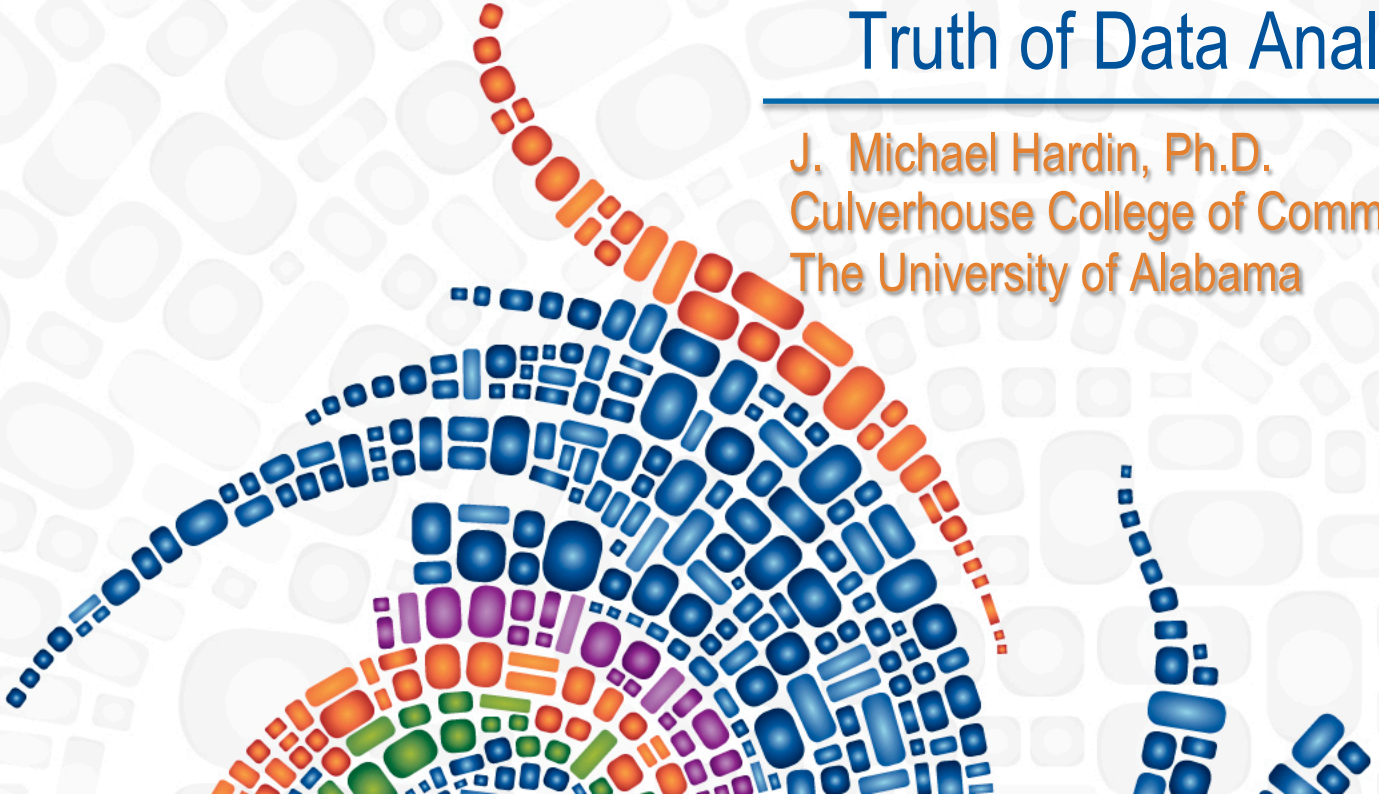
SAS® GLOBAL FORUM 2015
The Journey Is Yours
Paper 3423-2015

To Believe or Not To Believe? The Truth of Data Analytics Results

J. Michael Hardin, Ph.D.
Culverhouse College of Commerce
The University of Alabama



SAS GLOBAL FORUM

Introduction

- Not a technical talk
- Conceptual Issues

SAS GLOBAL FORUM

Current Interest

- Wall Street Journal articles
  - Data Crunchers are sexy!
- Harvard Business Review

Big Data

JOURNAL REPORT | UNLEASHING INNOVATION

BIG DATA

Meet The Experts: A New Online Conversation DETAILS, R2

© 2013 Dow Jones & Company. All Rights Reserved.

THE WALL STREET JOURNAL. Monday, March 11, 2013

HOW Big Data IS CHANGING THE WHOLE EQUATION FOR BUSINESS

BY STEVEN ROSENBUSH AND MICHAEL TOTTY

THERE'S A TON OF INFORMATION OUT THERE. And businesses are figuring out how to put it to work.

The experts call this state of affairs big data. The definition is squishy, but it usually boils down to this: Companies have access to vastly more information than they used to, it comes from many more different sources than before, and they can get it almost as soon as it's generated.

Big data often gets linked to companies that *already* deal in information, like Google, Facebook and Amazon. But businesses in a slew of industries are putting it front and center in more and more parts of their operations. They're gathering huge amounts of information, often meshing traditional measures like sales with things like comments on social-media sites and location information from mobile devices. And they're scrutinizing it to figure out how to improve their products, cut costs and keep customers coming back.

Shippers are using sensors on trucks to find ways to speed up deliveries. Manufacturers can travel through thousands of forum posts to figure out if customers will like a new feature on their product. Hiring managers study how candidates answer questions to see if they'd be a good match.

