an introduction to

VISUALIZING DATA by joel laumans

AN INTRODUCTION TO VISUALIZING DATA by Joel Laumans

Table of Contents

1	**Introduction**	**1**
	Definition	
	Purpose	
2	**Data visualizations**	**2**
3	**Examples**	**3**
4	**The data**	**5**
	Examine the data	
5	**Data visualization patterns**	**8**
6	**Revealing the data**	**12**
	Preattentinve variables	
7	**Multimedia**	**16**
	Screen resolutions	
	User interaction	
8	**Tools and further reading**	**20**

Foreword

The purpose of this document is to provide an introduction to the theory behind visualizing data. After studying the works of many talented people I decided to summarize the key points of information into this single paper. If you found this document interesting please take some time to look at the list of resources that I used (see Chapter 8) because I could never have created this without the excellent work done by others.

If you have any comments or feedback please feel free to contact me.

Joel Laumans

User Experience Designer

jlaumans@gmail.com

http://www.piksels.com/

http://creatinginspiration.net/

http://www.linkedin.com/in/joellaumans

1 Introduction

To create a truly powerful data visualization a combination of artistic, statistical, and mathematical skills are required - this is most likely the reason why the first multivariate statistical graphics only appeared late in the 18th century. Over time the use of data visualization has become continually more popular; partly because the tools to create data graphics are readily available, but also, because there is an urgency to communicate information both quickly and effectively as possible.

The purpose of this document is to provide an introduction to data visualization by exploring the purpose, the requirements, and methods of visualizing data.

2 Data visualizations

Definition

Data visualizations, also known as data graphics, can be best explained by quoting Edward Tufte:

> *"Data graphics visually display measured quantities by means of the combined use of points, lines, a coordinate system, numbers, symbols, words, shading, and color."*

A common misconception is that data visualizations are the same as information graphics (infographics). It is important to understand that data visualizations always communicate a message by visualizing quantifiable data objectively, while infographics can be used to communicate any information at all (usually with a specific goal); regardless of whether it is quantifiable or not.

Purpose

Creating a data visualization is more than simply translating a table of data into a visualization. Data visualizations should communicate data in the most effective way; to truly reveal the data they should be **quick**, **accurate**, and **powerful**. Creating visuals can easily summarize and communicate data to other people - making even the largest or most complicated sets of data understandable.

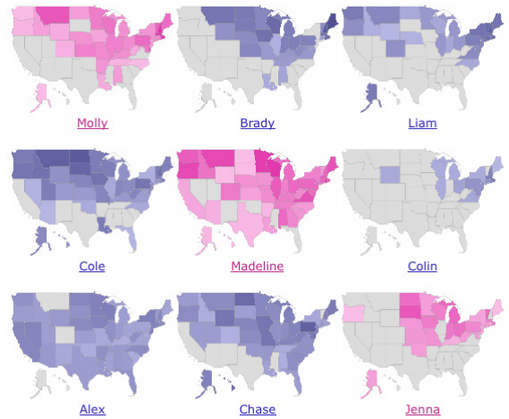
Usage

Data visualizations can take many different forms depending on the information that is being communicated - from simple bar charts that communicate rising oil prices to interactive applications that analyze website visitor data.

3 Examples

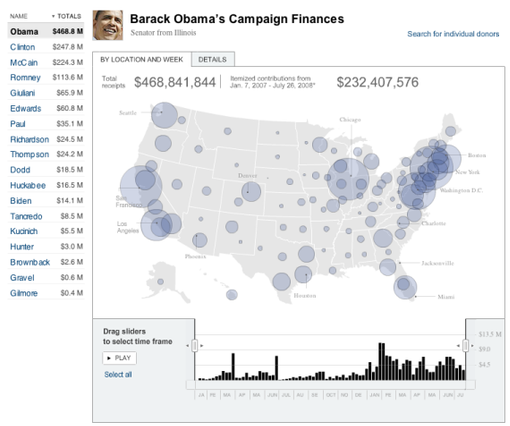
Name popularity in the United States of America

This is a collection of maps visualizing the popularity of names in different states of the United States. The color (pink vs blue) represents the gender of the name, and the saturation of the color represents the popularity of the name in the respective region. This is a very simple and powerful visualization because we can easily conclude that 'Alex' is popular through almost the entire country, while 'Colin' is more popular in the North Eastern states.



