8-13-2016

Human computer interaction and data visualization

Olivia Hsieh
Santa Clara University

ENGLISH 106: ADVANCED WRITING

HUMAN COMPUTER INTERACTION AND DATA VISUALIZATION

Olivia Hsieh

ENGL 105: Advanced Writing

13 August 2016

ENGLISH 106: ADVANCED WRITING
HUMAN COMPUTER INTERACTION AND DATA VISUALIZATION

Introduction

In 2013, SINTEF cited that 90% of the world's data had been created over the past two years.[1] With the onset of Big Data, the ability to analyze data at a rate directly proportional to its collection becomes quite near impossible, albeit nonetheless important. In "Play With Data – An Exploration of Play Analytics and Its Effect on Player Experiences," Ben Medler explains its pertinence: "Data can be given a different context through relating it to other data. A relation informs someone of how data can be correlated or combined with other data."[2]

To better make sense of those relations, one can represent data in different forms. Below is the first known line chart made by William Playfair:

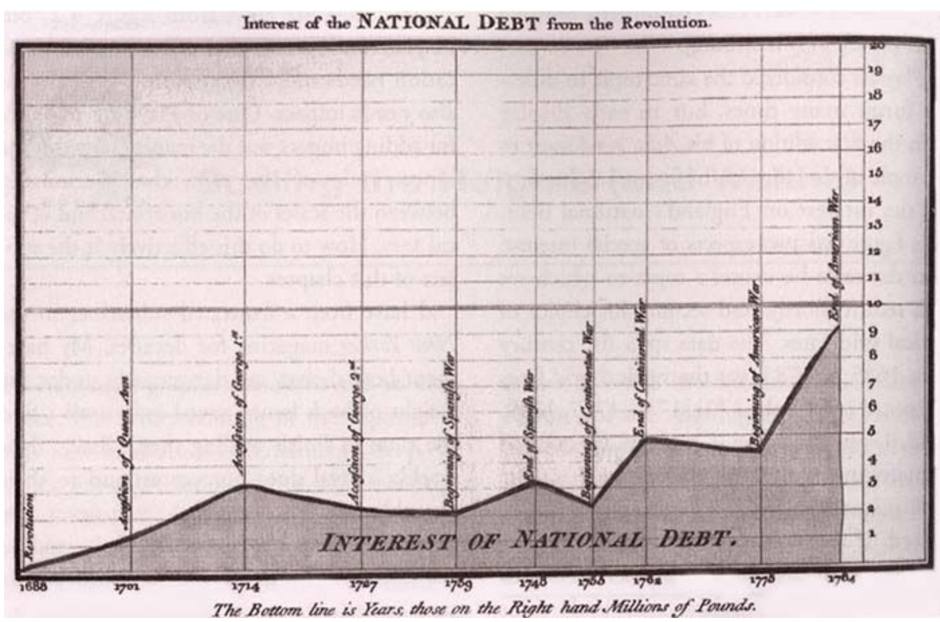


Fig. 1: Line Chart of National Debt from the Revolution[3]

[1] SINTEF. "Big Data, for better or worse: 90% of world's data generated over last two years."
[2] Medler, Ben, "Play With Data – An Exploration of Play Analytics and Its Effect on Player Experiences"
[3] Playfair, William, "Interest of national debt from the Revolution"

As opposed to looking at a table which lists the information in the graph, one can much more quickly arrive at the conclusion that national debt from the Revolution has generally risen over the course of a century. Due to the way the mind processes information, data visualization enables one to better recognize trends, patterns, and correlations and thus meaningfully analyze and provide context to data. For this reason, data visualization, or the creation of graphics to represent data, has seen a rise in popularity.

At times representing all information of interest in a static visualization can pose a challenge, particularly when one wishes to determine whether and how multiple factors affect an outcome. Yi et al. identify this fundamental issue in "Toward a Deeper Understanding of the Role of Interaction in Information Visualization": "While static images clearly have analytic and expressive value, their usefulness becomes more limited as the data set that they represent grows larger with more variables." [4] Consider any graph database: as the number of relationships rises exponentially, it proves increasingly difficult to locate a specific relationship without the need to zoom in and follow a trail of relationships. In an age where the sheer amount of data exceeds the ability to properly analyze and provide context to it, it becomes necessary to determine more optimal ways to understand data.

One way currently being explored is the addition of interaction in data visualization. As such, this paper serves to define the ways in which human computer interaction (HCI) strengthen data science. It first defines data and information and examines the purpose of data visualization via the effect of Gestalt grouping on the Stroop task. The second section outlines Ben Fry's process for understanding data and Ji Soo Yi and Youan ah Kang's taxonomic system of

[4] Yi et. al, "Toward a Deeper Understanding of the Role of Interaction in Information Visualization"

interaction types, laying the foundation for evaluating the efficacy of a visualization. The final section addresses the role of HCI in data science and, utilizing the frameworks defined in the second section, looks at why interaction is integral to modern data visualization.

Gestalt Grouping and the Importance of Visualization

Cognitive psychology lays the foundation for human perception; in particular, Gestalt grouping deals with the way the human mind associates elements and determines whether they belong to a whole or function independently. Consider their role, specifically concerning the laws of similarity and proximity, in the Stroop task in "Role of Gestalt Grouping in Selective Attention Evidence from the Stroop Task." In their experiment, Lamers and Roelofs conduct a modified version of the Stroop task to evaluate the effect of congruent and incongruent stimuli on reaction time.

