999999999 deaths per 10,000 people in 2016. Based on an estimated 3600 in the clinical trials.

Finally, the prototype performs a statistical power analysis by simulating the results of 10,000 clinical trials with 3600 patients in the drug and placebo groups. The histograms show the number of simulated trials where a certain number of patients (the *x* axis) died for the placebo (control) group and the Vioxx (drug) group.

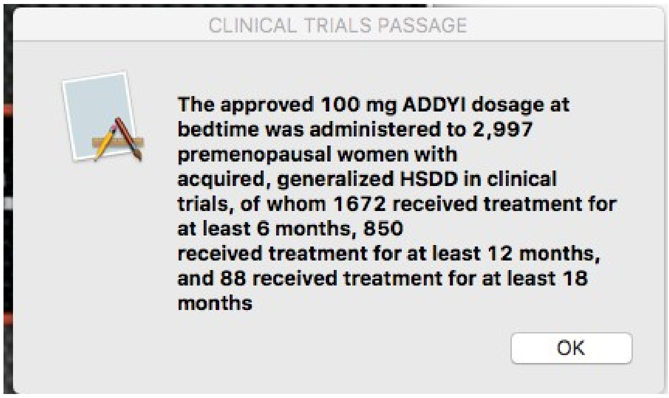
On average, more patients die in the Vioxx (drug) group than the placebo (control) group. However, there is considerable variation. The deaths are like flipping an unfair coin with only a one in one-thousand (1 in 1000/0.1 percent) chance of heads and getting heads.

In a few percent of the simulated trials, more patients die in the placebo group than the drug group. In over twenty-five percent (one in four) simulated trials, the excess of deaths in the drug group is too small to meet the ninety-five percent (95 %) confidence level (about two standard deviations) statistical significance test commonly used.

The prototype generates a warning message because over a quarter of the simulated clinical trials would have failed to detect an increase of 0.4 percent in the risk of death (4 in 1,000) per year compared to the placebo. In other words, despite FDA approval, the drug could easily have a substantial undetected risk of death.

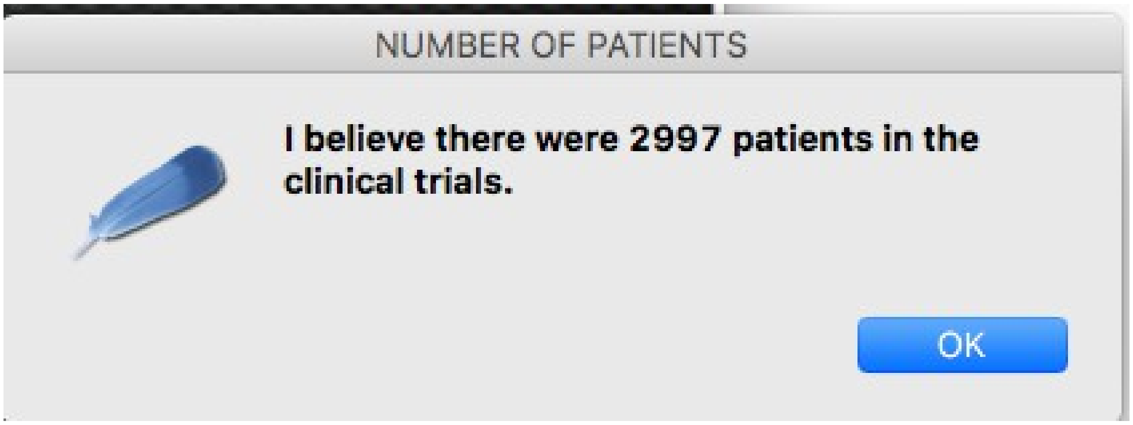
Analyzing the final printed label for ADDYI (approved in 2015)

These are the results of analyzing the FDA final printed label for the controversial drug ADDYI (*flibanserin*), approved in 2015[40,41,42,43]:



