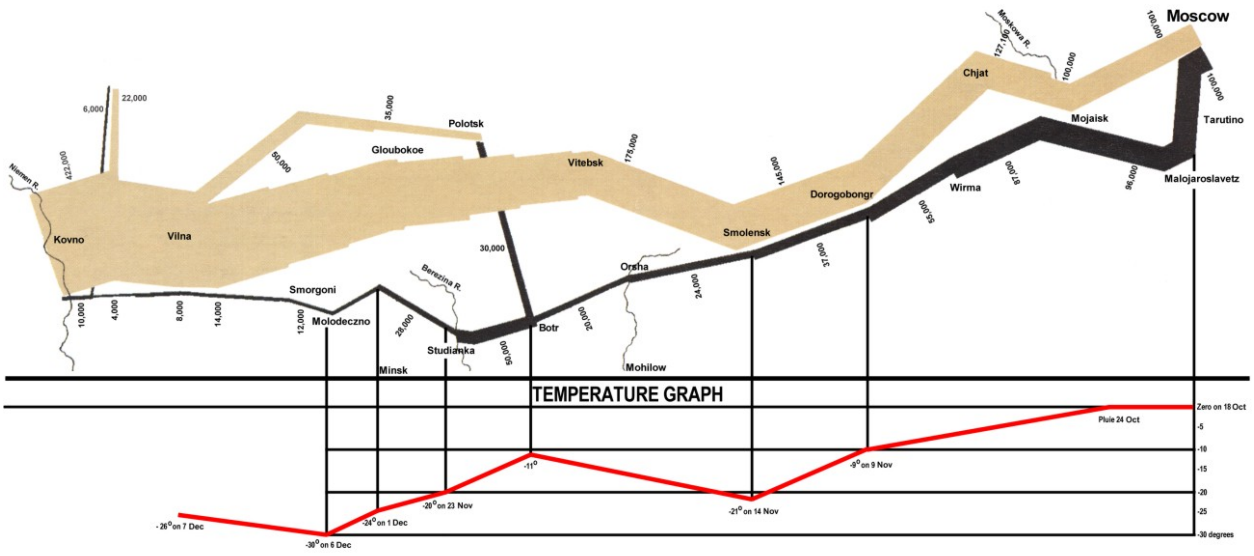
# Data Visualization
# CSCI 444 (UG)
# Fall 2022 Syllabus

> The minimum we should hope for with any display technology is that it should do no harm.
>
> –EDWARD TUFTE



## Instructor Details


| | |
|---|---|
| **Name:** | Professor Jesse Johnson |
| **Office:** | 417 Social Science Building |
| **Telephone:** | (406) 243-2356 |
| **Email:** | `jesse.johnson@umontana.edu` |
| **Web:** | Faculty Home Page |
| **Office Hours:** | MW 15:00–17:00 |
| | *Or, by appointment.* |


## Prerequisites

Students taking this course are expected to have:

- Experience with modern, complex quantitative software libraries to the degree that *independently* mastering several new software in a semester does not present a problem.
- Organizational skills and familiarity with computers sufficient to install new software, create a file system for the course, and execute programs and move files from the command line.
- Some programming experience with any language.
- Evidence of mathematical maturity as shown by successful completion of calculus and/or statistics.
- Maturity enough to show up for class, consistently.
- Maturity enough to offer constructive criticism to your peers.

## Course Objectives

This course emphasizes practice over theory, compelling students to master the tools required to produce high quality visualizations. As such, a majority of the student's time is spent in the creation of original visuals. Through the course, students will encounter data having different relations between members, each of which presents its own challenges in terms of visualization. The student driven process of creation will be complemented with in-class discussions of reading material that emphasize a framework for assessing the quality of visualizations, the technical skills required, and the mathematical concepts providing the structure for data visualization techniques.

## Student Outcomes

Upon successful completion of this course, student will be able to:

1. critically evaluate visualizations and suggest improvements and refinements.
2. design, implement, and evaluate a computer-based process to for the visualization of data.
3. use principles of human perception and cognition in visualization design.

4. precondition data sets to make them readily accessible to the visualization software being used.

5. create web-based interactive visualizations.

6. quickly adapt to any quantitative visualization programming environment.

## Textbook

This semester I'll be using the following text. You need to purchase a copy.

**Visualization: Analysis and Design**
*Tamara Munzner*
CRC Press
2014

You do not have to purchase the following, but it does inform much of what is discussed.

