Big Data: From Data to Metadata

by George Legrady

Content Set: Bloomsbury Encyclopedia of New Media Art – Connect Entries by Topic (custom-browse?docid=enmaarticlesbytopic) and Bloomsbury Encyclopedia of New Media Art Entries (custom-browse?docid=BloomsburyEncyclopediaofNewMediaArtEntries)

DOI: 10.5040/9781474207959.032

VIEW SECTIONS

Access to multivariate data in real time, in most cases through the internet, generated by systems from anywhere, and in combination with increasingly greater storage capabilities has resulted in an exponential growth in the production, accumulation, analysis, correlation and applications of data. The transition from analogue to digital data capture, much of it by autonomous recording systems, has become a fundamental resource in science, business, the medical industry, government and surveillance for prediction and to assist in decision making and guiding policy once it is processed into information.

Given the exponential growth of data production, its harvesting, processing, dissemination and storage, it is inconceivable to give a comprehensive overview of the impact of Big Data, a general term that has entered public discourse since the early 2000s and represents the increasingly growing accumulation of diverse sets of information through digital form. Data cannot be imagined today without considering it in relation to its delivery systems (the internet); the flow of dynamic production through the broad range of sensors; satellite information; real-time streaming; the interconnectedness of systems (Internet of Things), where data coming from diverse sources activates other devices and processes, further generating new information as data produces more data; and advances in storage technologies.

Not too long ago, new terms such as ‘data processing’, ‘metadata’, ‘algorithms’, etc. entered common language. In the 1960s the term ‘information’ was introduced as a cultural artefact through the writings of Marshall MacLuhan as a way in which to describe identifiable events in both nature and culture, identified by the term ‘systems’, formally analysed by Ludwig von Bertalanffy’s ‘General Systems Theory’, which proposes that complex systems share organizing principles. What are the fundamentals of data and Big Data? Survival for all living things is dependent on collecting, evaluating and learning from data by which to build up knowledge to predict and respond to situations. Data surrounds us, we perceive it, we acknowledge it, we evaluate it, we classify it, we act upon it and we store it for future use.

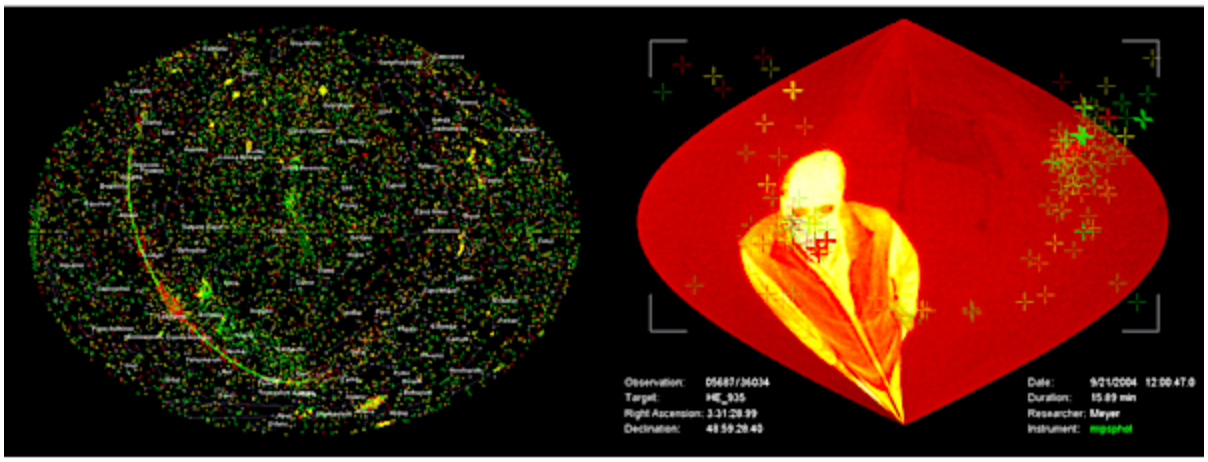

Figure 32.1. *We Are Stardust* (2008–15) by George Legrady. Interactive Installation using NASA Spitzer observation log data. Collection ZKM Museum, Karlsruhe, Germany. © George Legrady / ZKM Museum, Karlsruhe, Germany.

Artistic practice has also turned to data as raw material for aesthetic research and creation. Whereas scientific research is continuously fine-tuning methods to increase precision in data processing, artistic approaches tend to explore the full spectrum of possibilities from coherent messages to anomalies within the system to pursue unexpected correlations or outliers, or to embrace noisy data systems, potential for narrative through incoherent associations and imaginary predictions. Data in an artistic context becomes raw material for artistic expression, for cultural analysis and understanding, for applications by which to investigate and make visible and, given the open systems and free-form nature in art, to pursue initiatives whose purpose may explore the full gamut from the play of the imagination to political activism.

Classification, Data Mining and Knowledge Discovery

We initially learn to classify according to properties such as material, colour, scale, height, weight, texture and taste, and to order according to degree of similarity and difference. In the same way that the meaning of words can only be described by other words, the relative value of data is defined in relation to other data, resulting in the term ‘metadata’, which became widely used in the late 1990s. Digital data files in most cases rely on metadata, for instance the EXIF standard that specifies the properties of how a digital photograph was created giving date, time of creation, which camera and its settings, and location where the image was created. Relational databases, consisting of a table with rows of records matched against ranked attributes, have made it possible to sort, compare and record hierarchical relationships. For example, card catalogue systems in libraries required contextual and content descriptions to differentiate items in the collections. Libraries have been dependent on bibliographic classifications such as the Dewey Decimal class system or Library of Congress classification to record and also physically place items for retrieval access. Classification is therefore a critical component by which we understand and can retrieve data.