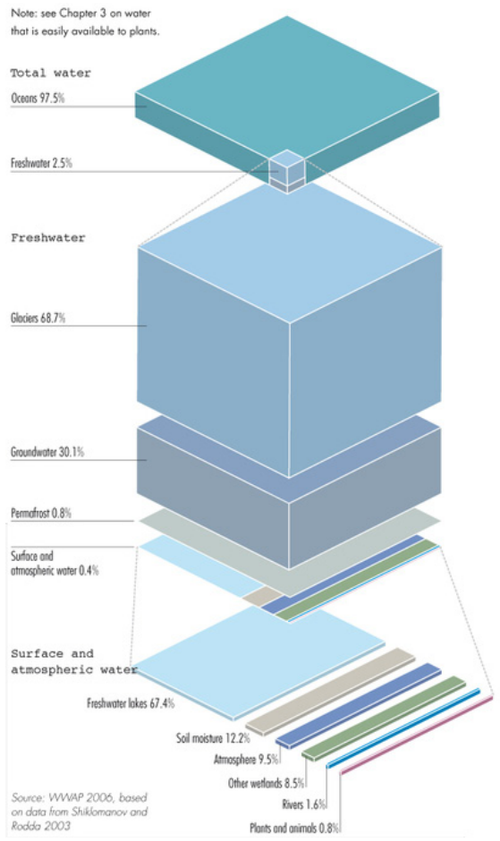
Campaign Finances

This is a screenshot of an interactive visualization made by The New York Time's to visualize the which regions were financing the US presidential candidates. Each circle represents a major city, and the size of the circle represents the amount of money donated. At the bottom is a bar chart which shows the amount of money (vertical) donated per week (horizontal). What makes it very powerful, was the element of interaction - users can select which candidate they want to view, the time span, and retrieve extra data by clicking on the circles.



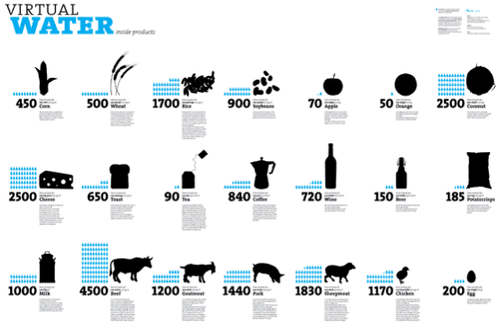
Distribution of the world's water

This a great example of how visualizing data can be used to communicate a very powerful message. Read from top to bottom, this visual tells us that only 2,5% of the world's water is fresh water, and of those 2,5% only a small portion is actually available for human use because the majority is frozen in glaciers; a fact that most people are unaware of, but communicated clearly in this visualization..



Virtual Water

This is another visualization created to raise awareness about excessive water usage. It visualizes the amount of water that is used in the production of certain products. The blue water drops represent the amount of water used per product: for example it takes 1000 liters of water to produce 1 liter of milk, or 4500 liters of water for a single piece of steak. Very simple and effective.



4 The data

There are countless methods of how to create data visualizations - they will vary greatly depending on the content of the data, as well as the purpose of the visualization. Whichever form it takes, the most important is to maintain graphical excellence, explained by Tufte as:

> *"Excellence in statistical graphics consists of complex ideas communicated with clarity, precision, and efficiency."*

According to Tufte's checklist, all data visualizations should:

- Show the data
- Be accurate (don't distort the data)
- Make large data sets coherent
- Serve a clear purpose
- Reveal the data at different levels (overview versus detailed)
- Encourage the viewer to compare different pieces of data

Examine the data

Every set of data can be visualized in multiple ways, some more effective than others. In order to create an effective data visualization the data first that to be understood. Therefore, the first step in creating a data visualization is to examine the data. Basic questions that need to be answered are:

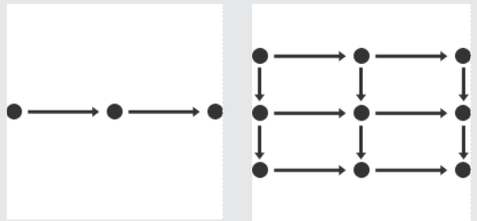
- What is the data?
- What are the relationships between the variables?
- How is the data organized?
- What needs to be communicated?

Before thinking about how the visualization should look, it is important to have answered these questions because they will determine the form of the data visualization.