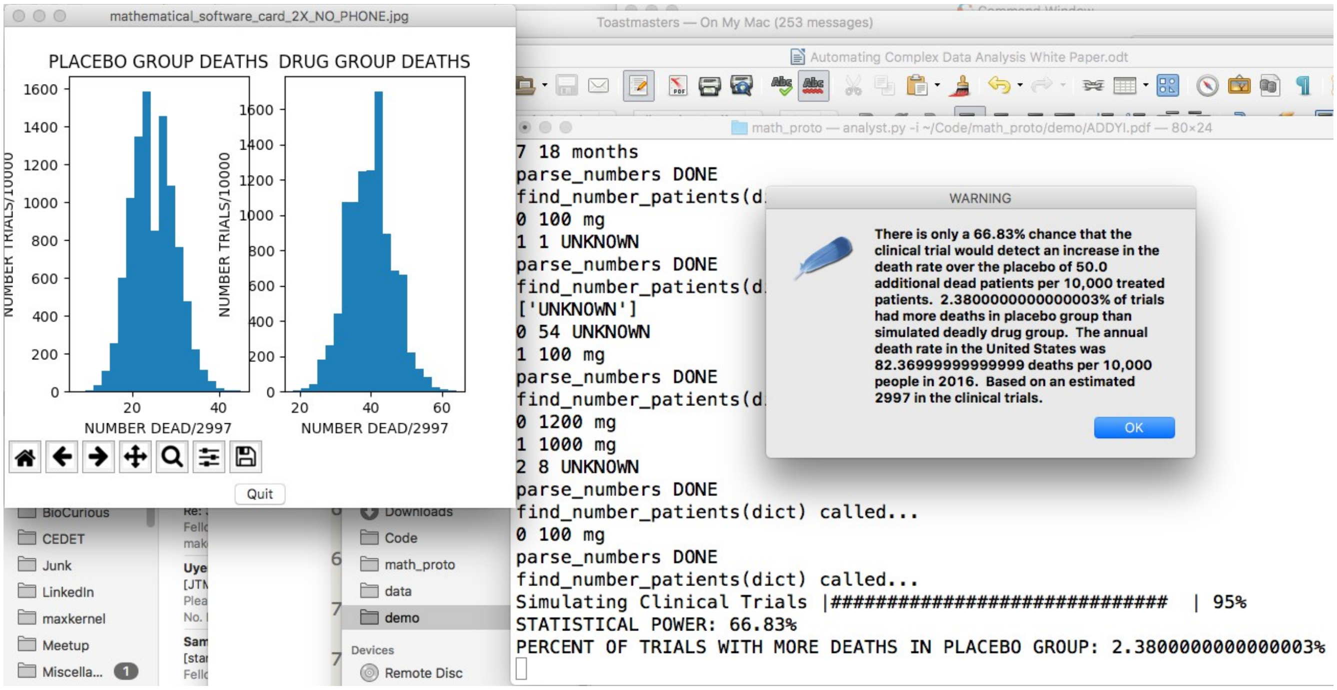
The approved 100 mg ADDYI dosage at bedtime was administered to 2,997 premenopausal women with acquired, generalized HSDD in clinical trials, of whom 1672 received treatment for at least 6 months, 850 received treatment for at least 12 months, and 88 received treatment for at least 18 months

First the prototype again identifies a passage referring to the number of patients in the clinical trial (above).



I believe there were 2997 patients in the clinical trials.

Then the prototype extracts the number of patients from the passage (above).



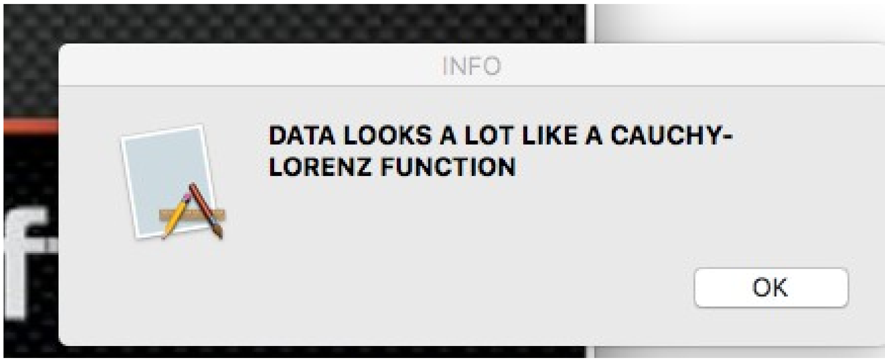
```
7 18 months
parse_numbers DONE
find_number_patients(d
0 100 mg
1 1 UNKNOWN
parse_numbers DONE
find_number_patients(d
['UNKNOWN']
0 54 UNKNOWN
1 100 mg
parse_numbers DONE
find_number_patients(d
0 1200 mg
1 1000 mg
2 8 UNKNOWN
parse_numbers DONE
find_number_patients(dict) called...
0 100 mg
parse_numbers DONE
find_number_patients(dict) called...
Simulating Clinical Trials |###############################  | 95%
STATISTICAL POWER: 66.83%
PERCENT OF TRIALS WITH MORE DEATHS IN PLACEBO GROUP: 2.3800000000000003%
```

There is only a 66.83% chance that the clinical trial would detect an increase in the death rate over the placebo of 50.0 additional dead patients per 10,000 treated patients. 2.3800000000000003% of trials had more deaths in placebo group than simulated deadly drug group. The annual death rate in the United States was 82.36999999999999 deaths per 10,000 people in 2016. Based on an estimated 2997 in the clinical trials.

