

CRAFTING A DATA VISUALIZATION COURSE FOR THE TECH INDUSTRY*

Ajay Bandi and Abdelaziz Fellah
School of Computer Science and Information Systems
Northwest Missouri State University
Maryville, MO 64468
ajay@nwmissouri.edu
afellah@nwmissouri.edu

ABSTRACT

This paper presents a new data visualization course for computer science and data science majors. We designed and developed the course with the collaboration of IT and data services professionals from the industry. We identified the need for students to acquire knowledge of data science concepts. The course emphasizes on both the technical and soft skills needs of the industry. We offered and implemented the course at Northwest Missouri State by aligning the School's learning outcomes with those of the tech industry. Also, we compared the top three visualization tools in the Gartner BI magic quadrant of 2016.

INTRODUCTION

Data visualization - the visual representation of information in a graphical format - is a broad area of study at the crossroads of a variety of fields and disciplines, including, computer science (i.e., machine learning, data mining), mathematics, engineering, sciences, and business intelligence. Data visualization covers a spectrum of visualization techniques and tools for working with different data sets which may exist within or across various lines of evidence.

Data visualization has gained its popularity in recent years and a large number of colleges and Universities have integrated a course in data visualization in their curricula under different names and across multiple disciplines. Hutchings and Squire [4]

* Copyright © 2017 by the Consortium for Computing Sciences in Colleges. Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the CCSC copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Consortium for Computing Sciences in Colleges. To copy otherwise, or to republish, requires a fee and/or specific permission.

introduced the VisMap, a web-based tool for the undergraduate course titled "Data science and visualization." The VisMap exposes students to various data science stages with an easy way to manipulate and visualize data. Robbins, Senseman, and Pate [8] taught data visualization for biologists using MATLAB. Urness [10] introduced a unit on data visualization in the computer graphics course for undergraduate students, which covers the basic applications of graphs and colormaps. Other disciplines focus on the aspects of data acquisition, processing, analyzing, and gaining insights from the visual information.

In computer science major, data visualization has become one of the most exciting courses among students at Northwest Missouri State University. The course is a specialized elective and a promising solution for the digital information age in today's industry. More than 2000 data visualization courses are being offered across campuses [3]. Data visualization is the graphical representation of information, the transformation of data into an image (i.e., chart, graph), the narrative and communication of data, and the analysis and the interpretation of data from the invisibility to visibility to gain insight into the data. In computer science, data visualization course may focus on good design practice for visualization, visual technical computing skills, and tools, algorithms and methods for display of data from a variety of fields, including 3D models, dissemination formats, computer graphics, image processing. Data visualization has its trends and challenges.

Data Visualization in the Visual Computing Era

Data visualization has emerged as a prominent subfield in computer science, enlightening traditional computer graphics era, and has become an active area of research. Importantly, data visualization provides a platform for software development visualization tools and has become a data science industry of its own.

Data Science is a growing field that used in various domains. It is all about the activities of preparing, processing, wrangling, exploring data, drawing conclusions, and making decisions. Data visualization uses the findings made by a data analytics team and translates them into a pictorial or graphical representation, which helps computer scientists to make inferences about the structured and unstructured data.

Industries require a set of skills and technical knowledge from computer science graduates who have given different roles, responsibilities, and titles, systems engineers, network engineers, UX designers, requirements analysts, software architects, software developers, programming analysts, and quality analysts. In these roles, for example, computer analysts have to keep track of the data throughout the software development life cycle. That is, the number of architecture or design violations [1], the number of functional and non-functional requirements, the number of change requests and issues resolved, the number of test cases passed or failed, and the time spent in developing software modules. In addition to the software industry, computer and data science majors work in several other domains requiring data science skills, such as health care, defense, banking, biotechnology, consulting firms, DNA sequencing labs, epidemiology research, internet search engines, and satellite imagery research [6, 7].

The School of Computer Science and Information Systems at Northwest Missouri State University is no exception to offering a data visualization course. However, the difference is that the course is not only aligned with the goals of the school but also shifted towards the industry needs in collaboration with the employers.

DATA VISUALIZATION IN THE VISUAL COMPUTING ERA

Data visualization has emerged as a prominent subfield in computer science, enlightening traditional computer graphics era, and has become an active area of research. Importantly, data visualization provides a platform for software development visualization tools and has become a data science industry of its own.