Whereas the Stroop task traditionally evaluates an individual's ability to recognize a color word , which can either be printed in the color named (congruent stimuli) or in a different color (incongruent stimuli), Lamers and Roelof add an additional variable: space. The following diagram represents the four test cases, with "Continuous" as the control and the grey bars representing color patches of the colors red, blue, and green.

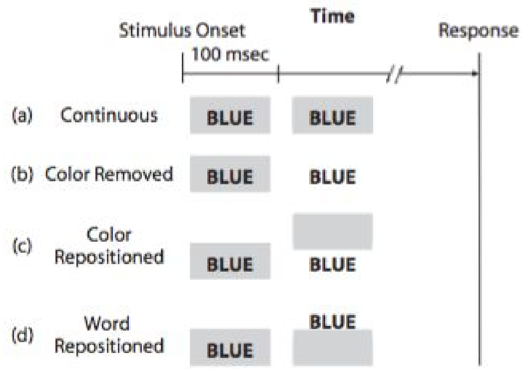


Fig. 2: Schematic of Four Example Conditions from Lamers and Roelof's Experiment[5]

The two find that separation of color and word temporally and spatially at least partially inhibits the individual's association of the color with the word. Their experiment highlights the basis for data visualization: the human mind better distinguishes elements with a distinctive enough separation and, generally, more accurately forms relationships between data presented in a way attuned to the intricacies of human perception.

Norman expands on the necessity of visualization:

> "The power of the unaided mind is highly overrated. Without external aids, memory, thought, and reasoning are all constrained. But human intelligence is highly flexible and adaptive, superb at inventing procedures and objects that overcome its own limits. The real powers come from devising external aids that enhance cognitive abilities. How have we increased memory, thought, and reasoning? By the invention of external aids: It is things that make us smart." [6]

---

[5] Martijn J.M. Lamers and Ardi Roelofs, "Role of Gestalt grouping in selective attention: evidence from the Stroop task"

[6] Norman, Donald A. "Cognition in the head and in the world"

In other words, while the human mind as is is capable of complex analysis, it further excels when information in need of analysis is transformed into a format accommodating of cognitive abilities, such as those provided through data visualization.

Defining Data and Information

To begin looking at data visualization, one must begin with a definition of data. As data as a whole encompasses a wide stretch of forms (field data, raw data, experimental data, etc.), for the purpose of evaluating it at a higher level, a general definition suffices: data is a set of quantitative and/or qualitative attributes about an event. However without some sort of transformation or processing, it remains simply so. Stahl cites William M. Ulrich in his definition of information: "Where data are the raw facts of the world, information is then data 'with meaning.' 'When "data" acquires context-dependent meaning and relevance, it becomes information.'" [7] In order to transition from data to information, or data with context, one must prime it for analysis. Floridi and Sanders explain, "The technical-functional notion of the relationship between data and information is that information is data that has been processed (or interpreted) in order to make it useful, sensible, or meaningful." [8] Visualization provides this bridge from raw data to analyzation.

A Framework for Understanding Data

Ben Fry breaks down the process of analyzing data into seven steps: acquire, parse, filter, mine, represent, and interact. In an initial example where he details the Unites States Postal Service (USPS) Zoning Improvement Plan (ZIP) numbering system, he explains: "Each step of

[7] Stahl, Bernd Carsten, "Information Systems: Critical Perspectives"
[8] L. Floridi and Sanders J.W., "On the Morality of Artificial Agents"

the process is inextricably linked because of how they affect one another" and that current attempts at understanding data fail to address the process as a whole, with transitions between each step leaving room for the process to become further convoluted.[9]

Fry begins with an interest in the correlation between ZIP codes and location, which leads to the creation of *zipdecode*, an application that attempts to find geographical meaning from a ZIP code. He first acquires 45,000 ZIP codes from the U.S. Census Bureau and turns it into a usable format for filtering and creating the visualization. After he refines the initial visualization, he adds a level of interaction by allowing users to replace the last letter of a query as well as zoom in and out. In doing so, he realizes the lack of a transition makes it difficult for the user to comprehend that a change in the query is being made. To help the user understand the occurrence of the change in the last letter, Fry adds a semi-slow color change to help the user maintain context.

Beyond the necessity of every step in his process, implicit in his example is the need for interaction in allowing for user directed exploration. Yi et. al expand on representation and interaction as the two major components of data visualization; they argue the two are not mutually exclusive and parallel Fry in his view on the ability of interaction to impact other stages in the process: "Interaction with a system may activate a change in representation." [10] Since interaction places more emphasis on user direction, if a user were to struggle with locating data, this would be immediately made clear and call for a design solution. In other words, the