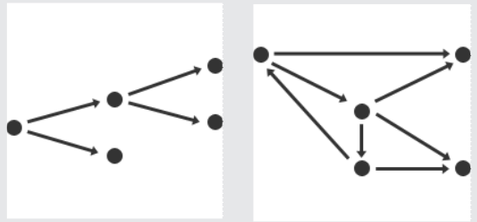
One of the most important parts is understanding how the data is organized and related to each other. For example, if you want to communicate company growth over time a simple line chart might be the most effective - but to communicate population density around the world it might be more effective to use a cartographic (map) visual.

There are countless possibilities of how to visualize data; but there are many design patterns which are proven to be effective for specific types of data. The following section provides examples of data visualization patterns and a short rationale for when to use them.

Examples of how data can be organized



linear *tabular*



hierarchical *networked*

geographic *other*

Ben Fry's "Seven Stages of Visualizing Data"

Visualizing complex data sets often requires insights from diverse fields of knowledge, such as statistics, data mining, graphic design, et cetera. Ben Fry suggests a seven stage design process, reconciling all stages into a single process.

1. Aquire

Obtain the data, whether from an Excel document, an XML feed, et cetera.

2. Parse

Data will not always be organized ideally for visualizing it. Give your data structure by ordering it into categories.

3. Filter

Be careful to prevent information overload, remove all but the data of interest

4. Mine

Apply methods from statistics or data mining as a way to find patterns or meaning in the data.

5. Represent

Choose a basic visual model to visual the data. (see Chapter 5)

6. Refine

Improve the basic representation to make it more clear and more visually engaging. (see Chapter 6)

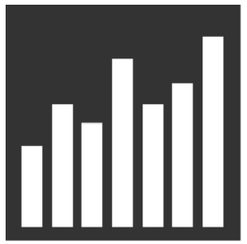
7. Interact

Add methods for manipulating the data. Allows users to control what they see or even possibly how they see it.

5 Data visualization patterns

Independent quantities

Comparing the values of independent variables



Bar charts

Simple bar charts are the most common form of data visualizations. Typically they only display different quantities of single-variable data. However other variations, such as stacked bar charts or multi-set bar charts can be used to compare multiple variables using bars.

Continuous quantities

For data that is continuous, for example when visualizing data over a period of time.

Line graphs

Line graphs are created by plotting points on a Cartesian grid, usually with the horizontal axis representing time. They are very powerful because without looking at the specific data, they show how a variable develops over time (from left to right).

Stacked area charts

Similar to line charts, however with the added value of filled areas. The data that is stacked adds up to a total of all variables combined. For example a business might use stacked area charts to visualize their total income, with each stacked area a different income channel.

Proportions

Proportions are used when the data represents parts of a whole.

Pie charts

Simple pie charts are the most common visual used to compare proporational data. They give viewers a very quick understanding of the distribution of the data. Pie charts are not useful when comparing many pieces of data with relatively close values.

Ring charts

Similar to pie charts; ring charts are used to visualize the distribution of a data set. The advantage is they compare similar data sets. The alternative would be to place multiple pie charts next to each other, this can also be viewed as a space-saver.

Correlations

These are used when each piece of data has two quantifiable variables which can be plotted on a grid.

Scatterplots

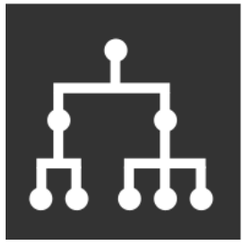
They are created by plotting independent points on a Cartesian grid. Scatter plots are often used to find the relationship between data or to reveal information such as trends within the data which are not easily visible when in a table. Only works with two dimensional data.

Bubble charts

Similar to scatterplots; however, bubble charts display more dimensions of data by varying size (or also color, texture, etc) of the bubbles. It therefore can display multiple dimensions of data in a two dimensional display.

Hierarchies

Used when the data has a strict hierarchy that needs to be communicated.



Tree diagrams

Tree diagrams are often used when wanting to represent the strict hierarchy of data. They are most often used to represent strict hierarchies such as family trees or how data is stored in a computer system.

Networks

Network visualizations are used when the most important feature of the visualization is to show which data is connected to each other, as opposed to how.

Diagram maps

These visualizations are used to primarily represent the connections between different nodes or points. Their purpose is to show which points are connected to each other. Common examples of diagram maps are metro maps and social network visualizations.