**The Visual Display of Quantitative Information**
*Edward Tufte*
Graphics Press
2001

## Course Logistics

### Hypothesis Driven Visualization

The course is project driven. All projects share the same approach, but differ in data sets used. This semester, expect to do 6-7 of these projects. Given a data set, your approach will always be as follows:

1. Form a hypothesis from the data. State the hypothesis as clearly as possible.

2. Produce a visual from the data to support your hypothesis. Your visual can and should have multiple panels, but can be no more than one 8.5 by 11 inch page when printed without rescaling.

3. Produce a second visual that has arrows and labels that indicate how and where your visual is consistent with best practices discussed in class. A way to do this is to create your visual, print it out, mark it up with a pen, photograph it with your phone, and then submit that. Others prefer to annotate with software. Either is fine.

Rubrics for grading projects will be provided to students before they are due.

### Required Software

This semester I am requiring students to complete their visualizations as web pages. I suggest you do this using a github account. Some software that will make this possible includes but is not limited to:

- D3 – a `javascript` library with some very insightful approaches to data visualization. Using this library can be a fair amount of work, but the results are strong.
- Vega-lite A grammar for visualization. Like a language. Super interesting and quite powerful. The biggest drawback seems to be the lack of a good editor for the `json` files used.
- Vega as above but slightly lower level.
- plotly I don't know a lot about this but it appears to satisfy the requirement of web based.
- Tableau (only for non-CS majors) - A commercial product that is like Excel with good visualization abilities. I like it, but licensing and use are restrictive. Student licenses are available.

All assignments will be submitted as a link to a web page displaying the information.

A good IDE can help a lot. Here's one that we've used: Webstorm. You should also install Microsoft Teams and login with your UM account.

### Online Resources

As in previous years, there is a Moodle supplement to the course.

In addition, I've created a Microsoft Teams supplement. I recommend we use Teams to discuss course related issues as it is a better forum than what is provided for Moodle. I also find it to be an easy way to talk to students that can not get in for office hours.

## Assignments

Much of the learning in this class achieved by doing. I've been developing assignments over the years and they break down about as follows. I'll continue to refine these through the semester, so the final form may be slightly different from what appears here.

Throughout, data are required. There are no constraints on the data you use for your assignments, provided the *data relation* is upheld. One excellent clearing house for data is awesome-public-datasets.

1. ## The scatter plot and histogram

   **Data Relation:** One to one relations, or ordered-pairs for the scatter plot, and categorical count data for the histogram.

   **English Language Description:** The most common types of data visualization for a reason, plots of ordered-pairs reveal trends in data that can only be seen in a graphical construction. Fundamentally, the relation is that of independent and dependent variables. A histogram reveals the distribution of data points, suggesting a mean and standard deviation.

**Example Data Sources:** The bureau of labor statistics provides a lot to think about: Bureau of Labor Statistics. If work and money isn't your thing, try disease Center for Disease Control.

**Example Hypotheses:** Montana's wages are higher for workers without college degrees than wages for similar workers in Washington, Oregon, and Idaho. Summer is the best season to contract an STD, except when the carnival comes to town. The period having the highest incidence of tuberculosis in the United States corresponded to the period of highest troop deployments overseas.

## 2. The geographic vector data

**Data Relation:** A set of points, distributed on the surface of Earth, or other planet. Each point registers the location of a feature of interest. In some cases, there is a relationship between the points, as they define a road, or a political boundary. Other times they are independent, like cities.

**English Language Description:** Here, ordered pairs are longitude and latitude. Special care must be taken to assure that the projection of points on the surface of a sphere (lat./long. pairs) to a plane (such as a Mercator projection) is done correctly and consistently.

**Data Sources:** Political boundaries are held within many (all?) states as part of their digital assets. Aggregations are also common at places like geoboundaries.org. Open street map is an amazing (overwhelming?) resource. There are numerous vector maps of interesting points, like cities, cell phone towers, mines, Walmarts, etc. Consider some of the historical databases too, for example, there are databases of ancient cities that include features for the cities.

**Example Hypotheses:** Rivers and cities are usually close. The largest rate of growth is currently in cities in the mountains. Coastal cities have more irregular networks of roads around them than do cities on the plains.

## 3. The raster plot

**Data Relation:** A matrix of values.

**English Language Description:** A two-dimensional plane is broken into cells. In each cell, there is a value. Such data are used to represent everything from satellite data to the output of a mathematical function. More abstractly, heat maps are used to show a relation between variables or explore a larger set of values, such as days of the year.

