DataRemix: Designing The Datamade Through ArtScience Collaboration

Ruth West, Roger Malina, John Lewis, *Member, IEEE*, Scot Gresham-Lancaster, Alejandro Borsani, Brian Merlo, and Lifan Wang

Fig. 1. Interacting with the ATLAS in silico installation on the Varrier™ 100-million pixel autostereographic display at Calit2 UCSD.

Abstract—ArtScience is emerging as one approach for creating novel ways of seeing and new ways of knowing. We propose a role for ArtScience research and creative work in contributing to the necessary shifts to go beyond the current crisis of representation. We specifically describe DataRemix, a recombination and reappropreation practice intended to trigger novel subjective experiences and associations. The narratives framing data creation and representation circumscribe what we can see and know, and how we see and know. How do we see and know beyond what our instruments, algorithms, representational schemas and training guide us to see and know? How do we look for what we don't know we're looking for when we can only examine at most a tiny fraction of the available data? Our argument is grounded in and will be illustrated by experience with several ArtScience collaborations spanning genomics, astrophysics, new media, and holographic sound design.

Index Terms—New Media, Multimedia, Artistic Visualization, Collaboration, Aesthetics, ArtScience, Representation, Sonification, Remix, Sound Distribution System, Holographic Sound.

1 MANIFESTO: A CRISIS OF REPRESENTATION

> *...Darwin faced a very basic visual problem: how could natural selection, a concept almost by definition impossible to illustrate directly, be illustrated, especially when the existing visual conventions of the natural sciences were associated in varying degrees with conceptions of species fixity?*
>
> Jonathan Smith, Charles Darwin and Victorian Visual Culture [1, p.1]

It is the sense of the authors of this paper that in an era of increasingly vast and abstract data we face a "crisis" in representation comparable to Darwin's problem.

- *Ruth West is with University of North Texas. E-mail: ruth.west@unt.edu.*
- *Roger Malina is with UT Dallas. E-mail: roger.malina@utdallas.edu.*
- *John Lewis is with Victoria University. E-mail: john.lewis@vuw.ac.nz*
- *Scot Gresham-Lancaster is with UT Dallas. E-mail: scotgl@utdallas.edu.*
- *Alejandro Borsani is with University of North Texas. E-mail: alejandro.borsani@unt.edu.*
- *Brian Merlo is with UT Dallas. E-mail: brian.merlo@utdallas.edu.*
- *LiFan Wang is with Texas A&M. E-mail: wang@physics.tamu.edu.*

The increasingly rapid pace of digitization of nature and culture will result in a "digital universe" by 2020 estimated at 40 trillion gigabytes [2]. What defines data as "big" is shifting. In some fields, such as the digital humanities, access to large-scale data enables new kinds of analyses but does not necessarily require use of supercomputers. boyd and Crawford [3] observe that big data is no longer solely characterized by its size, but by the "capacity to search, aggregate, and cross-reference large data sets[3, p. 663]." And as Gantz and Reinzel point out, ironically "as soon as information is created or captured and enters the digital cosmos, much of it is lost [2, p. 4]." As we accrue data of unprecedented resolution, size and scope, it comes with an aura of incompleteness.

In addition to its ephemerality, or uncertainty as to whether the data is complete or representative (e.g. partial data due to restrictions placed on access to social media data such as Twitter's data stream[3]), additional forms of incompleteness include perceptual limits, the location of meaning and the gap between data and their underlying phenomena. While the amount of available data is increasing there are constraints imposed by human cognitive and perceptual systems. The human visual system is generally the highest bandwidth input to the brain, but phenomena such as change blindness and memory prompting[4] illustrate how limited perception actually is. Visu-

alizations also necessarily involve subjective choices, though this is not always acknowledged or made explicit. Consider a data set with 100 dimensions. Restricting our scope to visualizations that involve only three of the 100 attributes, there are more than 100,000 possible combinations to choose from. Techniques such as principal component analysis can be meaningless if the attributes have different units. Factor analysis is more appropriate but is the wrong choice if subtle or nonlinear interactions are of interest. Of course these are some of the core issues in data visualization [5] – our point is simply that objectively testing every possible visualization is not tractable, and thus the role of subjective choices must be foregrounded.

There is yet a more fundamental issue: there is, and may always be, a gap between the data (and algorithmic reductions of the data) and the meaning ascribed to the data. Turing award winner Peter Naur described this issue[6]: while one can devise rules to assign meaning to the output of a rule-based system, the output of these rules must be interpreted in turn, leading to an infinite regress. Naur is the 'N' in the BNF formalism, and is thus well acquainted with rule-based systems.

Data is considered a resource, a raw material that can be manipulated and refined from information-to-knowledge-to-wisdom[7]. Whether generated by terrestrial observatories, automated genomic sequencing, social media, high-resolution sub-cellular imaging, surveillance video, financial transactions, or the emerging Quantified Self movement[8] the richness of these massive repositories is such that an enormous amount of interpretations can be generated by both the creators of the data and by others for a variety of unanticipated uses. Yet through choices such as what to sample, the sampling resolution, file formats, what gets stored, cloaked, or discarded when the data is too large to retain all of it, or the database schemas utilized in managing it, unspoken framing narratives arise that encode agreed upon assumptions about what the creators think they will find in the data, what they think they can know. The same processes of human cognition when dealing with the flux of direct sensory data are at work when confronted with large data streams generated by instruments.

In addition to these framing narratives of data creation, we make choices in how we process and represent data. Modern scientific practice is often regarded as being remote from experiential and aesthetic concerns. In an increasing number of scientific disciplines, hypotheses are formulated and evaluated as algorithmic queries applied over millions of data records representing the digitized data from instruments that detect phenomena to which our own senses are blind. Anticipating this trend, historian of science Daniel Boorstin identified "epistemological inversion," a shift from primarily hypothesis-driven to discovery-driven approaches[9] as one of the profound implications of big data for science; he joked that we were moving from a "meaning rich, data poor: to a data rich, meaning poor" situation. While transforming and accelerating discovery, algorithmic approaches often do not provide a direct or embodied experience of the data. The scientist does not necessarily see or hear the data and has little human experience of it. This dissociation brings some risks. Statistics texts give illustrations of false conclusions that can easily occur with such blind approaches: models applied to data that do not fit the models' assumptions silently give incorrect conclusions[5]. Often non-uniqueness problems occur when many different models can be matched to the same data. More generally, this "meaning blindness" may encourage orthodoxy.