Lots of obstacles remain. Some are technical, but business as usual also can stand in the way. In most companies, decisions are still based on HIPPO—the highest paid person's opinion—and persuading an executive that data trumps intuition can be a hard sell.

Please turn to the next page

From a Molehill To a Mountain

A timeline of digital data

Everyone knows that the amount of digital data has exploded in recent years. But it isn't always easy to put today's mountain of data in perspective.

To illustrate the data boom, here's a timeline with snapshots of some of the largest corporate data collections from each decade since the Univac computer was the hot new thing.

—*Matthew Kassel*

2010s | **100 Petabytes**
Facebook Is Facebook's data hoard bigger than Google's today? Some say yes, some say no. According to Facebook, its user content makes up more than 100 petabytes of stored photos and video. And analyzing that data generates about 500 terabytes of new information every day—more than 2½ times the size of a '90s Wal-Mart data cache.

2000s | **25 Petabytes**
Google The explosion of the Internet in the '90s set the stage for Web-based companies to emerge in the following decade as the global leaders in big data.

1990s | **180 Terabytes**
Wal Mart Wal-Mart in this decade became the largest American bricks-and-mortar retail operation, and, it is believed, had the biggest commercial data warehouse in the world. To put this number in perspective, experts have estimated that Amazon.com in the late '90s had single-digit terabytes of stored data.

1980s | **450 Gigabytes**
CitiCorp's NAIB In the 1980s, banks were at the forefront of data growth, with ATMs coming into vogue and banks focusing on collecting and analyzing transaction data for all their businesses.

1970s | **80 Gigabytes**
Federal Express Cosmos FedEx's Cosmos system, introduced in the 1970s, allowed the company to scan and track its huge volume of packages being shipped around the world.

1960s | **807 Megabytes**
American Airlines Sabre In the 1960s, American Airlines developed Sabre, a flight-reservation system built around one of the largest IBM computing systems available. Sabre was one of the first online computing systems, allowing the airline to keep track of an immense matrix of reservations, flight schedules and seat inventories.

1950s | **600 Megabytes**
John Hancock In the 1950s, insurance companies had the biggest data hoards as they collected and stored information about policyholders. John Hancock Mutual Life Insurance Co. was one of the pioneers in digitizing customer information, storing data from two million life-insurance policies on a Univac computing system acquired in 1955.

100,000,000,000,000,000 BYTES
90,000,000,000,000,000
80,000,000,000,000,000
70,000,000,000,000,000
60,000,000,000,000,000
50,000,000,000,000,000
40,000,000,000,000,000
30,000,000,000,000,000
20,000,000,000,000,000
10,000,000,000,000,000
0

1950 1960 1970 1980 1990 2000 2010

Contemporary empiricism and normative finance

- Is *Good to Great* great? "It is important to understand that we developed all of the concepts in this book by making empirical deductions directly from the data. We did not begin this project with a theory to test or prove. We sought to build a theory from the ground up, derived directly from the evidence."
- "Contemporary empiricism is like interstate driving by only looking in the rearview mirror." Dr. Bob Brooks, UA Professor of Finance
- "There are things we know, absent empirical data."

  Dr. Bob Brooks, UA Professor of Finance

---

8 J. Collins, *Good to Great* (New York, 2001), HarperCollins Publishers, Inc. See Kristine Beck and Bruce Niendorf, "Good to Great, or Great Data Mining?" *Journal of Financial Education* (Spring 2009), 80-95.

SAS GLOBAL FORUM

Let's look at the Paper

- Kristine Beck and Bruce Niendorf, "Good to Great, or Great Data Mining?" *Journal of Financial Education* (Spring 2009), 80-95.

- Hand, David J., "Data Mining: Statistics and More?", The American Statistician, (52), May 1998, 112-118

SAS GLOBAL FORUM

Who coined the term data mining?

- Michael Lovell's paper

Data Mining

“I originally titled the paper "Data Grubbing." But the editor was concerned that the paper was too pessimistic. I made some minor adjustments and he was still unhappy. I changed the title to Data Mining and he bought it. In retrospect that was a mistake, because data mining now refers to the use of a variety of different techniques for trying to extract conclusions from the gigantic data sets that are now generated by credit cards and so forth.”

-Mike Lovell

Wesleyan University
(personal communication, e-mail)

SAS GLOBAL FORUM

Outline

I. Explore some aspects of “analytic” approach versus traditional statistics.

II. Review of some Basic Philosophy.

III. Philosophical aspects of statistics.

IV. A philosophy for analytics.

V. Final comments.

Traditional Statistics—Early Years

- Randomization
- Experimental Design
- Generalizability
- Data collection expensive, data sets “small”
- Much emphasis on estimation
- Explicit hypotheses

What is Statistics

- Statistics is operational knowledge accumulation and as such is a the frontline of any discussion of the scientific method in particular and the philosophy of science in general.
- Thus, as Oscar Kempthorne (1976) has noted, statisticians (in particular applied statisticians) are involved in basic philosophical dilemmas. Unfortunately, however, neither statisticians nor scientist have recognized their involvement in these dilemmas. This lack of recognition has led to some very deep controversies in the field of statistics itself and in other scientific fields, especially where statistical issues such as the p-value or hypothesis testing has played a significant role in the controversy.

What is Statistics?

- …statistics refers to the methodology for the collection, presentation, and analysis of data, and for the uses of such data (p.1) (Neter, Wasserman and Whitmore, 1978).
- Statistics is the branch of scientific method which deals with the data obtained by counting or measuring the properties of populations of natural phenomena. In this definition 'natural phenomena' includes all happenings of the external world, whether human or not (p.2) (Kendall and Stuart, 1977).
- Statistical methods of analysis are intended to aid the interpretation of data that rare subject to appreciable variability (p.1) (Cox and Hinkley, 1974).