Finally, the prototype performs a statistical power analysis to determine the safety level of the drug. As with Vioxx, it determines that a substantial increase in the risk of death over the placebo group could not be reliably detected with the number of patients in the clinical trial.

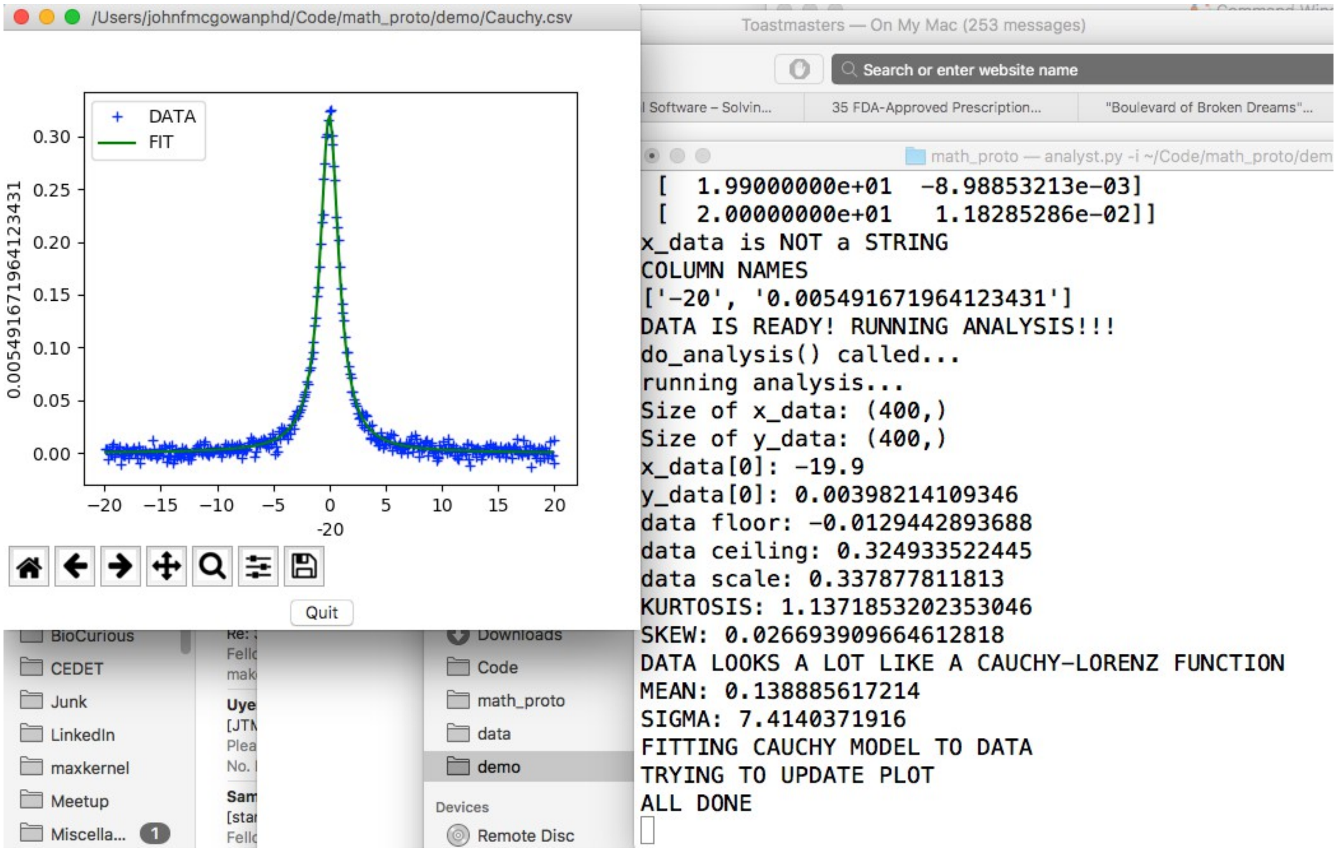
Here are some results using the prototype to identify the mathematics responsible for some data using some simple pattern and shape recognition methods.

