



## CSCI 4550/6550 Interactive Visualization

<https://www.cs.rpi.edu/~cutler/classes/visualization/S24/>

# Lecture 20: Additional Topics

## Today

---

- Quiz 2
- Scatterplots & Uncertainty
  - "Scatterplots: Tasks, Data, and Designs"
  - "Winglets: Visualizing Association with Uncertainty in Multi-class Scatterplots"
- Readings for Today
  - "Why Authors Don't Visualize Uncertainty"
  - "Beyond Generating Code: Evaluating GPT on a Data Visualization Course"
- Readings for Friday

## Quiz 2 on Tuesday, April 9th

---

- During normal class time, 2-3:50pm
- No laptops/phones/watches/etc.
- 1 page (double-sided) of notes allowed, handwritten or printed
- Sample problems on the calendar
- Crayons/colored pencils/markers will be provided

## Today

---

- Quiz 2
- Scatterplots & Uncertainty
  - "Scatterplots: Tasks, Data, and Designs"
  - "Winglets: Visualizing Association with Uncertainty in Multi-class Scatterplots"
- Readings for Today
  - "Why Authors Don't Visualize Uncertainty"
  - "Beyond Generating Code: Evaluating GPT on a Data Visualization Course"
- Readings for Friday

# “Scatterplots: Tasks, Data, and Designs”

Sarikaya & Gleicher, InfoVis 2017

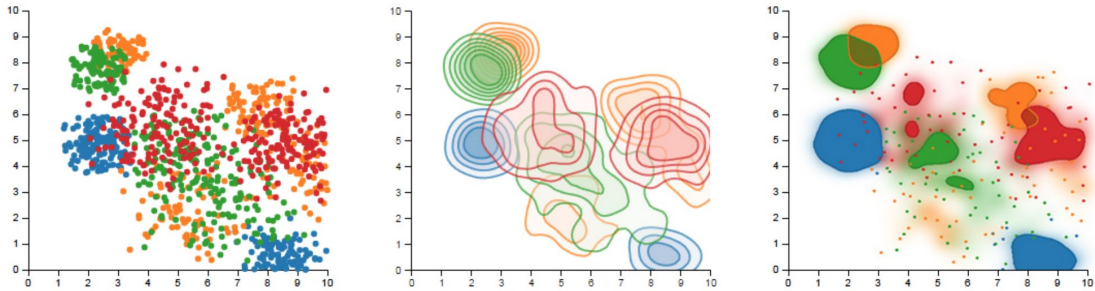
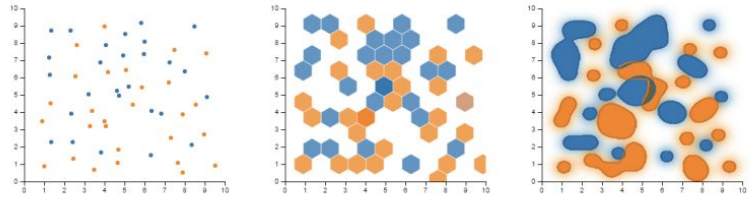


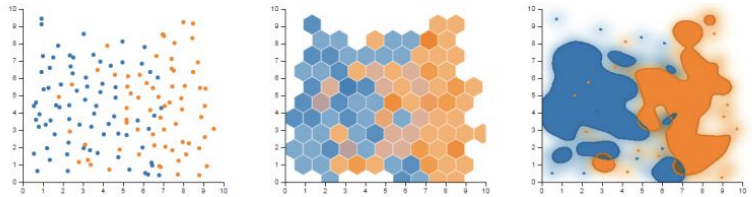
Fig. 3. Three different designs (left-to-right: traditional scatterplot, contour map [19], and Splatterplot [49]) display different information about the same 100 item, four class (mapped to color) dataset. While the traditional scatterplot exhibits some overdraw, the two alternative approaches use point grouping techniques to emphasize numerosity and distribution comparison tasks. The contour map conveys density gradients, while the Splatterplot uses thresholded regions to convey dense areas.