Data Science is a growing field that has been used in various domains. It is all about the activities of preparing, processing, wrangling, exploring data, drawing conclusions, and making decisions. Data visualization uses the conclusions made by a data analytics team, and translates them into a pictorial or graphical representation, which helps computer scientists to make inferences about the structured and unstructured data.

Industries require a set of skills and technical knowledge from computer science graduates who have been given different roles, responsibilities, and titles, systems engineers, network engineers, UX designers, requirements analysts, software architects, software developers, programming analysts, and quality analysts. In these roles, for example, computer analysts have to keep track of the data throughout the software development life cycle. That is, the number of architecture or design violations [1], the number of functional and non-functional requirements, the number of change requests and issues resolved, the number of test cases passed or failed, and the time spent in developing software modules. In addition to the software industry, computer and data science majors work in several other domains requiring data science skills, such as health care, defense, banking, biotechnology, consulting firms, DNA sequencing labs, epidemiology research, internet search engines, and satellite imagery research [6, 7].

The School of Computer Science and Information Systems at Northwest Missouri State University, is no exception to offering a data visualization course. However, the difference is that the course is not only aligned with the goals of the school but it is shifted towards the industry needs in collaboration with the employers.

Professional Advisory Team

The Professional Advisory Team (PAT) is a group of more than 40 professionals from various industries that participate and review annually Northwest's computer science curriculum. PAT serves in an advisory capacity to the school, make recommendations, update the school about new technology and ensure the curriculum at Northwest is current with technology and industry trends. It is a one full day meeting. Usually, it starts with listening and panel discussion as shown in Figure 1. Then, faculty members and PAT members are divided into working groups.



Figure 1. Professional Advisory Team meeting

<http://www.nwmissouri.edu/academics/undergraduate/majors/computer-science.htm>

DATA VISUALIZATION COURSE AT NORTHWEST MISSOURI STATE UNIVERSITY

In 2016, we have introduced data visualization as an elective course for both computer and data sciences undergraduate and Applied Computer Science graduate students. Before developing this course, we solicited inputs from our Professional Advisory Team (PAT) to determine and identify the appropriate technical topics and tools that should be covered in the course to reflect the needs of the workplace and industry. We contacted few recruiting/advertising tech agencies, in particular, Data Services Professionals from VML, Inc. and Garmin Ltd., located in Kansas City, MO. VML is a full-service global digital marketing and advertising agency with offices across six continents around the globe. Based on our continued discussions within the school, the best practices and advice of the Professional Advisory Team and the recommendations of VML and Garmin, we collectively have developed a data visualization course that have encompassed a broad range of basic visualization concepts and skills using different visualization tools and technologies that are expected from college graduates entering the tech industry. There is no required text book for this course. The course learning outcomes cover step-by-step the following abilities:

Understanding basic foundation of data visualization, identifying and applying data visualization techniques

Students were introduced to a wealth of different visualization techniques, ranging from scientific visualization, information visualization to visual analytics, and info graphics for broad and complex data. Besides, several visualization algorithms are covered assuming that students have no prerequisite in graphics and with no experience in graphics programming.

Generating word clouds to visualize qualitative data

Students were introduced to the analysis of qualitative data and asked to produce word clouds using any of the available online tools. The generated word cloud may contain any number of words that are related to a particular domain/area and occupies the suitable shape.

Cleaning data sets and generating various charts or graphs using Microsoft Excel

In this course, students were provided with some data sets in Microsoft Excel. During the initial days of the course, we had few assessments associated with Excel which include cleaning the data set, using formulas to calculate the required values, color coding the Excel based on the values. Finally, students were asked to generate various kinds of basic charts such as bar chart, column chart, radar chart, line graph, multi-line graph, scatter plot chart, pie chart, area chart, box whisker chart. They also have assignments on generating a combo chart which is the combination of any of the charts mentioned above. Our PAT members' suggested teaching MS Excel in this course as it has the core and essential features for visualizing data.