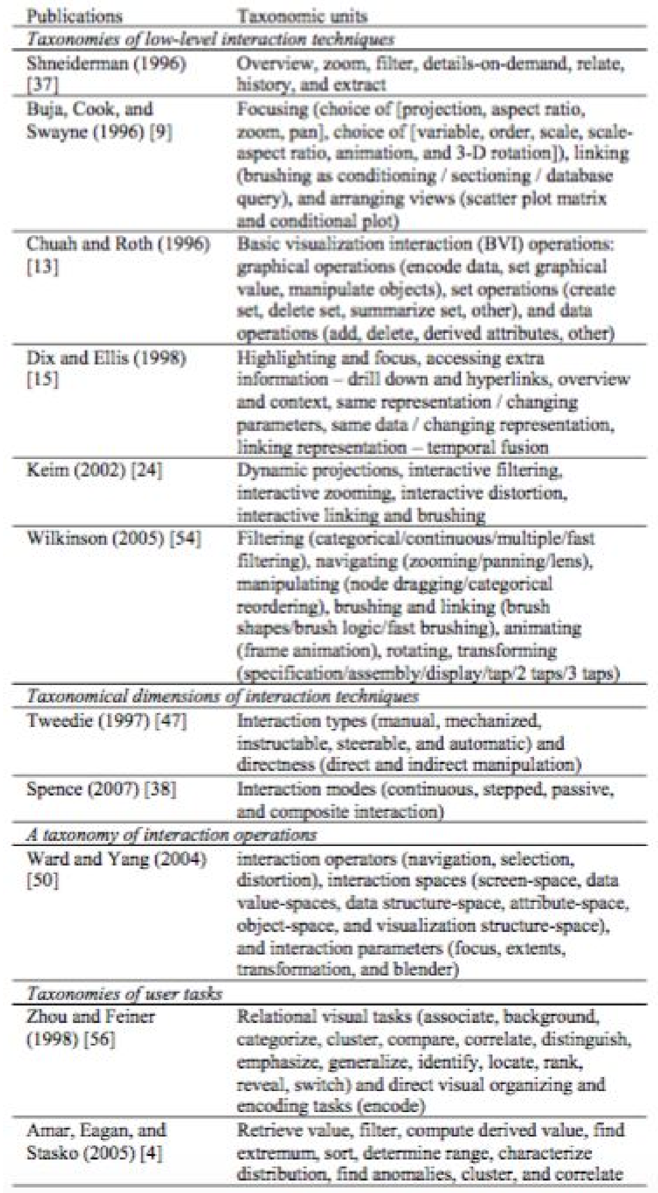
[9] Fry, Ben, "Computational Information Design"
[10] Yi et. al, "Toward a Deeper Understanding of the Role of Interaction in Information Visualization"

instantaneous nature of interactive data visualization makes identification of areas needing refinement much easier.

A Framework for Understanding Interaction

The presence of large amounts of data can inhibit a user's ability to observe distinct patterns. Dissatisfied with the inability of existing taxonomies to encompass both high and low level views of interaction and bridge interaction techniques with user objectives, Ji Soo Yi and Youn ah Kang offer a new framework to better gauge evaluate the efficacy of existing interaction techniques.

In their study, they uncover seven prominent categories of interaction: selection, exploration, reconfiguration, encoding, abstraction/elaboration, filtering, and connection. These seven categories provide a basis upon which one can understand potential ways a user engages with interactive visualizations. Understanding these concepts will enable one to understand interaction at a lower level and establish a common vocabulary for discussing how interaction strengthens data visualization.



Publications	Taxonomic units
Taxonomies of low-level interaction techniques	
Shneiderman (1996) [37]	Overview, zoom, filter, details-on-demand, relate, history, and extract
Buja, Cook, and Swayne (1996) [9]	Focusing (choice of [projection, aspect ratio, zoom, pan], choice of [variable, order, scale, scale-aspect ratio, animation, and 3-D rotation]), linking (brushing as conditioning / sectioning / database query), and arranging views (scatter plot matrix and conditional plot)
Chuah and Roth (1996) [13]	Basic visualization interaction (BVI) operations: graphical operations (encode data, set graphical value, manipulate objects), set operations (create set, delete set, summarize set, other), and data operations (add, delete, derived attributes, other)
Dix and Ellis (1998) [15]	Highlighting and focus, accessing extra information – drill down and hyperlinks, overview and context, same representation / changing parameters, same data / changing representation, linking representation – temporal fusion
Keim (2002) [24]	Dynamic projections, interactive filtering, interactive zooming, interactive distortion, interactive linking and brushing
Wilkinson (2005) [54]	Filtering (categorical/continuous/multiple/fast filtering), navigating (zooming/panning/lens), manipulating (node dragging/categorical reordering), brushing and linking (brush shapes/brush logic/fast brushing), animating (frame animation), rotating, transforming (specification/assembly/display/tap/2 taps/3 taps)
Taxonomical dimensions of interaction techniques	
Tweedie (1997) [47]	Interaction types (manual, mechanized, instructable, steerable, and automatic) and directness (direct and indirect manipulation)
Spence (2007) [38]	Interaction modes (continuous, stepped, passive, and composite interaction)
A taxonomy of interaction operations	
Ward and Yang (2004) [50]	interaction operators (navigation, selection, distortion), interaction spaces (screen-space, data value-spaces, data structure-space, attribute-space, object-space, and visualization structure-space), and interaction parameters (focus, extents, transformation, and blender)
Taxonomies of user tasks	
Zhou and Feiner (1998) [56]	Relational visual tasks (associate, background, categorize, cluster, compare, correlate, distinguish, emphasize, generalize, identify, locate, rank, reveal, switch) and direct visual organizing and encoding tasks (encode)
Amar, Eagan, and Stasko (2005) [4]	Retrieve value, filter, compute derived value, find extremum, sort, determine range, characterize distribution, find anomalies, cluster, and correlate

Fig. 3: Select Infovis Taxonomies Compiled by Yi et. al[11]

Selection enables users to track items of interest and perform additional actions including deletion, translation, and emphasis of a selected item. Exploration allows a user to discover more visual information potentially obscured by screen limitations via methods such as panning.