What is Statistics?

- Years ago a statistician might have claimed that statistics deals with the processing of data…today's statistician will be more likely to say that statistics is concerned with decision making in the face of uncertainty (p.1) (Chernoff and Moses, 1959).
- By [statistical] inference I mean how we find things out– whether with a view to using the new knowledge as a basis for explicit action or not—and how it comes to pass that we often acquire practically identical opinions in the light of evidence (p.1) (Savage, 1962).

Business Analytics/Data Mining

- No randomization
- No experimental design
- What about generalizability?
- Data is cheap, data sets large
- Much emphasis on prediction
- Perhaps, no hypotheses, or only loosely articulated

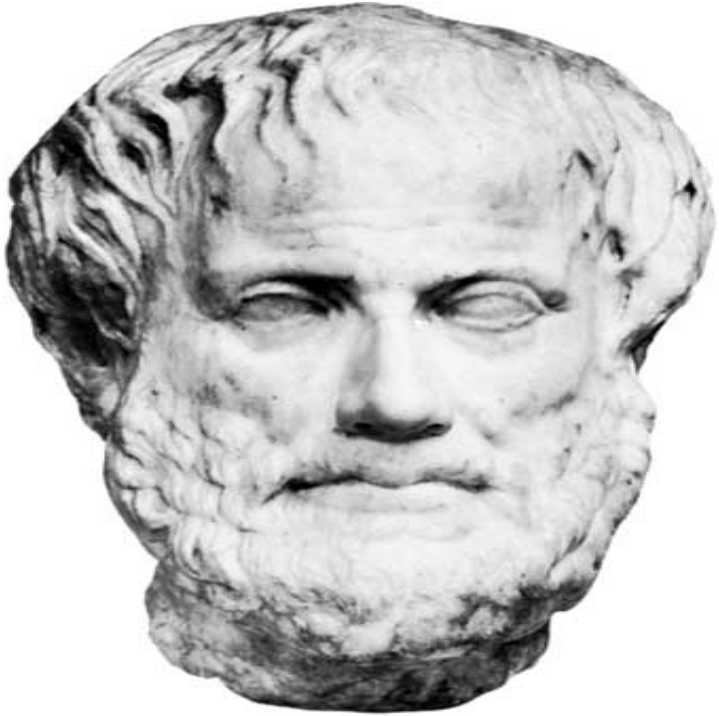
Brief Review of Philosophy

SAS GLOBAL FORUM

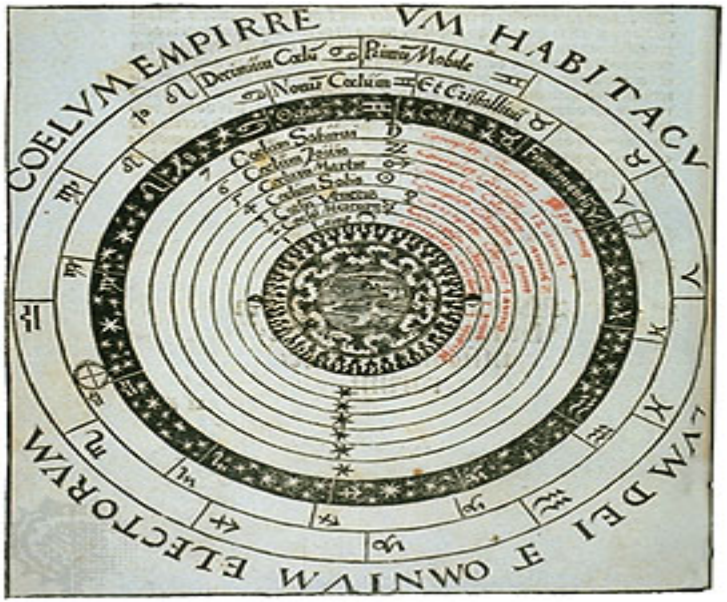
School of Athens

Plato vs. Aristotle

Aristotle



Medieval World View



Ptolematic System

