



Higher Education
Quality Council
of Ontario

An agency of the Government of Ontario

Information Visualization: A User Guide for Educational Data

Ajay Sivanand and Brian Frank
Queen's University



Published by

The Higher Education Quality Council of Ontario

1 Yonge Street, Suite 2402
Toronto, ON Canada, M5E 1E5

Phone: (416) 212-3893
Fax: (416) 212-3899
Web: www.heqco.ca
E-mail: info@heqco.ca

Cite this publication in the following format:

Sivanand, A. & Frank, B. (2018). *Information Visualization: A User Guide for Educational Data*.
Toronto: Higher Education Quality Council of Ontario.



The opinions expressed in this research document are those of the authors and do not necessarily represent the views or official policies of the Higher Education Quality Council of Ontario or other agencies or organizations that may have provided support, financial or otherwise, for this project. © Queens Printer for Ontario, 2018

Acknowledgement:

This report could not have been the quality it is, without the help of Dr. Jake Kaupp and David Waller. Thank you to Gregory Hicks for his technical writing experience and for helping with the structure and clarity of the report.

Table of Contents

| | |
|--|----|
| Introduction | 5 |
| Visualization Design Guide..... | 6 |
| Step 1: Understanding the problem | 6 |
| Step 2: Assigning Attributes..... | 13 |
| Step 3: Refining the Visualization | 16 |
| Guide Limitations and Best Practices..... | 17 |
| Guide Examples..... | 17 |
| Example 1: “Do underrepresented groups have the same graduating rate as highly represented groups?” | 17 |
| Example 2: “What are the most commonly selected courses for students of a particular demographic group?” | 20 |
| References | 23 |
| Glossary of Terms..... | 24 |

List of Tables

| | |
|---|----|
| Table 1: Summary of Data Abstraction Categories and Options | 7 |
| Table 2: List of Attribute Types with Examples..... | 9 |
| Table 3: Summary of Action Abstraction Levels and Corresponding Options for Each..... | 11 |
| Table 4: Summary of the Types of Targets and the Condition of Task to Which They Can Apply..... | 12 |
| Table 5: Ordered List of Channels (adapted from Figure 5.1 in Munzner, 2014) | 14 |
| Table 6: Abstracted Data Table for a Question about Graduation Rates | 18 |
| Table 7: Abstracted Data Table for Question about Course Selection | 20 |

List of Figures

| | |
|--|----|
| Figure 1: Examples of Items and Attributes in Tabular Format | 8 |
| Figure 2: Examples of Comparison of Similar Groups in a Data Set | 15 |
| Figure 3: Visualization for “Do underrepresented groups have the same graduating rate as highly represented groups?” | 19 |
| Figure 4: First Version of Visualization for “What are the most commonly selected courses for students of a particular demographic group?” | 21 |
| Figure 5: Updated Version of Visualization for “What are the most commonly selected courses for students of a particular demographic group?” | 22 |

Introduction

When interpreting or communicating information with data, information visualizations play a prevalent and possibly integral role. Data visualization tools and libraries such as [Raw Graph](#) and [Data Visualization Catalogue](#) can help users visually depict a set of data. Given the many different tools available, it is important to know which visual representations can be executed most efficiently while also maximizing their utility.

Postsecondary faculty and staff looking to increase their usage of educational data to drive decision-making regularly encounter the limitations of traditional tools and approaches for analyzing data. Data sets and the questions they are intended to answer have expanded and become more complex. This has made it harder to generate usable insights with conventional methods. There are few software tools available to incorporate complex educational data into program improvement decisions at the course, department or institution level. A full list of these tools can be found in our review of the current work in the area (Sivanand & Frank, 2017).

This guide was created as part of a larger research project whose goal was to find ways to promote the use of educational data at the postsecondary level through information visualizations. It is intended to help faculty and educational support staff design more information-rich visualizations that make better use of available educational data and provide insight into more complex areas of inquiry.

The research project identified the main stakeholders involved in the visualization creation process as well as the major challenges that they currently face. An online web application called [Easel: Visualization Recommendation System \(VRS\)](#), was developed to support several of these challenges, specifically those around determining both questions to ask with data and ways of answering those questions. Such questions include “How are students from other departments doing in courses we are offering?” and “Do quiz marks predict success to a similar question on the final exam?” The recommended visualizations were designed based on expertise gained from various resources (Tufte & Graves-Morris, 1983; Bertin, 1981; Börner & Polley, 2014) but primarily Munzner’s textbook on the subject: *Visualization Analysis and Design* (2014). Sivanand gives full details of the larger research project and Easel’s development (2017).

The intention of Easel: VRS was never to provide recommendations for all questions that a user might have. We distilled the relevant areas of Munzner’s textbook into a more streamlined methodology to create purpose-built visualizations for use with educational data. This guide is intended to assist someone looking to take their educational data from a question of inquiry and illustrate it as a visualization that would be most effective for their needs.

Following Munzner’s approach to visualization design, this guide focuses on abstraction of the type of data and task at hand. The abstraction process involves transitioning the factors surrounding the visualization design from terms and concepts specific to the education domain into ones that are common to visualization designs across domains. With the data and task abstracted along various criteria, the

consequences of individual design choices can be easily discerned one at a time. The resulting visualization is optimized to provide insight into the task it was designed for with the available data. The guide is made to be platform agnostic, and so does not detail how one would program the visualization into one of the many available software tools (e.g., Tableau, D3, ggplot2, InfoVis, VTK); just how to design it.

To summarize, this guide is intended to help you:

- Understand where visualizations might provide value in your data interpretation practice.
- Refine broad areas of inquiry into specific and answerable research questions that can be answered precisely and accurately.
- Design a purpose-built visualization that frames the data to make the information you are looking for more immediately apparent.
- Re-contextualize new areas of inquiry so that they can be more easily compared with ones that already have solutions.

Visualization Design Guide

This process for visualization design is adapted from the workflow outlined in Munzner's book *Visualization Analysis & Design* (2014). The kinds of data and tasks found in the education field only encompass a portion of what the book offers guidance on. The process presented here is therefore a streamlined version that can be more accessible to those in postsecondary education (PSE). For example, this guide focuses on the steps pertinent to tabular data, as it represents the majority of work done in education. For our purposes, tabular data is data sets that can be organized into rows and columns.