Introducing JavaScript libraries such as D3.js, Leaflet.js and Chart.js to visualize data in web pages and mobile devices

By using Excel to generate the basic charts, students got familiar with the course, and it is perfect timing to introduce some JavaScript libraries that are used to create charts and maps in web pages and mobile devices. Students were taught some of the leading JavaScript libraries to generate the required charts and graphs. Students used the Java programming language and NetBeans IDE to create charts using these libraries where the results are visualized in different browsers (i.e., Google Chrome). With the knowledge of JavaScript, students learned how a programming language would be used to visualize data in web pages and mobile devices.

Generating charts and maps using interactive dashboards to represent data in Tableau for decision making

Tableau is the interactive data visualization tool that we use in this course. Students download and install the free student version of Tableau. Students were taught to write SQL queries in Tableau to connect to external databases. We have created multiple assignments where students need to generate various kinds of charts such as bubble chart, tree map, histogram, box plot, bullet chart, radial stacked bar chart, and then various kinds of maps such as choropleth map, filled map, symbol map where Tableau identifies the location automatically in the data set. After generating all the required maps and charts, students combine them into an interactive dashboard which helps in creating some meaningful stories to make crucial decisions.

Enhancing written and oral communication skills

Each student was asked to select and explore a visualization tool and present it in the class. Students were also instructed to prepare a worksheet (a detailed step by step procedure on how to install a tool, connect data source, and visualize an appropriate chart for a given data set) where the rest of the students were asked to work on it. Some of the interesting tools that students presented include Microsoft Power BI, Raw, Plotly, Infogram, QlikView, QlikSense and some of the charts they generated such as Dendrogram, Sunburst chart, Bullet chart, Sankey diagram, Packed Bubble chart, Venn diagram, Heat map, Network diagram, etc. (See Appendix B). This helped students to get familiar with numerous tools and technologies and enhance their written and oral communication skills.

Enhancing teamwork

We have also included a final project where each team of 6 students worked together on a common theme. The first step involved is to select a huge data set which contains several hundred to thousands of records. In the second phase, after selecting the dataset, each of the team members will work collaboratively with other team members to form their goals related to the dataset. This would convey hidden stories from the raw data set using the tools most appropriate charts learned in the classroom. In the final step, the team has to present their work to the entire class.

STUDENT ASSESSMENTS

In our tool-based data visualization course, students completed six assignments and 14 worksheets in which each assignment is derived from tools and technologies discussed in class. Also, every student selected a visualization tool to demonstrate a 15-minute presentation to the class with a sample data set. Every student's presentation includes a PowerPoint presentation with the specific goals from the data using the following format (i) a worksheet explaining the installation of the visualization tool, (ii) the detailed steps for data source connection, (iii) the steps needed to represent data in an appropriate chart, (iv) and the results in a pdf, jpeg or URL format that they share with all students. This individual presentation helps assess each student's adaptability to a new tool or technology. It also helps to assess oral and written communication skills as well. Another in-class activity is identifying a "wrong" visualization chart or diagram and then re-visualizing the same data with an effective chart.

Students also complete a team-prepared term project. The teams consist of 6 students, and the project has three completion milestones. First, students need to select the data set and write a proposal document outlining at least six goals and the respective charts, the steps to clean the data, and identifying the visualization tools and technologies chosen for the project. By the end of this milestone, teams will be finishing the implementation of the project. The final milestone will be the team presentation and submission of project artifacts (raw data, cleaned data, and a detailed document of how to use the tools, and the presentation slides).

An example of a student work is presented in Figures 2 and 3. The steam graph shows the visualization of movies which have collected the highest box office in particular years and their respective months. However, the graph fails to show the right display that reflects which movie has hit the highest in the box office as shown in Figure 2a. It is not the right choice, and it is relevant to consider the tree map to be used in the individual application where the size of the rectangles reflects how the movies performed at the box office (See Figure 2b). Figure 3a, the Internet Marketing Tree does not visualize the correlation among attributes. Branding, Content and Online Advertising are not connected to the primary trunk. This leads to misinterpreting the data. Using a Clustered Dendrogram as shown in Figure 3b, we can illustrate the arrangement of the type of marketing, and it's sub categories that are easily understandable and ordered correctly. We can still increase the visualization by adding colors using a different tool other than RAW.