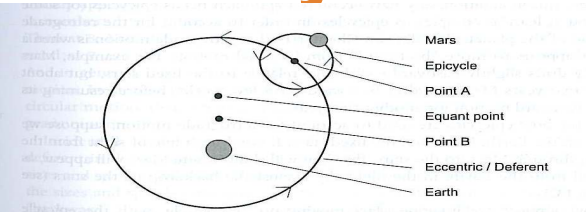


Figure 13.1 Treatment of Mars on the Ptolemaic system

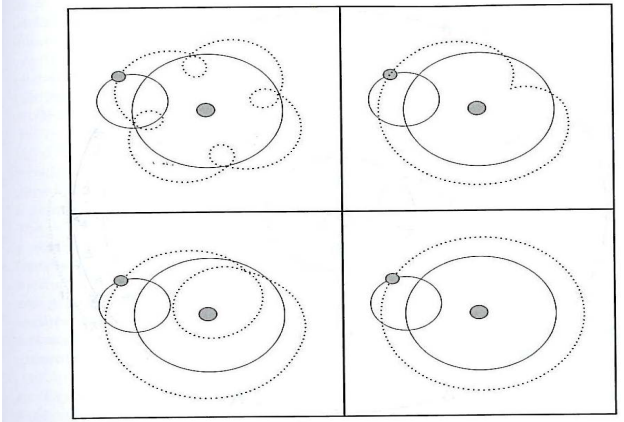


Figure 13.2 The flexibility of epicycle–deferent systems

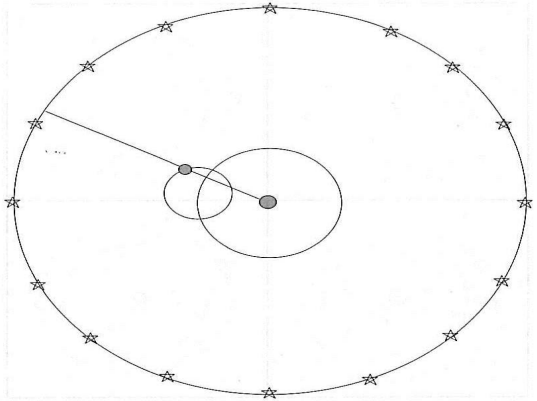
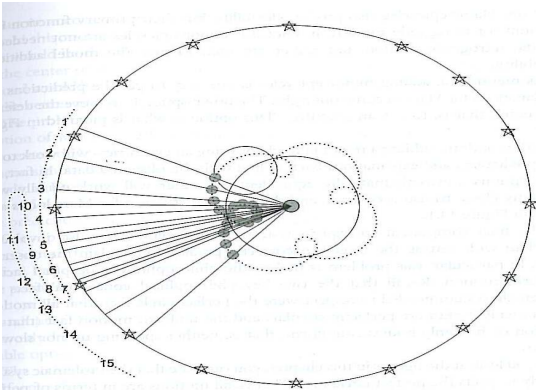


Figure 13.3 Position of Mars against the backdrop of the fixed stars



ure 13.4 Explanation of retrograde motion on the Ptolemaic system

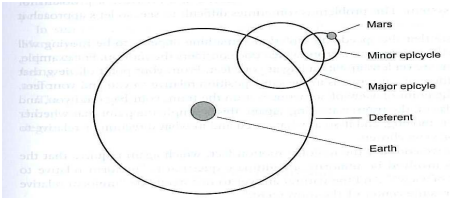


Figure 13.5 Minor and major epicycles

R. Dewitt, *Worldviews* (United Kingdom, 2010), Wiley-Blackwell Publishers, Page 115, 117-119.

Nicolas Copernicus (1473-1543)

The Copernican System

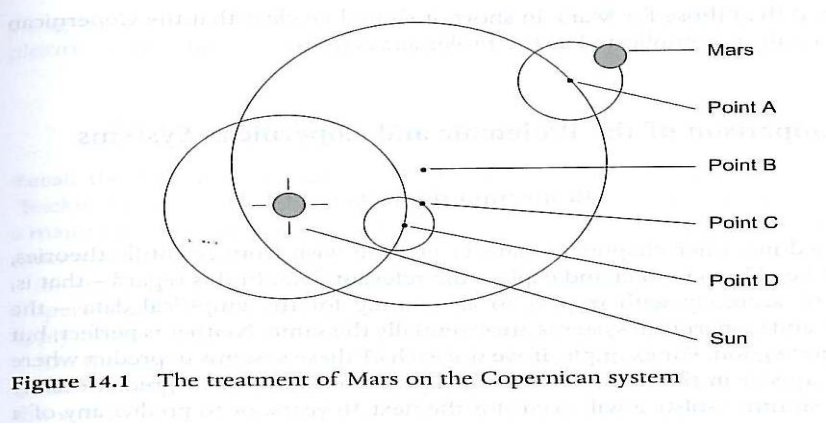


Figure 14.1 The treatment of Mars on the Copernican system

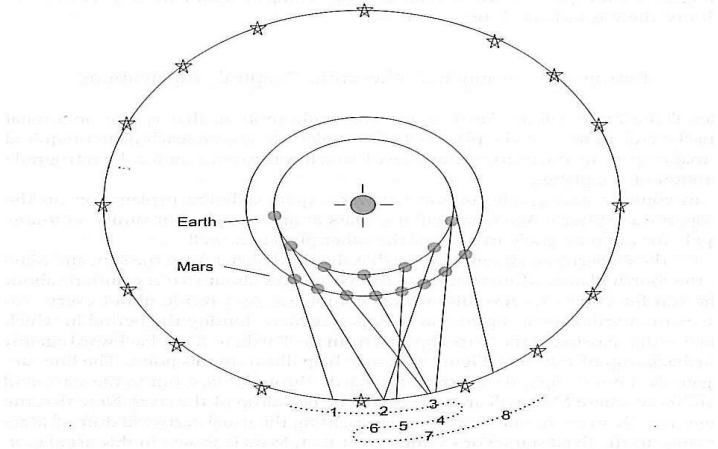


Figure 14.2 Explanation of retrograde motion on the Copernican system

R. Dewitt, *Worldviews* (United Kingdom, 2010), Wiley-Blackwell Publishers, Page 125,128.

Tycho Brahe (1546-1601)