The adapted process is separated into three general steps:

1. Understanding the problem
2. Assigning attributes
3. Refining the visualization

Step 1: Understanding the problem

A visualization should be designed for the kind of data it is being used to analyze as well as the kind of problem for which it is providing insight. Before taking steps to visualize the data, it is important to understand both the kind of data and the kind of problem in order to make designing the visualization more straightforward and help ensure that the intended goal for the visualization is best achieved. There are two parts to understanding the problem: knowing the type of data that is being used and defining the task the visualization is intended to help with. Both are done through a process of abstraction using categories

defined by Munzner. Though there is some minor overlap between these two parts, they are largely separate processes and can be done in any order.

Data Abstraction: What Data is Being Shown?

The first part in the design of a visualization in response to a question is identifying the nature of the data that is available. The available data sets and attributes determine how the data can be broken down through abstraction and pieced together to answer the question that has been posed. The process for this abstraction starts with considering what data may be available from the individual posing an educational question or what other data is needed for educational purposes. Each question requires careful consideration of the role of the individual at their institution and accordingly, what data they would have available to them from which to build a visualization. Initially, the data that will be needed to answer the question should be identified. This guide assumes that the data will already be available, but this process does help in determining what data would need to be collected.

Table 1: Summary of Data Abstraction Categories and Options

| Data Set Abstraction | | Attribute Abstraction | |
|-----------------------|---|-----------------------|--|
| Abstraction Category | Options | Abstraction Category | Options |
| Data Types | <ul style="list-style-type: none"> • Item • Attribute | Attribute Semantic | <ul style="list-style-type: none"> • Key attribute • Value attribute |
| Data Set Types | <ul style="list-style-type: none"> • Tabular Data | Attribute Types | <ul style="list-style-type: none"> • Categorical • Ordinal • Quantitative |
| Data Set Availability | <ul style="list-style-type: none"> • Static • Dynamic | Ordering Direction | <ul style="list-style-type: none"> • Sequential • Diverging • Cyclic |

The data abstraction process requires identification of the characteristics from each category in Table 1 for each question that requires a visualization. Each data set to be visualized is categorized into data types, which are the set of items or attributes to be visualized, the data set type (of which only tabular data is considered for this guide), and data set availability, which is dynamic if the data can change, or static if not. **Items** in the data set are the individual unit of analysis for each area of inquiry, corresponding to a row in a data set. **Attributes** are the data associated with, and containing characteristics of, each item and are divided into key or value attributes; they correspond to columns in a data set.

For example, in the case of a table detailing students and their assessment performances over a course, each student would be an item (an individual unit of analysis in a visualization), while their performance on each assessment would be an attribute. Such a case can be seen in Figure 1, where each column is an attribute, and each cell corresponds to an attribute of that student. If, instead, one was looking at summary

statistics (mean, standard deviation, etc.) of different assessments, each assessment might be an item while each statistic would be an attribute of that assessment.

Figure 1: Examples of Items and Attributes in Tabular Format

| | Key Attribute | Value Attributes | | | | |
|-------|---------------|------------------|-----------|-------------|------------|--------------|
| | Student_ID | GPA | Test_Time | Citizenship | Birthdate | Test_Mastery |
| Items | 830744 | 3.99 | 23 | Canada | 1993-12-12 | Proficient |
| | 612032 | 3.06 | 42 | Canada | 1993-10-12 | Below Basic |
| | 691492 | 3.48 | 38 | Canada | 1993-02-24 | Proficient |
| | 548130 | 3.13 | 27 | Sri Lanka | 1993-09-16 | Proficient |
| | 708006 | 4.06 | 4 | Canada | 1994-06-29 | Below Basic |
| | 889780 | 0.70 | 19 | Canada | 1994-02-17 | Proficient |
| | 603028 | 3.76 | 50 | Canada | 1993-12-22 | Proficient |

Note: This is a fictional example of tabular data demonstrating items as well as key and value attributes. Here, each student is an item along with six associated attributes—Student ID, GPA, Test Time, Citizenship, Birthdate and Test Mastery. Student ID is a key attribute while the rest are value attributes. The associated data for each student are corresponding values for each attribute.

Each item in a data set has attributes associated with it, which are identified by considering how the data relates to the real world in the context of each area of inquiry. Specifically, which attributes are needed as key attributes and which as value attributes? **Key** attributes are ones that can uniquely identify an item, while **value** attributes are characteristics about the item. In the example of the students and their assessment performance, a student number would be a key, while their grade on an assessment would be a value attribute.

The distinctions between key and value attributes are not intrinsic to the data and are often driven by the task abstraction, which helps clarify the purpose and goal of each question. Keys are always connected to the items of each visualization. For simple questions, at least one key is typically required for each visualization, however; multiple keys can be required to identify grouping attributes for more complex questions. When looking at groupings of students, their student number and the attribute used for grouping might be separate keys. In the example shown in Figure 1, the student ID column is the most obvious key. However, if the visualization also groups students according to their citizenship, that could also be a key.

A value attribute is determined by identifying what recorded measurements or observations are needed to answer each question. For example, determining if a class is improving in a skill requires measuring student performance.

Once the different data types are confirmed, the identification of the data set type is straightforward. Since tabular data is the main data set type found in PSE, that is the focus here. This process would look slightly different if network or geospatial data was being considered.

Data set availability is determined by the nature of the question. Questions are characterized as either static (if the data being visualized is available all at once), or dynamic (if the data is continuously being collected and added to the visualization). The dynamic characterization would mean either that attributes for an item would change with time, or that items would be added or deleted.

The attribute characteristics can be one of three types: categorical, ordinal or quantitative. **Categorical** data does not have any implicit ordering. As a result, the only operation that can be performed on this kind of data is determining whether two elements are the same or not the same. If the data does have an implicit ordering, it can be either ordinal or quantitative. The main difference between these two is that arithmetic can be performed on **quantitative** data to create new data, but not on **ordinal** data. For example, a ranked list is ordinal, but adding the first item to the third is not a meaningful concept. Table 2 provides an easy reference for assigning attribute types. Note that keys can only be categorical or ordinal, not quantitative.

