Chapter 20
Designing for Small and Large Datasets

Jan Willem Tulp

David Bihanic's Interview with Jan Willem Tulp

David Bihanic (DB): As a way of introduction, could you please introduce yourself and describe your background (e.g. training, work experience, career path, etc.)?

Jan Willem Tulp (JWT): My name is Jan Willem Tulp, I live in The Hague in The Netherlands and I work as a freelance information visualizer. I have a one-man company called TULP interactive. I have a BSc in interaction design. I was very much interested in the combination of software and design, which was the main reason I studied interaction design. It turned out I was a fairly good programmer, and after my graduation I found it hard to find a job where I could both work on software and on design. Looking back at my career I have switched quite a bit between various companies. I enjoyed all of my jobs, but reasons to get a job at another company were that at more creative companies (web-design, web-marketing, etc.) I missed the technical challenge, but at more technical companies I thought that the things we created were primarily exciting at the back-end, not necessarily at the front-end. Eventually I was a Software Architect leading a group of software developers at a client, and giving workshops for colleagues about the latest technologies and agile and lean ways of writing software.

In my spare time I was often creating graphics with software like Photoshop and 3D software. And eventually I learned about the field of data visualization, there was not a single epiphany for me, more a gradual awareness. At some point I realized that this was what I wanted to do, and so I made the decision to prepare for my freelancing business. I decided to work 4 days a week instead of 5, and use this one day a week to immerse myself into data visualization, and use my savings to compensate for the loss of income.

J.W. Tulp (✉)
TULP Interactive, The Hague, The Netherlands
e-mail: janwillem@tulpinteractive.com

© Springer-Verlag London 2015

D. Bihanic (ed.), *New Challenges for Data Design*,
DOI 10.1007/978-1-4471-6596-5_20

So I started reading the books from Stephen Few and Tufte, and became more and more excited about data visualization. For me it felt like all the pieces of the puzzle came together, the field I had been looking for all these years. With data visualization there are so many aspects, graphical, technical, statistical, interaction, etc. that to me it is challenging in many ways. I started contacting freelancers like Moritz Stefaner and Ben Hosken from Flinklabs to learn about working as a data visualization freelancer. I also participated in challenges, especially the ones organized by http://www.visualizing.org (and winning one as well!), and I read more books, like the ones by Colin Ware, but also about Processing, graphic design and interaction design.

When I had 3 clients lined-up who were interested in working with me I made the decision to start officially as a freelancer. That was September 2011. And ever since I started freelancing I have been overwhelmed with projects

DB: As many great data designers (who have strong artistic and technical skills), you explore the intersections of Art-Design and Science. You aim to discover relationships that may exist between different materials and technology in order to create data visualizations that inspire and engage people (…)

Today, it appears clearly that it's a mistake to separate these two disciplines (design and technology): technology and design perfectly complement each other (…) Technology has now entered the realm of design enabling a better understanding of its emotional aspects and implications (…)

Do you think that a closer alliance formed by design and technology could help to create better experiences through data?

JWT: Creating data visualization indeed involves both artistic and technical skills. But I think that there is more.

First there is the dataset itself. The quality and the potential of the data is of crucial importance. You can be the greatest visualization designer in the world, but with low-quality data with little potential you can design what you want but there the result will never be great.

Secondly, it's also what you do with the data. The dataset itself can already be interesting, but you can also do some statistical analysis, maybe enrich or combine the dataset with other datasets to make it more interesting or to create a richer context.

And finally the visual representation and interaction design play an important role as well. Even with a great dataset and some insightful results after a statistical analysis, if you as a designer make some choices so that the visual appearance is not appealing or effective, or both, it also does not work.

DB: You explore, transform and visualize data as images (or interactive visualizations) to gain insight into phenomena—of course, the purpose of data design is insight not images (…)

How do you find insights in data?

JWT: It is true that usually it's about images, but for me there are some other factors that play an important role as well. I make different choices if I create a visual analysis tool that is of crucial importance to a small group of specialists that use a

visualization on a regular basis inside a company for instance. Effectiveness of a visualization is very important here. I make other choices if the visualization is to inform and engage with a wide audience. Also other factors might play a role as well, like the available size of the canvas, the medium (print, web, poster, iPad, etc.). I have done one project for Nielsen where I have created visualizations that were going to be part of their new corporate identity where visualizations were, even though based on real data, only used for aesthetic purposes in brochures for instance. So, for me it is a bit more nuanced when you say that it's insight, not images.

