Forbes / Tech

MAY 28, 2013 @ 09:09 AM 264,133

A Very Short History Of Data Science

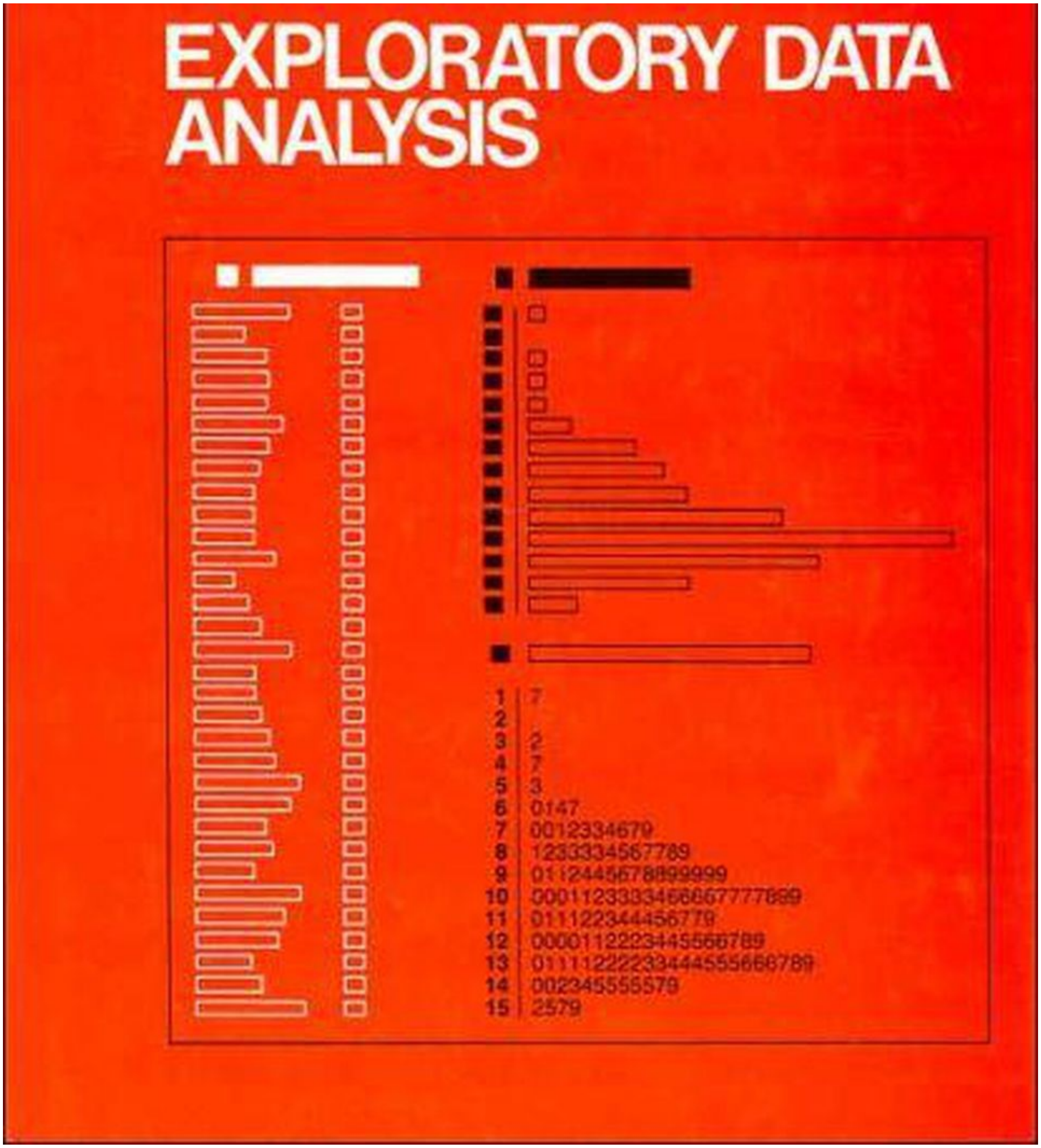
Gil Press, CONTRIBUTOR

I write about technology, entrepreneurs and innovation. **FULL BIO**

Opinions expressed by Forbes Contributors are their own.

