Baby Names, Visualization, and Social Data Analysis

Martin Wattenberg

IBM Research

ABSTRACT

The NameVoyager, a web-based visualization of historical trends in baby naming, has proven remarkably popular. This paper discusses the display techniques used for smooth visual exploration of thousands of time series and the application's simple keyboard-based mechanism for filtering the view. We also describe design decisions behind the application and lessons learned in creating an application that makes do-it-yourself data mining fun. The prime lesson, it is hypothesized, is that a web-based information visualization is fruitfully viewed not as a tool but as part of an online social environment. In other words, to design a successful exploratory data analysis tool, one good strategy is to create a system that enables "social" data analysis.

CR Categories and Subject Descriptors: Design Study, Time-Varying Data Visualization, Human-Computer Interaction

1 INTRODUCTION

In February of 2005, my wife published her first book, a guide to American baby names called *The Baby Name Wizard* [Wattenberg 2005] which used a data-analysis approach to understanding name styles. To help call attention to the book, I created a web-based visualization applet, the NameVoyager [6], which lets users interactively explore name data—specifically, historical name popularity figures. The gambit succeeded and without any advertising it drew more than 500,000 site visits in the first two weeks after launch. Two months afterwards it is maintaining an average of 10,000 visits a day. Perhaps more important is that evidence suggests many people are engaging deeply with the visualization, spending considerable time and discovering for themselves facts and insights about name trends.

The broad popularity and effectiveness of the NameVoyager is especially interesting because it is, in essence, an exploratory data analysis application for a data set of 6,000 time series. In many situations, ranging from education to retirement planning, it is important to encourage users to interact with complex data sets. Understanding the factors that led a statistical exploration program to become a minor fad may shed light on the broader problem of encouraging users to engage in their own personal data mining expeditions.

An important piece of the puzzle is the public nature of a web-based application. As of April 2005, Google finds more than 11,000 references to the NameVoyager, many of which turn out to be lengthy sequences of comments on blogs and discussion sites. These comments provide clues as to how and why users are spending time with the applet. This data is in no way a scientific survey, but it does represent a large body of field usage information in which patterns emerge.

In hundreds of spontaneous comments, users are seen to be engaged in extended exploratory data analysis, identifying trends and anomalies and forming conjectures. These self-reports also lead to an observation about the NameVoyager: usage patterns are strongly social, and seem more closely related to those of online multiplayer games than to a conventional single-user statistical tool. Indeed, users seem to fall neatly into Richard Bartle's well-known categorization of online game players [Bartle 1996] as explorers, achievers, socializers, or killers. This stands in contrast to the traditional view of information visualization as a task-oriented problem-solving activity. We hypothesize that the broad popularity of the NameVoyager stems from features that not only give it a game-like sense of fun, but that make it especially suitable for "social" data analysis. We then suggest some general properties which may encourage this type of usage of visualizations.

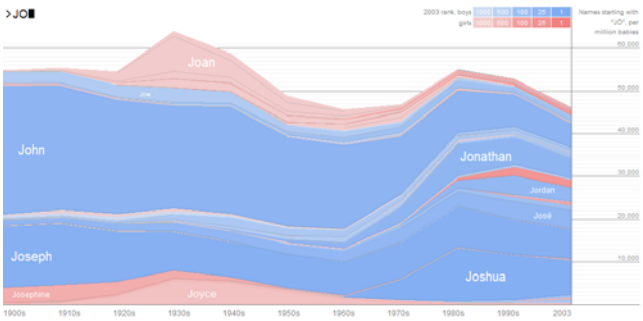

Figure 1. The NameVoyager

2 THE NAMEVOYAGER

2.1 Data

The NameVoyager is based on a data set, derived from public Social Security Administration (SSA) information that tracks baby name trends in the United States. For each decade since 1900, and each year since 2001, the SSA publishes separate lists of the most popular 1,000 boys and girls names, along with the exact number of babies given these names. These lists were downloaded, collated, cleaned, and normalized by the author of the *Baby Name Wizard* book to produce a data set containing popularity time series for roughly 6,000 distinct names.

