IQT QUARTERLY

VOL. 7 NO. 4 SPRING 2016

DATA by Design

IQT
IN-Q-TEL

IQT Quarterly is a publication of In-Q-Tel, Inc., the strategic investment organization that serves as a bridge between the U.S. Intelligence Community and venture-backed startup firms on the leading edge of technological innovation. *IQT Quarterly* advances the situational awareness component of the IQT mission, serving as a platform to debut, discuss, and debate issues of innovation in the areas of overlap between commercial potential and U.S. Intelligence Community needs. For comments or questions regarding IQT or this document, please visit www.iqt.org, write to iqtquarterly@iqt.org, or call 703-248-3000. The views expressed are those of the authors in their personal capacities and do not necessarily reflect the opinion of IQT, their employers, or the Government.

EDITORIAL
IQT Quarterly, published by In-Q-Tel, Inc.
Editor-in-Chief: Adam Dove
Theme Editor: Andrea Brennen
Contributing Editors: Carrie Sessine, Brittany Carambio, and Melissa Hayes

Design by Lomangino Studio LLC
Printed in the United States of America

IQT
QUARTERLY

Identify. Adapt. Deliver.

TABLE OF CONTENTS

On Our Radar: Data is a Design Problem

By Andrea Brennen

It's time to start thinking about data as a design problem

Today, our capacity to collect, process, store, and query data far exceeds our ability to understand the meaning of that data. Over the past few decades, there has been tremendous evolution in "big data" tools and technologies, but we are still not quite getting the insight we want from all of this data we now have.

Data becomes useful when it provides information that is meaningful to someone for some purpose. However, meaning doesn't come from tools and technology alone; it requires human interpretation. Often, this begins with *visualization* — translating data into a visual representation so that a human can make sense of it.

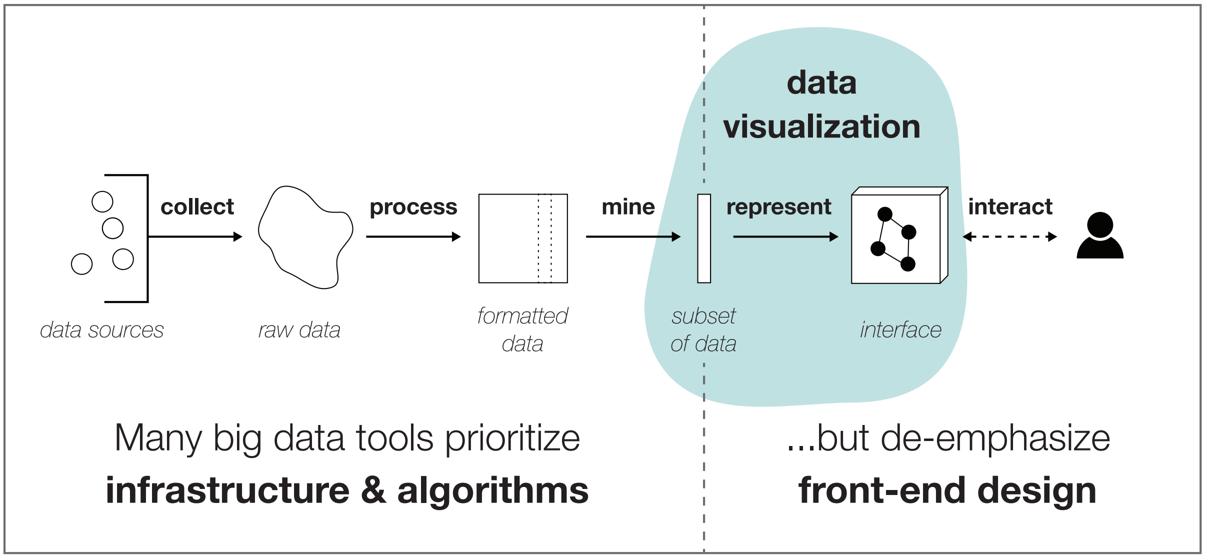
At least as far back as the 1940s, researchers extolled the benefits of visualizing data in support of analysis efforts. For example, neurologists argued that "better differentiation" between quantitative metrics did not come from a mechanical interpretation of data, but rather, from the interaction between the representation of data and the trained eye of the interpreter or analyst.[1]

When it comes to data, human interpretation is facilitated and mediated by an interface that is designed and built by someone. It is time to turn our attention to solving the front-end design challenges created by big data infrastructure, to ask how we can represent data in the most effective way and how we can build intuitive, usable interfaces that help people find the insight they seek. In other words, it's time to start thinking about data as a design problem.

Incidentally, the need for design isn't limited to visualization: it is part of a larger, consumer-oriented focus that's shifting the entire technology sector. As technologies become more ubiquitous, companies have to consider a wider audience for their products if they want those products to succeed. And appealing to a wider audience means providing a compelling user experience. As many of us spend more and more time with digital tools, the frustration associated with a bad user experience is magnified. And as we spend more time with better user interfaces in our personal technologies, we have less tolerance for poor design in our professional lives.

What is data visualization?

Within this issue of the *IQT Quarterly,* we separate the discussion of data *visualization* from related processes of collecting, cleaning, storing, and querying data. This should not be read as a dismissal of the importance of these processes, but rather, an attempt to highlight the challenges of visualization, a single — but often overlooked — step in a very complicated workflow.



One obstacle to progress in this area is that many people lack a common vocabulary to talk about data, visualization, and design. This article introduces some useful terminology and presents a framework for thinking about data visualization technologies, in order to facilitate a conversation about current capabilities, challenges, and future solutions.

Visualization can take many forms — bar chart or scatter plot, static illustration or interactive web graphic, animation or 3D model — and visualizations can be produced manually, through software, or generated automatically. The framework offered here focuses on the subset of visualization tools used to support data analysis; in other words, software tools that meet the following three criteria: 1) the input to those tools is abstract and non-visual, e.g., not image processing, video processing, or naturalistic 3D modeling tools; 2) the output is an image, e.g., not display hardware or virtual reality (VR) platforms; and 3) the result is readable and recognizable to a human. Data-related tasks that do not require a "human-in-the-loop" will probably not require visualization.

While data visualization needs vary widely, there are some general trends and common themes. The following discussion of challenges and trends is based on feedback collected from over 100 individuals across five Intelligence Community agencies, as well as commonalities noticed while evaluating 105 visualization companies and open source projects. I conducted these efforts with T.J. Rylander (Managing Partner, Investments) in support of IQT's 2015 Data Visualization Market Survey.

Exploration vs. Explanation

At a high level, it is helpful to think about two main uses of data visualization: often, we either want to use visualization to **explore** something about a dataset that we don't already know, or to **explain** something about that dataset to someone else. Regardless of whether we want exploratory or explanatory visualization, the goal is typically threefold:

1. Represent a large quantity of data coherently;
2. Do this in a way that emphasizes or reveals relationships in the data; and
3. Help someone digest a lot of information quickly.

Different uses and applications of visualization require different tools. Exploration and explanation are very different tasks, often performed by different people within an organization. In order to be effective, visualizations should be designed (or chosen) to support a specific audience and purpose.

	Exploration	Monitoring	Explanation
Purpose	Analyze data, identify patterns	Show changes over time	Communicate an idea, "storytelling"
Work Product	Analysis platform (forensic), plots & charts	Dashboards (time-based/ streaming), templates	Graphics / "infographics," (reports & presentations), interactive web graphics
User/Audience	**Expert user** (subject matter experts & data scientists)	**Managers** or someone responsible for taking action	**Non-expert audience** (decision-makers or collaborators)

MORE DATA
less known about it

LESS DATA
more known about it

Exploratory visualizations help an expert user — a data scientist or subject matter expert — analyze data. The expert user often has a large quantity of data, doesn't know very much about that data in advance, and needs an analysis platform that can provide a range of different "views" into various aspects of that dataset. Exploratory visualizations need to be extensible and fairly generic, leveraging chart and plot types common to statistical analysis (line graph, bar chart, scatter plot, etc.). Platforms offering access to these visualizations must be responsive, fast, and agile enough to accommodate varying analytic workflows.

Explanatory visualizations are visual aids that help communicate or illustrate information to an audience that often includes non-experts. Essentially, explanatory visualizations are used to convey or reinforce a particular story about the data. The process of making an explanatory visualization usually begins with less data, but more is known — or more assumptions are made — about that data in advance. As such, the visualization can be designed in a way that highlights or conveys a specific message. Sometimes, experts assume that the same charts and plots they use to do their analysis can be presented, unchanged, to a general audience. However, these experts bring with them a wealth of contextual information — about the data as well as the analytical methods used to decode it — to their reading of the visualization. Since a general audience usually lacks this background knowledge, explanatory visualizations typically require additional context that helps convey how a particular result was obtained and why it is meaningful.

System Monitoring is a third common use case that combines aspects of exploratory and explanatory visualization. Tools used to monitor the status of a system are often configured as "dashboards" or templates, providing a number of different visualization “widgets” or views into data. A dashboard will often be configured to ingest streaming data and the specific visualizations will highlight how various states of the system are changing over time. Two common examples are dashboards that administrators use to monitor the health and status of a complex system (such as an enterprise IT network) and “Business Intelligence” dashboards that managers use to understand the “health and status” of their team, department, or company.

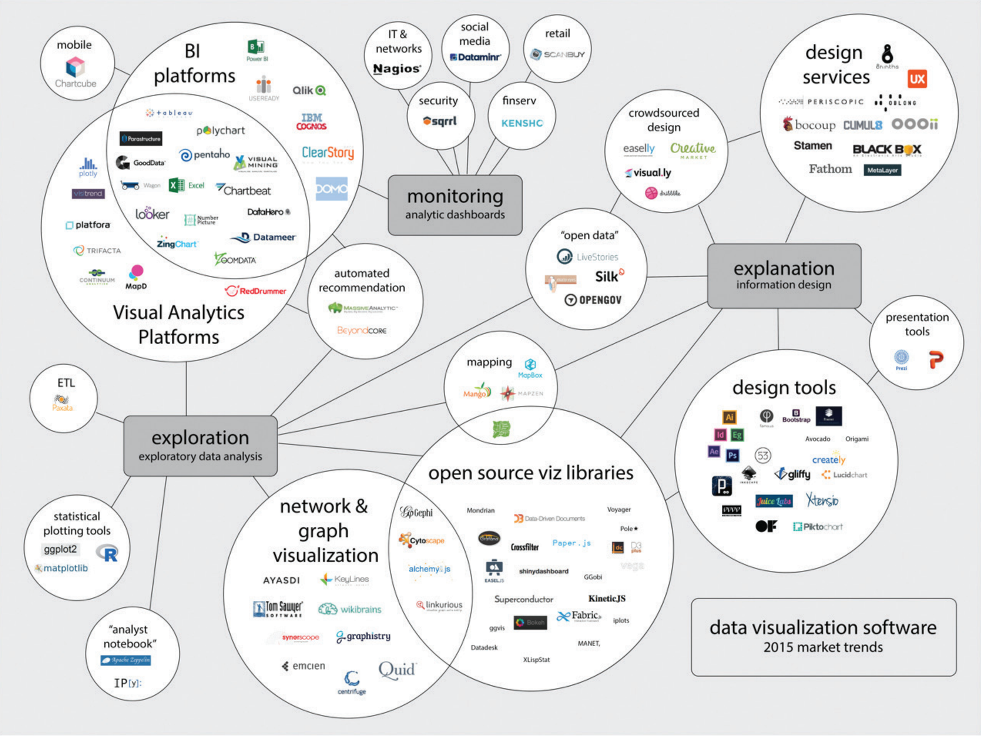
Technologies and Trends

Today, in large enterprises, the most commonly used visualization tools are Microsoft Excel and “Business Intelligence” products like Tableau and Domo, although “real-time dashboard” products are becoming increasingly popular. The majority of visualization-focused technology companies are building tools for *exploratory* visualization use cases. Most of the companies supporting explanatory visualization are either creative firms offering design services, or companies like Adobe that are providing tools that designers (often employed by design services companies) use to do their work.

Several recent technology trends among data visualization tools echo preferences and inclinations in software development, more generally. For example, there is increased focus on cloud deployments, browser-based tools, and incorporating innovation from open source libraries. Additionally, there has been a recent trend of companies leveraging hardware acceleration — both server and client-side graphics processing units (or “GPUs”) — to enhance the performance and scalability of analytic and visualization capabilities. Newer trends in the field include an increased focus on the importance of “storytelling” in data science[2], a discussion about the need for “visualization recommendation” and “automated reporting,” and commonly-expressed desires for data analysis tools equipped with an “easy button.” Additionally, in the spirit of the Uber-for-everything sharing economy, several companies are experimenting with ways to leverage crowdsourcing to scale access to design services.

Most current visualization tools can be broken down into two categories: integrated platforms like Excel and “roll-your-own” web-based components, including the popular JavaScript library, D3.js. Integrated platforms do not require users to write software code and are intended to be used by analysts or managers. These products prioritize ease of use, speed, and reusability of visualizations; however, they achieve these benefits at the expense of visualization specificity and customization. Web-based libraries, in contrast, offer much more variety and flexibility. Through these components, skilled users have access to a vast array of artisanal capabilities; however, leveraging these capabilities requires significant time, expertise, and custom code development.

Despite the availability of so many new tools, several visualization tasks remain challenging, especially for users who lack the time and/or skill necessary to build custom applications. Some of these challenges include visualizing data at scale; displaying relationships between data and correlating multiple types of information; visualizing errors, biases, and uncertainty;



visualizing algorithms and complex analysis techniques; creating effective graphics quickly; sharing and distributing interactive visualizations; finding the "best" visualization for a particular use case; evaluating the effectiveness of a visualization; and finding reputable, comprehensive information about visualization techniques and best practices.

Where do we go from here?

Several of the challenges mentioned above are active areas of development within the technology sector. For example, a number of companies are tackling the various problems associated with visualizing data at scale — from increasing processing speed to improving the legibility of displays. Visualization recommendation and automated reporting are areas of significant activity and some front-end design challenges are being addressed by including better visualization templates and graphic standards within popular data analysis platforms.

However, other data visualization challenges will not be solved by technology alone. Developing better methods to display errors, biases, and uncertainty, to visualize the behavior of algorithms, or to evaluate the effectiveness of a particular visualization technique will require interdisciplinary research efforts that incorporate an understanding of statistical analysis, digital storytelling techniques, human perceptual limitations, and visual and interaction design principles. Data scientists and visualization developers will be more effective if they have training in design and storytelling skills, or are able to collaborate with designers who have expertise in these areas. Additionally, to combat bias and misinformation, there is a need to develop review standards for data visualizations and better visual literacy among audiences, particularly when it comes to the representation of quantitative and statistical information.

As we rethink our relationship to data, we have a tremendous opportunity to reimagine the way we work, to experiment with new ways of producing and sharing knowledge, to question our own biases, how we draw conclusions, and what constitutes evidence. Data visualization — translating data into an accessible visual representation — can aid these discussions, allowing experts, analysts, and decision-makers to participate in the conversation, drawing on varying expertise to weigh the results of quantitative analysis. But this will only be possible if we design the tools, the interfaces, and the mechanisms that enable this experience. Q

Andrea Brennen *is an artist, designer, and visualization researcher. As IQT's Visualization Specialist, her focus is emerging technologies and techniques at the intersection of data visualization and user experience design. Prior to IQT, Brennen was a member of the technical research staff at MIT Lincoln Laboratory where she led a team of developers designing and building software to help analysts make sense of multi-dimensional communication network datasets. She received a M.Arch. in Architectural Design from MIT and bachelor's degrees in Mathematics and Studio Art from Grinnell College. Brennen's work has been published and exhibited in a variety of technical and creative forums around the world.*

REFERENCES

[1] Gibbs, Frederick A. and Erna L. Gibbs, preface to *Atlas of Electroencephalography* (Cambridge, MA: Cummings, 1941). As quoted in Lorraine Daston and Peter Galison's *Objectivity* (New York: Zone Books, 2010).

[2] Two active participants in this discussion of storytelling are academics who also have strong ties to industry: Jeffrey Heer runs the University of Washington Interactive Data Lab and is a cofounder of Trifacta and Robert Kosara was on the faculty at UNC Charlotte, but now works at Tableau.

A Look Inside

This issue of the *IQT Quarterly* examines technologies, trends, and challenges associated with visualization.

We open with a Q&A with John Maeda, an industry thought leader and Design Partner at VC firm Kleiner Perkins Caufield & Byers. He discusses how startup technology companies should be thinking about design and how organizations can align leadership and human capital around design.

Irene Ros of Bocoup walks through a user-centered visualization process that includes research, data analysis, information design, and validation. A case study on Bocoup's Stereotropes project demonstrates this process and its effectiveness for surfacing patterns and insights from data.

Anthony Vanky examines how visualization can be used to inform community decision makers and transform urban environments. His work at the MIT Senseable City Lab uses data visualization to gain insights into transportation patterns, energy consumption, and waste infrastructure.