Tychonic System

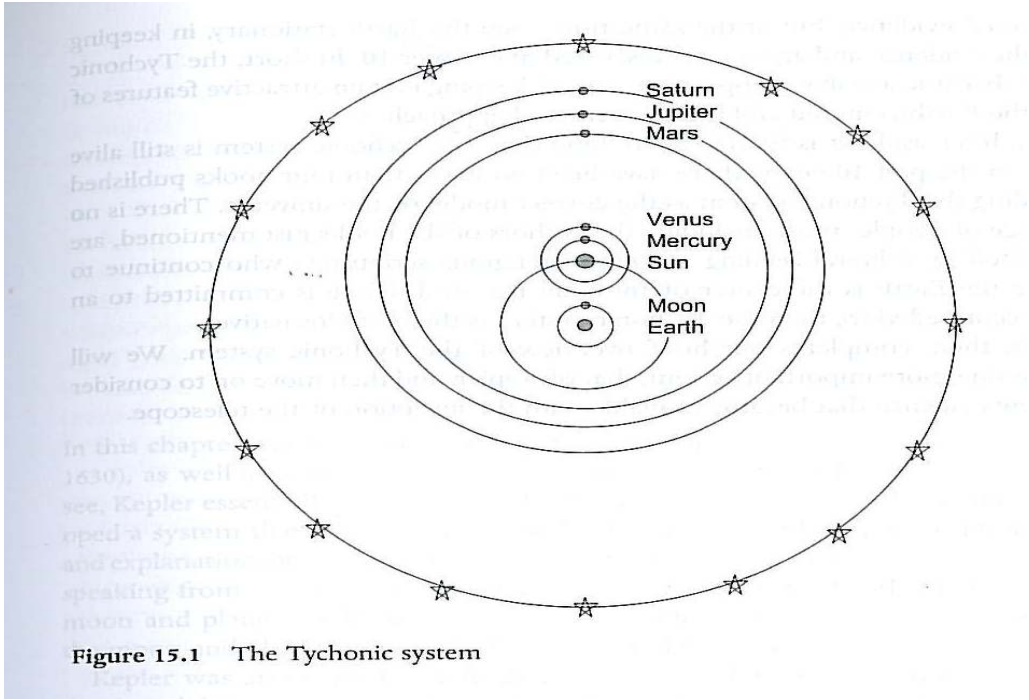


Figure 15.1 The Tychonic system

R. Dewitt, *Worldviews* (United Kingdom, 2010), Wiley-Blackwell Publishers, Page 135.

Johannes Kepler (1571-1630)

Kepler's System

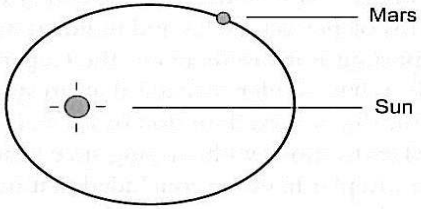


Figure 16.2 Orbit of Mars on Kepler's system

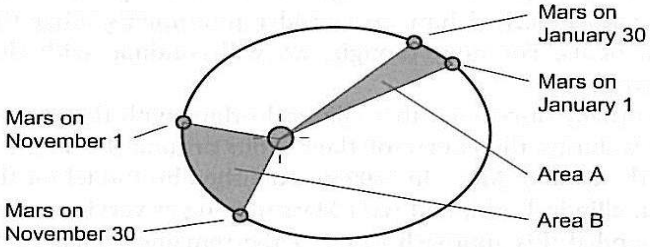


Figure 16.3 Illustration of Kepler's second law

Galileo Galileo

Evidence from the Telescope

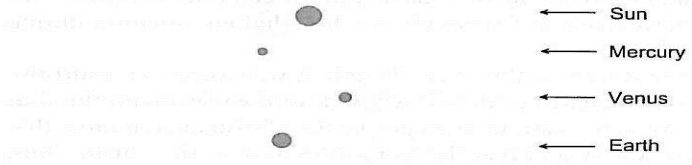


Figure 17.1 "Photo" of sun and planets

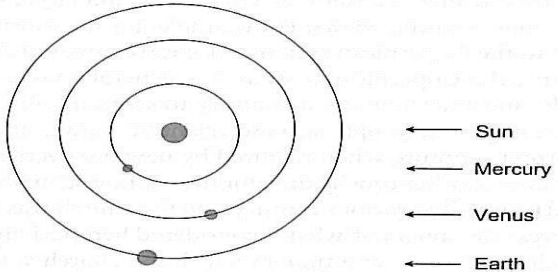


Figure 17.2 Sun-centered interpretation of "photo"

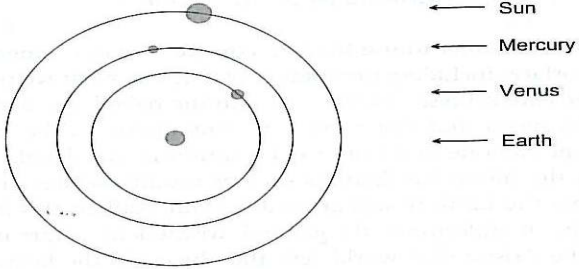
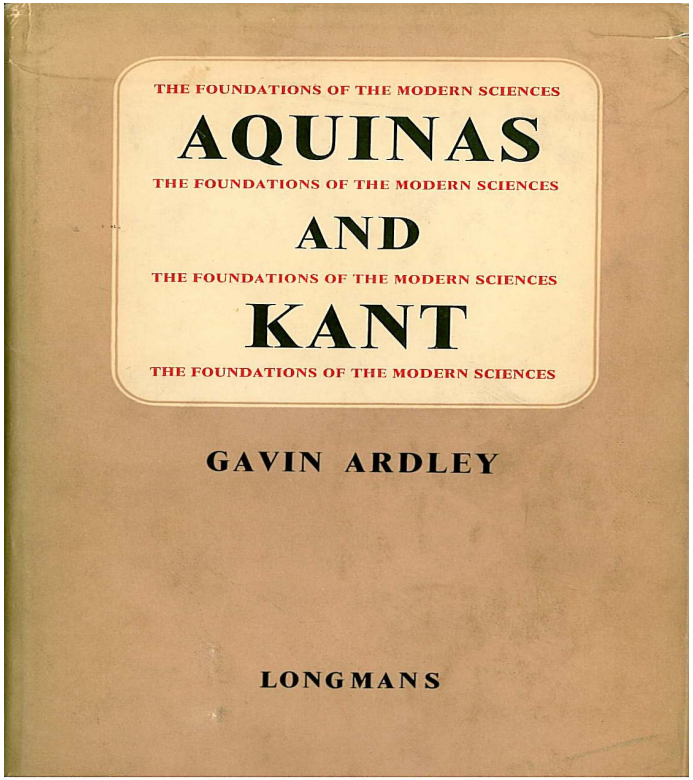