Complex relationships between the practices of science, communication about science, aesthetic experience and visual culture are longstanding and continue in to the present. In the time of Darwin, visual atlases produced by skilled artists, engravers and lithographers recorded and displayed subtle nuances and radical discontinuities in variation amongst biological specimens. Analogous to the on-going accrual of massive data, the great scientific expeditions, such as the voyage of the H.M.S. Challenger, resulted in collections of unprecedented numbers of specimens. And just as Darwin faced the challenge of representing natural selection in a 19th Century visual culture invested in the concepts of species fixity[1], we in the 21st Century face representational challenges in engaging with vast and abstract data. Often massive data, such as metagenomics collections that disrupt our organism-centric view of nature and reorient it towards communities of sequences, abstract phenomena outside the human perceptual range. Framing narratives then arise from our choice of algorithms, statistics, representational schemas, displays, interaction technologies, and metaphors used in processing and representing the data. These narratives reflect culturally prevailing ways of seeing and knowing that lack representations for data that engender concepts outside their scope. One particular shift in metaphor is occurring as we move from thinking of knowledge organized through a tree of knowledge, to within our networked society, of knowledge as a network structure. The shift in metaphor necessarily entails shifts in systems of representation. As with renaissance perspective, emergence of new ways of thinking embeds not only epistemological, but also social relationships and relationships to the world.

In combination, the narratives framing data creation and representation circumscribe what we can see and know, and how we see and know. In an era of data richness, we are convinced that we are faced with a crisis of representation. *How do we see and know beyond what our instruments, algorithms, representational schemas and prevailing culture enable us to see and know?* How do we make explicit the implicit assumptions in our systems of representation and their blind spots? In other words, how do we look for what we don't know we're looking for? Artists and scientists each bring their own experimental approaches to making sense and meaning. ArtScience is emerging as one approach for creating novel ways of seeing and new ways of knowing and exploring "hybrid" strategies[10]. *Our work asks: Can experiential, artistic or aesthetic approaches transcend these challenges and enhance our encounters with vast data collections in which the subject of study is high-dimensional, invisible, or otherwise abstract?* In this concept paper we propose a role for artistic and aesthetically impelled approaches developed through ArtScience research and creative work in contributing to the necessary shifts to go beyond the crisis of representation.

2 Aesthetically-Impelled Approaches: ATLAS in silico and Ecce Homology

We define aesthetically-impelled approaches as data-driven, yet neither hypothesis nor problem driven mappings between data and formal elements of its representation, such as shape, color, texture or line quality. Formal elements convey content that relates to or comments upon some aspect of the data or its context. The prior work of one of the authors (RW) in aesthetically-impelled approaches to working with vast and abstract data sets includes two ArtScience collaborations focused on hybrid processes and outcomes. Our conceptualization of DataRemix is informed by lessons learned in their development.

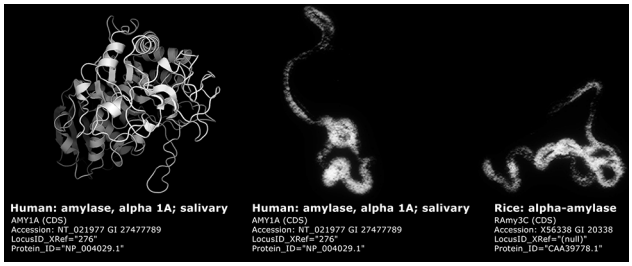

Fig. 2. Comparison of standard 3D ribbon diagram and calligraphic visualizations of human and rice amylase protein. (left) Human amylase, alpha 1A; salivary protein, standard 3D ribbon diagram structure. (middle) Human amylase, alpha 1A; salivary calligraphic protein stroke. (right) Rice alpha-amylase protein calligraphic stroke. Image orientation: The start (N-terminus) of each protein, including the standard 3D structure, is at the top-left of each image/stroke. 3D Representation: A standard representation of the 3D structure of amylase based on the atomic coordinates deposited in the protein data bank. Image rendered with Pymol [44].

2.1 Ecce Homology

Ecce Homology, a physically interactive installation named after Friedrich Nietzsche's Ecce Homo (a meditation on how one becomes what one is) explores human evolution by examining similarities – known as "homology" – between genes from human beings and a target organism, the rice plant[11] (http://insilicov1.org). It offers an aesthetic and meditative encounter with vast and abstract genetic data, visualizes a primary algorithm in comparative genomics as it is working at run-time (BLAST, the Basic Local Alignment Search Tool[12, 13]) and visualizes genomic data as calligraphic forms. Motivated by a desire to move beyond the traditional representations of genetic data as long text strings (A, C, T and G for nucleic acids and the additional twenty single-letter amino acid abbreviations) the unique sequence visualization is based on ideographic and pictographic languages and is reminiscent of Chinese calligraphy or Sanskrit writing.

The custom software developed for *Ecce Homology* visualizes both DNA and amino acid sequences as the strokes and radicals of a non-phonetic calligraphic "alphabet" in which stroke curvature, width, pressure and brush profile are varied according to the characteristics of each sequence. This visualization is not "problem or hypothesis driven" rather it is an aesthetically motivated, high-dimensional, and holistic mapping inspired by the parallels between the way in which protein structure specified by DNA reflects its function in an organism and the manner in which form and visual structure in pictographic languages is directly connected to their meaning. As developed for *Ecce Homology*, calligraphic forms provide a spatially compact visualization and afford the simultaneous display of many genes, enabling the recognition of similarity and homology via pattern recognition. Conventional tools cannot easily obtain this holistic view in a visceral, human-readable format.

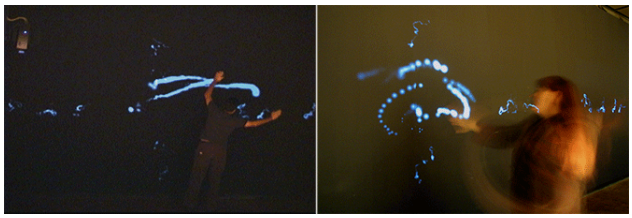

Fig. 3. Interaction within Ecce Homology is designed around an aesthetic of slowness. (left) Slow motion results in continuous and sustained gesture traces. (right) Swift motion results in scattered, rapidly fading gesture traces.

