Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Seminar Ausarbeitung

Truong Vinh Phan

Conveying Knowledge from Big Data with Visualization and Storytelling

Fakultät Technik und Informatik
Department Informatik

Faculty of Engineering and Computer Science
Department of Computer Science

Contents

1 Idea for the Master Thesis 1

2 Motivation 1

3 Approach & Risk Assessment 1

3.1 The 4x4 model for knowledge content 1

3.2 Set a theme for the story 2

3.3 Channel the audience 2

3.4 Data sourcing, cleaning and exploring 4

3.5 Define the narrative for storytelling 6

3.6 Experiment with visual designs/elements 6

3.7 Select the right technology for implementation 7

3.8 Overview of available technologies 8

3.9 Share, study and assess results 9

4 Conclusion 10

References 11

1 Idea for the Master Thesis

The idea for the master thesis is to explore how well and to what extent knowledge from a potentially large and complex data set can be conveyed to different audiences solely by means of visualization and storytelling. The result will help measure the importance of the role that data visualization and storytelling plays in solving big data problems.

2 Motivation

The most important reasons to get visualization involved in big data analysis had been discussed in AW1 and AW2. To quickly recap, while computers excel at some tasks, like spotting certain patterns, calculating predictive models and doing data mining, there are certain things human can still be far more effective, such as identifying visual patterns and anomalies, seeing patterns across multiple variables and groups, and most importantly, interpreting content of images.

Storytelling has been used since the beginning of history, in various ways and types, to convey information, events and knowledge. In the age of data, it is visualization that is used extensively as a storytelling medium in data journalism [10]. Combined with communication and good design for scanners, they form the core solution for the information overload problem. But, as shown in a study done at Northwestern University [12], while we sort of feel overwhelmed at times, we are actually generally empowered by the information we have available to us online and, at the same time, also questioning the accuracy of this information. The 4x4 model for knowledge content by Bill Shander is one of the approaches to this problem, of which good data visualization is a key part [14].

3 Approach & Risk Assessment

3.1 The 4x4 model for knowledge content

The 4x4 model consists of four key models and four components, in which content should be created, as shown in Fig.1. The first level, called the Water Cooler, describes content that is short, succinct and direct, usually represents ideas. In online content, they can be a headline or a tweet. This kind of content serves the purpose of engaging the audience and to grab attention. The Cafe level describes content that is a bit more longer and in-depth, such as a blog post or a short video . It is a progression from the first level that further explains the ideas and is the most difficult content to create, because in order to have good relatability and thus further engaging the audience, it must tell a compelling story. The Research Library