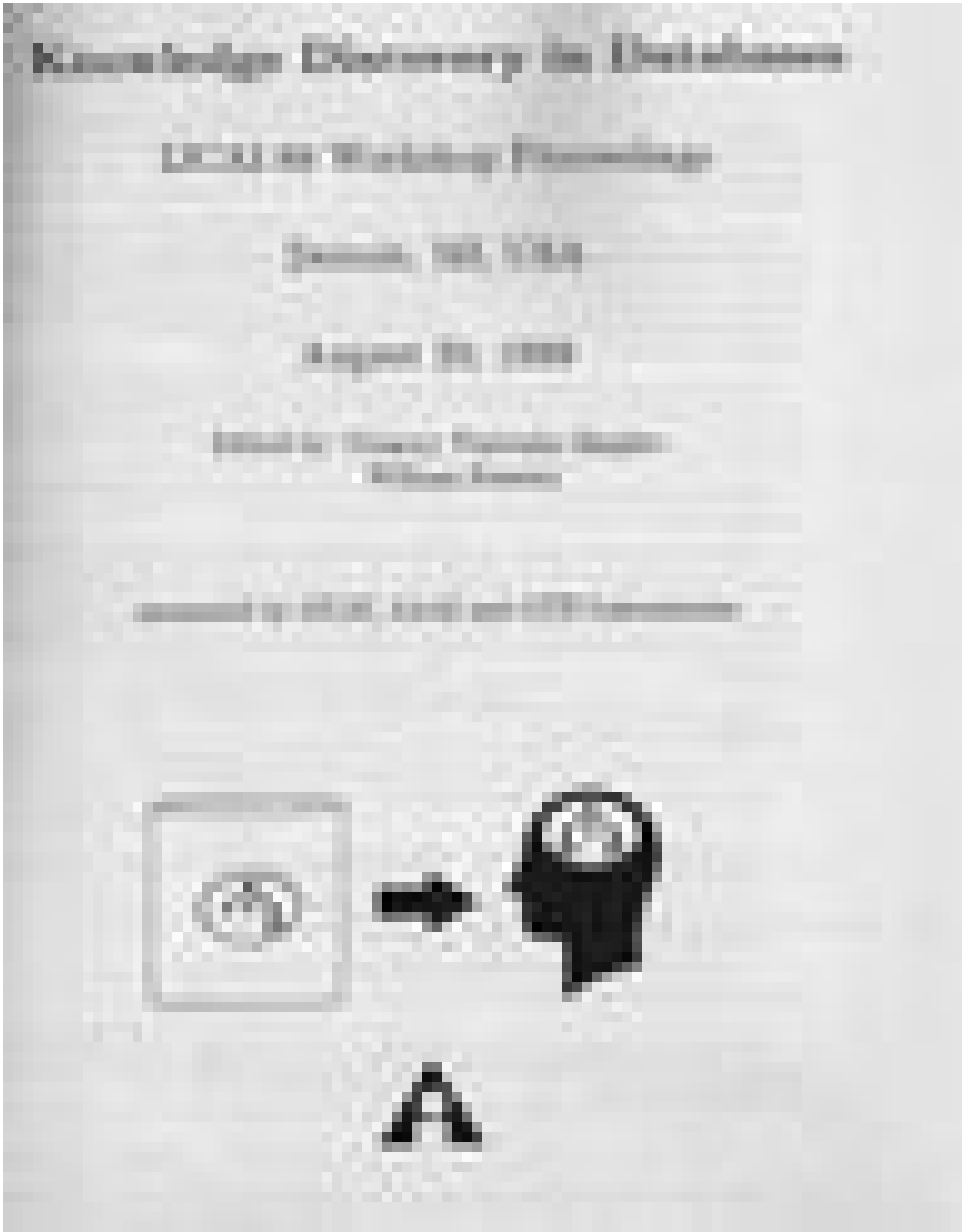
The story of how data scientists became sexy is mostly the story of the coupling of the mature discipline of statistics with a very young one--computer science. The term “Data Science” has emerged only recently to specifically designate a new profession that is expected to make sense of the vast stores of big data. But making sense of data has a long history and has been discussed by scientists, statisticians, librarians, computer scientists and others for years. The following timeline traces the evolution of the term “Data Science” and its use, attempts to define it, and related terms.