To create each character, a brush stroke generator (BSG) written in C++ reads the nucleotide sequences from GenBank files[14] for the genes that are to be visualized. These are the sole source of input data for the calligraphic strokes–each gene is processed by the same algorithm to generate a unique character. For a given gene, the amino acid sequence expressed by its nucleotides is determined by the BSG and then sent to a remote server[15] for secondary structure and turn prediction. This combined information is used by the BSG to determine the basic shape of the character. Each character is thus fundamentally a two-dimensional structure prediction for a gene, and is represented at this point in the process by a spline output as a text file. Features such as mass to volume ratio, pKa and hydrophobic effect are calculated along the sequence using a hamming window to smooth neighbouring data. These features control width and brush pressure of amino acid strokes. A similar process with different sets of data mappings determines the characteristics of DNA strokes. A naturalistic rendering of gene characters is achieved by modelling a brush depositing ink as it is drawn across a textured sheet of paper.

We visualized the BLAST algorithm as it operates on DNA and amino acid sequences. This required implementing our own BLAST code based on the NCBI distribution[16] and developing methods to visualize intermediary products of the algorithm as it was running and acting upon the calligraphic sequence visualizations. The self-organizing eight-member collaboration that produced *Ecce Homology* spanned the disciplines of new media arts, performance, computer science (graphics and vision), bioinformatics, proteomics, molecular biology and engineering.

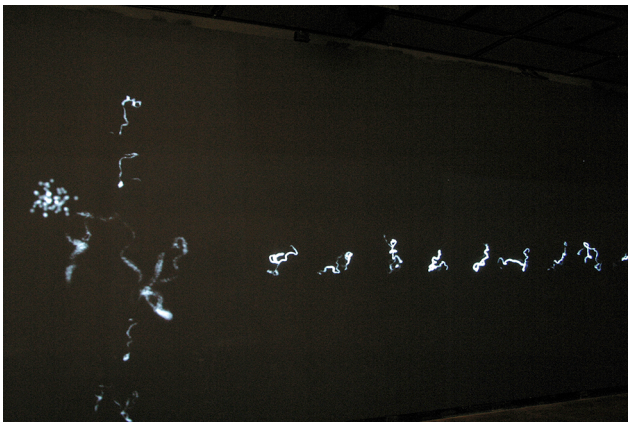

Fig. 4. Human gene character undergoing BLAST. The human gene (translated into protein) selected from the characters on the vertical axis is enlarged in the central area where the viewer's gesture traces had been. The collection of points at the upper left represents the query sequence being segmented into "words" that are compared to the target database sequences depicted on the horizontal axis.

2.2 ATLAS in silico

ATLAS in silico is an interactive virtual environment/installation that places vast and abstract data in a form and context that can be experienced[17] (http://atlasinsilico.net). It utilizes metagenomics data from the Global Ocean Sampling Expedition (GOS) (2003 - 2006) conducted by the J. Craig Venter Institute to study the genetics of communities of marine microorganisms throughout the world's oceans[18]. The GOS circumnavigation was inspired by the HMS Challenger and HMS Beagle expeditions. The microorganisms under study sequester carbon from the atmosphere with potentially significant impacts on global climate, yet the mechanisms by which they do so are poorly understood. Whole genome environmental shotgun sequencing was used to overcome our inability to culture these organisms in the laboratory resulting in a data set which, at the time of its release in 2007, was one of the world's largest metagenomics data set.

Within *ATLAS in silico*'s virtual environment users explore GOS data in combination with contextual metadata at various levels of scale and resolution through interaction with multiple data-driven abstract visual and auditory patterns at different levels of detail. The notion of "context" frames the experience and takes various forms ranging from structuring the virtual environment according to metadata describing the GOS data collection, to the use of socio-economic and environmental contextual metadata pertaining to geographical regions nearest GOS sampling sites to drive visual and auditory pattern generation, to playing a role in both data sonification and audio design that is responsive to user interaction. Participants experience an environment constructed as an abstract visual and auditory pattern that is at once dynamic and coherently structured, yet which only reveals its characteristics as the they disturb the pattern through their exploration.

Again, our approach for visualization and sonification of GOS data and contextual metadata is data-driven yet non-hypothesis nor problem driven. It is a hybrid multi-scale strategy that merges quantitative and qualitative representation with the aim of supporting open-ended, discovery-oriented browsing and exploration of massive multidimensional data collections in ways that do not require a priori knowledge of the relationship between the underlying data and its mapping. Data is presented visually within the virtual environment as dynamic and abstract patterns in different positions relative to the overall virtual world coordinate system and the user (real-world) coordinates. Data is also presented not only as audio objects within the virtual environment, but using spatialization strategies that position and move audio

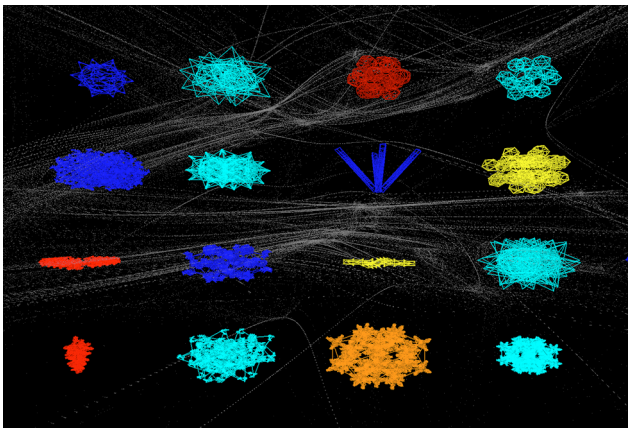

Fig. 5. User-selected meta-shape grammar objects for GOS records are displayed within the virtual environment. Glyphs contain scalable auditory data signatures. Head and hand-tracked tetherless interaction allows users to "play" the selection of glyphs for auditory comparisons and trigger interactive visual features for drill down/drill up operations.

objects relative to the user according to both their interaction with the patterns, and the relation between and within data objects themselves.