Next, we shift to "Design Strategies for Data Exploration," a series of tools, technologies, and tactics for scalable data visualization. Todd Mostak opens this section with an overview of MapD, describing the platform's ability to create a more seamless data exploration experience with faster hardware.

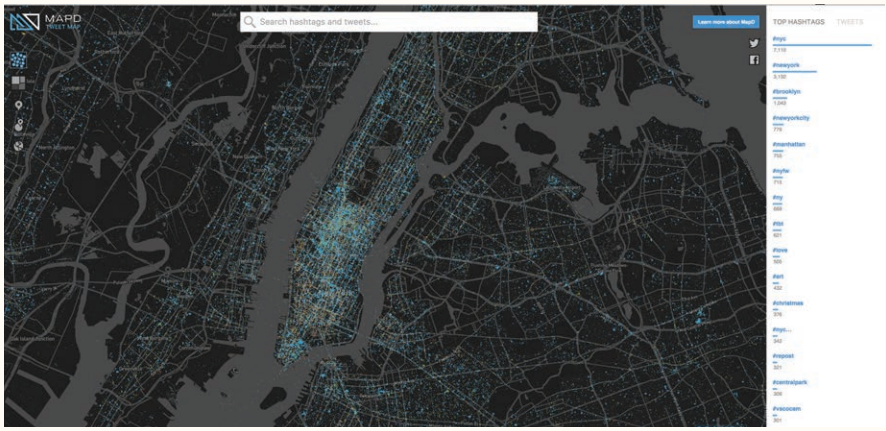
Nicki Dlugash of Mapbox explains the importance of a well-designed basemap using an example of geotagged tweet data. This visualization flips the typical hierarchy of data to map features, using a simple basemap to allow the tweet data to highlight points of interest.

Peter Wang and Hunt Sparra of Continuum Analytics consider the problem of overplotting, which occurs when the number of elements plotted exceeds the numbers of pixels available — for example, in a dataset with millions or billions of points. Continuum's Datashader library applies statistical transformations and sampling techniques to solve this problem automatically and dynamically at different zoom levels.

Leo Meyerovich and Paden Tomasello of Graphistry describe the pitfalls of legacy network and graph visualization tools that were built for smaller scale data. As networks become more complex, Graphistry's platform provides scalable visual analysis for incident responders.

We close the issue with commentary and visualization examples from the Plotly team. While many visualization workflows entail a variety of tools and languages, Plotly provides cross-language compatibility and social functionality to facilitate collaborative data visualization.

Beyond the technologies presented in this issue, there is a range of innovation taking place in the commercial and open source visualization communities. As datasets continue to expand and inform mission critical decisions, it will be key for the Intelligence Community to maintain evolving awareness of visualization tactics and solutions.



#nyc	7,118
#newyork	3,132
#brooklyn	1,043
#newyorkcity	779
#manhattan	755
#nyfw	715
#ny	669
#tbt	621
#love	505
#art	432
#christmas	376
#nyc...	342
#repost	321
#centralpark	309
#vscocam	301

This issue's cover image is a visualization of tweets in New York from October 2014 to February 2015. Colors represent different data sources (iOS, Instagram, Android, Foursquare, Windows, Blackberry, and others). The visualization was created with MapD and leverages a Mapbox basemap.

Design, Human Capital, and the Value of Leadership: A Q&A with John Maeda

For years it seemed that the solution to every problem in technology was to build a smaller, faster, better chip. But today, a growing number of companies distinguish themselves not only through efficiency and optimization, but with innovative design — by building products that are more intuitive and more portable, easier and more enjoyable to use.

In his widely read *#DesignInTech Report 2015,* John Maeda documents several trends that reflect the importance of design in technology.[1] He discusses the success of companies like Airbnb and Mint that were founded or co-founded by designers and the acquisition of creative firms by tech giants including Google and Facebook. Before joining top-tier venture capital firm Kleiner Perkins Caufield & Byers (KPCB) in 2014 as the company's first Design Partner, Maeda served as President of the Rhode Island School of Design (RISD) and, prior to that, was a tenured faculty member at the MIT Media Lab.

This Q&A is adapted from a conversation between John Maeda and Andrea Brennen on December 22, 2015.

Andrea Brennen: The design of products is becoming a much more important factor in the success of technology companies. Why is this happening now?

John Maeda: The first reason is mobile computing, which has enabled everyone to have a graphical supercomputer. More people are using computing devices and those people have needs that hadn't been invented 20 years ago. But now, computing is in the realm of the regular consumer and regular consumers' standards for usability are much higher.

The theme of the #DesignInTech Report was this notion that if you check email twice a day like we used to, when you have a bad experience, it's okay. But now we're unlocking our phones every few seconds; if the experience is bad, it hurts all day.

The second reason is the passing of Steve Jobs. That was a huge moment. Suddenly companies are asking, "Apple is so successful. Why aren't we that successful? What is this design thing?" [Business leaders] are thinking, "Now that he's gone, maybe I can be the one."

AB: With software, it's common to think about "good design" in terms of "usability," or making something easy to use. When we think about design in technology, what else should we keep in mind?

JM: I break up the world of design into three categories. One is physical design, another is social design, and the third is cognitive design. Physical design is the design of materials in relationship to our bodies and it's changing because of the Internet of Things, as physical devices are getting smarter. Social design is the design of environments, places, people, and organizations. Cognitive design of information was traditionally static, or "graphic," but has become dynamic with the computer.

The most interesting kind of design is the intersection of all three — physical, social, cognitive — but because no academic organization controls that, there are no seeds being planted. Industry is the most fertile opportunity, but because industry has the quarterly earnings problem, they can't address this head-on as a "problem." I think it's also a human capital problem — we need more people who understand how different phenomena are changing and who can leverage their connections to other people who are dealing with the same problems.

AB: What should companies do to bring design into their organizations?

JM: At Kleiner Perkins I've worked with over 80 companies. I now know how to work with them at all stages and it's really simple: it's about aligning the organization around design from top to bottom and bottom to top. When you create a culture where design is valued through the whole chain, design naturally gets better. But this requires leadership at all points; it requires talent at all points; and it requires people valuing the product to be designed versus not designed at all.

With startups, it's all about the speed of getting larger. It isn't an idea problem; it's an organization problem. And you don't solve organizational problems through technology. Instead, it's things like: how is the design culture growing? How does the CEO relate to design? Does the CEO see design as more than picking a color?

AB: Do you find that people who aren't familiar with design have misconceptions about what designers do?

JM: People come to me all the time and they say, you're going to love this because it's been *designed* — and I say I don't really care about that.

One challenge with design and tech is that design has gotten big, so people react by hiring designers. But then they hire the wrong kind of designers. I was talking to an Amazon person recently and I asked, "When Amazon hires engineers, what kind of engineers are they hiring? Civil engineers? Mechanical engineers? Chemical engineers? No. [They're hiring] computer science-y engineers." But when we talk about design, people don't know that there is a taxonomy of designers.

For example, a CEO reads the "design" issue of the Harvard Business Review and decides to hire a designer to improve the product's mobile experience. The CEO knows the designer who recently created an amazingly cool t-shirt for the company, so s/he hires that designer to fix the mobile app. A month goes by and the design of the mobile app doesn't improve at all, so the CEO gives up on design. But it's because the CEO hired the wrong kind of designer — s/he didn't know s/he needed a user experience (UX) designer specifically, and didn't know how to look for one. S/he assumed that all designers are the same.

AB: Why do you think this happens?

JM: There are so few designers that have the training to work in the technology industry. I don't think there's any design department in the world that is serving high tech's needs. Contrast that with engineering. Engineers can generally get a degree and work in high tech, but it's not the same for designers. Design's relationship to business has been a driving force behind all of industry's success with consumers, but in universities, the only place that is taught is in industrial design.

The second problem is with leadership. We have a shortage of designers in this industry and by virtue of that, we have a shortage of leaders [who understand design].

So, a shortage of designers, a shortage of leaders, and then mis-hiring — people who hire the [wrong type of] designer and then decide design wasn't really important. Or they hire business-thinker designers who can PowerPoint you to death, but who aren't actually problem solvers.

The value of human capital is only going to increase, and you have to hire people that can achieve growth. All you're doing is building a team.

AB: What advice do you have for startups that want to build a design culture?

JM: The best startups are going to have to have design executives on their Board of Advisors and on the Board of Directors because that will help them accelerate their company's culture formation. I look at it as my life goal now, to encourage that.

Being on the board of Sonos has been amazing. It's opened my eyes to a new way of working, but it's required having a really visionary CEO who wanted someone at the board level to be able to intervene on hiring decisions, on the design of the organization, on experience questions, and the approach and philosophy of the team.

AB: In the next 5-10 years, which technology sectors will need design the most?

JM: Healthcare and security — any kind of network security. I'm very actively promoting these kinds of companies so that people start to think [focusing on design is] normal. Design is most interesting when you introduce it to a place that doesn't have it.

I don't advocate for design for every possible thing, but if you're consumer oriented, you have to be design oriented.

AB: What mistakes do startup companies make when it comes to design?

JM: One thing I'm seeing is that post-Series B companies generally make the following mistake. At the Seed and Series A stages, everyone pitches in; engineers set up routers, designers make business cards. But the company keeps growing and at some point, post-Series B, they need a headquarters and the following happens:

The Executive team makes a call; they think, "We're moving into a new headquarters and someone has to do the interior — let's get the design team on to work on it." This is coming from a good place, but a month after they're done, the designers have lost their identity as the people making the product. Now they're the guys responsible for picking out the chairs.

So I tell everyone, when you get to this stage, don't let your product design team design your interior. You'll want to save money and the designers are more than happy to do it, but at that stage, you don't ask the engineers to install Office on everyone's computers. If you generally don't do that, then you need to apply the same rule to your designers. Otherwise their growth will be stunted for a curious reason that no one will understand.

AB: You were a tenured professor at the MIT Media Lab and after that, the President of the Rhode Island School of Design (RISD). What made you want to leave academia?

JM: In 2001, I was an MIT professor. Other professors were upset because they'd lost money during the dot com crash and there were budget cuts, but everyone would say to me "Don't worry about the money because you're a creative person." After the third person said that to me, I went to get my M.B.A. as a hobby.

When I ran RISD, I did it because I wanted to run a business. Although I had the misfortune of arriving as chief executive just when the global financial crisis of 2009 began, the upside was that I had to learn how to run a $150 million business really quickly, and at scale.

AB: I imagine it was still a big transition to go from university president to venture capitalist?

JM: People say that to me, but every world is the same. Whether it's government, a company, a university — every world is interfacing with the environment. Either it is doing a good job interfacing with the environment, or not. If it's doing a good job interfacing with the environment, it's successful. If it's very insular, it's generally less successful, and more political. Academia, research labs, startups, big companies — the common denominator is people. Organizations thrive by having high quality people.

AB: How do the people in venture capital compare to the people at MIT?

JM: At MIT, everyone is crazy. And everyone is brilliant. It's a meritocracy. Venture is the same — it's a meritocracy.

What is different...is that they move faster in different dimensions. For example, research labs at MIT can make new ideas quickly because they are so agile. Many of the things I see in VC now, I saw in research labs 20 years ago. In research you're making new ideas and the stakes are "low," because not everything has to work. But in venture, it's like you've got a money cannon. You shoot money at an idea and hope it grows faster.

> **The best startups are going to have design executives on their Board of Advisors and on the Board of Directors because that will help them accelerate their company's culture formation.**

AB: Whether we're talking about introducing design culture or new technologies or both, inertia can be so powerful. It's difficult to come into an organization and tell people that they should change something that they've been doing for a long time. How can we do a better job of bridging the gap between old, trusted ways of doing things and new tools or methods?

JM: The skill you're talking about is change management. And it's a core skill of a leader.

Have you read Switch?[2] I thought it was a good framing of change management. The Heath brothers give the example of a man going to get his hair cut and getting a card with 10 spaces on it; when he gets 10 punches, he'll get a free hair cut. If he gets a card with 10 spots and only one punch, he won't come back. But if he gets a card with 10 spots and three punches, he will, because he doesn't have to work as hard.

Tech people don't need a lot of change management, but regular people do.

[While I was at RISD,] I hired a political strategist and it changed my life. She listened to me for a half hour and then said, "Oh, I know your problem...you think if you do the right thing and work hard, things will go your way. And your whole life has proved that." And I said "Well, yea." And she said, "John, you are absolutely wrong."

She showed me a bell curve. Pointing to one end, she said, "That's the part that likes you. And the other end, that's the part that's anti-you. The middle is the undecided majority. What do you think their goal is?"

I said, "Well, they want to do the right thing." She said, "No. the undecided majority just wants to be on the winning side when the dust clears." And then it all became evident to me — to win the majority, I had to win everyone one by one. And to do that, I had to become a new person.

As a creative person, sometimes the only way to keep being creative is to become a new person. We have to become new people or we become stale.

I like the phrase by Ray Kroc: "I'd rather be green and growing than ripe and ready to rot." Q

***John Maeda** is an American executive spearheading a new convergence across the design and technology industries. He currently advises dozens of technology businesses as a partner at Kleiner Perkins Caufield & Byers — a world-leading venture capital firm in Silicon Valley. An internationally recognized speaker and author, Maeda's books include* The Laws of Simplicity, Creative Code, *and* Redesigning Leadership*. Maeda holds degrees in Electrical Engineering and Computer Science from MIT, an M.B.A. from Arizona State University, and a Ph.D. from the University of Tsukuba in Japan.*

REFERENCES

[1] Maeda, J. *Design In Tech Report 2015* [PDF document]. Retrieved from http://www.kpcb.com/blog/design-in-tech-report-2015

[2] Heath, Chip, and Dan Heath. Switch: How to Change Things When Change Is Hard. Waterville, Me: Thorndike Press, 2011. Print.

Visualization Design Process: Turning Numbers Into Insight

By Irene Ros

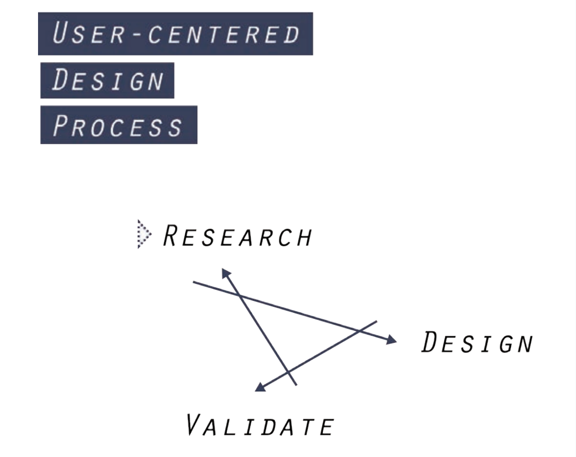
For many of us, working with numbers is a daily task and conveying their importance to others is a core responsibility. From spreadsheets to reports and presentations, we translate values, figures, and statistics into something our audience will care about. This is not a trivial process. Most numbers are insignificant or meaningless without context. And so, we need to weave a narrative around them to explain their importance and impact. For example, even the most basic of data, such as quarterly sales represented by a line graph, will lead to more questions. Have sales gone up or down? Is this different from last year? What does that actually mean to our bottom line? Creating data visualizations that can answer these questions involves a design process of inquiry and planning.

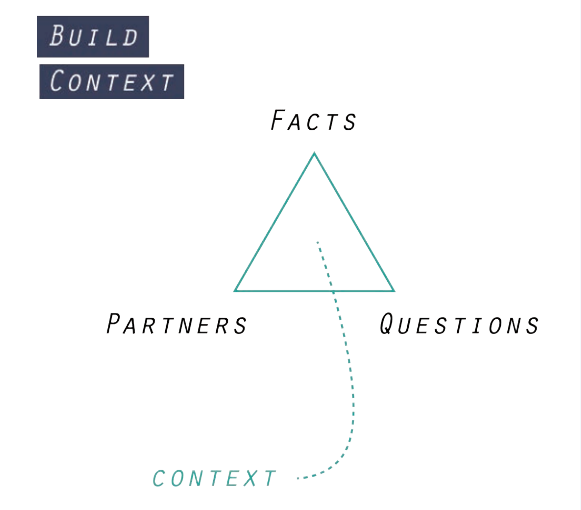
We understand the need for good visualization, but can we identify a process to follow when visualization is needed to explain data? At Bocoup, years of experience have led us to a data visualization design process rooted in a user-centered approach, a process at the core of which is an iterative cycle, focused on research, implementation, and validation. Each of these steps feeds the next, creating a cycle of feedback and improvement. From the outside, a design process may seem like a linear undertaking, starting with a question and ending with an answer. In reality, the process is much more cyclical in nature, enabling the kind of discovery needed to understand and sufficiently solve the issue at hand. We often think we know what we want to create or what questions we want to answer, but only deeper, iterative inquiry can lead us in the right direction.

Research and Data Analysis

The first step of any design process is to gather as much information as we can about the context we operate in. This context building is incredibly important to creating successful data visualization — the circumstances under which our readers will be interpreting our visualization work will impact the way they react to it. For example, given our simple sales chart mentioned previously, if there was no noticeable change from month to month in sales revenue, our readers could interpret it as “a good year” if our overall revenue was higher or “a bad year” if it was lower than expected. The context of the overall financial health of a company will be on the mind of anyone reading this chart. Some of our readers will have the information they need to interpret the charts in this context, but some will not. It is our job to provide all the information we can, including the context needed to understand the visualization.

