Belief-Driven Data Journalism

Francis Nguyen
Northwestern University
Evanston, Illinois
francis.nguyen@northwestern.edu

Samana Shrestha
Northwestern University
Evanston, Illinois
samana.shrestha@northwestern.edu

Joe Germuska
Northwestern University
Evanston, Illinois
joegermuska@northwestern.edu

Yea-Seul Kim
University of Washington
Seattle, Washington
yeaseul1@uw.edu

Jessica Hullman
Northwestern University
Evanston, Illinois
jhullman@northwestern.edu

ABSTRACT

Journalists often use visualizations and other interactive representations to support stories they convey in articles. While readers bring their prior beliefs to interpret these representations, typical models of designing them do not consider the readers' beliefs. We propose "Belief-driven data journalism" as a framework for integrating readers' beliefs in designing and supporting interaction with data-driven articles. We present four case studies to illustrate how belief-driven data journalism can serve journalistic goals and reflect on design considerations. We describe an authoring tool that we are developing to help journalists and others with varying technical expertise create belief-driven data journalism pieces.

CCS CONCEPTS

• **Human-centered computing** → *Human computer interaction (HCI)*; *Collaborative content creation*; *Visualization application domains*;

KEYWORDS

Belief-Driven data journalism, Visualization, Data representation

1 INTRODUCTION

Journalistic norms emphasize goals of communicating accurate information, conveying multiple perspectives, and engaging readers with the news. Visualizations and other interactive representations can support these goals in the context of data-driven reporting. In consuming these representations, it's natural for readers to bring their beliefs to bear on the new information. For example, imagine an article that visualizes the distribution of gun related crimes in the United States with regards to state and demographic information. In this article, a reader can query their state to see how various factors might impact gun related crimes. Readers might draw on their political beliefs to interpret the visualizations and results: a left-leaning reader might utilize the visualization and data to correlate the volume of guns to gun violence and advocate for gun-control, where a right-leaning reader might use the visualization to link gun violence to mental illness and advocate for better mental health care. Hence, the two readers might draw different conclusions from the same set of data.

Typical approaches to designing data representations like visualizations don't consider readers' prior beliefs. We ask, how might incorporating readers' beliefs change interactions with data journalism? Previous research shows that eliciting and presenting people's beliefs about data can have benefits for individual readers. Kim et al. [10] find that prompting users to express their prior beliefs about the data before seeing them, and then showing the data alongside their beliefs, helps them recall the data later. Explicitly showing the gap between prior predictions and data may evoke users' curiosity to engage with the data to fill the gap. Visualizing others' beliefs can socially engage readers to think more critically about the relationship between beliefs and data. When others' beliefs show a consistent trend, visualizing them can improve a reader's ability to remember the data and inform how they update their beliefs [11].

We propose "Belief-driven data journalism" as a framework in which readers' beliefs are elicited and depicted along with data in an article. We outline design goals for informative and engaging interactive experiences that belief-driven data journalism can support, and connect these to existing empirical research. We present four case studies to illustrate the design space possibilities. We describe an authoring tool that we are developing to make it easier for journalists to create belief-driven data journalism pieces.

2 BELIEF-DRIVEN DATA REPRESENTATIONS

We identify four goals of belief-driven data representations that help journalists develop engaging data journalism pieces.

Engage users by surprising them: Journalists can use belief-driven data representations when the data depicts unusual trends. Curiosity is piqued when one perceives a gap between what they know and what they want to know [14]. Journalists can stimulate users' curiosity by asking readers' to sketch predictions about a phenomena (e.g., job growth, election outcomes, climate change, etc.) in an interactive visualization emphasizing the gap between what readers believe and observed data. The user might consequently spend more time with the article to identify explanations for the discrepancy [1].

Provide an active learning platform: Journalists can use belief-driven data representations when data contains new facts that their readers may not know. Psychology research shows that a learner who generates an externalization (e.g., drawing) while consuming information have better performance than those who don't externalize on learning [17]. Belief-driven data representations give readers a chance to learn about facts by prompting them to draw their expectations for a phenomena in an interactive visualization, and generates self-explanations about the errors they can make.

Improve awareness of diverse beliefs: Journalists can use belief-driven data representations when they expect readers to hold a diverse range of beliefs about a topic, such as political events or gender expectations. A journalist asks the user to express their own beliefs either in text or by sketching a prediction in an interactive visualization. After, the journalist can unveil other readers' opinions

Computation + Journalism, Feb 2019, Miami, FL, USA
2018. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

and answers, exposing the user to a diverse set of viewpoints. Accordingly, users can become better informed of a variety of perspectives on the topic. Such exposure improves learning and decision making by corroborating or challenging their beliefs [8].