[11] Yi et. al, "Toward a Deeper Understanding of the Role of Interaction in Information Visualization"

Reconfiguration rearranges data to uncover a different perspective of the data; one further can compare data side by side. Encoding data encompasses “translating” the data into a different representation, e.g. topographical maps use blue to represent water, green to represent areas with vegetation, and tightly spaced lines to indicate rapid escalation in elevation. Abstraction simplifies data representation whereas elaboration reveals specific attributes of a data subset. Filtering shows data based on user specified filters, otherwise known as conditions; one can take further advantage of filtering by making slight visual modifications to “filtered out” elements, such as having elements nearby or a different shade/color to illustrate a loose relationship. Connection highlights relationships between data, showing different ways in which they are related. Having now established a common vocabulary, one can now begin a lower level discussion of the impact of interaction on visualizations.

The Role of Interaction

The arbitrariness of each type of interaction increases the efficiency of data analyzation and allows users to quickly view data selectively and manipulate it to account for personalized ways of learning. Jared Schiffman identifies an issue with current visualizations of computation: “The absence of a means to view and interact with computation directly is why programming systems are incomplete.” [12] As the world attempts to visualize more complex data, interaction underlies user directed manipulation of data, an attribute essential to revealing concealed information in visualizations of multivariate data.

Ben Fry: Visualizing Coding Sequences and Transcription

[12] Schiffman, Jared, “Aesthetics of Computation - Unveiling the Visual Machine”

Ben Fry exemplifies this in his section on coding sequences and transcription. In protein synthesis, amino acids are specified by a codon, or a set of three DNA nucleotides. Selection of the first two nucleotides more or less determines the amino acid produced and results in an event called *four-fold degenerate*, where less emphasis is placed on the third nucleotide. This stems from the nature of the transcription process, as the selection of each nucleotide impacts the remaining amino acid possibilities.

The following table represents a typical visualization of all 64 possible codons and their corresponding amino acids. As a static visualization, while it achieves its function of organizing all outcomes and makes it easy for a user to determine the order of the nucleotides via color encoding (as seen in the sequences, red is first, blue is second, and green is third), it fails to highlight four-fold degeneration successfully and even struggles somewhat with being too visually busy.

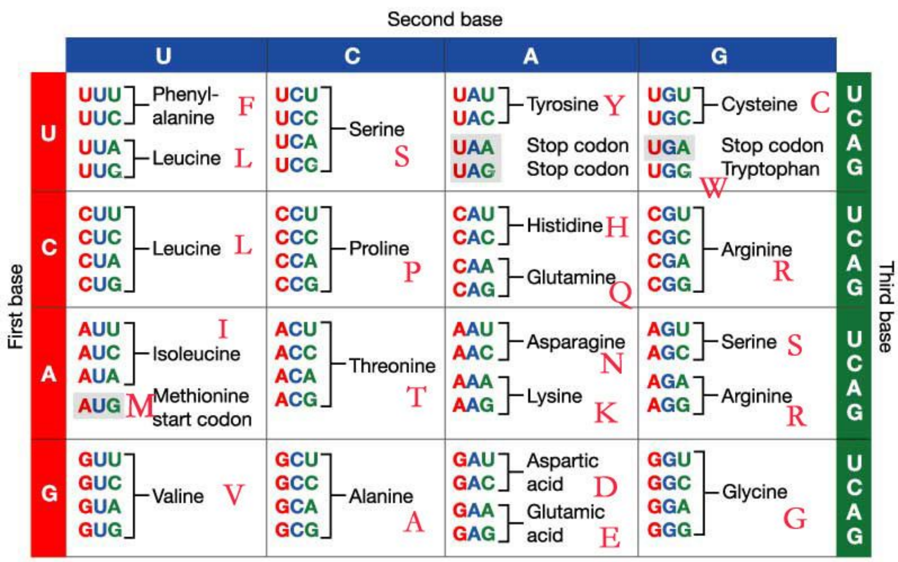


Fig. 4: Codon Table[13]

[13] Birla Institute of Scientific Research, "Codon Table"

Having identified these issues and more, Fry creates a more functional and aesthetic visualization, as seen in Fig. 5. To reduce the business of the visualization, Fry takes out repetitive amino acids and displays only the first three letters of each. He then visually showcases both two-fold and four-fold degenerates - two-fold by having the two possible amino acids divide the row and four-fold by having the sole possible amino acid take up the entire row. Where Fry's static visualization succeeds more so than the traditional display is in its display of the headers: using scale, the visualization guides the user's eye in a counterclockwise fashion and connotes the impact of each successive selection of nucleotides using decreasing size. Additionally he highlights hydrophilic amino acids by greying out the table cell and changing the text color to blue and uses a grey text color for the remainder of the amino acids, a detail lacking in the traditional visualization.