- Depending on viewer task:
  - Identify outliers
  - Comparing distributions
- Concerns:
  - Too much data → Overdraw (& drawing order)
- Possible Solutions:
  - Reduce data / sampling
  - Bin the data
  - Density estimation
  - SPLOMS (scatterplot matrix) & scagnostics (scatterplot diagnostics)
- Developed taxonomy of scatterplot analysis tasks & collected range of scatterplot design decisions (survey of visualization papers)
- Developed a table of recommendations
- Some open problems remain... handling large total data sizes & large # of categories (> # of visibly distinct colors)

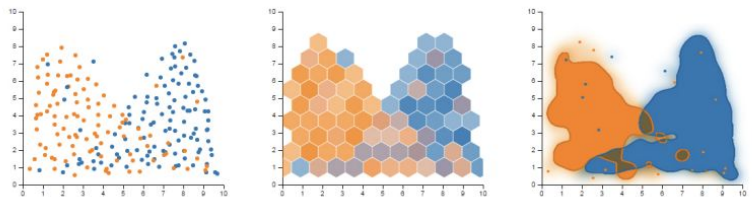
For random distributions with few points (top row), the traditional scatterplot (left) describes the data plainly.



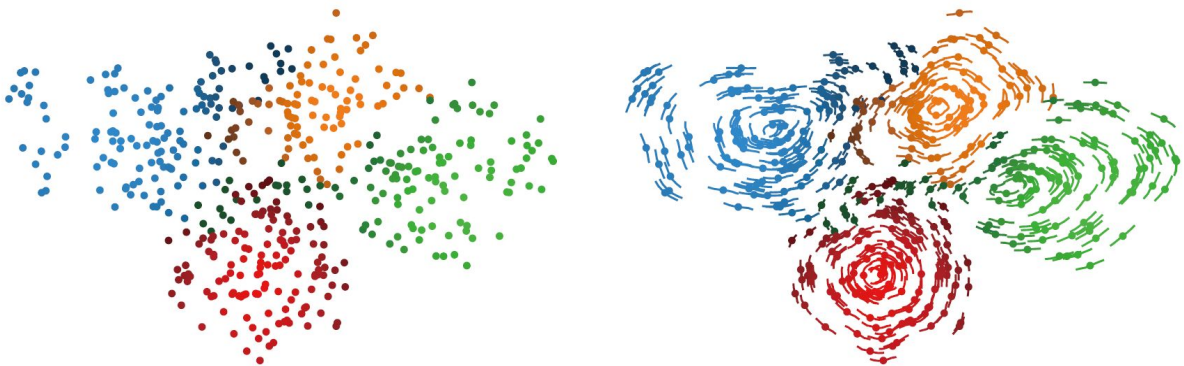
With increasing numbers of points (middle row), aggregation representations such as binning (center) communicate spatial density.



With overlapping distributions (bottom row), density-based representations communicate overlap and can also show outliers (right), which disappear in the binned representation (middle).



"Winglets: Visualizing Association with Uncertainty in Multi-class Scatterplots", Lu, Wang, Lanir, Fish, Yue, Cohen-Or, and Huang, IEEE InfoVis 2019



The Green cluster does not overlap with the two clusters in the original high dimensional space, while the Orange and Blue ones do. The wings help to perceive the global structure of the clusters, despite the uncertainty and overlap.