The production of inventories and creation of archives initially begins in the collection and the organization of data. An artistic approach may consider the classification process as a form of cultural narrative and aesthetic expression as in Hasan Elahi's series Tracking Transcience, which consists of systematic classification of things experienced, places visited, meals consumed, etc. Under pressure of having to report to counter-terrorism agencies because of his name mistakenly ending on a watch list, Elahi has extensively documented all of his activities, creating taxonomies through photographic documentation over time, a project that has greater ramifications in terms of the technological data generating the life we exist in today. Places we visit, photographs we take, ATMs we use, our shopping transactions all generate metadata which, as Elahi points out, can be correlated through time-stamped and geolocated information.

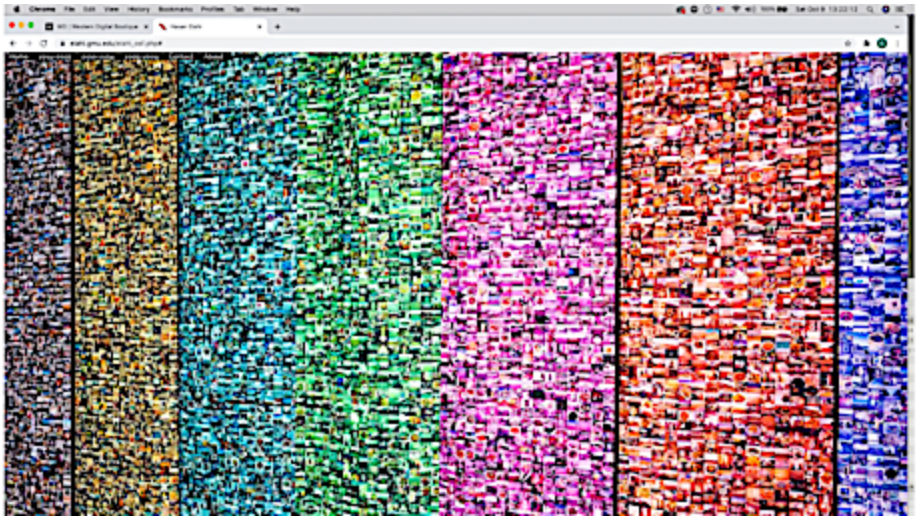

Figure 32.2. *Thousand Little Brothers* (2014) by Hasan Elahi. Pigment print on canvas, 27.5 feet x 16 feet, 838 cm x 487 cm, as installed at Open Society, Foundations, New York, New York. www.movingwalls.org/moving-walls/22/thousand-little-brothers.html (www.movingwalls.org/moving-walls/22/thousand-little-brothers.html) © Hasan Elahi / Open Society, Foundations, NY.

As data is produced, the analysis can only become meaningful through efficient techniques for retrieving information. Making sense of data involves having an idea, or intent, as to how to proceed in the data-analysis process. 'Data mining' is the term used in industry by which databases are searched to retrieve relevant, desired information. One of the key aspects of the process is to identify patterns through computational-based algorithms, for example, frequent pattern growth (FP-Tree) algorithms, otherwise also known as association-rule mining, which identify repeating patterns and correlations between unrelated datasets, the classic mythic example being the purchase in grocery stores of nappies and beer at the same time at a certain time of the day. The goal in this approach is to find inherent regularities in the data, in most cases driven by marketing interests to identify relationships which can then be used to increase sales.

The media theorist Lev Manovich discusses the analysis of data from the perspective of both aesthetic and humanities research in his publication *Cultural Analytics* (2020). Manovich also addresses the question of how to work with data-integrating methodologies bridging engineering computer science procedures with cultural analysis techniques from humanities research. His work since the mid-1990s has been to analyse the impact of global-scale visual culture through computational means given the challenges of scale, speed and diversity of information flow. Whereas strategies of data mining tend to identify statistically volume-based activities, a cultural analytics perspective may also reveal rich unexpected insights in the long tail of data distributions where data is differentiated due to unique features that limit it to singular or fewer statistical clustering.

If 'data mining' is a term used to describe extracting patterns from a large data corpus, possibly in many cases knowing in advance what the outcome may be, 'knowledge discovery' or KDD is an approach to data mining where one does not know in advance what is to be discovered, but which leads to results that lead to insights. The most immediate outcome of a KDD approach may be identification of anomalies within the system of things that do not fit, or are misclassified, or reveal irregularities within the classification system itself. Fraud detection is a critical component of data analysis in business applications. In most cases, patterns emerge over time because a system tends to operate in regular ways, with events repeating themselves. When a data event occurs that does not match historically evolved patterns, systems are designed to flag the unexpected events.