With regards to finding insights, for some client projects the insights are already known. The work I do for *Scientific American* is usually based on research that has already been completed and a paper has been written. So the conclusions and insights are already there. My job here is to create a visualization that communicates these insights in an appealing way.

For some other project I look for insights myself. This can partially be some statistical analysis I do, sometimes with tools like Tableau, or Gephi if it is a network based dataset. But quite often I write custom Python code to do some data analysis. I really like this because it also allows me to wrangle a dataset, restructure it, combine, and use any other package that is available in Python if necessary.

But another important part is also the fact that creating visualizations is an iterative process: I try to make a dataset as soon as possible so that I visually get a sense of a dataset. And then based on what I see, I start making choices, improvements, other approaches, refinements, etc. During this process I create lots of interim visualizations that I critically look at if it reveals some insight or not, and how well it reveals this insight. And based on my conclusions I make adjustments.

With interactive visualizations I create the means for the end-user to find insights by himself. In this case I have provided the required interaction possibilities that allows for filtering, clustering, changing encodings, changing configurations, etc. in order to reveal insights. The amount of interaction and analysis possibilities depend on the type of project. For general audience type projects you usually want some basic interactions, but visual analysis tools require more ways of interacting to reveal insights.

DB: Do you think that data visualization allows to gain insights that one could not get in any other way?

JWT: I do think so yes. I also do think that some insights can be communicated in different ways, but the famous example is of course the Anscombe's Quartet, where 4 small datasets have the exact same summary statistics. So, just based on this you would conclude that these 4 datasets are the same, but one you start visualizing the datasets, you see that each of them has a different shape. So, you would definitely have missed this information if you would not have visualized this dataset.

I also believe that especially complex patterns can be communicated visually very well. Take your own social network on LinkedIn or Twitter. You could describe all of your connections and how all of your connections might be linked together, but showing a network diagram, perhaps with some clusters gives you an overall sense of the network that is hard to get in another way.

I also believe it works in a very strong way if you have multiple visualizations that are linked, and an end-user can brush in one visualization and sees how that affects the other linked visualizations. That reveals some relationships that are also otherwise hard to communicate in an effective way.

DB: What is your personal method for finding new ways to visualize and make sense of data?

JWT: I am not sure if I have really found new ways to visualize or make sense of data. What I do tend to do is to constantly critically look at my own results and improve on that. And these improvements can just be combinations of other visualizations, or parts of other visualizations. You might look at The *Sociopatterns* project (Fig. 20.1)[1] I did for *Scientific American* as a radial network diagram, enclosed by a donut diagram with custom endings for the links (circles inside circles). I still don't know if this is a new way or not, but it is a custom visualization that combines different components or ways to visualize information into one representation.

I also believe very strongly that it is very important to pay attention to details, in general but also with regards to the visual appearance. I quite often spend quite some time making minor adjustments just to make sure that labels are optimally readable, elements are well positioned and have good proportions, transitions don't go too fast or too slow, picking the right colors, curves are curvy enough but not too much or too little etc. It involves a lot of tweaking of very small things, but this absolutely contributes to the overall appearance and makes the final result look much better.

DB: As a result of your data design projects, the audience gains the power of understanding data easily and quickly (…)

Do you also intend to improve public awareness on critical issues (e.g. *Urban Water Explorer and The Economic Value of Nature* projects, etc.)?

JWT: Sometimes yes, but that is mostly only a possibility when I initiate personal projects. *The Urban Water Explorer*[2] (Fig. 20.2) and *The Economic Value of Nature*[3] projects were both part of a challenge organized by http://www.visualizing.org and they provided the dataset of the partner they were working with at that time. And for most client projects there is also an intention or a goal a client has, and client projects are in many cases based on the dataset from that client.

I would like to do more projects to improve public awareness or any other way a visualization could make a contribution, but I must admit that so far my self-initiated projects were driven by other intentions, like an interesting question I had myself I wanted to answer with a visualization, or an interesting dataset I found and wanted to explore, an idea for a specific type of interaction, or just simply to try out a new technology or concept. And sometimes it is also driven by a dataset itself.