This work included generating strategies for structuring virtual environments according to the metadata describing the entire GOS data set and a strategy for multi-scale auditory display, the Scalable Auditory Data Signature[19], that encapsulates different depths of data representation through a process of temporal and spatial scaling. Our research also produced a novel strategy for hybrid audio spatialization and localization in virtual reality[17] that combines both amplitude and delay based strategies simultaneously for different layers of sound and a strategy for scalable visual data representations (n-dimensional glyphs) based on meta shape grammars[20]. The nineteen member collaboration that produced *ATLAS in silico* spanned the disciplines of virtual reality, auditory display and sonification, visualization, electronic music/composition, new media arts, metagenomics, computer graphics, and high-performance computing.

2.3 Polarities: Reproducibility vs. Non-Reproducibility, and Intuition vs. Validity

Developing both *Ecce Homology* and *ATLAS in silico* provided us direct experience with the multiple polarities that arise and need to be negotiated within art-science collaborations when engaging in artistic/aesthetic approaches to working with scientific data[21]. Key amongst these are reproducibility versus non-reproducibility, and intuition versus validity. Reproducibility, the hypotheses and protocols that can be tested against, evaluated, iterated and refined are perceived as "objective" and "rational" and set in opposition to artistic/aesthetic methods that undertake an intuitive process, perceived as "creative" "subjective" and "irrational." The emphasis on uniqueness and virtuosity that is highly valued in the arts is viewed as leading towards products that are non-reproducible, and therefore un-testable in the context of the sciences. We believe these tensions and the role of transdisciplinarity[22] must be acknowledged and become part of the dialog between artists and scientists.

We seek to create aesthetic/artistic approaches to meaningfully complement empirical methods being developed in visualization, information visualization, visual analytics, statistics and machine learning for addressing the crisis of representation. Here three aspects of artistic practice, the externalization of intuition by making; skill in creation and use of metaphor; and skill in the use of additional sensory modalities to make the abstract experiential, can serve to help bridge the objective-subjective divide.

The concept of making as a method for externalizing and recording intuition holds promise as mechanism for establishing a form of provenance to make explicit the assumptions used in the development of aesthetic/artistic approaches to working with scientific data. Artists' skill in the creation and use of metaphor, viewed as a predominant mode of thought and problem solving by Lakoff[23] can foster novel understanding of data and yield new concepts and insights. Emerging concepts of embodied cognition[24] recognize the role of body in regards to thought. Artists longstanding use of additional sensory modes (e.g. auditory, tactile, scent etc.) to engage embodiment can serve to create pan-sensory approaches that bring abstracted data in to an experiential realm and facilitate understanding.

The data-driven, yet non-hypothesis nor problem driven, representational strategies and the whole-body interactive systems developed for *Ecce Homology* and *ATLAS in silico* embody these three dimensions and are only a few examples of work that resides on the edge between objective and subjective. To generalize and evolve these ideas towards broader applicability, we propose the concept of DataRemix.

3 DataRemix: A Working Definition

We define DataRemix as: the reappropriation and recombination, (remixing) of data in any and all states along the continuum of its transformation from raw to processed. We are interested in the use of additional sensory modalities to make the abstract experiential, in combination with the cultural practices of reappropriation and recombination, the externalization of intuition by making, and the creation and use of metaphor. Grounded in the view of remix as a form of literacy for the 21st Century[25] it posits a mechanism to destabilize the narratives framing data creation and representation that contribute to the crisis of representation.

In the context of DataRemix, data is conceptualized as existing along a continuum from raw to processed and behaving akin to a, flux, flow, element, or fluid [26] having multiple transition states rather than as discrete objects. Transitions between raw, partially processed and processed states are triggered by algorithmic, analytic, statistical or other manipulations of the data and decisions by the designers of the system of representation. Novel representational schemas and interactive modalities can be created and explored by externalizing this continuum and making it available to be remixed. Partially processed data and intermediary states along the continuum can be remixed to create different outcomes from those created by the original domain or problem specific processing pipeline. For example, if noise is being removed from raw data with the final end state being "clean" data, then data remixing options would include analysis of the noise and re-incorporation of noise in to different intermediate stages to produce multiple outcomes and representations. DataRemix as a practice and a way to collaborate are linked and made possible by this concept of data as fluid.

Our concept of DataRemix aligns with, and potentially extends, Eduardo Navas's concept of Regenerative Remixes: "juxtaposing two or more elements that are constantly updated, meaning that they are designed to change according to data flow[27, p.8]" Navas distinguishes *regenerative* remixes which require continual updating from three basic types that include *extended* (remixing to extend the duration in time), *selective* (remixing through the process of addition or subtraction of content from the original while retaining its essence), and *reflexive* (creates an allegory, material is added and subtracted, the remixed work stands on its own yet is dependent on the original)[27]. It is likely that our work in developing DataRemix Engines for specific application domains will incorporate remix types that fit multiple, if not all of Navas's classifications.

DataRemix is also an approach to interdisciplinary collaboration characterized by the existence of multiple inputs and objectives, each relevant to one or more disciplines and collaborators, that result in a variety of outputs with different finalities. These types of ArtScience collaborations are facilitated by the view of data as existing along a continuum from raw to processed with transition states that can be accessed and utilized for various purposes, including multiple analyses of the same data. In contrast to ArtScience collaborations that culminate in one overarching hybrid outcome, such as a single artistic installation or publication, these collaborations are characterized by joint group goals being created collaboratively alongside work to progress individual goals as part of an explicit strategy. These may in-

clude creating works of art, new technologies, aesthetically-impelled scientific tools, education outreach, and multiple forms of disciplinary and interdisciplinary knowledge, simultaneously and at various temporal rates.

The collaborative process functions as a path to multiple finalities, each of them valid within some contexts and not within others. Looked at from a "30,000 foot view" this type of collaboration appears somewhat chaotic due to the multiplicity of intersecting trajectories coming in and out of the collaborative flow. But the underlying stream linking multiple inputs, outputs and processes gains coherence at different levels of resolution.

Our definition of DataRemix is a work in progress. We are externalizing it at this conceptual stage to engage in a broader conversation with multiple communities in the arts and sciences to inform its evolution. As the concepts are put in to practice, through this dialog and the evaluation of the collaborative process and its outcomes, we expect the definition to change.