Gibbons [2] showed evidence that inviting guest speakers to the classroom would be helpful and worth including in class time. By following the Kamoun and Selim [5] framework of attracting IT professionals to the classroom, we accommodated two guest lectures during class time by bringing in the IT and data services professionals from VML, Inc., and Garmin.

The guest lecturers discussed data analysis and visualization techniques, using various charts of their projects, and the challenges they face as data experts. In our post-course survey, 96% of the students agreed our guest lecturers benefitted students by identifying new trends in the IT industry. We investigated several tools, and our findings are presented by comparing the top three tools from the Gartner BI magic quadrant for 2016 [9] (See Appendix A).

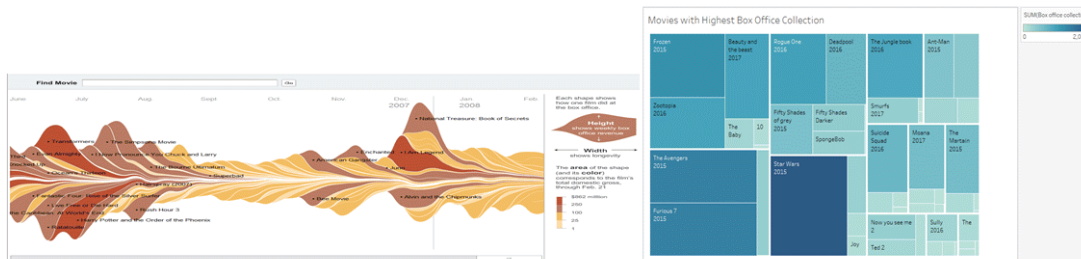


Figure 2(a) Steam graph

(b) Re-visualization of the steam graph using the tree map

- [6] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. Byers, A. H. Big data: The next frontier for innovation, competition, and productivity, 2011, www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Digital/Our%20Insights/Big%20data%20The%20next%20frontier%20for%20in%20novation/MGI_big_data_full_report.ashx , retrieved November 29, 2016.
- [7] Patil, D. J. & Davenport, T. H. Data scientist: The sexiest job of the 21st Century, 2012, www.hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century, retrieved November 29, 2016.
- [8] Robbins, K. A., Senseman, D. M., Pate, P. E. Teaching biologists to compute using data visualization. *Proceedings of the 42nd ACM Technical Symposium on Computer Science Education*, 335-340, 2011.
- [9] Sallam, R. L., Tapadinhas, J., Parenteau, J., Yuen, D. Hostmann, B. Magic quadrant for business intelligence and analytics platforms, 2016, www.gartner.com/doc/reprints?id=1-2XXET8P&ct=160204, retrieved November 29, 2016.
- [10] Urness, T. Incorporating data visualization in a course on computer graphics. *Journal of Computing Science in Colleges*, 31, (5), 147-154, 2016.

Appendix A

Criteria	Tableau	QlikView	Microsoft Power BI
BI Platform Administration	Server availability 99.99%, Disaster Recovery No DR on desktop	Perfect scaling on multi-processor, multi-core hardware	Slow: Large data sets. Disaster Recovery
Security and User Administration	Secure Dashboards Security: Filter	Desktop: No security features Administrator: cloud services.	BI service: Build on <u>Mircosoft's</u> cloud Azure Active Directory
Data Source Connectivity	Connectivity to many data sources like MySQL, Microsoft SQL Server, etc.	QlikView can be connected to several types of data sources like MySQL, Excel	Connectivity to different data sources like Excel, SAP HANA, Oracle etc.
Metadata Management	3-tier system. Abstraction Layer, Data Model Layer, Run Time Model	Metadata Management: Descriptive metadata Administrative metadata Structural metadata	Metadata management data source and access permissions
Data Storage	ETL mechanism No data storage links itself to the data	ETL mechanism No data storage links itself to the data	Inbuilt ETL tool Mechanisms and tools for data analysts
Data Preparation	Drag and drop options Flexibility /limitations in merging data sources	No drag and drop, no interface interactions Flexibility in data merging	Drag and drop options Flexibility in merging from data sources
Advanced Analytics	Flexible integration with other tools (ex: <u>snapLogic</u>)	Flexible integration with R language for prediction	No integration with other analytics engines
Analytic Dashboards	Interactive dashboards and storyboards	Effective dashboards for interactive visualizations	Combines titles to create dashboards
Interactive Visual Exploration	Generation, manipulation of Charts (coloring, labeling, resizing)	Generation, manipulation of charts (legend, colors, resizing)	Manipulation of title, data colors, background, labeling
Mobile Exploration	Fluidity in viewing, finger tap feature, filtering	Mobile Apps, interactive analysis, associative search	Mobile & desktop apps, touch-optimized interaction
Publishing Analytic Content	Storyboards, Data Refreshing, Server scheduling	Dashboards, Reloading data, Scheduling, Reconfiguration	Dashboards, Reports, Animations
Collaboration and Social BI	https://community.tableau.com/	https://community.qlik.com/welcome	http://community.powerbi.com/