Context building is done traditionally through stakeholders who can be our partners through the discovery and implementation process. These partners can help us gather valuable facts that inform our design choices and help us identify the questions we should be asking of our data. This first part of context building helps us understand the environment we are working in.





Context alone is rarely enough to answer some of the questions we identify with our stakeholders — there are new insights to be gleaned and communicated from the data we have. This is why visualization is important to begin with. Before we can visualize our data, we need to acquire, transform, and analyze it. Thanks to the guidance of our context, we know what data we should obtain, how exactly we'll need to manipulate it, and what analysis to perform — whether it is through traditional or Bayesian statistical analysis, machine learning, or other analytic methods. The insight gathering process, much like the user-centered design process itself, is a cyclical one. One insight can result in the need to obtain more data, or new data can result in the need to re-run or adjust our analysis methods.

This analysis will enable us to gain new insights that will either confirm, add to, or refute what we know. Data analysis doesn't always yield new answers, but insights gleaned can be valuable for gaining a more complete picture of context.

Information Design

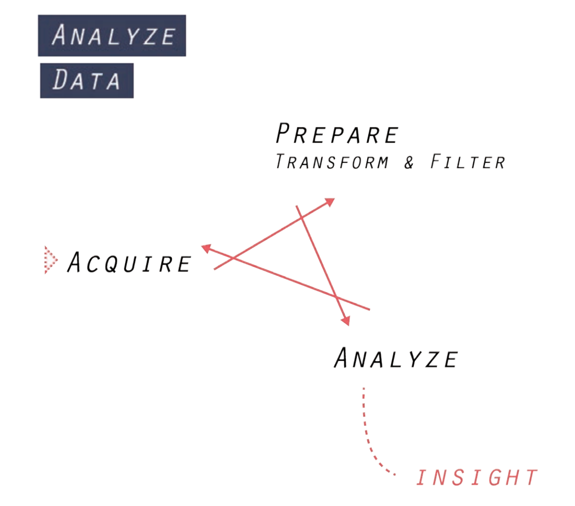
After we find insights in the data that are relevant to our audience, we begin the process of finding the right visual mapping. There are hundreds of ways to represent the same data and the right choice of visualization isn't always clear. This is where experimentation becomes key to finding the right mapping. We begin by choosing a visual mapping and prototyping it (on paper, spreadsheets, code, etc.). In the process of prototyping visualizations with actual data we often discover a narrative — we identify certain points or patterns of interest that are worth highlighting. We explore user interactions that can highlight those appropriately and validate with our stakeholders that our findings are indeed of interest, informative, or noteworthy. More often than not this process is repeated several times until we zero in on the narrative that matters most and the interactions that do it justice.

Much like our other processes so far, this process is cyclical, informing our understanding of our data, the expectations of our audience, and narrative we are conveying.

Validation

A common thread across all aspects of the design process is the continuous need for validation. During the research phase, validation enables us to ensure that our understanding of stakeholder needs is accurate. Building context is done through activities such as user interviews, defining journey-maps of current and expected user behaviors, and understanding the user stories of the system being designed. Many of these activities lead to consensus or highlight divergence in our understanding of the system needs. They are also meant to elicit new information, but over time that information begins to coalesce into a common narrative, at which point, we are validating what we know more than eliciting new context.

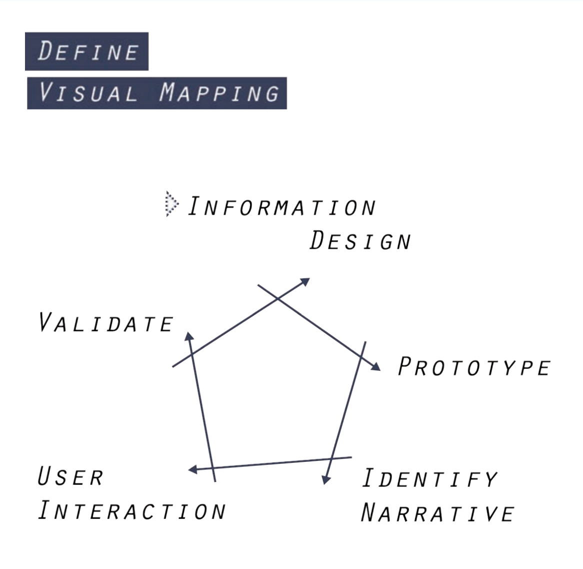
Validating the results of our data analysis with our stakeholders allows us to check whether the insights



we are finding are significant and make sure we are making proper use of the data we have. At times, data can be incomplete or incorrect and validation allows us to identify that through conversations. While we perform our analysis, we frequently communicate our findings and ask our stakeholders for a "gut-check reaction" to what we find. More often than not, their insight will add a layer of narrative or context we didn't detect and they didn't think to share with us prior to our findings.

Lastly, during the information design phase, validation allows us to make sure we are designing the right system for the questions at hand. There are many visual methods to choose from, and finding the one that targets our needs most directly isn't always easy — it is a combination of our experience and our stakeholder needs that gets us to the right answer. Not surprisingly, we've represented very similar data in very different ways over time, because the system goals were different.

Most importantly, validation enables us to develop and maintain a culture of feedback throughout the process between us and our stakeholders. Visualization displays are living and breathing artifacts, and the art of designing them is often as informative as the displays themselves. Q



Case Study: Stereotropes

***Irene Ros** discusses how Bocoup used the data visualization design process to produce Stereotropes.*

Complex characters in films often fall into familiar patterns called "tropes". Tropes are devices and conventions that a writer can reasonably rely on as being present in the audience's minds and expectations. Some tropes are cliché (stereotyped and trite), while others put into plain sight what is otherwise hidden or unspoken.

Stereotropes is an interactive visualization experiment built by the Bocoup Data Visualization team, exploring a set of tropes authored and tracked by TV Tropes, a collaborative wiki-based community.

Stereotropes begins with a specific subset of tropes identified as being 'Always Male' or 'Always Female.' This vast collection of data allow us to ask interesting questions about the representation of gender in films. Are certain adjectives more likely to describe a female or male trope? How are these adjectives used together? Through our work on Stereotropes, we were able to launch a data-driven conversation about these questions.

User Research

Working with gender-related data is a challenging undertaking. We worked with stakeholders — colleagues, partners on the Open Web, film aficionados, and gender equality activists — to inform the way we portrayed this gender data. These stakeholders asked valuable questions about the dataset, which we were then able to translate into data exploration tasks, for example "How are the two genders described differently?" or "What traits do characters share?"

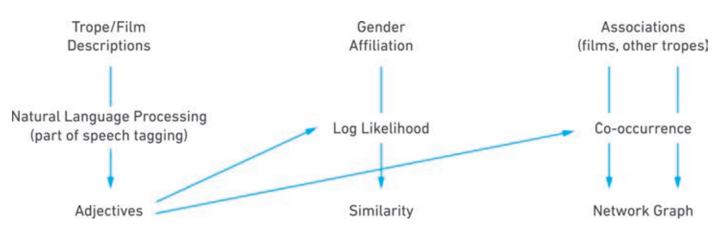
Our stakeholders also proved to be invaluable throughout the project as they validated our designs and analysis results. We wanted to remain as impartial as possible, so anything from color choices to visualization techniques had to be representative without exaggeration, engaging without being sensational, and provoking but only if true.

Data Research & Discovery

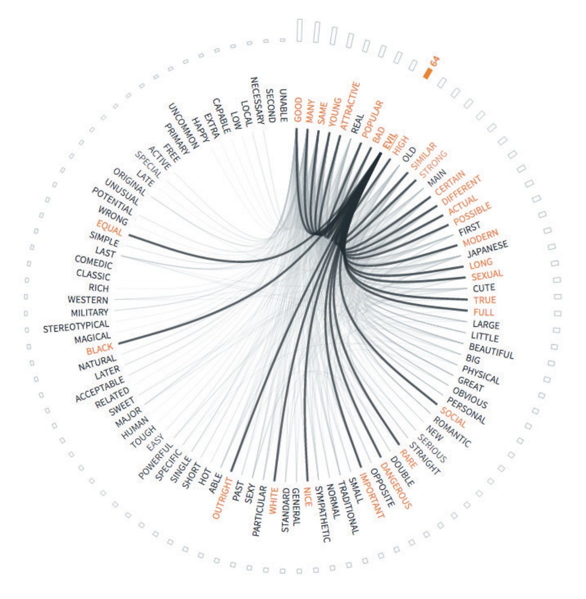
TV Tropes is structured as a collaborative wiki, where community members describe the tropes they see in colorful and highly vivid language, creating a rich database. Our data for Stereotropes came from an intermediary source called DBTropes, offering a Linked Data version of the TV Tropes wiki. We also augmented our data with film metadata from Rotten Tomatoes and the Open Movie database.

We were able to focus on a few different elements of our data, specifically the descriptions of the tropes, the associations between tropes and the films they appear in and the affiliation of specific tropes with a primary gender. From the descriptions we extracted all adjectives used across all tropes, which enabled us to aggregate tropes, consider the likelihood of any one adjective appearing in one trope or another and how those adjectives were used across genders. The adjectives also allowed us to build a similarity network of tropes that resembled each other through the language that was used to describe them. Lastly, because films had a release date associated with them, we were able to inquire into how the usage of tropes has evolved over time, and whether their frequency had changed.

Thanks to this data we were able to explore many questions such as, "Are there specific adjectives that are commonly used to describe one but not the other gender?" (yes) and "Are there any interesting patterns in the way language is used in tropes?" (yes). Our explorations yielded interesting results. For example, many of the top adjectives associated with female tropes relate to youth, physical appearance, traditional gender roles, and sexuality. On the other hand, many words associated with physicality or heroism (and anti-heroism) are associated with male tropes. While many of the tropes matched challenging stereotypes, others were indicative of a more positive shift. The use of the trope *Damsel Out of Distress*, describing a female character who doesn't wait for her hero to save her, is on the rise while many of the more negative tropes, such as *Femme Fatale* are on the decline.



Adjectives extracted from trope descriptions allowed us to form connections between other adjectives, tropes and films.



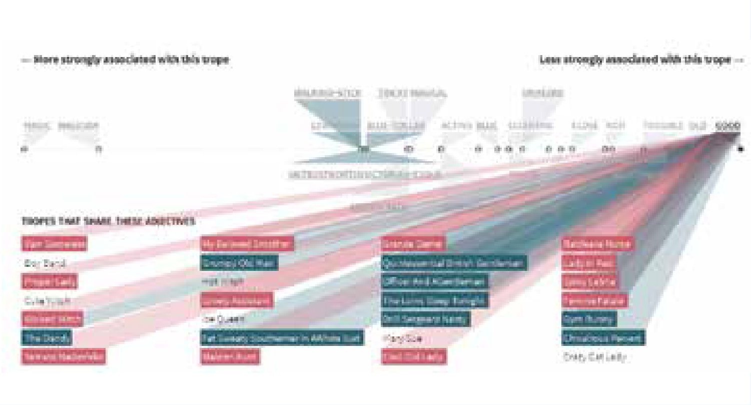
The top 100 adjectives appearing in trope descriptions. Adjectives that occur together in more than 10 tropes are connected using grey lines. Hovering on a word shows the adjectives that occur together most often.

Information Design

Our findings in the data spurred many visual explorations of how one might best represent these insights. Visualizing text is a fairly new field, and we explored relationships and links in a way that required novel visual techniques. We started by creating a quick visual way to browse through our trope/adjective collection, which very clearly alluded to the importance of exploring links between tropes through shared adjective usage. This became a core metaphor throughout Stereotropes.

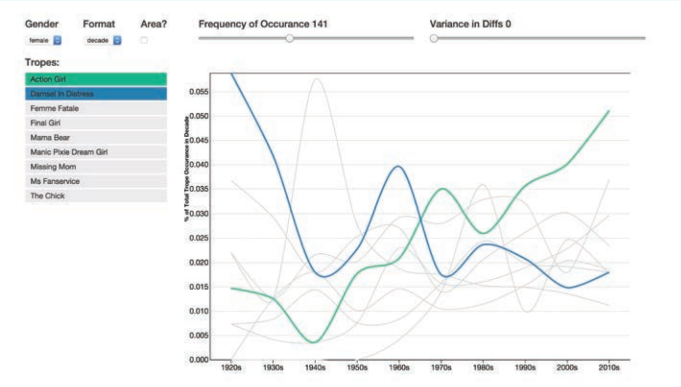
We wanted to combine linking of tropes through shared adjective usage with a representation of which adjectives were specifically utilized in a trope's description, and how likely they were to appear in that trope versus the other tropes in our dataset (for example, an adjective like "good" or "bad" was fairly common and thus not any more likely to appear in any one trope). Our explorations eventually led us to a novel visualization technique, combining a linear timeline-based view for our likelihood score with a network visualization.

The prototyping process is a great way to eliminate candidate visualizations quickly. Initially, we wanted to make trope usage over time a core aspect of Stereotropes, but we ended up focusing on trends. Each trope in our dataset listed all the films it appeared in. The films thus belonged to a decade, which allowed us to connect the tropes over time more directly. We developed a prototype tool to allow for comparing the change in usage of multiple tropes at once, but there were several challenges with this design. For example, we had a different number of films in each decade. Modern films received better coverage, so we couldn't use raw counts as a basis for comparison. However, when we looked at the percent of trope use per decade instead of counts, we saw three patterns: a trope was becoming more popular, becoming less popular, or its use has stayed about the same. That's what our stakeholders were looking for.

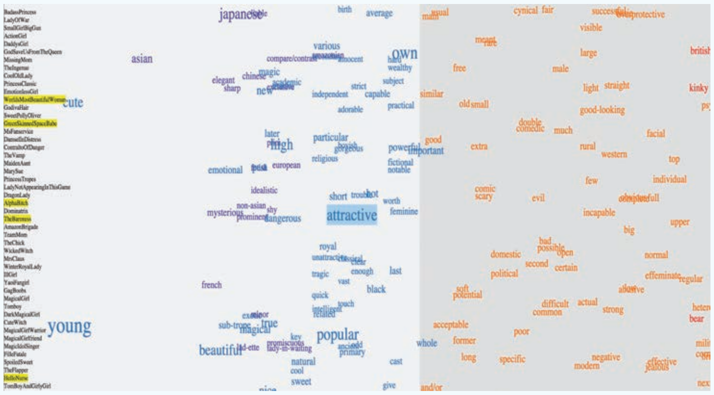

More strongly associated with this trope
Less strongly associated with this trope
TROPES THAT SHARE THESE ADJECTIVES
Boy Band
The Dandy
Grumpy Old Man
Quintessential British Gentleman
Officer And A Gentleman
Drill Sergeant Nasty
Lady in Red
Femme Fatale
Gym Bunny
Crazy Cat Lady

Our trope adjective "timeline", showing the adjective use in a trope, the likelihood of those adjectives being more strongly associated with this trope in particular, and the related adjectives in our data through shared language use.

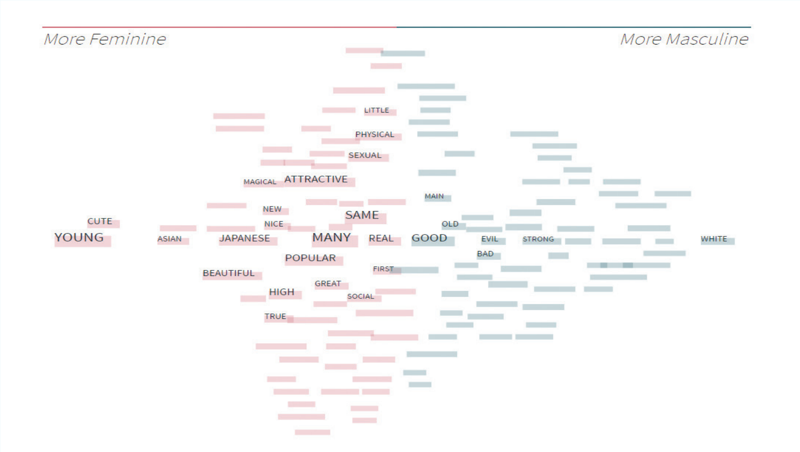
Lastly, some of our prototypes enabled us to see patterns in our data that we simply couldn't discern using analysis alone. The gender visualization started as a prototype for evaluating adjectives across the entire dataset. Specifically we wanted to see if we could find adjectives that were particularly male or particularly female. By looking at the full collection of adjectives, we could start asking questions like, "What adjectives are more likely to describe male tropes but not female tropes, and vice versa?" Or, "What adjectives are fairly gender neutral?" Our initial prototype was incredibly effective and engaging to our users. Our final version resembled our prototype quite a bit as a result.



Example: decline in Damsel In Distress and the increase in Action Girl over time



Prototype version of adjective-gender associations



Final version of our adjective-gender association.