Understand beliefs at scale: There are limitations to how well news organizations currently understand their online readers; typically, only basic demographic information or interaction logs are recorded. Deepened knowledge of readers allows journalists to design better presentations that improve engagement and connection to articles. Authors can use belief-driven journalism to understand how readers interact with an article and what prior beliefs they have to develop a more holistic understanding of their reader. For example, journalists can frame the data-driven article as a survey platform that solicits and exposes readers' beliefs on topics like political events, gender expectations or climate change. The generated belief collections can be used by news organizations as a more semantically rich source of information relative to reader demographic profiles. This data can support designing for maximum surprise or belief change on future data pieces in the same domain. Furthermore, if journalists want to model readers' beliefs, they can collect beliefs in addition to how uncertain a reader is about the topic for use in a Bayesian cognitive model [12]. This approach provides a formalized method for journalists to deepen understanding of how readers' final beliefs are formulated based on their prior beliefs.

3 CASE STUDIES

We present four case studies that vary in how information is visually represented and the goals of using a belief-driven framework.

3.1 Voter Demographic Trends

In the U.S. Presidential election of 2008, more than than 90% of black voters and more than 65% of Hispanic voters voted for Barack Obama. A journalist wants to do a piece reflecting on the voting behaviors of different ethnicities. The journalist obtains data describing the percentage of voters who voted for the Republican candidate in the 2008 presidential election. The dataset includes the voter's ethnicity (Hispanic, White, Black) and income level (under $75K, over $75K) (see [7] for a related NYT presentation). The visualizations show the percentage of voters who voted for the candidate by ethnicity and by income for all states (Fig. 1(a)). Ethnicity is encoded by color (red, green, blue), and the percentage value is encoded by the height of dots.

The journalist wants to prompt readers to consider how income and ethnicity together influence which party a voter supports by asking them to predict the trend in support for the Republican candidate among the different ethnicities across both income levels. However, asking the reader to sketch or provide their prediction about every state is overwhelming. The journalist decides to use personalization by detecting the reader's IP address, prompting them for a prediction about only their state. To elicit their beliefs, the journalist can provide an empty canvas, and walk the reader through drawing each prediction one at a time using a mouse.

In a different scenario, the journalist might want to draw attention to how the influence of income is different for one of the ethnicities. For example, imagine that readers have a mental model that states people from a higher income bracket are more likely to vote for Republican candidates than people from a lower income bracket. Expecting this, the journalist might choose to elicit predictions only for data that depicts the opposite trend for an ethnicity, like the Hispanic group in Figure 1(b), to ensure that readers notice this information (Fig. 1(c)).

Once a reader submits their beliefs, the visualization can show actual data for each predicted ethnicity alongside the reader's prediction. They may choose to emphasize the gap between the two, such as by annotating with reasons why the Hispanic group might deviate from the expected pattern. If most other readers also had false beliefs for this trend, the journalist can present the set of collected beliefs to emphasize the pervasiveness of the expectation or simply to allow the reader to reflect on how their predictions compared to others. The lines drawn by other readers can either be depicted raw as individual lines to convey variance among the other readers when fewer samples are collected (Fig. 1(d) left) or aggregated to convey the representative value predicted by all other readers (Fig. 1(d) right). Finally, consider the case where the journalist elicits readers' predictions only for their state, detected using IP address. Here, the journalist may choose to present social information only for the reader's state, or to contrast the degree of consensus in the reader's own state with that in other states, for example by exposing those states where beliefs are most divergent.

