THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF SUPPLY CHAIN AND INFORMATION SYSTEMS

DATA VISUALIZATION

FARA DELITSKY
Spring 2010

A thesis
submitted in partial fulfillment
of the requirements
for a baccalaureate degree
in Management Information Systems
with honors in Management Information Systems

Reviewed and approved* by the following:

John Jordan
Senior Lecturer in Supply Chain and Information Systems
Thesis Supervisor

John C. Spychalski
Professor Emeritus of Supply Chain Management
Honors Adviser

* Signatures are on file in the Schreyer Honors College.

ABSTRACT

The amount of data being collected and stored today is growing at an exponential rate. Our ability to turn this raw data into useful information and knowledge is not only key to making the collection of this data worthwhile, but also in preventing this data from having a slowing effect on communication and decision making. Currently, most literature details strengths and weaknesses of already-created visualization, but lack rules to follow as we create data visualization from the beginning. Creating a visualization requires an understanding of the data type at hand, as well as a specific purpose and audience for this data. Designing an informative data visualization also requires that we continually question possible causality until we have gathered enough data to reveal a story that the raw data would not have otherwise told.

TABLE OF CONTENTS

LIST OF FIGURES

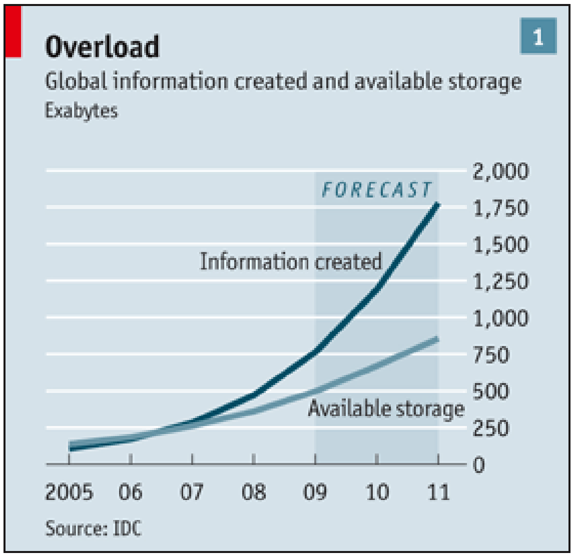
LIST OF TABLES

Chapter 1

Introduction

There are about 6.69 billion people living in the world. Today each item we purchase, each highway toll collection point we drive through, and each e-mail we send, generates new data that is recorded and stored. In 2008 alone, all of these transactions and activities generated 487 exabytes of data-- a number unfathomable to most (Gantz and Reinsel, 1). To put this number in perspective, in order to hold all of this data, each of the 6.69 billion people in the world would need to own 4.5 iPods, each of which holds 16GB of data. By 2012, it is expected that five times as many bits will be created as in 2008 (Gantz and Reinsel, 1). Figure 1-1 below, a graph presented in “Data, data everywhere,” published by *The Economist*, shows this forecasted growth in data creation. Additionally, Figure 1-1shows that the amount of data being created far surpasses available storage, a trend that will only worsen in the years to come.

Figure 1-1 Information and storage



Source: *The Economist*, "Data, data everywhere"

Where is all of this data coming from? According to Keim, et al., "Virtually every branch of industry or business and any political or personal activity nowadays generate vast amounts of data" (154). One company that is known for its data collection and storage is Wal-Mart. According to an article in *The Economist*, "Wal-Mart, a retail giant, handles more than 1 million customer transactions every hour, feeding databases estimated at more than 2.5 petabytes" ("Data, data everywhere"). While the author attributes some of this growth in data creation to corporations' participation in large-scale data collection, the article also says that this increase has been driven by the over one billion people who entered the middle class between 1990 and 2005 ("Data, data everywhere"). As more people enter the middle class, more people become literate, which in turn "fuels information" ("Data, data everywhere"). With this change, more information is not only being created, but also shared across the globe.

While the data we have today empowers us in many ways, and is now required for most businesses to compete successfully, simply having the data is not enough. According to O'Grady and O'Grady, "A huge portion of the world's knowledge lies at our fingertips, but limited competencies in terms of comprehension and communication often block our access" (75). One problem we face with simply collecting vast amounts of data, is that raw data- lists of names, numbers, and other statistics- has little meaning. It is nearly impossible to look at lists of raw data and recognize patterns or trends, make comparisons, or understand reasoning or causality. Without turning raw data into information by means of charts and graphs, the data is useless for both informing an audience, and as a basis for decision making. According to *The Economist*, while the data we have is plentiful, "what is scarce is the ability to extract wisdom from them" ("Data, data everywhere"). With the amount of data available, there is the potential for too much data to make decision making and daily processes slower and more difficult. If we don't know how to use the data, or where to look for answers, the data is no longer empowers us, but in fact slows us down. This is where effective data visualizations come into play. Because "good design has the power to prevent poor user experiences and lost opportunities," extracting knowledge from data, not simply collecting it, is essential. (O'Grady and O'Grady, 75). Today, as the amount of data we have grows exponentially, so does our need for effective visualizations to create information and knowledge from the data we have.

Scope and Paper Overview

The scope of this research does not include theories of design in terms of memory, font, symbols, meaning of color, or cultural considerations. Instead, my research is focused on effectiveness of a visualization in relation to how well it tells the story behind the data. In other words, in order to be “effective,” a visualization must inform a user of meaning and knowledge from the data more than a spreadsheet or list of numbers otherwise would.

Chapter 2

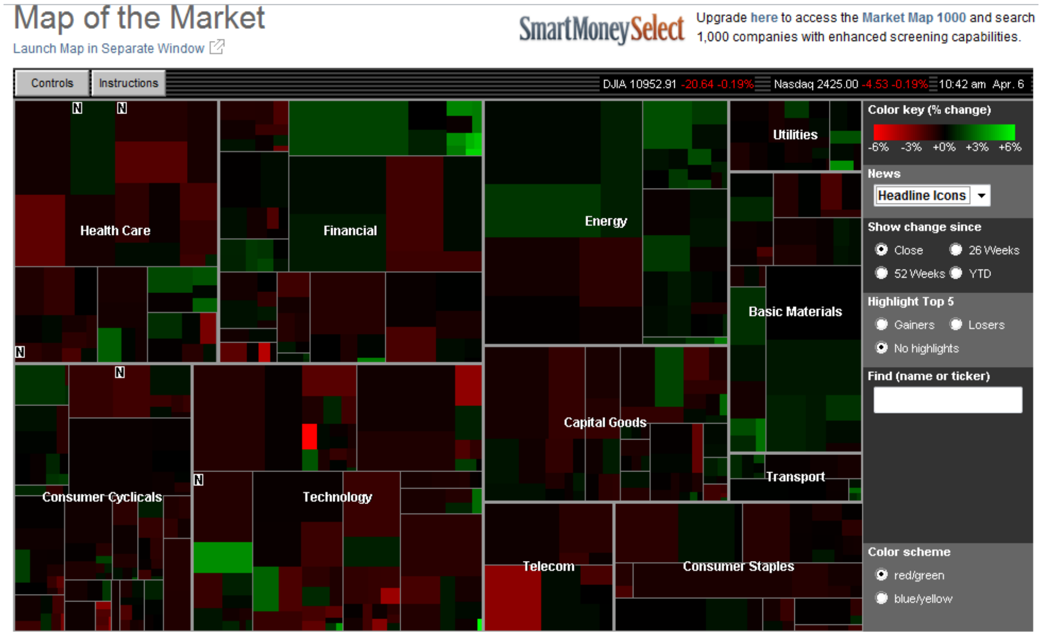
Problem

The question arises as to how to create visualizations that provide real meaning. As described in *The Information Design Handbook,* "information delivery is dependent on clarity of communication to retain its relevance to a global audience. Designers provide that context by turning statistics into stories, providing meaning for the end user" (O'Grady and O'Grady, 16). A review of current literature on this topic finds books replete with already-created data visualizations, followed by in-depth critiques. These post-assessments often detail which aspects of the visualization lack clarity, and which aspects meet the needs of the end user. *However, none of these analyses dictate a set of rules to follow while creating the visualization.* This process of turning data into a meaningful story requires a consideration of the purpose of communicating the data, the audience's needs, and the data type being shared. As the visualization is being created, we must constantly question causality, and continue to add data until the story is complete (or as complete as possible).