Figure 17.3 Earth-centered interpretation of "photo"

R. Dewitt, *Worldviews* (United Kingdom, 2010), Wiley-Blackwell Publishers, Page 150-151.

What was the Copernican Revolution?



Cardinal Bellarmine (On Galileo, his former teacher)

- If there were a real proof that the sun is in the centre of the universe, that the earth is in the third heaven, and that the sun does not go round the earth but the earth round the sun, then we should have to proceed with great circumspection in explaining passages of Scripture which appear to teach the contrary, and rather admit that we did not understand them than declare an opinion to be false which is proved to be true. But as for myself, I shall not believe that here are such proofs until they are shown to me. *Nor is a proof that, if the sun be supposed at the centre of the universe and the earth in the third heaven, the celestial appearances are thereby explained, equivalent to a proof that the sun actually is in the centre and the earth in the the third heaven.*

Saving the Appearances

- σω’ζεν τα` φαινομενα
- (sozein ta phaeinomena)
- Common idea from the time of Heraclitus to Plato to Aristotle

John Milton

Paradise Lost (Book 8)

Or if they list to try
Conjecture, he his fabric of the heavens
Hath left to their disputes, perhaps to move
His laughter at their quaint opinions wide
Hereafter, when they come to model heaven,
And calculate the stars; how they will wield
The mighty frame; how build, unbuild, contrive,
*To **save appearances**; how gird the sphere*
With centric and eccentric scribbled o'er,
Cycle and epicycle, orb in orb.

Cited in , O.Barfield, *Saving the Appearances, A Study in Idolarty* (Hanover, 1988) University Press of New England, 48.

SAS GLOBAL FORUM

What changed?

- How Reality would be determined based on appearances!

Pre-Modern Philosophy

Francis Bacon

Knowledge is Easy: Bacon's Methodology

- Bacon may have been the first data miner. His ideas may be summarized as:

1. Collect all relevant data without presuppositions
2. Analyze the data to uncover suggestive correlations among them
3. Experiment to test possible correlations

Bacon
The New Organon (1620)

- Human mind is an obstacle to knowledge of nature. It is the problem, not the solution
- Idols of Mind
  - Idols of the tribe
  - Idols of the cave
  - Idols of the theater
  - Idols of the marketplace

What is Science?

- Bacon (Simple Empiricism)
  - Science: Accumulation and Classification of Observations
  - Induction is the “easy road to knowledge”
    - Make observations
    - Summarize them
    - Generalize
- Discovery can be a routine and automatic process. Carried out “as by machinery”; only patience is needed, not difficult or abstract thought.
- Hume, 19th Century empiricism

SAS GLOBAL FORUM

What is Science?

- Galileo (rationalistic)
  - Type of Concept.
  - The combination of theory and experiment.
  - Goal of expressing laws of nature as mathematical relationships among measureable variables.
- Newton (rationalistic)
  - Alliance of mathematics and experimentation.
  - New concepts are the product, not of observation, or mathematical deduction, or the two together, but creative imagination.
  - Interaction of observation, theory, mathematical deduction and imaginative new concept.

SAS GLOBAL FORUM

Modern Philosophy

Rene’ Descartes

Empiricists

David Hume

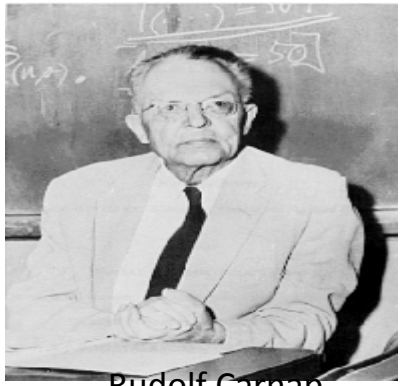
John Locke

George Berkeley

Gottfried Leibniz

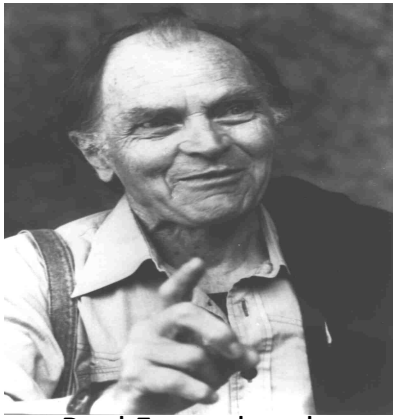
Immanuel Kant

20th Century



Rudolf Carnap

Ludwig Wittgenstein



Paul Feyerabend

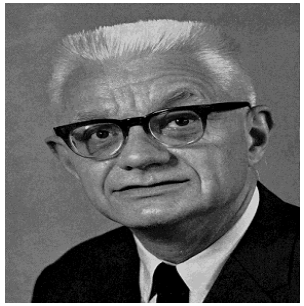
Thomas Kuhn

20th Century

I. Lakatos

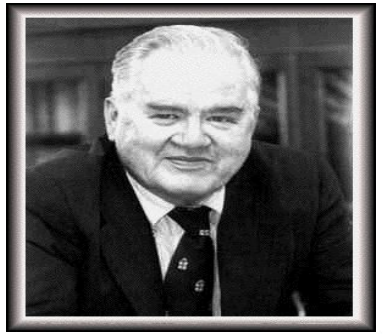
Michael Polyani

Karl Popper



Carl Hempel

I. J. (Jack) Good



John Tukey

Views of Science

- Positivists
- Instrumentalist
- Idealists
- Realists

SAS GLOBAL FORUM

20th Century positivism

I. Science starts from publicly observable data which can be described in pure observation-language independent of any theoretical assumptions.