Table 2: List of Attribute Types with Examples

| Attribute Type | Definitions | Examples |
|----------------|---|---|
| Categorical | <ul style="list-style-type: none"> Not ordered Can only be compared | Faculty: Engineering, Humanities, Law Degree Program: bachelors, bachelor's with honours, master's |
| Ordinal | <ul style="list-style-type: none"> Ordered Cannot be part of arithmetic calculations | Year of Study: First < Second < Fourth Grade on Assessment: A > B+ > D |
| Quantitative | <ul style="list-style-type: none"> Ordered Can be part of arithmetic calculations Often measurements | Assessment Mark: 61% < 80% < 95% Class Attendance: 558 of 620 (90%), 12 of 16 (75%) |

In Figure 1, student ID and citizenship are categorical attributes. Notice that even though student ID is a number, individual numbers have no relation to one another. It does not make sense to say that student 603028 has a lower ordering than 830744. Test mastery is an ordinal attribute since there is a clear ordering

from below basic to basic to proficient. GPA and test time are both quantitative attributes, since it would make sense for this data to be subtracted, averaged etc. The birthdate attribute however, is not as simple to categorize since it depends on how the attribute is being used. If the students are being grouped by year of birth then it could be considered an ordinal attribute. But if birthdate was instead being used as part of some calculation, it would be considered a quantitative attribute.

The **ordering direction** applies to ordinal and quantitative data and can be sequential, diverging or cyclic. An attribute is **sequential** when it is a range that has a clear minimum and maximum. The mark of an assessment graded out of 10 would be sequential, ranging from a low of 0 to a maximum of 10. **Diverging** attributes are ranges that have two sequences that can increase in opposite directions that meet at a common zero point. How far a student's grade is from the average would be a diverging attribute. A **cyclic** attribute is one whose values wrap around to a starting point. Common examples would be the days in a month, or terms in a school year. All the ordered attributes in Figure 1 are sequential.

The semantics of the data make both the attribute type and ordering direction easy to identify. For example, assessment results such as percentages or GPAs can be clearly recognized as quantitative and sequential, semesters can be recognized as ordinal and sequential, and program type is categorical with no ordering direction.

Some of the attributes required by the visualization are not provided by the raw data. Instead, they need to be derived using calculations. Common examples of derived attributes (including differences, averages, categories, etc.) and should be considered additional value attributes when applicable, or less commonly as key attributes. The nature of the derivation is linked to the task abstraction and discussed further in the next section. These derived attributes should be categorized accordingly as well.

Task Abstraction: Why is this Being Shown?

The task abstraction process clarifies the intentions and goals of each area of inquiry that the data is being used to provide insight into. The fundamental purpose for the visualization can be more easily discerned by moving each area of inquiry from domain-specific terms, such as grades or students, to an abstracted domain of visualization terms and concepts, such as a distribution of items. This process also enables recognition of similarities between questions and the designs. Furthermore, task abstraction helps identify the necessary data transformations which guide the data abstraction process. The task abstraction categories can be seen in Table 3.

Table 3: Summary of Action Abstraction Levels and Corresponding Options for Each

| Level of Abstraction | Options |
|----------------------|--|
| Analyze | <ul style="list-style-type: none"> • Present • Discover: exploration/discussion |
| Search | <ul style="list-style-type: none"> • Known target with unknown location • Known target and location • Unknown target with known location • Unknown target and location |
| Query | <ul style="list-style-type: none"> • Identify one target • Compare few targets • Summarize all targets |

The task abstraction begins with considering how the question is posed and the utility of creating a visualization to deliver a response. Examples of utility include improving visualizations for common questions, or providing deeper insight into a class, program, institution, etc. Once a question is concise and focused, the task is relatively simple to classify. However, when the question is not focused, the task abstraction process can narrow down the query to what it is that is being asked.

The task abstraction is based on two primary concepts: actions and targets. An **action** is what a visualization enables a user to do with data and is categorized into three levels. A target is an aspect of the data being analyzed. The purpose of a visualization design can be defined in these terms as letting the user perform an action on a target. These actions are what the visualization primarily enables the user to do with the data and can be categorized into three levels: analyze, search and query.

At the highest level, **analyze**, a visualization is either meant to discover new information or present known information to others. Discover is further subdivided into exploration (when the analysis is being done by a single party) and discussion (when the visualization is used as a tool to facilitate group analysis).

The two other levels are identified when we can specify what the target is. A **target** is what a user is trying to analyze from a visualization. With all data, the target can be a trend, an outlier, or when more specific, a feature. When dealing with a single value attribute, the target could be the distribution of the data or the extreme items in the data. When there are multiple attributes, the target can be a dependency or a correlation. The types of targets can be seen in Table 4.

Table 4: Summary of the Types of Targets and the Condition of Task to Which They Can Apply

| Target Type | Condition |
|---|-------------------------------------|
| <ul style="list-style-type: none"> • Trend • Feature | Can apply to all data |
| <ul style="list-style-type: none"> • Outlier • Extreme | Can be a single item or a few items |
| <ul style="list-style-type: none"> • Distribution | Data only has one value attribute |
| <ul style="list-style-type: none"> • Dependency • Correlation | Data has multiple value attributes |

After the target type is identified, actions can be categorized according to the next level, its search classifications. The **search** level requires consideration of whether the identity of the target is known prior to looking at the visualization, as well as if the location of the target among the data is known. The location often depends on the specificity of the question. For example, questions that focus on a particular data subset have known locations while generic questions tend to have unknown locations.

The lowest level, **query** partially depends on the number of targets. The action could be to identify one target, compare multiple targets, or summarize all targets. For example, if you were trying to see if there is an increase or decrease in graduation rates over time, you would be identifying (action query) a trend (target). Alternatively, if you were trying to see how one group of students were performing compared to the overall class, you would be comparing (action query) the distributions (targets).

As mentioned previously, this task abstraction method has the added benefit of gauging the specificity of the current task. If a visualization is being created for a question that does not neatly fit into these task abstraction categories, it is likely that the question needs to be more precise. A question like “Are there cliques within a student team?” could not be neatly abstracted since a ‘clique’ would not be a readily identifiable feature target from a set of data. The question could then be refined to “Which student teams are demonstrating inconsistent perceptions?” The new question is more specific and more easily answered with the help of a visualization.