Recognizing simulated Cauchy-Lorenz data



DATA LOOKS A LOT LIKE A CAUCHY-LORENZ FUNCTION

First the program analyzes the data and determines that it looks like the Cauchy-Lorenz function.

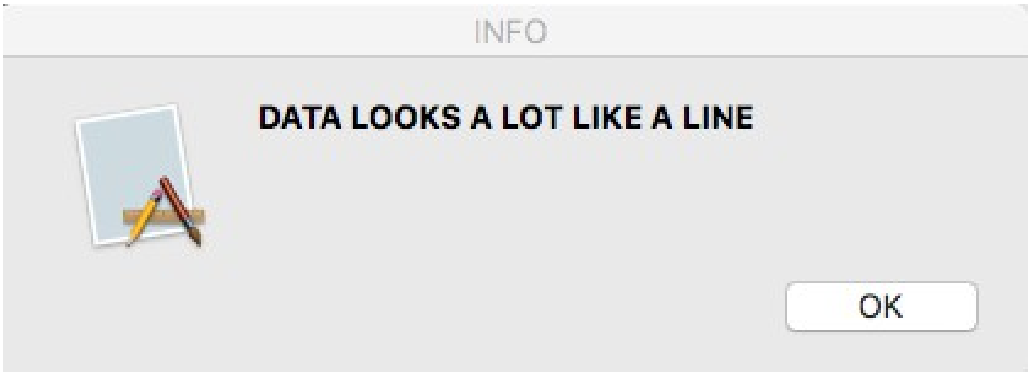


```
 [  1.99000000e+01  -8.98853213e-03]
 [  2.00000000e+01   1.18285286e-02]]
x_data is NOT a STRING
COLUMN NAMES
['-20', '0.005491671964123431']
DATA IS READY! RUNNING ANALYSIS!!!
do_analysis() called...
running analysis...
Size of x_data: (400,)
Size of y_data: (400,)
x_data[0]: -19.9
y_data[0]: 0.00398214109346
data floor: -0.0129442893688
data ceiling: 0.324933522445
data scale: 0.337877811813
KURTOSIS: 1.1371853202353046
SKEW: 0.026693909664612818
DATA LOOKS A LOT LIKE A CAUCHY-LORENZ FUNCTION
MEAN: 0.138885617214
SIGMA: 7.4140371916
FITTING CAUCHY MODEL TO DATA
TRYING TO UPDATE PLOT
ALL DONE
```

The program fits a Cauchy-Lorenz model tot the data and shows a plot comparing the data and the fit result. In this case, the agreement is excellent.