These time series turn out to be meaningful in many ways. A graph of the popularity of a given name reveals a great deal about its overall cultural connotations and "feel," and names whose popularity is correlated over time tend to seem similar. (For more information, see *The Baby Name Wizard*.)

2.2 Visualization method

The method used to visualize the data is straightforward: given a set of name popularity time series, a set of stacked graphs is produced, as in Figure TK. The x-axis corresponds to date, and the y-axis to total frequency for all names currently in view, in terms of names per million babies. Each stripe represents a name, and the thickness of a stripe is proportional to its frequency of use at the given time step.

In keeping with American tradition, the stripes are colored either pink for girls or blue for boys. The brightness of each stripe varies according to the most recent popularity data, so that currently popular names are darkest and stand out the most. The idea behind this color scheme is twofold. First, names that are currently popular are more likely to be of interest to viewers—many people will probably want to know statistics on Jennifer, but few are looking for Cloyd. Second, the fact that the brightness varies provides a way to distinguish neighboring name stripes without relying on visually heavy borders.

2.3 Interaction

The NameVoyager follows Shneiderman's mantra of "overview first, zoom and filter, details on demand." (Shneiderman, 1996) When the applet starts, the viewer sees a set of stripes representing all names in the database. Filtering this data is achieved via an extremely simple mechanism. A user may type in letters, forming a prefix; the applet will then visualize data on only those names beginning with that prefix.

The applet reacts directly with each keystroke, so it is not necessary for the user to press return or to click a submit button. Not only does this instant interaction save the user some work, but it helps demonstrate how to mine the data. A user might not think that searching the data set by prefix would be interesting, but seeing the striking patterns for single letters like O or K could encourage further exploration. In addition, the applet moves smoothly between states, so that when a letter is typed, an animated transition helps preserve context.

Figure 1 shows an example: typing "JO" will yield a graph with prominent stripes for popular names such as John, Jonathan, Joseph, and Joyce, along with many thinner stripes for less popular names like Josette. Because the initial letters of a name contribute strongly to its sound, names that start with the same letters often have similar graph patterns. As a result, the simple mechanism of filtering by prefix is effective in highlighting interesting name trends. Typing "O" produces the graph in Figure 2, with an easily identifiable pattern of popularity of O names at the beginning and end of the 1900s, but a significant dip mid-century. Typing "LAT" highlights a trend in the African-American community in the 1970s, comprising names such as LaToya, LaTanya, LaTisha, and so on, as in Figure 3. Name stripes are ordered alphabetically on the screen from top to bottom to aid in identifying such prefix-based cultural clusters.

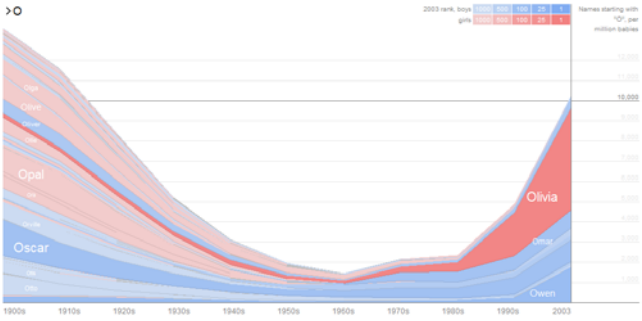

Figure 2. Names beginning with O

To learn details of a name, a viewer can use the mouse. Hovering over a name stripe will produce a pop-up box with numerical details for a given name at a given point in time. Clicking on a name stripe produces a graph of the popularity of that name alone.

This interaction technique may be compared to dynamic query systems such as starfield displays [Ahlberg & Shneiderman, 1994b] or TimeSearcher [Hochheiser & Shneiderman, 2004]. The keyboard interaction may be viewed as an alternative to the Alphaslider of [Ahlberg & Shneiderman, 1994a]. A key distinction between the graphical display of the NameVoyager and the visualization used in TimeSearcher, is the NameVoyager's use of a graph that sums all the time series. This technique seems likely to be of use in many other situations where summing is a natural operation, such as investigating product sales data.