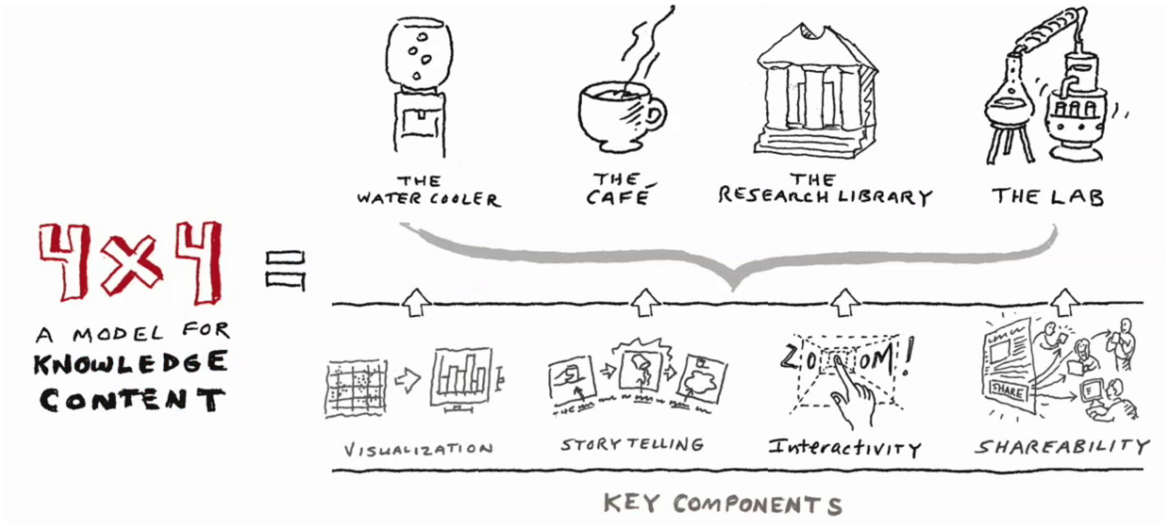


Figure 1: The 4x4 model for winning knowledge content

level describes content that will give the audience a true in-depth knowledge of the topic or idea presented in the first two levels. This content is often in the form of research papers and data that back up what is asserted in the first two levels. The last level of content, the Lab, empowers the audience to interact with the Research Library level content. This level of content is not as common as the others, but is the most powerful. The audience can adjust and filter the already massaged and analyzed data to their interests.

The four components are visualization, storytelling, interactivity and shareability, which help turn the abstract into something relatable, engaging the right audience at the right content level based on their needs.

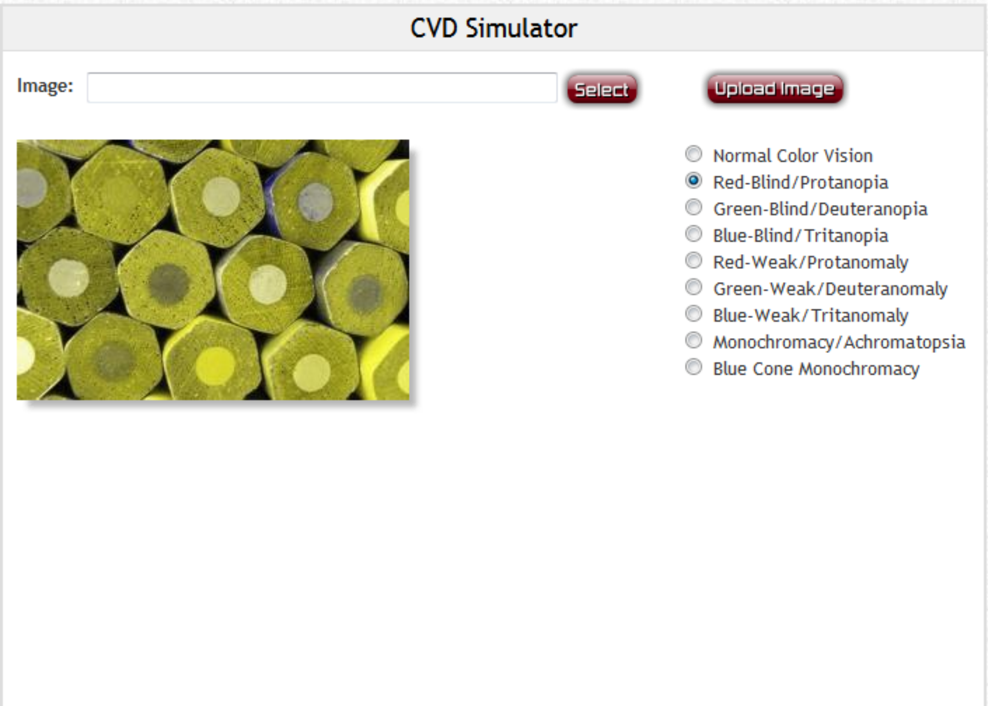
3.2 Set a theme for the story

The first level of content, the Water Cooler, is what people use to present ideas and concepts. In content online, it could be a news headline, a tweet or a thirty-second video. This can be used to get the first inspiration for the story. Theme can varies from sports to politics, to healthcare and finance, or to climate and education, which will be filtered based on interest and domain knowledge of the content author. The main question in this step is “What topic will story have?”.

3.3 Channel the audience

As in any communications, whether it is a website, a video or a blog post, visualization also requires knowing the audience to adjust the presentation to them. Among other things, the following key points outline the most important aspects that we need to understand about our audience.