Exploratory analysis of a database can result in the discovery of patterns that can reveal unexpected information. Correlating data with another external information source or superimposing data from different knowledge domains further facilitate discovery. An additional critical component to information discovery is the method of representation. Combining a variety of expertise can result in the most effective delivery. Eric Cadora (the Justice Mapping Center) collaborated with Laura Kurgan (Center for Spatial Research, Columbia University), Sarah Williams (Civic Data Design Lab, MIT) and graphic designer David Reinfurt to realize the Million Dollar Blocks project, which is in the permanent collection of the Museum of Modern Art in New York City. Through the overlapping of datasets, urban spatial maps and data visualization, the team intersected information and data-processing methods to create a visualization presentation bridging social justice, design and architecture. Million Dollar Blocks consists in the overlaying of data from the criminal justice system, specifically the home addresses of incarcerated individuals as they entered the prison system, with the costs of keeping them in prison. This is superimposed on a visual map of Brooklyn gathered through satellite data and geographic information systems, revealing a concentration of locations many of the incarcerated individuals resided. The data was further juxtaposed with census data that provided additional information about poverty and race. The outcome was to identity that it cost $359 million dollars to imprison people from a small area of Brooklyn. For example, Community District 16 has 3.5 per cent of Brooklyn's population but 8.5 per cent of its prison admissions. Eleven blocks in 2003 cost $11,839,665 (see

https://c4sr.columbia.edu/projects/million-dollar-blocks (https://c4sr.columbia.edu/projects/million-dollar-blocks)). Given the project's revelatory visualization, Kurgan asks what if more money were spent on the neighbourhood rather than in displacement?

Online and Streaming Data

With the rise of the internet, information has become accessible anywhere, anytime and continuously reformulates itself (much of it through automated processes) as new data enters the data stream. The internet is a 'knowledge space' we collectively engage with and contribute to, and, according to French philosopher Pierre Levy, who reviews the historical conditions in his 1994 publication *Collective Intelligence: Mankind's Emerging World in Cyberspace*, the internet reformulates our social interactions. Ken Goldberg and Joseph Santarromana's *The Telegarden* was an early (1995) internet-based work that engaged collaborative community participation. A limited-space table-scale garden environment was installed at the Ars Electronica Museum from 1995 to 2004. The installation consisted of a multipurpose robotic arm that could plant seeds, water plants and do various garden-related activities accessible online 24 hours a day. The system was made available to the general public, who could access online, plant, water and monitor the controlled environment. The installation raised the challenge of to what degree telematic presence, being able to have impactful action at a remote distance, would also be managed by a sense of responsibility for one's own actions.

Streaming data from sensors and other sources – such as GPS (Global Positioning System), which consists of near real-time information – enhances our interaction with our environment. Through digital sampling, all things such as the weather, temperature, geological and oceanic activities, stock exchange, traffic, etc. result in measurements that can be technologically harvested, stored and therefore analysed, and consequently formalized and applied in any way, to assist in decision making, meaning to predict what may come next. *Gnomon*, created in 1996 by artist/architect Bruce Tomb, graphic artist Tom Bonauro and designer John Randolph, was exhibited at the San Francisco Museum of Art in the Architecture and Design Department. It consisted of a large 18-foot-long potato-like translucent fibreglass structure on wheels that adjusted its positioning based on GPS signals. The sculptural device also included infrared collision detection and other sensors. Local visitor interaction would be used to activate additional positioning responses. The system also generated video, and the audio was separated into low, mid and high frequencies spatially directed with high frequencies towards the ceiling and low, through a subwoofer, towards the floor. The installation was created at a time when GPS signals were degraded by 'selective availability' implemented for national security reasons.

Big Data is irrevocably linked with the internet as the delivery platform. The Photosynth software now discontinued by Microsoft, while not an artistic work in itself, but rather a service that revealed the interrelationship of data autonomously connecting with online data, consisted of two unique features. If we consider the still-photographic image as a multidimensional data structure, data patterns from photographic images can potentially be matched with others and, through this process, the digital photographic image, previously a single visual record, can be endlessly enhanced for detail and

panoramic extensions through pattern matching with other crowdsourced data found online. A unique by-product revealed by Photosynth was the ability to identify the point-of-view locations of where each photograph was taken.

The monitoring of natural phenomena transmitted as data online in near real time such as weather conditions, traffic, ocean temperatures and seismic waveforms has been at the core of scientific and military research since the 1940s. Streaming data can be applied to artistic explorations in any number of ways. Paul DeMarinis's *The Messenger* (1998) translates character data from email messages into electrical signals which trigger dancing skeletons. *Mori: An Internet-Based Earthwork* (1999) by artist/robotics engineer Ken Goldberg and multimedia artist/electronic composer Randall Packer with Wojciech Matusik and Gregory Kuhn is an installation that receives a seismic data stream and translates it into a multi-modal experience through a resonating vibrating platform, wave visualizations and surround low-frequency sounds by which to sense the unpredictable fluctuations of the Earth's movements.

The *Listening Post*, a data visualization and sonification installation realized in the early 2000s and still operational, is one of the more significant projects that uses streaming data to create a multi-modal, aesthetic experience. Created collaboratively by statistician Mark Hansen and sound media artist Ben Rubin, the installation integrates complex engineering, real-time data processing, statistical analysis and sonification to explore the aesthetic potential of data as raw resource. Texts culled from chat rooms are sampled, statistically compiled, sonified and sequenced to produce a continuously changing, complex layered set of sounds and text arrangements. The project began with accessing chat room data in real time on the internet. This required custom software and implementation of various protocols by which to identify the source material, and then to harvest the data. The next step involved analysing the data to classify the contents for topics and other associated metadata. The various texts were then sonified using text-to-speech software and synthesized voices. Various other sonic components were added. The visualization involved featuring the texts on 231 small displays in a 2D matrix of 11 x 21 positions. Additional custom software processed all the various multi-modal content components to activate different animations in a sequence of movements by which to vary the viewing experience. Challenges with presenting a multi-component installation of this type are that changes take place over time in communication protocols, software dependencies, online interaction, cultural behaviour (do people still chat?) and audience responses.