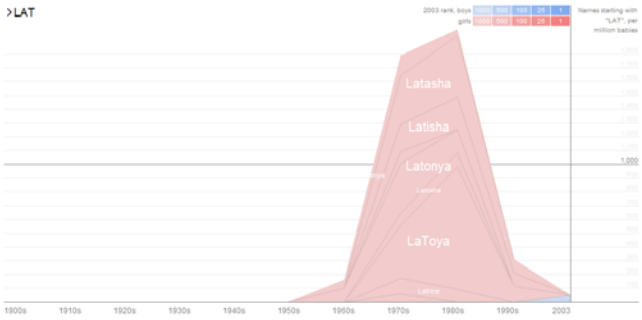

Figure 3. Names Beginning with LAT

2.4 Technical Implementation

The NameVoyager is a Java applet, written using JDK 1.1 so that it may run in a wide variety of browsers. All the name data (a 60K zip file) is loaded at startup and parsed into Java objects, so that it may be accessed rapidly.

To make the animated transitions run smoothly, not all 6,000 stripes are drawn; instead, a simple level-of-detail calculation is performed so that only stripes wider than 2 pixels are rendered to the screen. As a result, in practice the applet only draws about 200 or fewer stripes per frame. In an initial version of the applet, this culling of names caused prominent and irritating white stripes in the graph, where the background would "show through" the undrawn stripes. Replacing the white background with a neutral gray, halfway between the blue and pink tones of the name stripes, proved effective in removing this annoying effect.

3 Results

3.1 Traffic and Web Comments

As mentioned in the introduction, the NameVoyager received a remarkable number of visits within weeks of launch. The applet has been downloaded more than 900,000 times as of mid-April. It has also been extensively discussed on the web, in blogs, discussion forums, and similar sites. This web-based conversation is important for two reasons. First, it is further evidence that users were engaging deeply with the applet and of its widespread popularity. It is not uncommon to find discussions in the comments section of a blog that contain dozens of posts. Such long discussions occur even when it is not related to the topic of the web site—for instance, one of the most extensive sets of comments was found on a forum in a well-known libertarian magazine site.

The second reason these comments are important is that they provide a window into the user experience, and we quote them extensively below. Comments that have been posted to the web are clearly not a scientific sample, since only the most enthusiastic users will comment. Nonetheless, examining these comments

suggests some interesting hypotheses regarding the source of popularity of the NameVoyager.

3.2 The Target Audience and the Surprise Factor

As one might expect, there are many positive comments from people in the target audience for the visualization—users who have a strong interest in names and therefore might be interested in buying the book. Two examples (all quotes in this paper are taken from public web sites) illustrate this:

"*This is perfect, as baby names weigh heavily on my mind these days.*"

"*Useful fodder for historical fiction, too, if you're looking for typical names for a given age and time period*."

A surprising observation is that many people outside the target audience found themselves enjoying the applet. The surprise here is not the author's, but of the users themselves. Some sample quotes:

"*Surprisingly addictive*"

"*This rules, even though it's about baby names*"

"*Cool… by the way, I don't like babies or children*."

This "surprise factor" is a reason for optimism. It is common to want users to explore a set of data that they may have little inherent interest in. A good example is the amount of effort and money that American companies spend to encourage their employees to understand 401(k) plans. It is therefore worthwhile to look for clues to what made the NameVoyager appeal to people who profess inherent boredom with the topic of baby names.

4 Social Data Analysis

One of the most consistent themes seen in comments about the NameVoyager is that exploring the data has become a social activity. Many people mention group usage, for instance:

"*I happened upon it at work today and it affected the productivity of our entire department*."

Of special interest, however, is that when a group of people uses the applet, they often do so in a social, collaborative fashion, engaging in a dialogue as they mine the data. This is true even for loosely knit groups of web users. For example, here are some quotes from the comments section of one blog:

"*For a challenge, try finding a name that was popular at the beginning of the sample (around 1900), went out of style, then came back into vogue recently*"

Another person responds, "*Take a look at Grace, #18 in the 1900s, #13 in 2003, and down in the 200s and 300s during mid-century*".

A third writes "1900's comeback: Porter. Another one, with a mini-peak in trough: Caroline," and then adds, "More challenges: which is the steadiest popular name? Victor?" and "Which letter has gone down most consistently? W? Observation: Note the recent upsurge in Y; basically all due to Hispanic (and some Middle Eastern) names"

The original poster responds, "*You're right, W has gone most consistently down, although F is pretty close (if it weren't for Faith…)*"

These quotes, which are just a small part of the full exchange, illustrate two points. First, they show how a group of people is using the NameVoyager as a stimulus to conversation and repartee.