**Example Data Sources:** Students are often interested in `geoTIFF` or `netCDF` formatted data because these datasets are geographic in nature - each cell has a specific location on the planet. As such, you'll find many potential sources, but a few that are interesting include National Snow and Ice Datacenter, Socioeconomic Data and Applications Center, and NASA Earth Observatory. State and local municipalities have also started to provide a lot of geographic data, check out government web sites. Also feel free to generate a surface representing a mathematical functions you are interested in, or a heat map that describes a relationship between variables.

**Example Hypotheses:** The extent of glaciated areas has decreased more in the 90s that the period from 2000 to 2010. The fastest growing populations are in equatorial regions of the world. The Mandelbrot set's self similarity breaks down quickly if single precision numbers are used.

## 4. The network

**Data Relation:** A graph, or numerous one-to-one and one-to-many relations within the same data set. The relations may imply a hierarchy, for example in the case of folders and files on a computer, or not, such as a network of friends on social media.

**English Language Description:** Networks are graphs of complex relations between nodes. A node might be a person, and edge would denote a friendship. Evaluation of these relations is an open problem and useful for a addressing questions ranging from marketing to radicalization of terrorists.

**Example Data Sources:** Sometimes smaller is easier in this assignment, try this site. There are many large networks available for analysis, but only aggregate data can be visualized in a meaningful way.

**Example Hypotheses:** Networks often have a highly connected node that, if removed, would result in several disjoint networks. Network centrality is a good indicator of how to disrupt networks. The flow of information through a network depends on the degree distribution.

**Suggested Software:** Gephi can help compute several network statistics that might inform your visualization. We will cover some of the formula in class too.

## 5. High-dimensional data

**Data Relation:** Data having at least three dimensions.

**English Language Description:** Display data that includes three or more dimensions. Sometimes this is time plus two dimensional spatial data, but there are many other options.

**Example Data Sources:** All the above, and beyond. It's also possible to take simulation output as a source in this assignment. If you go that route, then I recommend Paraview as a visualization tool.

**Example Hypotheses:** Completely unconstrained.

## 6. Multi-view display or data dashboard or film

**Data Relation:** Multiple, some span of data used in previous assignments. Students may also explore the use of film to document a scientific phenomena.

**English Language Description:** This assignment will be directed towards a web service that displays multiple views of related datasets and automatically updates the view. Student that elect to produce a short film will have to document a scientific phenomena an justify their camera choices such as focal length, aperture, shutter speed, frame rate, exposure, autofocus, and noise reduction.

**Example Data Sources:** Any above and beyond.

**Example Hypotheses:** This doesn't fit as well here. Films do not have hypothesis. A dashboard may test several and should have an overarching hypothesis.

## Graduate Increment

This is a **"UG"** course, meaning it can be taken for graduate credit.

If you are a graduate student taking the course, then you'll be required to do one additional assignment. You can do use any dataset you like for this assignment. You will present the results of this assignment to the class during the final examination meeting. Expectations for this assignment will be high, but the grading rubric used will be the same as that used for the other six projects.

## Meeting Times/Place

**Times:** Monday, Wednesday, Friday 14:00–13:50
**Place:** Social Science 362

## Final Exam Time and Place

To be clear, this is the time when students receiving the graduate increment will present their final projects. There are no examinations in this course. All students are required to attend.

*Due to my travel schedule this is likely to be remote.*

**Time:** 13:10-15:10, Thursday, December 15
**Place:** Social Science 362

## Grading Policy

### Grading scale

| | |
|---|---|
| A | 94-100 |
| A- | 90-93 |
| B+ | 87-89 |
| B | 83-86 |
| B- | 80-82 |
| C+ | 77-79 |
| C | 73-76 |
| C- | 70-72 |
| D+ | 67-69 |
| D | 63-66 |
| D- | 60-62 |
| F | 0-59 |

Students achieving the numerical scores above are guaranteed the associated letter grade. However, if average performance is low, I may decide to assign a higher letter grade for a lower score; e.g. a B+ for a numerical score of 84.

Students taking the course pass/no pass are required to earn a grade of D or better in order to pass.