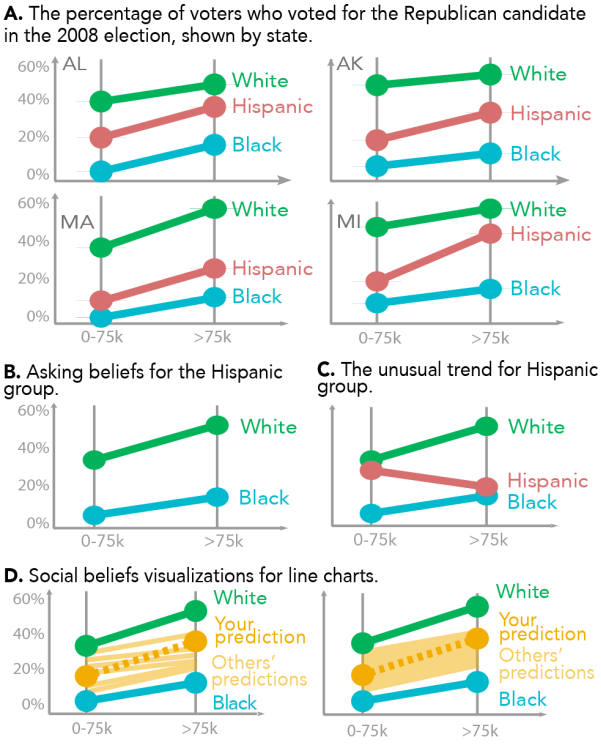

Figure 1: Plots showing various stages of engagement in belief-driven data representations. All plots show income by percentage of voters who voted for the candidate by ethnicity.

3.2 Health Conditions among the Elderly

Imagine that a journalist is featuring a story about how many elderly people in assisted living centers suffer from chronic diseases, especially Alzheimer's and other forms of dementia.

The journalist wants to emphasize the proportion of residents in the center who have the condition (42% in this case) to enhance

awareness of the disease. The journalist has to decide which visual representation to use. Should the journalist use a grid visualization (Fig. 2) as used in the New York Times [4]? If so, how many icons should they use to represent the set of respondents? Or, should they use a more conventional visualization like a pie chart?

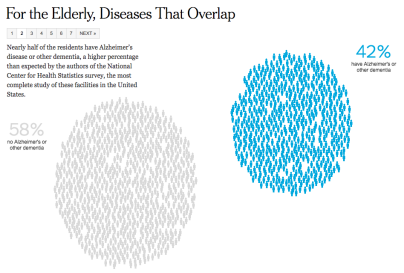

Figure 2: The visualization shows the proportion of assisted living residents who have Alzheimer's disease or dementia.

A standard approach to evaluating these design ideas might involve the journalist showing the visualization to a few colleagues and asking them about the proportion they see, or for their opinions on which visualization seems most impactful. A belief-driven paradigm however, when combined with a notion of Bayesian inference, can provide a rational standard for testing this question. In particular, Kim et al. [12] demonstrate that a Bayesian inference framework can explain how to derive posterior beliefs (i.e., the reader's updated beliefs after interacting with the visualization) given a set of prior beliefs and the observed data. These posterior beliefs represent what the reader should believe if they updated their beliefs based on the new data in a statistically optimal way (e.g., using Bayes rule). By eliciting a sample of readers' prior beliefs and beliefs after using each of the three visualization designs, the journalist can answer the question, which visualization results in the most rational inferences? Assuming the goal of the visualization is to help readers understand the data itself as well as the statistical implications (e.g., the relative amount of sampling error) of the data accurately, a good visualization will produce posterior beliefs that are closer to the (normative) posterior beliefs calculated using Bayesian statistics. A good visualization in this scenario is likely to the one that best conveys to readers the reliability of the statistic about dementia, which is based on a very large sample (n=750,000).

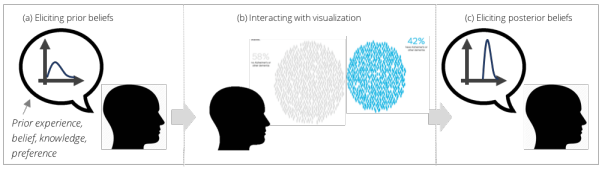

Figure 3: The Bayesian framework considers one's prior beliefs into the process of formulating the posterior beliefs.

3.3 Who will win?

Journalists often utilize choropleth maps to convey the geographical context of events, or present differences and similarities across geographical regions. For example, FiveThirtyEight used a choropleth to show the predicted chances of winning for Clinton and Trump in the 2016 U.S presidential election by state (Fig. 4). The predicted value is encoded in color using a diverging color scale.

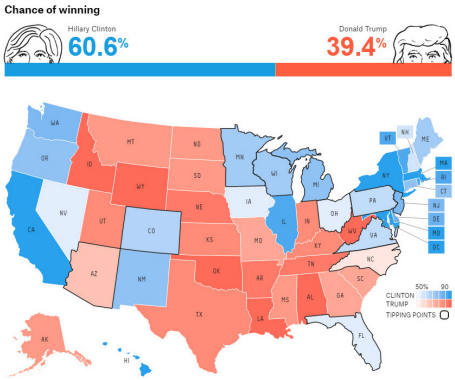

Figure 4: Choropleth showing winning chances for Clinton and Trump in the 2016 U.S presidential election.