## Today

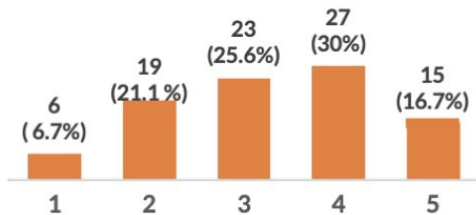
---

- Quiz 2
- Scatterplots & Uncertainty
  - "Scatterplots: Tasks, Data, and Designs"
  - "Winglets: Visualizing Association with Uncertainty in Multi-class Scatterplots"
- Readings for Today
  - "Why Authors Don't Visualize Uncertainty"
  - "Beyond Generating Code: Evaluating GPT on a Data Visualization Course"
- Readings for Friday

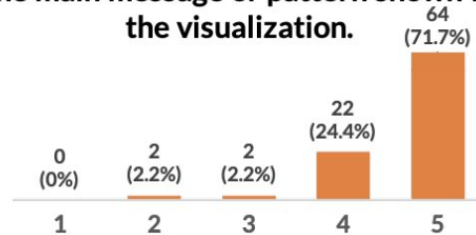
## Reading for Tuesday *pick one*

- "Why Authors Don't Visualize Uncertainty", Hullman, IEEE VIS 2019

(c) Most audiences cannot understand uncertainty information.



(d) Uncertainty information should be presented when it calls into question the main message or pattern shown in the visualization.



- It is technically difficult to visualize uncertainty effectively & intuitively
- Uncertainty is viewed negatively, might be rejected
  - expectation that visualizations will deliver a clear message
- Worried the readers won't understand
  
- It's irresponsible to not be transparent about uncertainty
- Uncertainty is key part of the story of the data

# Today

- Quiz 2
- Scatterplots & Uncertainty
  - "Scatterplots: Tasks, Data, and Designs"
  - "Winglets: Visualizing Association with Uncertainty in Multi-class Scatterplots"
- Readings for Today
  - "Why Authors Don't Visualize Uncertainty"
  - "Beyond Generating Code: Evaluating GPT on a Data Visualization Course"
- Readings for Friday

# Reading for Tuesday *pick one*

- "Beyond Generating Code: Evaluating GPT on a Data Visualization Course", Chen, Zhang, Wang, Troidl, Warchol, Beyer, Gehlenborg, & Pfister, IEEE EduVis 2023

**a) Data Cleanup and Exploration**

Please clean this dataset and create an interactive visualization for users to explore it.

Sure! This is the python code to clean the dataset:

```
python
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

You can explore the cleaned dataset using this interactive scatterplot created in D3.js:

```
import('https://d3js.org').then(data => {
  const margin = { top: 20, right: 20, bottom: 20, left: 20 };
  const width = 100 - margin.left - margin.right;
  const height = 100 - margin.top - margin.bottom;
  // ...
});
```

**b) Read and Interact with Vis**

This is the visualization you created for the dataset. What insights do you read from it? You can interact with it programmatically.

```
html
<g class="context"><path class="area"
d="M0,0,62354,..."
```

This appears to be an SVG representation of a data visualization, specifically a scatterplot. From this visualization, I find the following insights:

1. ....

Please run this code to interact with it:

```
svg.selectAll('rect').on('mouseover', function() {
  // ...
});
```

**c) Present and Explain Insights**

These are the insights you discovered. Please create a scrollytelling with visualizations to present the findings for general audiences:

1. ....
2. ....

Sure! This is a scrollytelling, which tells....

**Insight 1**

There appears to be a positive correlation between log(gdp\_per\_capita) and healthy\_life\_expectancy. The visualization also shows that the correlation is stronger for Central and Eastern Europe, South America and Caribbean, Middle East and North Africa, Sub-Saharan Africa, and Sub-Saharan Africa (excluding high income).

Healthy Life Expectancy

log(gdp\_per\_capita)

Central and Eastern Europe  
South America and Caribbean  
Middle East and North Africa  
Sub-Saharan Africa  
Sub-Saharan Africa (excluding high income)

Figure 1: Our experiments show that GPT can a) clean and explore CSV datasets, b) read visualizations in SVG format, interact with visualizations through dispatching Javascript events, and c) create explanatory visualizations to present data insights.

- Prior work with ChatGPT
  - “GPT makes coherent and contextually appropriate text”
  - Can “pass” graduate, medical, & law school standardized tests
  - Has been used to generate code for visualizations...
- Let’s have ChatGPT take a Visualization Course -
  - Evaluate on tasks of data interpretation, visualization design, data exploration, & insight communication
- Quizzes & tests
  - 91 fully-autogradeable quiz questions
  - 9 homeworks - essays & coding activities with sample data
  - Scored 80% on quizzes & homework
- Graded by 3 human TAs
  - ChatGPT answers mixed with student answers from prior terms
  - TAs could distinguish human vs GPT with 70% accuracy

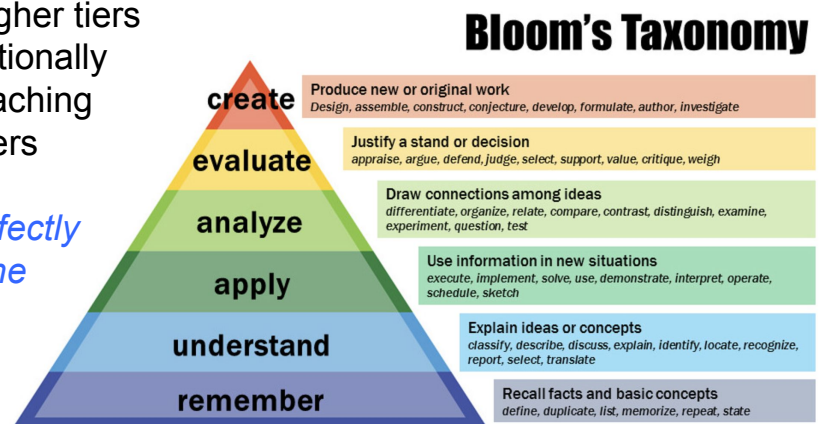
- What are the pros of ChatGPT?
  - Helps people learn a new programming language or toolkit
  - Helps people learn new-to-them functions or features
  - Generates tutorial-like examples of usage
- What did ChatGPT fail at?
  - Cannot read images - *presumably this will be added in future*
  - Confused by format of ‘matching problems’ - *presumably this will be fixed in future*
  - Couldn’t do homework with lengthy input or multiple files (limits on max tokens/lines of code) - *presumably will be expanded in future*
  - Wasn’t as “creative” as human students
- Final Project – *had to break open-ended final project into parts*
  - Can collect & clean & merge data
  - Can read SVG images and interact with JS visualizations
  - Can interpret information and write about it in different styles



- Should we be worried about ChatGPT generating visualizations?
  - Where is the data coming from?
  - Did it “hallucinate” the data?
  - Is the data correct, but the visualization is biased or misleading?
  - Ongoing concerns with AI about ethics, equity, & inclusion
- “Need to redesign Visualization Education”
  - To prevent students from cheating by using ChatGPT?
  - Because people with careers in visualization will need different skills?
- Emphasize areas where humans are irreplaceable (& ChatGPT fails)
  - E.g., open-ended questions, sketching, unique information from class, group projects that promote teamwork and collaboration
- Integrate ChatGPT into the classroom
  - Teach students to learn from ChatGPT and to become prompt engineers, to coax creativity out of ChatGPT

## Bloom’s Taxonomy for Education (1956, 2001)

- Like the original food pyramid...
- Starting at the foundation tier (bottom) and build skills at each higher tiers
- Should spend proportionally more time & effort teaching & learning at lower tiers
- *ChatGPT can do the bottom tiers near-perfectly*
- *Does this eliminate the need (or ability) to teach, learn, & test students on skills in lower tiers?*