Purpose

While some data visualizations are created for the purpose of informing or persuading an audience, others are created to empower the user to make a decision. If a visualization is being created for the purpose of decision making, the type of decision being made dictates the level of precision necessary. For example, deciding which new markets would be most profitable requires data pertaining to competitors, consumer trends, laws, logistics, and so on. Because this data is dense, it is best displayed in an interactive form, allowing a viewer to filter or drill down as necessary. If this is a decision related to something more clear cut, the viewer just needs a general idea as opposed to detailed information. This is best displayed in a way where glancing at a chart or visualization reveals enough of an answer to decide. A well known example of this is the Map of the Market, found on Smartmoney.com, shown in Figure 2-1 below:

Figure 2-1 Map of the Market



Upgrade here to access the Market Map 1000 and search 1,000 companies with enhanced screening capabilities.

DJIA 10952.91 -20.64 -0.19% Nasdaq 2425.00 -4.53 -0.19% 10:42 am Apr. 6

News: Headline Icons

Show change since: ● Close ○ 26 Weeks ○ 52 Weeks ○ YTD

Highlight Top 5: ○ Gainers ○ Losers ● No highlights

Find (name or ticker)

Color scheme: ● red/green ○ blue/yellow

Source: Smartmoney.com

The Map of the Market, created by Martin Wattenberg, is a real-time, interactive heat map. A heat map uses a range of colors to represent different data values. The map is divided into different segments, each of which represents one of eleven industries (Health Care, Financial, Technology, Communication, Energy, Capital Goods, Communication, Utilities, Basic Materials, Transport, and Consumer Staples). Within each of these industries, each rectangle represents an individual company, where the relative size of each rectangle represents the market cap for the company. The industries' companies are clustered together so that rectangles in close proximity to one another all have similar recent stock performance. An algorithm places these rectangles such that they are clustered by performance, yet also fit inside the confines of the rectangular industry segment. The colors on the map indicate recent stock price changes: shades of red indicate negative changes in stock prices, while shades of green show positive ones. According to its creator, Wattenberg, "the goal was to give a quick answer to the question, 'how is the market doing today?'" (bewitched.com). While the map is interactive and enables a user to access industry and company specific data if they want, a glance at the map allows a viewer to gauge the changes and trends of a market comprised of 500 companies in 11 different industries based simply on color and shade.

Audience

Considering the intended audience of visualization is closely tied to considering its purpose because the specific audience dictates how the data will be used. Depending on the audience of the visualization, the designer must cater to the viewers' needs, in terms of choosing the right type of data representation for the message being communicated.

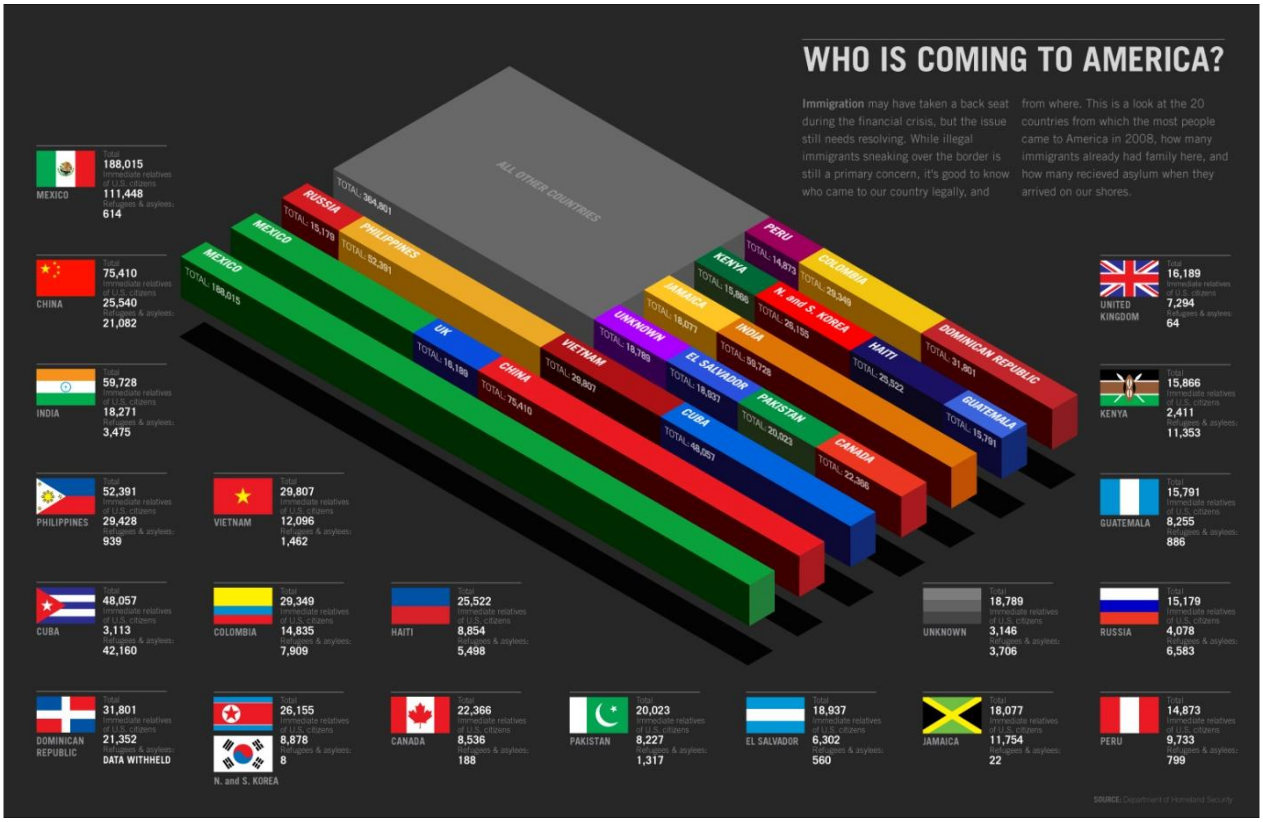
Data Type

Data type plays a key role in the basis of design. Keim, et al. categorizes data into the following two "sub-communities": the spatio-temporal data sub-community, or the network and graph data sub-community (163). Data that falls into the spatio-temporal data category is "data with references to time and space" (Keim et al., 163). Geospatial data is data connected to specific real-world locations, and temporal data is data connected to time, and often has a specific order (Keim et al., 163). On the other hand, network and graph data is data that naturally maps to real life relationships and situations such as "transportation networks, electric power grids," and "interactions between people" or entities (Keim et al., 164).

Whether there is time series data, geo-spatial data, or network data available, it is important to represent data in a logical way for the end user. Data with a geospatial aspect, for example, lends itself well to a geographic map, because users intuitively know where to look for data pertaining to specific countries, cities, or landmarks. On the other hand, using space when representing data that has no logical geo-spatial mapping often proves to be confusing and ineffective to the user, although there are some exceptions.

While placing geospatial data on a map may seem like an obvious design choice to some, many designers fail to use space appropriately, as in the following example. Figure 2-2 below, a data visualization from Vizworld.com, entitled "Who is Coming to America?," presents data on the number of people immigrating to America from each of 20 countries (Hand).

Figure 2-2 “Who is Coming to America?”



WHO IS COMING TO AMERICA?

Immigration may have taken a back seat during the financial crisis, but the issue still needs resolving. While illegal immigrants sneaking over the border is still a primary concern, it's good to know who came to our country legally, and from where. This is a look at the 20 countries from which the most people came to America in 2008, how many immigrants already had family here, and how many recieved asylum when they arrived on our shores.