While this visualization was useful in allowing people to understand each candidate's chances of winning, the accuracy of the poll predictions that were used to produce these model predictions were later questioned based upon the surprising nature of the 2016 election outcome for many pollsters and Americans [6]. One reason cited for why the poll results, and hence the model predictions, may not have been very accurate is that people were less likely than in other elections to honestly answer questions about how they intended to vote. We speculate: Could using belief-driven data journalism to elicit peoples' predictions about how *others* would vote throughout the election cycle have reduced some of the surprise of this outcome for readers?

Collective intelligence applications aim to improve decision making by aggregating many individuals' intellectual knowledge [13]. Arrow et al. claim that a prediction made by a crowd can be more accurate than a model's prediction even when an individual has a slightly inaccurate intuition [2]. Imagine getting a contract that gives you $1 if candidate A wins the election in 2020 presidential election, then selling the contract in a market open to the public. The final price of this contract (e.g., $0.23) reflects the crowd's beliefs about this election, and (under proper incentives) represents an accurate predicted probability that candidate A actually wins the election (e.g., 0.23/1.23*100%).

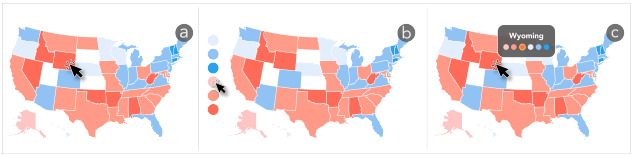

Figure 5: Different ways of eliciting beliefs. (a) Setting the value by clicking on the region repeatedly until the desired value of color is shown, (b) using a legend to select the desired value, and then brush the target region, or (c) using a tool-tip.

Belief-driven data journalism can adapt this idea of utilizing readers' aggregated beliefs to predict future events, such as elections. Imagine Figure 4 as a belief-driven visualization that prompts a user for their prediction of who will win in a state. Different elicitation interactions might be used depending on how many regions the journalist wants to elicit (Fig. 5). After eliciting readers' beliefs, journalists can show the aggregated responses. For example a dot density display could use colored dots to represent the number of readers that provided a value associated with the color (Fig. 6).

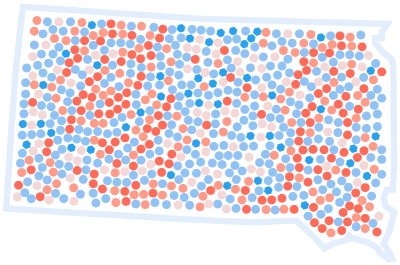

Figure 6: Visualizations for social beliefs.

3.4 Did Trump Pay Hush Money or Not?

People don't believe all the stories presented in the media. Belief-driven data journalism can encourage readers to reflect their beliefs about events in the news, including why exactly they trust or don't trust a certain perspective on a story.

Consider how Michael Cohen, Donald Trump's former lawyer, pleaded guilty to violating campaign finance laws to pay alleged former lovers of Trump to keep quiet, as directed by Trump [3]. Trump denied that he directed Cohen to do so [5]. The probability that a reader ultimately believes to whether Trump did in fact direct Cohen to pay the money can be modeled as a function of the reader's prior beliefs regarding each individuals' intentions and actions. A reader might suspect that Trump directed Cohen to pay the money. Belief-driven journalism can help them understand what other beliefs feed that suspicion.

For example, readers might have a prior about Trump's belief state and how it plays into the events. If the reader were to assume that Trump did tell Cohen to pay the money, what do they think is the probability that Trump believes that he didn't? If they assume that Trump didn't tell Cohen to pay the money, what do they think is the probability that Trump believes he did tell him to? Similarly the reader can consider their prior beliefs over Cohen's belief state. In addition, other beliefs about character can play a role, such as Trump's and Cohen's likelihood to lie in different situations.

By defining and eliciting relevant prior beliefs for a controversial news event, a journalist can help a reader understand how their beliefs about the people and situation predict their belief state regarding whether the event happened. For example, after prompting all statements above about the story, a system can quantify the probability of the reader's beliefs about the topic: "According to your prior beliefs, you believe there is an 87% chance that Trump did in fact tell Cohen to pay the money."

