Essays on Data Science

Rafael C. Alvarado

9/20/23

Table of contents

Preface

Part I

The History of Data Science

Chapter 1

Abstract

Consensus on the definition of data science remains low despite the widespread establishment of academic programs in the field and continued demand for data scientists in industry. Definitions range from rebranded statistics to data-driven science to the science of data to simply the application of machine learning to so-called big data to solve real world problems. Current efforts to trace the history of the field in order to clarify its definition, such as Donoho's "50 Years of Data Science" (Donoho 2017), tend to focus on a short period when a small group of statisticians adopted the term in an unsuccessful attempt to rebrand their field in the face of the overshadowing effects of computational statistics and data mining. Using textual evidence from primary sources, this essay traces the history of the term to the early 1960s, when it was first used by the US Air Force in a surprisingly similar way to its current usage, to 2012, the year that *Harvard Business Review* published the enormously influential article "Data Scientist: The Sexiest Job of the 21^{st} Century" (Davenport and Patil 2012), even as the American Statistical Association acknowledged a profound "disconnect" between statistics and data science. Among the themes that emerge from this review are (1) a continuous and consistent meaning of data science as the practice of managing and processing scientific data from the 1960s to the present, (2) a long-standing opposition between data analysts and data miners that has animated the field since the 1980s, and (3) the phenomenon of "data impedance"—the disproportion between surplus data, indexed by phrases like "data deluge" and "big data," and the limitations of computational machinery and methods to process them. This persistent condition appears to have motivated the use of the term and the field itself since its beginnings.

Chapter 2

Synopsis

Historically, there are two primary usages of “data science”: the classical and the statistical. Today, these two usages have collided in the context of the academy where a new form of data science is being forged. This process can only be enhanced by an understanding of the contexts and motivations behind these two usages.

The classical usage began in the 1960s in the US military-industrial sector and continues to this day in that area and in the data-driven sciences. Among various public and private organizations that adopted the term at that time, the Data Sciences Lab of the Air Force Cambridge Research Laboratories formed in 1963 stands out(AFCRL 1963). This lab established a paradigm that grew out of the need to dynamically process vast amounts of sensor data—dubbed “data deluge” since the 1950s—in real-time to support decision-making and modeling of complex phenomena. This lab pioneered the use of artificial intelligence and computer visualization to extract patterns from new forms of data. Over time this usage evolved to include working with large, complex, and dynamic data sets in a variety of scientific fields, including physics, environmental science, and biology.

A secondary development of the classical usage began in Silicon Valley around 2008 and was famously promoted by *Harvard Business Review* in 2012 when it called data scientist the “sexiest job of the 21st century” (Davenport and Patil 2012). In this usage, the paradigm of classical data science jumped from the military-industrial and scientific sectors to the commercial sector of so-called surveillance capitalism (Zuboff 2019). In this context the usage became wildly popular, prompting a high demand for data scientists in industry as well as for authoritative explanations of the nature of the role. This in turn motivated the formation of numerous degree programs throughout the academy in the US, typically in the form of masters degrees. It also elicited a strong reaction from academic statisticians alarmed by the perceived disconnect between data science and their own field (Davidian 2013).

The statistical usage began in Japan in the early 1990s among academic statisticians concerned with the threats and opportunities associated with computational statistics, data mining, and other developments relating to the rise of available computing and surplus data. Chikio Hayashi coined this usage in 1992(Ohsumi 2002); by 1996 the International Federation of Classification Societies (IFCS) adopted it and began a long engagement with data science that continues to this day in Europe and Japan(Hayashi 1998a). A parallel but shorter-lived version of this usage emerged in the US between 1996 and 2001 when a small group of academic statisticians unsuccessfully exhorted the field to brand itself as data science and embrace new developments in computational statistics and machine learning (Kettenring 1997b; C. F. J. Wu 1997; Cleveland 2001). A third iteration of this usage was developed by traditional statisticians who viewed the rise of second-wave classical data science as a threat to their field that must be "owned"(Rodriguez 2012; Davidian 2013; Yu 2014; "ASA Statement on the Role of Statistics in Data Science" 2015; Donoho 2017).

The classical usage developed in the context of an historically unique assemblage of networked data-generating and computational machinery that characterized post-war military and scientific endeavors. The primary problem faced in this context is what we can call data impedance—the endemic disproportion between the production of data by signal-generating instruments—from satellites and particle accelerators to smart phones and credit cards—capturing a wide range of natural and behavioral phenomena, and the ability to process and interpret these data, often rapidly, by means of computational machinery. In this context, data science emerged as an eclectic form of expertise associated with the pipeline of activities that begins with the acquisition of data, often from non-experimental sources, to its reduction and modeling, to its visualization and interpretation.

By contrast, the statistical usage developed in the context of an established academic field reacting to the very developments within which classical data science emerged. Each iteration of this usage was motivated by the perceived threats and opportunities opened by computational methods and the growing surplus of data being captured by databases and shared over networks. A recurrent theme in this usage is an attraction to machine learning methods developed primarily by computer scientists and a repulsion to data mining, perceived as an unprincipled collection of methods lacking in experimental design and mathematical grounding. In this context, data science emerged as a rebranded form of statistics that would encompass and enhance the good parts of the new computational methods while throwing away the bad.

Chapter 3

Introduction

The interests of data scientists—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—lie in having their creativity and intellectual contributions fully recognized.

National Science Board, "Long-Lived Digital Collections: Enabling Research and Education in the 21st Century" (Simberloff et al. 2005: 27).

Data science today is characterized by a paradox. The large number and rapid growth of job opportunities and academic programs associated with the field over the past decade suggest that it has matured into an established field with a recognizable body of knowledge. Yet consensus on the definition of data science remains low. Members and observers of the field possess widely variant understandings of data science, resulting in divergent expectations of the knowledge, skill sets, and abilities required by data scientists. Definitions, when they are not laundry lists, range from a rebranded version of statistics to data-driven science to the science of data to simply the application of machine learning to so-called big data to solve real world problems. These differences cannot be reduced to so-called semantics; they reflect a range of deep-seated institutional commitments and values, as well as variant understandings about the nature of knowledge and science. The lack of shared understanding poses a significant problem for academic programs in data science: it inhibits the development of standards and a professional community, confounds the allocation of resources, and threatens to undermine the authority and long-term prospects of these programs.

This essay approaches the problem of defining data science by describing how the collocation "data science," and its grammatical variants "data sciences" and "data scientist," have been used historically.[1] The primary method employed is the close reading and precise seriation of textual evidence drawn from a representative collection of primary sources, including organizational reports, academic articles, news stories, advertisements, and other contemporary forms of evidence. These are used to trace the history of the term's social and institutional contexts of use as well as its denotative and connotative meanings. Extensive extracts are often presented, rather than paraphrased, as these in many cases provide the reader with direct and illuminating evidence for the meanings in question.[2]

This historiography is presented as a series of decades in which the term takes on a new meaning, beginning with its initial usage in the 1960s and ending around 2012, when the phrase becomes a commonplace. It is shown that the phrase has a continuous and consistent usage throughout this history. As usage of the phrase evolved, its meanings were always additions to and inflections on prior meanings; in no case did newer usages completely contradict what preceded them, nor did they appear as cases of random independent invention.