ALL OTHER COUNTRIES
TOTAL: 364,801
MEXICO TOTAL: 188,015
RUSSIA TOTAL: 15,179
PHILIPPINES TOTAL: 52,391
UK TOTAL: 16,189
CHINA TOTAL: 75,410
VIETNAM TOTAL: 29,807
UNKNOWN TOTAL: 18,789
JAMAICA TOTAL: 18,077
KENYA TOTAL: 15,866
PERU TOTAL: 14,873
COLOMBIA TOTAL: 29,349
N. and S. KOREA TOTAL: 26,155
INDIA TOTAL: 59,728
EL SALVADOR TOTAL: 18,937
CUBA TOTAL: 48,057
PAKISTAN TOTAL: 20,023
CANADA TOTAL: 22,366
HAITI TOTAL: 25,522
GUATEMALA TOTAL: 15,791
DOMINICAN REPUBLIC TOTAL: 31,801
MEXICO Total 188,015 Immediate relatives of U.S. citizens 111,448 Refugees & asylees: 614
CHINA Total 75,410 Immediate relatives of U.S. citizens 25,540 Refugees & asylees: 21,082
INDIA Total 59,728 Immediate relatives of U.S. citizens 18,271 Refugees & asylees: 3,475
PHILIPPINES Total 52,391 Immediate relatives of U.S. citizens 29,428 Refugees & asylees: 939
VIETNAM Total 29,807 Immediate relatives of U.S. citizens 12,096 Refugees & asylees: 1,462
CUBA Total 48,057 Immediate relatives of U.S. citizens 3,113 Refugees & asylees: 42,160
COLOMBIA Total 29,349 Immediate relatives of U.S. citizens 14,835 Refugees & asylees: 7,909
HAITI Total 25,522 Immediate relatives of U.S. citizens 8,854 Refugees & asylees: 5,498
UNKNOWN Total 18,789 Immediate relatives of U.S. citizens 3,146 Refugees & asylees: 3,706
RUSSIA Total 15,179 Immediate relatives of U.S. citizens 4,078 Refugees & asylees: 6,583
DOMINICAN REPUBLIC Total 31,801 Immediate relatives of U.S. citizens 21,352 Refugees & asylees: DATA WITHHELD
N. and S. KOREA Total 26,155 Immediate relatives of U.S. citizens 8,878 Refugees & asylees: 8
CANADA Total 22,366 Immediate relatives of U.S. citizens 8,536 Refugees & asylees: 188
PAKISTAN Total 20,023 Immediate relatives of U.S. citizens 8,227 Refugees & asylees: 1,317
EL SALVADOR Total 18,937 Immediate relatives of U.S. citizens 6,302 Refugees & asylees: 560
JAMAICA Total 18,077 Immediate relatives of U.S. citizens 11,754 Refugees & asylees: 22
PERU Total 14,873 Immediate relatives of U.S. citizens 9,733 Refugees & asylees: 799
UNITED KINGDOM Total 16,189 Immediate relatives of U.S. citizens 7,294 Refugees & asylees: 64
KENYA Total 15,866 Immediate relatives of U.S. citizens 2,411 Refugees & asylees: 11,353
GUATEMALA Total 15,791 Immediate relatives of U.S. citizens 8,255 Refugees & asylees: 886

SOURCE: Department of Homeland Security

Source: Vizworld.com

The description on this visualization reads as follows:

> Immigration has taken a back seat during the financial crisis, but the issue still needs resolving. While illegal immigrants sneaking over the border is still a primary concern, it’s good to know who came to our country legally, and from where. This is a look at the 20 countries from which the most people came to America in 2008, how many immigrants already had family here, and how many received asylum when they arrived on our shores (Hand).

The creator of this visualization used a bar graph in the shape of an American flag, using each stripe as a bar, or a collection of bars. This American flag serves the purpose of creative design, and not effectiveness. The flag does not work to take raw data and transform it into knowledge. Instead, this visualization still displays raw data, now in a less organized form than a simple list.

The following describes the use of the American flag, as well as many other common design mistakes that were made in this visualization:

- **Use of space** Instead of placing these numbers on their respective countries on a geographical map, the creator of this visualization placed each country on the stripes of an American flag. While most people know directly where to look on a geographical map to find a specific country, there is no sensible way to find these countries on this visualization. The countries have been placed on this flag only in such an order so that they fit within the confines of the 7 available stripes.

- **Readability** Further details about each country surround the American flag bar graph. Unlike the countries' placements on the bar graph itself, the information surrounding the graph is ordered by how many total immigrants came to America (largest to smallest). Because the countries are listed in order of immigration numbers on the side, while the countries' bars were arranged only to fit the stripes, the information and the bars do not coordinate well. Reading this visualization requires work on the part of the user to match bars with data.

- **Representation** While each stripe on an American flag is equal in length, the countries, or groups of countries do not add up to equal numbers of immigrants. This implies that the bars on the graph are not accurately scaled to their relative size.

- **Comparisons** With staggered placement of bars it is nearly impossible to compare sizes. One cannot visually determine if the blue bar that represents Haiti is larger or smaller than the dark red bar that represents Vietnam without reading the numbers (in which case, the bars are useless).

- **Causality** While this map presents raw data-- countries and numbers-- there is no explanation for the vast differences in numbers among the different countries. For example, why is that of 188,015 of Mexico's total immigrants, only 614 are listed under Refugees/Asylees, where has 42,160 of Cuba's total 48,057 immigrants fall under this category (Hand)?

In order to understand how an alternate design could improve the effectiveness for the user, I extracted the data from this visualization, and placed it into a spreadsheet. Table 2-1 lists the countries in the same order (largest to smallest), with the associated data in adjacent cells.

Table 2-1 “Who is Coming to America?” Data

Country	**Total**	**Immediate Relatives of US citizens**	**Refugees and Asylees**
Mexico	188,015	111,448	614
China	75,410	25,540	21,082
India	59,728	18,271	3,475
Phillipenes	52,391	29,428	939
Cuba	48,057	3,113	42,160
Dominican Republic	31,801	21,352	N/A
Vietnam	29,807	12,096	1,462
Columbia	29,349	14,835	7,909
North and South Korea	26,155	8,878	8
Haiti	25,522	8,854	5,498
Canada	22,366	8,536	188
Pakistan	20,023	8,227	1,317
El Salvador	18,937	6,302	560
unknown	18,789	3,146	3,706
Jamaica	18,077	11,754	22
United Kingdom	16,189	7,294	64
Kenya	15,866	2,411	11,353

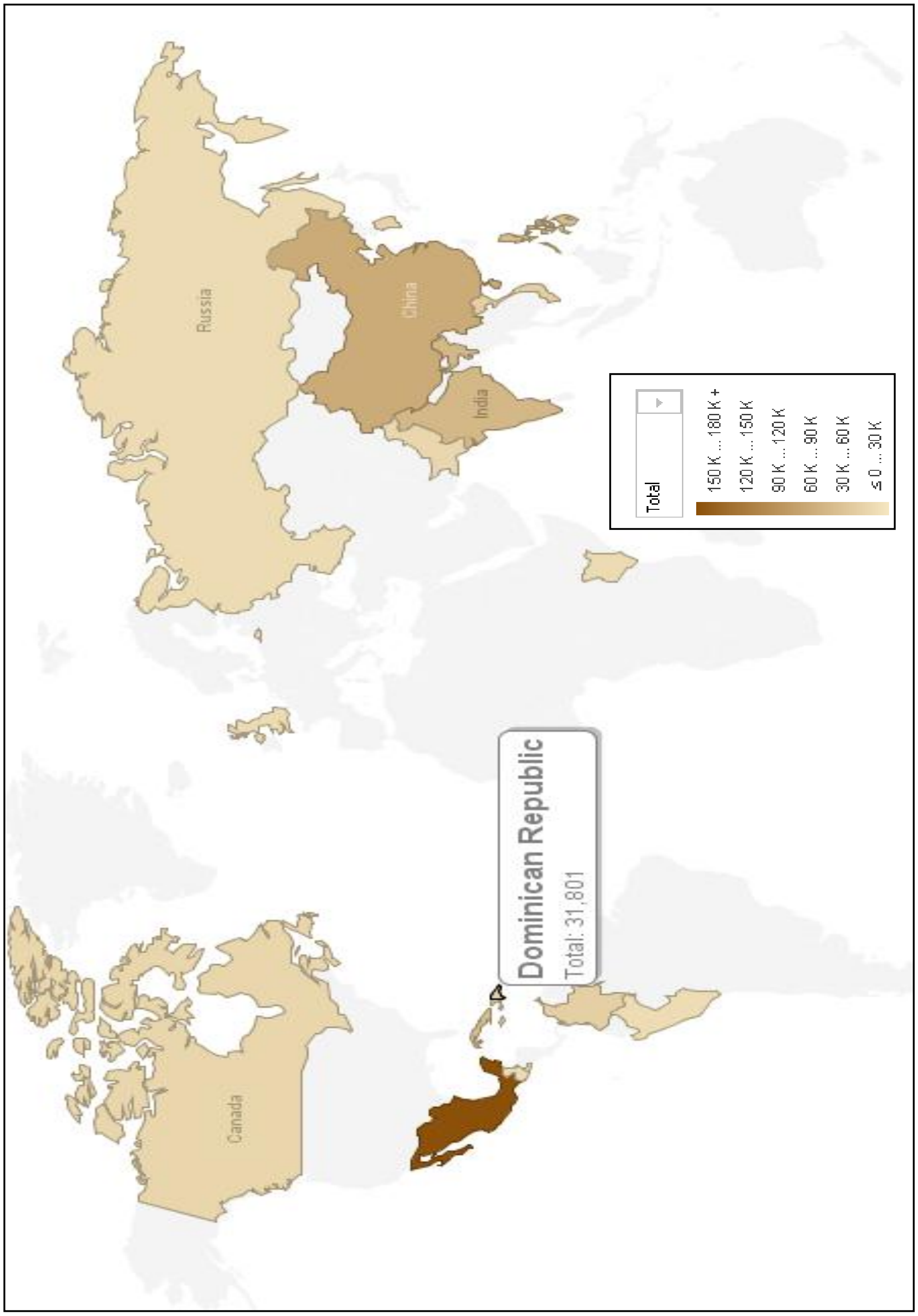
Guatemala	15,791	8,255	886
Russia	15,179	4,078	6,583
Peru	14,873	9,733	799