- Culture: has an effect on language, perspective, context, etc. Color is the first important factor that culture has a big effect on. For instance, the color of a wedding dress may vary drastically among different cultures. And what looks odd in one culture might be completely normal in another one. Another important factor that varies among cultures is the narrative context. For example, a visualization about hockey statistics for people in northern countries, with whom hockey is a more familiar sport, certainly needs less context and is more comprehensible than if it is for people in other parts of the world. The question to be asked is “Does our audience know the underlying story of what we are talking about?”.
- Level of expertise: affects the amount of context, the type of language used, etc. If the audience are mostly experts in an industry, the language of the story certainly contains more lingo and less context. Otherwise, more background information needs to be provided and the story might be shallower with less detail.
- Consumption context/channel: The environment, in which the story is going to be published, also affects the approach to visualization. If it is a more serious environment, a higher standard of excellence will be required, with more statistical integrity for example. Otherwise, a less detail-oriented approach and lower journalistic standards with, for example, fewer decimal places in numbers, might be acceptable.
- Accessibility: The required level of accessibility is also an important factor that affects color, contrast, font size, among other things. The visualization project is going to target mainly sighted people, so it might not need much effort to ensure a good level of accessibility. However, color-blindness (more accurately, color vision deficiency) is an important issue that needs to be addressed. Roughly 8% of men and 0.5% of women have color vision deficiency [9], with the most common form being Deuteranopia/Deuteranomaly (red-green color blindness) [11]. There are tools to help address this problem, for example, the Color Blindness Simulator enables the simulation of the color perception of various color-blindness forms, as shown in Fig.2 [1].
- True believers/Skeptics among the audience is an important factor to consider. Understand the skepticism and argument of the audience against the story and the data behind it helps reduce bias, which might affect the credibility of the visualization and the story. Whether the visualization project aims to change minds, to convince people, or only to provide “facts” also has an effect on the level of interactivity and detail.
- Action: The action/reaction of the target audience needs to be planned to design better outcomes. Questions to be asked would be “Do we need our audience to perform a specific action (e.g. share on social media, answer a poll question, etc.) after seeing the visualization and learning of our story?” and “Does our design lead toward this outcome?”.



CVD Simulator

Image: Select Upload Image

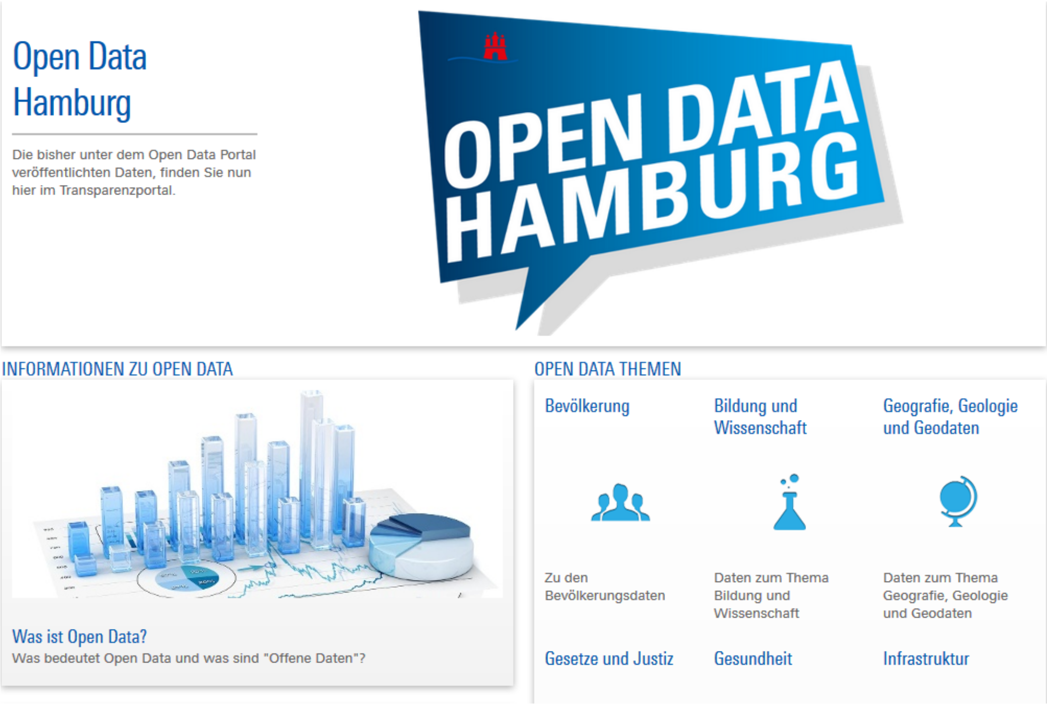
- ○ Normal Color Vision
- ● Red-Blind/Protanopia
- ○ Green-Blind/Deuteranopia
- ○ Blue-Blind/Tritanopia
- ○ Red-Weak/Protanomaly
- ○ Green-Weak/Deuteranomaly
- ○ Blue-Weak/Tritanomaly
- ○ Monochromacy/Achromatopsia
- ○ Blue Cone Monochromacy