Table I. Comparison of data visualization tools

Appendix B

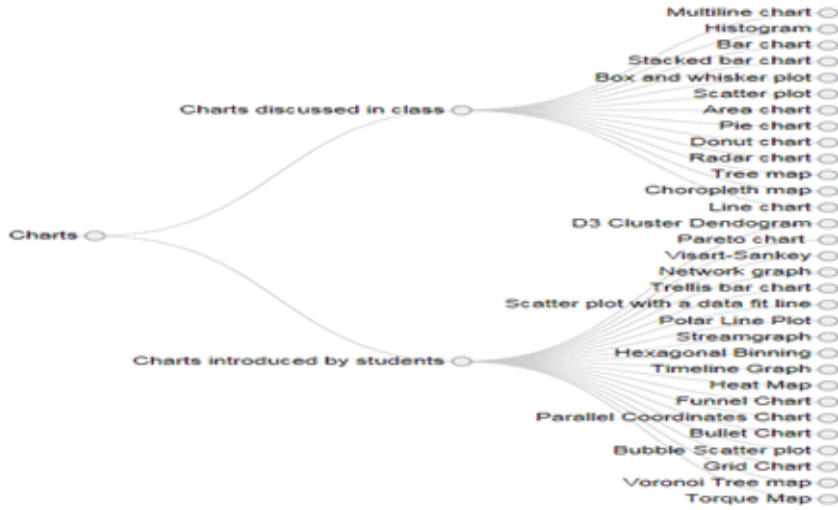


Figure 4. A cluster dendrogram representing charts introduced in the Data Visualization course

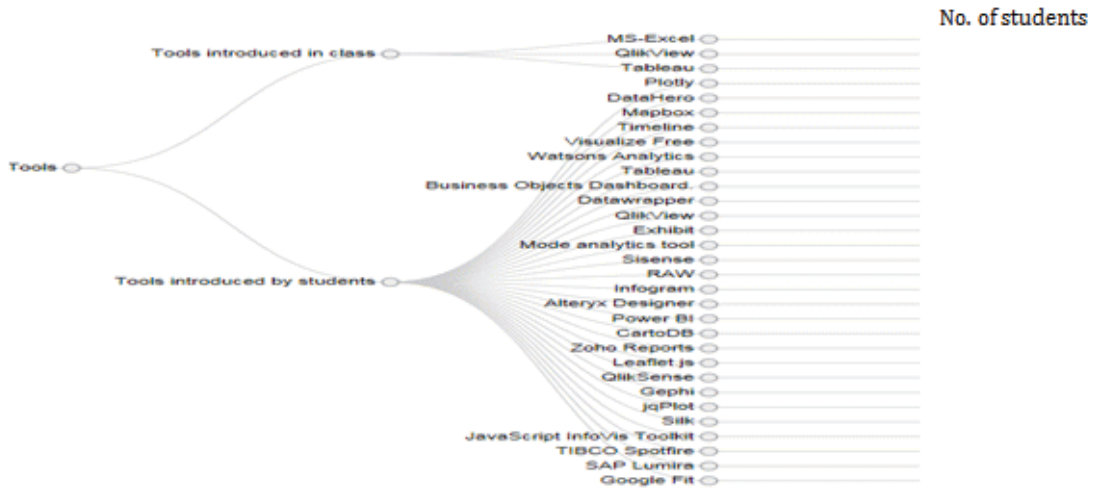


Figure 5. A cluster dendrogram representing Tools introduced in the Data Visualization course