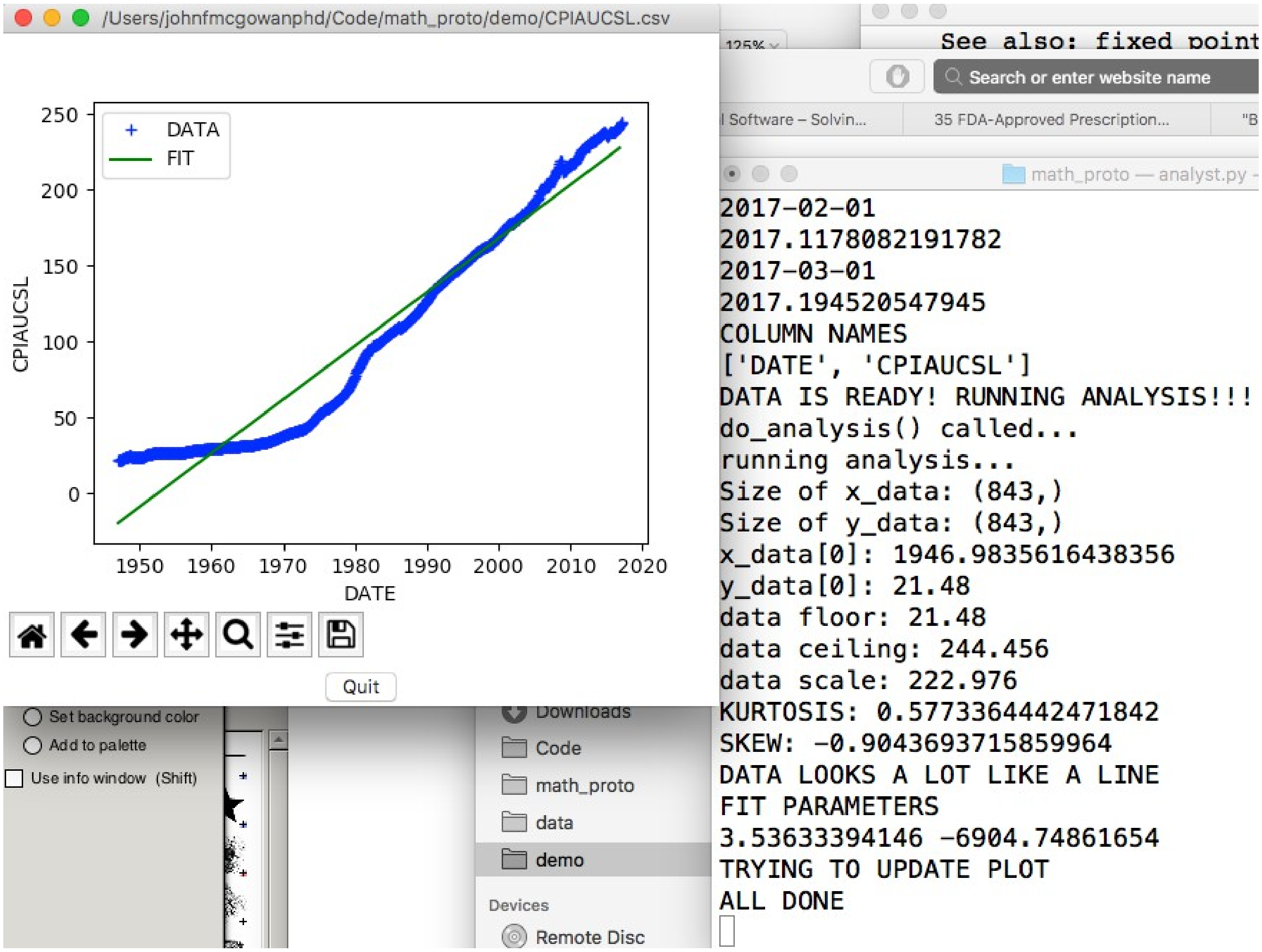
Recognizing real data

Next, the prototype is given some data on the consumer price level from the Federal Reserve:



DATA LOOKS A LOT LIKE A LINE

The prototype thinks the data looks a lot like a line.



```
2017-02-01
2017.1178082191782
2017-03-01
2017.194520547945
COLUMN NAMES
['DATE', 'CPIAUCSL']
DATA IS READY! RUNNING ANALYSIS!!!
do_analysis() called...
running analysis...
Size of x_data: (843,)
Size of y_data: (843,)
x_data[0]: 1946.9835616438356
y_data[0]: 21.48
data floor: 21.48
data ceiling: 244.456
data scale: 222.976
KURTOSIS: 0.5773364442471842
SKEW: -0.9043693715859964
DATA LOOKS A LOT LIKE A LINE
FIT PARAMETERS
3.53633394146 -6904.74861654
TRYING TO UPDATE PLOT
ALL DONE
```

The prototype tries fitting a line to the data. It gets moderate agreement. The data is a lot like a line but clearly different. In a more advanced system, the system would then try to analyze the residuals, the differences between the data and the line model and try to find a model for the residuals, exploring the space of possible functions until it found a good match or several good matches. There are many mathematical models that can match a given data set.

Current Attempts to Automate Complex Data Analysis

There are a number of other current attempts to automate complex data analysis, primarily within the context of "machine learning" (ML) and "deep learning." Most machine learning and deep learning, formerly known as artificial neural networks (ANN), involves fitting extremely complex mathematical models with very large numbers, sometimes hundreds of thousands, of adjustable/fitted parameters to data, often very large amounts of data collected by Internet giants such as Google and Facebook.

The Automated Statistician is a research project from scientists and Cambridge and MIT that has received funding from Google. It appears to attempt automate the identification of the appropriate mathematical/statistical model from data.

Google and Facebook have both announced projects both known as AutoML to use artificial intelligence to automate the construction of machine learning models.

Machine learning startup Skytree markets a technology called AutoModel to automate selection of the algorithms and parameters for machine learning models.

It is probable with the current Artificial Intelligence/DeepLearning/Machine Learning/Data Science craze that a number of other established companies and startups are attempting to automate complex data analysis in various ways.

Conclusion

State of the art complex data analysis is slow, expensive, error-prone, and often unconvincing. Automation of complex data analysis can save time, save money, reduce or eliminate errors, save lives in cases like FDA drug approvals, increase persuasiveness, and enable third-party auditing of results.