### Assessments and weights

The following assessments will be used and weighted according to the values in the table to determine final grades.

| Component | Description | Weight (UG/G) |
|---|---|---|
| Projects | A total of 6 projects that use various sets of data and software tools. | 70/50% |
| Final project | A more comprehensive final project based on data students identify. | 0/20% |
| Group/classroom | Some grades will be developed from participation in classroom and online activities. A major component will be presentation of academic papers. | 30/30% |

## Attendance Policy

Students absent when called up to do in-class work will be given a grade of 0%. Students informing the instructor of a valid reason for missing class *in advance*, via email, will not be called upon. Valid reasons include family emergencies and illness. I may ask for documentation of absence (doctors note, obituary, etc.). Cell phone photos are useful for this - a selfie in the doctors office, or next to a car that won't start...

## Academic Integrity

All students must practice academic honesty. Academic misconduct is subject to an academic penalty by the course instructor and/or a disciplinary sanction by the University. All students need to be familiar with the Student Conduct Code. I will follow the guidelines given there. In cases of academic dishonesty, I will seek out the maximum allowable penalty. If you have questions about which behaviors are acceptable, especially regarding use of code found on the internet or shared by your peers, please ask me.

## Disabilities

Students with disabilities may request reasonable modifications by contacting me. The University of Montana assures equal access to instruction through collaboration between students with disabilities, instructors, and the Office of Disability Equity (ODE). "Reasonable" means the University permits no fundamental alterations of academic standards or retroactive modifications.

# A guess at a schedule:

| MONDAY | WEDNESDAY | FRIDAY |
|---|---|---|

| MONDAY | WEDNESDAY | FRIDAY |
|---|---|---|
| Aug 29th **1**<br>Visualization Zoo | 31st **2**<br>Munzner Chapter 1 | Sep 2nd **3**<br>The value of visualization |
| 5th<br>*Labor Day* | 7th **4**<br>Webstorm, HTML, JavaScript, Vega-lite | 9th **5**<br>Perception in Visualization |
| 12th **6**<br>Munzner Chapter 2 | 14th **7**<br>Assignment 1 Due: Ordered pairs and histograms | 16th **8**<br>DOM manipulation and git |
| 19th **9**<br>On the theory of scales and measurement | 21st **10**<br>Munzner Chapter 3 | 23rd **11**<br>Crowdsourcing graphical perception: using mechanical turk... |
| 26th **12**<br>Munzner Chapter 10 | 28th **13**<br>Raster plots, and geographic projections in Vega-lite | 30th **14**<br>How to pick more beautiful colors for your data visualizations |
| Oct 3rd **15**<br>Assignment 2 Due: Rasters | 5th **16**<br>JSON data | 7th **17**<br>Mike Bostock's Let's make a Map |
| 10th **18**<br>Munzner Chapter 4 | 12th **19**<br>The good, the bad, and the baised: five ways visualizations can mislead (and how to fix them) | 14th **20**<br>USA Temperature: can I sucker you? |
| 17th **21**<br>Assignment 3 Due: Geographic Vectors | 19th **22**<br>Interactive Content | 21st **23**<br>Munzner Chapter 5-6 |
| 24th **24**<br>Jerome Cukier's tutorial on scales | 26th **25**<br>Munzner Chapter 9 | 28th **26**<br>Assignment 4 Due: Networks |
| 31st **27**<br>Interactive dynamics for visual analysis | Nov 2nd **28**<br>Munzner Chapter 7 | 4th **29**<br>D3 and JavaScript |
| 7th **30**<br>Taggle: Scalable Visualization of Tabular Data through Aggregation | 9th **31**<br>Surprise! Bayesian Weighting for De-Biasing Thematic Maps | 11th<br>*Veterans Day* |

| MONDAY | WEDNESDAY | FRIDAY |
|---|---|---|
| 14th **32**<br>Munzner Chapter 8,11 | 16th **33**<br>The State of the Art in Visualizing Multivariate Networks | 18th **34**<br>Assignment 5 Due: High-dimensional data |
| 21st **35**<br>Munzner Chapter 12,13 | 23rd<br>*Thanksgiving Travel Day* | 25th<br>*Thanksgiving Holiday* |
| 28th **36**<br>Narrative Visualiztion: Telling Stories with Data | 30th **37**<br>Munzner Chapter 14 | Dec 2nd **38**<br>Points of view: Sets and intersections |
| 5th **39**<br>A nested model for visualization design and validation | 7th **40**<br>Assignment 6 Due: Dashboard | 9th **41**<br>Flexibility in schedule |
| 12th **42**<br>Wrap up/Course evaluation | 14th **43**<br>Finals Week: Presentations on Tuesday | 16th **44**<br>Finals Week |