1962 John W. Tukey writes in “The Future of Data Analysis”: “For a long time I thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and doubt… I have come to feel that my central interest is in *data analysis*… Data analysis, and the parts of statistics which adhere to it, must…take on the characteristics of science rather than those of mathematics… data analysis is intrinsically an empirical science… How vital and how important… is the rise of the stored-program electronic computer? In many instances the answer may surprise many by being ‘important but not vital,’ although in others there is no doubt but what the computer has been ‘vital.’” In 1947, Tukey coined the term “bit” which Claude Shannon used in his 1948 paper “A Mathematical Theory of Communications.” In 1977, Tukey published *Exploratory Data Analysis*, arguing that more emphasis needed to be placed on using data to suggest hypotheses to test and that Exploratory Data Analysis and Confirmatory Data Analysis "can—and should—proceed side by side."



1974 Peter Naur publishes *Concise Survey of Computer Methods* in Sweden and the United States. The book is a survey of contemporary data processing methods that are used in a wide range of applications. It is organized around the concept of data as defined in the *IFIP Guide to Concepts and Terms in Data Processing*: "[Data is] a representation of facts or ideas in a formalized manner capable of being communicated or manipulated by some process." The Preface to the book tells the reader that a course plan was presented at the IFIP Congress in 1968, titled "Datalogy, the science of data and of data processes and its place in education," and that in the text of the book, "the term 'data science' has been used freely." Naur offers the following definition of data science: "The science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences."

1977 The International Association for Statistical Computing (IASC) is established as a Section of the ISI. "It is the mission of the IASC to link traditional statistical methodology, modern computer technology, and the knowledge of domain experts in order to convert data into information and knowledge."



Knowledge Discovery in Databases

IJCAI-89 Workshop Proceedings

Detroit, MI, USA

August 20, 1989

Edited by Gregory Piatetsky-Shapiro
William Frawley

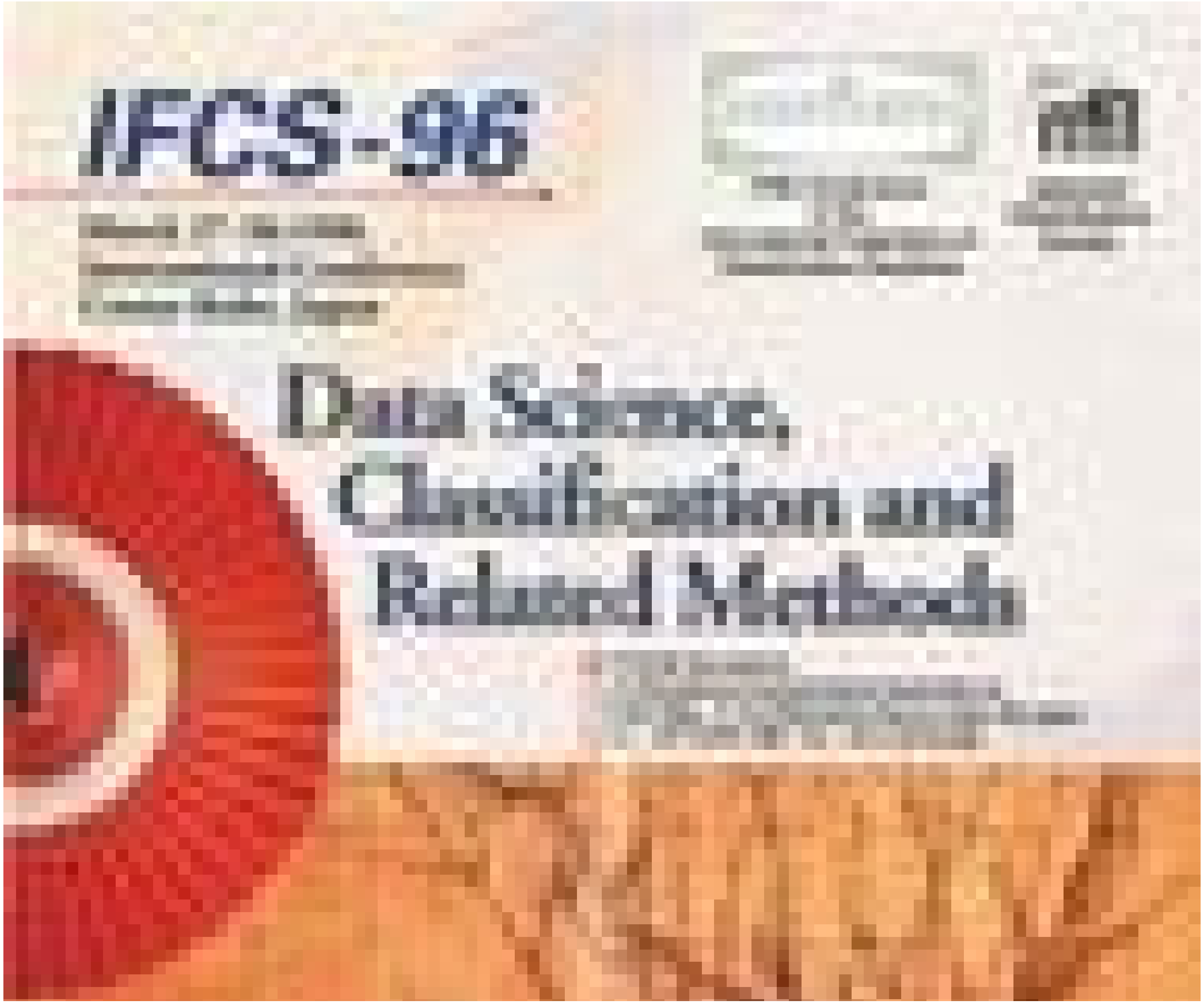
sponsored by IJCAI, AAAI and GTE Laboratories