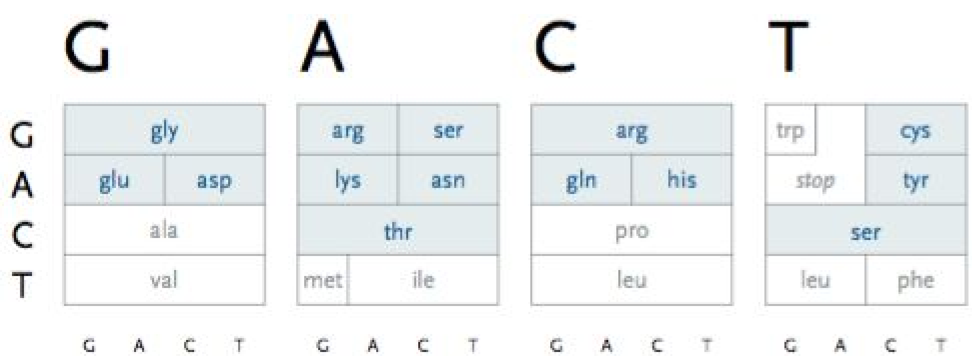


Fig. 5: Ben Fry's Static Visualization of Codons[14]

Fry then allows users to interact with the visualization via selection. As seen in Fig. 6, selection of a nucleotide results in the "stepping forward" of the amino acid. The addition of selection poses two direct benefits: the ability to filter and learning via experimentation. Looking

[14] Fry, Ben, "Computational Information Design"

at the diagram, since the selection of a nucleotide results in a step forward, a user could identify all amino acids with codons beginning with C or perhaps CA at a glance. In other words, whereas the traditional visualization requires a user to look up the nucleotide in the table and then match the corresponding possible amino acids, Fry's animation allows the user to instantly recognize amino acid possibilities based on user specified nucleotides. Moreover since the stepping occurs real time, users are able to learn about codons by experimenting with different combinations.

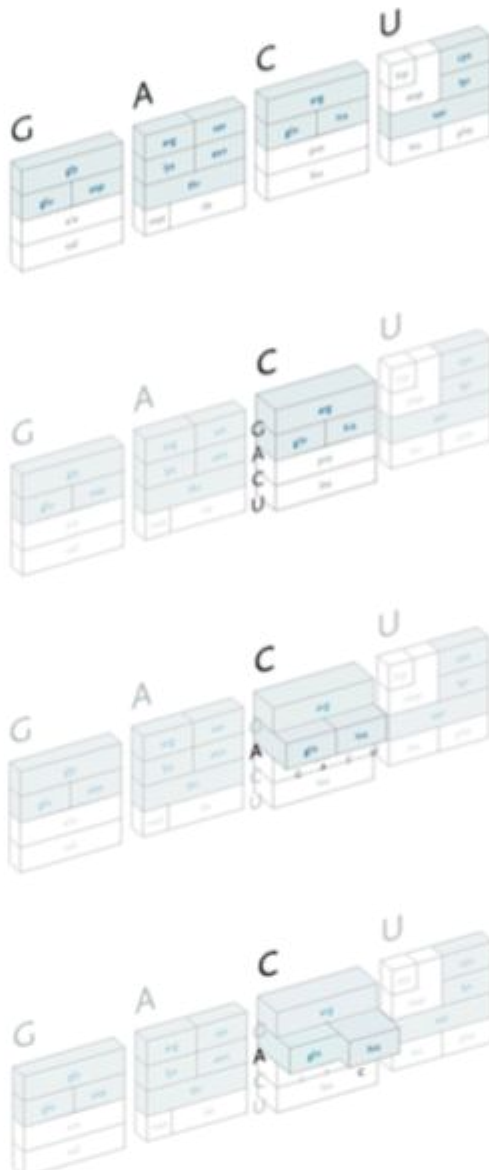

G
A
C
U

Fig. 6: Ben Fry's Dynamic Visualization of Codons[15]

Visualizing Forbes' Top 50 Sports Franchises by Longevity and Success

Column Five creates a visualization of 50 sports franchises and attempts to explore the correlation between their monetary worth, time, and championships played. Although utilizing scale to represent a quantitative attribute typically struggles with precision, Column Five successfully employs it to present an abstract representation that allows for swift visual comparison.