They also reveal an effective style of data analysis: this group of people is diving very deeply into the data set! They are setting each other pattern-finding challenges, noting outlying data points, and making guesses about causal relations. Each person seems to be building on the findings of the others, making the group as a whole extremely effective at mining the data—and having fun at the same time. Strange or surprising pieces of information serve as a kind of trophy for the finder. We refer to this process of data mining through dialogue, one-upmanship, and repartee as social data analysis. It is a version of exploratory data analysis that relies on social interaction as source of inspiration and motivation.

We hypothesize that viewing exploratory data analysis as a social activity may explain much of the reaction to the NameVoyager. Its popularity among people who do not find the data intrinsically interesting, for instance, could partly be due to the fact that these users are enjoying the social activity surrounding the applet. In the next sections, to understand better the social structure of this type of exploratory data analysis, we consider the different roles that users may play.

4.1 Roles in Social Data Analysis

As in any social system, it seems that people using the NameVoyager have a wide range of styles of interaction with each other. Comments on the web suggest that there are four distinct types of users. Interestingly, these types seem to align closely with a taxonomy developed by Richard Bartle [Bartle, 1996] in the context of an early class of online social environment called a MUD.

Bartle suggested that denizens of such online multiplayer environments typically fall into one of four types: achievers, socializers, explorers, and killers. Below we describe how each of these roles corresponds to a particular type of NameVoyager user.

4.2 Achievers

The context of the NameVoyager is a site designed to help expectant parents name their babies, so the stated "goal" of the applet is to find a good name. As described in Section 3.3, many people do exactly that:

"*We want something slightly retro, nice, and not too popular, and this visualization gives us all that.*"

Such users correspond to the Achievers in Bartle's classification: people who try to "achieve within the game's context."

4.3 Socializers

A second class of NameVoyager users consists of people whose main concern is their interactions with others, and who place their data exploration in a personal social context. These people, corresponding to Bartle's "Socializers," use screenshots and data from the applet as a catalyst for conversation and storytelling about themselves and their friends and family. A common sight on a blog is a person posting a screenshot of the graph of their

own name's popularity, or a friend's, with humorous comments. A typical quote of this type is:

"*Runes name doesnt show up at all… but my name has suddenly gotten popular … I HAD IT FIRST! heh*"

Often people talk about family members as they speculate about names, and see the changing popularity numbers as a kind of personal plotline:

"*my grandmother was named Coral and from what I can tell the name appeared out of nowhere in 1880…is it from a celebrity or something*?"

"*I got: 'No names starting with LINUS were in the top 1,000 names in any decade.' Translation: Your son's name will NEVER be cool*."

"*Woo! Emily (being me) was number 1 in 2003! go me*!"

Such relationship-oriented and storytelling behavior in the context of information visualization has been observed before in depictions of email archives [Viégas & Donath, 2002].

4.4 Explorers

Many users of the NameVoyager seemed to delight in unearthing odd names or unusual clusters. One person posted a screenshot created after typing "ETH": it showed the name Ethel being gradually and completely eclipsed by the trendy name Ethan. Another found the dramatic cluster of names starting with "LAT" (Latisha, Latoya, etc.) described in section 2.3. A well-known pundit used the NameVoyager to comment on the changing statistical distribution of names over the past century

These users were certainly not using the NameVoyager to name children, but rather were mining for nuggets of information that they could show to others as trophies of their expedition. They are directly analogous to Bartle's Explorers, people who want to learn as much as possible about the environment and who delight in discovering odd or unexpected features.

4.5 Killers

The last type in Bartle's taxonomy is the Killer, someone who enjoys imposing themselves on others and causing distress. One might think that there would be no Killers in the gentle world of baby names, but one would be wrong. A common theme is that certain users take pleasure in singling out names for ridicule. For these people the NameVoyager is a delightful source of fresh targets:

"*It is also damn entertaining to me (and the real reason why I am writing this) that I can type in Lexus and find that people actually name their kids Lexus*."