It is not possible to mention all of the projects that have engaged streaming data but the following have contributed different perspectives. *Emotional Forecast*, realized in 2010, is an installation by artist Maurice Benayoun that explores the potential of data accessed on the internet, which is then resituated through visualization and animation to read as data expressing emotions. Benayoun's installation *Emotion Wind* (2014) integrates online data with on-site gallery interactivity.

At the San José International Airport, *eCLOUD* (2010) is a permanent installation by Aaron Koblin, Nik Hafermaas and Dan Goods that consists of a large cloud-like display of multiple rectangular liquid crystal tiles hanging from the ceiling in the open space of the main passageway between two gates. The tiles are

clustered in a semi-random formation to simulate the shape of a cloud and each panel can be individually activated between two states – either transparent or opaque. The artwork harvests descriptive weather conditions, temperature, humidity, wind speed, wind direction and visibility from the OpenWeatherMap API every ten minutes. The custom software cycles through the list of cities, changing every thirty seconds. The weather condition ('Light Rain', 'Cloudy', etc.) determines which animation mode will be used, and the rest of the data values are mapped to parameters that drive the animation. A screen display also features a real-time 3D rendering of the sculpture/animation, along with the city weather data.

One of the major activities in scientific research involves the continuous sampling of our environment at the global or planetary levels through incoming data from sensors placed in all types of locations where activities in natural phenomena occur and/or monitoring is critical, for instance as in industrial and radioactive sites. 'Amphibious Architecture', realized in 2009 by architect David Benjamin and artist Natalie Jeremijenko, collects sensor data in New York Harbor. It is an example of a sociological and artistic approach whose goal is defined by discipline-specific interests from architecture and environmental studies, but without the constraints of scientific research principles. An artistic approach allows for significant exploration, which can integrate real science with poetics, social concerns and the play of the imagination. Installations by the British artist Steve Tanza (Stanza) integrate real-time tracking data to activate a mass of electronic devices, bringing attention to the interconnectedness of collected data streams to create an experiential installation environment.

The Artistic Approach

The artistic approach to working with data, Big Data and streaming data may be summarized in the following ways: (a) an emphasis on alternative modes of enquiry, (b) innovation and complexity in visualizations and representation and (c) play of the imagination. Artistic practice provides a level of flexibility much greater than standard research approaches as the practice is not constrained by conventions defined by the discipline. There are trends that influence how artworks may come to be, but as innovation in aesthetic form is a key criteria, projects aim to be unique. Artists can ask questions that at first glance may not have any purpose or relevance as the arts discipline is an open-ended system. Any approach is a valid one, any method by which to proceed is acceptable, and the key goals are to arrive at results that are insightful or expressive, challenge previously defined assumptions and stimulate new ways of imagining and representing how we understand the world.

Artists, designers, architects and other practitioners in the arts will in most cases approach a problem from a combined empirical and formalized perspective. Constraints in conceptual development and on 'what is possible' or interesting in most cases result from the limits of technological possibilities, audience conditions and market realities and what at any moment are the predefining cultural and ideological values. Experienced practitioners bring to data representation a set of advanced skills and a conviction in the grammar of visual and formal vocabularies. Data in itself does not determine how it is to be represented, and the formal and visually syntactic grammar of representation, which follows a set of

principles, is open to variations from precise order to near-chaos. The design process tends to be recursively evolving with the intent to achieve the emergence of unforeseen relationships, correspondences and new outlooks. The combining of differing discipline-specific methodologies is valued, and not dependent on any prior validation. 'What if I take from here and apply in this unexpected way?' in many cases may result in unpredicted results. In comparison to the more formal disciplines, such as the sciences, the medical field and other academic specializations where the research process follows structured conventions, the artistic process is guided by speculation, unexpected juxtapositions, insights gained through perception, in essence an approach that the philosopher Kant defined as 'the play of the imagination'.

Advanced knowledge of the grammar and aesthetics of visualization is an important aspect of artistic approaches to data representation. In his book *The Visual Display of Quantitative Information* (1983), the statistician, computer scientist and designer Edward Tufte provides a range of examples with discussion of challenges and solutions of representing data through visualization. The following are a selection of examples of artists and data visualization designers whose projects demonstrate various ways of engaging with data through visualizations and interactivity. Lisa Jevbratt's *1:1* (1999) is a visualization map of the IP numerical addresses of the internet viewed on one screen. The design consists of colour-coding the IP addresses, which results in an abstract image of horizontally drawn lines. Through five different interfaces (Migration, Hierarchical, Every, Random, Excursion), the project visually reveals transitions and changes in how the various IP addresses were repurposed between 1999 to 2001. *The Secret Lives of Numbers*, an interactive data visualization realized in 2002 by Golan Levin and collaborators Jonathan Feinberg, Shelly Wynecoop and Martin Wattenberg, provides a statistically determined graph of the relative popularity of each integer between 0 and 1 million. Collecting data from a broad range of sources, the visualization reveals the relative performance of numbers in relation to each other. Even though the project is a statistical demonstration of cumulative results, the overall effect is poetic and narrative in form. The nature of the changes in data and data access over time suggests that the accumulation most probably may vary each year, resulting in an interesting transitional difference through the years. Starting in the early 2000s, designer Nicholas Feltron has each year produced an annual report that documents his personal activities in the year based on collected data (see http://feltron.com (http://feltron.com)). Each annual report has a unique visual identity similar to corporate annual reports, and compiled data and graphs provide summaries of his activities such as meals eaten, drinks, restaurants and places visited, subways and taxis taken, etc. Somewhat similar to Hasan Elahi's documentation of everyday activities, Feltron gives us an overview of his actions of the past year within the design framework we associate with corporate information.