The result is a picture of the transformation of a semantic complex that indexes a consistent set of technical, social, and cultural realities that constitute what may be called the situation of data science, a situation that motivates the writing of this essay. Anticipating follow-up research to this essay, this situation has been described repeatedly by data scientists of all stripes as a kind of data processing *pipeline*, a sequence of operations that begins with the consumption of data and ends with the production of data products, ranging from research results and visualizations to software services employed by various sectors of society.

[1] In this essay, a collocation is defined as a combination of two or more words that function as a lexical unit. In contrast to a mere n-gram, its usage tends to be idiomatic and non-random. Etymologically, the usage of a collocation often begins as a marked construction, by means of quotes and hyphens, before eventually becoming idiomatic. Often, a collocation becomes so common that it becomes a single word. For example, the word "database" began as "data base" and "data-base" before evolving into its current form (after beating out "data bank"). Throughout this essay, the collocation "data science" is referred to as a term or phrase, reflecting its unitary semantic status.

[2] The arguments and observations made by the authors in each case are represented in historical tense, not the textual present, which is the usual custom in writing about the history of ideas. For example, instead of saying that "Tukey argues P" in an essay from the 1960s, the evidence is presented as "Tukey argued P." This is done in order to ground the evidence in its social and historical setting.

Chapter 4

The 1960s

4.1 First Uses

The first uses of the phrase "data science," as a recognizable collocation, appeared in the early 1960s in both singular and plural forms. Two main uses are found in the written record almost simultaneously, one in a military context, the other industrial. In both cases, the phrase functioned as an organizational rubric for a new kind of labor associated with the rise of large-scale data generating and processing technologies that were the hallmark of the postwar era.

The military use first appeared in a series of reports covering the period from July 1962 to June 1970 on research carried out by the Data Sciences Laboratory (DSL). The DSL was founded in 1963 as one of several labs associated with the US Air Force Cambridge Research Laboratories (AFCRL).[1] These reports do not provide an explicit definition of data science or a rationale for choosing the expression over others, but its meaning is clear from context. Consider the stated motivation for the lab—which, as one of the first attested uses of the phrase, is worth quoting at length:

> The most striking common factor in the advances of the major technologies during the past fifteen years [i.e. since WWII] is the increased use and exchange of information. *Modern data processing and computing machinery, together with improved communications*, has made it possible to ask for, collect, process and use *astronomical amounts* of detailed data. …
>
> But in the face of this progress there is impatience with *the limitations of existing machines.* …

[1] The DSL was formed by combining the Computer and Mathematical Sciences Laboratory and the Communications Sciences Laboratory in the 1963 reorganization of the AFCRL (Venkateswaran 1963: 628). Within the AFCRL, the lab was noted for its "research on speech patterns [which] dated back to the 1940's [sic]" (Altshuler 2013: 27-28).

> A large number of military systems—for example, those concerned with surveillance and warning, command and control, or weather prediction—deal in *highly perishable information.* Few existing computers are capable of handling this information in "real-time"—that is, processing the data as they come in. Higher speed is one way to a solution. But increased speed will not overcome fundamental shortcomings of existing computers. These shortcomings arise from the fact that existing machines, having essentially evolved as numerical calculators, are not always optimally organized to perform the tasks they are called upon to do. …
>
> … *A considerable amount of the data to be processed is not numerical.* It is in audio or visual form. Immense amounts of visual data—for example, TIROS satellite pictures or bubble chamber pictures of atomic processes—remain unevaluated for lack of processing capability. In part this is due to the fact that, from the data processing point of view, the information content of pictorial inputs is highly redundant, *demanding excessive channel capacity* in transmission and compelling processing machinery to handle vast amounts of meaningless or non-essential information. Similar considerations prevail for speech. …
>
> In real-life situations *data are almost never available in unadulterated form*, but are usually distorted or masked by spurious signals. Examples are seismic data, radio propagation measurements, radar and infrared surveillance data and bioelectric signals. …
>
> An increasing amount of *data processing research* is aimed at the creation of machines or machine programs that incorporate features of *deductive and inductive reasoning, learning, adaptation, hypothesis formation and recognition.* Such features are commonly associated with human thought processes and, when incorporated in machines, are frequently termed "artificial intelligence." Artificial intelligence is of utmost importance in decision situations where not all possible future events can be foreseen [AFCRL (1963); emphasis added].

The two later reports are more succinct:

> The program of the Data Sciences Laboratory centers on the processing, transmission, display and use of information. Implicit in this program statement is an emphasis on computer technology (AFCRL 1967: 13).
>
> Broadly defined, the program of the Data Sciences Laboratory involves the automatic processing, filtering, interpretation and transmission of information (AFCRL 1970: 318).

Based on these excerpts alone, one could be forgiven for inferring that data science was invented by the US Air Force around 1963 with the formation of the

DSL. Most of the elements currently considered central to the field were brought together there: a concern for processing what is later called “big data,” clearly defined in terms of volume, velocity, and variety (and volatility); a recognition of the fundamental messiness of data; and a focus on artificial intelligence as an essential approach to extract value from such data. The lab produced significant research on pattern recognition and classification, machine learning, neural networks, and spoken language processing in the service of processing the novel forms of data described above.

More important than locating a precise time and place for the origin of the field—a task doomed to fail, given the complexity and multithreaded nature of historical phenomena—is the work of describing the historical situation within which the phrase data science developed and which it indexes. A clue to this context is the repeated emphasis on data and information processing we find in the DSL’s descriptions of its work. Specifically, the phrases “the data processing point of view” and “data processing research” index a set of military projects and concerns associated with the early Cold War.

The AFCRL was originally established in 1945 as the Cambridge Field Station, a unit created to hold onto the Harvard and MIT scientists and engineers who performed significant research on radar and electronics in WWII. During the 1950s, the lab focused on Project Lincoln, which led to the creation of the Semi-Automatic Ground Environment (SAGE), a real-time command-and-control system developed to counter to perceived threat of an airborne nuclear attack by the Soviet Union. As a continental air defense system, SAGE was designed to collect, analyze, and relay data from a vast array of geographically distributed radars in real-time, in order to initiate an effective response to an aerial attack. At the heart of the system was a network of large digital computers that coordinated the data retrieved from the radar sites over phone lines and processed them to produce a single unified image—literally displayed on a monitor—of the airspace over a wide area.