Validation

Throughout the Stereotropes process, we revealed our findings to our colleagues and requested input where possible. Even simple questions like "Should we stick to a familiar pink and blue color scheme or introduce a more neutral one?" resulted in iterations and test cycles that allowed us to evaluate the efficacy of our designs and implementation. We also derived some novel visualizations that needed to be tested for efficacy, readability, and ease-of-use. Prototyping allowed us to do this before we were too invested in any one design.

Stereotropes is an excellent experiment in finding meaning in unstructured text, revealing structures and patterns in natural language and looking for opportunities to reflect on a cultural conversation with data-driven insights. Through our design process we were able to find these meaningful structures and patterns, explore the best ways to represent them, and ensure repeatedly that we were on track to creating something meaningful with the help of our users.

__Irene Ros__ is the Director of Data Visualization at Bocoup. Her team transforms raw data into compelling, visually engaging experiences using Open Web technologies. From research and data analysis to visualization design and development, Bocoup partners with organizations in academia and industry to build custom tools and visualizations that inspire and inform.

Make Data Make Sense: The Importance of Visualization in Data Analytics

By Anthony Vanky

"The variables are many, but they are not helter-skelter; they are interrelated into an organic whole," said urbanist Jane Jacobs. Jacobs was referring to Hudson Street in 1960s Manhattan, but the same could be said of any street in any city today.[1] The key difference is that today, big data allows us to describe and analyze Jacobs' "interrelated variables" better than she could have imagined.

In the current big data paradigm, data is king. Every additional layer, dimension, and dataset promises to bring us closer to a "perfect" representation of reality. Visualization provides fertile opportunities for analysis, helping to situate data in social, economic, and political contexts; it also changes the way we understand those contexts. In James Corner's words, mapping (a specific application of visualization) "unfolds potential; it re-makes territory over and over again, each time with new and diverse consequences."[2] By adding context to data, visualization helps us move beyond a focus on finding specific answers and enables us to ask richer questions that ultimately lead to greater insights. Too often, data analysis is focused on achieving narrow outcomes. But how do we know if we are even asking the right questions?

The MIT Senseable City Lab studies how data can reveal underlying patterns of cities and their citizens. With an abundance of data, question-driven exploration can be challenging in general, but this is especially true in the complex environment of cities. The Lab's work is an example of how visualization can be used as part of a research methodology, as we explore how different representations of data can transform how we understand and communicate city dynamics. This article discusses two projects developed at the Senseable City Lab in an effort to show how visualization can be used to support data analysis and to instigate critical innovation in cities.

LIVE Singapore! Asking Better Questions

Singapore is perhaps the city that is most aggressively leveraging technology and computing to transform an urban area into a more livable community. Data is produced by and collected from an increasingly large and diverse set of infrastructures, forming an urban living laboratory. However, the current data ecosystem is segregated. And while segregated data is useful for individual operations, data contains much more value when it is combined from different sources and made accessible to users through creative new applications.

The LIVE Singapore! Project aims to close the feedback loop between citizens and the digital, real-time data collected across the multiple urban infrastructure networks that sustain urban life. The project seeks to develop an open platform for the collection, fusion, and distribution of real-time data that originate from a large number and variety of different sources.

With current capabilities, investigating this data is difficult. There are limited tools that allow for quick data visualization and intuitive analysis of discrete datasets from various spatial and temporal sources. The goal of the LIVE Singapore! project is not only to offer new capabilities, but to inform the analytical and modeling process — to make it question-driven, instead of answer-oriented. Through LIVE Singapore! the MIT Senseable City Lab is creating several visualization tools that will allow researchers and policymakers to quickly and intuitively survey the city's data. Overlaying these different types of data, within the specificities of their contexts, allows for a type of exploration that will shape questions for further analysis.

For example, Singapore's population is heavily reliant on public and shared transit. And due to its location on the equator, tropical rainstorms are often heavy, fast-moving, and short. By viewing the map in Figure 1 — showing rain data overlaid with real-time location and availability of taxis — researchers could quickly form new questions, test hypotheses and gain insight into how the country's transportation infrastructure might be impacted by heavy rains during the wet season. Through this exploratory analysis, we ultimately found that taxis tended to avoid areas with heavy rainstorms despite the ease of picking up a quick fare. We then confirmed this finding analytically and shared it with the Singaporean government for consideration in future policy decisions.

In another study, the Lab investigated correlations between electricity usage and ambient temperature to better understand (and demonstrate) how exhaust from air conditioning increases the urban heat island effect, thus creating a viscous cycle that drives further energy consumption.

Trash Track: Driving Individual Behavior

In Trash Track, the Lab studied refuse data from Seattle to explore the potential of making an otherwise invisible infrastructure — our waste infrastructure — visible. Due to complex sorting processes, regulatory environments, and fragmented ecosystems of waste handers, it is difficult to generate a global understanding of the waste stream. While many industries thoroughly understand the processes by which their products are created, surprisingly little is known about what happens when we discard an object. As a response, Trash Track can be viewed as an inversion of the supply chain by examining the "removal chain."

A dramatic drop in the cost of sensors is creating a new paradigm in technological development — the Internet of Things is connecting objects of everyday life to the Internet, enabling these objects to contribute their data to tell their own stories. Researchers in the Lab leveraged the dramatic drop in the cost of sensors in order to gain insight into this otherwise opaque removal chain.

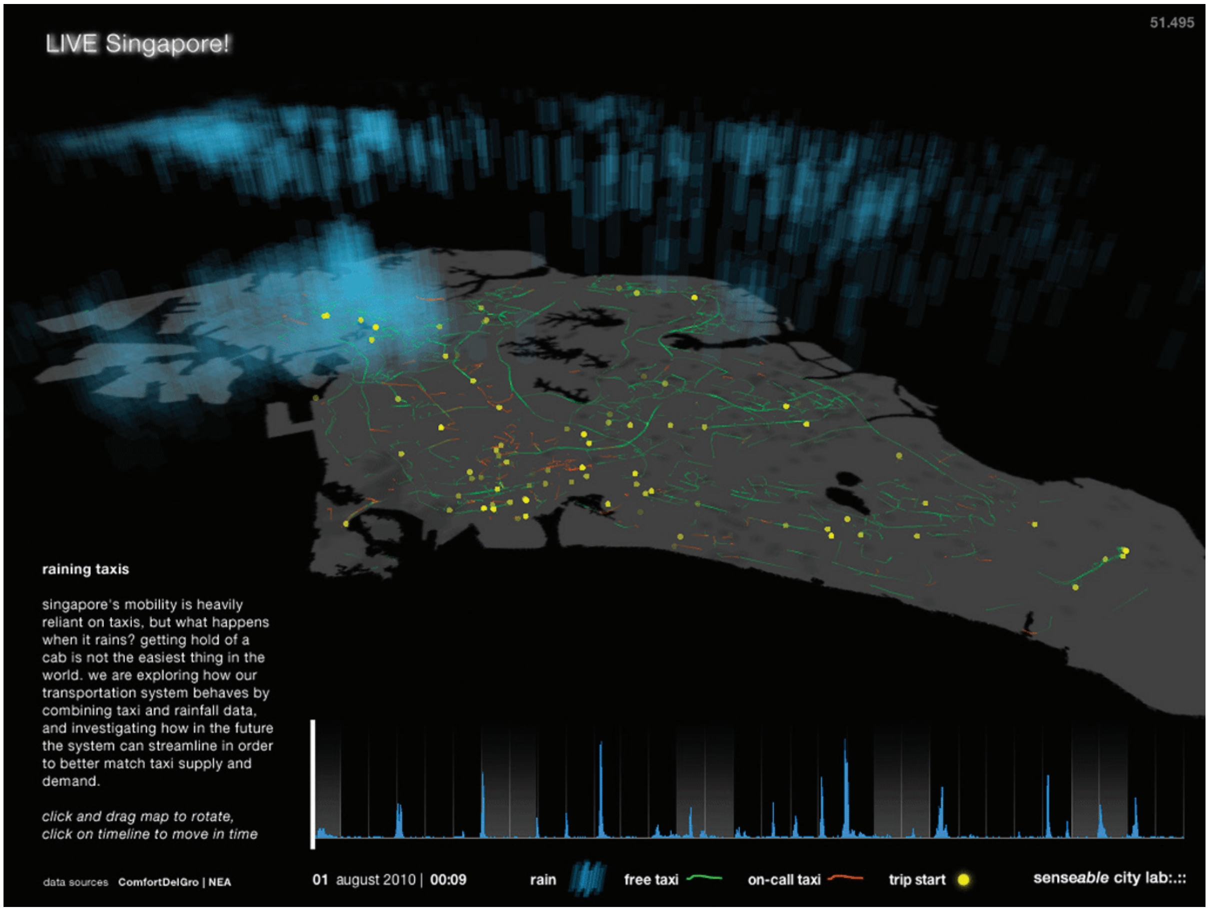
The researchers designed and placed small sensors — including GPS sensors that communicated over a cellular phone network — into the refuse of 500 volunteers from Seattle. Residents disposed of the tracked items as they normally would, allowing the researchers to record the movement of various types of household waste. While a bulk of the waste went to local disposal and recycling centers, over time, the waste footprint grew to span the breadth of the continent. From Washington to Georgia, trash traveled to the most efficient — measured by cost or by regulation — resting place.

While the data itself is powerful in providing the first characterization of this removal chain, the map in Figure 2 reveals the complexity of waste infrastructure. Individuals can visually trace objects as they travel from city to city, state to state, or in one case from Seattle to the Midwest, only to come back to the Pacific Coast. Without the visual map, some of these paths would remain hidden from view and awareness.

Implicit in the act of visualizing is the desire to transform abstract data into relatable stories that engage and inform conversation. The true power of this visualization is how it can incite change. By seeing her own personal waste map, one volunteer decided to swear off bottled water after seeing how quick decisions made out of convenience led to long-lived negative consequences — plastic bottles sitting for eternity in a landfill.

The Promise of Visualization

Visualization allows for more equitable conversations about data — therein lies its power and vulnerability. Opinions and beliefs can be changed and swayed by



singapore's mobility is heavily reliant on taxis. but what happens when it rains? getting hold of a cab is not the easiest thing in the world. we are exploring how our transportation system behaves by combining taxi and rainfall data, and investigating how in the future the system can streamline in order to better match taxi supply and demand.

click and drag map to rotate, click on timeline to move in time

Figure 1 | The LIVE Singapore! project shows how visualizing different types of data in context can lead to innovative ideas about city operations.

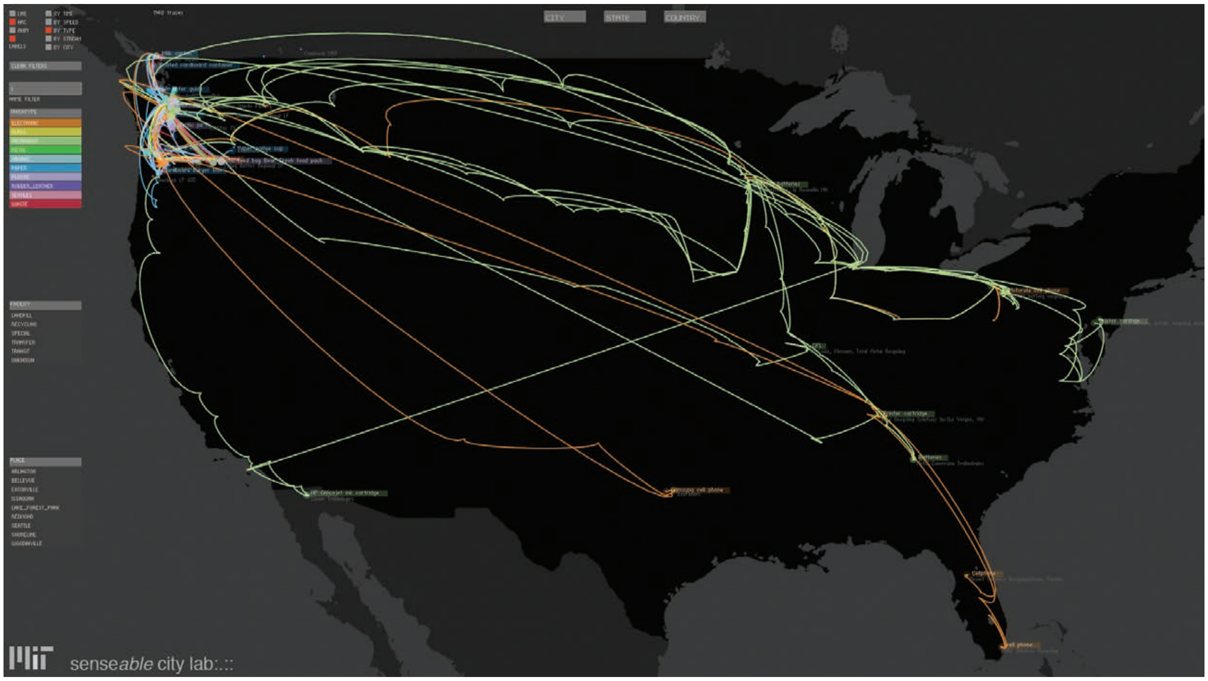


Figure 2 | The Trash Track project visualizes the "removal chain" of waste infrastructure.

these representations, but it is important to keep in mind that the audience is often at a disadvantage, lacking expert knowledge about the topic at hand. Further, as each layer, dimension, and dataset bring a more robust representation of reality, these representations are still imperfect, rife with problems and biases.

The simulated reality is never perfect, but representations of simulated reality — of data — still impact decisions about society, sometimes in negative ways. In the 1970s, New York City burned because of imprecise models that were intended to save the city millions of dollars per year by making the fire department more efficient. The promise of a streamlined, non-partisan technocracy — informed and supported by analysts and modelers — led to lost livelihoods for many of its citizens. However, the possibility of negative outcomes shouldn't prevent us from exploring the opportunities to be had from data. In a way, the current data paradigm is like the weather — despite general distrust of the weatherman, we all listen to the forecast because it's better than nothing.

In some ways, research and design tests the many potential futures that lay ahead of us. Herbert Simon, the noted scientist, believed that design was a unique process as it is concerned not with how things are, but with how the world could be. But the testing of these visions requires dialogue with a broader audience. Without accessible means to do so, these potential futures are left to be forgotten, consensus always out of reach.

By evolving how we use and share data, a new generation of informed decision-makers will be able to transform our communities. Armed with accessible visual representations of abstract data, we will understand more about the world in which we live. Q

Anthony Vanky *is a researcher and former partner strategist at the Massachusetts Institute of Technology (MIT) Senseable City Lab and a Ph.D. candidate in Urban Studies and Planning at MIT. His research considers the use of new, pervasive sensing technologies in measuring and evaluating the built environment. He has widely presented on topics of technology and urbanism, including at EmTech, Harvard University, Skanska Worldwide Research & Development, the Fraunhofer Institute, and the New York Center for Architecture. Trained as an urban designer and architect, he has worked on projects across the U.S., and his design work has been exhibited widely, including at the Venice Biennale, the Dutch Design Week, the Gwangju Design Biennale, and New Orleans DesCours. He previously served in national leadership capacities on each of the five governing organization of the architecture profession in the U.S., including the American Institute of Architects (AIA), and has launched two non-profit advocacy organizations.*

REFERENCES

[1] Jacobs, J. (1969). *The economy of cities*. New York: Random House.

[2] Corner, J. (1999). The Agency of Mapping: Speculation, Critique and Invention. In D. Cosgrove (Ed.), *Mappings* (pp. 213-252). London: Reaktion.

[3] Simon, Herbert A. 1996. *The Sciences of the Artificial - 3rd Edition*. The MIT Press.

Scaling Tufte: Design Strategies for Data Exploration

By Andrea Brennen

In his canonical book *The Visual Display of Quantitative Information,* Edward Tufte notes how powerfully and effectively visualizations can convey data at scale. He writes, "The most extensive data maps place millions of bits of information on a single page before our eyes. No other method for the display of statistical information is so powerful."[1]

In his widely read books, Tufte outlines several design principles that can be applied to large-scale data. For example, he advocates removing *chartjunk,* "interior decoration...that does not tell the viewer anything new" and maximizing *data-ink,* the "proportion of...ink devoted to the non-redundant display of data-information."[2] He discusses scalable techniques such as small multiples[3] and sparklines[4], and recommends using smaller point sizes and lighter line weights.

Throughout his work, Tufte illustrates these recommendations with a rich set of examples. However, a majority of these examples are individually-crafted static images printed on paper. In practice, it is not always obvious how to adapt Tufte's guidance to the design of extensible, interactive visualizations that are intended to be displayed and consumed digitally.

The need for scalable visualization capabilities often coincides with the goal of data exploration — using visualization to explore data and discover new insights. But designing and developing scalable exploratory visualization tools represent a unique set of challenges, many of which have not yet been solved.

Visualizations must scale in several ways

Borrowing the popular notion that big data can be characterized by "4 Vs" — volume, velocity, variety, and veracity — it's worth noting that different aspects of scale impact the design of visualizations in very different ways. Exploratory visualization tools must scale functionally, to accommodate large volumes of data (volume), but this is rarely enough. Analysts often also want to highlight how and how quickly data changes over time (velocity), show how different aspects of a dataset are interrelated (variety), and provide indications of uncertainty, risk, error, and bias (veracity).