The journalist can then use the article to inform readers of the beliefs that other readers' bring to the article. For example, visualizations could show the probability with which other readers believed that Trump would lie about what he did if he had in fact told Cohen to pay the money, as well as the variation across readers' predictions of this probability. Reporting on beliefs as well as events themselves provides an intriguing social lens on the story. A recent interactive related to the Cavanaugh trial demonstrates some of these possibilities [16].

4 AUTHORING TOOL DEVELOPMENT

Many data journalists may lack sufficient web programming skills to author interactive belief-driven data journalism pieces. To aid journalists in realizing belief-driven data journalism, we present an authoring tool that we are developing. Our tool currently supports line charts. Initially supporting line charts benefits journalists by allowing them to elicit individual data points as well as trends in data. Additionally, the interactions involved in the elicitation (e.g., drag to draw) and visual representations of social information (e.g., showing aggregated lines, raw lines) are relatively intuitive. We report on the intended usage scenario and design as well as our development process.

4.1 Configuration & Authoring Workflow

We draw on our case studies to define the minimum requirements that allow journalists to explore the possibilities of belief-driven data journalism. By considering decision points during authoring, we designed a declarative configuration for belief-driven data journalism pieces consisting of line charts. The configuration records an author's choices regarding the data they choose to plot, data the user has to predict, and how to plot others' predictions. Specifically, our prototype tool requires specification of the following configuration parameters: the *dataset*, *encodings*, *elicitation*, and *social information*.

The journalist first loads a *dataset* through a URL entry field. The URL represents a Google Sheet where the journalist stores their dataset. The tool infers columns' data types from the dataset (e.g., categorical, quantitative). After validating the dataset, the journalist is prompted to select *encodings*. The *encodings* map columns in the dataset to visual attributes of the data representation, most important of which are encoding variables to plot on the x and y axes. Because a dataset may contain many quantitative variables, the tool relies on the journalist's domain knowledge to select meaningful axis encodings. Furthermore, the journalist can specify whether to encode additional data in the *line* configuration, marking the *multiple* parameter as true. When *multiple* is specified, an additional encoding is required: *color*, which encodes column names in the dataset to colors. The tool then produces a preview of the visualization based on given parameters to provide visual feedback to the journalist, enabling iteration until a design is complete.

The next step requires the journalist to make their choice of *elicitation*. The journalist is prompted to select a *reference*, a single column or query representing a subset that they wish to elicit (e.g. "Germany"). After selection, the tool requires specification of a *range* of values to elicit so that journalists can engage readers' in multiple scenarios, such as asking readers to make predictions for a missing section of data given the rest, or to predict future trends given prior data. After the information to elicit is specified, the journalist can curate *hints* based on their needs. *Hints* are provided to the reader while they are creating a prediction. *Hints* can be graphical (e.g., a point the prediction contain) or textual (e.g., "The line should spike near the beginning"). The journalist previews and selects specific points used as hints that readers' predictions should include or writes a brief helpful message.

The last set of considerations an author makes in the workflow is choosing options for presenting *social* information. Currently, *social data* is managed using a separate Google Sheet, however we acknowledge the limitations of using sheets as a system to store data. A challenge in future iterations will be to develop a robust yet flexible system to manage social information. The next consideration a journalist must make is to specify how to view their social data (*social visualization type*). The journalist can choose the *minimum* threshold of data points before presenting social data to the user; a small number of social data points won't be representative and may lead to negative social influence [9]. In the case where not much data

has been collected, the tool defaults to displaying raw data from readers. When many predictions are plotted, however, lines can quickly become too dense, rendering any useful visual feedback from the social information unintelligible. In this case, the tool aggregates the data as a heatmap by default. Once the display of social information has been chosen, the journalist can choose options to "export" the visualization as an easily embeddable *iFrame*.

5 FUTURE WORK

Evaluating Prototype: We plan to evaluate our prototype by using journalists' feedback on our design choices and selection of implemented features for the tool. By conducting usability tests with journalists, we hope to gather various perspectives that will inform the development of future features.

Researching Impact of Social Cues: Psychological research indicates that presenting other people's beliefs can affect an individual's beliefs in many ways. In the context of designing a belief-driven representation, design decisions must be made about whether to separate beliefs from data views, leading to questions about how to design a visual vocabulary that can be reused so that readers can rely on consistent presentation norms for beliefs versus data [15], whether to show or remove outliers, or when to show all beliefs versus only those that don't align with the reader's. We hope to explore the many possibilities for belief curation and design in collaboration with journalists.