Although responsibility for research on such military surveillance systems was moved out of the lab in 1961, just before the Data Sciences Lab was formed, it is plausible that the SAGE project influenced the mission of the lab by providing a concrete paradigm for a new kind of information processing situation. This was the situation of using of advanced computational machinery and state-of-the-art data reduction and pattern recognizing methods to process vast amounts of real-time signal data, coming from geographically distributed radars and satellites, in order to represent a complex space of operations and guide making decisions about how to operate in that space. The paradigm was also applied to the problem of weather forecasting and other geophysical domains. (If we replace radars with smart phones and the Internet of Things, it is not difficult to draw a parallel between this arrangement and that of social media corporations today.)

Evidence for the influence of radar and satellite-based real-time command and control systems on the conceptualization of data science may be found in the idioms we currently associate with the field, such as the use of “signal and noise” to refer to the presence and absence of statistically significant patterns and the use of Receiver Operator Characteristic (ROC) curves—first used by military radar operators in 1941—to measure the performance of binary classifiers. Other idioms, such as "data deluge," also emerge in this context. A history of the expression data deluge is worth its own study, but it is clear that its provenance was the situation described above. The term gained currency in the 1960s in reference to satellite data collected by NASA and the military. Consider this passage from the NASA publication *Scientific Satellites*:

> The data deluge, information flood, or whatever you choose to call it, is hard to measure in common terms. An Observatory-class satellite may spew out more than 10^{11} data words during its lifetime, the equivalent of several hundred thousand books. Data-rate projections, summed for all scientific satellites, prophesy hundreds of millions of words per day descending on Earth-based data processing centers. These data must be translated to a common language, or at least a language widely understood by computers (viz, PCM), then edited, cataloged, indexed, archived, and made available to the scientific community upon demand. Obviously, the vaunted information explosion is not only confined to technical reports alone, but also to the data from which they are written. In fact, the quantity of raw data generally exceeds the length of the resulting paper by many orders of magnitude (Corliss 1967: 157).[2]

Work on such projects generated an enormous amount of research on the problems arising from the processing and interpreting data. In the preceding text, the author describes this work in some detail, specifying a series of stages in which data are transformed into a form suitable for scientific analysis. We would recognize this work today as data wrangling. It is reasonable to infer that the concept of data science emerged to designate this kind of work, which, in any case, is consistent with the published mission of the Data Sciences Lab.

Prior to the formation of the DSL, the phrase "data-processing scientist" was in use to designate the work involved in data reduction centers, such as the one built at the Langley Aeronautical Lab in Virginia to process the enormous amounts of data generated by wind tunnel experiments and other sources associated with the nascent space program. Data reduction was essential to projects like SAGE, in which vast amounts of real-time signal data had to be reduced prior to analysis. In a House appropriations hearing in 1958, the following description of this kind of work was provided by Dr. James H. Doolittle,[3] the last chairman of NACA before it became NASA:

> The data processing function is much more complex than the mere production line job of translating raw data into usable form. Each

[2] Preceding the usage of data deluge and in a wider context is "information explosion." Both expressions conjure images of disaster and have been remarkably persistent up to the present era. Only with the coining of "big data" have they been displaced by a more positive term.

[3] This is the very same General Doolittle of Doolittle's Raid.

> new research project must be reviewed to determine how the data will be obtained, what type and volume of calculations are required, and what modifications must be made to the recording instruments and data-processing apparatus to meet the requirements. *It may even be necessary for the data-processing scientist to design and construct new equipment for a new type of problem.* Some projects cannot be undertaken until the specific means of obtaining and handling the data have been worked out. In some research areas, on-line service to a data processing center saves considerable time by allowing the project engineer to obtain a spot check on the computed results while the facility is in operation. This permits him to make an immediate change in the test conditions to obtain the results that he wants [U. S. C. H. Appropriations (1958); emphases added].

4.2 Impedance

The kind of work conducted by NASA and the Air Force in this period provides a context for understanding the meaning of data science when the phrase first appeared. In this context, data science designated a kind of research focused on what we may call the *impedance* that arises from the ever-growing requirements of data produced by an expanding array of signal generating technologies (e.g. radars and satellites), scientific instruments, and reports on one hand, and the limited capacity of computational machinery to process these on the other. It is concerned specifically with the development of computational methods and tools to handle the problems and harness the opportunities posed by surplus data. In this context, *data science is the science of processing and extracting value from data by means of computation.* Although the specific technologies have changed continually, the condition of data impedance, the disproportion between data abundance and computational scarcity relative to the need to extract value from the data, has been constant since this time, and defines the condition that gives rise to data science in this sense.

This interpretation of the meaning of data science is corroborated by other contemporary usages. A report on a US Department of Defense program to define standards "to interchange data among independent data systems" refers to a "Data Science Task Group" established in 1966 "to formulate views of data and definitions of data terms that would meet the needs of the program" (Crawford, Jr. 1974: 51). Crawford, a fellow student of Claude Shannon at MIT under Vannevar Bush, was affiliated with IBM's Advanced Systems Development Division, a group that had developed optical scanners to recognize handwritten numbers in 1964. In addition, the term appeared in the trademarked name of at least two corporations in the United States: Data Science Corporation, formed in 1962 by a former IBM employee ("Robert Allen Obituary (2014) - St. Louis, MO - St. Louis Post-Dispatch," n.d.), and Mohawk Data Sciences, founded in 1964 by a three former UNIVAC engineers ("Mohawk Data Sciences" 1966).

Both companies provided data processing services and lasted well into the era of personal computing. In the late 1960s and 1970s, many other companies used term as well, such as Data Science Ventures (Mort Collins Ventures, n.d.) and Carroll Data Science Corporation (Office 1979).[4]

4.3 Meaning of "data" and the information crisis

Let us consider the meaning and significance of the word "data" in these examples, especially given the DOD's concern to define it, as a clue for the motivation of the term "data science" when other candidates, such as computer science and information science, might have sufficed at the time. The choice of the term appears to be motivated by a concern to define and understand *data* itself as an object of study, a surprisingly opaque concept that is thrown into sharp relief in the context of getting computers to do the hard work of processing information in the context of impedance, as a result of their commercialization and widespread use in science, industry, and government. Thus although the term "data" has a long history—deriving from the Latin word for that which is *given* in the epistemological sense, either through the senses, reason, or authority—in this context it refers to the structured and discrete representation of information sources so that these may be processed by computers. In other words, *data is machine readable information.*[5] It follows that the data sciences in this period are concerned with understanding machine readable information, in terms of how to represent it and how to process it in order to extract value.