[1] Project homepage, URL, March 12, 2014: http://tulpinteractive.com/projects/sociopatterns/.

[2] Project homepage, URL, January 12, 2014: http://tulpinteractive.com/projects/urban-water-explorer/.

[3] Project homepage, URL, January 12, 2014: http://tulpinteractive.com/projects/the-economic-value-of-nature/.

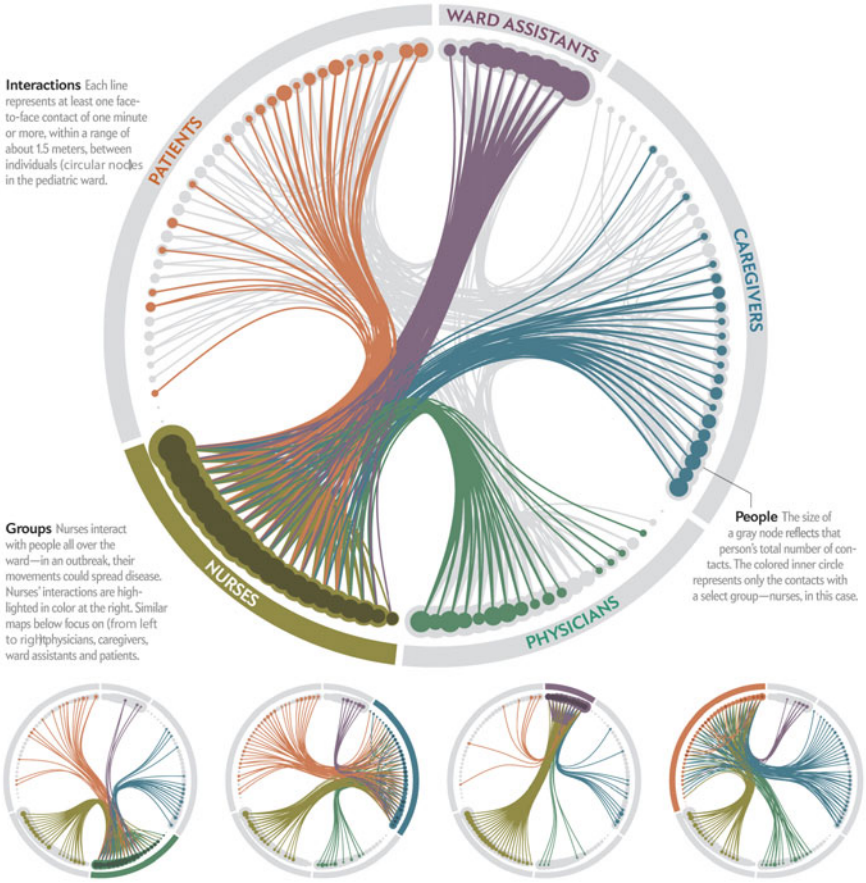

Fig. 20.1 *Sociopatterns*, Jan Willem Tulp, 2011

DB: Do you wish to communicate the value of behavior change and help inspire your audience to action? Do you consider that designers must have active citizen participation and encourage more people to also participate?

JWT: I think that would be great, yet at the same time I realize that my personal passion for data visualization itself is stronger than an ambition to encourage active citizen participation. I think it would be fantastic if that is the effect of a visualization for some people, but usually I am more driven by other motives, as mentioned in the previous question.

In The Netherlands when there are elections there are always a few websites where you can fill out a few question and get some advice on what party you should vote on (the last elections there were over 20 political parties in The Netherlands to choose from!). A few years ago before I started working as a freelancer I created a different visual interface to answer those questions and to get some other types of