Cartographic

For data that is relevant to specific locations or regions which can be plotted on a map.

Maps

Maps are used when the data is related to a specific location (for example a city, or country). The advantage is that their spatial representation directly relates to a real-world situation. However at times can be difficult to read.

Flows

When the data is part of a process, it can be visualized using flow diagrams.

Sankey diagrams

Sankey diagrams are composed of several smaller 'arrows' or channels, which merge together into one large channel. For example, a sankey could be used to visualize a movie's revenue, first movie tickets, then merchandise, then DVD sales. Each smaller arrow would be representing a source of income.

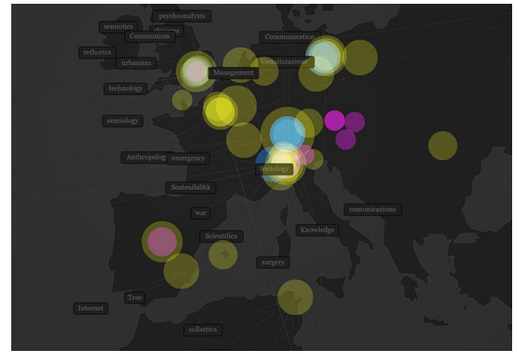
Data visualization patterns (cont.)

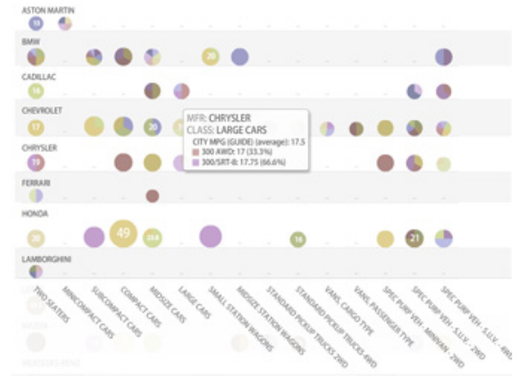
The aforementioned examples represent only a small number of the most popular types of visualizations. To find more information about the different types of data visualizations refer to the references listed at the end of this document.

There are many advantages to using existing patterns for data visualization. They have been proven to work effectively for a specific type of dataset. Furthermore, people are familiar with reading these types of visualizations, making them easier to understand.

Unfortunately, many data sets have unique characteristics which will force you to come up with new ways of communicating the data in the most effective and powerful way possible. When one single data visualization pattern will not be adequate, the solution many times is to create a combination of several data patterns to create a single message. Each pattern can then be used to represent specific variables in the data.

The two visualizations on the right are good examples of how to combine visualization patterns. The top visualization is a combination of a cartographic, bubble chart, and network diagram. The bottom is a combination of bubble chart and pie charts plotted onto a matrix.





6 Revealing the data

Creating a powerful data visualization is not about simply translating a table of data into a visual graphic, it is about communicating the meaning of the data. Choosing the most adequate visualization design pattern is an important step because it will immediately tell users about how the data is organized and what you are trying to communicate about the data. However, it is not only the type of visualization pattern chosen - but also the design of the individual elements that play an important role in communicating information to others.

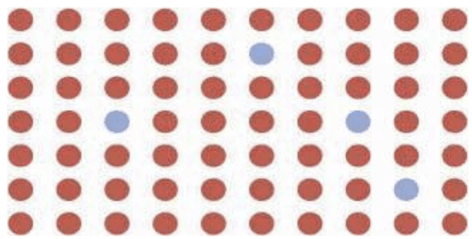
Preattentinve variables

There are certain visual features in design that will work preattentively; they will communicate something about the design before the user pays conscious attention to it. Jennifer Tidwell has coined these as preattentive variables.

A powerful data visualization should work quickly and effectively; therefore the visual design should help reveal the data to the users.

In Tidwell's book, Designing Interfaces, she provides the following examples.

Take a look at the following group of dots and find all of the blue dots.



Quick, right? Even if we increased the number of red dots, finding the blue dots will be just as quick, because color works preattentively!

Now take a look at the values in the table and find all numbers greater than 1.0.