1989 Gregory Piatetsky-Shapiro organizes and chairs the first Knowledge Discovery in

Databases (KDD) workshop. In 1995, it became the annual ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD).

September 1994 *BusinessWeek* publishes a cover story on "Database Marketing": "Companies are collecting mountains of information about you, crunching it to predict how likely you are to buy a product, and using that knowledge to craft a marketing message precisely calibrated to get you to do so… An earlier flush of enthusiasm prompted by the spread of checkout scanners in the 1980s ended in widespread disappointment: Many companies were too overwhelmed by the sheer quantity of data to do anything useful with the information… Still, many companies believe they have no choice but to brave the database-marketing frontier."

1996 Members of the *International Federation of Classification Societies (IFCS)* meet in Kobe, Japan, for their biennial conference. For the first time, the term "data science" is included in the title of the conference ("Data science, classification, and related methods"). The IFCS was founded in 1985 by six country- and language-specific classification societies, one of which, *The Classification Society*, was founded in 1964. The classification societies have variously used the terms data analysis, data mining, and data science in their publications.



1996 Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth publish "From Data Mining to Knowledge Discovery in Databases." They write: "Historically, the notion of finding useful patterns in data has been given a variety of names, including data mining, knowledge extraction, information discovery, information harvesting, data archeology, and data pattern processing… In our view, KDD [Knowledge Discovery in Databases] refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process. *Data mining* is the application of specific algorithms for extracting patterns from data… the additional steps in the KDD process, such as data preparation, data selection, data cleaning, incorporation of appropriate prior knowledge, and proper interpretation of the results of mining, are essential to ensure that useful knowledge is derived from the data. Blind application of data-mining methods (rightly criticized as data dredging in the statistical literature) can be a dangerous activity, easily leading to the discovery of meaningless and invalid patterns."

1997 In his inaugural lecture for the H. C. Carver Chair in Statistics at the University of Michigan, Professor C. F. Jeff Wu (currently at the Georgia Institute of Technology), calls for statistics to be renamed data science and statisticians to be renamed data scientists.

1997 The journal Data Mining and Knowledge Discovery is launched; the reversal of the order of the two terms in its title reflecting the ascendance of "data mining" as the more popular way to designate "extracting information from large databases."

December 1999 Jacob Zahavi is quoted in "Mining Data for Nuggets of Knowledge" in

Knowledge@Wharton: "Conventional statistical methods work well with small data sets. Today's databases, however, can involve millions of rows and scores of columns of data... Scalability is a huge issue in data mining. Another technical challenge is developing models that can do a better job analyzing data, detecting non-linear relationships and interaction between elements... Special data mining tools may have to be developed to address web-site decisions."