(Lest there be any doubt about the pugilistic nature of the author of this quote, note that it was found on a site called www.youandwhosearmy.com.)

"*Britney, Brittney, Britany, Brittany, Brittani, Britannie, Britni. Enough already.*"

This quote shows how Killers are often willing, as Bartle noted, to do a fair amount of exploration to uncover targets. Thus they too have a useful role to play in the setting of social data analysis.

5 Design Hypotheses for Social Data Analysis

The evidence above suggests that a large part of the power and popularity of the NameVoyager derives from the fact that it encourages a social style of data analysis. What leads users to approach data analysis as a social activity? Certain factors are obvious. The NameVoyager is easily accessible on the web so that a large group of people can see it. The interaction design, referred to on the web with such terms as "cool," "fantastic," and "whizzy," means that applet is something that people may be eager to associate themselves with, like a fashionable piece of clothing.

These factors, however, would apply to anything trendy on the web, whether a funny Flash animation or witty personality quiz. Are there any aspects of the NameVoyager's popularity that are specific to information visualization? We present three hypotheses below.

5.1 Common Ground

The first hypothesis is that some degree of common knowledge of the underlying data set is necessary for social data analysis to take place. The data set used in the NameVoyager pertains to names that are largely familiar to its users. Almost everyone in the U.S. has a sense of the connotations of a large number of names: although people may differ in their tastes, most Americans would agree on the likely ethnicity of a Rodrigo or a LaTanya, or the likely age of an Ethel versus a Heather. Similarly, many names relate to celebrities, pop culture icons, or historical figures.

This common ground is what makes conversation about the data possible and interesting. Some sample quotes:

"*Look what the Simpsons did to the name Bart.*"

"*Roosevelt has two spikes right about where you'd expect them.*"

"*I love the fact that Xander and Willow show up on the list in the 90s, thereby confirming the existence of Buffy fans as hardcore as me*."

The authors of these comments are sharing results of their data mining because they know that their readers will understand the cultural references.

The fact that the data is presented as a timeline over a standard period, 1900 to present, also provides a common context. A time period can serve as a shared grid on which users overlay personal and cultural knowlege.

So how might one design a visualization to emphasize common ground? In some cases, of course, the data set is fixed. But in many situations there is some flexibility. For example, in an educational setting a teacher might want to use a cartographic visualization of pollution data, but have a choice of possible locations. The common-ground principle suggests that it might be helpful to choose some area that many students are familiar with already, so they would want to share discoveries they make as they explore.

5.2 Personal Perspective

The next hypothesis is that, once common ground is established, it is helpful for each person to have a naturally unique perspective on the data. This unique perspective can serve as a kind of icebreaker in the conversation. It also means that, because each person is approaching the data in a different way, a group may collectively explore more pieces of the data. Evidence for

this hypothesis comes from [Ludford et al, 2004], which described a system that encouraged community participation by highlighting unique pieces of knowledge that an individual might have.

In the case of the NameVoyager, each person has one obvious point of entry: their own name. Names of relatives and close friends are also common conversation starters. Some sample comments illustrate this:

"*I was appalled to note that my name is now in the top 100, while it was about 700th when I was born…*"

"*My given name peaked in 1900 (or earlier) and has been on the slide ever since. Seems to be off the radar now. Elmer is more popular these days*!"

"*It also confirmed my suspicion that our eight-month-old son's name, Jackson, was rapidly gaining in popularity. Dangit, and we thought he would avoid having 4 kids in kindergarten with the same name*!!!"

"*We spent hours typing in the names of everyone we know*."

Applying the personal perspective principle may require some flexibility in the data set, but it may also be possible to guide people without modifying the data. For instance, imagine a visualization tool designed to help people understand different stock market investment strategies. There are several unique perspectives that people might take: for instance, looking at how their own company's stock has performed, or how the market as a whole did at significant points in their life. It is possible that the visualization could be tailored to bring out these perspectives.