0.103	0.176	0.387	0.300	0.379	0.276	0.179	0.321	0.192	0.250
0.333	0.384	0.564	0.587	0.857	1.064	0.698	0.621	0.232	0.316
0.421	0.309	0.654	0.729	0.228	0.529	0.832	0.935	0.452	0.426
0.266	0.750	1.056	0.936	0.911	0.820	0.723	1.201	0.935	0.819
0.225	0.326	0.643	0.337	0.721	0.837	0.682	0.987	0.984	0.849
0.187	0.586	0.529	0.340	0.829	0.835	0.873	0.945	1.103	0.710
0.153	0.485	0.560	0.428	0.628	0.335	0.956	0.879	0.699	0.424

Table 1: Find the values larger than 1.0

There is no visual aid to help us find the values greater than 1.0, forcing us to read all of the data and understand it. Now look at the following table.

0.103	0.176	0.387	0.300	0.379	0.276	0.179	0.321	0.192	0.250
0.333	0.384	0.564	0.587	0.857	1.064	0.698	0.621	0.232	0.316
0.421	0.309	0.654	0.729	0.228	0.529	0.832	0.935	0.452	0.426
0.266	0.750	1.056	0.936	0.911	0.820	0.723	1.201	0.935	0.819
0.225	0.326	0.643	0.337	0.721	0.837	0.682	0.987	0.984	0.849
0.187	0.586	0.529	0.340	0.829	0.835	0.873	0.945	1.103	0.710
0.153	0.485	0.560	0.428	0.628	0.335	0.956	0.879	0.699	0.424

Table 2: Find the values larger than 1.0

Finding the values has become much easier by simply changing some of the visual features we can significantly improve the search times for users.

On the following pages all eight of Tidwell's preattentive variables will be demonstrated.

In essence, all display patterns use any single or combination of these variables to visualize data. Properly applying these visual variables is what allows large sets of data to be quickly and accurately understood in data visualizations.

7 Multimedia

Until recently, all data visualizations were static and predefined; however, many modern data visualizations are created using multimedia interfaces that expand the possibilities of data visualizations.

When working with multimedia interfaces (as opposed to print), there are new factors which can influence the effectiveness of communicating information to users. The most limiting factor of multimedia is the relatively small amount of information that can be displayed on a screen, but the main advantage of using multimedia is that visualizations can be dynamic, animated, and allow for user interaction.

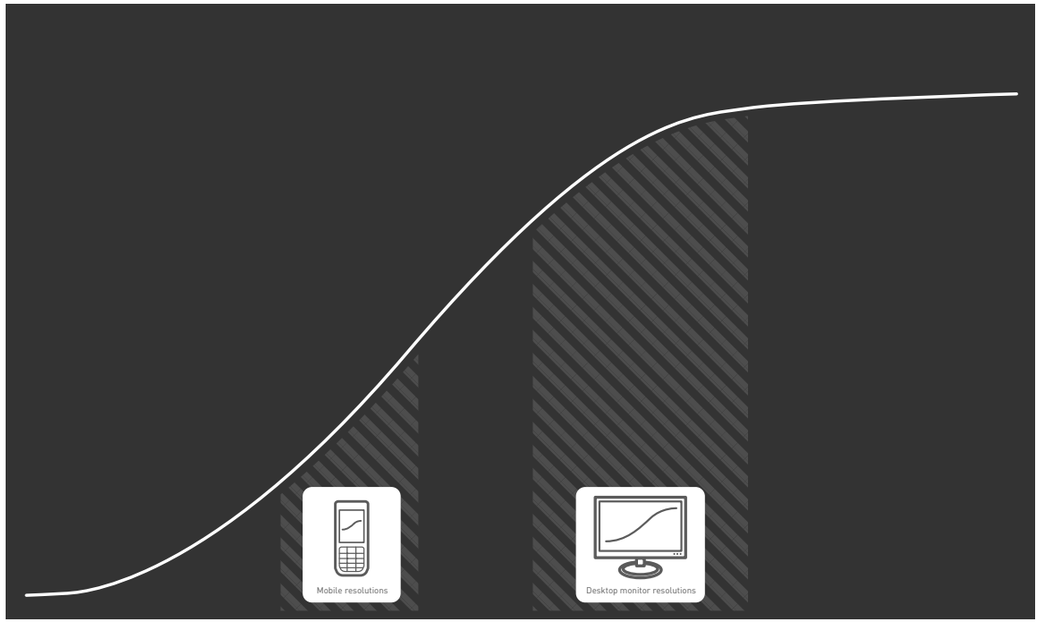
Screen resolutions

The most common screens where users will come across data visualizations are desktop monitors and mobile phones.