As datasets increase in complexity, there may be a desire to visualize individual data points instead of displaying a summary of data, as is common with many familiar chart types (bar, pie, etc.). Summary graphics are extremely useful once one knows what to summarize, but this is not always the case with initial, exploratory analysis. While many big data technologies are capable of querying millions of rows of data, far fewer provide the ability to render or draw millions of points on a screen. And the visualization research community is only beginning to tackle the design challenges associated with representing, navigating, and interacting with millions of points in ways that make the data legible and interpretable to human users.

We don't know what we're looking for

Effective exploratory visualizations should produce or allow for new insights into a dataset, but when these visualizations are being designed and developed, the nature of the desired insight is often unknown. After all, if analysts already had the insight, one could argue they wouldn't need the visualization. As a result, visualization designers and tool-builders are often faced with the challenge of designing solutions intended to satisfy unknown (or even unknowable) requirements. Since it is not possible to customize or optimize a design for all possible use cases, visualization designers and developers must find a productive balance between generalizability and specificity.

In response, many exploratory visualization tools aim to provide multiple ways of viewing complex datasets. The popular idea of a dashboard with visualization widgets offers multiple simultaneous views into a dataset, allowing users to choose and customize what they want to see. With this approach, individual visualizations must be generalizable and extensible enough to satisfy a variety of needs. In some cases, it is also important to allow for additional views that have not yet been anticipated.

Not all visualizations are equally insightful for all data types

With more variety and choice of visualizations, some notion of recommendation or curation becomes increasingly important. However, this is a challenge because a visualization's effectiveness is contingent on a variety of factors that vary in subjectivity and predictability. These factors include the type, amount, and resolution of data being represented, the specific application or use case for the visualization, the audience viewing the data, the form factor of the display, human perceptual limitations, cultural influences and biases, and the desired emphasis. Several companies and open source projects are beginning to tackle the problems associated with "visualization recommendation" and "automated reporting" but currently, this problem is far from solved.

We need to correlate and synthesize multiple types of data

Additionally, the availability of so many analytic and visualization tools creates a new problem for analysts: the need to synthesize information across a variety of disparate displays and user interfaces. As these tools are often designed and developed by different people, groups, and companies, there is little incentive to coordinate design efforts. This can create a difficult challenge for analysts who must constantly switch between multiple screens, tools, or browser tabs to monitor a variety of different simultaneous factors.

We need to visualize more than "results"

Today, data scientists and subject matter experts use a variety of tools and methods to conduct their analysis.

> **The need for scalable visualization capabilities often coincides with the goal of data exploration — using visualization to explore data and discover new insights. But designing and developing scalable exploratory visualization tools represents a unique set of challenges.**

Typically, at the end of their workflow they produce a visualization that summarizes their results or highlights a key finding. While visualization is certainly helpful in this capacity, common visualization types do not always convey important factors about how the analysis was conducted. For example: Where did the data come from? Who collected it and how? What was the sample size and understood margin of error? What algorithmic techniques were used during the analysis? Results are just one part of a larger data story. Visualization can be much more effective if it is used to convey not only the results of analysis, but also relevant methodological and data provenance factors.

To address these challenges, visualization designers, developers, and tool-builders are applying the foundational tenants of information design — such as those outlined by Tufte — but they are also inventing additional strategies for scalable data visualization. In the following pages, we present a series of these tools, technologies, and tactics. This is by no means a comprehensive list, but rather, a snapshot of emerging ideas. Framed as answers to common questions about the practicalities of visualizing big data, they can be viewed as design strategies for data exploration at scale. Q

REFERENCES

[1] Tufte, E. R. *The Visual Display of Qualitative Information.* Cheshire, CT: Graphics, 1984. Print.

[2] Ibid.

[3] "A series of graphics, showing the same combination of variables, indexed by changes in another variable" (Tufte).

[4] "*Datawords:* data-intense, design-simple, word-sized graphics" or "small, high-resolution graphics usually embedded in a full context of words, numbers, images" (Tufte).

Strategy #1: Use Faster Hardware

***Todd Mostak**, CEO of MapD, explains how we can answer questions faster with GPUs.*

Ask Questions Faster

As datasets grow ever larger in size, it is important that we retain the ability for analysts to interactively explore data without waiting seconds or minutes for answers. Immediate results spur the analyst's process of exploration, deduction, and discovery. But immediacy has become hard to achieve across large datasets using traditional CPU-based computing. MapD is developing software to exploit the new paradigm of GPU computing and deliver on instantaneous analysis across multi-billion row datasets.

The 50-year history of supercomputers is full of massive, warehouse-sized machines accessible only to the most elite scientists and researchers. More recently, there has been migration from supercomputers toward large clusters of commodity servers, particularly for data analytics. While these clusters are typically less expensive and less exotic than traditional supercomputers, they still require a significant amount of rack space and electricity, as well as trained technicians. Even with all of these advances, current systems continue to exhibit latencies that prevent true interactivity. While the advent of cloud computing is lessening barriers for heavy compute resources, datasets must often remain behind a company's firewall due to privacy issues or to protect a competitive advantage embodied in the data.

Supercomputer in a Server

Originally designed to render video games, GPUs (graphics processing units) have evolved into general purpose computational engines that excel at performing repetitive tasks in parallel. Buttressed by high-speed memory, GPUs can perform thousands of calculations simultaneously, making them an excellent fit for data queries and many machine learning algorithms; the native graphics pipeline of the cards means they can be used to render large datasets in milliseconds. While GPU performance and memory capacity are still rapidly increasing, in the current generation, up to eight GPU cards with 192GB total GPU memory can be installed in a single server, allowing data to be queried at rates approaching three terabytes per second by almost 40,000 cores. This is like having a supercomputer in a single server.

MapD was founded by scientists, engineers, and data analysts on a mission to make data exploration an immersive experience. The focus is not only to make queries faster, but to create an immersive data exploration experience that removes the disconnect between analyst and data.

Figures 1 and 2 provide examples of some of MapD's visualization capabilities. By default, MapD's platform is built on the cross-filter paradigm, which means that any filter applied to a chart will also be applied instantly to all other visible charts. This helps to highlight correlations between variables, which may be shown in different charts.

Figure 1 shows 500,000 taxi pickups in the Lower East Side of Manhattan for the two weeks around Christmas 2014. The heatmap in the upper right shows that many people stayed home on Christmas Day (Thursday), but the district lived up to its late night party reputation on Christmas Eve and over the weekend.

Figure 2 shows political donations to federal office holders from 2007-2012. In the timeline in the lower left, one can see that Democrats raised more money than Republicans towards the end of the 2008 presidential election cycle. Q

__Todd Mostak__ is CEO of MapD. Mostak developed a prototype of MapD while working on his Harvard thesis on the Arab Spring. While searching for patterns in a dataset of hundreds of millions of tweets, he waited hours (or sometimes days) to process a single query. Frustrated that he couldn't access a cluster of computers to perform these computations, he instead paired off-the-shelf video game GPU cards with a new design for parallel databases and later pursued this technology as a researcher at MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL), before launching MapD in 2013.

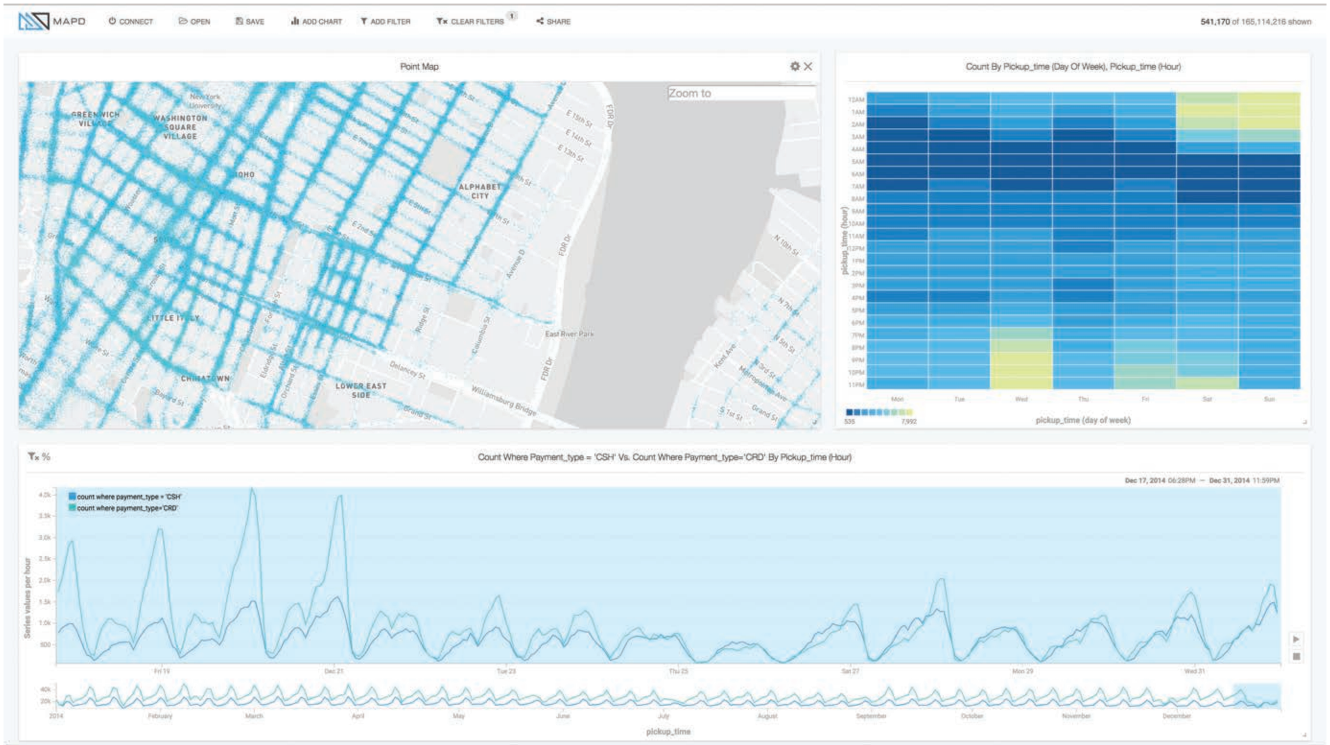


Figure 1 | 500,000 taxi pickups in the Lower East Side of Manhattan for the two weeks around Christmas 2014.

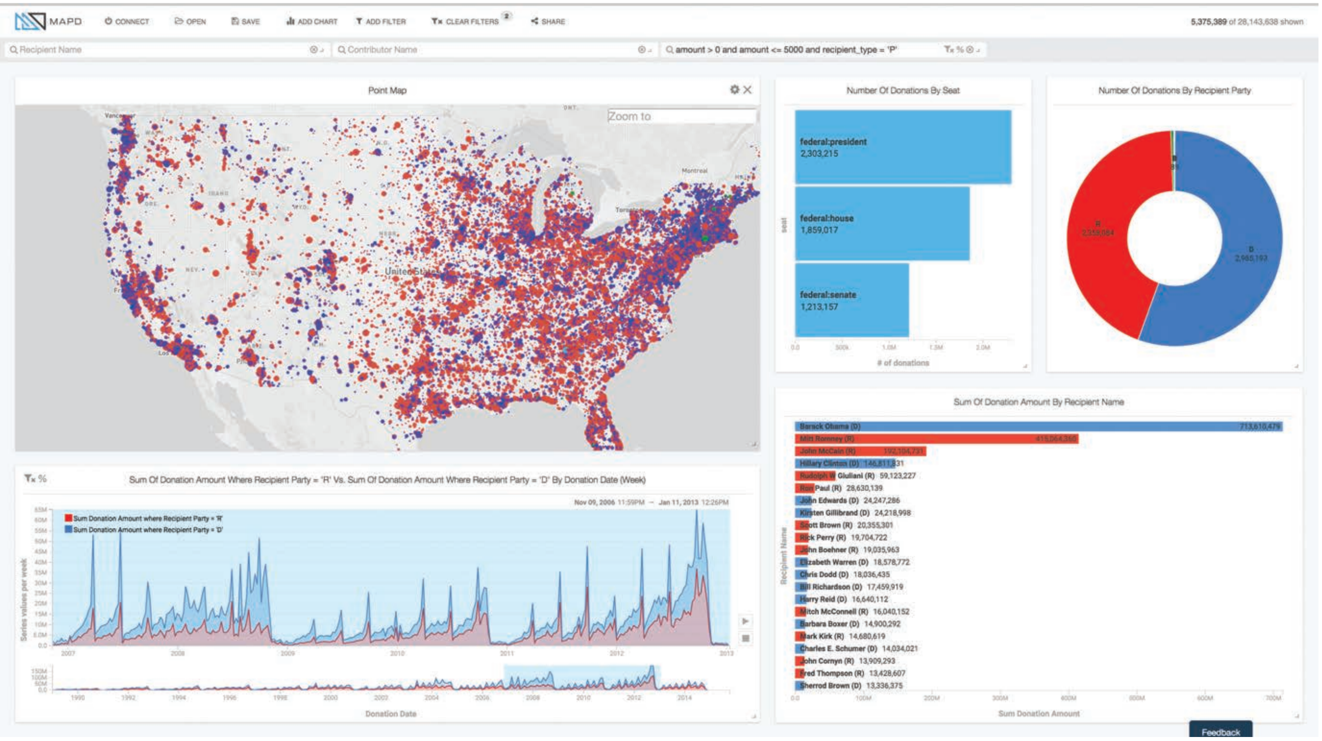
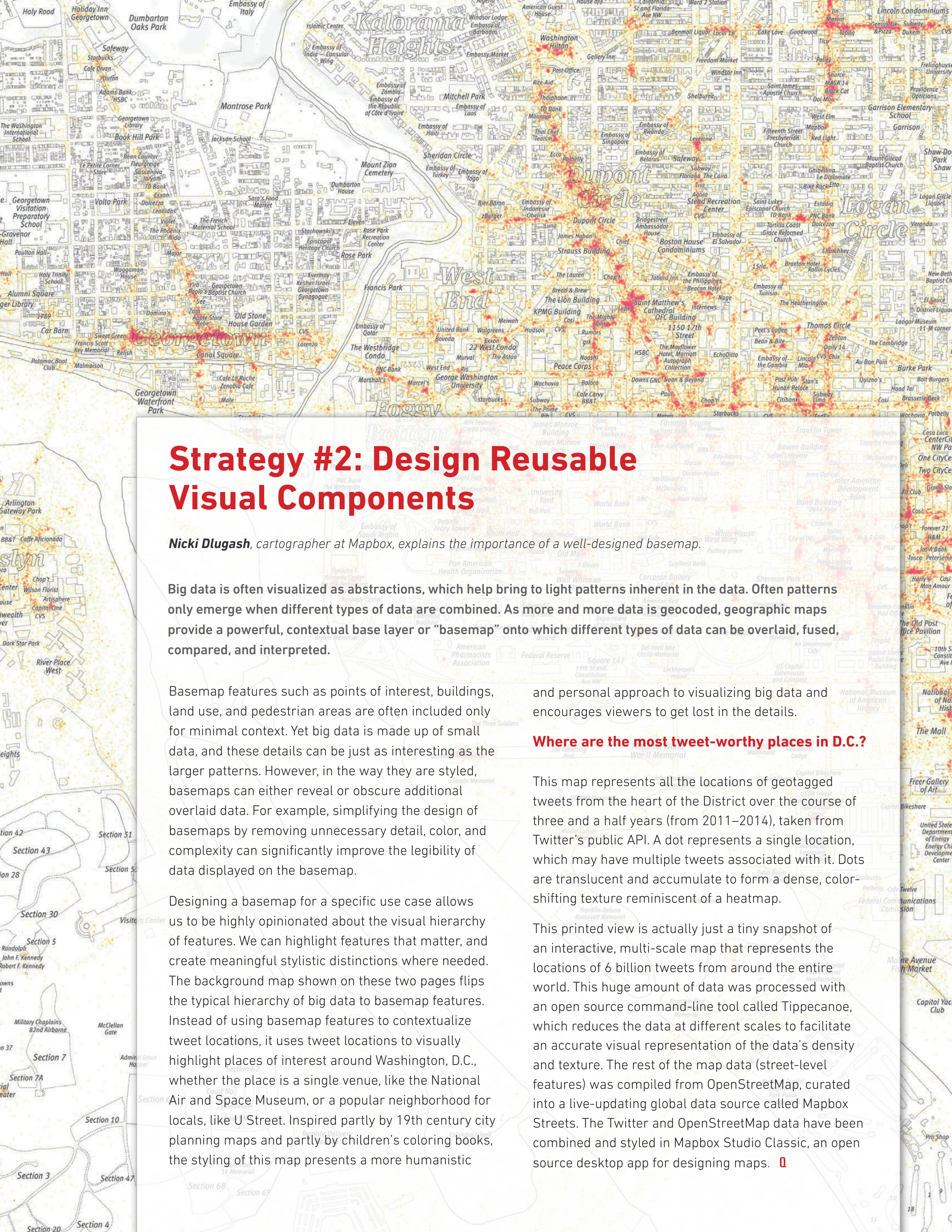

MAPD
CONNECT
OPEN
SAVE
ADD CHART
ADD FILTER
CLEAR FILTERS
SHARE
5,375,389 of 28,143,638 shown
Recipient Name
Contributor Name
amount > 0 and amount <= 5000 and recipient_type = 'P'
Point Map
Zoom to
United States
Montreal
Number Of Donations By Seat
seat
federal:president
2,303,215
federal:house
1,859,017
federal:senate
1,213,157
0.0
500k
1.0M
1.5M
2.0M
# of donations
Number Of Donations By Recipient Party
R
2,358,984
D
2,980,193
Sum Of Donation Amount Where Recipient Party = 'R' Vs. Sum Of Donation Amount Where Recipient Party = 'D' By Donation Date (Week)
Nov 09, 2006 11:59PM — Jan 11, 2013 12:26PM
Sum Donation Amount where Recipient Party = 'R'
Sum Donation Amount where Recipient Party = 'D'
Series values per week
2007
2008
2009
2010
2011
2012
2013
1990
1992
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
Donation Date
Sum Of Donation Amount By Recipient Name
Recipient Name
Barack Obama (D) 713,610,479
Mitt Romney (R) 415,064,360
John McCain (R) 192,104,731
Hillary Clinton (D) 146,811,831
Rudolph W Giuliani (R) 59,123,227
Ron Paul (R) 28,630,139
John Edwards (D) 24,247,286
Kirsten Gillibrand (D) 24,218,998
Scott Brown (R) 20,355,301
Rick Perry (R) 19,704,722
John Boehner (R) 19,035,963
Elizabeth Warren (D) 18,578,772
Chris Dodd (D) 18,036,435
Bill Richardson (D) 17,459,919
Harry Reid (D) 16,640,112
Mitch McConnell (R) 16,040,152
Barbara Boxer (D) 14,900,292
Mark Kirk (R) 14,680,619
Charles E. Schumer (D) 14,034,021
John Cornyn (R) 13,909,293
Fred Thompson (R) 13,428,607
Sherrod Brown (D) 13,336,375
0.0
100M
200M
300M
400M
500M
600M
700M
Sum Donation Amount
Feedback

Figure 2 | Political donations to federal office holders from 2007-2012. Interactive versions of these visualizations can be found at mapd.com/demos.