Aaron Koblin's *Flight Patterns* (2011) uses flight pattern data of over 140,000 airplanes to create aesthetic, abstract visual animations. *Wind Map* (2012) by Fernanda Viégas and Martin Wattenberg visualizes the flow of wind patterns over a map of the US based on data collected hourly from the National Digital Forecast Database. Agnieszka Kurant's *Conversions* (2019) receives Twitter data from the

internet, which is aggregated by sentiment-analysis software to activate temperature changes on a panel that transforms in colour. Ken Rinaldo's sculpture *Abiopoiesis Microbiome* (2016) translates wind and weather data to drive the motion of sculptural forms.

The ultimate aestheticization of data staged within a visual, interactive and sonified multimedia presentation is the work of the Japanese artist and composer Ryoji Ikeda. Consisting of visual data in black-and-white forms, numeric streams and text data that are activated in animations of varying intensity, the visualizations are driven by audio consisting of pure waves (sine, triangle, rectangular, sawtooth) to randomized noise, organized in complex but rhythmic patterns played at loud volume. Projects such as *data tron*, *datamatics* and *data flux* use data from sources that seem scientific, for example mathematics, DNA and molecular structures. No matter how technical or complex the data information, Ikeda repurposes them to generate an aesthetic experience.

Digital photographs in themselves can be considered as high-dimensional datasets as they consist of grids of pixels, each of which has assigned numeric values that represent their location in the 2D matrix of the image, and also each pixel has numeric values that represent their colour and brightness. These tonal values which represent the pixels' colours can be statistically processed and averaged computationally, for instance allowing for multiple images to be easily sandwiched together. Nancy Burson, Jason Salavon and Idris Khan have all created images based on statistical averaging of pixel values to create artistic works.

Nancy Burson, with the aid of MIT engineers Richard Carling and David Kramlich produced a series of pioneer photographic images in the early 1980s that were portraits resulting from the statistical averaging of then current politicians' portraits. As an example, the image titled *Big Brother* in 1983 is a composite of portraits of dictators known at the time. Portraits of Stalin, Mussolini, Mao, Hitler and Khomeini were blended together pixel by pixel to result in a portrait that mixed features of each individual.

Jason Salavon and Idris Khan are two recent artists whose works have also explored in detail the potential of statistical averaging in the visual domain. In Salavon's Portrait series (2009–10) portraits by major Dutch artists such as Hals, Rembrandt, Van Dyck and Velasquez are averaged to result in blurry but nonetheless recognizable positioning of persons within the canvas space. The intent of the project may have been to identify how each of those painters individually explored where to position their subject within the rectangular canvas. Salavon's *Every Playboy Centerfold 1988–1987*, *The Class of 1988* and other such works also proceed in this way, averaging the visual content of a collection of images that follow a pre-defined syntax of visual ordering. The *Playboy* centrefold represented a specific visual ordering, a formula by which the magazine determined how the figure of the centrefolds would be positioned within the image space. The British artist Idris Khan also works in this way, averaging collections as digital collages, in his case with an emphasis on historical source material such as formally staged photographs of industrial structures by the German couple Bernd and Hilla Becher, and sheet music such as Six Suites for the solo cello by Bach, which are precisely registered to overlap in visual detail so that the lines of music blend into each other to create a complex visual texture.

Noise, Ideology and Challenges

The intersections of biometric data collection, data distribution, issues of personal identity and the potential of inaccurate data representation are topics of critical interest to artists who recognize the need to make visible the means by which citizens are classified and tracked within the larger bio-info government management systems. All data repositories inherently record or introduce noise within the system. This results from incomplete data, noisy data, human or system errors in transactions, modelling errors due to mathematical processes or else classification errors. Sterling Crispin's series Data Masks (2013) are fictional data portraits produced by positioning face recognition systems onto noise patterns. Yoon Chung Han takes biometric data from fingerprints (*Digiti Sonus*, 2012–13) and iris data (*Eyes*, 2018) transforming it into a multi-sensory audio-visual experience. Judith Donath's *Data Portraits* (2010) uses data collected online which is then presented in various graphs as information visualization, and participants willingly provide their identifying data for digital media arts pioneer Lynn Hershman Leeson's *Shadow Stalker* (2018), which showcases the fact that we live in an era of permanent data tracing. Researcher Kate Crawford and the artist Trevor Paglen have collaborated on investigating the critical problematics of biased evaluation encoded in data, in particular when decisions are made by autonomous data evaluation systems. Facial recognition systems may wrongly connect identities stored in data, or else intentional matches can be realized, either through disrupting data security or faking the data. Critical data studies is a research area where computer science, communication studies and art practice can intersect to develop projects that respond to the operational forces in how data functions. Forensic Architecture at Goldsmiths College, University of London, is such a research centre. It integrates architectural analysis with digital modelling to introduce a critical discourse delivered through artistic artworks.