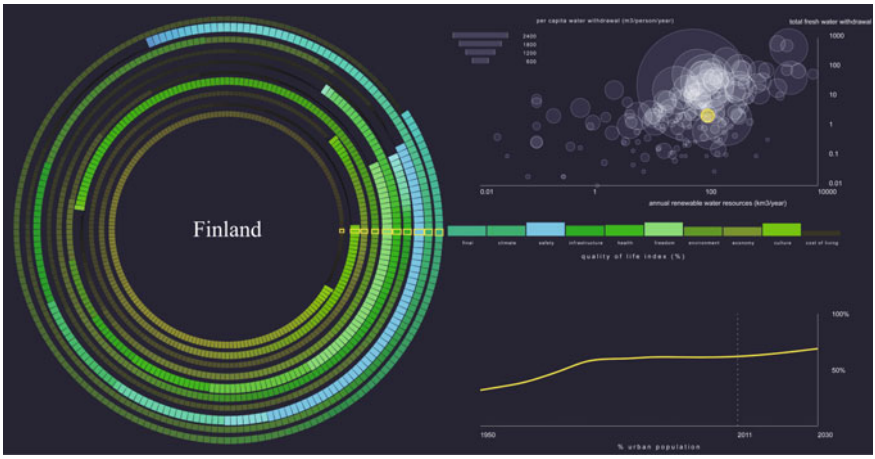

Fig. 20.2 *Urban water explorer*, Jan Willem Tulp, 2011

answers, like overlap between political parties. Also, my version gave direct feedback of your answer to a question instead of showing the final result at the end which was done on the original website. For this project I was driven by the fact that I thought that the user interface was really lacking context, and not giving you direct feedback so you were not really aware of the effect of your choice. So, I just wanted to make something that would give you better information, initially just for myself, but then publish it and see if other people could benefit from it too. And I did get some fantastic results, people were making screenshots and posting that on forums to show that the found insights they were not aware of. A friend of mine even decided to vote on a different political party than he has done so far, so he really made a drastic change in his political preference based on this tool. I think it is fantastic that this can be the effect, but I was primarily driven by the fact that I could make a better user interface, not necessarily to change as many people's mind on what to vote.

DB: Does data visualization provide the opportunity to pass knowledge to the masses?

JWT: I think it does, but there are some crucial factors here if you want to do something for the masses. It really has to be about a subject that is relevant to people at the time they look at it. And I think you can improve the effect if it an original and/or an engaging design that people enjoy playing with (and learn about the insights on the side!).

In general I think the general public might just have a quick look and then go on. I recently created a network visualization of ingredients that go well together for *Scientific American*. This visualization received a lot of attention, and was written about on quite a few websites with large number of visitors, like *FastCo Design* and *LifeHacker*. Of course, attention from these websites helps, but one differentiating

factor compared to other visualization I created for *Scientific American* is that food and ingredients speaks to almost everybody, makes a meal every now and then. It also helped that the dataset really revealed some surprising combinations of ingredients that appear to go well together. Some people said that they were really amazed by the insights, or that they were going to use the visualization to try out some new recipes. I think this is an example of a visualization that was using a dataset and showed insights that are relevant, surprising and fun to almost anybody, and that's really crucial if you want to pass knowledge to the masses.

DB: You manage to create simple visualizations that communicate complex data (…) Every designer knows that keeping visualization simple lets the data tell its own story (…)

Is it difficult to keep it simple?

JWT: I had to think about this question, what do I do to keep it simple?

One of the things that I can think of is keep asking yourself if the visualization still serves its purpose or communicates the intended insights. Quite often you can communicate many more types of insights based on the same dataset, but actually keeping the communicated message simple helps.

I also tend to use a limited number of colors, or gradients of colors. Using too many colors not only becomes harder to understand for people, it can also become overwhelming, especially if the colors don't really go very well together.

What's also really important here is that you pay attention to details to make it look simple. For instance, if elements or components are not neatly aligned, the overall appearance becomes noisy.

DB: Let's keep it simple to increase cognition (as said, in substance, Ben Shneiderman). May data visualization have a significant role in the amplification of human capabilities (perception, sensation, intelligence, memory, etc.)?

JWT: It at least amplifies cognition in the sense that you externalize information that you otherwise had to keep in your head. So in that sense a visualization truly contributes to amplification of human capabilities.

Also, representing information in a visual way, for instance properly using pre-attentive variables, allows for really quick identification of exceptional elements. This is also a much more effective and faster way than for instance reading textual information.

DB: In a recent conference, Ben Fry was referring to a collaborator's comment who said about a new data visualization in progress [I quote him]: "this data visualization proposal is too pretty, it doesn't make sense (…)" [it's a very telling comment, isn't it?]