Source: Data from “Who is Coming to America?” (Figure 2-2)

This simple list is more effective than the creative graph because one does not have to match bar graph with corresponding flags to find information on Immediate Relatives, and Refugees and Asylees. In the case that a simple list of raw data more effectively communicates information to a user, we can conclude that the visual aspect has likely added confusion, not analytical clarity to the data at hand.

While this list of data is clearer than the American flag bar graph, this does not mean that the list of data is the most effective mode of communication. Again, because this data has a geospatial aspect, the data could be best displayed on a map, as Figure 2-3 shows.

Figure 2-3 Map version of “Who is Coming to America?” Data



Source: Data from “Who is Coming to America?” visualization (Figure 2-2)

Map created with Many Eyes beta

“Who is Coming to America?” incorporates geospatial data that was not placed on a geographic map. On the other hand, for SmartMoney.com’s Map of the Market, for which no geographic data exists, the use of space has proven to be effective. Why is it that without any geographical mapping or logical placing, that the dimension of space works so well for the Map of the Market? The success of this map can be attributed to the following:

- **Organization** The map visually organizes a complicated market, organizing company by industry, and placing like companies close together.
- **Glanceability** Without reading any numbers (percentages or dollar values), a user can visually understand relative market cap of the companies based on size, and stock price performance based on color.
- **Consistency** The consistency of the placement of these industries has made it almost like a map in the sense that regular viewers know just where to look for industry- or company-specific information.

Chapter 3

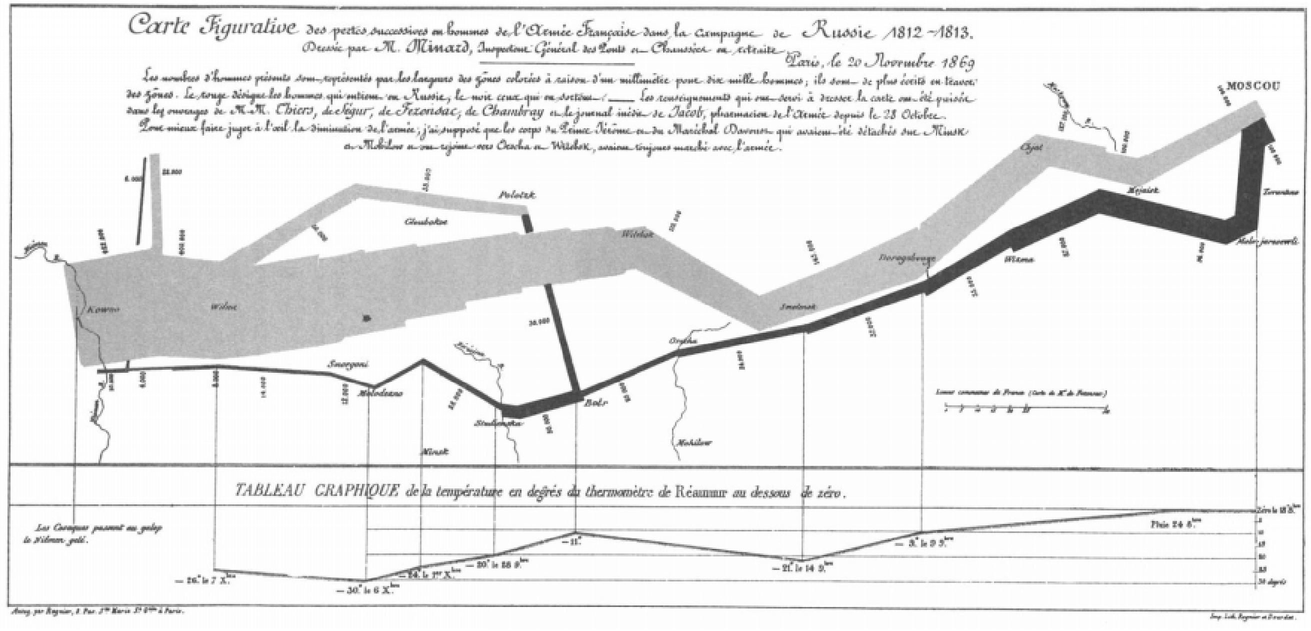
Charles Joseph Minard's Map

Many professionals point to one map as encompassing all of the most important principles of analytical design. This well known visualization is Charles Joseph Minard's map of Napoleon's army invading Russia in 1812, which effectively tells a story of Napoleon's Russian campaign. Edward Tufte, a Yale professor and expert statistician, whose work has defined the field of information design, calls this "one of the best statistical graphics ever" (*Beautiful Evidence*, 122). Minard included many dimensions of data to capture multiple aspects of the Russian invasion: he uses a narrowing line to represent the thinning of Napoleon's army, longitude and longitude coordinates, geographical features (rivers), and temperatures that coincide with the retreat of the army. Had Minard designed a map only including geographical positions and the size of the army at each point, it would have been enough to communicate that Napoleon's army was quickly tapering. However, Minard's decision to include temperatures coordinating with the army's retreat, as well as geographical features that the army encountered, allows us to understand the reasoning behind the loss of soldiers. The many dimensions of data included in this map allow us to see the army's campaign, from beginning to end, on a static map.

Figure 3-1 is an image of of Minard's graph. This version was printed in *Towards a Rational Historiography*, a book by Lionel Gossman, in the series *Transactions of the American Philosophical Society.*

Figure 3-1 Map of Napoleon’s Army

Figure 1



Napoleon’s March to Moscow The War of 1812

Charles Joseph Minard

This classic of Charles Joseph Minard (1781–1870), the French engineer, shows the terrible fate of Napoleon’s army in Russia. Described by E. J. Marey as seeming to defy the pen of the historian by its brutal eloquence, this combination of data map and time-series, drawn in 1861, portrays the devastating losses suffered in Napoleon’s Russian campaign of 1812. Beginning at the left on the Polish-Russian border near the Niemen River, the thick band shows the size of the army (422,000 men) as it invaded Russia in June 1812. The width of the band indicates the size of the army at each place on the map. In September, the army reached Moscow, which was by then sacked and deserted, with 100,000 men. The path of Napoleon’s retreat from Moscow is depicted by the darker, lower band, which is linked to a temperature scale and dates at the bottom of the chart. It was a bitterly cold winter, and many froze on the march out of Russia. As the graphic shows, the crossing of the Berezina River was a disaster, and the army finally struggled back into Poland with only 10,000 men remaining. Also shown are the movements of auxiliary troops, as they sought to protect the rear and the flank of the advancing army. Minard’s graphic tells a rich, coherent story with its multivariate data, far more enlightening than just a single number bouncing along over time. *Six* variables are plotted: the size of the army, its location on a two-dimensional surface, direction of the army’s movement, and temperature on various dates during the retreat from Moscow. It may well be the best statistical graphic ever drawn.

