1 Introduction

Technology in modern society is ubiquitous, economically explosive, and culturally male. The stereotype of the young white male software developer is still the norm in information technology (IT) fields, even after decades of intervention to increase diversity in tech. According to the National Center for Women in Tech, in 2015 women made up 57% of the U.S. workforce but held just 25% of professional computing jobs; they likewise received more than half of all U.S. bachelor's degrees but only 17% of degrees in computer science, down from 37% in 1985 (National Center for Women & Information Technology, 2016). Despite the large body of research documenting the underrepresentation of women in IT, technology has not seen the same progress toward gender parity observed in the other STEM fields. This project uses interactive visualization to present this research to a general audience, highlighting the persistent gender imbalance within IT.

2 Prior Work

2.1 Women in IT

Modern computing is overwhelmingly male, according to a long, robust, and surprisingly consistent literature on gender in technology. Sanders (2005) summarizes research from the 1980s through the mid-2000s to show that women and girls regularly have less exposure to computers, especially programming; that they are less confident in their own computer skills, despite often being more proficient than their male peers; and that they are uncomfortable and uninterested in stereotypically macho tech culture.

Recent research focuses on the educational and professional consequences of those attitudes. Fewer women than men study computer science, and more women switch majors or drop out of the program (Cohoon, 2006). Similarly, fewer women enter careers in computing (Bartol and Aspray, 2006), and they change careers or leave the workforce at much higher rates than women in any other career path (Glass et al., 2013). For most researchers, this is a causal relationship: fewer women study computer science, so fewer women find and keep computer science jobs. These studies employ the metaphor of a “leaky pipeline” (Camp, 1997) and suggest that recruiting girls to computer science earlier in their education is the key to diversifying the IT workforce. More recent analyses, however, highlight the many different career paths that lead to IT work, arguing that a focus on the pipeline unduly ignores these alternate paths. Both perspectives agree that the lack of women in IT is a problem both for women and for IT as a field; the following sections summarize the two views.

2.1.1 The Leaky Pipeline

In the pipeline view, women are underrepresented in technical fields because they are neither encouraged to enter them nor given the support they need to remain. Studies investigating the leaky pipeline typically explore why girls never develop a serious interest in computer science (Barker et al., 2006), why female computer science majors frequently change departments (Katz et al., 2006), or specific techniques to improve female student retention (for instance, see the large literature on pair programming, e.g., Werner et al., 2005; Porter et al., 2013).

Gendered patterns in computer use appear well before high school: middle school boys are more likely to use computers for gaming, while girls use computers primary to communicate (Barker and Aspray, 2006). Increased exposure is correlated with greater confidence and interest in computers, so boys, with more varied computer experience, typically have much higher confidence in their own skills (Frantom et al., 2002). Both gaps, in experience and in confidence, only widen as children get older (Fitzpatrick and Hardman, 2000). These patterns have been documented since home computers were fairly uncommon (Jones and Clarke, 1995), but to my knowledge have not been investigated since the advent of ubiquitous mobile computing.

Formal computing education is also predominantly male, with fewer women than men studying information technology even as other STEM fields move toward gender parity. In 2004, just 16% of high schoolers who took the Advanced Placement (AP) computer science exam were female; Physics B, the next-most disparate AP exam, had more than twice as many female test-takers at 34% (Barker and Aspray, 2006). In a national study of computer science departments, Cohoon (2006) found that women made up 24% of undergraduate computer science majors, but comprised 32% of the computer science students who switched to another major. Programs with higher female enrollment saw less-gendered attrition rates, corresponding with higher overall graduation rates and a more collaborative culture seen in departments that actively recruit and mentor female students (Margolis and Fisher, 2002).

2.1.2 Problems with the Pipeline Metaphor

Unlike work on IT education, research on women in IT careers often rejects the notion of a pipeline and highlights instead the many possible paths to a career in technology. Bartol and Aspray (2006) argue that the distinction between education and career implicit in the pipeline metaphor is rarely valid for knowledge workers, and note that jobs in IT are being created and filled much faster than traditional students are earning IT degrees. Many workers are entering IT from other pathways rather than from a computer science degree program. Women in particular are more likely to transition into tech from another field, whether by taking on more technical tasks within an existing role (Von Hellens et al., 2001), by returning to school after some time in another career (Jesse, 2006), or by receiving vocational tech training (Chapple, 2006).