- ChatGPT & Education
  - Can instructors use ChatGPT to write & deliver lectures?
  - Can instructors use ChatGPT to write homeworks & quizzes?
  - Can TAs use ChatGPT to grade?
  - Can instructors use ChatGPT to give end-of-term grades?
  - Can AI replace instructors & TAs?
  - Should you be paying for an education or just use ChatGPT?
- ChatGPT & your future career
  - Can you use ChatGPT in an interview?
  - Can ChatGPT decide whether you should be hired?
  - Do you want colleagues who just use ChatGPT to do their work?
  - Or to manage employees who use ChatGPT to do their work?
  - Will an improved AI take your job in the future?

## Scatterplots

Definition:

A scatterplot is a type of data visualization that displays the relationship between two numerical variables. Each data point in the plot represents an observation, with one variable on the x-axis (horizontal axis) and the other variable on the y-axis (vertical axis).

Key Components:

1. **X-axis:** Represents one variable (usually the independent variable).
2. **Y-axis:** Represents the other variable (usually the dependent variable).
3. **Data Points:** Each data point on the plot represents a single observation from the dataset.
4. **Trend Line:** In some cases, a trend line may be added to the plot to show the overall trend or relationship between the variables.

#### Purpose:

- Identify relationships: Scatterplots help in visualizing the relationship between two variables. They can reveal patterns such as linear, non-linear, or no relationship.
- Outlier detection: Outliers are data points that deviate significantly from the general pattern in the data. Scatterplots can help in identifying outliers.
- Clustering: Scatterplots can reveal clusters or groups within the data if they exist.

#### Interpretation:

- **Positive Relationship:** If the data points generally trend upwards from left to right, it indicates a positive relationship between the variables. This means that as one variable increases, the other variable also tends to increase.
- **Negative Relationship:** If the data points generally trend downwards from left to right, it indicates a negative relationship between the variables. This means that as one variable increases, the other variable tends to decrease.
- **No Relationship:** If the data points appear scattered without any clear pattern, it suggests that there is no significant relationship between the variables.

#### Example:

Consider a dataset containing information about students' study hours and exam scores. A scatterplot of study hours (x-axis) versus exam scores (y-axis) can help visualize whether there is a relationship between the two variables. If there is a positive relationship, we would expect to see higher exam scores for students who study more hours.

#### Tips for Effective Scatterplots:

1. Choose appropriate scales for the axes.
2. Label the axes clearly to indicate the variables being represented.
3. Use different colors or markers to distinguish between different groups or categories within the data.
4. Add a title to the plot to provide context and explanation.
5. Include a legend if necessary, especially when multiple groups are represented in the plot.

### Quiz: Scatterplots

1. What is a scatterplot?
  - A) A type of pie chart
  - B) A data visualization that displays the relationship between two numerical variables
  - C) A bar chart with horizontal bars
  - D) A line plot showing changes over time
2. What do the data points in a scatterplot represent?
  - A) Categories in the dataset
  - B) Observations or individual data points
  - C) Mean values of the variables
  - D) Standard deviations of the variables
3. How can outliers be identified in a scatterplot?