Strategy #2: Design Reusable Visual Components

***Nicki Dlugash**, cartographer at Mapbox, explains the importance of a well-designed basemap.*

Big data is often visualized as abstractions, which help bring to light patterns inherent in the data. Often patterns only emerge when different types of data are combined. As more and more data is geocoded, geographic maps provide a powerful, contextual base layer or "basemap" onto which different types of data can be overlaid, fused, compared, and interpreted.

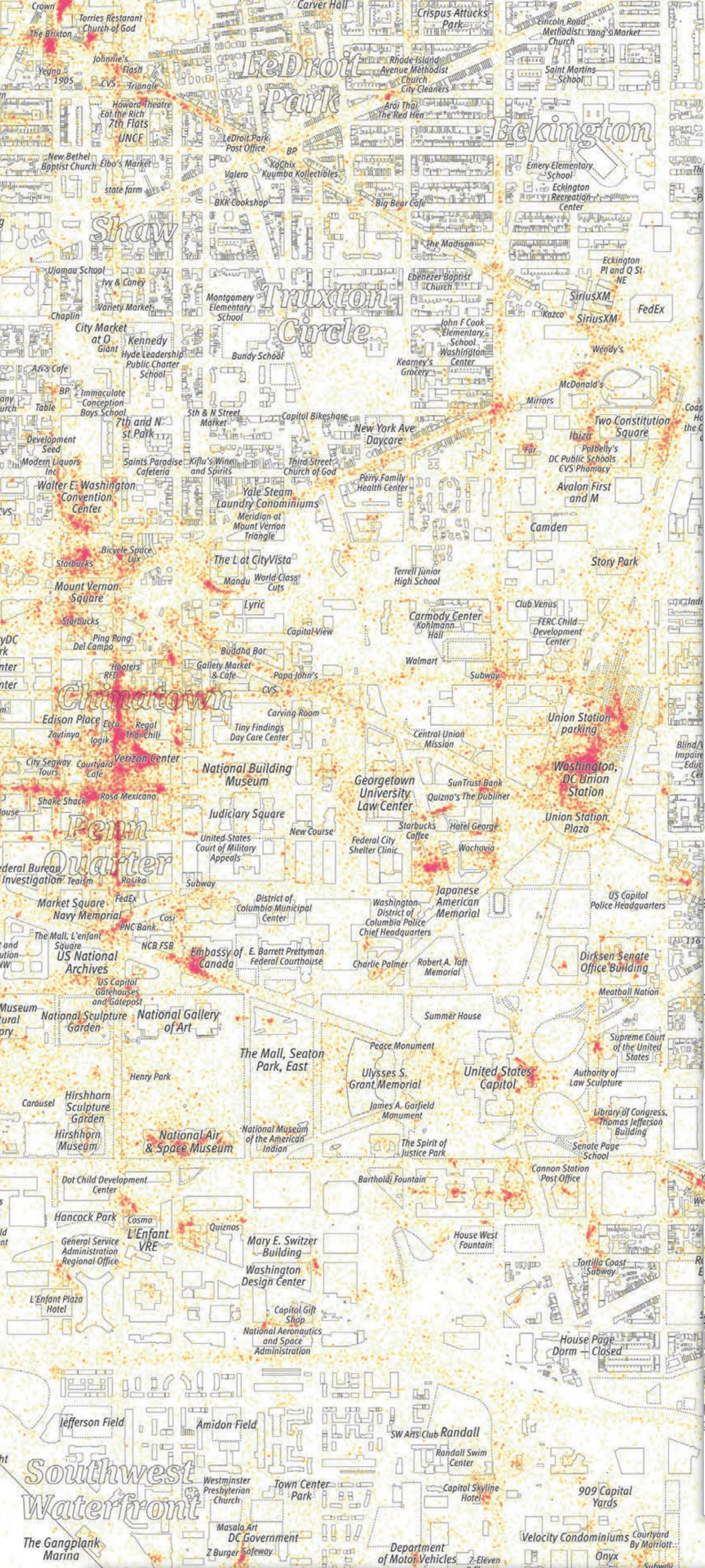
Basemap features such as points of interest, buildings, land use, and pedestrian areas are often included only for minimal context. Yet big data is made up of small data, and these details can be just as interesting as the larger patterns. However, in the way they are styled, basemaps can either reveal or obscure additional overlaid data. For example, simplifying the design of basemaps by removing unnecessary detail, color, and complexity can significantly improve the legibility of data displayed on the basemap.

Designing a basemap for a specific use case allows us to be highly opinionated about the visual hierarchy of features. We can highlight features that matter, and create meaningful stylistic distinctions where needed. The background map shown on these two pages flips the typical hierarchy of big data to basemap features. Instead of using basemap features to contextualize tweet locations, it uses tweet locations to visually highlight places of interest around Washington, D.C., whether the place is a single venue, like the National Air and Space Museum, or a popular neighborhood for locals, like U Street. Inspired partly by 19th century city planning maps and partly by children's coloring books, the styling of this map presents a more humanistic and personal approach to visualizing big data and encourages viewers to get lost in the details.

Where are the most tweet-worthy places in D.C.?

This map represents all the locations of geotagged tweets from the heart of the District over the course of three and a half years (from 2011–2014), taken from Twitter's public API. A dot represents a single location, which may have multiple tweets associated with it. Dots are translucent and accumulate to form a dense, color-shifting texture reminiscent of a heatmap.

This printed view is actually just a tiny snapshot of an interactive, multi-scale map that represents the locations of 6 billion tweets from around the entire world. This huge amount of data was processed with an open source command-line tool called Tippecanoe, which reduces the data at different scales to facilitate an accurate visual representation of the data's density and texture. The rest of the map data (street-level features) was compiled from OpenStreetMap, curated into a live-updating global data source called Mapbox Streets. The Twitter and OpenStreetMap data have been combined and styled in Mapbox Studio Classic, an open source desktop app for designing maps.



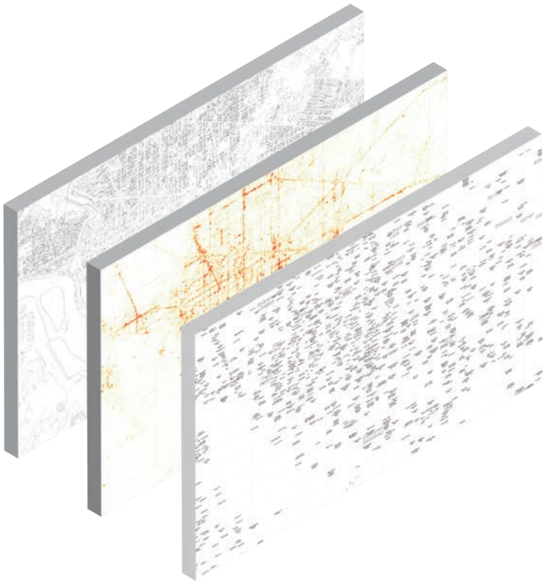


Figure 1 | There are three layers to the background map: a minimal basemap, geotagged tweet data, and feature labels.

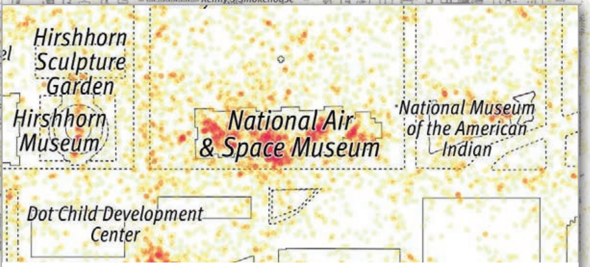

Hirshhorn Sculpture Garden
Hirshhorn Museum
National Air & Space Museum
National Museum of the American Indian
Dot Child Development Center

Figure 2 | Geotagged tweets near the National Air & Space Museum.

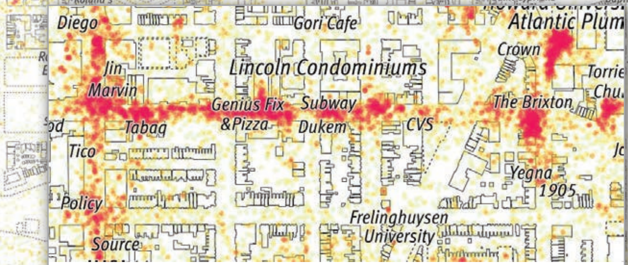
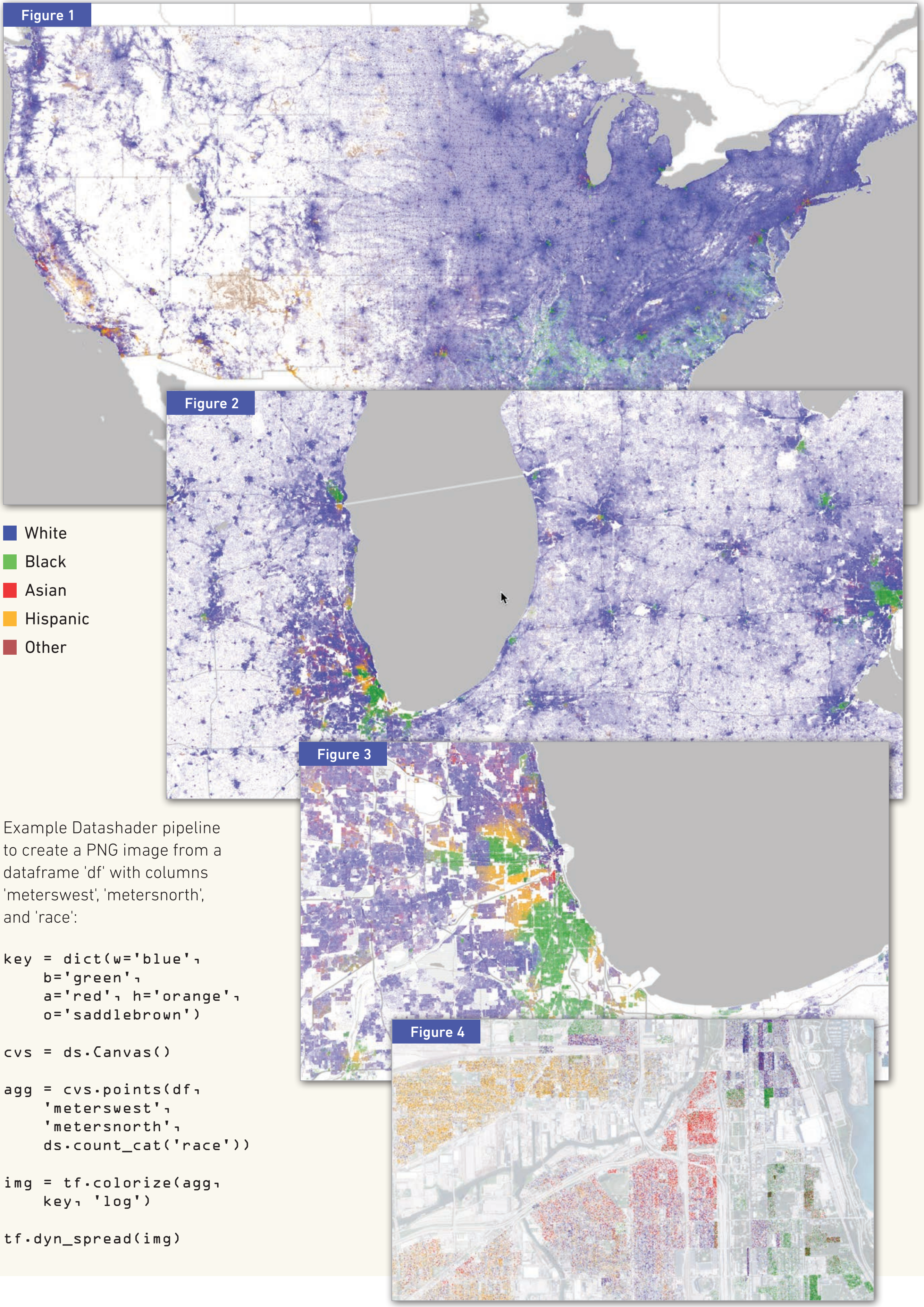


Figure 3 | Geotagged tweets near points of interest on U Street, Northwest.

***Nicki Dlugash** is a cartographer at Mapbox, where she designs basemaps and explores methods of visualizing geodata.*

Data © Mapbox © OpenStreetMap contributors and from the Twitter Streaming API. Twitter data compiled and processed by Eric Fischer. Map designed by Nicki Dlugash.



- 🟦 White
- 🟩 Black
- 🟥 Asian
- 🟨 Hispanic
- 🟫 Other

Example Datashader pipeline to create a PNG image from a dataframe 'df' with columns 'meterswest', 'metersnorth', and 'race':

```
key = dict(w='blue',
    b='green',
    a='red', h='orange',
    o='saddlebrown')

cvs = ds.Canvas()

agg = cvs.points(df,
 'meterswest',
 'metersnorth',
    ds.count_cat('race'))

img = tf.colorize(agg,
    key, 'log')

tf.dyn_spread(img)
```

Strategy #3: Fix Overplotting

Peter Wang *and* ***Hunt Sparra*** *of Continuum Analytics describe how we can do better than one point per pixel.*

One of the challenges of visualizing large datasets is that plots that work well for a few hundred data points become visually uninterpretable for millions or billions of points. Overplotting — when the number of elements to be plotted exceeds the number of pixels available — is a significant, but often overlooked problem with many visualization tools. When many points are plotted directly on top of one another, it is difficult for analysts to interpret data accurately through the visualization.

To avoid overplotting, it is necessary to down-sample, cluster, or filter data before it is visualized; however, few visualization tools provide support for sophisticated downsampling techniques. One common strategy is to change the transparency or "alpha value" of data points. However, with this approach, outliers in the dataset are often not legible. As Joseph Cottam, Andrew Lumsdaine, and Peter Wang explain in *Overplotting: Unified Solutions Under Abstract Rendering,* "The highest peaks are still over-saturated and the lowest troughs are indistinguishable from empty."[1]

Bokeh's Datashader library provides intelligent downsampling for large server-side data, allowing analysts to apply sophisticated statistical transformations and sampling techniques to the way data is displayed. Applications of this technique make it computationally and perceptually practical to analyze data at scale and offer analysts new ways to distinguish gradation, variation, error, and uncertainty within large datasets.