Some Early 1990s Explorations of Data and Interaction

The introduction of recording, publishing and duplicating technologies throughout the twentieth century led to alternative explorations and uses by the artistic communities parallel to those in mainstream industry applications. Artistic repurposing of existing data became a creative process in itself as evidenced by art movements that included Dada, Fluxus, Musique Concrète, Book Arts, collage and assemblage, where pre-existing cultural information in the form of newspaper clippings, found or appropriated images, texts, audio and video samples became raw material for creative purposes.

During the late 1980s and early 1990s, new digital technologies were introduced to the market through which images, sounds, texts, that is, cultural information, could be sampled and reformatted into digital form thereby levelling out the differences between official documents stored in archives and personal, family documents consisting of snapshots, home movies, birth certificates, etc. stored at home in shoeboxes and closets.

Macromedia Director, a software initially developed for the gaming industry in the early 1990s, was one of the first available multi-linear interactive authoring tools that allowed for the organization of data, including text, still images, video and audio. Because it could be programmed by the Lingo scripting language, and function as a controller of LaserDisc video by which to integrate high-quality video and audio, the commercially focused Macromedia Director software became an ideal platform for creating artistic-oriented works. One of the great privileges of artistic practice is that it is an open system, allowing for invention, configuration and juxtaposition of resources, ideas and references without significant constraints imposed by the history of the discipline.

Authoring in the Director environment had its challenges. Multi-directional, combinatory narrative sequences were a new form of expression that had to be thought through, and, as examples were limited, for any of the works one had to conceptually and aesthetically classify, cluster and define hierarchies of relationships with which disparate sets of data could be placed in relation to each other.

I was initially inspired by the Fluxus artist Daniel Spoerri's heuristic-based publication *An Anecdoted Topography of Chance* published in 1966. A set of items on a dinner table, residue from a dinner, were visually mapped in a drawing according to their position, with each recorded as a numbered inventory entry. Each item in the catalogue had its own page, with title, small doodle, followed by an anecdotal description of what the time was, where it came from, followed by literary and historical references. A few who had attended the dinner then added footnotes through which the items were cross-referenced. Using the Director technologies and the model of Spoerri's book, I produced three interactive artworks – the *Anecdoted Archive from the Cold War* (1992), *the [clearing]* (1994) and *Slippery Traces* (1995) published by the ZKM in their Artintact series – each of which consisted of topic-specific multivariate data such as texts, images, videos, sounds, etc. with the intent to creatively explore the potential of interactivity for narrative and visualization sequencing. Each of the projects consisted of cultural data that had to be organized in some way, and the challenge was to arrive at a system by which the disparate data could be clustered and sequenced.

Macromedia Director quickly became the authoring tool of choice for many artists with which to compile digital data for narrative. In 1993 Christine Tamblyn, a media artist, writer and theorist at San Francisco State University, produced the widely exhibited interactive CD-ROM *She Loves It, She Loves It Not: Women and Technology*, a digital media artwork that consisted of found material which she combined based on a cultural analytic feminist approach to address the representation of the relationship of women and technologies.

The collection of data as a potential for narrative, cultural critique and discourse is exemplified by *The File Room* by the conceptual media artist Antoni Muntadas. Initially exhibited in 1994, the goal of the artwork has been to function as an evolving metadocument on the topic of censorship. Each of these examples underscores the fact that one of the key features of data is that it continuously accumulates and at some point requires classification.

Making Visible the Invisible at the Seattle Central Library

Making Visible the Invisible is a commissioned permanent artwork located at the Seattle Central Library that I realized with computer scientist and media artist Rama Hoeztlein. It has been in continuous operation since September 2005. It may be the longest-operating data visualization in existence. Designed as an aesthetic artwork that also functions as an information system that represents the reading, viewing and listening interests of the downtown Seattle community, it also makes visible to patrons and participants the results and real-time operations of the data infrastructure by which a library operates through visualizing the circulation of data in near real time, information that is not normally accessible for public view. The installation receives data from the library's main servers on all the items checked out in the previous hour. The installation stores the data on a server located at the library which then distributes parsed data to three other computers, each of which has specific tasks to produce the animations which are displayed on six large horizontally positioned screens above and across the length of the main information desk. The four animations are continuously updated with new data of the checkout activities of the previous hour. The data is aggregated by custom software to statistically compile current topics of interest that patrons have retrieved from the library in various media, such as books, DVDs, CDs and, in the early operations of the installation, also VHS video tapes and music on audio tape.

The software design and visualizations were guided by the spirit of 'knowledge discovery', providing the audience with the opportunity to perceive, in various ways through the different animations, the topics of interest circulating while they are on site. The first animation provides the statistical overview, giving literal, numeric data that compares how many movies, books and CDs have circulated. The next animation features the same data but this time through a chronological by-the-minute streaming of titles with some metadata, colour-coded to differentiate media and titles. The following animation gives an overview in a two-dimensional matrix form of all checkouts, revealing which categories are the most active in relation to the others. The fourth animation parses all the titles to create a keyword index of words from the most active titles. These words are then colour-coded and spatially distributed on the six screens according to their categorical classification. For instance, if a word shows up in titles that are 23 percent in religion, 37 percent in philosophy, 5 percent in science, 12 percent in literature and 23 percent in history, the statistics determine the colour mix and position of the word on the screen.