Edward R. Tufte, *The Visual Display of Quantitative Information* Graphics Press Box 430 Cheshire, Connecticut 06410

Source: *Towards a Rational Historiography*

In his book, *Beautiful Evidence*, Edward Tufte provides an analysis of Minard's work. He attributes the effectiveness of Minard's graph to the following:

- **Comparisons** The juxtaposition of the thick line representing the army invading Russia in June of 1812, with the thinner, darker line of the army's retreat, allows a viewer to gauge the magnitude of change and loss of soldiers (*Beautiful Evidence,* 126).
- **Causality** In order to explain a possible cause of the death of Napoleon's soldiers, Minard includes a scale that portrays the temperature at 8 points during the retreat from October 18th to December 7th. Tufte points out that while the temperatures themselves are of no "evidential value," Minard references sources that include "eye-witness accounts of the ghastly frozen soldiers" (*Beautiful Evidence*, 129).
- **Multivariate analysis** As already discussed, Minard included multiple relevant variables on the map (direction, temperature, size, etc.) to capture the complete story of the French army's invasion (*Beautiful Evidence,* 129).
- **Integrated evidence** Minard combines text, graphics, charts, and numbers. According to Tufte, "rarely is a distinction among the different modes of evidence useful for making sound inferences. It is all information after all" (*Beautiful Evidence*, 131).
- **Documentation** Detailed documentation regarding sources and authors lends credibility. Tufte explains that scales of measurement, labels, and sufficient data are all necessary. Minard's map includes three different scales, labels for geographic features and dates, and sufficient longitude and latitude coordinates--

all of which allow for a complete, credible visualization (*Beautiful Evidence*, 132).

- **Content** Tufte's flow map of Napoleon's army was ultimately so effective because of the content that he included in this flow graph. Tufte explains this as an important aspect of design: "this is a content-driven craft, to be evaluated by its success in assisting thinking about the substance" (*Beautiful Evidence*, 136)

Tufte's analysis explains that all of these aspects of Minard's map of Napoleon's army are what make it possible for the map to explain multiple aspects of the Russian campaign. These principles are ones that should be incorporated into the data collection process and design in order to create an effective visualization.

Chapter 4

Creating an Informative Visualization

Visualization Example

Just as Minard's graph encompasses all of the dimensions of data necessary to reveal the story of Napoleon's army, when creating a data visualization, we must be sure to include all available data to communicate a story to the end user. In order to build up to a complete and effective visualization, we must question what other data is relevant to show probable cause at each stage of the design. Tufte explains that visualizations often "rely solely on one type of data or stay at one level of analysis" (*Beautiful Evidence*, 78). He says that instead, the diagram or map should include "whatever evidence it takes to understand what is going on" (*Beautiful Evidence*, 78). In one of his other books, *The Visual Display of Quantitative Information*, Tufte discusses the use of relevant data and content as more important than the creative design of a visualization. When it comes to designing a visualization, Tufte explains that "allowing artistic-illustrators to control the design and content of statistical graphs is almost like allowing typographers to control the content, style, and editing of prose" (*Visual Display*, 87). While design is important, without significant data to use, the design is irrelevant.

The following example walks through the process of creating a visual display of data pertaining to birth places of Major League Baseball players playing in 2009, using a visualization tool called Many Eyes Beta, and a collection of other tools. Many Eyes Beta was created by a group of researchers in IBM's Collaborative User Experience research group (ManyEyes). The tool allows a user to upload data and choose a visualization type (world map, tree diagram, etc.) for that data. The different combinations and uses of these tools further emphasize our need for effective visualizations and visualization tools today.   There was not one specific tool that allowed all of the data in one set to be mapped together, or multiple data sets to be mapped simultaneously. *While this example demonstrates the importance of including all relevant data and information to tell a complete story, it also details the challenges and limitations of information design tools.*

Example: Birthplaces of MLB players in 2009

According to a report recently completed by The Institute for Diversity and Ethics in Sport (TIDES), "The total population of Major League players of color (39.6 percent) was comprised of Latino (27 percent), African-American (10.2 percent) or Asian (2.4 percent). MLB has been remarkably consistent in terms of the percentage of white players" (Lapchick). This statement made in the report is a statement of raw data, with no visual element, and no explanation explaining these differences in percentages present in the MLB.

Step 1: Map data (MLB players & birthplaces)

In order to delve deeper into these statistics, I researched data of not only the race of the players, but of the birthplaces of the players by country. One website, baseball-alamanac.com, has the data shown in Table 4-1 documented for the MLB baseball players of 2009:

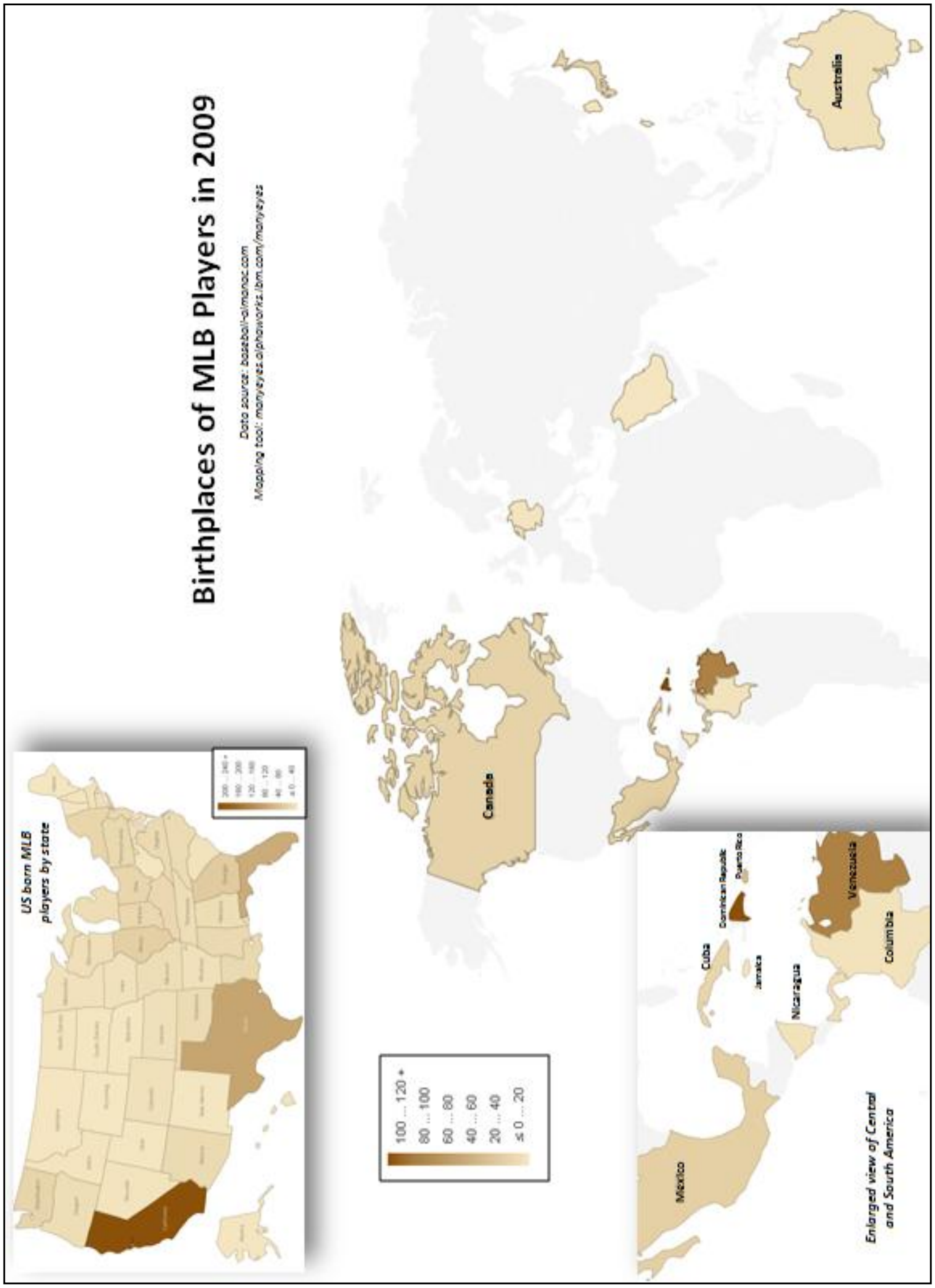
Table 4-1 MLB Players by State and Country Data

State	#	State	#	Country	#
Alabama	15	Nebraska	3	Aruba	1
Alaska	2	Nevada	6	Australia	6
Arizona	20	New Hampshire	4	Canada	19
Arkansas	9	New Jersey	20	Colombia	4
California	225	New Mexico	3	Cuba	16
Colorado	9	New York	26	Curacao	2
Connecticut	9	North Carolina	19	Dominican Republic	139
Delaware	2	North Dakota	4	Germany	2
Florida	88	Ohio	26	Jamaica	2
Georgia	42	Oklahoma	19	Japan	18
Hawaii	5	Oregon	15	Mexico	22
Idaho	4	Pennsylvania	22	Netherlands	1
Illinois	35	Rhode Island	4	Netherlands Antilles	4
Indiana	23	South Carolina	7	Nicaragua	4
Iowa	8	South Dakota	3	Panama	7
Kansas	11	Tennessee	15	Puerto Rico	39
Kentucky	23	Texas	99	Saudi Arabia	1
Louisiana	22	Utah	3	South Korea	2
Maine	1	Vermont	2	Taiwan	4
Maryland	12	Virginia	20	Venezuela	94
Massachusetts	5	Washington	27	West Germany	3
Michigan	16	Washington, D.C.	3		
Minnesota	8	West Virginia	3		
Mississippi	22	Wisconsin	12		
Missouri	15	Wyoming	1		
Montana	2				

Data Source: baseball-almanac.com

Because this data involves geographic locations, we know that the likely way to approach visualizing this data is with a map. Figure 4-1 is a visualization made using multiple data maps created with Many Eyes beta program.