Datashader builds on a technique called abstract rendering, which essentially applies statistical binning techniques to the rendering process, creating a synthetic data space that sits, conceptually, between the representation of data and the colored pixels within an image and functions as an abstraction layer. When effective, this abstraction layer allows an analyst to manipulate aspects of the data "that neither the raw data, pixels, or geometric constructions do by themselves...[including] the ability to compute the full range of overlap."[2]

Figure 1 shows a scatterplot of 300 million points from the 2010 U.S. Census, with one point per person, colored by racial category. The color of each pixel is calculated not only to show category values (here, race) but also to emphasize point density. Datashader makes these calculations by breaking up the rendering pipeline into separate, well-defined stages: initial data selection; projection from data space into display space; aggregation of values at each point in display space; and transformation from aggregated data into visible pixels. Optionally, this data may be overlaid onto an existing basemap or satellite image.

With a few lines of code, the visualization can be tuned to reveal or highlight specific aspects of the data that are of interest. When zooming in (Figures 2, 3, and 4), each of these computations are re-run to reflect the very different numbers of points now falling into each zoomed pixel, automatically revealing more detail in the datasets, without manual intervention. Figure 3 shows that, like most U.S. cities, Chicago is highly racially divided by neighborhood and district.

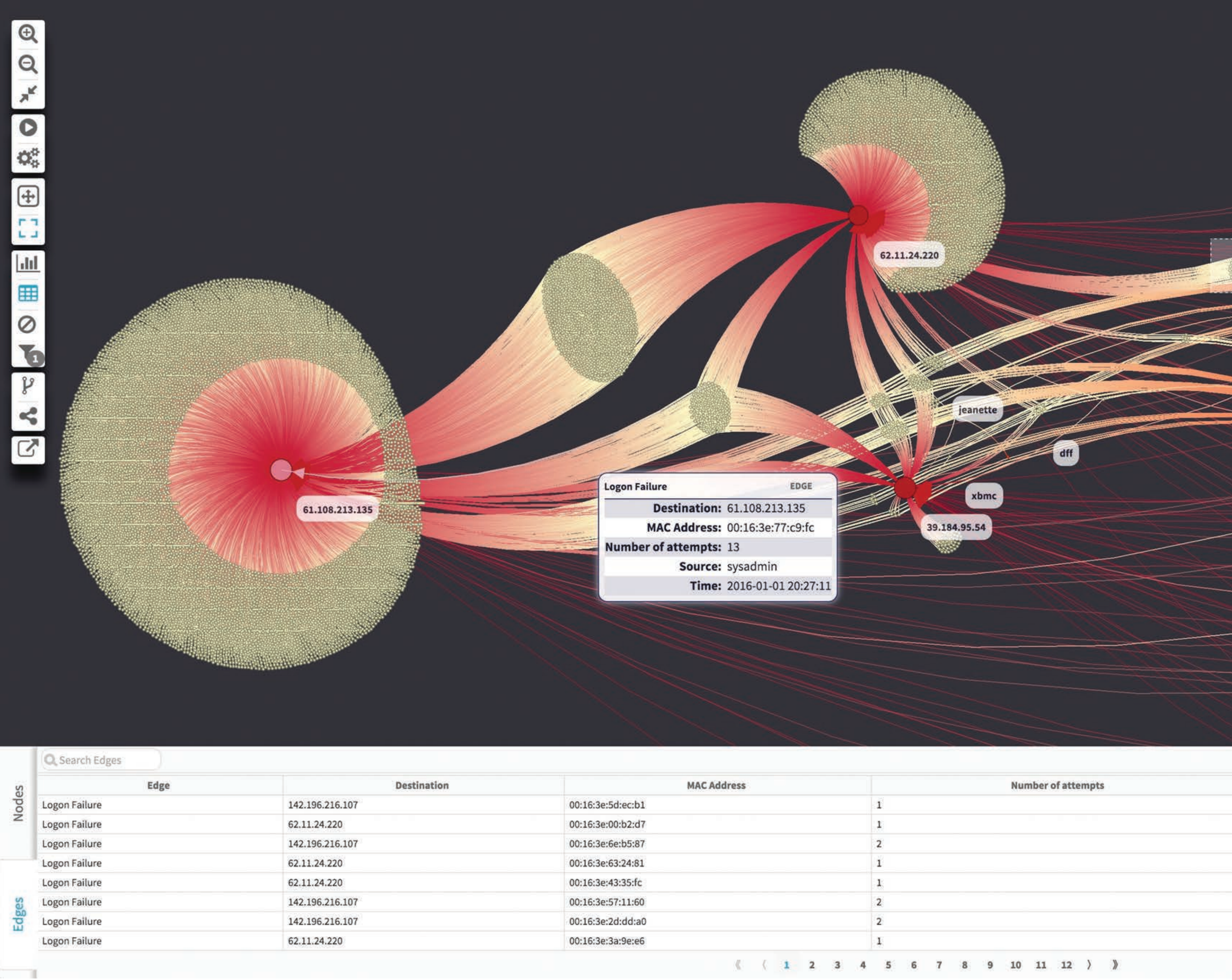
At very high zoom levels, individual data points (in this case, people) become visible, with dot sizes automatically increased to show the population density and the racial makeup of each region (such as the Asian population in Chicago's Chinatown, shown in Figure 4 in red). Other patterns become clear only when the data is related to other datasets, or augmented with additional context. For example, the abstracted Google satellite image shown in Figure 4 reveals a non-residential area that acts as a border between different racial groups.

Datashader saw its first public release in February 2016. It is a companion to the Bokeh plotting package, but can be used with any system that can process bitmap images. Q

Continuum Analytics *develops Anaconda, the leading modern open source analytics platform powered by Python. Continuum's Python-based platform and consulting services empower organizations to analyze, manage, and visualize big data — turning massive datasets into actionable insights and business value.*

[1] A. Cottam, A. Lumsdaine and P. Wang , "Overplotting: Unified solutions under abstract rendering," The First Workshop on Big Data Visualization (in conjunction with IEEE Big Data).

[2] Ibid.



Logon Failure EDGE

Destination: 61.108.213.135
MAC Address: 00:16:3e:77:c9:fc
Number of attempts: 13
Source: sysadmin
Time: 2016-01-01 20:27:11

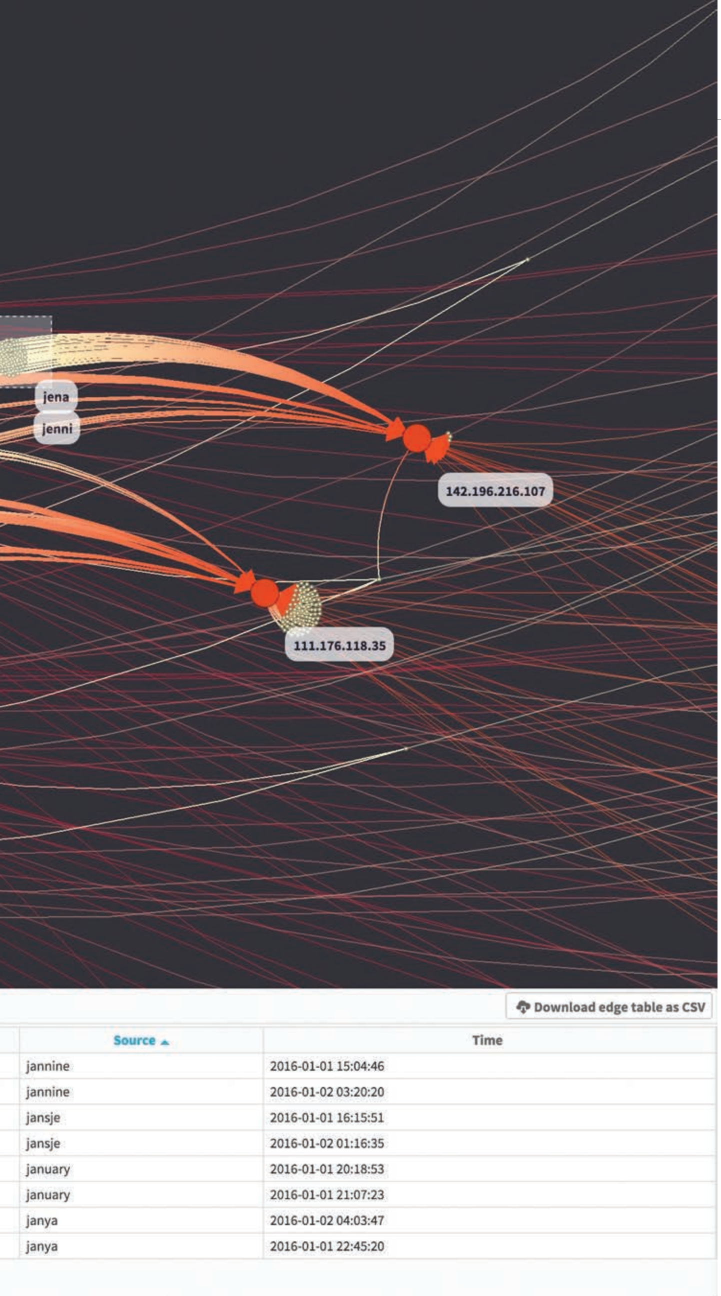
Edge	Destination	MAC Address	Number of attempts
Logon Failure	142.196.216.107	00:16:3e:5d:ec:b1	1
Logon Failure	62.11.24.220	00:16:3e:00:b2:d7	1
Logon Failure	142.196.216.107	00:16:3e:6e:b5:87	2
Logon Failure	62.11.24.220	00:16:3e:63:24:81	1
Logon Failure	62.11.24.220	00:16:3e:43:35:fc	1
Logon Failure	142.196.216.107	00:16:3e:57:11:60	2
Logon Failure	142.196.216.107	00:16:3e:2d:dd:a0	2
Logon Failure	62.11.24.220	00:16:3e:3a:9e:e6	1

Strategy #4: Display Relationships Between Data

Displaying relationships between data at scale is a challenge. ***Leo Meyerovich*** *and* ***Paden Tomasello*** *explain how Graphistry can help.*

Computing systems and networks are becoming increasingly complex. An expanding range of connected devices, a move towards the virtualization of network resources, and a growing reliance on public and private cloud infrastructure all complicate the task of understanding of network assets, topologies, and anomalies. As more users, software, and hardware plug into computer systems, incident responders need to maintain situational awareness.

A vast range of companies are developing new ways to monitor and extract data from network infrastructure, or to leverage advanced data science and machine learning techniques to provide new insights into the operation, use, and potential misuse of these networks. However, the power of these tools will ultimately depend not only on the quality and the extent of the raw data collected, but also analysts' ability to interpret, make sense of, and act on insights derived from that data.



Source ▲	Time
jannine	2016-01-01 15:04:46
jannine	2016-01-02 03:20:20
jansje	2016-01-01 16:15:51
jansje	2016-01-02 01:16:35
january	2016-01-01 20:18:53
january	2016-01-01 21:07:23
janya	2016-01-02 04:03:47
janya	2016-01-01 22:45:20

Figure 1 | Clustering failed logins immediately highlighted two "slow" distributed dictionary attacks against users whose account name starts with "j."

Visualization can help provide insight into networks and network data, but today's network and graph visualization tools were built for smaller scale data. With current capabilities, security analysts are left working blind — faced with tasks like correlating billions of alerts or combing through months of transaction data, analysts lack the tools to display and navigate relationships between data at scale.

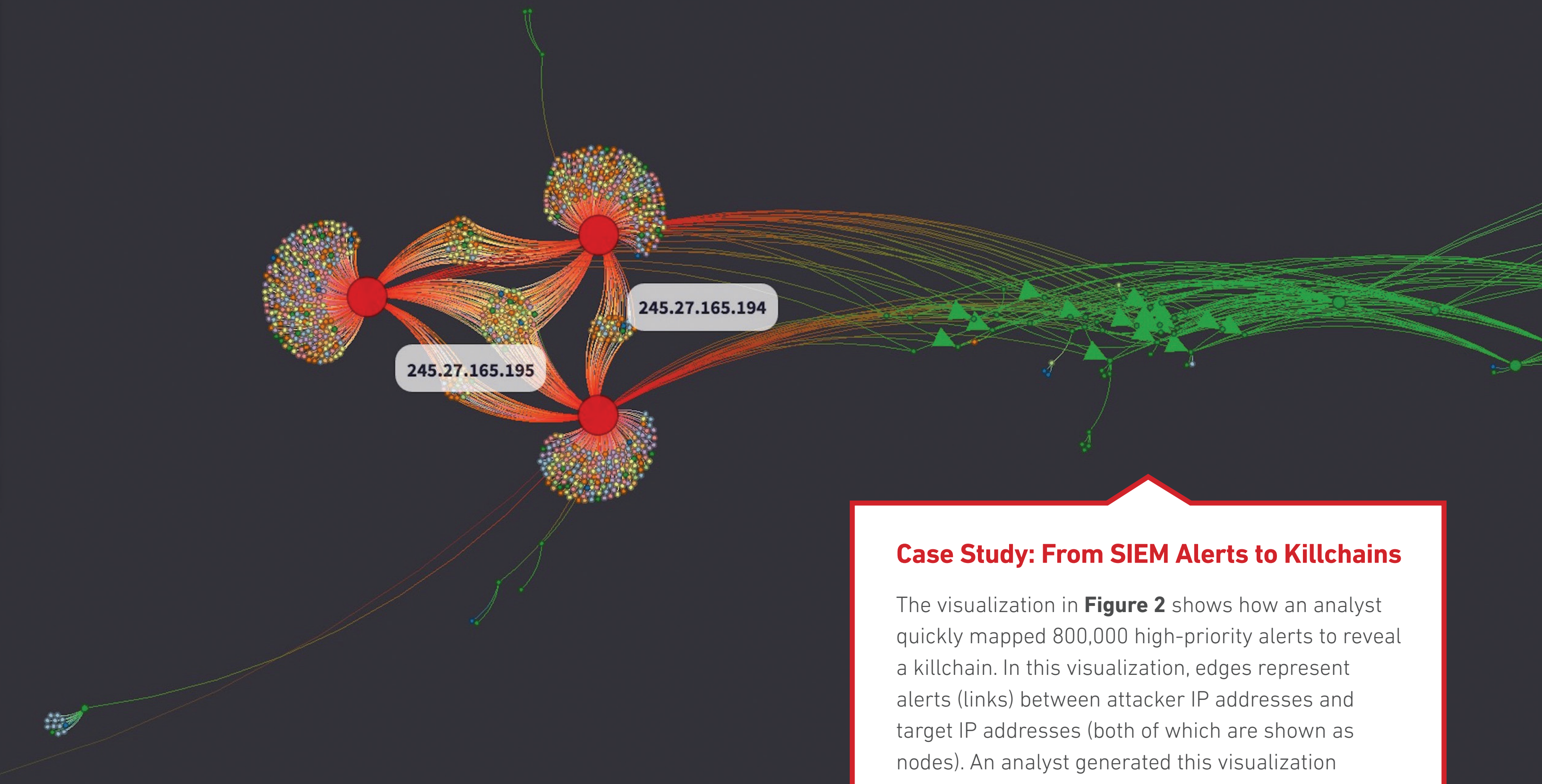
The power of graph visualization becomes clear when contrasted with traditional diagrams. For example, a bar chart is great at summarizing data, but it does so by hiding individual entities and how they relate to each other. In contrast, graphs show connections between individual entities. For example, whereas a bar chart would enable an analyst to see the number of ICMP Pings within a particular time window, a graph visualization can reveal that the ICMP Pings co-occurred with IP spoofing and failed logins across similar hostnames. A good graph visualization tool will assist an analyst in finding underlying patterns and statistically meaningful correlations. Through responsiveness and interactivity, these tools can permit analysts to drill into particularly interesting or suspicious areas, to refine their queries, and to iterate their search.

Graphistry is bringing the intelligence and scalability of supercomputing to network and graph visualization. By using GPU computation — which lets us execute thousands of commands in parallel — we are able to use techniques that were previously too computationally expensive. For the first time, analysts can see every device and alert in their system. Graphistry's early users have already tracked lateral movement, hunted for data exfiltrations, spotted command-and-control beaconing, and caught misbehaving firewalls. Our combined server-client GPU architecture promises to unlock a new generation of visual analytics for security by providing access to scalable visualization techniques.

Visually Exploring Events for Linked Attacks

Figure 1 provides an example of how Graphistry improved workflows around auditing failed logins at a Fortune 500 company. The company's incident responders were using ArcSight to collect over 100 millions alerts a day from various security vendors. They could already search billions of events for failed login attempts; however, they were unable to understand the returned result of over 1 million alerts. Graphistry allowed the incident responders to load and display this entire dataset, visually correlate the alerts, and further explore and drill down without coding.