  - A) By their position relative to the origin
  - B) By their color
  - C) By their position relative to the overall pattern of the data
  - D) Outliers cannot be identified in scatterplots
4. In a scatterplot, what does a positive relationship between variables indicate?
  - A) As one variable increases, the other variable tends to decrease.
  - B) There is no relationship between the variables.
  - C) As one variable increases, the other variable tends to increase.
  - D) The variables are negatively correlated.
5. Which of the following is NOT a tip for creating effective scatterplots?
  - A) Choose appropriate scales for the axes.
  - B) Label the axes clearly.
  - C) Use different colors or markers to distinguish between groups.
  - D) Remove all data points that do not fit the expected pattern.

### Quiz: Scatterplots

1. What is a scatterplot?
  - A) A type of pie chart
  - B) A data visualization that displays the relationship between two numerical variables
  - C) A bar chart with horizontal bars
  - D) A line plot showing changes over time
2. What do the data points in a scatterplot represent?
  - A) Categories in the dataset
  - B) Observations or individual data points
  - C) Mean values of the variables
  - D) Standard deviations of the variables
3. How can outliers be identified in a scatterplot?
  - A) By their position relative to the origin
  - B) By their color
  - C) By their position relative to the overall pattern of the data
  - D) Outliers cannot be identified in scatterplots
4. In a scatterplot, what does a positive relationship between variables indicate?
  - A) As one variable increases, the other variable tends to decrease.
  - B) There is no relationship between the variables.
  - C) As one variable increases, the other variable tends to increase.
  - D) The variables are negatively correlated.
5. Which of the following is NOT a tip for creating effective scatterplots?
  - A) Choose appropriate scales for the axes.
  - B) Label the axes clearly.
  - C) Use different colors or markers to distinguish between groups.
  - D) Remove all data points that do not fit the expected pattern.



ChatGPT

Here are the correct answers to the quiz:

1. **B) A data visualization that displays the relationship between two numerical variables**
2. **B) Observations or individual data points**
3. **C) By their position relative to the overall pattern of the data**
4. **C) As one variable increases, the other variable tends to increase.**
5. **D) Remove all data points that do not fit the expected pattern.**

# Today

- Quiz 2
- Scatterplots & Uncertainty
  - "Scatterplots: Tasks, Data, and Designs"
  - "Winglets: Visualizing Association with Uncertainty in Multi-class Scatterplots"
- Readings for Today
  - "Why Authors Don't Visualize Uncertainty"
  - "Beyond Generating Code: Evaluating GPT on a Data Visualization Course"
- Readings for Friday

## Reading for Friday *pick one*

- "When Red Means Good, Bad, or Canada: Exploring People's Reasoning for Choosing Color Palettes",

Each State's Largest Foreign Trading Partner



“ This option has the most diverse color palette and the diverse colors are neutral in terms of symbolism; whereas green shades connote ecology and red, white and blue are US/politics related. ”

“ Canada is the trading partner that occurs on the map the most and I associate red with Canada because of the flag. The color choice makes the most sense to me. ”

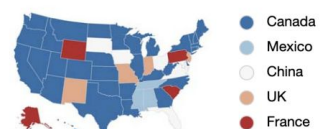
Each State's Largest Foreign Trading Partner



“ Green seems appropriate because it is the color of money, and the subject matter is trading with the majority being Canada. ”