Task and data abstraction overlap when considering derived data. This is data that is calculated in addition to the raw gathered data. Examples of derived data include averages of grades across assessments or students, counts for bin ranges, and differences between student scores or evaluations. The overlap between the two abstractions occurs when the derived data reduces many value attributes down to a few, such as when calculating the average of the results of an assessment; the many grades of assessments are reduced to a single average value. It is important to be clear about what targets are being identified or compared so that those targets do not get reduced. For example, if the target of a visualization was a student, then any reduction of students would want to be avoided. Examples might include calculating the

average for an assessment or binning students for a histogram. In tabular data, such as a table where each student is a row and each column is an assessment measure, one useful way to think about reduction is whether rows (students) or columns (assessments) are the ones being reduced.

Overall, the main steps one should complete for a data type and task are as follows:

Data Abstraction

1. Determine data items and attributes.
2. Define the associated keys and values for the attributes.
3. Determine the attribute type and ordering direction (if applicable) for each attribute.
4. Identify what data needs to be derived for the task.

Task Abstraction


1. Decide if the purpose for the visualization is to present, explore or create discussion.
2. Identify the target.
3. Determine whether you are looking for a specific target and whether its location is known.
4. Finalize whether you are trying to identify a target, compare a few targets or summarize all the targets.

Step 2: Assigning Attributes

Once the data abstraction process is completed, the result should be a list of attributes; each that is either a key or value, and of a specific type (categorical, ordinal or quantitative). Additionally, based on the task abstraction, there might be an order to the precedence of the attributes. For example, when looking at means and standard deviations of a set of assessments, the task might either be to compare averages or the spread of results. If looking at averages, then the mean would have a higher precedent, while if the spread of results was the primary focus then that attribute would have a higher precedent. The other attribute might still be worth visualizing, but in a secondary manner.

With such a table, it can be somewhat mechanical to come up with a first pass at a visualization with the use of visual marks and channels. A **mark** is a basic geometric element that depicts an item such as points on scatter plots or lines in bar charts. **Channels** are different ways that you can control a mark's appearance such as position, shape and colour. Channels are used with marks to represent magnitude (quantitative data) or identity (categorical or ordinal data). Some channels have been shown empirically to be read more or less accurately. Table 5 lists channels for magnitude and identity from most to least readable.

Table 5: Ordered List of Channels (adapted from Figure 5.1 in Munzner, 2014)

| | Magnitude | Identity |
|---|--|---|
| More Accurate  Less Accurate | Position on a common scale (horizontal or vertical) Position on an unaligned scale Length Tilt/angle Area (2D size) Colour luminance Color saturation Volume (3D size) | Position (spatial region) Color hue Shape |

Notice that position is at the top of both lists in Table 5. This understates how much better position is than many of the other channels and that it should be given special consideration and priority. Position can be an accurate way to represent quantitative data, and effectively distinguish categories by separating marks into regions. A common example of both of these in use is in a categorized bar chart, where the position on a common scale is used to represent the quantities value of each bar while placing bars together shows that they are part of the same group.

Some channels may also interfere with each other and should be treated with care. For example, if a horizontal position channel is used on the marks in conjunction with a vertical position channel, it might look like an area channel which would have unintended consequences on the perception of the data.

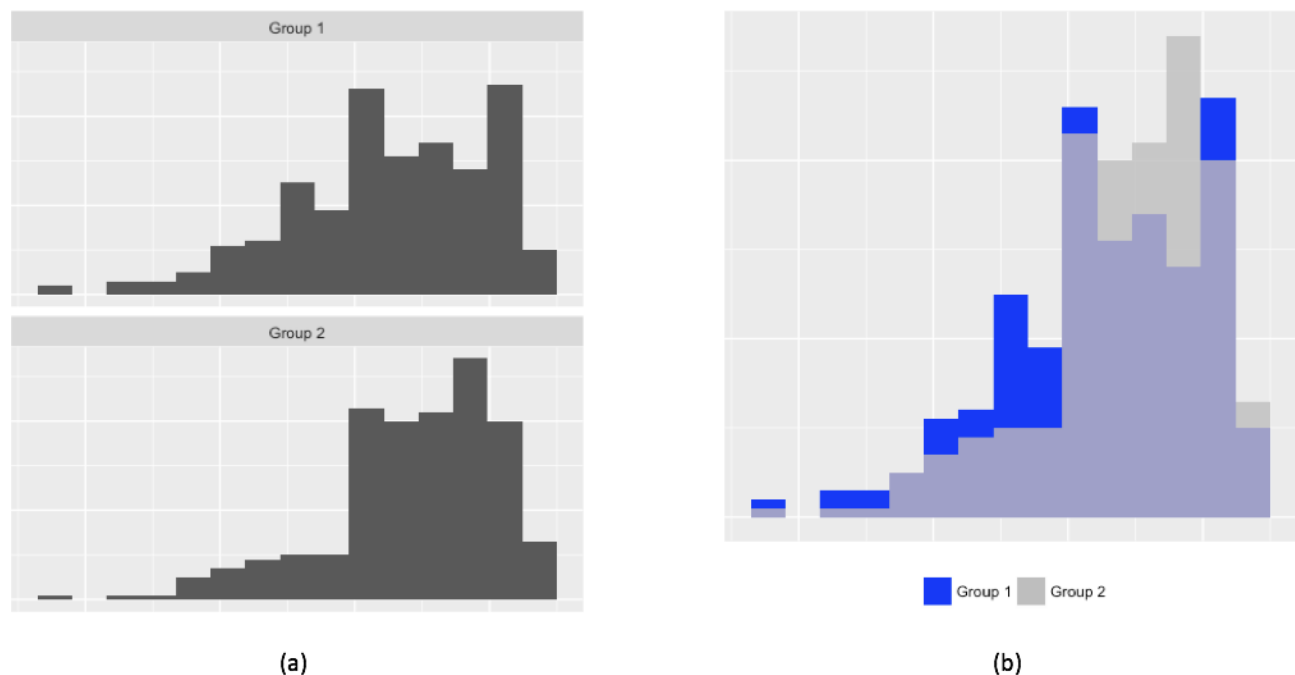
Colour should also be given special consideration. The colour channel can be separated into hue, luminance and saturation. A colour's hue (blue, red, yellow etc.) can be used with great effect to distinguish categories but is poor for representing quantitative information. A colour's luminance and saturation (its darkness and brightness) are useful ways of conveying quantitative information, but should be used with caution given their low ranking in Table 5. While colour can be used well for comparison between marks, it does not work well for a user to accurately read quantitative information. Also, when choosing colours, pay attention to whether your colour palette works for people with colour-blindness. [ColourBrewer](#) is a great online resource for colour palettes. Additionally, a good rule of thumb is that a visualization can still work for its intended purpose when being viewed in greyscale (without hues). This ensures that the three aspects of colour are sufficiently distinguishable and accessible regardless of the medium in which the visualization is presented.