4 INSTRUMENT: ONE ANTARCTIC NIGHT

To operationalize DataRemix and explore its potential we are developing a DataRemix Engine for various sources of scientific data, with a first test bed using astrophysics data. Our collaboration is developing *INSTRUMENT: One Antarctic Night* as a performative and reconfigurable remix installation. *INSTRUMENT* will engage participants in creating, performing and sharing aesthetic multi-modal remixes of astrophysics data using data-driven yet non-hypothesis nor problem driven visualization and sonification strategies. For example, as they remix in "noise" that has been removed by the scientific process, participants "scratch" their own DataRemix versions of the universe and explore how meaning and knowledge is created in science and culture. The art-science collaboration blends new media art, astrophysics, electronic music, sonification, visualization, immersive systems, online and mobile participatory media and emerging technologies in holographic sound environments.

As a DataRemix collaboration our shared objective is to develop an artwork that makes use of scientific data and has multiple reconfigurable footprints and instantiations, while simultaneously producing individually focused outcomes. These include research in experimental data sonification and visualization, test-bedding holographic sound environment technologies, and developing aesthetically/artistically impelled scientific tools for use in astrophysics research and education outreach. Below we detail our rationale for working with astrophysics data, and in the following section discuss our approach to the design and development of the DataRemix engine for *INSTRUMENT*.

Astronomy was one of the first sciences to make the transition to primary reliance on digital data and experience the benefits of big data. The international astronomy community invested heavily in the 1980s and 1990s in developing international data standards, open access middleware and open access databases. The resulting "virtual observatories" have proved to be remarkably productive, allowing scientists who were not the originators of the data to use the data for new kinds of research questions. The International Virtual Observatory Alliance (http://www.ivoa.net/) coordinates strategies that allow data from different observatories, types of telescopes, and time periods to be combined for scientific research purposes. Astronomical funding agencies have in many cases required that data be made public and as a result Virtual Observatories have allowed astronomers outside of the observatory groups to analyze the data and make discoveries both in the original data through different analyses or by combining data seamlessly from different instruments (sometimes called "multi-wavelength" astronomy). As a result research questions not posed by the data takers have been addressed by other scientists.

In the US, NASA was one of the early funding agencies to encourage education outreach for scientists in NASA projects to develop innovative work to present their work to different kinds of publics, a policy which has developed internationally in the astronomy research community[28]. Together with the open data policies, this has resulted is an extensive body of work of "space-art" by artists using astronomical data (see the Space and the Arts bibliography at http://www.olats.org/space/biblio/bibliography.php).

Finally the virtual observatory movement also enabled many of the very early citizen science applications by interested publics. For instance the NASA exoplanet search Kepler mission has established a "planet hunter" program (http://www.planethunters.org/) that makes it possible for anyone to sieve through the data taken. The Kepler spacecraft takes brightness measurements, or "light curves," of over 150,000 stars every 30 minutes. People can then hunt for planets by looking for a brief dip in brightness that occurs when a planet passes in front of the star. Many published results (https://www.zooniverse.org/publications) now result from the involvement of non-scientist collaborators.

Given the very open situation with respect to astronomical data and their heterogeneous nature, such data sets offer ideal test beds for the DataRemix approach.

4.1 The Chinese Small Telescope Array Data

Within this context, we contacted one of the co-authors Lifan Wang, co-investigator of the Chinese Small Telescope ARray, CSTAR, in Antarctica (http://mcba11.phys.unsw.edu.au/ plato/cstar.html). Deployed in 2008 on Dome A, the highest point in Antarctica, this robotic telescope takes advantage of the unique observing conditions in the infrared. The four 14.5 cm diameter telescopes have a field of view of 4.5 x4.5 degrees, The cameras take images every 30 seconds during the entire Antarctic night of more than 160 days duration. In the results reported in 2011[28] there were approximately 100,000 stars detected in the master reference images, reaching a depth of i = 20.4 mag. One resulting catalogue of over 10,000 individual sources reaches a limiting magnitude of 16 located in a 23 square-degree region centered on the south celestial pole. Each source monitored allows the derivation of detailed light curves that are then used to detect transient events[30]. The photometric catalogue can be used for studying any variability in these sources, and for the discovery of transient sources such as supernovae, gamma-ray bursts, minor planets and occultation searches of transits by exoplanets. Additional data sets are available from four whole sky webcams and video cameras which both monitor the instrument conditions and the sky conditions. A typical data set from one observing season for instance consisted of 287,800 frames over 1615 hours with a total data set size of 350Gb.

Algorithmic analysis iteratively refines raw CSTAR images into progressively processed data. The raw data is manipulated repeatedly by the scientists to put the data in a form that can be analysed for the intended purpose; technical procedures are carried out step by step, such as "flatfielding" or removing the effect of sensitivity variations, the subtraction of the "background," corrections for the offset or zero point of the camera, exposure times are calibrated and normalized, brightnesses are calibrated and uncertainties rescaled, satellite trails and areas of saturation are masked, and master images are registered to create an optimal balance of "signal to noise. Because of the interference caused by atmospheric conditions 3000 of the best images obtained during a 24-hour period of exceptionally good conditions were used to form a master reference image. In one discussion of the data processing pipeline over 9 separate steps are itemized and checked. The large pixel scale of the camera and the relatively high stellar density of the field makes blending a problem. A blend is defined as two or more stars located within 30 arcseconds (2 CSTAR pixels) of each other. A difference-imaging photometry pipeline has been implemented that will enable a more robust detection of variables in crowded environments and will deliver higher photometric precision. Because of the complexity of the data analysis processing for some data sets two groups have carried out independent analyses of the data; one at the National Astronomical Observatories of the Chinese Academy of Sciences and another at Texas A&M University and Beijing Normal University. A comparison of the photometric precision of the two reductions allows processing errors to be detected. The concept of multiple analyses of the same data is one that is central to the DataRemix concept.

4.2 Developing a DataRemix Engine for CSTAR Data

To create *INSTRUMENT* we are developing a DataRemix Engine for CSTAR data. Similar to our work in *Ecce Homology* that externalized the functioning of the BLAST algorithm at run-time and visualized the ranking of high-scoring pairs and alignment of sequences, design of a remix engine for *INSTRUMENT* requires getting "under the hood" of the original CSTAR data processing pipelines.