5.3 Deep Pointers

The final hypothesis about how the NameVoyager encourages social data exploration is that it allows people to share the state of the visualization at any point in their explorations. Because the interaction model is so simple—just a matter of typing a few letters—it is very easy to guide other people to the same state. And indeed, many comments on the web are written in the imperative voice:

"*Take a look at K and see how it exploded in the last decade or two*"

"*Type in Adolph for example*"

"*You want some real fun, run 'Hillary'*"

What people are doing here, by hand, is creating a kind of "deep pointer" into the application—that is, making a reference into a particular state following interaction. The ability to do this may be critical to the conversation surrounding the NameVoyager, since it means that people can quickly share not just the results of their investigations, but the exact state of the visualization. Solitary, asychronous usage can easily become a shared experience. The ease of "showing off" discoveries also fosters a motivating sense of pride and competitiveness.

Thus a natural design principle might be that information visualization software ought to provide "deep pointers" if it is intended to support collaborative analysis. Such pointers could involve creating special URLs for later reference or some other technology. Note that this principle may impose some subtle constraints. Some graph layout algorithms, for example, involve random numbers, or depend on a long history of user manipulations; these algorithms might need to be modified to allow different people to see consistent views.

6 Conclusion and Future Directions

The NameVoyager is a visualization of baby name popularity data. This visualization uses keyboard-based interaction and smooth animation to allow users to explore a set of 6,000 time series. The applet has proven extremely popular, with hundreds of thousands of users in the space of two months. In addition, thousands of comments about the visualization have been written on the web.

This paper has explored the reaction to the NameVoyager, using these web comments as evidence. This methodology is somewhat unusual, but the sheer amount of online discussion of the NameVoyager provides a useful source of detailed descriptions from real users, and is a fruitful source of hypotheses about how and why the NameVoyager is effective.

The comments reveal that the NameVoyager is popular even among people who have no vested interest in looking for names—the applet is somehow appealing to people even when it is not solving an immediate problem. Moreover, users seem to be doing extensive data mining with the application, finding for themselves new patterns in the data. These facts make it all the more interesting to understand the NameVoyager's popularity, since it may serve as a model for other situations, especially in education, where the goal is to impart insight into a set of data that may not be immediately relevant to a user.

A central observation made from comments found on the web is that usage of the NameVoyager often involves a high degree of dialogue between users. It seems, at least in some cases, to be a social activity in which users discuss findings, set each other puzzles, and draw inspiration from one another. We believe this type of activity, which we term social data analysis, is the key to the efficacy and popularity of the applet. The collaborative, distributed nature means that people can join forces and share knowledge; the social aspect, because it is intrinsically enjoyable may explain the applet's appeal to users who state that they do not like babies or are not interested in baby names.

Understanding the patterns of social data analysis seems like a promising area for future research. This paper makes use of Bartle's taxonomy of players in multi-user online games as a starting point for understanding the different roles of people interacting with the NameVoyager. Identifying further frameworks and design principles related to social data analysis may be a fruitful avenue of investigation.

References

[1] Ahlberg, C. and Shneiderman, B. (1994) The alphaslider: a compact and rapid selector. *ACM Conference on Human Factors in Computing System*

[2] Ahlberg, C. and Shneiderman, B. (1994) Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays. *ACM Conference on Human Factors in Computing System*s.

[3] Bartle, R. (1996) *Players Who Suit MUDs*, Journal of MUD Research, 1:1. Available at http://www.mud.co.uk/richard/hcds.htm

[4] Hochheiser, H., Shneiderman, B., (2004) Dynamic Query Tools for Time Series Data Sets, Timebox Widgets for Interactive Exploration, *Information Visualization* 3, 1.

[5] Ludford, P., Cosley, D., Frankowski, D., and Terveen, L. (2004) Think different: increasing online community participation using uniqueness and group dissimilarity. *Proceedings of the SIGCHI*

[6] *NameVoyager*: http://babynamewizard.com/namevoyager/lnv0105.html

[7] Shneiderman, B. (1996) The eyes have it: A task by data type taxonomy for information visualizations, *Proc. 1996 IEEE, Visual Languages*.

[8] Viégas, F. and Donath, J. (2002) PostHistory: Visualizing Email Networks Over Time. *Sunbelt Social Network Conference XXII*. New Orleans, USA.

[9] Wattenberg, L. 2005. *The Baby Name Wizard*. New York: Broadway *conference on Human factors in computing systems.*