Channels can be assigned to value attributes by determining if they are quantitative or categorical and going down the prioritized list incrementally. However, choices should ensure that the keys are still easily distinguishable to the extent that the task requires. Generally, if you have one key, your visualization will

resemble a list. A bar chart can be thought of as a list of bars, each representing a quantitative value. With two keys, the visualization is likely to look like a matrix. A heat map is the most common example, where there are rows and columns for identifying keys and the value in each cell is represented by a quantitative channel, often luminance. With three keys, you can use a 3D visualization, but it is not recommended. Another choice might be to re-use certain channels, such as spatial region but in a different direction; while noting any extra side-effects. More than three keys requires more advanced techniques like recursive subdivision, where smaller sub-visualizations are subdivided within the overall visualization. An example of this might be a histogram or pie chart within the bars of a bar chart.

Another aspect to consider when assigning value attributes is whether the visualization should be faceted or layered. Faceting displays similar visualizations beside each other separated by a key. When you have multiple keys, this is a good way to use a one-key or two-key visualization multiple times. Figure 2a illustrates an example of a visualization with one key used multiple times for another key with values Group 1 and Group 2. This can be especially useful when comparing a specific value attribute along a key. Layering is another useful approach, especially when the key task is a comparison of two targets. Layering displays one visualization over another using the same scale and space, often by reducing the transparency of the visualization on top. Figure 2b illustrates an example where values of another key are layered for the purposes of comparison.

Figure 2: Examples of Comparison of Similar Groups in a Data Set



Note: In (a), the data sets are faceted while (b) illustrates how data sets might be layered.

If using faceting or layering, find an order that makes sense for that grouping. There might be an implicit ordering, but if not an ordering can be imposed. For example, when comparing non-ordinal groups of students, the number of students in each group might be a possible ordering.

Overall, the main steps one should complete when assigning attributes are:

1. Determine whether attributes are keys or values, and if value attributes are quantitative, ordinal or categorical.
2. Assign channels in order based on value attributes that are most important to the visualization while accounting for interference between channels.
 - Decide whether the visualization requires any faceting or layering.

Step 3: Refining the Visualization

Once a visualization has been designed, the final step is to evaluate and refine it. Once channels have been assigned to the respective data, the resulting visualization may or may not resemble a familiar visualization. At this point, it is important to check whether the created visualization is well suited for the task it was intended to help provide insight for. Common visualizations are prolific because they have found use in many tasks, but they are not always the best visualization for a given purpose.

Often, when a visualization designed for a specific task is built and displayed, the information that most readily presents itself might not suit the task that it was intended to bring insight into. If this is the case, there are two possible solutions. The first is to revisit step 2 and find other ways to represent the data that more accurately and readily present the information you are looking for while accounting for what is known to not work. This might involve some trial and error.

The other solution is to repurpose the visualization for a different task and return to step 1. This visualization may provide insight for an alternate question that is also worth asking. The advantage with this second route is that the visualization has already been created. Building a visualization is a highly iterative process and it is common to redo previous steps once you see a final visualization of the data.

If the visualization does provide the right kind of information, make sure to provide all the necessary details for a user to interpret the visualization quickly. This includes legends, labels, titles, tick marks on axes, text annotated on marks and any other details that would make interpretation of the visualization quicker (as seen in Figures 3 and 5 in the examples). The formatting of the visualization, such as background colour, axis grid lines and other aspects can have a significant impact on the readability of information. It is sometimes appropriate to use channels more than once in the same visualization. For example, a bar chart with different colours might already have a legend, but it might be useful to add text to each bar for ease of interpretation. Often this step can mean trying different options and seeing what works best, or even giving the eventual visualization consumers some options about how they consume the data.

Guide Limitations and Best Practices

As mentioned earlier, this guide focuses on tabular data. Other kinds of data, such as network data or geographical data would require slightly modified approaches. Within tabular data too, the visualization might be restricted by the tool being used to create the visualization. Many of the dedicated software tools for visualization can handle very large data sets (e.g., looking at data for whole institutions). However, visualizations of that size might not function as well in live or interactive formats. More widely used visualization tools such as Microsoft Excel can also struggle with more intricate visualizations or with data that changes frequently. This was a large reason why this guide was made without any specific tool in mind, since this process for designing a visualization applies to data sets of any size. Though, depending on the available tools and skillset, only certain visualizations designs might be attainable.

Additionally, visualizations work best to provide information that can help users make data-driven decisions. These decisions cannot be made with visualization alone. There is an interpretation stage that occurs between using the visualization and making a decision, where the user applies the insight from the visualization to their needs. For example, when a program director looks at the courses most commonly selected among their students, they are now better informed about whether there need to be any program changes. The information gained from the visualization might even suggest that no changes are needed.

As such, when trying to find where visualizations might help your practice, do not look for challenges that visualizations can solve. Instead, focus on what kind of information can better help you tackle those challenges. A question like “Is the program teaching and assessing the outcomes we think it is?” is worth answering but could instead be reconsidered as “Where are we teaching and assessing the intended outcomes?” A visualization designed around the latter question can then help answer the former with the additional insight of the faculty and staff involved.

Guide Examples

Example 1: “Do underrepresented groups have the same graduating rate as highly represented groups?”

Step 1: Data and Task Abstraction

The task here is to compare trends of student graduation rates. The target then would be trends, and the action would be to compare. In this case the location is known since the relevant data would be known beforehand, but the target would be unknown since there are multiple groups that might be worth comparing. The abstracted data table would be as seen in Table 6. Note that we chose year of graduation to be an ordinal value since it was a key. Time is treated as discrete here given the context that a graduating class would have little effect on successive years, though time could be classified as quantitative for a different task.