2.2 Narrative Data Visualization

The most memorable data visualizations do not simply present data; they tell stories. Most iconically, Charles Minard's famous map, pictured in Figure 1, evocatively depicts Napoleons Russian campaign of 1812 by mapping the army's journey and its size as 422,000 soldiers set out confidently, just 100,000 reach Moscow, and only a tenth of those make it home through the frigid winter weather. Many modern analysts, designers, journalists also use data to tell their stories. Established news sources like the New York Times and the BBC feature interactive, data-driven stories, as do dedicated data journalism outlets like FiveThirtyEight. Even personal correspondence can contain data stories, as with Giorgia Lupi and Stefanie Posavec's Dear Data project (Lupi and Posavec, 2016): the two sent each other weekly postcards visualizing an aspect of their lives that week, from the doors they opened in week 24 to the times they said goodbye in week 52. When they shared their project online, so many people connected with their data stories that their website now hosts mailing groups for other would-be data diarists.

Narrative visualizations combine the engagement of storytelling with the authority of data analysis, creating a distinct medium that blends the author's message with the viewer's interactions. Not all visualizations either seek or achieve this balance, as Lee et al. (2015) note; many visualizations are analytical tools that deliberately avoid narrative framing. But many visualizations are carefully crafted to illustrate the author's point. Segel and Heer (2010) identify four narrative structures commonly used to balance narrative and exploration in data stories: the checklist, which lets users explore a labeled storyboard of the interactive story; the interactive slideshow, which follows a traditional slideshow format but allows users to explore the data found on each slide; the drill-down story, which presents a summary graphic that illustrates the theme and lets users explore the underlying data; and the martini glass, which guides users through a tightly structured narrative before opening up into free exploration.

2.2.1 Visual Narratives as Activism

The combination of compelling story and convincing data makes visual narratives a natural tool for social or policy change. As early as the 1850s, John Snow, one of the fathers of modern epidemiology, used an annotated map to convince London officials to close a cholera-contaminated well (Johnson, 2006). In the same spirit, Radio (2016) published an interactive map of lead levels in 4,000 households in Flint, MI to demonstrate that the problem affected the entire city.

Visualizations that address smaller topics are also persuasive. Pandey et al. (2014) presented audiences with graphs or numerical statistics about a range of social issues; among viewers without strong prior opinions, significantly more were convinced by the simple graphics than by the numbers alone. However, Kim et al. (2010) demonstrate that visualizations without context—that is, visualizations that are not part of a narrative—do not show the same persuasive effects. They used graphical displays to encourage users to conserve energy by turning off their computers when not in use. One group saw a display of the time their computers sat idle, without no mention of energy conservation; the other saw an animation of a coral reef that got healthier as they reduced idle time. Although the coral visualization did not directly display their usage data, participants found it far more compelling than the standard idle tracker because it told a story by connecting their actions to environmental change.

Iconic representations are not the only way to provide context. Zuckerman and Gal-Oz (2014) evaluate the effectiveness of two different step trackers, one that includes a game and one that simply displays the user's daily steps as a bar chart filling toward their step goal. Their bar chart is visually quite similar to Kim et al's ineffective idle time tracker, and yet it was just as effective as the game in motivating participants to walk more. The bar chart in Zuckerman and Gal-Oz's study represents progress toward the user's own goals; it represents a simple but compelling narrative that motivates users to act.

3 Project Overview

This resource connects prior research on women's representation across IT, from high school to employment, to present an interactive visualization of the state of women in technology. It combines several publicly available datasets, spanning the education pipeline from high school to employment in computing, over the last several years, allowing viewers to explore not only the current state of gender diversity in computing but recent trends across the stages of the pipeline. Short narrative

Table 1: Summary of Data Sources Used

Data Source	**Years Included**	**Subset**	**Variables**
College Board AP Statistics	2010–2016	AP Computer Science Exam	Gender, Exam Score
CRA's Taulbee Survey	2010–2015	Graduate & Undergraduate degrees awarded	Gender, Program

explanations provide context for each visualization, suggesting features of the dataset that viewers may want to explore.

3.1 Data Sources

The data behind the visualization comes from three publicly available sources:

- The College Board's AP Computer Science test taker demographics
- The Taulbee Survey's annual reports of diversity in computing education
- The Bureau of Labor Statistics' employment report by occupation

Where feasible, I collected all data going back to 2010; the Bureau of Labor Statistics changed their reporting categories in 2011, so for consistency I excluded their 2010 data. Only the College Board had released 2016 data at the time of collection. Because not all groups contain the same years' data, visualizations including multiple groups use percentage breakdowns by gender, rather than absolute comparisons.