Do you think on the contrary that the aesthetic consistency of a visualization (which determines its beauty) can improve its understanding?

JWT: Based on my own experience I think it does. Research should show if this is actually the case or not. I do know that Nick Cawthon and Andrew Vande Moere published a paper called "The Effect of Aesthetic on the Usability of Data

Visualization"[4] in which they conclude: "More specifically, the results illustrate that the most aesthetic data visualization technique also performs relatively high in metrics of effectiveness, rate of task abandonment, and latency of erroneous response. We argue that these results show that aesthetic should no longer be seen as a cost to utility."

DB: Do you think that data visualization opens new windows to see the world in new ways?

JWT: Yes, data visualization is a very good way to make the abstract and complex understandable, as it communicates information in a way that is very natural to us as human beings. We are very good at processing visual information, and are much better at pattern recognition than computers currently are.

Personally I find it one of the most rewarding moments the first time you visualize a dataset. I am always really curious what a dataset 'looks like'. This is something you cannot comprehend just by looking at the raw numbers. It is even more rewarding if it is a dataset that has never been looked at, for instance because you have collected the data from many sources and merged it into one, or you have enriched the dataset.

DB: Can it help people to better understand/interpret the world that they live in?

JWT: I think it does, for the same reasons mentioned in the previous question. However, the fact that a visualization can open a new window to see the world in new ways does not necessarily mean that it is better understood. The role of the visualization designer becomes important to guide the user though the visualization in order to understand it, depending on the prior knowledge of the user.

DB: What impact can data visualization have on our lives? Can it affect our daily lives? If so, how?

JWT: It can. I mentioned the alternative interface that helps you pick a political party before. That has really helped some people make a better informed decision. But key here is that the visualization should be relevant at the time the user looks or interacts with it. And of course, scientific visualization, for instance MRi scan visualizations have an impact for sure. And the more traditional diagrams that are often used in Business Intelligence tools also help make analysts better decisions. But both the MRi scan visualizations and Business Intelligence dashboards provide information that can be acted on directly, and is relevant at that moment, which are very important factors in order to have an impact.

DB: You constantly invent new visual metaphors, new visual figures and patterns that expose the data in creative and effective ways (…)

[4] Cawthon N, Vande Moere A (2007) The Effect of Aesthetic on the Usability of Data Visualization. In Proceedings of IV'07. Available online, URL, January 12, 2014: http://infoscape.org/publications/iv07b.pdf.

Is it essential for you to create new adapted visualization ways to new types of incoming data?

JWT: It is of course not essential to create new visual metaphors, figures and patterns. But it is part of the creativity that I enjoy very much about the way that I approach my work. In some cases traditional statistical diagram could also suffice, but the fun is to create creative approaches to represent a dataset. And that's also one of the reasons why clients approach me, because I give a unique twist to a visualization with my own style.

DB: Is it fair to say that a visualization pattern applies only to a specific set of data?

JWT: I think you can call it affordances of the data if that's what you mean. So, timestamps in a dataset would suggest a line graph, geographical coordinates would suggest a map, etc. I do believe however that it's not only about the affordances of the data that matters. In the *Close Votes* project (Fig. 20.3), I have visualized cities in The Netherlands. Size and color of each circle that represent a city changes based on the similarity of voting result with the selected city. Now, I have 2 different configurations of the visualization, one is a map of The Netherlands. This naturally allows you to reveal geographical or location based patterns. But I have also included a radial configuration. The visualization still shows cities, but not positioned on a map but radially. The reason is that for this layout I wanted to focus on similarity but on a slightly different way, and also allow the user to more easily compare cities based on number of people, which is easier to spot in this view than the geographical view.

So even if the data itself has some affordances and suggests some way of visual representation, it is also important to keep in mind what you want to show, and if that affordance is the primary type of insight you want to reveal or if it should be context only, or perhaps not even relevant to the types of patterns you want to show.

DB: You use various tools and programs to design your data design projects. Among them is the famous JavaScript library for manipulating documents based on data using HTML, SVG and CSS. You count today among one of the best experts in this solution.

Could you please tell us more about it? Could you please indicate what its main advantages are? Why choose D3.js for presenting complex data into perceivable information?