Table 6: Abstracted Data Table for a Question about Graduation Rates

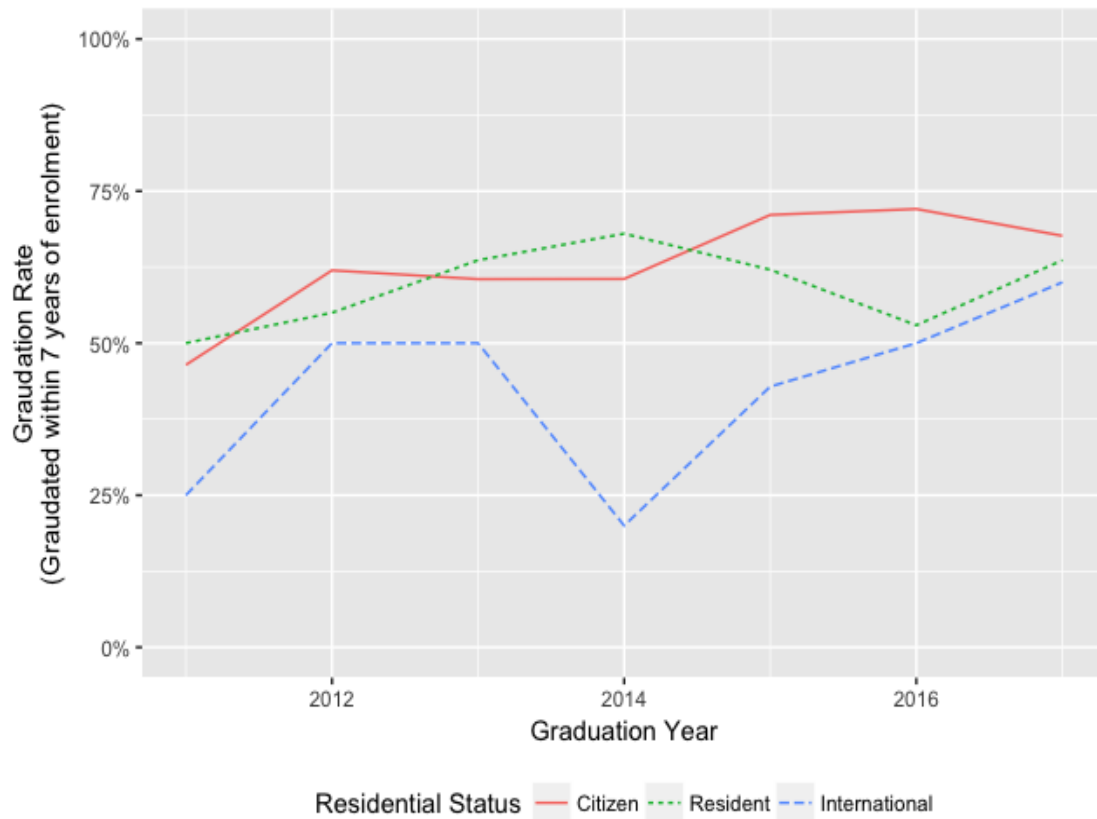
| Attribute | Attribute Semantic | Attribute Type |
|---------------------------|--------------------|--------------------------|
| Demographic group | Key | Categorical |
| Year of graduation | Key | Ordinal, sequential |
| Graduation rate (derived) | Value | Quantitative, sequential |

Step 2: Assigning Attributes

Channels can then be assigned by consulting Table 4. There are two keys: demographic group and year of graduation. Spatial region, colour or shape can be assigned to them since these are identity attributes. Since the task is to compare trends over years, the different demographic groups should be viewed in the same region for easy comparison. Therefore, colour is used for demographic group region for year of graduation. There is one value attribute, so position can be used on a common scale. The result is coloured marks that are positioned in 2D space according to a year and graduation rate.

For the visual mark, either a point mark or line mark can be chosen to represent the data. This would mean the difference between a scatter plot and a bar graph, respectively. With the use of points, the marks can be linked with lines to emphasize that marks within the same group are related. For the two position channels, there is no implicit reason in the data abstractions that dictate whether graduation rate or year of graduation should occupy the horizontal or vertical channels. However, because of convention, time was chosen to occupy the horizontal channel. Otherwise the visualization could be unnecessarily obtuse to viewers who are used to a certain visualization standard. Figure 3 shows the visualization these choices might produce.

Figure 3: Visualization for “Do underrepresented groups have the same graduating rate as highly represented groups?”



Step 3: Refining Visualization

Since this question deals specifically with underrepresented groups, it might also be worth including an extra quantitative attribute for the percentage of the population that group represents. As one would expect, the proportion would change from year to year. Another channel can be assigned for this attribute if it does not interfere with the vertical channel occupied by graduation rate. The current mark could be augmented, or more can be added. One way to augment the current mark would be to use the angle channel so that each point is a slice of a pie chart. Another way might be to use the horizontal position channel where the mark in each year is horizontally placed proportionally to each year region. Or an altogether new set of marks could be used where a bar chart in each year-region represents the proportion of students. The options are plentiful, but the process provides a structured way of trying different, possibly viable approaches. The best way of finding the most effective visualization is to try different options to see what does and does not work.

Example 2: “What are the most commonly selected courses for students of a particular demographic group?”

Step 1: Data and Task Abstraction

In this case, the task is to identify courses that are extremes or outliers of the course selection rates. The target here being outliers, and the action being to identify. Here, both the target and location are unknown. The set of abstracted data would be as seen in Table 7. The course selection rate here is taken as the percentage of the demographic group that took the course. In this case, we also have some information about how the courses are distributed based on year of study.