3.1.1 Data Cleaning

All three data sources are available either as spreadsheets or in PDF tables. The College Board spreadsheets include demographic information for all AP tests, broken out by gender and by exam score, so I extracted the data for the Computer Science exam and discarded the others. The Bureau of Labor Statistics similarly includes all occupations in their spreadsheet, so I extracted the data for all of the "Computer and Mathematical Occupations" as seen in the literature. The Taulbee Survey publishes its data as part of a PDF report; because the data is already aggregated (and consequently fairly small), I manually copied the data from their PDFs into a CSV file.

Both the College Board and the Taulbee Survey give the aggregated counts of male and female students for each relevant category, so the only further processing required was to format each row consistently. The Bureau of Labor Statistics, on the other hand, provides a total count of employees for each category, rounded to the nearest thousand, along with the percentage of women employed in each category. They do not provide diversity statistics for categories with fewer than 50,000 total employees, so I excluded these categories from the visualization. For the categories included, I used the percentages provided to calculate an approximate female employee count, to correspond with the data for educational stages.

3.2 Users

My project is intended for a general audience, not simply for womens advocates within technology. It will consequently require a strong narrative thread, since viewers may not be particularly interested in the topic and may not know how to interpred analytical visualizations. To aid in designing for a primarily non-specialist audience, I will be designing the narrative, visualization, and interactions with four personas in mind:

- Gabriella Griffin, 19, undergraduate psychology major
- Ryan Mitchell, 28, software engineer
- Laurie Woods, 34, front-end developer/womens coding meetup organizer
- Monica Pierce, 47, accountant/mother of two (Ian, 18; Julie, 16)

Gabriella, profiled in the sidebar on the next page, is an excellent example of the projects target audience: a young woman comfortable using computers, with some programming experience but no exposure to career opportunities in IT.

3.3 Tasks

The narrative portion of the project will guide users through some of the key insights in the data; a full exploratory analysis of the data is required to determine what questions the narrative will address. Since the visualization will have four main components (overview, high school, higher education, career), the narrative should introduce each of these areas. This could be accomplished with any of Segel and Heer (2010) narrative structures, depending on the relationship between the final dataset and the users needs.

The exploratory portion of the project is shaped by the users needs much more directly. It should present the data clearly enough to let them explore their own questions, such as:

- What kinds of technical careers are open to women? What kinds are still relatively closed?
- When do women get involved in tech? Is a computer science degree required?
- Is there a stage when more women seem to leave IT fields?

- Has the focus on recruiting and retaining women in computer science been effective?
- Why is it so important to get women involved in the first place?

This range of questions requires at least three types of data views:

1. An overview, showing women entering or leaving computer science at each stage of the traditional IT pipeline.
2. Detailed views, providing more details or supplementary information about each stage.
3. A time series view, indicating the trends over time for at least one data series.

Users should be able to navigate easily between the three types of views, as well as between the detailed views.

3.4 Project Development

3.4.1 Process and Tools

After fleshing out secondary personas and a few use cases, the development process will begin with exploratory data analysis both to clean the data and to shape a narrative for the final visualization. This exploratory analysis will primarily be done using the R software package, with additional tools like Google Refine to speed data cleaning.

Next, low-fidelity storyboards and paper prototypes will allow me to quickly collect feedback on designs for the final visualization. The visualization itself will be created in Javascript, using the D3.js visualization library.

3.4.2 Evaluation

Because this project focuses on design and development rather than user testing, there will not be a formal usability evaluation of the final visualization. However, I will be designing with usability principles in mind, and I intend to solicit informal feedback to incorporate into the final visualization.

References

Barker, L. J. and Aspray, W. (2006). The state of research on girls and it. In Cohoon, J. M. and Aspray, W., editors, *Women and Information Technology*, chapter 1, pages 3–54. MIT Press, Cambridge, MA.

Barker, L. J., Snow, E., Garvin-Doxas, K., and Weston, T. (2006). Recruiting middle school girls into it: Data on girls' perceptions and experiences from a mixed-demographic group. In Cohoon, J. M. and Aspray, W., editors, *Women and Information Technology*, chapter 4, pages 115–136. MIT Press, Cambridge, MA.

Bartol, K. M. and Aspray, W. (2006). The transition from the academic world to the it workplace. In *Women and Information Technology*, chapter 13, pages 377–419. MIT Press, Cambridge, MA.