II. Theories can be certified or falsified comparison with this fixed experimental

III. The choice between theories is rational, objective and in accordance with specifiable criteria.

So, what is science?

- Norman Campbell (1953)
  - Two aspects:
    I. Science is a body of useful and practical knowledge and a method of obtaining it.
    II. Science is pure intellectual study.
- Comments
  - Definition depends on ones philosophy
  - Theory verses data (facts)
  - ***Is science in the facts or the theories?***

SAS GLOBAL FORUM

So, what is science?

- Scientific Theory
  I. Expressed in only naturalistic terms
  II. Using if-then propositions
  III. Testable by experimentation
  IV. Always corrigible

Philosophies of Science
(What is it that we can know?)

- Empiricism: scientific knowledge is “wholly and entirely limited to descriptions-observation statements, generalizations, and the like-which are developed from pure sensory experience.”
- “…the only legitimate starting point for scientific knowledge is sensory experience, “i.e. the data.
- “…in a fundamental sense, the experiment (human experience) happens first, and scientific knowledge is distilled, induced as it were, from the experiment.”

SAS GLOBAL FORUM

Philosophies of Science

- Rationalism: - it is possible, by pure unaided reason, first, to conceive and comprehend certain very general features of the universe, and then, from these conceptions, to deduce mathematically a description of what the actual empirical world was like, prior to any experiment. The role of experiment…is a decision procedure for testing between alternative deduced results. If one reasoned mathematically and came to the conclusion that *x* would be the actual situation in the world, then an experiment could be designed to check whether or not *x* really did occur. *Gale (1979), Theory of Science*

Instrumentalism

- Main task of science is to “explain” / predict the relevant data (and further observations). It simply is not important whether or not a theory (or parts of a theory) reflect the way things “really are”.

Realism

- Science ought to explain and predict the relevant data, but additionally a good scientific theory (model) should reflect (refer) to things that really are (exist).

Prediction and Explanation (What is going on in science?)

- Description
- Predicting
- Explanation
- Understanding

SAS GLOBAL FORUM

Applying Philosophy to Statistics

- Two “Types” of Statistics
- Confirmatory-a *priori* theory
  - Hypothesis testing (and estimation in many cases)
  - “Traditional” statistics
  - Neyman-Pearson
- Exploratory-applied to observational data collected without well defined hypotheses for the purpose of generating hypotheses.—Knowledge Discovery
  - Revels *patterns*, the merits of which are determined introspectively by the researcher’s asking whether he or she finds the pattern explicable, given the context in which it is obtained (I.J. Good)
  - Achieves simplicity by reducing data or by smoothing data
  - “We can claim only to be groping toward the truth.” (Cochran, 1972)
  - Good, Tukey

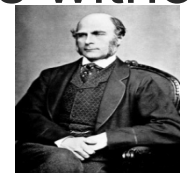
Origin of Exploratory statistics in 19th Century Empiricism

- Mulaik (1985)
  - Empiricist Thought
  - Baconianism-knowledge of the world can (and should be) attained without using systematic methods of inductive inquiry.
    - J.S. Mill (1891) refute to Whewell (1847)

      - Whewell-extended work of Gauss
        - Stresses the use of hypotheses
        - "Realist" view of science
        - Presumption: scientist had already correctly identified independent and dependent variables
        - Least squares interpretation

Origin of Exploratory statistics in 19th Century Empiricism (continued)

- Four methods of Inductive Inference—could be employed to discover causes without the use of hypotheses.



  - Causation=Association
  - Galton (1871)—regression line
  - Yule (1897)—correlation multiple regression, factor analysis, canonical correlation.
  - Karl Pearson-to be scientific, one has to be quantitative and use statistics in one's research. Statistics, however, is basically descriptive.

Origin of Exploratory statistics in 19th Century Empiricism (continued)

2. Associationism and cognitive calculi
   - Associationism—knowledge is obtained from the associations of impressions in phenomenal experience.
   - Cognitive Calculi-Cognitive processes may be modeled mathematically and may be augmented by mathematical devices.
   - Connections we come to perceive between impressions are supplied by the associative processes of the mind—modern paradigm for EDA.
   - Built strongly on the idea of causality=association
     - Pearson, Galton Yule—correlation coefficients, regression
   - French probabilists Condorcet and Laplace-mathematization of mental processes-early roots of Bayesian statistics.