Modern mobile phones resolutions vary between 144 x 176 pixels up until 320 x 480 pixels, with the most popular being 240 x 320 pixels. Only simple data visualizations can be displayed at this resolution. Comparing 20 different pieces of data becomes difficult.

Typical desktop computer resolutions are much higher, ranging between 1024 x 768 pixels and 1600 x 1200 pixels, allowing for richer visualizations that communicate sets of data that are larger and more complex.

The graph on the following page illustrates the effectiveness of a data visualization versus the screen resolution of the medium. Low resolutions are a true limiting factor for displaying data. The effectiveness of a data visualization increases quickly as the resolution increases, however plateaus once reaching the upper limit of desktop monitor resolutions. The reason for the plateau is simply because, at this point, more information can be displayed than a person can see at any single point in time. This means that by being able to display more information than a user can see will not increase the effectiveness of the data visualization. When working with multimedia data visualizations, the most effective resolutions will be those of standard desktop monitors.



User interaction

Traditionally, data visualization was about choosing the correct design pattern to visualize the respective data to create the most meaning; however when we allow the user to interact with the data, it is not only about how the data is displayed, but also about how it behaves. When users interact with the data visualization, they can control and manipulate how and what is displayed. There are many forms of user interaction that can be applied to data visualizations. These forms are in two categories:

- *Data selection and filtering* – The user can control which data is displayed
- *Data arrangement and navigation* – The user can control how the data is displayed or viewed.

Data selection and filtering will help users control precisely which data is being visualized. This will help users find data only relevant to what they are looking for, and help prevent information overload.

Data arrangement and navigation can help

users find new meaning in the data. Simply displaying the same data in a different fashion can help people come to new conclusions and see different relationships between the data.

> *"Each set of data has particular display needs, and the purpose for which you're using the data set has just as much of an effect on those needs as the data itself." (Ben Fry)*

Allowing users to control these two variables instantly makes data visualizations more powerful, simply because it makes them more particular to a single user.

When using screens as a medium for displaying data visualizations, they can be organized into four different categories: static, animated, interactive, and direct manipulation

The most powerful type are those that support direct manipulation because they will let users immediately see how the variables they adjust influence the data that is being displayed. These visualizations that which support direct manipulation, have the advantage that:

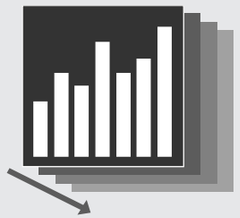
- Users can quickly learn the relationship between different variables.
- Users can immediately see if their actions are furthering their goals, and if not, they can simply change the direction of their activity.

In conclusion, when you take a traditional data visualization and add the combination of user interaction and direct manipulation, it creates a powerful formula for visualizing data. Giving the user control over the visualization makes it more meaningful to the user, and helps them

Categories of data visualizations

Static

The data visualization does not change over time, and offers no user interaction.



Animated

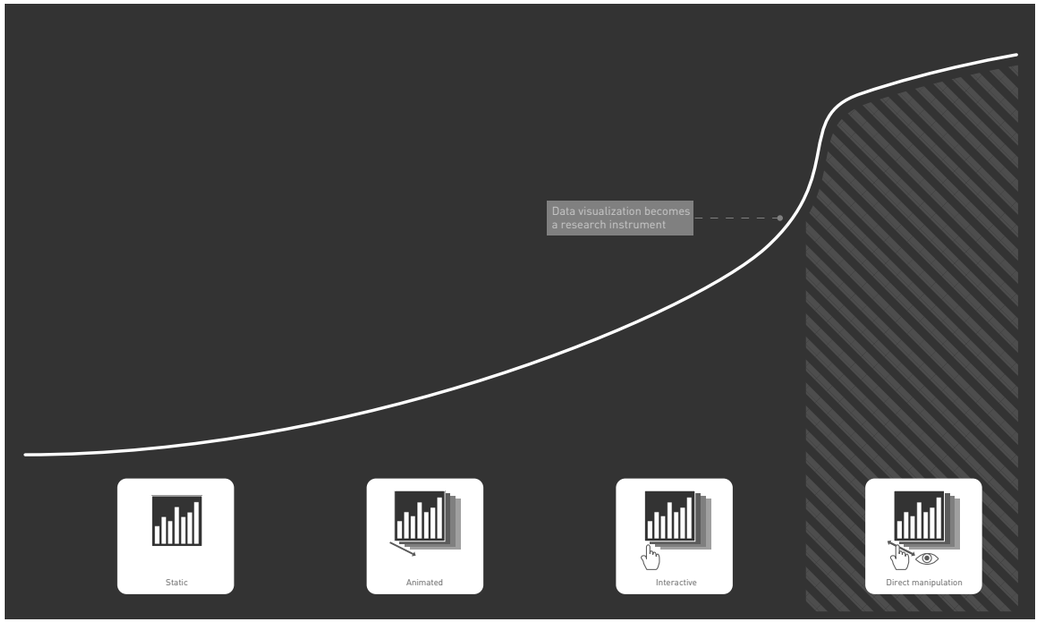
The data visualization changes over time - using time as another dimension or variable.