Our initial work will map algorithmic and technical elements of the CSTAR pipeline, as well as the inflection points in the pipeline that represent state transitions as the data flows from raw to processed. This mapping will drive the next level of design for formulating and delineating auditory, visual, and interactive structures within the remix engine. An additional layer will map these elements to remix algorithms, user interface elements and user experience flows.

The result will be an initial array of remix tools to allow users to perform traditional remix operations such as "sampling" "cut" "copy" and "paste." To accomplish this will require work to define basic elements such as what constitutes an auditory "sample," a "beat," or a visual "mashup" element relevant to astrophysics data. In addition to developing data-driven yet non-hypothesis nor problem driven experimental visualizations, we will explore reappropriating algorithms utilized in processing CSTAR imaging data to develop alternative representational strategies. We will also explore the application of quantitative and qualitative listening strategies[31, 32, 33]. These strategies allow listening for quantities of specific values in sound or for abstract and subjective qualities of a sound. Their combination allows for communicating specific data values or comparative differences. And we will explore the application and extension of prior research in scalable auditory data signatures[19] developed as a part of *ATLAS in silico*. Scalable Auditory Data Signatures are entities of auditory representation that encapsulate different depths of data display, achieved through a process of temporal and spatial scaling and which create an audio-visual perceptual synthesis which Michel Chion calls *syncretism,* establishing a perceptual coherence between the different layers of sound and the visual aspects of the data representation[34].

As an example of individually focused research that proceeds as part of a DataRemix collaboration yet progresses at a different rate and has discrete outputs with unique finalities, we will undertake the design, development and test-bedding of holographic sound environments. The concept of holographic sound is that sound sources are delocalized and then re-localized as discrete sound objects while moving dynamically with listeners in the space. We will conduct research in to development of an electroacoustic control system building upon a principle called, Boundary Surface Control (BSC)[35]. BSC is mathematically expressed as different interpretations of the Kirchhoff-Helmholtz Integral Equation. These equations provide a framework to create the discrete sound objects, i.e., "sound holograms," and localize them at arbitrary points in a volume space. To efficiently and accurately control wave-fronts, a new form of loudspeaker is needed to generate sound. Previous three-dimensional sound installations require the use of hundreds, if not thousands of loudspeakers to synchronously distribute sound by using time-delays and/or customized power distribution layers, such as the geodesic sound projection map, SpaceMap[36]. Our research will explore the use of solid-state loudspeakers made of Multi-Walled Carbon Nanotubes (CNT)[37] assembled in sheets of varying size and shape, and positioned along the geometry of the space giving a planar surface of which sound can be generated. This makes them an ideal candidate for synthesizing wave fronts. The public-facing remix installation serves as an application use case for research in development of CNT speakers and holographic sound, and provides real-world contexts for comparative benchmarking with commercially available sound distribution systems.

As we develop the remix engine, our interactions with Lifan Wang and the CSTAR researchers will allow us to iterate in a direction that we hope will result in new aesthetically-impelled tools for working with their data. These represent a first step towards evaluating the potential of DataRemix in creating strategies to represent concepts that are not easily or readily conveyed by existing approaches, such as dark energy or matter. In addition to the tools, we will develop interfaces and user experience suitable for multiple formats onsite in museum spaces, in online and mobile contexts, and as science-outreach.

The strategy that we propose in DataRemix, of remixing or reprocessing in other ways at different stages of the data pipeline, departs from most art-science projects which more often than not use the 'cleaned' or processed data that is usually the form provided for education outreach or science information purposes (including scientific publications). ArtScience projects which have intentionally used raw or partially processed scientific data include the work of Semi-Conductor (http://semiconductorfilms.com/). During an artists residency at the Space Sciences Laboratory at UC Berkeley they produced films that exploited defects and noise in the unprocessed data of solar observations to produce a work "Brilliant Noise" (http://semiconductorfilms.com/art/brilliant-noise/) which intentionally assigns meaning to parts of the data cleaned out by the scientists in their processing; this work could be described as an early form of Glitch art. In the context of New Media Art, Glitch is a genre that makes use of unexpected results of technical malfunctions. It creates art from imperfection to create new meanings. This resonates with the sampling strategies of DataRemix. Elements proposed for this remix engine include analysing the raw CSTAR camera data in ways that were not used for the scientific studies (e.g. analyses that do not respect the time sequence), analysing data that was cleaned from the data for scientific purposes at different steps of the processing pipeline, and the ability to remix in data from other wavelengths.

DataRemix is situated in digital culture where remix strategies are common and often drive products that treat data as data flow rather than data objects, and displays the work in performative and interactive forms that are not common for the presentation of scientific data. Education Outreach oriented art-science projects sometimes reify the final presentation as definitive in some way; the DataRemix reinforces an understanding that meaning emerges from a complex set of decisions and processes that may not be visible in the final product.

5 But what does it mean?

Remixing and appropriation have a long and rich tradition in the arts. From traditional collage to DJ culture, rearranging pre-existing visual and sonic information in order to create new meaning has been a long time preoccupation for artists of all backgrounds.

The term collage is attributed to Pablo Picasso and derives form the French "coller" meaning, "to glue." In his *Still Life Chair Caning* (1912) the Spanish painter pasted a printed image of chair caning onto his painting. By appropriating and combining this material into a larger whole Picasso showed a path that artists would explore for decades to come.

Around the same time that Picasso's first collage was being created, the French artist Marcel Duchamp had an idea that would radically change what was considered to be art. *Bicycle Wheel* (1913) is Duchamp's first *readymade*, a set of objects he conceived to challenge assumptions about what is a work of art. Duchamp combined two mass-produced parts - a bicycle wheel and a kitchen stool - to create a type of non-functional machine. By simply selecting prefabricated items and calling them art, he subverted established notions of the artist's craft and the viewer's aesthetic experience.

In literature, the recombination of pre-existing material was started by Dadaist poet and Duchamp's friend Tristan Tzara. He composed a poem by randomly pulling out of a bag various words that were previously cut out from a newspaper. The technique had been published in an issue of 391 along with the poem by Tzara called, *dada manifesto on feeble love and bitter love* (1920).

This random combination of words was a precedent to the *cut up* and *folds in* techniques. The first is performed by taking a finished and fully linear text and cutting it in pieces with few words on each piece; the resulting pieces are then rearranged into a new text. The fold-in technique works by taking two sheets of linear text folding each sheet in half vertically and combining with the other, then reading across the resulting page.