In this visualization, each failed login alert is represented as an edge that "links" an alert's username (white nodes) to its IP address (red or orange nodes). Many alerts share the same username or IP, so the result is a giant connected graph, which our tool automatically interprets in order to cluster the alerts. The static visualization shown here is only a single snapshot of a fully interactive tool that provides a range of filtering options through the user interface. For example, an analyst could search for brute force attacks

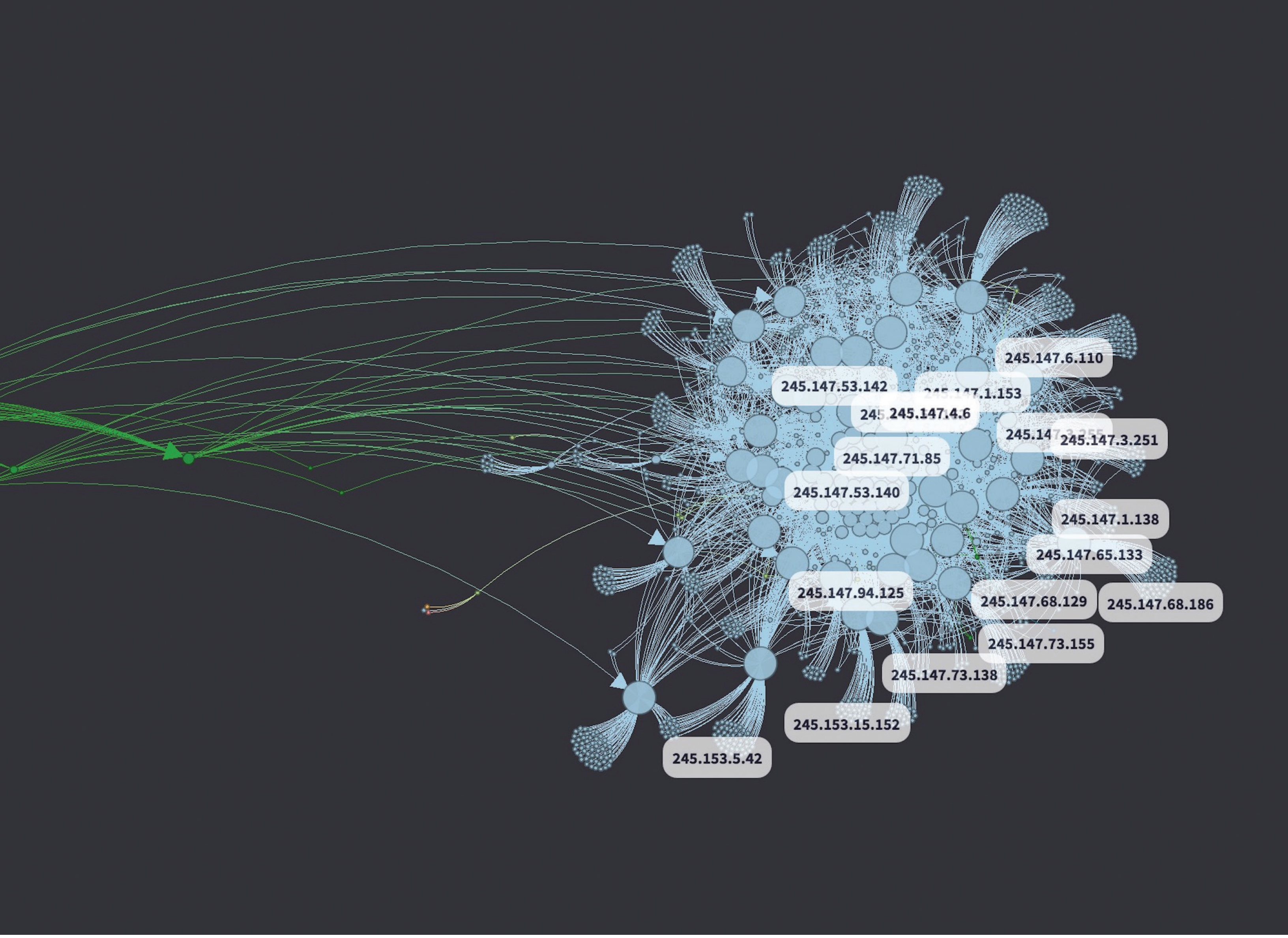


— where an attacker tries to log into a single machine in several different ways — by filtering for edges with high repetition counts (representing the multiple attempts). Or, to search for more sophisticated attackers, the analyst could filter for username nodes with many outgoing edges (representing one attacker, but several attacks). As we say internally, “a slider is worth a thousand queries.”

Case Study: From SIEM Alerts to Killchains

The visualization in **Figure 2** shows how an analyst quickly mapped 800,000 high-priority alerts to reveal a killchain. In this visualization, edges represent alerts (links) between attacker IP addresses and target IP addresses (both of which are shown as nodes). An analyst generated this visualization with only two queries: the first query filtered for all high-priority alerts; the second query added all low-priority alerts that involved the suspicious IPs identified by the first query. Graphistry automated the visualization tasks such as clustering behaviorally similar computers together and giving them meaningful colors and sizes. The output is a clear and interactive view of a killchain:

1. Insider portscan (identification) by three red attackers
2. Two steps of logins across systems (lateral movement) as green edges
3. Data transfers (data exfiltration) as blue nodes and edges
4. From there, the analyst used the interactive controls to drill into individual machines and alerts, and thereby determine follow-on investigations and remediations



__Leo Meyerovich__ co-founded Graphistry, Inc. in 2014 to scale visual graph analysis: the platform enables analysts to explore billions of security alerts by connecting browsers to GPU clusters. Graphistry builds upon the founding team's work at UC Berkeley on the first parallel web browser and the Superconductor language for declarative GPU-accelerated data visualization. Some of Meyerovich's most referenced research is in language-based security and policy verification. His broader programming language design efforts received awards for the first reactive web language (OOPSLA, NSF GRFP), automatic parallelization and parallelizing the web browser (PLDI, Qualcomm Innovation Fellow), and sociological foundations (OOPSLA, SIGPLAN).

__Paden Tomasello__ is a platform engineer at Graphistry, where he splits his time between developing new security analyses for hunting through hundreds of millions of daily alerts, and optimizing GPU kernels to hit 100ms response times. Tomasello graduated from UC Berkeley, where he first became passionate for GPU computing. He also enjoys programming competitions and has been awarded achievements in topics such as optimizing matrix multiplication and machine learning using neural nets and decision trees.

Strategy #5: Collaborate

Analyzing data at scale requires multiple people with different kinds of expertise. ***Plotly's*** *platform for creating charts and dashboards facilitates collaboration.*

Data scientists, analysts, and developers use a variety of web-based open source tools, libraries, and languages in order to create visualizations. Toolkits provide varying capabilities, and users with different types of expertise often prefer to use different programming languages. For example, many front-end visualization libraries require familiarity with JavaScript, but data scientists, engineers, statisticians, and computer scientists often have more experience with scientific computing tools such as Python, Matlab, and R. Bridging the gap between these toolkits can be a significant obstacle to collaboration. Additionally, with current workflows it is difficult to share and distribute live, interactive versions of visualizations. In many enterprises, it is common to embed static screenshots of visualizations in PowerPoint and send them as email attachments, but sending someone a static image of a visualization does not allow him or her to edit the file or to understand how it was generated. This is a serious barrier to data science collaboration.

Plotly believes that two key features — cross-language compatibility and native social functionality — can fix what is currently a fractured data visualization workflow. Through Plotly, users can create visualizations in their preferred computing language (Python, R, Matlab, JavaScript, etc.) or by uploading data and selecting a chart type through a point-and-click user interface. Interactive visualizations — along with their underlying source code and source data — can then be shared, published, and distributed through Plotly's platform, where other users can view, comment on, and edit "forked" versions of those visualizations in the programming language of their choice (even if it is different from the language in which the visualization was created). Plotly's 100 percent web-based collaboration platform also introduces version control into visualization workflows, providing viewers the means to verify the data associated with a visualization. This verification is not possible when visualizations are distributed in PowerPoint files or attached to emails. We like to think of it as a collaborative platform for heterogeneous data environments, or, more simply, a tool that lets you work with other people to visualize data, together.

The images in Figures 1-8 were created by Plotly users around the globe. They highlight just a few of the many types of visualizations users are making and sharing with Plotly. Clockwise, from top left:

Figure 1 | Daily temperature visualized by season from 1997-2001.

Figure 2 | This plot was created from a MySQL database and tracks life expectancy vs GNP. Bubbles are sized for country population and the colors show the continent.

Figure 3 | Bubble chart showing master painter color choices throughout history.

Figure 4 | Capital bikeshare distribution of trip duration by day (2012).

Figure 5 | Diamond prices by carat size.

Figure 6 | Interactive chord diagram showing the flow of visitors between cities.

Figure 7 | Communities in a Facebook network, defined as an igraph Graph Data.

Figure 8 | Randal Olson noticed that most of his classes had a paltry gender ratio, and used Plotly to chart the percentage of bachelor's degrees conferred to women. He uncovered the fact that, particularly compared to other departments, there are proportionally few women in engineering and technological fields. While the tech gender gap is widely covered, there's value to visualizing it — the gender ratio in Computer Science is back down to what it was in the 1970s.

Plotly *offers a collaborative, web-based platform for analyzing and visualizing data. It provides a standard interface to a variety of graphing and statistical tools such as Python, R, Matlab, Excel, and Julia.*

Figure 1

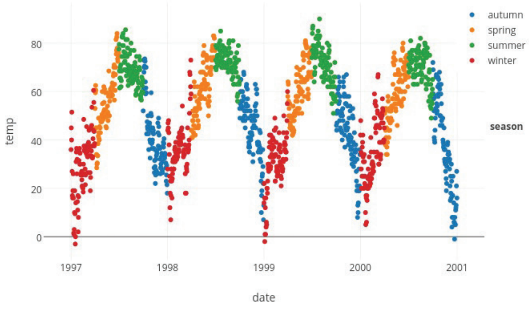


Figure 2

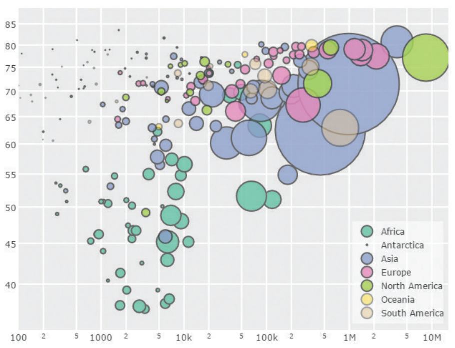


Figure 3

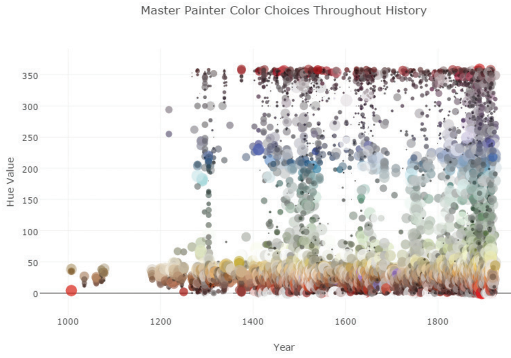


Figure 8

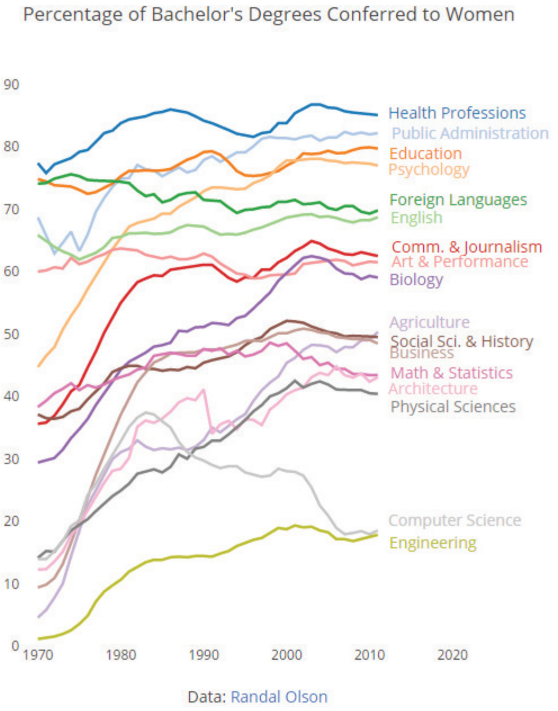


Figure 4

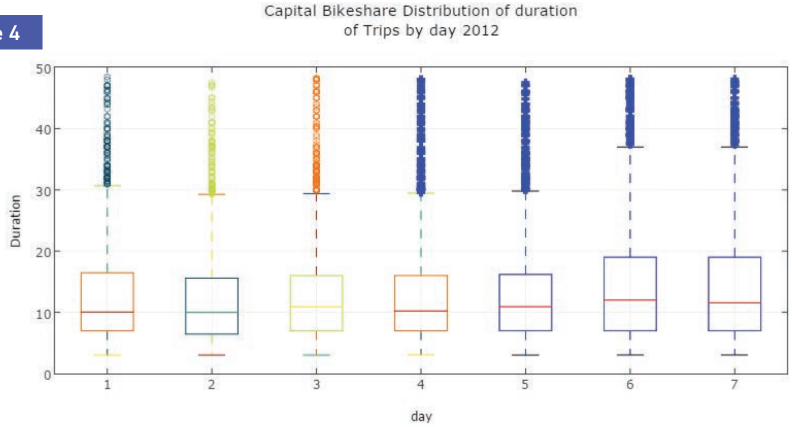


Figure 5

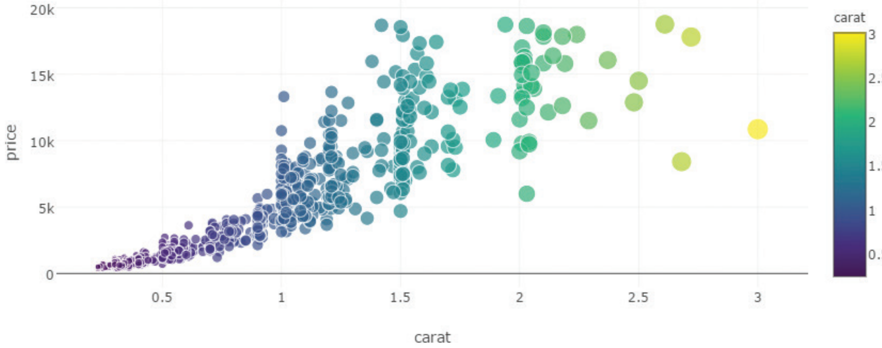


Figure 7

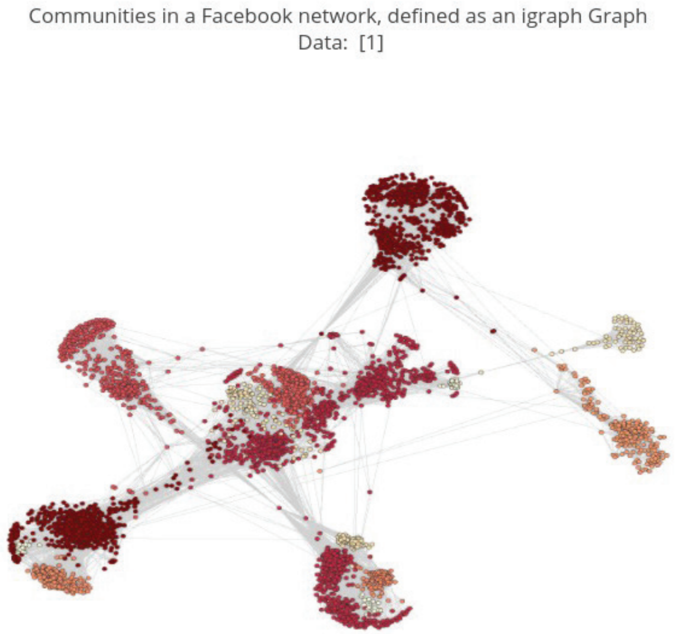
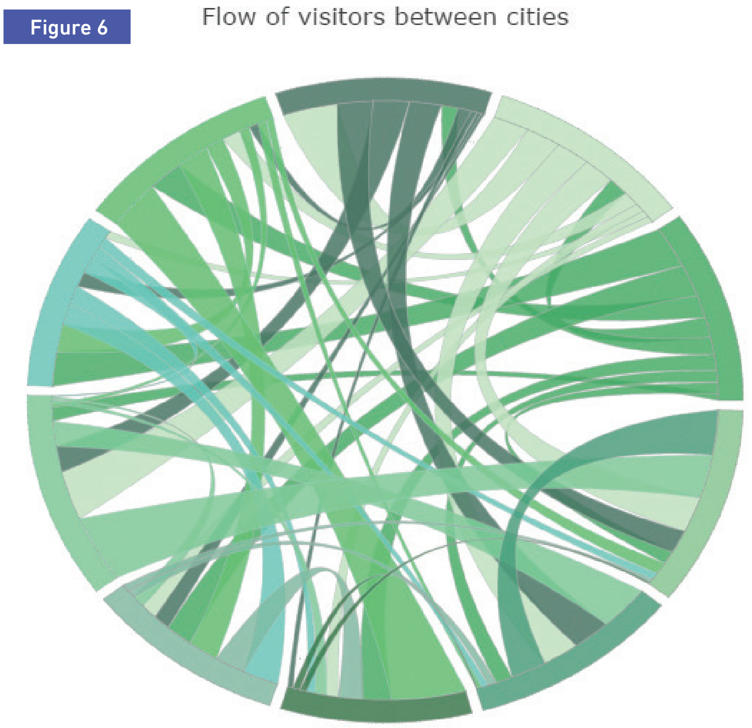


Figure 6



IQT®

IN·Q·TEL