Figure 4-1 MLB Players by Country and State



Source: data from baseball-almanac.com

Individual maps created with Many Eyes beta

Challenge:

As we go through the process of creating this visualization, we begin to discover that there are not many tools that map data just as we need it, they only map aspects of the necessary data. This is one of the obstacles that designers encounter while trying to create truly effective data visualizations—*the "perfect" tools simply don't exist*.

This tool is able to map either country-specific data on a map, or state-specific data on a separate map, but cannot map the data by U.S. state on the same map as the data listed by country. As a result, the map of the United States has been mapped separately, and the number of players from the U.S. (999 players) was excluded from the key used to read the global data—the United States has its own range of colors to represent state data.

Observation:

The U.S. map above shows that the states with the largest number of MLB player birthplaces are California, Florida, and Texas. The world as a whole has players clustered most densely in South and Central America, specifically in the Dominican Republic.

Step 2: Map next set of data (average yearly temperatures)

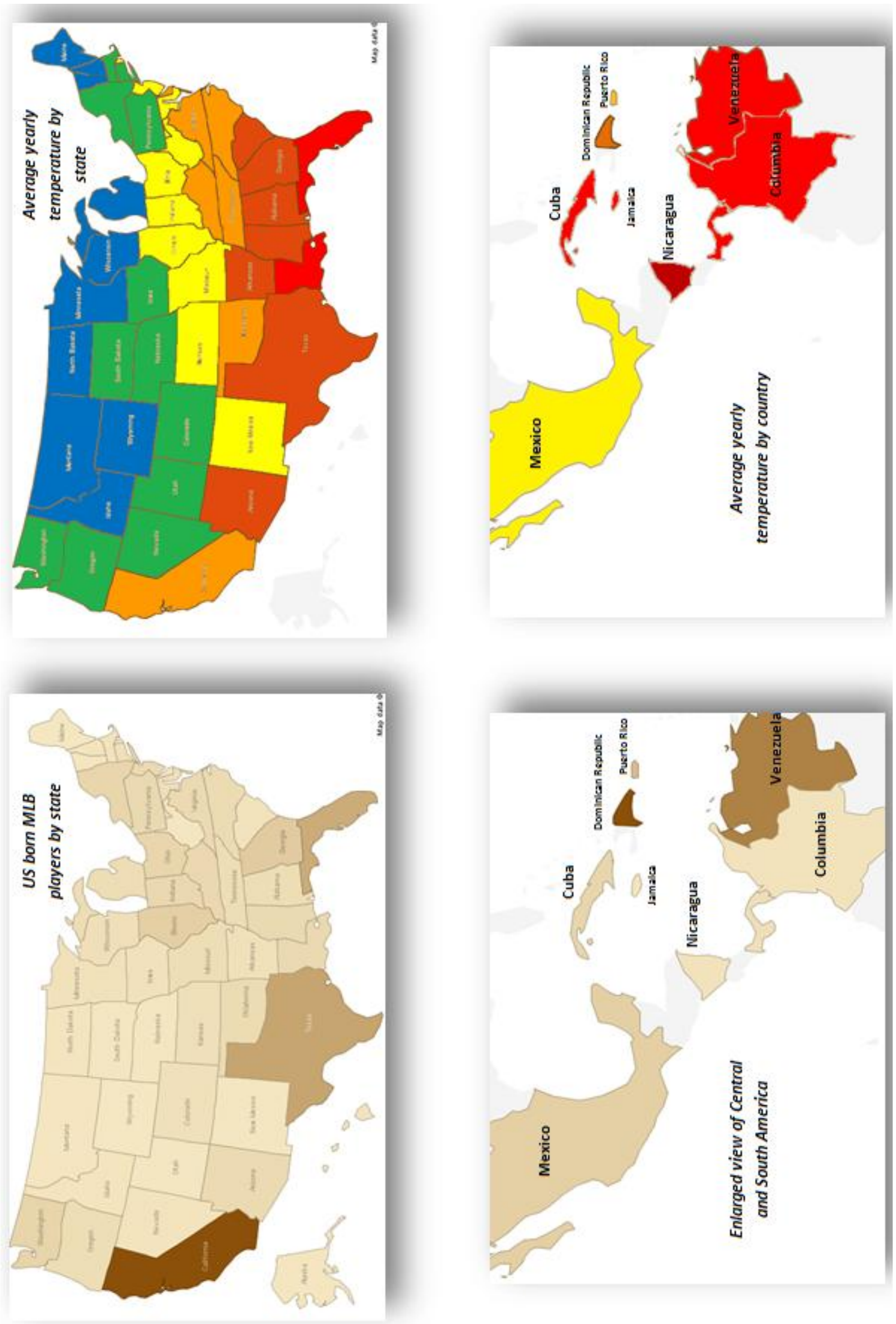
Research on these areas reveals similarities in their climates -- commonalities that could help to explain the above results. In an attempt to show that climate is one explanatory variable, average yearly temperature data in Table 4-2 is mapped in Figure 4-2 using Many Eyes beta (to map the data), and other applications to alter the colors:

Table 4-2 Average Yearly Temperature Data

Location	Average Yearly Temperature (°F)	Location	Average Yearly Temperature (°F)
Florida	70.73	New Jersey	52.65
Louisiana	66.39	Illinois	51.74
Nicaragua	78.63	West Virginia	51.72
Venezuela	77.92	Indiana	51.64
Jamaica	77.875	Ohio	50.68
Colombia	77.7931	Rhode Island	50.07
Cuba	77.571	Nevada	49.87
Dominican Republic	77.057	Connecticut	49.05
Puerto Rico	75.68	Nebraska	48.77
Mexico	69.95	Pennsylvania	48.77
Texas	64.83	Utah	48.64
Georgia	63.51	Oregon	48.41
Mississippi	63.35	Washington	48.26
Alabama	62.77	Massachusetts	47.86
South Carolina	62.42	Iowa	47.81
Arkansas	60.42	New York	45.35
Arizona	60.31	Colorado	45.15
Oklahoma	59.54	South Dakota	45.14
California	59.4	Michigan	44.41
North Carolina	59.01	Idaho	44.39
Tennessee	57.57	New Hampshire	43.8
Kentucky	55.59	Wisconsin	43.12
Delaware	55.27	Vermont	42.88
Virginia	55.11	Montana	42.74
Missouri	54.45	Wyoming	41.98
Kansas	54.25	Minnesota	41.16
Maryland	54.22	Maine	40.97
New Mexico	53.44	North Dakota	40.43

Source: US states data from esrl.noaa.gov Source: non-US data from weatherbase.com

Figure 4-2 Map of Average Yearly Temperature



Sources: data from baseball-almanac.com, esrl.noaa.gov, weatherbase.com

Individual maps created with Many Eyes beta

Challenge:

This specific tool is not able to map both numbers (average temperature and number of players on a single graphic). Ideally, the tool would display population by using height as another dimension of the state, while the color of the state would represent the average yearly temperature. Additionally, this tool does not allow us to use colors other than the scale of brown shades, so we cannot use differences in color to split the states and countries in half and map both sets of data on the map.

Observation:

As we look at the results, we are convinced of a strong association: areas with warmer temperatures are likely to produce more baseball players. It isn't hard to understand that in a place like Florida or the Dominican Republic, baseball can be played almost year round, whereas places further north have limited baseball seasons.