Figure 2: Coblis - The Color Blindness Simulator

3.4 Data sourcing, cleaning and exploring

Data visualization is often the end artifact, after multiple steps - finding reliable data sources, formatting and cleaning the data, and finding the story it tells. Sourcing a large and interesting data set in the age of data is made easier by open-data movements with many resources readily available:

- Government, city-specific and political data: data portals made available by many governments and city as part of the open-data initiative (Data.gov, Socrata, Transparenzportal Hamburg, DeStatis, etc.)
- Data aggregators: house data from various sources, which help finding category-specific data easier (Programmable Web, Infochimps, Google Public data explorer, etc.)
- Social/news data: using APIs provided by social and news sites (Instagram, Foursquare, Twitter, Facebook, The New York Times, The Guardian, etc.), it is possible to access and explore data on each particular platform (news feeds, articles, etc.)
- Weather/sport data: using APIs from weather and sport sites, it is possible to access detailed weather/sport stats (temperature, wind, precipitation, etc. for weather, players, teams, coaches, leaders by season for various sports).



Open Data Hamburg

Die bisher unter dem Open Data Portal veröffentlichten Daten, finden Sie nun hier im Transparenzportal.

OPEN DATA HAMBURG
INFORMATIONEN ZU OPEN DATA
OPEN DATA THEMEN
Bevölkerung
Bildung und Wissenschaft
Geografie, Geologie und Geodaten
Zu den Bevölkerungsdaten
Daten zum Thema Bildung und Wissenschaft
Daten zum Thema Geografie, Geologie und Geodaten
Gesetze und Justiz
Gesundheit
Infrastruktur

Was ist Open Data?

Was bedeutet Open Data und was sind "Offene Daten"?

Figure 3: The Hamburg open-data portal [8]

- Research/scientific/academic data: academic work often contains interesting data, but usually requires contacting the research authors directly.

Raw data almost always contains errors, like false or missing values/characters and often not comes in the right format to be parsed and processed. As such, data adjustments, conversions and cleaning need to be done. Common data adjustments include calculating indices and ratios, aggregating and regrouping data, converting data from one format to another, such as from Spreadsheet/CSV to JSON or XML, etc. Using basic tools and spreadsheet applications (Excel, Google Refine, Data Wrangler, etc.) with advanced functions like Pivot Table, extra data parameters can be easily calculated, while many online tools and resources as well as built-in functions in most programming languages offer a quick and easy way to convert data. These steps are inevitable and part of the data exploration process, and might risk consuming a large amount of time due to data complexity. The inability to find a reliable data source and the lack of domain knowledge also contribute to the risks in this phase as well.