Further evidence of this concern for what might be called the information crisis in scientific research—and for the idea that the solution to this crisis hinges on refining the concept of data—can be found in the formation of the International Council for Science (ICSU) Committee on Data for Science and Technology (CODATA) in 1966. This organization was established by an international group of physicists alarmed that the "deluge of data was swamping the traditional publication and retrieval mechanisms," and that this posed "a danger that much of it would be lost to future generations" (Lide and Wood 2012). Importantly, CODATA still exists and currently identifies itself with the field of data science. In 2001 it launched the *Data Science Journal,* focused on "the management, dissemination, use and reuse of research data and databases across all research domains, including science, technology, the humanities and the arts" ("Data Science Journal," n.d.). Aware that the definition of the field had changed significantly since its founding, the journal provided the following clarification in 2014:

> We primarily want to *specify* our definition of "data science" as the

[4] This continues into the 1980s, with Gateway Data Sciences Corp and Vertex Data Science, Ltd.

[5] Preceding the usage of data deluge and in a wider context is "information explosion." Both expressions conjure images of disaster and have been remarkably persistent up to the present era. Only with the coining of "big data" have they been displaced by a more positive term.

> classic sense of the science of data practices that advance human understanding and knowledge—the evidence-based study of the socio-technical developments and transformations that affect science policy; the conduct and methods of research; and the data systems, standards, and infrastructure that are integral to research.
>
> We recognize the contemporary emphasis on data science, which is more concerned with data analytics, statistics, and inference. We embrace this new definition but seek papers that focus specifically on the data concerns of an application in analytics, machine learning, cybersecurity or what have you. We continue to seek papers addressing data stewardship, representation, re-use, policy, education etc.
>
> Most importantly, we seek broad lessons on the science of data. Contributors should generalize the significance of their contribution and demonstrate how their work has wide significance or application beyond their core study [Parsons (2019); emphasis in original].

This retrospective definition supports the idea that data science in the 1960s—which we may call, following this note, classical data science—was concerned with understanding data practices, where data is understood to be a universal medium into which information in a variety of native forms, from scientific essays to radio signals from outer space, must be encoded so that it may be shared and processed by computational machinery. Data science as “the science of data practices that advance human understanding and knowledge” is concerned with defining and inventing this medium, its structure and function.

4.4 A Note on Tukey

Tukey’s famous essay on data analysis, which appears during the same time period, touches on some of the drivers noted here, such as the high volume and spottiness of real data and the impact of the computer, but from the perspective of advanced mathematical statistics (Tukey 1962). One difference between his view and that adopted by the AFCRL is of interest here: whereas Tukey appears to have regarded the computer as a more or less fixed technology, replaceable in many tasks by “pen, paper, and slide rule” but irreplaceable (he conceded) in others, the Data Sciences Lab viewed the computer as a fluid technology, one that needed to be pushed beyond its original design envelope as a numerical calculator. In fact, the AFCRL and similar groups appear to have provided the impetus to move computer science beyond a concern for abstract algorithms and to include the study of data structures and technologies, specifically databases. It is, as we shall see, a difference that continues to underlie current disputes over the meaning and value of data science.

4.5 The 1970s

4.5.1 Data Science ...

It is clear that by the early 1970s the term data science had been in circulation in several contexts and referred to ideas and tools relating to computational data processing. Importantly, these usages were not obscure—the AFCRL was one of the premier research laboratories in the world and closely connected with Harvard and MIT (Altshuler 2013), an international cross-roads of intellectual life where many would have come into contact with the term. Similarly, IBM and UNIVAC, the sources of the founders of two self-proclaimed data science companies, were the two largest computer manufacturers at the time.[6]

4.5.2 Naur's *Datalogi*

Although the AFCRL closed the Data Sciences Lab by 1970,[7] the term continued to be used, most notably by the Danish computer scientist Peter Naur, who suggested that computer science, a relatively new field, be renamed to data science. His argument, consistent with previous usage, was that computer science is fundamentally concerned with data processing and not mere computation, i.e. what the AFCRL derided as numerical calculation. Earlier, in the 1960s, Naur had coined the term "datalogy" (Danish: *datalogi*) for this purpose, but later found the term data science to be a suitable synonym, perhaps due to its currency or to his familiarity with the DSL, which shared his research interest in developing programming languages (Naur 1966, 1968). In contrast to the AFCRL, Naur provided an explicit definition of data science:

> The starting point is the concept of *data*, as defined in [0.7]: DATA: *A representation of facts or ideas in a formalized manner capable of being communicated or manipulated by some process.* Data science is the science of dealing with data, once they have been established, while the relation of data to what they represent is delegated to other fields and sciences.
>
> The usefulness of data and data processes derives from their appli-

[6] As evidence for the visibility of the AFCRL as this time, here is an add that appeared in a 1964 issue of the British weekly *Nature*:

A SYMPOSIUM on "Models for the Perception of Speech and Visual Form", sponsored by the Data Sciences Laboratory of the Air Force Cambridge Research Laboratory, will be held in Boston during November 11-14. Further information can be obtained from Mr. G. A. Cushman, Wentworth Institute, 550 Huntington Avenue, Boston, Massachusetts 02115 ("Announcements" 1964).

[7] Altshuler writes that the lab was "abolished" in June 1972 "in response to a large reduction in manpower authorizations" (Altshuler 2013: 27). However, the unit is not mentioned in the July 1970 to June 1972 research report (AFCRL 1973). The lab's closure may have been a consequence of the passing of the Mansfield Amendment in 1969, which prohibited the military from carrying out "any research project or study unless such project or study has a direct and apparent relationship to a specific military function" (U. S. C. H. C. on Appropriations 1970: 348).

> cation in building and handling models of reality.
>
> ...
>
> A basic principle of data science is this: The data representation must be chosen with due regard to the transformation to be achieved and the data processing tools available. This stresses the importance of concern for the characteristics of the data processing tools.
>
> Limits on what may be achieved by data processing may arise both from the difficulty of establishing data that represent a field of interest in a relevant manner, and from the difficulty of formulating the data processing needed. Some of the difficulty of understanding these limits is caused by the ease with which certain data processing tasks are performed by humans [Naur (1974): 30-31; emphasis and citation in original; the reference "0.7" refers to Gould (1971)].

Clearly, Naur's definition inherits the classical definition described above; it locates the meaning of the term in the series of practices associated with the larger activity of data processing. These practices include establishment, choice of representation, conversion and transformation, the modeling of reality, and the guiding of human actions. One difference is that Naur is keen to locate data science within a division of labor implied by this general process, separating data science *per se* from the work of data acquisition (establishment) and the domain knowledge required to acquire data effectively. In this view, data science is more specifically concerned with the formal representation of data (i.e. with data structures and models), a practice that must be done in light of how data are to be transformed downstream, and with which tools (i.e. algorithms and programming languages). As we shall see, the weighting that Naur assigns to this kind of work is not inherited by later theorists. However, the general image of a sequential process with distinct phases in the life cycle of data is. Here we see the appearance of the image of a pipeline, unnamed but implied by the concept of *process*, which dominates the mental representation of the field from its origins in the 1960s.