The data analysis and access requirements were collaboratively worked out with the library's IT and legal departments over a year-long discussion. The metadata eventually agreed upon included the following for each record of checkouts: the item checkout time stamp, check-in time stamp, the unique acquisition ID of the item, the bibliographic record set by the Library of Congress, a Dewey classification number if the item is classified according to the Dewey system, the media type, a unique barcode for each object, the title, keywords and assigned call number, which is used to place the item on the shelves.

As mentioned, the installation has been in continuous operation since September 2005, a date that represents a 10 year mark following the start of Amazon, Yahoo, eBay, Internet Explorer, Windows 95, etc. and the wide adoption of the internet as the go-to for knowledge information. The installation at the library only visualizes activities of the previous hour but the artwork was designed to store all transactions, its full history of checkouts over time for future analysis. In the process of monitoring the performance of the software, the data has been regularly analysed in my research lab to check its validity, look for anomalies and to also identify and track cultural trends. At the end of each day, the artwork also receives data on the returned items with matching metadata to allow comparisons with checkout time stamps, which facilitate tracking of the performance of items over time. A question of interest that influenced the software design was to gain insights through tracking long-term performance into how the library in the twenty-first century had to reinvent itself to maintain community usefulness. What noticeable shifts could be identified by analysing the circulation of data to see emerging patterns that could indicate the impact of the internet in this transitional period to the digital Information Age? And at the granular level, one could identify the performance of an item at a particular time and date. As an example, a children's book titled *Stinky Cecil in Mudslide Mayhem* volume I, has the honour of being the 100 millionth checkout, which took place on June 12, 2021 at 5:05 p.m.

The artwork was conceived to provide information in the moment, and to also collect data over a long period, extending its function into a research resource for cultural change and long-term analysis as the accumulated data allows for research of granular historical detail over the 16 year period of its existence. The data also allows for insights into how an urban downtown community has interacted with a library based on the choice of checked-out topics, and for comparative studies between media and content in terms of their circulation history. As each item's multiple copies are also recorded individually, one can study an object's life expectancy, review correlations between checkout and return times, identify co-occurrence patterns between topics and media, make assumptions about community interests in terms of correlation to external world events, and potentially lead to prediction analysis in terms of both content and testing the robustness of prediction algorithms. Additionally, analysis can be carried out to test the classification coherence and system functionalities, and reveal anomalies in these areas.

Data Collection as Artistic Process and *Pockets Full of Memories*

In the early 1970s, the artist Hans Haacke, motivated by analysing operational systems within socio-economic and political contexts, introduced a number of installation works such as *MoMA Poll* (1970), where visitors were asked to answer a politically oriented question by placing their votes into transparent collection boxes that revealed how the majority of museum goers voted. For *Visitors' Profile* at the John Weber Gallery in 1972, exhibition visitors had the opportunity to fill out questionnaires contributing their demographic data during the first half of the exhibition. In the second half, the aggregated data was exhibited, functioning as a feedback system by which to reveal and inform the demographic breadth of the visiting community.

Pockets Full of Memories was commissioned for the Centre Pompidou by exhibition coordinator Boris Tissot, to address the combined themes of the public contributing data to construct and interact with a cultural data archive, and the aesthetic potential of interacting with such an archive as a form of narrative that would evolve throughout the length of the exhibition. The design of the installation took over a year to plan, to result in an installation where museum attendees would contribute data of their choice by digitally capturing an image of an object, and describing it through a questionnaire. The design and production had a strong interdisciplinary approach, requiring the expertise of individuals from diverse fields. Haacke's two mentioned artworks, which involved the collection of data to sample the cultural perspectives of the audience, served as a precedent for an installation I collaborated on in 2001 with Timo Honkela, a computer scientist with expertise in natural language processing and the Kohonen self-organizing artificial network, an early artificial intelligence algorithm. Other key contributors included Projekttriangle, a design team whose task was to give the installation and the interaction data collection station a unique unifying visual identity. Dr Brigitte Steinheider, a psychologist from the Fraunhofer Institute, assisted in the development of the questionnaire, and a team of hardware and software developers at the C^3 Centre for Media and Communication in Budapest realized the fabrication and software operation of the system. The three core components of the installation consisted of 1) a data collection/questionnaire station, 2) dynamic data processing by the Kohonen unsupervised, artificial neural-network, self-organizing map algorithm, and 3) a large projection which featured the organization of the public contributed objects, with their placement determined by the Kohonen algorithm according to the semantic descriptions requested by the questionnaire. The audience had the opportunity to also post messages online to any of the objects both on and off site of the exhibition space.

Figure 32.3. *Pockets Full of Memories* (2001–7) by George Legrady/Timo Honkela. Interactive installation. Contributions of digitized images of objects are classified by the Kohonen self-organizing artificial neural-network map algorithm based on semantic analysis of their descriptions. © George Legrady/Timo Honkela.