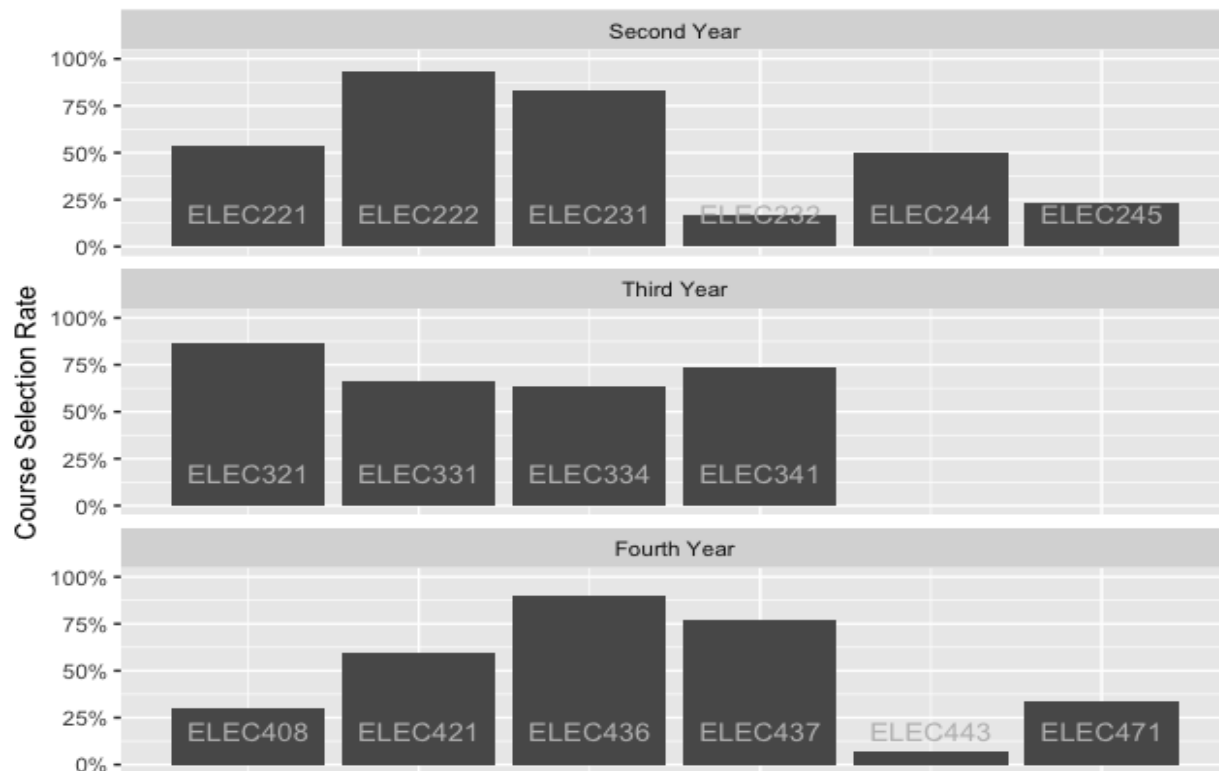
Table 7: Abstracted Data Table for Question about Course Selection

| Attribute | Attribute Semantic | Attribute Type |
|---|--------------------|--------------------------|
| Course selection rate among demographic group (%) | Value | Quantitative, sequential |
| Year of study | Key | Ordinal |

Step 2: Assigning Attributes

The simplest way to visualize this would be to make a list of selection rates, grouped by year of study. The position channel is used for the value attribute, while the bars are sectioned into regions for the year of study, ordered from second to fourth year. This initial version of the visualization can be seen in Figure 4.

Figure 4: First Version of Visualization for “What are the most commonly selected courses for students of a particular demographic group?”



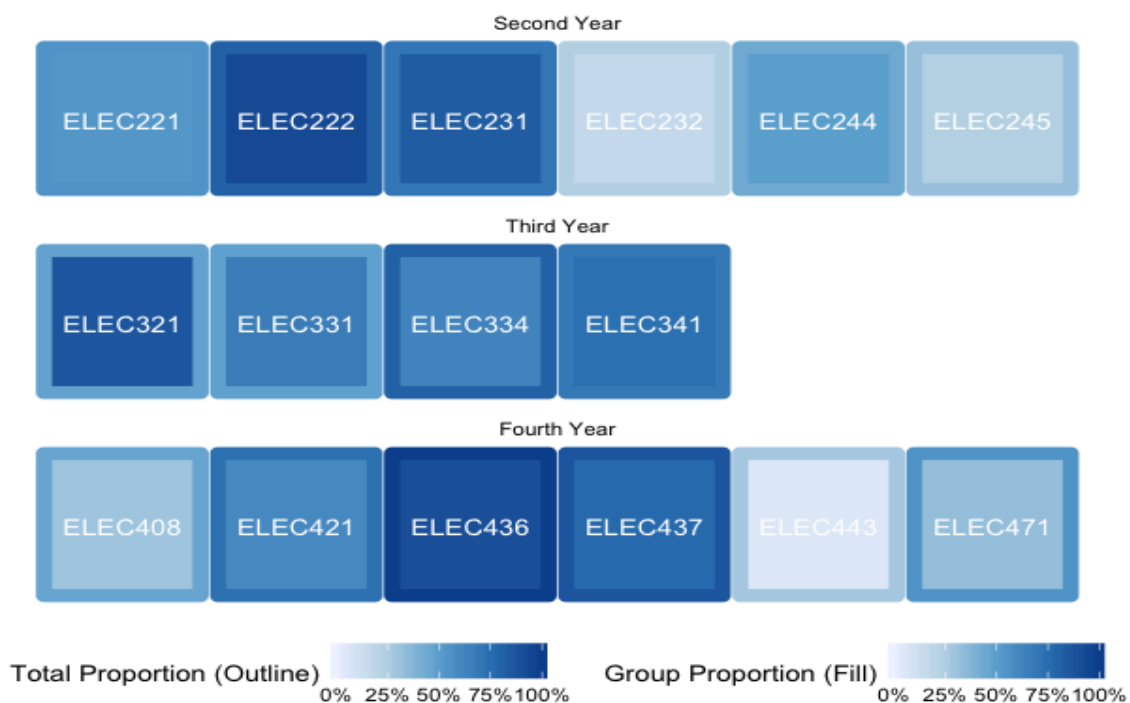
Step 3: Refining Visualization

After creating the initial version of the visualization as seen in Figure 4, two issues are apparent. First, while a user can tell what courses are taken most or least among this demographic group, they have no way of knowing whether that is different than the larger group. As such, there should maybe be a way to compare that, but in a way that is less salient than the demographic group’s selection rates. Secondly, while the selection rates are easy to compare against other courses in each year of study, it is less obvious which are the most and least selected courses among all years.