Far from being a fluke, Naur's usage developed the classical definition of data science initiated by NASA and the Air Force, intentionally or not. The fact that his attempt to rename computer science failed outside of his native country (and Sweden) is not important; his understanding of computer science sheds light on how closely the concept of data was (and is) related to computation and process.

It is worth noting that Naur's definition implies a familiarity with the real-world provenance of data processing in industry and government. Indeed, by this time computational data processing had penetrated all sectors of society, and the pressure to improve tools and methods to represent and process data had increased as well. As a result of this pressure, two important data standards were developed in this period: Codd's relational model, which laid the foundation for SQL and commercially viable relational databases in the 1980s, and Goldfarb's SGML, which would become a standard for encoding so-called unstructured textual data (such as legal documents) and later the basis for HTML and XML Goldfarb (1970). This focus on the human context of data processing is reflected in his later work; a volume of selections of his writing from 1951 to 1990, which includes his essay on data science, is entitled "Computing: A Human Activity" (Naur 1992).

4.5.3 Other Uses

- (*Computers in Education: Proceedings of the IFIP ... World Conference* 1975) Computers in Education, Proceedings of the IFIP World Conference
  - "The case is easily obscured, alas, by an instinct to confuse mathematics and data science" (p. 756)
  - "Recommending data science and data technology as areas of intellectual endeavor, and data studies as a subject for school learning;" (p. 758)

4.6 The 1980s

Chapter 5

The 1990s: Statistical Data Science

Interestingly, in 1977 a prefixed variant of the term appeared in the title of the technical report, “Non-Parametric Statistical Data Science: A Unified Approach Based on Density Estimation and Testing for ‘White Noise’ ” (Parzen 1977). However, Parzen later published a version of this work as “Nonparametric Statistical Data Modeling” (Parzen 1979), which indicates that his original word choice was mistaken. Yet the original choice may not have been entirely unmotivated: Parzen’s work attempted to unify parametric and non-parametric methods under one umbrella. Given the natural inclination of mathematical statistics for the former and data analysis for the latter, his choice of the term data science may have signaled an attempt to encompass both approaches to data. It is also worth noting that he later used to the term to introduce a “new culture of statistical science called LP MIXED DATA SCIENCE” (Parzen and Mukhopadhyay 2013), after the term became popular.[1] Whether or not this unifying goal was his motivation, statisticians later became quite interested in the term for precisely this reason.

5.1 1994: The Tokyo School

5.1.1 Ohsumi and Meta-Stat

In the early 1990s, the term resurfaced in the context of statistics. It appeared in the title of a 1994 essay by the Japanese statistician Noburu Ohsumi on the application of hypermedia to the problem of organizing data, “New Data

[1] Parzen’s use of the term “culture” here echoes his comments on Breiman’s famous essay on two cultures of statistical models, where he suggested that there are in fact several cultures, including his own, to which he devoted the majority of his response (Breiman 2001: 224–226).

and New Tools: A Hypermedia Environment for Navigating Statistical Knowledge in Data Science," an elaboration of an essay published two years earlier (Ohsumi 1992, 1994).[2] In these essays, Ohsumi described the by now familiar litany of problems associated with data impedance, although this time the focus was on the production of data resulting from its analysis and storage, not its consumption in so-called raw form:

> In research organizations handling statistical information, the volume of stored information resources, including research results, materials, and software, is increasing to the point that conventional separate databases and information management systems have become insufficient to deal with the amount. Increasing diversification in the media used these days interferes with the rapid retrieval and use of the information needed by users. A new system that realizes a presentation environment based on new concepts is needed to inform potential users of the value and effectiveness of using the vast amount of diverse data (Ohsumi 1992: 375).

For research facilities around the world, the products of classical data science—the database and data processing software—had become a sorcerer's apprentice, creating new problems with each solution. Organizations were drowning in the data sets they produced or acquired, the software used to process them, the print and digital libraries of reports and articles resulting from their analyses, and a host of other materials. The requirements, approach, and design goals of Ohsumi's proposed system, the Meta-Stat Navigator, are strikingly similar to those of a contemporary system designed to solve the information problems of another scientific organization: Berners-Lee's World Wide Web, famously developed at CERN in 1989 (Berners-Lee and Fischetti 2008). Of course, the latter quickly obviated Ohsumi's proposal and become synonymous with the Internet, invented decades earlier.

The significance of Meta-Stat for our purposes is that this kind of work was understood clearly as data science at this point in history. Data science continued to be connected with the processing and representation of data, and was distinct from data analysis, but with this important development: statisticians had become embedded in these technologies, and their work had changed significantly as a result. And, as a result of this change in working conditions, the connection between data analysis and data science became closer.

Here we may locate with some precision a crucial transformation in the meaning of the term, associated with its adoption by a new set of users. One clue to this change is the opportunity Ohsumi observed amid the challenges posed by data deluge:

[2] According to Ohsumi, "the term 'data science' appeared for the first time" in 1992, at a research exchange meeting between French and Japanese data analysts (so-called) at Montpelier University II in France (Ohsumi 2000: 331). He also claims to have "argued the urgency of the need to grasp the concept 'data science' " in 1992 (329).

> … the information handled by the statistical sciences lies on the boundaries of various other sciences and clarifies the relationships and nature of information that joins these sciences. Development of a system that fully organizes and integrates strategic information is essential (Ohsumi 1992: 375).

The Meta-Stat "system" was designed to realize the opportunity opened up by the central position statisticians had come to occupy among the prolifically data-generating sciences and the computational environment in which these data were made available. Data science, in this view, is *meta-statistics*, an encompassing concern for understanding data, understood as a universal medium, and its relationship to knowledge.

This perspective was shared by Ohsumi's senior compatriot and fellow statistician, Chikio Hayashi, whom Ohsumi described as "the pioneer and founder of data science" (Ohsumi 2004: 1).[3]

5.1.2 Hayashi's Usage

5.1.2.1 1993

In 1993, at a round table discussion during the fourth conference of the International Federation of Classification Societies held in Paris (IFCS-93), Hayashi uttered the phrase "Data Science" and was then asked to explain it. At the next conference (IFCS-96), he presented an answer, in addition to having the conference named to emphasize the importance of the term—"Data Science, Classification, and Related Methods." His definition is as follows:

[3] In his eulogy for Professor Hayashi, Ohsumi explained in some detail the motivation for his mentor's first usage of data science:

> In the last ten years of his life, Professor Hayashi asserted the importance of "data science as a theory of scientific methods." The starting point was the meaning of the words "data science;" this arose in 1992, when the professor was discussing the titles and introductions of a collection of papers to be published at the 2nd Japanese-French Scientific Seminar. *The professor used a Japanese phrase that translates as "data science" and for the rest of his life explored the concept behind this phrase.* When the papers at the seminar were published, it was proposed for the first time that the term referring to quantification be standardized as "quantification method" and the subsequent English terms were standardized as such. I can recall the professor reprimanding us with the words "you are talking about different concepts" when we referred to "deta kagaku" while another group called it "deta no kagaku" (data science).
>
> The basis of data science is an extremely straightforward concept; the fact that it needed to be expressed is an indication of the chaos that exists today in statistical data analysis and the deteriorating research environment. I am certain that the professor had intended to lead the way in achieving a breakthrough on these problems [Ohsumi (2002): 5; emphasis added].