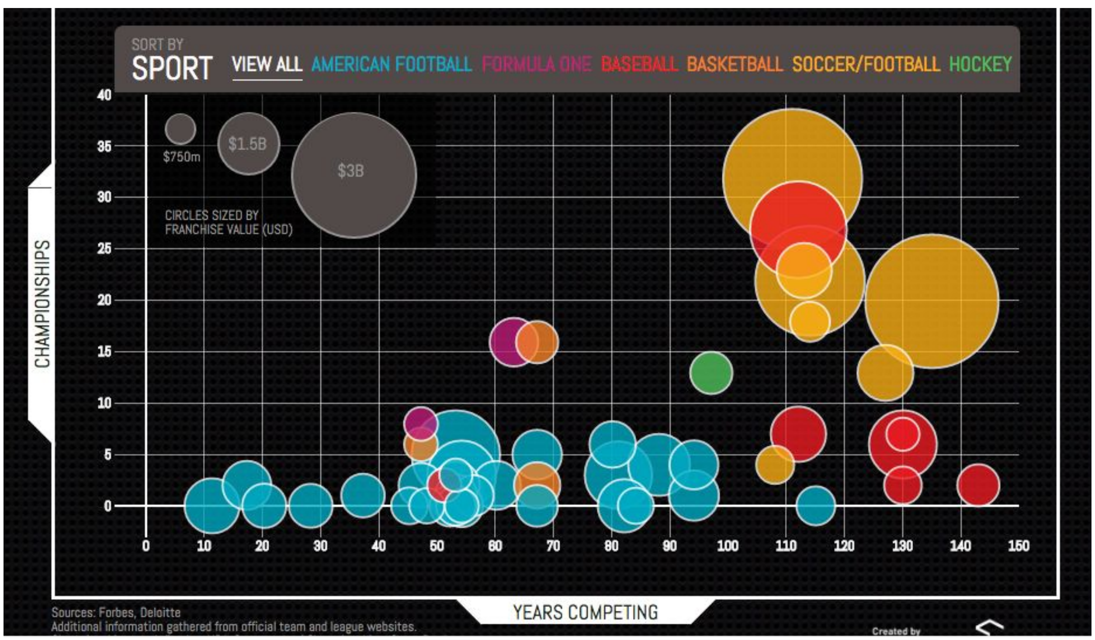


Fig. 7: Column Five's Interactive Visualization of Most Valuable Sports Franchises[16]

Column Five further strengthens the visualization through selection, elaboration, and filtering. Selection is accomplished via the "lights out" feature, where hovering over a specific

[15] Fry, Ben, "Computational Information Design"
[16] Column Five, "Interactive: Most Valuable Sports Franchises"

circle causes all other circles to dim and enables a user to visually focus on the circle of interest. In addition, hovering allows for a window to appear and elaborate on the exact franchise value, Forbes' rank, championships won, and years competing. This succession guides the user to first arrive at the conclusion that franchise value for soccer has a directly proportional relationship to the amount of championships played and years competed. From there, the user can explore specifics by hovering to reveal concealed elements or filtering the sport to determine whether the sport played has any impact on franchise value.

Compared to a static visualization, Column Five's design choices reduce visual clutter and promote exploration. Moreover a user can better form relationships across multiple variables, filter out franchise value by sport, and abstract/elaborate on data as needed.

A Visual Introduction to Machine Learning

Created by Stephanie Yee and Tony Chu, a Visual Introduction to Machine Learning is a storyline visualization that showcases how machine learning can be used to process data to determine whether a subset of houses is located in New York or San Francisco. The website limits user input to scrolling that triggers the aforementioned interaction types, with scrolling down moving the user forward in the storyline and scrolling up moving the user backward. The model first classifies home-elevation data and determines that the majority of homes in San Francisco have a higher elevation compared to New York. It then filters and visually encodes the data according to city: green denoting San Francisco and blue New York; this immediately makes identification of data by city easier as it takes on various reconfigurations. Following this, the visualization moves toward a multivariate representation, condensing the bars from the home-elevation and adding the cost per square foot. Based off of the clusters, regions in the plot are filtered/highlighted according to the color encoding of the cities: Yee and Chu represent houses ranging from 239.5 ft to 744.8 ft and having a price range of $293.0 to $4601.0 per square foot with a green rectangle and houses that had less than 239.5 ft and tended to range from $1776.0 to $4601.0 per square foot with a blue rectangle. The visualization then reconfigures the data into a scatterplot matrix of seven dimensions: elevation, year built, bathrooms, bedrooms, price, square feet, and price per square foot.

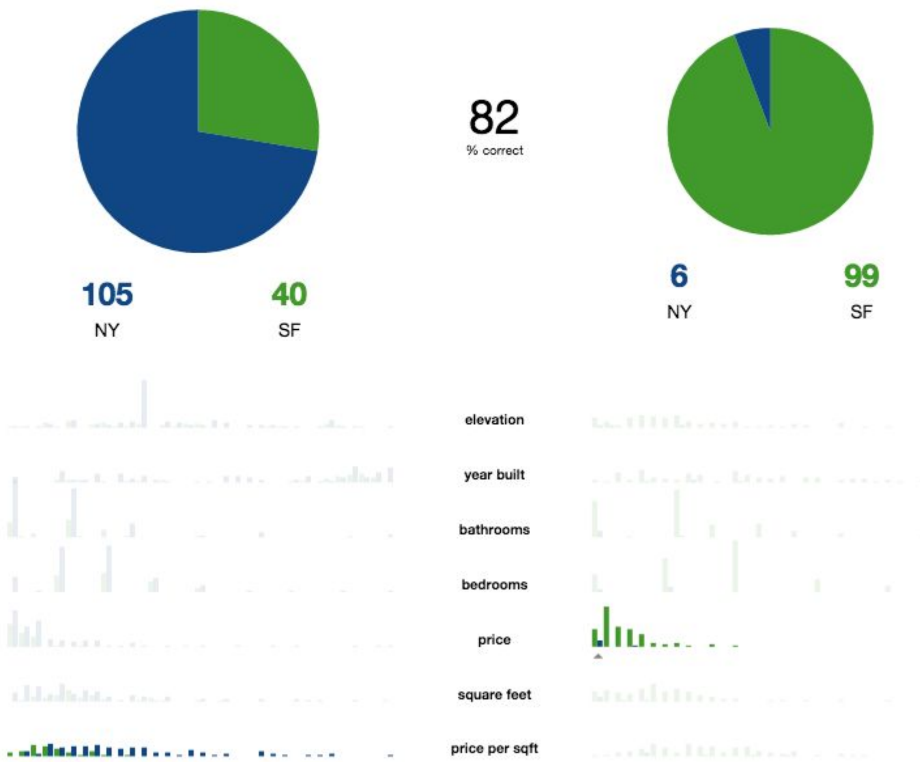


Fig. 8: Repetition of Split Point Categorization Across Seven Categories[17]

From there the visualization moves toward training the model; it revisits the elevation data and reconfigures data according to a split point, a point of comparison used to categorize

[17] R2D3, "A Visual Introduction to Machine Learning"

data. As can be seen in the above figure, multiple histograms show the splits. These are used in the next section to show the "growing" of a decision tree, where data is recursively split and assessed for accuracy. This is visually shown, with pie charts accompanying each fork. The final section shows the training of the decision tree model and stacks the test accuracy on top of the training accuracy, which displays number of guesses correct for each city as well as the percentage of test accuracy. Data points are also shown in their respective colors and are on the left side if they were guessed to be in New York and on the right if they were guessed to be in San Francisco.