JWT: Yes, D3 is currently the most popular framework for creating data visualizations. Even though the learning curve might be a bit steep for some, it has some very strong features. One of them is that it is a framework that has very important visualization concepts, functions and utilities built into the framework, while still offering flexibility to make custom visualizations. The fact that it is web-based is also great, because it allows to run the visualization easily in the browser, and it allows you to make combinations with other Javascript frameworks (for instance underscore.js which I use a lot too) very easily. Moreover at this moment there is a

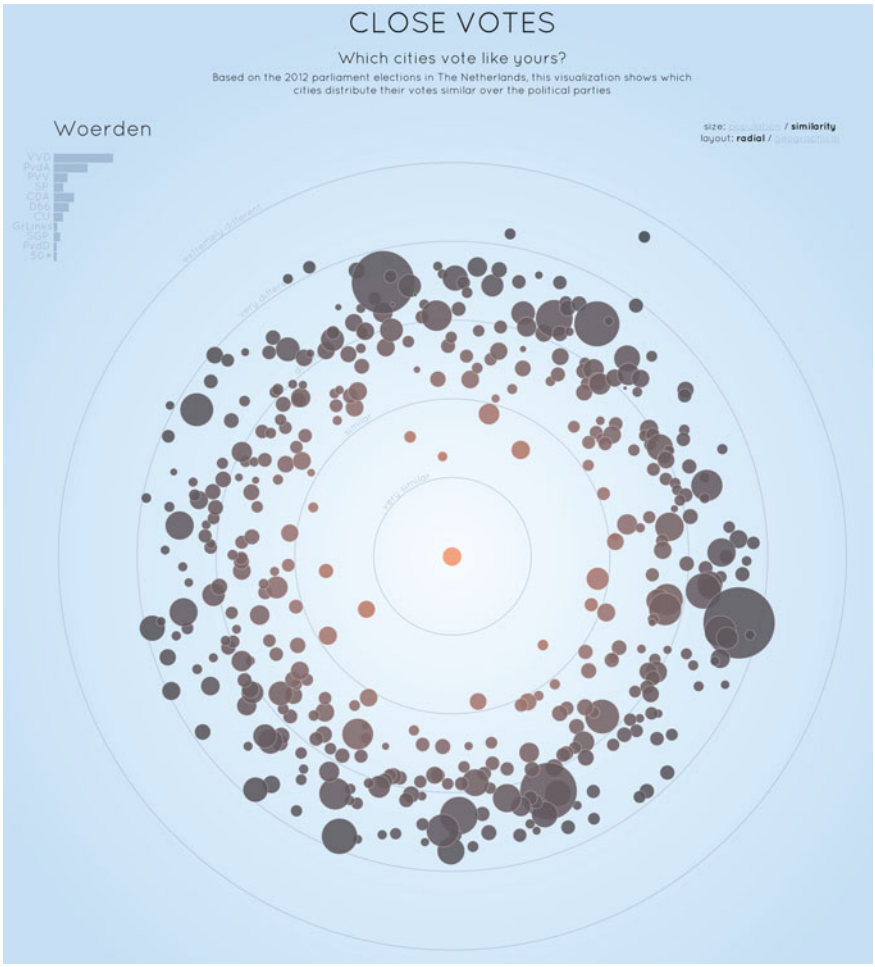

Fig. 20.3 *Close votes*, Jan Willem Tulp, 2012

very large user base for D3, there are many inspirational examples freely available, some people have written books, there are a lot of questions asked and answered on StackOverflow, etc. Finally, since D3 primarily works with SVG, you are essentially creating vector graphics. These can easily be extracted from the HTML page, and saved, so that you can open them in vector graphic tools like Illustrator to make the final tweaks.

DB: Today, you are one of the leading experts in Big Data visualization (visualization of very large amounts of data)?

What is the specific design approach and process for extremely large datasets?

JWT: I don't necessarily consider myself a Big Data visualization expert. Especially when creating a web-based visualization, you have to make sure that your data files are eventually small enough so that performance is good enough in the browser (in case of an interactive visualization). I do work with large datasets, and even though storing them in a database could sometimes be an option, my approach is usually writing custom Python code to analyse the data, but also optimize and simplify the dataset in various ways (aggregating, filtering, geometry simplification, pre-calculation, removing unneeded data, etc.).