5.1.1.1

> Data science is not only a synthetic concept to unify [mathematical] statistics, data analysis and their related methods but also comprises their results. It includes three phases, design for data, collection of data, and analysis on data. Data science intends to analyze and understand actual phenomena with "data." In other words, the aim of data science is to reveal the features of the hidden structure of complicated natural, human and social phenomena with data from a different point of view from the established or traditional theory and method. This point of view implies multidimensional, dynamic and flexible ways of thinking (Hayashi 1998b: 41).

Hayashi went on to describe the sequence of design, collection, and analysis as a primary and iterative "structure finding" process in which data are transformed from a state of "diversification," given the inherent "multifariousness" of the phenomena they represent, to one of "conceptualization or simplification" (41). The discovery of structure is accomplished with what we would recognize today as the methods of exploratory data analysis and unsupervised learning. In effect, Hayashi's definition abstracted the design goals of Ohsumi's Meta-Stat system and presented them as "a new paradigm" of science, one that would encompass statistics, data analysis, and their vast output of data within in a unified, process-oriented framework—data science (40).

In addition to Hayashi's own definition, it helpful also to see how the field was defined by the editors (who included Hayashi) of the proceedings of IFCS-96:

> The volume covers a wide range of topics and perspectives in *the growing field of data science*, including theoretical and methodological advances in domains relating to data gathering, classification and clustering, exploratory and multivariate data analysis, and knowledge discovery and seeking.
>
> It gives a broad view of the state of the art and is intended for those in the scientific community who either develop new data analysis methods or gather data and use search tools for analyzing and interpreting large and complex data sets. Presenting a wide field of applications, this book is of interest not only to data analysts, mathematicians, and statisticians but also to scientists from many areas and disciplines concerned with complex data: medicine, biology, space science, geoscience, environmental science, information science, image and pattern analysis, economics, statistics, social sciences, psychology, cognitive science, behavioral science, marketing and survey research, data mining, and knowledge organization [Hayashi (1998a): v; emphasis added].

Of interest here is use of "data science" as a big tent, an inclusive rubric under which to group a series of domains (which match roughly to a process) as well as a broad range of disciplines and levels, from tool builders to scientists and practice to theory. This passage is also significant for including within the scope of data science the methods of machine learning as well as data mining among the list of sciences concerned with "complex data," suggesting the prominence of these approaches at that time. We will see that not all definitions proceeding from this community were as inclusive.

Hayashi assigned a revolutionary and almost messianic role to data science. In his vision, the statistical sciences had lost their way. Mathematical statisticians had come to overvalue abstract inference and precision, and by choosing to work with the artificial data required to pursue these goals were "prone to be removed from reality" (40). Data analysts, although working with real data, had "come to manipulate or handle only existing data without taking into consideration both the quality of data and the meaning of data … to make efforts only for the refinement of convenient and serviceable computer software and to imitate popular ideas of mathematical statistics without considering the essential meaning" (40). As a result of these divergent attitudes toward data, and the disregard of both for the scientist's engagement with the primary, existential relationship between data and phenomena, the field had become stagnant and lacking in innovation. Data science emerged as a savior, unifying a divided people, showing their way out of the wilderness, and restoring prosperity and prestige to their community.

If Hayashi's criticisms of data analysis sound familiar to those leveled today against data scientists, it is because the issues data science was meant to resolve are recurring and systemic. So too is the separation between data analysis and mathematical statistics, which was recognized by Box, and later Tukey, in the 1970s. In his response to Parzen—who, we noted, sought to overcome a methodological split between the two subfields—Tukey wrote:

> I concur with the general sentiments expressed by George Box in his Presidential Address … that we have great need for the whole statistician in one body—for the analyst of data as well as for the probability model maker—and the inferential theorist/practitioner. One cannot, however, make a whole man by claiming that one can subsume one important class of mental activity under another class whose style and purposes are not only different but incompatible. To be "whole statisticians" as Box might put it, or to be "whole statistician-data analysts" as I might, means to be single persons who can take quite different views and adopt quite different styles as the needs change. As the title of my paper of yesterday put it, "we need both exploratory and confirmatory"! The twain can—and should—meet, but they need to remain a pair (or two distinct parts of a larger team) if they are to do what they should and can (Tukey 1979: 122).

Tukey implies a solution to the schism, later observed by Hayashi, in better organization, not in a utopian "new man" or in a synthetic science *per se*, recalling the division of labor proposed by Naur, but here focusing on different roles within that division. Implicit in this approach is the view that the problem with statistics was not epistemic but organizational.

Here it is helpful to recall a property of Kuhn's concept of paradigm—an obvious lens through which to observe our topic—which is often overlooked by those who use the term: it refers no to an abstract body of ideas that succeed on the basis of their intrinsic rationality or truth value, but to the successful practical application of ideas by means of novel methods and tools in a way that they may be imitated. The concept has both epistemic and social dimensions. Viewed in this light, the question of whether data science is in fact a science—our main question—becomes a matter of determining whether it solves important problems in new ways, by means of an assemblage of ideas, methods, and tools that may be grasped and imitated by others. Hence, although Tukey and Hayashi may appear to be divergent in their approaches to overcoming the problems, they represent the two aspects of a scientific paradigm, the one conceptual, the other practical. This should not be viewed as contradictory.

5.1.3 A Canary the has Forgotten to Sing

Following the IFCS meetings, as well as two meetings of the Japan Statistical Society that held "special sessions on data science,' ' Ohsumi developed Hayashi's definition as well as its rationale (Ohsumi 2000: 331). We call this the Tokyo school of data science, given the association of both Hayashi and Ohsumi with the Institute of Statistical Mathematics in Tokyo, Japan. In a paper that explicitly addressed the relationship between data analysis and data science, and which is perhaps the first of several to claim the flag of data science for statistics, Ohsumi declared that because of its privileging of"mathematical methodologies" over an engagement with data acquisition, data analysis had become "a canary that has forgotten to sing," referring to a Japanese children's song that contemplates a silent bird's fate (332).[4] Amplifying Hayashi, he asserted that "[h]ow data are gathered is the key to defining the relevant information and making it easy to understand and analyze" (331). In making this point, Ohsumi referred to a new figure on the scene, one that contradicted the principles he proposed:

> In my opinion, this viewpoint on the meaning of data science is fundamentally different from data mining (DM) and knowledge discovery (KD). These concepts are not of practical use because they neglect the problems of "data acquisition" and its practice (332).