The author is developing software tools and algorithms for automating complex data analysis. The author is looking for real-world data and use cases for automating complex data analysis and testing these tools and algorithms. Please contact the author at jmcgowan79@gmail.com

About the Author

John F. McGowan, Ph.D. solves problems using mathematics and mathematical software, including developing gesture recognition for touch devices, video compression and speech recognition technologies. He has extensive experience developing software in C, C++, MATLAB, Python, Visual Basic and many other programming languages. He has been a Visiting Scholar at HP Labs developing computer vision algorithms and software for mobile devices. He has worked as a contractor at NASA Ames Research Center involved in the research and development of image and video processing algorithms and technology. He has published articles on the origin and evolution of life, the exploration of Mars (anticipating the discovery of methane on Mars), and cheap access to space. He has a Ph.D. in physics from the University of Illinois at Urbana-Champaign and a B.S. in physics from the California Institute of Technology (Caltech).

E-Mail: jmcgowan79@gmail.com
Web Site: http://www.mathematical-software.com/

Video of Presentation

A video of the author's presentation on “Automating Complex Data Analysis” is available on YouTube.

Credits

The image of Tycho Brahe is from Wikimedia Commons and is in the public domain.

The image of Johannes Kepler is from Wikimedia Commons and is in the public domain.

The image of the frontispiece and title page of Galileo's Dialogue is from Wikimedia Commons and is in the public domain.

End Notes (Below)

1 “IBM TARGET SPSS REPORTS SECOND-QUARTER REVENUE DOWN 8%” Brian Tarran, RESEARCH LIVE, August 5, 2009 *SPSS Inc. reported quarterly revenues of $69.7 million in the second quarter of 2009, extrapolating to annual revenues gives about $280 million in revenues.*

2 “IBM to Acquire SPSS Inc. to Provide Clients Predictive Analytics Capabilities” IBM Press Release, July 28, 2009 *IBM (NYSE: IBM) and SPSS Inc. (Nasdaq: SPSS) today announced that the two companies have entered into a definitive merger agreement for IBM to acquire SPSS, a publicly-held company headquartered in Chicago, in an all cash transaction at a price of $50/share, resulting in a total cash consideration in the merger of approximately $1.2 billion.*

3 The Wolfram Research Company Background page reports “close to 700” employees. (Accessed June 19, 2017): *In our focus on long-term objectives, we have chosen to remain a closely held private group of companies, and our consistent business success has allowed us to build a strong organization capable of pursuing a broad range of research and development. With a tightly knit but geographically distributed group of* ***close to 700****, we are able to take a unified approach to a remarkable range of interdisciplinary projects, efficiently developing major innovations and quickly implementing them in our products.* Most software companies have revenues around $100-125,000 per employee (personal experience), suggesting Wolfram Research has total annual revenues in the neighborhood of $70-90 million.

4 “How did Vioxx debacle happen?” By Rita Rubin, *USA Today*, October 12, 2004 *The move was a stunning denouement for a blockbuster drug that had been marketed in more than 80 countries with* ***worldwide sales totaling $2.5 billion in 2003***.

5 Several estimates of the number of patients killed and seriously harmed by Vioxx were made. Dr. David Graham's November 2004 Testimony to the US Senate Finance Committee gives several estimates including his own.

6 See, for example, “Why is SAS preferred in clinical trials over other software?” by Adrian Olszewski, Quora

7 “35 FDA-Approved Prescription Drugs Later Pulled from the Market”, http://prescriptiondrugs.procon.org/view.resource.php?resourceID=005528

8 Singh, G, Triadafilopoulos, “Epidemiology of NSAID induced gastrointestinal complications.” *J Rheumatol*1999;**26** (suppl):18–24.

9 “Gastrointestinal Toxicity of Nonsteroidal Antiinflammatory Drugs,” M. Michael Wolfe, M.D., David R. Lichtenstein, M.D., and Gurkirpal Singh, M.D., *New England Journal of Medicine* 1999; 340: 1888-1899, June 17, 1999 DOI: 10.1056/NEJM199906173402407

10 “Arthritis Study Finds Risk in Antacid Use,” by Thomas H. Maugh II, Times Medical Writer, *Los Angeles Times*, July 23, 1996

11 Tom Nesi's book *Poison Pills: The Untold Story of the VIOXX Drug Scandal* gives a critical overview of the claims for a “hidden epidemic” of NSAID-induced deaths in *Chapter 9: The Epidemic Blossoms* (pages 80-87) and has several quotes from physicians doubting the claims. Nesi, Tom, *Poison Pills: The Untold Story of the VIOXX Drug Scandal*, St. Martin's Press, New York, 2008

