CSE 442 - Data Visualization

The Value of Visualization

Jeffrey Heer  University of Washington
How much data (bytes) did we produce in 2010?
2010: 1,200 exabytes and exponential growth...

Gantz et al., 2008, 2010
Data Created & Consumed

Source: IDC Digital Universe

~2x every 2 years

1.2 ZB

4.4 ZB

44.0 ZB

Data (Zettabytes)
But *what* is in all this data?

Data Created & Consumed

Source: IDC Digital Universe

~2x every 2 years

Data (Zettabytes)


1.2 ZB 4.4 ZB 44.0 ZB
Records of Human Activity
The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that’s going to be a hugely important skill in the next decades, ... because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it.

Hal Varian, Google’s Chief Economist

*The McKinsey Quarterly*, Jan 2009
The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that’s going to be a hugely important skill in the next decades, because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it. …to whose benefit?

Hal Varian, Google’s Chief Economist

The McKinsey Quarterly, Jan 2009
High potential for data abuse...
We move from data to information to knowledge to wisdom. And separating one from the other, being able to distinguish among and between them, that is, knowing the limitations and the danger of exercising one without the others, while respecting each category of intelligence, is generally what serious education is about.

Toni Morrison, American Novelist
*The Source of Self Regard*, 2019
How might we use visualization to empower understanding of data and analysis processes?
What is Visualization?

“Transformation of the symbolic into the geometric”
[McCormick et al. 1987]

“... finding the artificial memory that best supports our natural means of perception.” [Bertin 1967]

“The use of computer-generated, interactive, visual representations of data to amplify cognition.”
[Card, Mackinlay, & Shneiderman 1999]
<table>
<thead>
<tr>
<th>Set A</th>
<th>Set B</th>
<th>Set C</th>
<th>Set D</th>
</tr>
</thead>
<tbody>
<tr>
<td>(X)</td>
<td>(Y)</td>
<td>(X)</td>
<td>(Y)</td>
</tr>
<tr>
<td>10</td>
<td>8.04</td>
<td>10</td>
<td>9.14</td>
</tr>
<tr>
<td>8</td>
<td>6.95</td>
<td>8</td>
<td>8.14</td>
</tr>
<tr>
<td>13</td>
<td>7.58</td>
<td>13</td>
<td>8.74</td>
</tr>
<tr>
<td>9</td>
<td>8.81</td>
<td>9</td>
<td>8.77</td>
</tr>
<tr>
<td>11</td>
<td>8.33</td>
<td>11</td>
<td>9.26</td>
</tr>
<tr>
<td>14</td>
<td>9.96</td>
<td>14</td>
<td>8.1</td>
</tr>
<tr>
<td>6</td>
<td>7.24</td>
<td>6</td>
<td>6.13</td>
</tr>
<tr>
<td>4</td>
<td>4.26</td>
<td>4</td>
<td>3.1</td>
</tr>
<tr>
<td>12</td>
<td>10.84</td>
<td>12</td>
<td>9.11</td>
</tr>
<tr>
<td>7</td>
<td>4.82</td>
<td>7</td>
<td>7.26</td>
</tr>
<tr>
<td>5</td>
<td>5.68</td>
<td>5</td>
<td>4.74</td>
</tr>
</tbody>
</table>

**Summary Statistics**

- \(u_X = 9.0\)
- \(\sigma_X = 3.317\)
- \(u_Y = 7.5\)
- \(\sigma_Y = 2.03\)

**Linear Regression**

- \(Y = 3 + 0.5X\)
- \(R^2 = 0.67\)

[Anscombe 1973]
Set A

Set B

Set C

Set D

[Anscombe 1973]
Wikipedia History Flow [Viega & Wattenberg]
Why Create Visualizations?
Why Create Visualizations?

Answer questions (or discover them)
Make decisions
See data in context
Expand memory
Support graphical calculation
Find patterns
Present argument or tell a story
Inspire
Record Information
Gallop, Bay Horse “Daisy” [Muybridge]
E.J. Marey’s sphygmograph [from Braun 83]
You Draw It: How Family Income Predicts Children’s College Chances