Camp, T. (1997). The incredible shrinking pipeline. *Communications of the ACM*, 40(10):103–110.

Chapple, K. (2006). Foot in the door, mouse in hand: Low-income women, short-term job training programs, and it careers. In *Women and Information Technology*, chapter 15, pages 439–470. MIT Press, Cambridge, MA.

Cohoon, J. M. (2006). Just get over it or just get on with it: Retaining women in undergraduate computing. In Cohoon, J. M. and Aspray, W., editors, *Women and Information Technology*, chapter 7, pages 205–238. MIT Press, Cambridge, MA.

Fitzpatrick, H. and Hardman, M. (2000). Mediated activity in the primary classroom: Girls, boys and computers. *Learning and Instruction*, 10(5):431–446.

Frantom, C. G., Green, K. E., and Hoffman, E. R. (2002). Measure development: The children's attitudes toward technology scale (CATS). *Journal of Educational Computing Research*, 26(3):249–263.

Glass, J. L., Sassler, S., Levitte, Y., and Michelmore, K. M. (2013). What's so special about STEM? a comparison of women's retention in STEM and professional occupations. *Social Forces*, 92(2):723–756.

Jesse, J. K. (2006). The poverty of the pipeline metaphor: The AAAS/CPST study of nontraditional pathwasy into IT/CS education and the workforce. In *Women and Information Technology*, chapter 8, pages 239–278. MIT Press, Cambridge, MA.

Johnson, S. (2006). *The Ghost Map: The Story of London's Most Terrifying Epidemic, and How it Changed Science, Cities, and the Modern World*. Riverhead Books, New York.

Jones, T. and Clarke, V. A. (1995). Diversity as a determinant of attitudes: A possible explanation of the apparent advantage of single-sex settings. *Journal of Educational Computing Research*, 12(1):51–64.

Katz, S., Aronis, J., Wilson, C., Allbritton, D., and Soffa, M. L. (2006). Traversing the undergraduate curriculum in computer science: Where do students stumble? In Cohoon, J. M. and Aspray, W., editors, *Women and Information Technology*, chapter 12. MIT Press, Cambridge, MA.

Kim, T., Hong, H., and Magerko, B. (2010). Designing for persuasion: Toward ambient eco-visualization for awareness. In Ploug, T., Hasle, P., and Oinas-Kukkonen, H., editors, *Persuasive Technology: 5th International Conference, PERSUASIVE 2010, Copenhagen, Denmark, June 7-10, 2010. Proceedings*, pages 106–116, Berlin, Heidelberg. Springer Berlin Heidelberg.

Lee, B., Riche, N. H., Isenberg, P., and Carpendale, S. (2015). More than telling a story: Transforming data into visually shared stories. *IEEE Computer Graphics and Applications*, 35(5):84–90.

Lupi, G. and Posavec, S. (2016). *Dear Data*. Princeton Architectural Press, Princeton, NJ.

Margolis, J. and Fisher, A. (2002). *Unlocking the Clubhouse: Women in Computing*. MIT Press, Cambridge, MA.

National Center for Women & Information Technology (2016). Women in IT: By the numbers. Online at http://www.ncwit.org/bythenumbers.

Pandey, A. V., Manivannan, A., Nov, O., Satterthwaite, M., and Bertini, E. (2014). The persuasive power of data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2211–2220.

Porter, L., Guzdial, M., McDowell, C., and Simon, B. (2013). Success in introductory programming: What works? *Communications of the ACM*, 56(8):34–36.

Radio, M. P. (2016). Map: Take a closer look at flint lead testing results. Web. Accessed December 1, 2016.

Sanders, J. (2005). Gender and technology in education: A research review. *Seattle: Center for Gender Equity*, 20:2006.

Segel, E. and Heer, J. (2010). Narrative visualization: Telling stories with data. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1139–1148.

Von Hellens, L. A., Nielsen, S. H., and Trauth, E. M. (2001). Breaking and entering the male domain. women in the IT industry. In *Proceedings of the 2001 ACM SIGCPR conference on Computer personnel research*, pages 116–120. ACM.

Werner, L., Hanks, B., McDowell, H., Bullock, H., and Fernald, J. (2005). Want to increase retention of your female students? *Computing Research News*, 17(2):2.

Zuckerman, O. and Gal-Oz, A. (2014). Deconstructing gamification: Evaluating the effectiveness of continuous measurement, virtual rewards, and social comparison for promoting physical activity. *Personal and Ubiquitous Computing*, 18(7):1705–1719.