Understanding the data is the next vital part of the process, which helps reduce errors and increase accuracy. Basic mathematics and statistics knowledge is required for this step to calculate important parameters, such as mean, median, actual/rank indices, percentile, etc. Meta data regarding the data set also needs to be investigated. Sample size and the methodology, with which the data was generated, are valuable in evaluating the quality and reliability of the data. Being able to establish accurate relationships (correlation/causation) between data points and sum the data in a few main ideas/headlines help avoid making false claims and delivering false knowledge to the audience. Failing to build a good understanding of the data will risk succumbing to bias and delivering false knowledge.

Part of the process to understand the data also includes exploring the data visually. Using software packages and basic tools, for example: spreadsheet applications, Tableau, R, Gephi, etc. for statistical and network data analysis, or MapBox or CartoDB for mapping, quick visualizations can be made which would help deliver good insights through different visual forms.

3.5 Define the narrative for storytelling

Interactive visualizations are not necessarily consumed in a linear way, and thus should not control how the audience processes the information. Instead, a story will be structured in a narrative way, with a narrative process. The goal is to encourage but not forcing the audience to walk through the information in a linear, progressive way while exploring the data at the same time using sorting and filtering mechanisms. The basic structure of the story will include a beginning (headlines, introduction), a middle (call-outs, main ideas/theses, data, details) and an end (conclusion, data sources, follow-ups). Throughout the course of the story structure, imagery and metaphors will be used to increase relatability to complex data facts, thus giving the audience deeper impressions and better comprehension.

3.6 Experiment with visual designs/elements

Before going into actual design work, it is important to first experiment with different visual designs and elements through mock-ups, i.e. with wireframes and sketches. The advantages they offer include speed, flexibility and scale, which are vital to get to ideas and iterate on things quickly without having to know how to implement them technically or their feasibility. Important design aspects and visual elements to consider include:

- Illustration and iconography: used to capture attention, reinforce themes/linear storytelling structure and make content more relatable, therefore must be content relevant and theme-based. Imagery should be uniform and clear, as to not obstruct the reading of data values and content. Risks include difficulty in design/sourcing and overuse of imagery.
- Typography: is also used to capture attention, emphasize content and can change perception and understanding of the audience. Depending on the type (axes, legends, labels, infographics, call-outs, etc.), the approach will differ in term of typeface, font

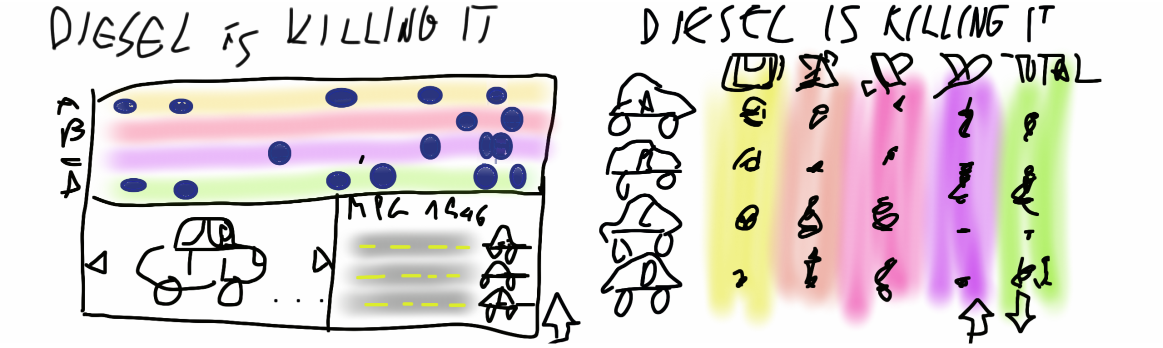


Figure 4: Wireframes and sketches

weight, etc. and should be uniform across the project. Risks include difficulty in maintaining balance between accuracy, readability, storytelling, granularity of the data and aesthetics, which might lead to false perception and understanding for the audience.