Despite this strong correlation, there is more to the story than average yearly temperature; there are exceptions to this association that cannot be ignored. While Haiti and the Dominican Republic are located on the same island, *why is it that baseball is the national sport in the Dominican Republic, while Haiti is a non baseball playing nation, where soccer prevails?*

In an article written by Dave Zirin in 2005, he explains the disproportionate number of Domincans playing in the MLB. While the statistics are from over five years ago, similar statistics still hold true today.

> “[In 2005 Major League Baseball was] increasingly dependent on talent born and bred in Latin America. Twenty-six percent of all players in the major leagues now hail from Latin America, including some of the game's most popular stars, like David Ortiz, Pedro Martinez and Sammy Sosa. Leading the way is the tiny nation of the Dominican Republic. Just five years ago there were sixty-six Dominican-born players on baseball's Opening Day rosters. This year, there were more than 100. This means roughly one out of every seven major league players was born in the Dominican Republic, by far the highest number from any country outside the United States” (Zirin).

Step 3: Add further relevant evidence (qualitative information)

How can this be explained? It is possible that the answer to this question cannot be found in the data we have already seen. To understand these differences, we need to incorporate other “modes of evidence” (Tufte, 130). A common challenge in the process of designing a data visualization is that there isn’t always data for everything. At some point, data runs out, and we are left with other “modes of evidence” that are more qualitative in nature, or sometimes no evidence at all (Tufte, 130).

Challenge:

When it comes to the history of baseball, not only is data on this topic rare and difficult to come across, but many of the stories and accounts on the topic differ. Of course, the most objective way to understand the movement of people between countries is to look at data pertaining to immigration and population changes. Unfortunately, there is no easy way to find data pertaining to movement of people between two specific countries, especially if it was not legal immigration. For a game like baseball, whose history of spreading throughout the world dates back to the late 1800s, the raw data cannot be easily found, and data that is found is often inaccurate.

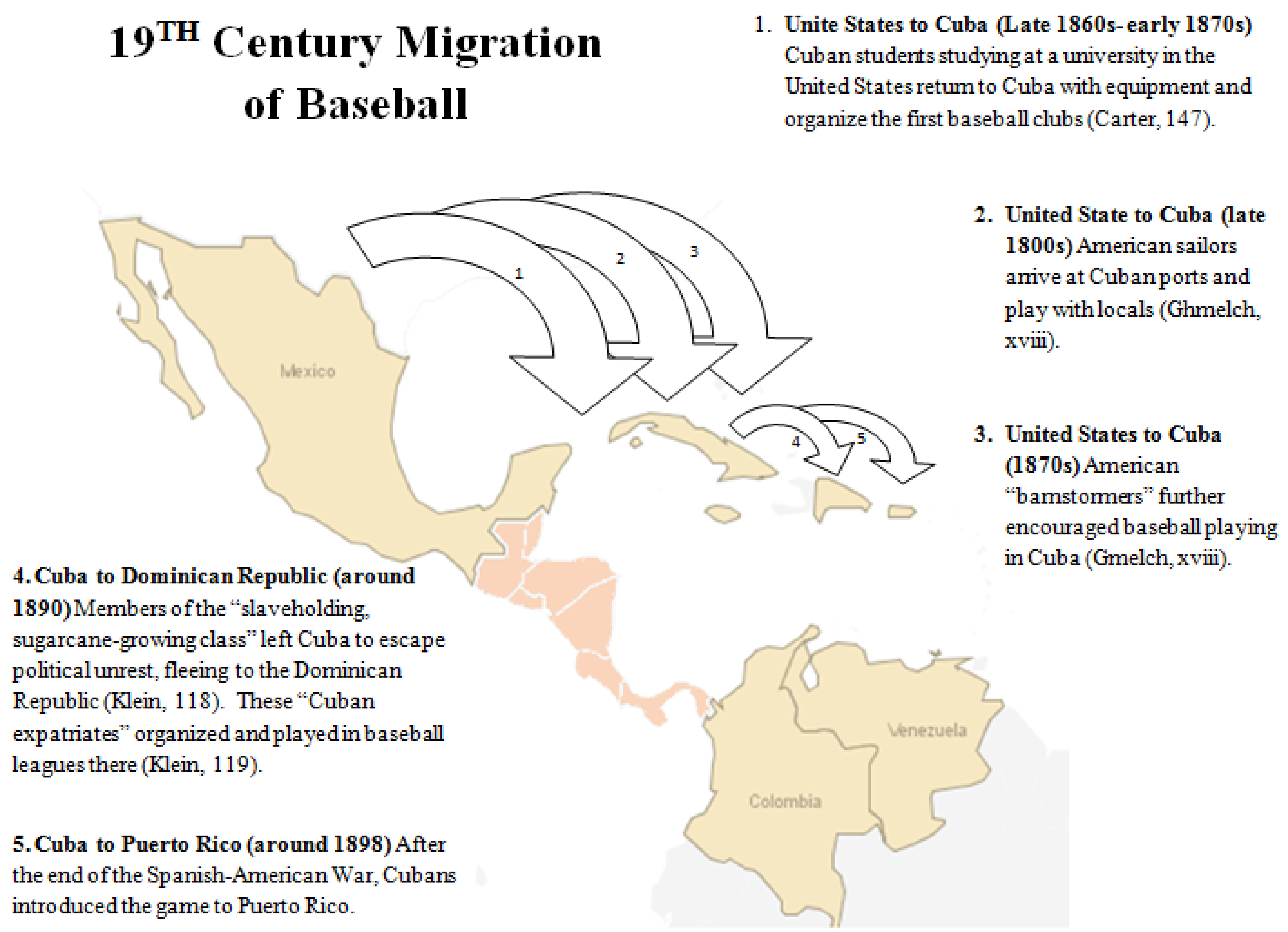
The game of baseball made its way to different countries by movement of people who introduced the game to others as they travelled. Sports and entertainment become a part of life that cannot be separated out from the culture of the people who play and watch. We look to stories, incidents, laws, and wars to understand the movement of people abroad who spread baseball to different countries along the way.

In a collection of essays put together in a book entitled *Baseball without Borders*, edited by George Gmelch, the history of baseball in 14 different countries is explained. As described in Thomas Carter's essay in *Baseball without Borders*, baseball first appeared in Cuba when students studying in the United States returned to the country in the 1860s (147). The book explains that "Folklore credits Nemiso Guillo for bringing the game to Cuba, when he returned from Springfield College in Mombile, Alabama, with a bat and a baseball in his trunk" (Gmelch, xviii). Also spreading the game to Cuba were American sailors that played the game with locals in Cuban ports, as well as American "barnstormers" (Gmelch xviii). The students that first brought the game of baseball to Cuba and organized the first clubs were wealthy men, and playing baseball soon became a sign of status, and games were elite social events (Carter, 148).

.

According to an essay by Allen Klein included in the book, Cubans were responsible for bringing the game of baseball to the Dominican Republic (*Baseball without Borders*,118). In the late 1800's, Cuba was experiencing political turmoil, and the fear of an end to slavery forced many Cuban slaveholders to leave Cuba for the Dominican Republic (Klein, 118). The game was further spread throughout the center of the country by other Cubans (Klein, 119). Besides bringing baseball to the Dominican Republic, Cubans also introduced the game to Puerto Rico after the Spanish-American War, around 1898. (Van Hyning and Otto, 161). All of this information is visually presented in Figure 4-3 below (the map is originally from Many Eyes beta, while the arrows and captions were added separately).

Figure 4-3 Nineteenth Century Migration of Baseball



19TH Century Migration of Baseball

1. **Unite States to Cuba (Late 1860s- early 1870s)** Cuban students studying at a university in the United States return to Cuba with equipment and organize the first baseball clubs (Carter, 147).

2. **United State to Cuba (late 1800s)** American sailors arrive at Cuban ports and play with locals (Ghmelch, xviii).

3. **United States to Cuba (1870s)** American "barnstormers" further encouraged baseball playing in Cuba (Gmelch, xviii).

4. Cuba to Dominican Republic (around 1890) Members of the "slaveholding, sugarcane-growing class" left Cuba to escape political unrest, fleeing to the Dominican Republic (Klein, 118). These "Cuban expatriates" organized and played in baseball leagues there (Klein, 119).

5. Cuba to Puerto Rico (around 1898) After the end of the Spanish-American War, Cubans introduced the game to Puerto Rico.