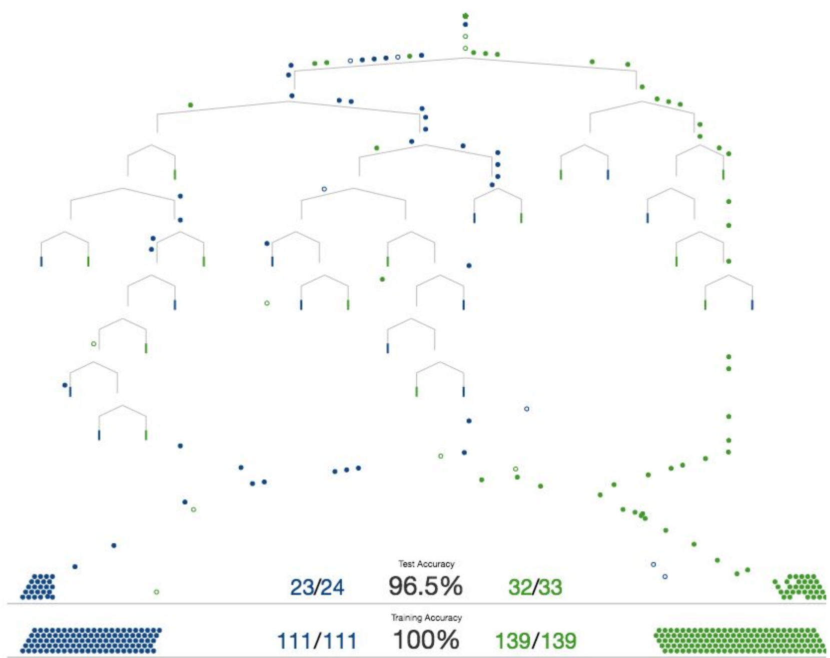


Fig. 9: Transition from Decision Tree to Training Accuracy[18]

In this case, interaction adds two direct benefits: the ability to maintain context between different stages in machine learning and user guided exploration. The transitions between each

[18] R2D3, "A Visual Introduction to Machine Learning"

step help to bridge the oftentimes complex steps of machine learning and allow the user to see exactly how comparison of different attributes impacts the representation. Going back to Fry's representation, the added interactivity enables one to iterate and thus determine whether another representation would more appropriately showcase a relationship. The user guided exploration via scrolling allows for the user to learn at a personalized pace; in other words, the user is able to slow down, speed up, go forwards, or go backwards based on whichever concepts he/she may need to review for clarification. Moreover compared to a static visualization, a user can now keep better track of where in the process he/she is at.

Spectrogram by Chrome Experiments

Created by Chrome Experiments, Music Lab offers a toolkit for understanding the way music works. In particular, Spectrogram provides user with a way to interact with varying frequencies and see in real time how higher frequency results in a higher pitch and lower frequency results in a lower pitch. Spectrogram presents several pre-recorded samples for initial understanding of frequency and further promotes exploration with the leftmost button, which allows users to create their own samples.

Whereas a static visualization would present a similar color encoding for power, the added interaction associates the auditory experience with the visualization. As music is very much an interactive experience, it makes sense that this mode of visualization would help users better learn about frequency and power.