- Position, size, shape, color and contrast: are five key variables to show variance in data and create differentiation between objects. It is more about creating rules to allow data to breathe form into geometric abstractions than about designing [13]. Color and contrast are additional useful variables to enhance emphasis and highlights, but pose a challenge for visual-impaired audience. Resources such as Color Brewer or Colorblindness Simulator provide color-blind safe color palettes as well as tools to safeguard against color-blind issues [2].
- Scales: have a big impact on perception and must be chosen carefully to reflect the relationships in the data accurately. Bias could lead to choosing the wrong scales, thus delivering false impressions and knowledge to the audience.
- The right paradigm: depends on the number of variables, the type of data (hierarchical, network, geographical, etc.) and the level of aesthetics and uniqueness needed, choices must be made between various visual paradigms to represent the data. Whether it is basic graphs, charts and maps, or something creative and innovative, or a combination of those, the balance between accuracy, readability and aesthetics must be maintained, which could be difficult and thus poses an additional risk for this step.

3.7 Select the right technology for implementation

Interactive visualization requires technical implementation. There are many technologies for creating visualization with different features and benefits. The most important criteria for picking the right combination are outlined below:

- Platform vision: whether the visualization project is a short-term or long-term one affects the choice of platform. In case it is short-term and does not require reusability, the opted platform should offer simplicity and speed. Otherwise it must be scalable, modular and offer reusability and robustness.
- Audience: can be categorized into tech-savvy and general, less tech-savvy people. For modern, techno-driven audience, implementing modern technologies should not pose any challenges/problems. Otherwise, device compatibility could be an issue. For instance, Flash technology is not compatible with iOS devices, or older versions of Internet Explorer browser do not support SVG format well. In case of a broader, mixed audience, cross-browser/platform technologies can be used with fall-back mechanisms (browser/platform detection, alternatives, etc.).
- Visual/conceptual goals: the complexity of the project from a visual standpoint also plays a role in technology choice. Out-of-the-box software only offer limited features and visualization capabilities to a certain extent. Complex visual shapes and ideas require more technical and versatile platforms.

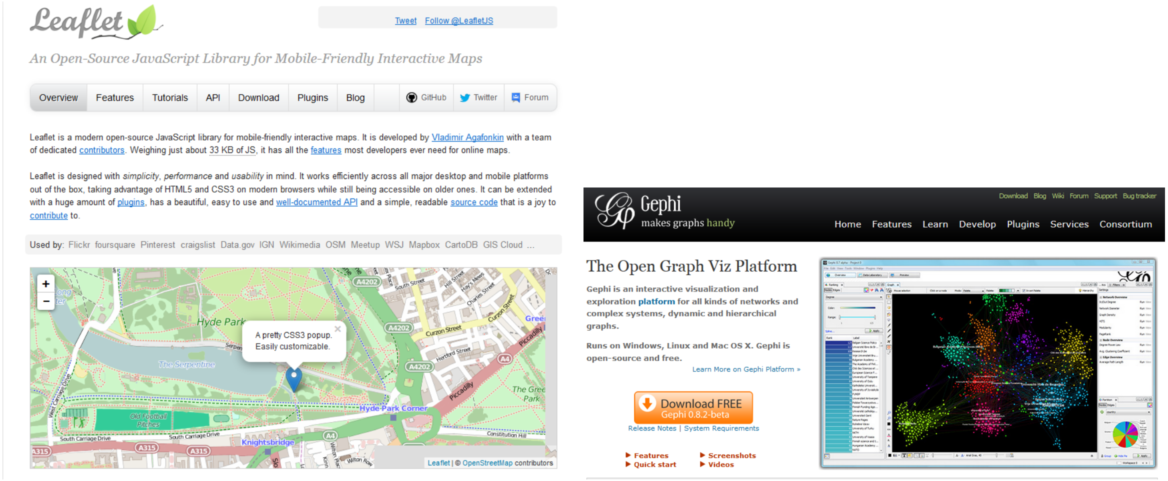
Risks at this phase include the time cost to study and the complexity of the required technologies.

3.8 Overview of available technologies

Typical software packages, off-the-shelf solutions like Tableau, Qlikview or Highcharts do not require deep technical skills and can import data and quickly create standard forms of visualization in a fairly packaged ways and also offer limited customizations. There are also open-source platforms like Gephi for creating network visualizations with many nodes and links.