You Draw It: How Family Income Predicts Children’s College Chances

Support Reasoning
**History of O-Ring Damage on SRM Field joints**

<table>
<thead>
<tr>
<th>No.</th>
<th>SRM No.</th>
<th>Erosion Depth (in.)</th>
<th>Perimeter Affected (deg.)</th>
<th>Nominal Dia. (in.)</th>
<th>Length Of Max Erosion (in.)</th>
<th>Total Heat Affected Length (in.)</th>
<th>Clocking Location (deg.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>22A</td>
<td>61A LH Center Field**</td>
<td>None</td>
<td>None</td>
<td>0.280</td>
<td>None</td>
<td>None</td>
<td>36° - 66°</td>
</tr>
<tr>
<td>22A</td>
<td>61A LH CENTER FIELD**</td>
<td>None</td>
<td>None</td>
<td>0.280</td>
<td>None</td>
<td>None</td>
<td>338° - 18°</td>
</tr>
<tr>
<td>15A</td>
<td>51C LH Forward Field**</td>
<td>0.010</td>
<td>154.0</td>
<td>0.280</td>
<td>4.25</td>
<td>5.25</td>
<td>163</td>
</tr>
<tr>
<td>15B</td>
<td>51C RH Center Field (prim)**</td>
<td>0.038</td>
<td>130.0</td>
<td>0.280</td>
<td>12.50</td>
<td>58.75</td>
<td>354</td>
</tr>
<tr>
<td>15B</td>
<td>51C RH Center Field (sec)**</td>
<td>None</td>
<td>45.0</td>
<td>0.280</td>
<td>None</td>
<td>29.50</td>
<td>354</td>
</tr>
<tr>
<td>13B</td>
<td>41D RH Forward Field</td>
<td>0.028</td>
<td>110.0</td>
<td>0.280</td>
<td>3.00</td>
<td>None</td>
<td>275</td>
</tr>
<tr>
<td>11A</td>
<td>41C LH Aft Field*</td>
<td>None</td>
<td>None</td>
<td>0.280</td>
<td>None</td>
<td>None</td>
<td>--</td>
</tr>
<tr>
<td>10A</td>
<td>41B LH Forward Field</td>
<td>0.040</td>
<td>217.0</td>
<td>0.280</td>
<td>3.00</td>
<td>14.50</td>
<td>351</td>
</tr>
<tr>
<td>2B</td>
<td>STS-2 RH Aft Field</td>
<td>0.053</td>
<td>116.0</td>
<td>0.280</td>
<td>--</td>
<td>--</td>
<td>90</td>
</tr>
</tbody>
</table>

---

*Hot gas path detected in putty. Indication of heat on O-ring, but no damage.
**Soot behind primary O-ring.
***Soot behind primary O-ring, heat affected secondary O-ring.

Clocking location of leak check port - 0 deg.

Other SRM-15 field joints had no blowholes in putty and no soot near or beyond the primary O-ring.

SRM-22 forward field joint had putty path to primary O-ring, but no O-ring erosion and no soot blowby. Other SRM-22 field joints had no blowholes in putty.

---

**Blow By History**

- **SRM-15 Worst Blow-By**
  - 2 case joints (90°, 110°) arc
  - Much worse visually than SRM-22

- **SRM-22 Blow-By**
  - 2 case joints (30°-40°)
  - Nozzle blow-by

---

**History of O-Ring Temperatures (Degrees F)**

<table>
<thead>
<tr>
<th>Motor</th>
<th>MHT</th>
<th>AMB</th>
<th>O-Ring</th>
<th>Wind</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM-1</td>
<td>68</td>
<td>36</td>
<td>47</td>
<td>10 MPH</td>
</tr>
<tr>
<td>DM-2</td>
<td>76</td>
<td>45</td>
<td>52</td>
<td>10 MPH</td>
</tr>
<tr>
<td>GM-3</td>
<td>72.5</td>
<td>40</td>
<td>48</td>
<td>10 MPH</td>
</tr>
<tr>
<td>GM-4</td>
<td>76</td>
<td>48</td>
<td>51</td>
<td>10 MPH</td>
</tr>
<tr>
<td>SRM-15</td>
<td>52</td>
<td>64</td>
<td>53</td>
<td>10 MPH</td>
</tr>
<tr>
<td>SRM-22</td>
<td>77</td>
<td>78</td>
<td>75</td>
<td>10 MPH</td>
</tr>
<tr>
<td>SRM-25</td>
<td>55</td>
<td>26</td>
<td>29</td>
<td>10 MPH</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>27</td>
<td>25 MPH</td>
</tr>
</tbody>
</table>
Make Decisions: Challenger

But wait! What is an appropriate “damage index”? Which temperatures, O-ring or outside air?

Chart of temperatures vs. O-ring damage [Tufte 97]
In 1854 John Snow plotted the position of each cholera case on a map. [from Tufte 83]
Data in Context: Cholera Outbreak

Used map to hypothesize that pump on Broad St. was the cause. [from Tufte 83]
**Answer Questions: Brain Power?**