12 “http://www.nejm.org/doi/full/10.1056/NEJM199910283411813,” Letter to the Editor from Robert W. Morgan, M.D., S.M.Hyg, Exponent Health Group, Menlo Park, CA 94025, N Engl J Med 1999; 341:1397-1399 October 28, 1999 DOI: 10.1056/NEJM199910283411813 *In his letter, Dr. Morgan sharply questions the 16,500 estimate of death from NSAIDs and the extrapolation from the ARAMIS arthritis database at Stanford to the general population. The correspondence also includes a response from Professor James Fries at Stanford, the Principal Investigator (PI) for the ARAMIS database project.*

13 United States Centers for Disease Control (CDC) MMWR (Morbidity and Mortality Weekly Report) May 7, 1999, Vol. 48, No. 17 *“May is National Arthritis Month. Arthritis and other rheumatic conditions are among the most common chronic conditions and constitute the leading cause of disability, affecting an*

estimated 42.7 million persons in the United States."

14 Traffic Safety Facts 1999: A Compilation of Motor Vehicle Crash Data from the Fatality Analysis Reporting System and the General Estimates System, US Department of Transportation, National Highway Traffic Safety Administration

15 Drug safety assessment in clinical trials: methodological challenges and opportunities
Sonal Singh and Yoon K Loke
Trials 2012 13:138
DOI: 10.1186/1745-6215-13-138© Singh and Loke; licensee BioMed Central Ltd. 2012
Received: 9 February 2012 Accepted: 30 July 2012 Published: 20 August 2012
The premarketing clinical trials required for approval of a drug primarily guard against type 1 error. ***RCTs are usually statistically underpowered*** *to detect the specific harm either by recruitment of a low-risk population or low intensity of ascertainment of events. The lack of statistical significance should not be used as proof of clinical safety in an underpowered clinical trial.*

16 Dorothy Hamill told her story in an appearance on *Larry King Live* on August 29, 2000 in a joint appearance with fellow 1976 Olympian Bruce Jenner, both paid by Merck. Nesi, Tom, *Poison Pills: The Untold Story of the VIOXX Drug Scandal*, St. Martin's Press, New York, 2008, pages 22-23

17 Prakash, Snigdha, *All the Justice Money Can Buy: Corporate Greed on Trial,* Kaplan Publishing, New York, 2011, page 160: "*As you can see in the attached, our Dorothy Hamill campaign (assessed post-JAMA) might be the highest impact ad we've had for Vioxx. We estimate that our campaign is generating at least 4:1 confirmed this summer (measured based on the number of new patients starts driven by DTC.)*" from internal Merck documents entered into evidence in one of the personal injury lawsuits. JAMA is *Journal of the American Medical Association*. DTC is Direct to Consumer advertising.

18 "Figure skating is beautiful on the ice, brutal on the body," by Joy Sewing, *The Chronicle* (San Francisco), February 17, 2010

19 Expression of Concern: Bombardier et al., "Comparison of Upper Gastrointestinal Toxicity of Rofecoxib and Naproxen in Patients with Rheumatoid Arthritis," N Engl J Med 2000;343:1520-8.
Gregory D. Curfman, M.D., Stephen Morrissey, Ph.D., and Jeffrey M. Drazen, M.D.
N Engl J Med 2005; 353:2813-2814 December 29, 2005 DOI: 10.1056/NEJMe058314

20 Comparison of Upper Gastrointestinal Toxicity of Rofecoxib and Naproxen in Patients with Rheumatoid Arthritis
Claire Bombardier, M.D., Loren Laine, M.D., Alise Reicin, M.D., Deborah Shapiro, Dr. P.H., Ruben Burgos-Vargas, M.D., Barry Davis, M.D., Ph.D., Richard Day, M.D., Marcos Bosi Ferraz, M.D., Ph.D., Christopher J. Hawkey, M.D., Marc C. Hochberg, M.D., Tore K. Kvien, M.D., and Thomas J. Schnitzer, M.D., Ph.D., for the VIGOR Study Group
N Engl J Med 2000; 343:1520-1528 (November 23, 2000 DOI: 10.1056/NEJM200011233432103)

21 Nesi, Tom, *Poison Pills: The Untold Story of the VIOXX Drug Scandal*, St. Martin's Press, New York, 2008, page 224, from data presented to the FDA – the raw numbers were not reported in the *New England Journal of Medicine* article on the VIGOR study published in November of 2000.