“ Since Canada is the largest partner to many of the states it should have the darkest color. It is easy to see the color changes when it is shading and not completely different colors. ”

Each State's Largest Foreign Trading Partner



“ I think maybe it seems fair to represent the closest countries by geographical distance using the blue colors, and it just so happens these statistics support that with this color scheme, for the most part. ”

“ Since this is a map of the USA it is nice to have a color scheme in line with the countries official colors. ”

# Reading for Friday *pick one*

- "Exploring D3 Implementation Challenges on Stack Overflow", Battle, Feng, & Webber, IEEE Visualization 2022

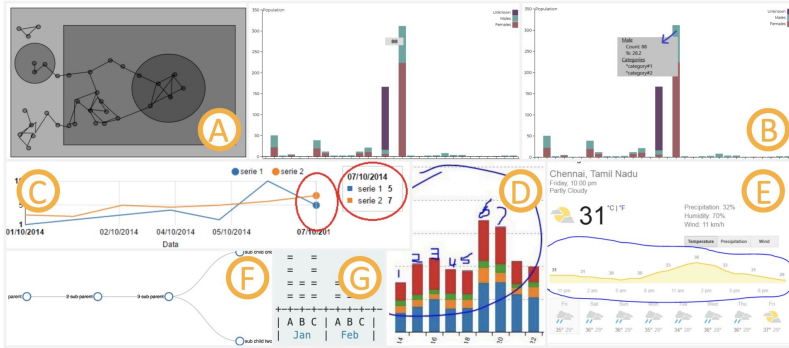


Figure 1: Example images shared online to convey desired visualizations (A, E, F, & G), interactions (B), and modifications (C & D).

# Reading for Friday *pick one*

- “Guidelines for Effective Usage of Text Highlighting Techniques”, Strobel, Oelke, Kwon, Schreck, Pfister, IEEE InfoVis 2015

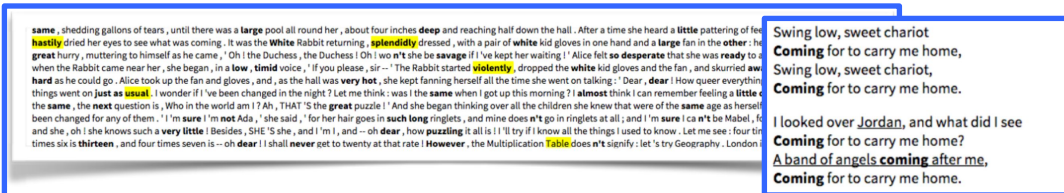


Fig. 1: Text highlighting techniques are commonly used to mark text features in documents. In this excerpt of “Alice in wonderland” all occurrences of adjectives and adverbs derived from part-of-speech tagging are highlighted in bold font, while words with typical adjective/adverb endings are highlighted with yellow background.

, and seemed to her to wink with one of its little eyes, but it said nothing. 'Perhaps it does not understand English,' thought Alice; 'I daresay it's a French mouse, come over with William the Conqueror.' (For, with all her knowledge of history, Alice had no very clear notion how long ago anything had happened.) So she began to gain: 'Où est ma chatte?' which was the first sentence in her French lesson-book. The Mouse gave a sudden leap out of the water, and seemed to quiver all over with fright. 'Oh, I beg your pardon!' cried Alice hastily, afraid that she had hurt the poor animal's feelings. 'I quite forgot you didn't like cats.' 'Not like cats!' cried the Mouse, in a shrill, passionate voice. 'Would YOU like cats if you were me?' 'Well, perhaps not,' said Alice in a soothing tone: 'don't be angry about it. And yet I wish I could show you our cat Dinah: think you'd take a fancy to cats if you could only see her. She is such a dear quiet thing.' Alice went on, half to herself, as she swam lazily about in the pool, and she sits purring so nicely by the fire, licking her paws and washing her face—and she is such a nice soft thing to nurse—and she is such a capital one for catching mice—oh, I beg your pardon!' cried Alice again, for this time the Mouse was bristling all over, and she felt certain it must be really offended. 'We won't talk about her any more if you'd rather not.' 'We indeed!' cried the Mouse, who was trembling down to the end of his tail. 'As if I would talk on such a subject! Our family always HATED cats: nasty, low, vulgar

Fig. 11: Example of combining techniques letter spacing and italics – according to our analysis this is not an effective combination for highlighting two equally important text features.

# Reading for Friday *pick one*

- “What Makes a Visualization Memorable?”,  
Borkin, Vo, Bylinskii, Isola, Sunkavalli, Oliva, & Pfister, INFOVIS 2013.

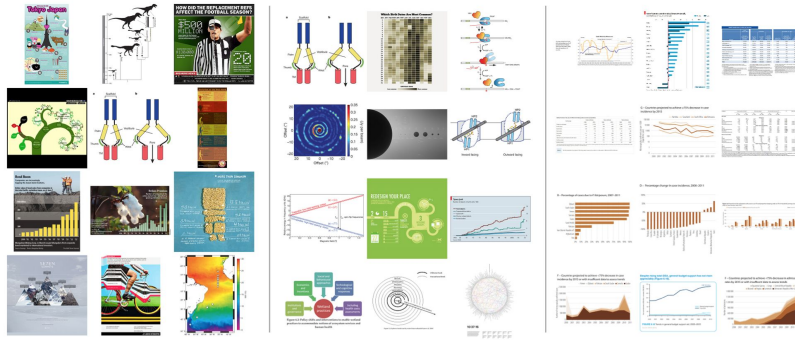


Fig. 1. **Left:** The top twelve overall most memorable visualizations from our experiment (most to least memorable from top left to bottom right). **Middle:** The top twelve most memorable visualizations from our experiment when visualizations containing human recognizable cartoons or images are removed (most to least memorable from top left to bottom right). **Right:** The twelve least memorable visualizations from our experiment (most to least memorable from top left to bottom right).