Interactive

The user can interact with the visualization by altering what data is viewed and / or how it is viewed - however the results of interactions are not immediately visible.

Direct manipulation

The user can interact with the visualization by altering what data is viewed and / or how it is viewed - and the result of all interactions are immediately visible.

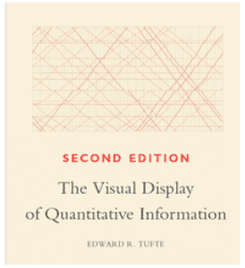


discover new meaning in the data. By adding user interaction, the data visualizations have been transformed from a static display of quantitative data into a tool for discovering new meaning and relationships in the data.

8 Tools and further reading

Books

Here is a short list of books for anyone interested in learning more about visualizing data.



The Visual Display of Quantitative Information

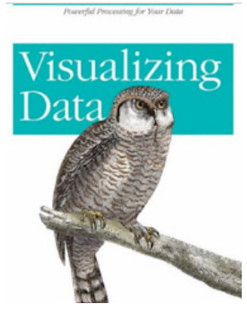
by Edward Tufte

This is the classic book about data graphics. Tufte explains the history and theory behind data graphics using more than 250 great examples. This is a must have book for all people interested in visualizing data.

Designing Interfaes

by Jenifer Tidwell

This is a brilliant book on User Interface Design with handy information for people interested in data visualizations. Tidwell has a whole chapter dedicated to visualizing complex data in which she explains how to organize and display data effectively.



Visualizing Data

by Ben Fry

For those interested in creating complex data visualizations, this is book is for you. This book explains how to create data visualizations using Processing (an opensource platform for data visualization). Each chapter is a tutorial teaching you about data visualization as well as how to use Processing.

Tools

Although simple data visualziations can be made using software such as Microsoft Excel, Here are tools you can use to create more complex data visualizations.

Many Eyes

http://manyeyes.alphaworks.ibm.com/manyeyes/

A research group at IBM created Many Eyes, an online platform for data visualizations. Using existing design patterns and data you can experiment and create your own data visualizations. This is a great place to start for people working with data visualizations for the first time.

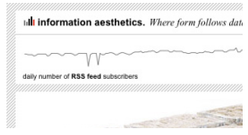
Processing

http://processing.org/

Processing is one of the most popular platforms for visualizing data. This this the must have tool for people who are interested in creating custom data visualizations with their own data. The initial learning curve is quite high because basic programming knowledge is required. The best way to get started is to buy the Ben Fry's book *Visualizing Data* which provides tutorials for people interested in learning Processing.

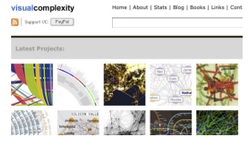
Websites

These are two websites that are frequently updated with great examples of data visualizations and infographics. They showcase the best examples of data visualization out there..



Information aesthetics

http://infosthetics.com/



Visual Complexity

http://www.visualcomplexity.com/vc/

Credits and thanks

A large portion of this booklet is based on work done by Edward Tufte, Ben Fry, and Jennifer Tidwell; three people which have done remarkable work for data visualization. Thank you.

If you find this booklet interesting, please consider looking into the work of these people.

I would also like to thank the following people for reviewing previous versions of this booklet and giving me helpful feedback, I appreciate it.

Nathan Verril
User Experience Consultant

http://www.linkedin.com/pub/1/150/135

Bas Leurs
Lecturer Interaction Design at the Rotterdam University of Applied Siences

http://www.linkedin.com/pub/0/330/308

Peter van Waart
Lecturer and Researcher at the Rotterdam University of Applied Siences

http://www.linkedin.com/in/petervanwaart