Visualization libraries, on the other hand, require more technical skills but support near limitless level of customization and creativity and can be grouped into HTML5 (e.g. Chart.js, Fabric.js, p5.js, etc.) and SVG (e.g. D3.js, Raphael.js, Snap.svg, etc.) libraries, both of which have its own pros and cons.

- SVG is vector-based with each shape is an object added to the DOM. This offers great device portability without any noticeable loss of quality and automatic view update by the browser in case any object attribute is changed. HTML5 solutions rely on the new canvas object, which in turn is raster-based, thus susceptible to loss of details due to pixelation.
- In case of shape or color change, the whole scene needs to be redrawn. It is also not possible to add event-listeners to Canvas shapes, thus behavior of mouse-clicks, for example, must be associated with single pixels on the canvas rather than shapes.



An Open-Source JavaScript Library for Mobile-Friendly Interactive Maps

Leaflet is a modern open-source JavaScript library for mobile-friendly interactive maps. It is developed by Vladimir Agafonkin with a team of dedicated contributors. Weighing just about 33 KB of JS, it has all the features most developers ever need for online maps.

Leaflet is designed with *simplicity*, *performance* and *usability* in mind. It works efficiently across all major desktop and mobile platforms out of the box, taking advantage of HTML5 and CSS3 on modern browsers while still being accessible on older ones. It can be extended with a huge amount of plugins, has a beautiful, easy to use and well-documented API and a simple, readable source code that is a joy to contribute to.

Used by: Flickr foursquare Pinterest craigslist Data.gov IGN Wikimedia OSM Meetup WSJ Mapbox CartoDB GIS Cloud ...

A pretty CSS3 popup.
Easily customizable.

The Open Graph Viz Platform

Gephi is an interactive visualization and exploration platform for all kinds of networks and complex systems, dynamic and hierarchical graphs.

Runs on Windows, Linux and Mac OS X. Gephi is open-source and free.

Learn More on Gephi Platform »

Figure 5: Open-source visualization packages for maps/graphs [6, 5]

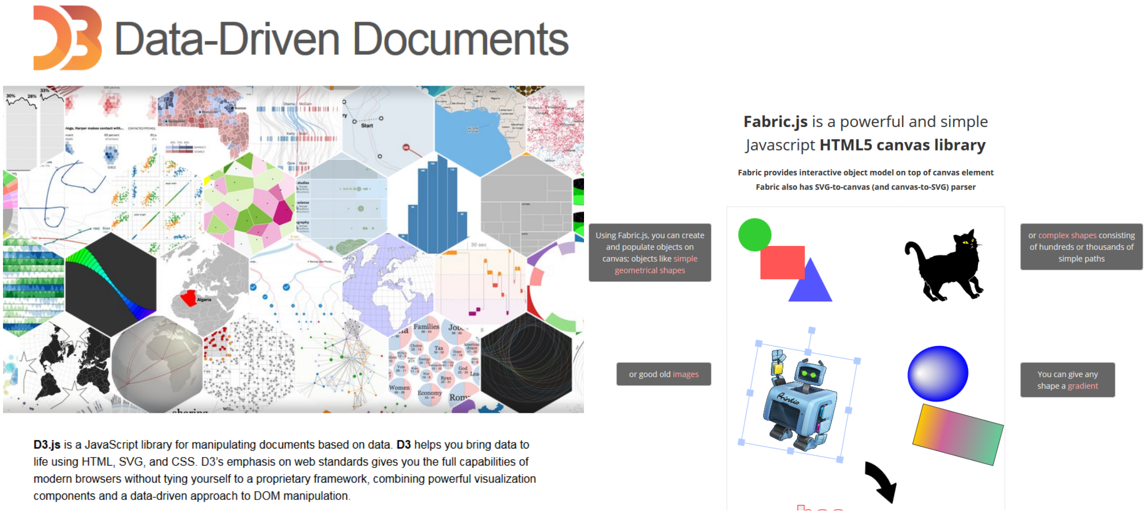
- SVG solutions also don't scale well in visualizations with many thousands of elements, in contrast to HTML5 solutions.