Both techniques were popularized in the late 1950s and early 1960s by writer William S. Burroughs. His preoccupation with the decon-

struction of words and language took him to borrow from the collage technique of visual artists to explore a fragmentary and non-linear approach to narrative. His method linked fragments of texts in surprising juxtapositions, offering unexpected leaps into uncharted territories that attempt to shake and ultimately transform the consciousness of the reader. For Burroughs, narrative operates as a vast, multi-threaded network that reflects the associative tendencies of the mind, underlying undetected connections, drawing attention to the links between disparate ideas and elements.

Remix, as we now know it, comes from the contemporary cultural phenomenon of recombination and appropriation in popular music and DJ Culture. Since the early stages of recorded sound in the late 19th century, technology has allowed individuals to reorganize the normal listening experience. With the arrival of easily editable magnetic tape in the 1940s and 1950s and the subsequent development of multi-track recording, such alterations became more common. Remix exists in a variety of forms with differing structures and emphasis in their relation to the original work. In its most simple form a remix is a song that has been edited to sound different from the original version. This can be achieved by rearranging the multiple tracks that constitutes a song or by the juxtaposition of bits and parts of different songs. Similar to the collage, the remix can be made of the combination of original artworks with other materials.

Lessig has argued that digital remix is the 21st Century literacy and as fundamental as reading[25]. To *copy and paste* data as we surf the web while doing research or just looking for a recipe has become so ubiquitous that we hardly think about it. To copy is to separate a piece of information from any original context or meaning so that it might be free to function otherwise. To paste is to reinscribe, to place the floating sample into a new chain of signification.

The idea of remixing sonic and visual information has given birth to a plethora of sub-entities that inhabit the digital biosphere, from mashups to animated gifs these events of ephemeral creativity prove that the joy of recombination rests not only in the final result of a specific mix, but in the performative aspect of remixing. The potential of endless reorganization for the emergence of new meaning is at the core of the idea of the remix. Appropriation and recombination are creation, following Paul D. Miller (aka DJ Spooky): "Give me two records and I will make you a universe..."

DataRemix proposes that we have all become information DJs.

6 DataRemix as Datamade

In both process and outcomes, the *"datamades"* resulting from DataRemix are envisioned to function analogously to Duchamp's *readymades*. Their ultimate objective is to destabilize the framing narratives of data creation and representation in order to generate the possibility for new forms to arise in hopes of allowing us to see and know beyond what our instruments, algorithms, representational schemas and prevailing culture enable us to see and know. Yet, reappropriation and recombination also bring with them the framing narratives of artistic traditions from the early 20th Century that continue to evolve in our digital culture. These carry an aura of arbitrariness that runs counter to the functioning of science which requires reproducibility and validity. This very contradiction is at the heart of our working definition of DataRemix. In proposing DataRemix we hope to contribute to the dialog about arbitrariness already ongoing in the visualization community. Maintaining the dichotomy of artistic approaches as devoid of meaning, decorative or subjective and non-artistic approaches as meaningful, valid and objective eschews the practical reality that, as Monroe observes, visualization is inherently aesthetic and created for an intended audience, and iterates towards the audience as part of the analytic process[21]. Additionally, familiarity with a representational schema enables us to forget that at one point elements of its design were also based on arbitrary yet repeatable mappings that lead to their utility and meaning. Stylistic and aesthetic concerns are increasingly a subject of study in the VIS and HCI communities [38, 39, 40, 41]. As Viegas and Wattenberg reflect, the power of artistic data visualization arises from artists "committing various sins of visual analytics" and directly engaging and guiding an audience towards a point of view[43]. They remind us that even with dispassionate analysis as its goal, creating a visualization that is truly neutral is "generally impossible" and propose further exploration of the value of artistic explorations[43]. In this light, we propose to explore DataRemix as a mechanism for artistic approaches to engage empirical approaches in creating new ways of seeing and knowing.

Acknowledgments

The authors thank the Chinese Small Telescope ARray, CSTAR, in Antarctica for making available CSTAR data. We thank the University of Texas Dallas ATEC program for its support of this work. We thank Frank DuFor for his invaluable advice and encouragement. We also thank the University of North Texas xREZ lab for its support of this work. Development of *ATLAS in silico* was supported in part by NSF award IIS – 0841031. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] Smith, J. (2006) Charles Darwin and Victorian Visual Culture. Cambridge University Press

[2] Gantz J.F., Reinsel D. (2012) The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East – United States, IDC. Online: http://www.emc.com/leadership/digital-universe/iview/index.htm Last accessed: 8/31/13

[3] danah boyd & Kate Crawford (2012) Critical Questions For Big Data, Information, Communication & Society, 15:5, 662-679

[4] Zaragoza, M. S., Belli, R., & Payment, K. E., (2007). Misinformation effects and the suggestibility of eyewitness memory. In M. Garry & H. Hayne (Eds.), Do Justice and Let the Sky Fall: Elizabeth F. Loftus and Her Contributions to Science, Law, and Academic Freedom (pp. 35-63). Mahwah, NJ: Lawrence Erlbaum Associates

[5] Cleveland, W, (1993) Visualizing Data, Hobart Press.

[6] Naur p. (1995) Knowing and the Mystique of Logic and Rules, Kluwer Academic

[7] Ackoff, R.L. (1989) "From Data to Wisdom", Journal of Applied Systems Analysis, Volume 16, 1989 p 3-9

[8] Quantified Self Movement. Online: http://quantifiedself.com/ Last accessed: 8/31/13

[9] Boorstin, Daniel J. (1994) Cleopatra's Nose: Essays on the Unexpected, New York: Random House

[10] Malina R.F., Strohecker S, LaFayette C, and Ione A. (2013) Steps to an Ecology of Networked Knowledge and Innovation: Enabling new forms of collaboration among sciences, engineering, arts, and design" http://seadnetwork.wordpress.com/draft-overview-of-a-report-on-the-sead-white-papers/ Last accessed: 8/31/13

[11] West R, Burke J, Kerfeld C, Mendelowitz E, Holton T, Lewis JP, Drucker E, Yan W. (2005) Both and Neither: in silico v1.0, *Ecce Homology*. Leonardo, 38 (4): 286-293

[12] Altschul, S. F., et al. (1990) "Basic local alignment search tool." J. Mol. Biol. 215: 403-410

[13] Altschul, S. F., et al. (1997) "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Research 25(17):3389-3402.