It is significant that Ohsumi excludes these new fields—or field, since the two so frequently co-occur, along with the variant KDD, "knowledge discovery in databases"—from his definition of data science, since many today would consider the two synonymous.

[4] The specific reference is to a poem, later set to music, written by the Japanese poet Saijoo Yaso ( ), who lived from 1892 to 1970. According to Miriam Davis, "The moral of the song is that if the canary loses its song it is not worth its existence so it should make the most of the gift of song it has been given." (Davis, n.d.)

5.2 Data Mining and Knowledge Discovery

The paradox is instructive: the name "data mining," as used here, made its appearance in the late 1980s and early 1990s as a rubric that included a set of practices motivated by precisely the same conditions that led the Tokyo school to propose the field of data science in the first place. Among these conditions was the relatively sudden appearance of vast amounts of data stored in databases—one of the fruits of classical data science—owing to the rise of relational databases and personal computing in the 1980s, and a suite of tools to work with data, from spreadsheets to programming languages to statistical software packages. Whereas many statisticians viewed these developments with alarm, being acutely aware of the epistemic disruptions they produced for the received workflow of data analysis, the data mining community embraced them as an opportunity to convert data into value. Coming mainly from the field of computer science, data miners developed a set of methods that included the application of machine learning algorithms to the data found in databases in various contexts, from science to industry (such as point-of-sale records generated as by-product of computerized cash registers and credit card use). The relationship between machine learning and data mining was also mutually beneficial—data miners supplied machine learning projects with the large sets of data required for this class of algorithms to perform well. This relationship was greatly reinforced with the rise and development of the Web and social media platforms, which generated enormous amounts of behavioral data.

Although the two fields—for simplicity, let's call them data analysis and data mining—were responding to the same conditions of data surplus and impedance, their philosophical orientations could not have been more opposed. This difference is clearest in their respective evaluations of data *provenance*, the source and conditions under which data are produced. For the data miner, data provenance is largely irrelevant to the possibility of converting data into value. Data are data, regardless of how they are generated, and the same methods may be applied to them regardless of source, so long as their structure is understood. Indeed, for the data miner data exists much as natural resources do, as a given part of the environment, which helps explain the success of the metaphor of *mining* over competing variants, such as *harvesting*, which implies intentional creation. For the data analyst, as Hayashi and Ohsumi took such pains to emphasize, provenance is, or should be, everything, echoing the statistician's orthodox preference for experimental over observational data.

This difference was expressed by Ohsumi in his definition of data science in opposition to data mining:

> Owing to the qualitative and quantitative changes in data [produced by the conditions described above], it is, indeed, becoming increasingly difficulty to grasp all aspects of a dataset in explaining various phenomena. Therefore, new techniques, such as DM, KD, complexity, and neural networks, are being proposed. However, the poten-

> tial of these methods to solve any of these problems is questionable (332).

Ohsumi went on to characterize the way data has changed by listing the new kinds of data with which the statistician is confronted. These include prominently data sets found in databases as by-products of various processes, such as passive accumulation (e.g. from point-of-sale devices), unstructured data (included in text fields), and aggregated data generated "spontaneously and accumulating automatically in the electronic data collection environment" (332-333). He explained his concern with data mining:

> When it comes to analyzing these datasets, people discuss DM and related techniques. However, the important questions to answer are: what dataset is necessary to explicate a certain phenomenon, why is it necessary, how to design its acquisition, and how difficult the whole process is. *This is more important than the dataset itself.* Books on DM do contain terms such as "data preparation", "getting the data", "sampling procedures", and "data auditing", but there is an assumption that the dataset is given and the procedure may start with analysis. Fiddling with a dataset once it is collected is merely a self-contained play of data handling [Ohsumi (2000): 333; emphasis added].

Although his evaluation of data mining seems to be woefully off base—a great deal of Google's success, to take one example, was founded on their embrace of data mining at the time of Ohsumi's essay—in fact his concern is not with the success of predictive analytics *per se*, but with solving what he considered to be the central problem of data science, that of understanding how data are generated in the first place. Given some of the issues that classifiers have encountered with respect to racial bias, for example, he cannot be said to have been wrong.

5.3 A Note on the IFCS

5.3.1 Highlights

- Founded in 1985.
  - Grew out of the German Classification Society, formed in 1977 — see (Bock 2001).
- In 1996 "first scientific society to use the now trendy term Data Science as a title of a conference" IFCS Newsletter 52, October 2015, p.2
  - Adopted the term "data science" from Hayashi.
- In 2015 "GfKl decided to add Data Science Society to the name of the Society" IFCS Newsletter 52: October 2015 p. 7
- Usage persists into current era (Rizzi, Vichi, and Bock 2013; Windham 2001; Baier and Wernecke 2006; Schader, Gaul, and Vichi 2012)

  - Programs at the University of GÖTTINGEN, host of German Classification Society Meetings.
    - ∗ Summer School "Data Science"https://www.uni-goettingen.de/en/data+science+2019/611949.html
    - ∗ Mathematical Data Science (B.Sc.)https://www.uni-goettingen.de/en/582230.html — Fit for the digital era: In the degree programme "Mathematical Data Science", students learn modern mathematical and statistical methods of data analysis and recognition of structures, and acquire tools from computer science to transform these methods into algorithms. Graduates will be able to deal scientifically with large quantities of data and thus gain access to attractive career opportunities.
    - ∗ Applied Data Science (B.Sc.)https://www.uni-goettingen.de/en/640719.html —The field of data science includes aspects of mathematics, computer science and statistics. Data science deals with data analysis and the knowledge gained from data, as well as the techniques required for processing large and partially unstructured data volume. In the Bachelor's degree programme "Applied Data Science", detailed knowledge of data analysis is taught on the basis of computer science and mathematics. In an application subject, students are also introduced to the practical application of the data analysis methods they have learned. The data analysis courses include aspects of machine learning, statistics, pattern recognition and the infrastructures required for effective analyses. Students may choose economics, biology, digital humanities, medical computer science, breeding informatics or physical modeling and data analysis as application subjects. ... Anyone choosing Applied Data Science as their study subject should be interested in formal mathematics as well as application-related practical work. The ability to work in a team is a vital prerequisite for daily professional work later on. English language skills are required. Special subject-related knowledge, especially in programming, is not required.
- Conferences refer to "statistics under one umbrella," echoing the theme of unification seen in Parzen 1977, the Tokyo School, and the US.
-

5.3.2 Narrative

In the January 1999 issue of the *IFCS Newsletter*, the Japanese Classification Society (JCS) reports that "a special seminar about 'Data mining in Data Science' will be held at March 1999" at the JCS Annual Meeting (DeBoeck 1999: 2).