Exploring input modalities to elicit beliefs: Natural language interactions are increasingly becoming a topic of interest as more people use news on phone or on conversational agents (e.g., Siri, Alexa). Belief elicitation can be expanded to accommodate more intuitive options for input, such as voice input. Future work could explore elicitation of beliefs about data or events via natural language.

REFERENCES

[1] Gregor Aisch, Amanda Cox, and Kevin Quealy. 2015. You Draw It: How Family Income Predicts Children's College Chances. *The New York Times*, May 28, 2015, *http://nyti.ms/1ezbuWY,*. (2015).

[2] Kenneth J Arrow, Robert Forsythe, Michael Gorham, Robert Hahn, Robin Hanson, John O Ledyard, Saul Levmore, Robert Litan, Paul Milgrom, Forrest D Nelson, et al. 2008. The promise of prediction markets. *Science-new york then washington-* 320, 5878 (2008), 877.

[3] BBC. 2018. Michael Cohen trial: Trump accused of directing hush money. *BBC*, Aug 22, 2018, *https://www.bbc.com/news/world-us-canada-45265546,*. (2018).

[4] Matthew Bloch and Hannah Fairfield. 2013. For the Elderly, Diseases That Overlap. *The New York Times*, Apr 15, 2013, https://archive.nytimes.com/www.nytimes.com/interactive/2013/04/16/science/disease-overlap-in-elderly.html,. (2013).

[5] JORDAN FABIAN. 2018. Trump denies having prior knowledge of Cohen hush-money payments. *The Hill*, Aug 22, 2018, *https://thehill.com/homenews/administration/403059-trump-denies-knowledge-of-cohen-hush-money-payments,*. (2018).

[6] Andrew Gelman. 2016. A 2% swing: The poll-based forecast did fine (on average) in blue states; they blew it in the red states. *Statistical Modeling, Causal Inference, and Social Science*, Nov 9, 2016, https://andrewgelman.com/2016/11/09/polls-just-fine-blue-states-blew-red-states/,. (2016).

[7] Andrew Gelman and Avi Feller. 2012. Red Versus Blue in a New Light. *Statistical Modeling, Causal Inference, and Social Science*, Nov 12, 2012, https://campaignstops.blogs.nytimes.com/2012/11/12/red-versus-blue-in-a-new-light/,. (2012).

[8] William Hart, Dolores Albarracín, Alice H Eagly, Inge Brechan, Matthew J Lindberg, and Lisa Merrill. 2009. Feeling validated versus being correct: a meta-analysis of selective exposure to information. *Psychological bulletin* 135, 4 (2009), 555.

[9] Jessica Hullman, Eytan Adar, and Priti Shah. 2011. The impact of social information on visual judgments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1461–1470.

[10] Yea-Seul Kim, Katharina Reinecke, and Jessica Hullman. 2017. Explaining the Gap: Visualizing One's Predictions Improves Recall and Comprehension of Data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM.

[11] Yea-Seul Kim, Katharina Reinecke, and Jessica Hullman. 2018. Data Through Others' Eyes: The Impact of Visualizing Others' Expectations on Visualization Interpretation. *IEEE transactions on visualization and computer graphics* 24, 1 (2018), 760–769.

[12] Yea-Seul Kim, Logan Walls, Pete Krafft, and Jessica Hullman. 2019. Bayesian Models of Everyday Data Interpretation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM.

[13] Pierre Lévy and Robert Bononno. 1997. *Collective intelligence: Mankind's emerging world in cyberspace*. Perseus books.

[14] George Loewenstein. 1994. The psychology of curiosity: A review and reinterpretation. *Psychological bulletin* 116, 1 (1994), 75.

[15] Zening Qu and Jessica Hullman. 2018. Keeping multiple views consistent: Constraints, validations, and exceptions in visualization authoring. *IEEE transactions on visualization and computer graphics* 24, 1 (2018), 468–477.

[16] Eric Saund. 2018. Did Kavanaugh Do It? *Medium Math*, Oct 2, 2018, https://medium.com/@saund/did-kavenaugh-do-it-9fb3e08bb2a3,. (2018).

[17] Steffen P Schmidgall, Alexander Eitel, and Katharina Scheiter. 2018. Why do learners who draw perform well? Investigating the role of visualization, generation and externalization in learner-generated drawing. *Learning and Instruction* (2018).