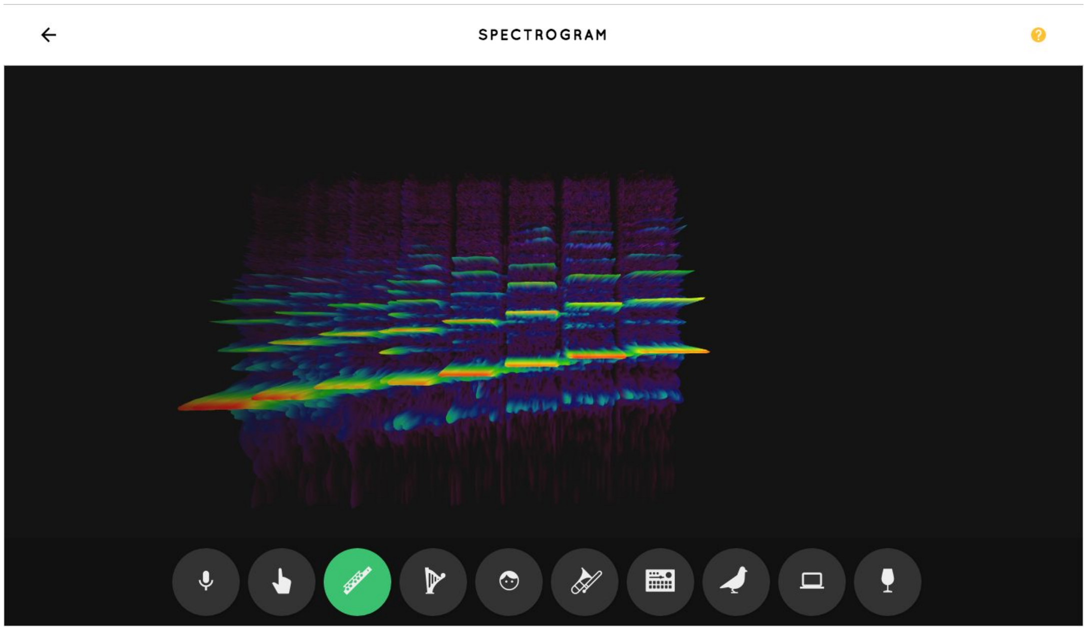


Fig. 10: Music Lab's Spectrogram Flute Sample[19]

Conclusion

With traditional methods of analyzation struggling to keep up with the overwhelming amount of data currently present, it is necessary to seek out more efficient ways to complement the human mind's analytical skills. Having identified the effect of Gestalt grouping principles on perception and established Fry and Yi's classification systems, it becomes clear that the future of data visualization and interaction are inextricably tied.

One area that could continue to use exploration is that of interaction spaces. In my research, I came across a white paper entitled, "Interaction Spaces in Data and Information Visualization" by M. Ward and J. Yang, which pointed at the possibility that efficiency of data visualization could be increased based on how it is stored. As the applications for interaction in data visualization are endless, further research into the topic could uncover even more ways of

[19] Chrome Experiments, "Music Lab: Spectrogram"

optimizing the way in which data is perceived and transformed into information. But at present, the benefits of incorporating interaction into data visualization remain clear.

Through Fry's visualization of coding sequences, the user could more quickly identify the possibilities of amino acids and better grasp the concept of a four-fold degenerate. Furthermore due to the lack of visual clutter present in previous representations, Fry could further incorporate another attribute: hydrophilicity. Through Column Five's visualization of the top 50 sports franchises, the user could isolate elements and elaborate and abstract franchise value for each company via selection. Through Yee and Chu's visualization of introductory machine learning concepts, users were led through a high level storyline of machine learning concepts in the context of housing and could manipulate the timeline via scrolling. Smooth transitions helped to showcase reconfiguration of housing data and bridge each step to the next. Through Chrome Experiments' Spectrogram, color encoded frequencies coupled with audio samples enhanced understanding of frequency and power beyond a static visualization.

Time and time again, dynamic visualizations compacted more information in a more breathable format and gave a higher degree of control to the user, whether through elaboration of concealed information or the ability to hone in on elements of interest via selection. Interaction is no longer a luxury in data science - it is a necessity.

Sources

Birla Institute of Scientific Research, "CRAT: Codon Restriction Analysis Tool." Bioinfo.bisr.res.in. Accessed August 28, 2016. http://bioinfo.bisr.res.in/project/crat/pictures/codon.jpg

Column Five, "Interactive: Most Valuable Sports Franchises." Columnfivemedia.com. Accessed August 20, 2016. http://www.columnfivemedia.com/workitems/interactivemostvaluablesportsfranchises

Donald A. Norman. "Cognition in the head and in the world." *Cognitive Science*, 17 (1993): 1-6.

Floridi, L. and Sanders J.W.. "On the Morality of Artificial Agents." *L. Introna and A. Marturano, eds., Proceedings Computer Ethics: Philosophical Enquiry – IT and the Body, Lancaster* (2001): 84–106.

Fry, Ben. "Computational information design." Massachusetts Institute of Technology. Dept of Architecture: Program in Media Arts and Sciences. *The Aesthetics + Computation Group* (2004).

Martijn J.M. Lamers and Ardi Roelofs, "Role of Gestalt grouping in selective attention: Evidence from the Stroop task." *Perception & Psychophysics*, Volume 69, Issue 8 (2007): 1305-1314.

Medler, Ben. "Play With Data - An Exploration of Play and Its Effect on Player Experience." *Parsons Journal for Information Mapping* (2012).

Ji Soo Yi, Youn ah Kang, John T. Stasko, Julie A. Jacko. “Toward a Deeper Understanding of the Role of Interaction in Information Visualization.” *IEEE Transactions on Visualization and Computer Graphics,* Volume 13, Issue 6 (2007).

Music Lab, “Spectrogram.” Chrome Experiments. Accessed August 20, 2016. https://musiclab.chromeexperiments.com/Spectrogram

Playfair, William. “Interest of the national debt from the Revolution.” *The Commercial and Political Atlas* (1786).

Schiffman, Jared. “Aesthetics of Computation - Unveiling the Visual Machine.” Massachusetts Institute of Technology. Dept of Architecture: Program in Media Arts and Sciences. *The Aesthetics + Computation Group* (2001).

SINTEF. "Big Data, for better or worse: 90% of world's data generated over last two years." *ScienceDaily*. ScienceDaily. 2013. Accessed August 20, 2016. www.sciencedaily.com/releases/2013/05/130522085217.htm

Stahl, Bernd Carsten. *Information Systems: Critical Perspectives*. (New York: Routledge 2010). 71.

Stephanie Yee and Tony Chu. “A Visual Introduction to Machine Learning.” R2D3.us. Accessed August 15, 2016. http://www.r2d3.us/visualintrotomachinelearningpart1/