Origin of Exploratory statistics in 19th Century Empiricism (continued)

- 3. Phenomenalism-the only reality is that which is perceived, and for statistics this means the only reality is the DATA!!
  - Opposite of realism, the early basis of science and statistics, e.g astronomers. (realism holds that scientific theories describe a universe of objects that actually exist independently of the scientist's efforts to know them. Thus, statistics in this view would try to distinguish the true value of quantities in their theories from the fallible, error-containing measures of these values.)
  - Lockean empiricists (existence of independent reality on which the impressions of the mind depended)-Laplace, Galton versus Pearson-therefore, statistics refers to nothing more real than summaries and resumes of data.
    - Denied Galton's Lockean attempt to see theoretical implications for regression
  - The regression line "is purely a statistical result and has no relation to any biological theory or hypothesis….. It is based on no data whatever except the actual statistics; *it is merely a convenient statistical method of expressing the observed facts.*

Summary

- Statistics, which had begun in a framework of scientific realism with astronomy, was transformed to fit a phenomenalist and instrumentalist empiricist framework by the end of the 19th and early in the 20th Century.
- Statistical developments, however , did follow two paths:
  - The one led by the “strong” empiricist philosophy, e.g. Pearson’s School.
  - The hypothetico-deductive method and confirmatory statistics of R.A. Fisher and his followers, Snedecor, Kempthrone, Rao, and Box.
- Comments:
  - In recent years realism has seen a resurgence
  - Interest in Popper-Is there a theory of Knowledge Discovery possible?

Conflict

- If our life is some variant of scientific realism, we will be concerned with the degree to which the reduced forms of data in EDA represent things which we believe exist in the world. Pearson's phenomenalism with data will just not do, and pity and Lockean empiricist with realist inclinations who gravitates to Pearson's phenomenalistically oriented form of Baconian statistics and attempts to discover things in the world beyond the statistical data. *Mulaik (1985, p.427)*

An Example

- Analysis: Bansal and Gupta (1978)
  - Probabilistic model for the survival of human lymphocytes following irradiation.
  - Standard paradigm within probability theory.
  - Three compartment model
  - Deduce a Poisson process and a pair of partial differential equations.
  - Probability that a given cell is in the normal state, P(t), is

Only Equation in Talk

SAS GLOBAL FORUM

Critique

1. “Saves” the empirical appearances, but lacks connection with biological or physical knowledge.
2. Poor content: it does not say much.
   - No attempt to say how the parameter may be related to properties of cell radiation
   - Nor how μ may be related to recovery characteristics of the cell
   - Does not risk any conjecture concerning the values of the parameters or even possible ranges, *a priori*
   - Parameters are estimated from the data, and than the conjecture is tested with the same data-circularity
   - No consequences of the model are deduced-when this is done the model is found to be qualitatively defective, experimental curves yield inflexion points, while the deduced consequences of the model show that it is incapable of yielding curves with points of inflection.

A Popperian Approcah

- Science proceeds from particular to universal
- No such thing as induction, only deduction
- Thus, in science one:
  1. Guesses the laws underlying our experiences
  2. Deduces their consequences
  3. Tests a suitable consequence
- Those guesses which survive become “laws”: they are not “proved” or “verified”—they are merely not yet falsified.
- Idea—we can refute a general statement by a single particular, we can never prove a general statement from any number of particulars.

A Popperian approach (continued)

Therefore

1. The prepared mind is the source of conjectures and hypotheses; we should not rely on data analysis for hypothesis formulation.
2. The proper role of statistics is in the enunciation of the scientist's conjectures, its translation into mathematical language, and the deduction of a variety of particular statistical hypotheses for attempted falsification.
3. Statistics should be a deductive tool.
4. Formulation of conjectures should not necessarily be tied to data, but data should be used to challenge discipline-based ideas, not to generate them.

A Popperian approach

- "The problem with the 'inductive outlook' is not so much that it is pretentious in claiming a new way of reasoning, but that it leads to the view that *hypotheses naturally emerge phoenix-like from data*. The inductivist outlook suggests that in any body of data there is 'information', and if only the right way of extracting it can be found then a hypothesis may be generated. This, in turn, motivates the devising of a diversity of computerized algorithms for processing large bodies of data, 'associations' being automatically produced every time the algorithm is run. The worst effects of such mechanical attempts at hypothesis generation are usually ameliorated by an intuitive good judgement in interpretation,but the techniques foster a research methodology hampered by ambiguous and weak conclusions." *Dolby G.R. (1982), "The role of statistics in the methodology of the life sciences", Biometrics, (38), 1969-1983.*

SAS GLOBAL FORUM

Some Observations at this point

1. One's views as to the appropriate statistical analyses for a given data set depends on an implicit philosophy of science.
2. Statisticians should develop methods that are congruent with their philosophy of science.
3. Statisticians should be concerned with philosophy. "I feel that statistics need philosophical thinking rather desperately…and…the philosophy of knowledge needs statistics." (Kempthrone, 1976). "One would not expect a book on scientific method to do the work of science itself…The purpose of an analytic or methodological study…is always indirect. It hopes to send others to their task *with clearer heads and less wasteful habits of investigation.* This necessitates a continual scrutiny of what these others are doing, or else analysis of meanings proceed in a vacuum." (Stevenson, 1959).

So??? … What does this imply about Business Analytics?

- Based on how the statistical/mathematical theories and developments have taken place, it is very reasonable to assume that Analytics is justified on a philosophy of science that is at least a weak form of instrumentalism.

SAS GLOBAL FORUM

A few Conclusions

- The ability to construct tuning (validation) and test data sets due to the "large data" environment in which we now live provides an ability to assess how well "models" predict.

- The goals of the business analyst may be (and often are) much different from that of the economic (or other) researcher. An instrumentalist view seems permissible, if not desirable, for "fast paced" decision making in business, especially for as it relates to "knowledge discovery."

- No philosophical theory of science for statistics (analysis) is without its problems or criticisms

Questions?

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.

SAS GLOBAL FORUM

SAS® GLOBAL FORUM

April 26-29
Dallas, TX