Hayashi, then president of the IFCS, wrote in the Sept 1999 Newsletter column "From the President":

> The potential energy of the IFCS is increasing over the years, and it keeps increasing since the last conference. We can say with firm confidence that the results of data science (including statistics, data analysis, classification and related methods in different respects) really contribute to the development of science and more in general to our civilization and global environment.
>
> Further, I expect that new approaches, and new methodologies and methods in data science will be developed for the coming 21st century, and that the presentations of IFCS meetings will reflect these new ideas. I believe that signs of new movements will be found in the Namur conference in 2000.

IFCS Newsletter Number 23 June 2002. News from the JCS 2002:

> (4) Statistics, Data Analysis, Classification and Data Science Chikio Hayashi (The Institute of Statistical Mathematics)
>
> Data design, data collection and data quality evaluation are crucial to data analysis if we are to draw out useful relevant information. Analysis of low-information data never bears fruit; however, data analytic methods can be refined. In spite of the importance of this issue in actual data mining and data analysis, *I am forced to ask why these problems cannot be discussed at its most essential level.* Perhaps it is a matter of the laborious practical work involved or the otherwise plodding pace of research. Indeed, *these problems are rarely addressed because in academic circles it is regarded as unsophisticated.* In the present talk, I dare to touch these problems with the fundamental concept of data science.

2002, the GfKl

> The German Classification Society (GfKl) will hold its 26th Annual Conference at the University of Mannheim, July 22–24, 2002 under the title "Between Data Science an Everyday Web Practice". This meeting will take place immediately after the 8th IFCS-Conference at Krakow.

In 2004, the president wrote:

> ... the society should foremost be a forum on what has rightly been dubbed "Data Science", dealing with *all aspects of collecting and analyzing data.* This not only covers classical mainstream statistics, but also statistically informal approaches such as Cluster Analysis, Multidimensional Scaling and *Data Mining.* The IFCS is an outstanding platform for dealing with *data analysis techniques that are traditionally ignored in mainstream statistics.* With its broad view on Data Science, *all approaches to data analysis are welcome*, and hence new approaches and classic approaches can be confronted but also synthesized, in an effort to create the best of both worlds. *The IFCS could*

> *and should play a leading role in these new developments in Data Science.* An important aspect here is that IFCS should foster the development of techniques with a keen mind on the applications for which the techniques are mentioned. In particular, *techniques must be made usable for the researcher who collects the data (not only for the statistician or classification expert).* Therefore, *techniques should be user-friendly and transparent, ensuring that the researcher who analyzes the data keeps a clear idea on what the analysis does and what the results display, and the interpretation of results should be crystal clear.* This implies, last but by all means not least, that IFCS should pay attention to how to teach our techniques to students and other prospective users. Thus, IFCS should promote an integrative view on developments in research and teaching of Data Science. One step in this direction is the round table session on this topic, organized by Helena Bacelar at IFCS2004. Other steps will be the organization of further IFCS sponsored courses on classification and data analysis.

Note the definition embraced "data mining" and the emphases on practices that go beyond what is perceived to be the more narrow purview of "mainstream statistics."

5.3.2.1 Conferences / Newsletters

- IFCS-2006: Data Science and Classification. July 25-29, 2006. University of Ljubljana, Faculty of Social Sciences, Ljubljana, Slovenia. 10th meeting.
- 32nd GfKl Annual Conference, "Advances in Data Analysis, Data Handling, and Business Intelligence" at the Helmut-Schmidt-University of Hamburg (Germany) from July 16 – 18, 2008. Section on Data Science and Innovative Tools.
- The 2013 Joint Conference of the German Classification Society (GfKl) and the French Classification Society (SFC) will take place in July 10-13, in Luxembourg. The title of the conference will be "European Conference on Data Analysis". Section on Data Science, including Data Pre-Processing, Text and Web Mining, Information Extraction and Retrieval, Personalization and Intelligent Agents.
- IFCS Newsletter 29
  - Issue 29 (2006) describes on a conference in Germany … "30th GfKl Annual Conference The German Classification Society GfKl (Gesellschaft für Klassifikation) will hold its 30th Annual Conference at the Free University of Berlin (Germany), March 8 – 10, 2006. The conference is titled 'Advances in Data Analysis' and focuses on data analysis, learning of latent structures in datasets, and unscrambling of knowledge."
  - Session on Data Science and Innovative Tools:

    Data Science and Innovative Tools:

    Data Cleaning and Pre-Processing; Data, Text and Web Mining; Information Extraction and Retrieval; Personalization and Intelligent Agents; Tools for Intelligent Data Analysis; New Challenges in Data Science.

    Applications: Subject Indexing and Library Science; Marketing and Management Science; e-commerce, Recommender Systems and Business Intelligence; Banking and Finance; Production, Controlling and OR; Biostatistics and Bioinformatics; Genome and DNA Analysis; Medical and Health Sciences; Archaeology and Geography; Engineering and Environment; Administrative Record Census; Linguistics and Statistical Musicology; Image and Signal Processing.
  - The earmarks: broad definition to include the gamut of processing and an ambivalent attitude toward data mining
- IFCS Newsletter 49
  - Paolo Giudici is (full) Professor of Statistics in the Department of Economics and Management at the University of Pavia in Italy. He currently teaches courses in statistics (undergraduate level), financial risk management (Master's level), and data science (PhD level). .
  - The next CLADAG meeting will be held at the Flamingo Hotel in Santa Margherita di Pula, Cagliari, on October 8-10, 2015. The meeting will take place under the auspices of the IFCS and of the Italian Statistical Society (SIS). The chosen location will facilitate exchange of ideas and networking, as well as an easier access for young researchers. Includes the theme Data Science under the category Multivariate Data Analysis.
  - News from GfKl. Brief report on 50th Anniversary Celebration Meeting Classification Society, Wednesday, July 9, 2014. FirstSite, Colchester, hosted by British Classification Society (BCS) and Department of Mathematical Sciences, University of Essex in cooperation with: Big Data and Science Week hosted by Faculty of Science & Health, University of Essex; Gesellschaft für Klassifikation (GfKl); and European Conference on Data Analysis (ECDA2014). Scientific programme included:
    - Fionn Murtagh, De Montfort University, Leicester, UK: "Data science in psychoanalysis: a short review of Matte Blanco's bi-logic, based on metric space and ultrametric or hierarchical topology."
    - David Wishart, Hans-Hermann Bock, David Hand, Maurizio Vichi, Peter Flach, Sabine Krolak-Schwerdt, and Berthold Lausen: "50th Anniversary resume, outlook and discussion: Classification society and data science."
- IFCS Newsletter 52: October 2015
  - "The Federation, in Kobe in 1996 and, again, Rome in 1998, was the first scientific society to use the now trendy term Data Science as a title of a conference." (2)