I could say that in my case many times when working with large datasets it is all pre-processing and optimization of the data. The final dataset is usually rather small, about 5 Mb is the maximum I recently had for a web-based project.

DB: How to use visualization in order to support the understanding of very large data sets?

JWT: I think visualization for big or small data is not fundamentally different. Human perception still works the same way, regardless of the size of the data you are visualizing. I do think that there can be technical challenges, like how to display large amounts of data in a limited space, or how to keep a good performance with large data sets. But these are more technical challenges than visualization challenges. The human visual system is able to see patterns really well, and that works for both smaller datasets and larger datasets.

DB: Do you sometimes see yourself as pioneering into previously unknown data territories, facing vast areas of undiscovered data landscapes to explore (e.g. *The Data Centric Universe* project)?

JWT: Yes I do, both with regards to data has not been visualized before, and with regards to domains that I am unfamiliar with. Today almost every organization works with data, and scientists have been working with data for much longer. This means that there are a lot of potential clients for me, and my clients also belong to very diverse market segments. So, for one project I am working with data from Amsterdam Airport, the next project I am working with dataset of everything that has been discovered in the Universe so far (the data centric universe).

DB: Do you intend to uncover the secrets of Big Data?

JWT: Not necessarily. I do think that visualization can really help to get a deeper understanding of a dataset. At the same time, as mentioned before, I do think that visualizing data is not necessarily fundamentally different for large datasets or small datasets, except for the technical challenges perhaps. Visualization in general can help to communicate in find insights in complex data, and this can work very well for Big Data as well.

DB: Does visualisation uncover the big picture of Big Data?

JWT: In a way yes, as it can provide an overview of a dataset. But this is also true for smaller datasets. But as mentioned before, visualization is a good companion for Big Data (Fig. 20.4), as it helps to find and communicate insights and patterns.

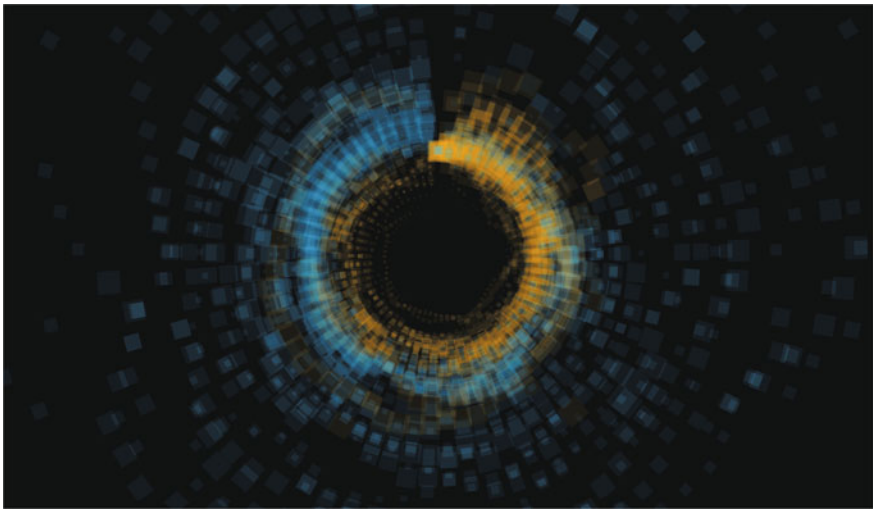

Fig. 20.4 *Nielsen identity design*, Jan Willem Tulp, 2012

Perhaps with statistical analysis you have to know a bit more what you're looking for, and visualization can have a more serendipitous effect, where you can be surprised by things you did not look for in the first place. Visualization can therefore be a nice companion in the visual analytics process where the visualization can generate some additional questions you didn't think of before.

DB: How to extract precious insights from Big Data?