[14] National Center for Biotechnology, GenBank : http://www.ncbi.nlm.nih.gov/genbank/

[15] Kaur, H. and Raghava, G.P.S. (2002) BetaTpred: Prediction of beta-turns in a protein using statistical algorithms. Bioinformatics 18:498-9. http://www.imtech.res.in/raghava/betatpred

[16] NCBI BLAST Code: http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE= BlastDocs&DOC_TYPE=DeveloperInfo

[17] West R, Gossmann J, Margolis T, Schulze JP, Lewis JP, Tenedorio D. (2009). Sensate abstraction: hybrid strategies for multi-dimensional data in expressive virtual reality contexts. Proceedings of the 21st Annual SPIE Symposium on Electronic Imaging, The Engineering Reality of Virtual Reality, 18-22 January 2009 San Jose, California, Volume 7238, pp. 72380I-72380I-11

[18] Shibu Yooseph, Granger Sutton, Douglas B Rusch, et al. (2007) The Sorcerer II Global Ocean Sampling Expedition: Expanding the Universe of Protein Families, PLOS Biology: 13 Mar 2007 — info:doi/10.1371/journal.pbio.0050016

[19] Gossman J, Hackbarth B, and West, R with Margolis T, Lewis J.P., and Mostafavi I. (2008) Scalable Auditory Data Signatures for Discovery Oriented Browsing in an Expressive Context. Proceedings of the 14th International Conference on Auditory Display, June 24 - 28, 2008, Paris

[20] Lewis, J.P., Rosenholtz, R., Fong, N., Neumann, U. VisualIDs: Automatic Distinctive Icons for Desktop Interfaces, ACM Trans. Graphics Volume 23, #3 (August 2004), pp. 416-423

[21] West R., Monroe L, Ford Morie J, Aguilera J. (2013) Art, science, and immersion: data-driven experiences. Proc. SPIE. 8649, The Engineering Reality of Virtual Reality 2013 86490H

[22] Punt M. (2010) Boundaries and Interfaces: Transdisciplinarity and the Big Data Problem, Leonardo Quarterly Reviews, 1.02, p 5

[23] Lakoff, G. and Johnson, M., (1999) Philosophy In The Flesh: the Embodied Mind and its Challenge to Western Thought, Basic Books

[24] Wilson, M., (2004) Six views of embodied cognition, Psychonomic Bulletin and Review, 9, 625-636

[25] Lessig L. (2008) REMIX Making Art and Commerce Thrive in The Hybrid Economy. Bloomsbury Academic, London

[26] Malina R. (2010) Big Data, Citizen Science and the Death of the University, Leonardo Quarterly Reviews, 1.02, p 2

[27] Navas, E. (2009) Regressive and Reflexive Mashups in Sampling Culture, in Sonvilla-Weiss, Stefan (Ed.) Mashup Cultures, Springer.

[28] Hawkins, I., Battle, R., Christian, C., & Malina, R. (1994) Science On-Line: Partnership Approach for the Creation of Internet-Based Classroom Resources. Astronomy education: current developments, future coordination Astronomical Society of the Pacific Conference Series, Volume 89, Proceedings of an ASP symposium held in College Park, MD, 24-25 June 1994, San Francisco: Astronomical Society of the Pacific (ASP), —c1996, edited by John A. Percy, p.215 Bibliographic Code: 1996ASPC89215H

[29] Wang, L. et al. (2011) Photometry Of Variable Stars From Dome A, Antarctica The Astronomical Journal 142

[30] Xu Zhou et al., The First Release of the CSTAR Point Source Catalog from Dome A, Antarctica Published by: The University of Chicago Press on behalf of the Astronomical Society of the Pacific

[31] Bruce N. Walker and Joshua T. Cothran, (2008) Sonification Sandbox: A Graphical Toolkit for Auditory Graphs, in Proc. 2008 Int. Conf. On Auditory Display (ICAD), Boston, MA, U.S.A

[32] William W. Gaver, (1993) Synthesizing Auditory Icons, in Proc. INTERCHI 93, Amsterdam, Netherlands, April

[33] Gossmann, J. (2005) Toward an Auditory Representation of Complexity in Proc. 2005 Int. Conf. On Auditory Display (ICAD), Limerick, Ireland,

[34] Chion M. (1994) Audio-Vision. Columbia University Press, New York

[35] Ise, S. (2005) The Boundary Surface Control Principle and It's Applications IECICE TRANS. Fundamentals, vol.E88-A, no.7, July

[36] Ellison, S. (2013) SpaceMap: 20 Years of Audio Origami, Lighting and Sound America, April.

[37] Kozlov M, Haines C, Oh J, Lima M, Fang S. (2009) Sound of Carbon Nanotube Assemblies. J. Appl. Phys.106, 124311

[38] Cawthon N., and Vande Moere A. (2007) The Effect of Aesthetic on the Usability of Data Visualization. IEEE Conference on Information Visualization (IV'07), pp. 637-648

[39] Tractinsky, N. (2004) Towards the Study of Aesthetics in Information Technology. Conference on Information Systems, pp. 771-780

[40] Salimun C., Purchase H., Simmons D.R., and Brewster S. (2010) The Effect of Aesthetically Pleasing Composition on Visual Search Performance. Nordic Conference on Human-Computer Interaction (NordiCHI'10), ACM, pp. 422-431

[41] Tractinsky N., Katz A.S., and Ikar D. (2000) What is Beautiful is Usable. Interacting with Computers, vol. 13, no. 2, pp. 127-145

[42] Vande Moere A., Tomitsch M., Wimmer C., Boesch C., and Grechenig T. (2012) Evaluating the Effect of Style in Information Visualization, Visualization and Computer Graphics, IEEE Transactions on , vol.18, no.12, pp.2739,2748

[43] Viegas FB., Wattenberg M. (2007) Artistic Data Visualization: Beyond Visual Analytics in D. Schuler (Ed.): Online Communities and Social Computing., HCII 2007, LNCS 4564, pp. 182–191, 2007. Springer-Verlag Berlin Heidelberg

[44] W.L. DeLano, The PyMOL Molecular Graphics System (San Carlos, CA: DeLano Scientific, 2002)