3.9 Share, study and assess results

The end visualization artifacts will be shared at the end of the project, accompanied by a study in the form of a small survey. The questionnaire will be limited to approximately 5 - 10 questions and sample size will be limited to 10 - 20 participants. Possible goals for the survey are to measure visualization against visualization and/or visualization against raw data in term of effectiveness, usability and overall user experience by assessing:

- How fast knowledge (facts/data attributes, etc.) can be conveyed to the audience (amount of time to answer questions regarding facts/attributes).
- How much knowledge can the audience gain within a fixed time frame (the number of questions correctly answered).
- Audience feedback (overall opinions about the data story/how much credibility would the audience rate the data story/knowledge).

The risk in this final phase of the project include the difficulty in forming meaningful survey questions, the insufficient number of survey respondents and the quality (accuracy/objectivity) of user's feedback.



Data-Driven Documents

D3.js is a JavaScript library for manipulating documents based on data. **D3** helps you bring data to life using HTML, SVG, and CSS. D3's emphasis on web standards gives you the full capabilities of modern browsers without tying yourself to a proprietary framework, combining powerful visualization components and a data-driven approach to DOM manipulation.

Fabric.js is a powerful and simple Javascript **HTML5 canvas library**

Fabric provides interactive object model on top of canvas element

Fabric also has SVG-to-canvas (and canvas-to-SVG) parser

Using Fabric.js, you can create and populate objects on canvas; objects like simple geometrical shapes

or complex shapes consisting of hundreds or thousands of simple paths

or good old images

You can give any shape a gradient

Figure 6: SVG vs HTML5-based solutions [3, 4]

4 Conclusion

The idea and approach outlined above include risks at every steps and phases of the project. Depending on the risks that actually come up, the details at each phase could be changed, but the general approach and idea would remain the same throughout the project.

References

[1] : *Coblis - Color Blindness Simulator*. – URL http://www.color-blindness.com/coblis-color-blindness-simulator/

[2] : *Color Brewer 2.0*. – URL http://colorbrewer2.org/

[3] : *D3.js - Data-Driven Documents*. – URL http://d3js.org

[4] : *Fabric.js Javascript Canvas Library*. – URL http://fabricjs.com/

[5] : *Gephi - The Open Graph Viz Platform*. – URL http://gephi.github.io/

[6] : *Leaflet.js - a Javascript library for mobile-friendly maps*. – URL http://leafletjs.com/

[7] : *The Power of Graduated Complexity in Presenting Knowledge Contentqe*. – URL https://www.linkedin.com/pulse/20141207143955-719430-the-power-of-graduated-complexity-in-presenti

[8] : *Transparenzportal Hamburg*. – URL http://transparenz.hamburg.de/open-data/

[9] : *Colorblind Population*. 4 2006. – URL http://www.color-blindness.com/2006/04/28/colorblind-population/

[10] : *Journalism in the Age of Data*. 2010. – URL http://datajournalism.stanford.edu/

[11] : *Red-Green Color Blindness*. 3 2010. – URL http://www.color-blindness.com/2010/03/16/red-green-color-blindness/

[12] HARGITTAI, Eszter ; NEUMAN, W. R. ; CURRY, Olivia: Taming the Information Tide: Perceptions of Information Overload in the American Home. In: *The Information Society* (2012), Nr. 28, S. 161–173. – URL http://www.tandfonline.com/doi/abs/10.1080/01972243.2012.669450

[13] HIDALGO, Cesar ; ALMOSSAWI, Ali: The Data-Visualization Revolution. In: *Scientific American* (2014), 3

[14] SHANDER, Bill: *The 4x4 Model for Winning Knowledge Content Online*. 4 2011. – URL http://inspiredm.com/winning-knowledge-content/

[15] TRAVERS, Michael: A visual representation for knowledge structures. In: *ACM Hypertext '89 Proceedings, Implementations and Interfaces* (1986)