JWT: In situations where you have to deal with many dimensions, you will probably have to do some statistical analysis or dimensionality reduction before visualizing. With hundreds or thousands of dimensions, you are simply overwhelmed, and you cannot just pick a few dimensions to visualize and hope for the best. After all, there are only a few visual mappings you could apply, like size, color, length, position, etc. And if you have many more dimensions than you can visually encode, then you have to prepare your dataset first, and also perhaps create some designs that allow for combinations of various dimensions. Scatter plot matrices or parallel coordinates are traditional examples of visualizations that allow for visualizing multiple dimensions at the same time, but you can also think of linking several visualizations to interactively explore a dataset. But still, statistical analysis might be very useful here.

DB: Are you specifically interested in 'data deluge' visualization (the Big Data-flow)—some, like the datajournalist Simon Rogers, called the 'Tsunami of data'?

JWT: Not necessarily. I think I am more interested in other factors that might make a dataset interesting, not necessarily the fact that it is Big Data. Personally I find it a challenge to come up with interesting and new ways to display a dataset

and its insights. Especially if there is some interactivity where a user can explore a dataset a little bit, those are fun.

DB: You create aesthetically appealing visualizations that do not distort data and distract the viewer from real information (…)

How can aesthetics qualities help to enhance understanding of data? Can beauty promote understanding in one way without undermining it in another? In a few words: is beauty (sometimes) useful?

JWT: I think it is. I think that when a visualization is aesthetically pleasing, looks original, it might speak to the emotional side as well, not just to the analytical or logical side. So, it becomes more enjoyable to experience a visualization. At the same time a visualization can become more memorable if it has a unique, and aesthetic appearance, which might be a great benefit if you want users or readers to remember the insights or stories you want to communicate based on the story. I also am convinced that if you pay attention to details and to aesthetics, that it will result in a less noisy visualization. Paying attention to alignment, balance, color, proportions, etc. will make it look nice and clean. And a clean visualization is easier and more pleasurable to comprehend than a noisy one.

DB: In your opinion, what are the big challenges for data design?

JWT: The process of creating a visualization requires you to switch between data, software and design continuously. You constantly have to verify visually if your visualization shows what it needs to show, and if it does so in a clear (and aesthetic) way. But also if there is enough context to understand the message, and if the interaction makes sense. So you're constantly checking various aspects of the visualization. And for me, working on my own, or at least specifically on the visualization part of a project, this is doable, since I do everything. But I am really curious if this is a scalable process where multiple people can work on the same visualization at the same time. Two people, both sitting behind the same screen, can possible, but when it becomes more than that, or you are not behind the same screen, it becomes difficult to continue this process of continuously evaluating your design, changing code, improving data, etc. And in many situations you cannot delegate the work of a part of the visualization to someone else (except for user interface controls or if there are multiple linked visualizations), because it is the entire visualization that needs to work as whole. So, how to scale up this process?

DB: What is your vision for the future of data design?

JWT: I think visualization will become more common. More people will be creating visualizations, and people will be more familiar with visualization. I also think that the need for visualization will increase, as Big Data, open data, quantified self, data driven journalism are all data related trends that could benefit from visualizations. You can see this already today, but it's still in its infancy right now.

DB: How do you intend to trackle these challenges?

JWT: To be honest, scaling up the process is something I haven't figured out. I do think it is great that more and more people are becoming interested in becoming a data visualization designer that has knowledge of coding, data and design. But that would mean that they could work in individually or in small teams. But working on a visualization with a large team remains a challenge I think.

Author Biography

Jan Willem Tulp is a data visualizer from The Netherlands. Since 2011 he works as a freelancer (TULP interactive) for clients all over the world. His visualizations do not just provide insight, but are well crafted and beautifully designed. His work ranges from custom visual exploration tools that help a client explore a dataset to visualizations that are more explanatory in nature. Some of his projects can be considered data-art.

Every now and then Jan Willem initiates a project by himself, just to explore new ideas, technologies and to have some fun. His work has appeared in magazines, books and exhibitions. Jan Willem works for clients such as *Scientific American*, *Popular Science*, Nielsen, Amsterdam Airport, Philips and *World Economic Forum*.

For more information and contact:

http://www.janwillemtulp.com
https://twitter.com/JanWillemTulp.

http://www.springer.com/978-1-4471-6595-8

New Challenges for Data Design
Bihanic, D. (Ed.)
2015, XIV, 447 p. 283 illus., 248 illus. in color.,
Hardcover
ISBN: 978-1-4471-6595-8