There are a number of ways we can reassign channels to accommodate these issues. We could add separate bars for each course to compare with the demographic group’s selection rates. However, this might distract from the visualization’s primary purpose. An alternative approach would be to use the colour channel for the total selection rates, but then it would be extra work to compare the group’s selection rates with the total’s, since one would be on the position channel while the other would be colour. Instead, we can use colour for both of the channels while making one less salient, as seen in Figure 4. Here, each bar is filled with

the colour corresponding to how much of the demographic group took the course, while outlined by how much of the total population took the course according to the legend below. In this version, we can still clearly identify which courses are taken the most or least, and further, we can then compare to see whether that course selection pattern is abnormal when compared to the total proportion. For example, ELEC 221 is taken by around half of the students, whether they are in the demographic group or not. ELEC 232 is not taken by many students, also regardless of demographic group. However, we can also see that ELEC 408 is less popular among this demographic group than the total population. In contrast, ELEC 321 is taken by a much larger portion of the demographic group than the total population and might be worth further inquiry.

Figure 5: Updated Version of Visualization for “What are the most commonly selected courses for students of a particular demographic group?”



Notice, that it is not as easy to compare exact values of the demographic group’s selection rates. That is sufficient however, since comparison is not the purpose of the visualization. The main purpose is to identify which ones are most and least selected, which is apparent through this method. However, if that information is required it could be added as smaller text elements on each block.

There are other items that might be considered as well. If certain courses among the set are prerequisites for each other, it might be sorting them into regions to make those connections easily apparent.

References

- Bertin, J. (1981). *Graphics and graphic information processing*. Berlin: Walter de Gruyter.
- Börner, K. & Polley, D. E. (2014). *Visual insights: A practical guide to making sense of data*. Boston: MIT Press.
- Munzner, T. (2014). *Visualization Analysis and Design*. CRC Press.
- Sivanand, A. & Frank, B. (2017). Information Visualisation in Education: A Review of Current Tools and Practices. *Proceedings of the Canadian Engineering Educational Association*.
- Sivanand, A. (2017). Supporting post-secondary educational data usage in the assessment process with information visualization. Doctoral Dissertation, Queen's University.
- Tufte, E. R. & Graves-Morris, P. R. (1983). *The visual display of quantitative information Vol 2, No 9*. Cheshire, CT: Graphics Press

Glossary of Terms

Action: An action is what a visualization enables a user to do with data. Actions are categorized into three levels: analyze, search and query.

Analyze: One of the three levels of action, analyze is the level at which a visualization is either meant to present known information to others or discover new information. Discovering new information is further subdivided into *exploration* (when the analysis is being done by a single party) and *discussion* (when the visualization is used as a tool to facilitate group analysis).

Attribute: The data associated with, and containing characteristics of, each item. Attributes are divided into key attributes and value attributes; they correspond to columns in a data set. Key attributes can be further divided into categorical or ordinal. Value attributes can be categorical, ordinal or quantitative.

Categorical: One of the three subtypes of attributes along with ordinal and quantitative, categorical data does not have any implicit ordering.

Channel: A channel is one of the ways that you can control a visual mark's appearance such as position, shape and colour. They are used with marks to represent magnitude (quantitative data) or identity (categorical or ordinal data).

Item: An item is the individual unit of analysis for each area of inquiry. Items correspond to rows in a data set.

Key attribute: One of the two types of attributes, a key attribute is one that can uniquely identify an item. Keys are always connected to the items of each visualization. At least one key is typically required, but multiple keys can be required to identify grouping attributes for more complex questions. Keys can only be categorical or ordinal, *not* quantitative.

Mark: A mark is a basic geometric element that depicts an item such as points on scatter plots or lines in bar charts. A mark's appearance can be controlled by channels.

Ordering direction: Ordering direction applies to ordinal and quantitative data and can be sequential, diverging or cyclic.

Sequential: One of the three categories of ordering direction, an attribute is sequential when it is a range that has a clear minimum and maximum.

Diverging: one of the three categories of ordering direction, an attribute is diverging when it has two sequences that can increase in opposite directions and meet at a common zero point.

Cyclic: One of the three categories of ordering direction, an attribute is cyclic if its values wrap around to a starting point.

Ordinal: One of the three subtypes of attributes along with categorical and quantitative, ordinal data is implicitly ordered data that *cannot* have arithmetic performed on it to create new data.

Quantitative: One of the three subtypes of data along with categorical and ordinal, quantitative data is implicitly ordered data that *can* have arithmetic performed on it to create new data.

Query: The lowest level of defining an action in the task abstraction process, querying involves either identifying one target, comparing multiple targets or summarizing all targets, a determination that depends on the number of targets.

Search: One of the three levels of defining an action in the task abstraction process, at the search level, consideration of whether the identity of the target is known prior to looking at the visualization, as well as if the location of the target among the data is known. The location often depends on the specificity of the question.

Target: A target is an aspect of the data being analyzed or what a user is trying to analyze from a visualization. With all data, the target can be a trend, an outlier, or when more specific, a feature. When dealing with a single value attribute, the target could be the distribution of the data or the extreme items in the data. When there are multiple attributes, the target can be a dependency or a correlation.

Value: The datum associated with an attribute of a certain item.

Value attribute: One of the two types of attributes, a value attribute is one that is a characteristic of an item. Values are determined by identifying what recorded measurements or observations are needed to answer a question. Value attributes can be categorical, ordinal or quantitative.

Summary of Data Abstraction Categories and Options

| Data Set Abstraction | | Attribute Abstraction | |
|-----------------------|---|-----------------------|--|
| Abstraction Category | Options | Abstraction Category | Options |
| Data Types | <ul style="list-style-type: none"> • Item • Attribute | Attribute Semantic | <ul style="list-style-type: none"> • Key attribute • Value attribute |
| Data Set Types | <ul style="list-style-type: none"> • Tabular Data | Attribute Types | <ul style="list-style-type: none"> • Categorical • Ordinal • Quantitative |
| Data Set Availability | <ul style="list-style-type: none"> • Static • Dynamic | Ordering Direction | <ul style="list-style-type: none"> • Sequential • Diverging • Cyclic |



Higher Education
Quality Council
of Ontario

An agency of the Government of Ontario