2001 William S. Cleveland publishes "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics." It is a plan “to enlarge the major areas of technical work of the field of statistics. Because the plan is ambitious and implies substantial change, the altered field will be called ‘data science.’" Cleveland puts the proposed new discipline in the context of computer science and the contemporary work in data mining: “…the benefit to the data analyst has been limited, because the knowledge among computer scientists about how to think of and approach the analysis of data is limited, just as the knowledge of computing environments by statisticians is limited. A merger of knowledge bases would produce a powerful force for innovation. This suggests that statisticians should look to computing for knowledge today just as data science looked to mathematics in the past. … departments of data science should contain faculty members who devote their careers to advances in computing with data and who form partnership with computer scientists.”

2001 Leo Breiman publishes “Statistical Modeling: The Two Cultures” (PDF): “There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.”

April 2002 Launch of *Data Science Journal*, publishing papers on “the management of data and databases in Science and Technology. The scope of the Journal includes descriptions of data systems, their publication on the internet, applications and legal issues.” The journal is published by the Committee on Data for Science and Technology (CODATA) of the International Council for Science (ICSU).

January 2003 Launch of *Journal of Data Science*: “By ‘Data Science’ we mean almost everything that has something to do with data: Collecting, analyzing, modeling…… yet the most important part is its applications--all sorts of applications. This journal is devoted to applications of statistical methods at large.... The *Journal of Data Science* will provide a platform for all data workers to present their views and exchange ideas."

May 2005 Thomas H. Davenport, Don Cohen, and Al Jacobson publish "Competing on Analytics," a Babson College Working Knowledge Research Center report, describing "the emergence of a new form of competition based on the extensive use of analytics, data, and fact-based decision making... Instead of competing on traditional factors, companies are beginning to employ statistical and quantitative analysis and predictive modeling as primary elements of competition. " The research is later published by Davenport in the *Harvard Business Review* (January 2006) and is expanded (with Jeanne G. Harris) into the book *Competing on Analytics: The New Science of Winning* (March 2007).

September 2005 The National Science Board publishes "Long-lived Digital Data Collections: Enabling Research and Education in the 21st Century." One of the recommendations of the report reads: "The NSF, working in partnership with collection managers and the community at large, should act to develop and mature the career path for data scientists and to ensure that the research enterprise includes a sufficient number of high-quality data scientists." The report defines data scientists as "the information and computer scientists, database and software engineers and programmers, disciplinary experts, curators and expert annotators, librarians, archivists, and others, who are crucial to the successful management of a digital data collection."

2007 The Research Center for Dataology and Data Science is established at Fudan University, Shanghai, China. In 2009, two of the center's researchers, Yangyong Zhu and Yun Xiong, publish "Introduction to Dataology and Data Science," in which they state "Different from natural science and social science, Dataology and Data Science takes data in cyberspace as its research object. It is a new science." The center holds annual symposiums on Dataology and Data Science.

July 2008 The JISC publishes the final report of a study it commissioned to "examine and make recommendations on the role and career development of data scientists and the associated supply of specialist data curation skills to the research community. " The study's final report, "The Skills, Role & Career Structure of Data Scientists & Curators: Assessment of Current Practice & Future Needs," defines data scientists as "people who work where the research is carried out--or, in the case of data centre personnel, in close collaboration with the creators of the data--and may be involved in creative enquiry and analysis, enabling others to work with digital data, and developments in data base technology.”

January 2009 *Harnessing the Power of Digital Data for Science and Society* is published. This report of the Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council states that “The nation needs to identify and promote the emergence of new disciplines and specialists expert in addressing the complex and dynamic challenges of digital preservation, sustained access, reuse and repurposing of data. Many disciplines are seeing the emergence of a new type of data science and management expert, accomplished in the computer, information, and data sciences arenas and in another domain science. These individuals are key to the current and future success of the scientific enterprise. However, these individuals often receive little recognition for their contributions and have limited career paths.”

January 2009 Hal Varian, Google’s Chief Economist, tells the *McKinsey Quarterly*: “I keep saying the sexy job in the next ten years will be statisticians. People think I’m joking, but who would’ve guessed that computer engineers would’ve been the sexy job of the 1990s? The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that’s going to be a hugely important skill in the next decades… Because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it… I do think those skills—of being able to access, understand, and communicate the insights you get from data analysis—are going to be extremely important. Managers need to be able to access and understand the data themselves.”