Individual map created with Many Eyes beta

Once all of this information comes together, we can start to understand the origins of baseball, why it is so popular in Cuba, and why it is played in the Dominican Republic, but not in Haiti. We see from graphing the data and information found in *Baseball without Borders* that a few different "waves" of people brought and played baseball to Cuba, and Cubans ultimately brought the game elsewhere when they fled their country.

While we now understand the reasons why baseball became the most popular sport in the Dominican Republic, this visualization still provokes further questions. It is not that the visualization is lacking. Instead, it answers the question we first asked, which leads to a desire for further understanding of the topic. For example, we might question why the Dominican Republic still produces more MLB players than any other country on the map outside of the United States. Popularity of a sport does not directly imply that the people who live there have the potential to become some of the best players in the world. What data do we need to answer this question? Data pertaining to education and baseball academies? Data pertaining to training levels and intensity at each age? This question can be answered if we continue to add more layers of relevant data to our collection, and question new observations.

This tool, like many other tools, lacked the capacity to create a visualization that reflected exactly the story we were trying to understand. Because this is often the case, it is important to remember that no matter what tool is being used, and no matter how much data we have, the ability to communicate information remains in the hands of the designer. The creator of the visualization must have a full understanding of the data in order to know *what* data to use, and *how* to graph it. Minard, for example, truly understood the story of Napoleon's army and the story he was trying to communicate when he was creating his flow map. This is why his map effectively communicated how and why Napoleon's soldiers didn't survive.

Chapter 5

Conclusion

The vast amount of data being collected and stored today is growing exponentially. This amount of data multiplies with every transaction that occurs, whether it's in the world of retail, transportation, government activity, or online commerce. While the common notion is that more data leads to more information and more well-informed decisions, many people don't understand the complexity of the data we collect. This complexity has implications on the use of all this data:

- It is difficult to extract knowledge from large sets of raw data: pattern recognition is nearly impossible and causality cannot be identified.
- Working through this complex data to try to find answers, when we have no tools to do so, can actually inhibit our decision making process.
- We often have an idea of what we want to know, but frequently don't know *what* data is relevant to look at, and what combinations of data will explain an event, statistic, or condition.

With all of the data being collected, we need effective visualizations to help us visually recognize what the data is telling us. While there are some effective data visualizations that have been created, they are rare, and often the result of luck or a highly creative/artistic designer. In other words, most of these effective visualizations have not been made with tools that can be used again for other data sets, and are often not accessible to others. Tools that are accessible are often low quality.

Given that extracting knowledge from complex data is difficult, and that effective visualizations are rare, the best we can do until an "ideal" tool is created, is follow a few steps when we start the design process.

1. Map the first set of data, in a way that is logical for the specific data type at hand.
2. Once this data shows a trend or pattern, question the observation. Ask, "why is this the case?"
3. Add another "layer" of relevant data that helps to show possible causality. Continue to question observations and add layers of relevant data.
4. At some point, there are no longer standard data sets available. If the story is still not told in its entirety, meaning that the original question remains unanswered, apply relevant qualitative data. This includes observations, written accounts, photographs, etc.
5. Incorporate these into the visualization-- just because it isn't a set of raw data, doesn't mean it should be disregarded. Again, as Tufte explained in *Beautiful Evidence*, there is no need to distinguish between quantitative and qualitative evidence, because "It is all information after all" (131).

The best data visualizations have been created by designers who are constantly questioning patterns they see, collecting and illustrating relevant data until it tells a story. However, by no means is a visualization expected to answer every question surrounding the topic at hand. In fact, effective visualizations are often the ones that provoke the most questions on the topic. Because the visualization has been mapped so well, the viewer is informed enough to further question causality.

The amount of data today is overwhelming to deal with, organize, and make use of. Even though there is no set of rules to create the perfect visualization, or a tool that maps data just as we need it, following those few rules can create a visualization effective enough to turn data to knowledge for an interested audience.

Works Cited

"A special report on managing information: Data, data everywhere." *The Economist* 25 Feb. 2010: n. pag. Web. 29 Mar. 2010. <http://www.economist.com/specialreports/displaystory.cfm?story_id=15557443>.

"Average Mean Temperature Index by month." *NOAA Earth System Research Laborator*. National Oceanic & Atmospheric Administration, n.d. Web. 16 Apr. 2010. <http://www.esrl.noaa.gov/psd/data/usclimate/tmp.state.19712000.climo>.

baseball-almanac.com. N.p., n.d. Web. 29 Mar. 2010.

Carter, Thomas. "Cuba: Community, Fans, and Ballplayers." *Baseball without Borders*. Ed. George Gmelch. Lincoln: University of Nebraska Press, 2006. 147-159. Print.

Gantz, John, and David Reinsel. "As the Economy Contracts, the Digital Universe Expands." *IDC-Multimedia White Paper*. EMC Corporation , May 2009. Web. 29 Mar. 2010. <http://www.emc.com/digital_universe >.

Gmelch, George. "Introduction: Around the Horn." *Baseball without Borders: The International Pastime*. Ed. George Gmelch. Lincoln: University of Nebraska Press, 2006. xiii-xxii. Print.

Hand, Randall. "Who is Coming to America." *Vizworld.com*. VizWorld, LLC, n.d. Web. 29 Mar. 2010. <http://www.vizworld.com/2009/05/who-is-coming-to-america/>.

Keim, Daniel, and Gennady Andrienko, et al. "Visual Analytics: Definition, Process, and Challenges." *Information Visualization: Human-Centered Issues and Perspectives*. .p.: Springer-Verlag , 2008. 154-175. Print.

Klein, Alan. "Dominican Republic: Forging an International Industry." *Baseball without Borders*. Ed. George Gmelch. Lincoln: University of Nebraska Press, 2006. 147-159. Print.

Lapchick, Richard, Alejandra Diaz-Calderon, and Derek McMechan. "The 2009 Racial and Gender Report Card: Major League Baseball." *TIDES: The Institute for Diversity and Ethics in Sport*. UCF College of Business Administration, n.d. Web. 29 Mar. 2010. <http://www.tidesport.org/RGRC/2009/2009_MLB_RGRC_PR_Final_rev.pdf>.

Many Eyes Beta. IBM, n.d. Web. 16 Apr. 2010. <http://manyeyes.alphaworks.ibm.com/manyeyes/>.

"Map of the Market." *Smartmoney.com*. Dow Jones & Company, Inc, n.d. Web. 29 Mar. 2010. <http://www.smartmoney.com/map-of-the-market/?hpadref=1>.

O'Grady, Jenn Visocky, and Ken Visocky O'Grady. *The Information Design Handbook*. Cincinnati: HOW Books, an imprint of F+W Publications, Inc., 2008. Print.

Towards a Rational Historiography. New Series ed. Vol. 79. : American Philosophical Society, 1989. 1-68. Transactions of the American Philosophical Society. Ser. 3. *JSTOR*. Web. 16 Apr. 2010.

Tufte, Edward. *Beautiful Evidence*. Cheshire: Graphics Press LLC, 2006. Print.

Tufte, Edward. *The Visual Display of Quantitative Information*. 2nd ed. 2001. Cheshire: Graphics Press, 2007. Print.

Van Hyning, Thomas E and Franklin Otto. "Puerto Rico: A Major League Steppingstone." *Baseball without Borders*. Ed. George Gmelch. Lincoln: University of Nebraska Press, 2006. 147-159. Print.

Wattenberg, Martin. "Map of the Market (1998)." *Bewitched.com*. N.p., n.d. Web. 29 Mar. 2010 <http://www.bewitched.com/marketmap.html>.

Wattenberg, Martin. "Visualizing the Stock Market." Conference on Human Factors in Computing Systems, 1999. Print.

Zirin, David. " How Baseball Strip-Mines the Dominican Republic." *CommonDreams.org*. N.p., 28 Oct. 2005. Web. 29 Mar. 2010. <http://www.commondreams.org/views05/1028-25.htm>.

ACADEMIC VITA of Fara Delitsky

Fara Delitsky
2122 Oliver Way
Merrick, NY 11566
ftd5000@psu.edu

Education: Bachelor of Science Degree in Management Information Systems,
Penn State University, Spring 2010
Schreyer Honors College
Thesis Title: Data Visualization
Thesis Supervisor: John Jordan

Related Experience:
Internship with Digital Tempus, Inc.
Summer 2009

Awards:
Dean’s List