One of the unique features of the artwork was the integration of an early unsupervised self-organizing artificial neural-network algorithm (the Kohonen Map), through which the public's contributed data was autonomously classified and visually clustered dynamically on a large projection screen. The algorithm's training set consisted of the multidimensional metadata associated with each scanned object contributed by the participants during the scanning process at the data collection station. This involved filling out a multi-screen questionnaire with which the public described the object through keywords and other information including a rating scale for each of the following eight attributes: Old-New, Soft-Hard, Natural-Synthetic, Disposable-Long Use, Personal-Non-Personal, Fashionable-Not Fashionable, Useful-Useless, Functional-Symbolic. This Osgood semantic differential sliding scale has been used extensively in the measurement of attitudes in surveys and research. Whereas most applications are limited to a range of five to ten positions between the polar opposites, a high sampling rate of 128 positions was implemented in the questionnaire to allow for a detailed, granular scaling by which subtle measurements could be made at each of the eight attribute layers. The algorithm processed the data 120 times every two minutes, each time further fine-tuning its calculation to eventually arrive at a fully ordered state where every object in the collection was precisely positioned topographically in relation to every other object. Each time a new object entered the collection, it disrupted the order, forcing the algorithm to recalculate.

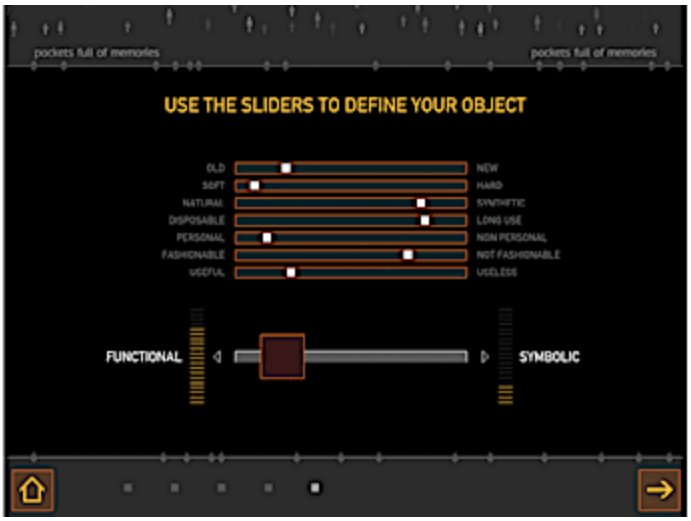

Figure 32.4. ***Pockets Full of Memories*** **(2001–7) by George Legrady/Timo Honkela.** Questionnaire. Contributions of digitized images of objects are classified by the Kohonen self-organizing artificial neural-network map algorithm based on semantic analysis of their descriptions. One of the data sources was a screen in the questionnaire that used a semantic differential rating scale by which contributors could describe their submitted object. © George Legrady.

After the initial exhibition in the summer of 2001, the installation travelled to seven other venues over a period of seven years, each with its own specific audiences bringing different perspectives to what such a data collection artwork could mean. The other venues included media arts festivals, a gallery, a museum dedicated to communication and art museums for the general public, with the final installation presented

at the Museum of Contemporary Art in Taipei in a Chinese-language version in 2007. The artwork was repurposed in 2015 for the Bogota International Book Fair to celebrate the fiftieth anniversary of Gabriel Garcia Marquez' epic novel *One Hundred Years of Solitude*. The installation, now titled *Imagining Mocando*, was revised in collaboration with artist/researcher colleagues Andrés Burbano at the Universitat Oberta de Catalunya (Barcelona) and Angus Forbes at Purdue University. It included a number of technological upgrades. Instead of a scanning station, the data was contributed through iPhones and iPads from both on and off site. The contributed data was collected in Bogota, but processing took place in Chicago and was visualized on a very large curved screen in Bogota, all taking place in near real time.

In summary, the conceptual premise of the installation was to explore the impact of semantic descriptions through a selected set of attributes by which a recognizable visual object may change in meaning based on how it was described. This artwork had in common with the Seattle library project that all the contributed data was kept as the training set for the artificial neural-network algorithm, but also for future analysis as each exhibition's and venue's contributed data can be thought of as an assemblage of cultural artefacts, ideal for an archaeological analysis by which to evaluate differences in cultural perceptions over time, venue-specific audiences and geographic locations between the various exhibitions.

Conclusion

The idea of Big Data, the increased collection of data over time, the flow of continuous data, the multivariate structure of the collected data, increased bandwidth, reduced storage costs, the interconnectedness of things (IofT), etc. are the fundamental conditions that are currently transforming our lives and global society as a whole. Big Data per se provides historical modelling, but data itself is not useful without analysis, and today with machine learning, autonomous processing through black boxes, results are determined outside of human supervision. Whereas such approaches can advance medical and scientific solutions, and are used by business and politics to make predictions, an artistic approach to Big Data provides an alternative perspective that makes visible possibly unusual relationships through visualizing, audifying and revealing them. The arts are a space, as-of-yet not driven by conventionalized function demands, where critical questions may be asked to assist understanding of the continuously changing conditions of culture, much of which is now defined by Big Data.

References

Levy, P. 1994. *Collective Intelligence: Mankind's Emerging World in Cyberspace*. Reading, MA: Perseus Books.

Manovich, L. 2020. *Cultural Analytics*. Cambridge, MA: MIT Press.

Spoerri, D. 1966. *An Anecdoted Topography of Chance*. New York: Something Else Press.

Tufte, E. 1983. *The Visual Display of Quantitative Information*. Cheshire, CT: Graphic Press.