22 The count of heart attacks is from other reports than Tom Nesi's *Poison Pills* book; it is not in the table on page 224 of *Poison Pills*. The exact number of deaths and other adverse events in the VIGOR study jumps around by a few events in different reports.

23 "Merck Critic Linked to Hedge Fund," by Geeta Anand and Gregory Zuckerman, *Wall Street Journa*l, Updated December 3, 2004

24 "Dr. Eric Topol denies conflict of interest over Vioxx condemnation," by Shelley Wood, *Medscape*, December 2, 2004

25 Risk of acute myocardial infarction and sudden cardiac death in patients treated with cyclo-

oxygenase 2 selective and non-selective non-steroidal anti-inflammatory drugs: nested case-control study

Graham, David J et al.

The Lancet , Volume 365 , Issue 9458 , 475 – 481, 5 February 2005

26 *N Engl J Med.* 2005 Mar 17;352(11):1092-102. Epub 2005 Feb 15.
Cardiovascular events associated with rofecoxib in a colorectal adenoma chemoprevention trial.
Bresalier RS1, Sandler RS, Quan H, Bolognese JA, Oxenius B, Horgan K, Lines C, Riddell R, Morton D, Lanas A, Konstam MA, Baron JA; Adenomatous Polyp Prevention on Vioxx (APPROVe) Trial Investigators.

27 M. Schumann, "Top 100 Megabrands: Chevrolet Leads the Race along Megabrand Road, but Marketers Hit 2001 Potholes," *Advertising Age* 72, no. 1 (2001): 1.

28 Ventola CL. Direct-to-Consumer Pharmaceutical Advertising: Therapeutic or Toxic? *Pharmacy and Therapeutics.* 2011;36(10):669-684.

29 "Merck Agrees to Settle Vioxx Suits for $4.85 Billion," by Alex Berenson, *New York Times*, Business Day section, November 9, 2007

30 "Analysts See Merck Victory in Vioxx Settlement," by Alex Berenson, *New York Times*, Business Day section, November 10, 2007

31 Vioxx National Class Action Canada

32 "Vioxx settlement to cost Merck up to $37M US: Deal would settle all lawsuits in Canada," The Associated Press, January 19, 2012

33 "Merck Australia wins appeal in Vioxx lawsuit," The Associated Press, October 11, 2011

34 German Class Action Lawsuit on Behalf of All German Citizens Regarding Vioxx, Oct. 4, 2005

35 "Merck to Pay $950 Million over Vioxx," by Duff Wilson, *New York Times*, Business Day section, November 22, 2011

36 "Merck to Pay $830 Million to Settle Vioxx Shareholder Suit: Settlement moves drug company closer to resolving litigation surrounding pulled painkiller," by Peter Loftus, *Wall Street Journal*, updated January 15, 2016

37 "Merck to Fund $4.85 Billion Vioxx Settlement," CBS News/Associated Press, July 17, 2008 *"The Vioxx case has cost Merck at least $6.38 billion,* ***including more than $1.53 billion through March 31 on legal costs for defense research and individual trials****, most of which it has won."*

38 "Merck's stock plunges after recall of drug Vioxx" By Linda A. Johnson, Associated Press, Published: Oct. 1, 2004 12:00 a.m

39 Friedhoff, Lawrence, *New Drugs: An Insider's Guide to the FDA's New Drug Approval Process for Scientists, Investors and Patients*, Pharmaceutical Special Projects Group, LLC (PSPG) Publishing, New York, 2009, page 125: "*The actual cost of doing the studies required to get one new drug approved is much lower. For a new chemical entity, a cost between $45 and $70 million is typically adequate.*"

40 "Another Controversy for the 'Female Viagra'?" by Kaitlin Bell Barnett, *Scientific American*, October 17, 2015

41 "FDA approves first treatment for sexual desire disorder: Addyi approved to treat premenopausal women," FDA News Release, August 18, 2015

42 "US Food and Drug Administration Approval of Flibanserin: Even the Score Does Not Add Up," Editorial by Steven Woloshin, MD, MS, and Lisa Schwarz, MD, MS, *JAMA Intern Med.* 2016;176(4):439-442. doi:10.1001/jamainternmed.2016.0073

43 "Efficacy and Safety of Flibanserin for the Treatment of Hypoactive Sexual Desire Disorder in Women: A Systematic Review and Meta-analysis," by Loes Jasper, MD; Frederik Feys, MS, PhD; Wichor M. Bramer BS; et al, *JAMA Intern Med.* 2016;176(4):453-462. doi:10.1001/jamainternmed.2015.8565