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Body Weight</th>
<th>Brain Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Lesser Short-tailed Shrew</td>
<td>5</td>
<td>0.14</td>
</tr>
<tr>
<td>2</td>
<td>Little Brown Bat</td>
<td>10</td>
<td>0.25</td>
</tr>
<tr>
<td>3</td>
<td>Mouse</td>
<td>23</td>
<td>0.3</td>
</tr>
<tr>
<td>4</td>
<td>Big Brown Bat</td>
<td>23</td>
<td>0.4</td>
</tr>
<tr>
<td>5</td>
<td>Musk Shrew</td>
<td>48</td>
<td>0.33</td>
</tr>
<tr>
<td>6</td>
<td>Star Nosed Mole</td>
<td>60</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>Eastern American Mole</td>
<td>75</td>
<td>1.2</td>
</tr>
<tr>
<td>8</td>
<td>Ground Squirrel</td>
<td>101</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>Tree Shrew</td>
<td>104</td>
<td>2.5</td>
</tr>
<tr>
<td>10</td>
<td>Golden Hamster</td>
<td>120</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>Mole Rate</td>
<td>122</td>
<td>3</td>
</tr>
<tr>
<td>12</td>
<td>Galago</td>
<td>200</td>
<td>5</td>
</tr>
<tr>
<td>13</td>
<td>Rat</td>
<td>280</td>
<td>1.9</td>
</tr>
<tr>
<td>14</td>
<td>Chinchilla</td>
<td>425</td>
<td>6.4</td>
</tr>
<tr>
<td>15</td>
<td>Desert Hedgehog</td>
<td>550</td>
<td>2.4</td>
</tr>
<tr>
<td>16</td>
<td>Rock Hyrax (a)</td>
<td>750</td>
<td>12.3</td>
</tr>
<tr>
<td>17</td>
<td>European Hedgehog</td>
<td>785</td>
<td>3.5</td>
</tr>
<tr>
<td>18</td>
<td>Tenrec</td>
<td>900</td>
<td>2.6</td>
</tr>
<tr>
<td>19</td>
<td>Arctic Ground Squirrel</td>
<td>920</td>
<td>5.7</td>
</tr>
<tr>
<td>20</td>
<td>African Giant Pouched Rat</td>
<td>1000</td>
<td>6.6</td>
</tr>
<tr>
<td>21</td>
<td>Guinea Pig</td>
<td>1040</td>
<td>5.5</td>
</tr>
<tr>
<td>22</td>
<td>Mountain Beaver</td>
<td>1350</td>
<td>8.1</td>
</tr>
<tr>
<td>23</td>
<td>Slow Loris</td>
<td>1400</td>
<td>12.5</td>
</tr>
<tr>
<td>24</td>
<td>Genet</td>
<td>1410</td>
<td>17.5</td>
</tr>
<tr>
<td>25</td>
<td>Phalanger</td>
<td>1620</td>
<td>11.4</td>
</tr>
</tbody>
</table>
The Elements of Graphing Data
[Cleveland]
Convey Information
“to affect thro’ the Eyes what we fail to convey to the public through their word-proof ears”

1856 “Coxcomb” of Crimean War Deaths, Florence Nightingale
Communicate, Inform, Inspire

Visualizing Black America, Du Bois et al. 1900

Bones in hand, Gray’s Anatomy 1918 ed.
New deaths attributed to Covid-19 in European Union, United States, Brazil and United Kingdom

Seven-day rolling average of new deaths, by number of days since 3 average daily deaths first recorded

Source: Financial Times analysis of data from the European Centre for Disease Prevention and Control, the Covid Tracking Project, the UK Dept of Health & Social Care and the Spanish Ministry of Health.
Data updated September 25 2020 12.46pm BST. Interactive version: ft.com/covid19
The coronavirus crisis is different

Job growth (or loss) since each recession began, based on weekly earnings

**1990 recession**

**2001 recession**

**2008 recession**

**Coronavirus crisis**

Notes: Based on a three-month average to show the trend in volatile data.
Source: Labor Department via IPUMS, with methodology assistance from Ernie Tedeschi of Evercore ISI

THE WASHINGTON POST
The Value of Visualization

**Record** information
- Blueprints, photographs, seismographs, …

**Analyze** data to support reasoning
- Develop and assess hypotheses
- Find patterns / Discover errors in data
- Expand memory

**Convey** information
- Communicate, inform, inspire
- Collaborate and revise
Goals of Visualization Research

1 **Understand** how visualizations convey information
   - What do people perceive / comprehend?
   - How do visualizations inform mental models?

2 **Develop principles and techniques** for creating effective visualizations and supporting analysis
   - Leverage perception & augment cognition
   - Improve ties between visualization & mental model
Course Topics
Data and Image Models

Sémiologie Graphique [Bertin 67]
Visualization Design

SlicerDicers' Sales Compared to Other Products

Problematic design

Sales of SlicerDicers Compared to Sales of Other Products
July - December, 2011

Redesign
Visualization Software

**D3**: Data-Driven Documents

**Vega-Lite / Altair**
Animation

Animated transitions in statistical data graphics [Heer & Robertson 07]
CIE 1931 xy chromaticity diagram showing the gamut of the sRGB and Adobe RGB color spaces including the Planckian locus, with temperatures indicated. Wavelengths of monochromatic light are shown in blue.
Graphical Perception

The psychophysics of sensory function [Stevens 61]
Hierarchies

Degree-Of-Interest Trees [Heer & Card 04]
Uncertainty
Interactive querying of 1.7B stars (1.2TB) in Falcon [Moritz et al. 2019]
Course Mechanics
You should expect to:

1. Evaluate and critique visualization designs
2. Learn visualization techniques & theory
3. Implement interactive data visualizations
4. Develop a substantial visualization project
Lectures, Activities & Office Hours

Course sessions will alternate among lecture and in-class activities. Thursdays will typically be activity days, but not always.

All lectures will be in-person and recorded. Please attend in-person but **NOT** if you feel ill.

Office hours are a mix of in-person and Zoom. Links for virtual office hours are on Canvas.

Use Ed to post questions and seek help!
Readings

There is no one universal textbook on visualization! So we will draw on books, notebooks, and linked articles...

Material in class will loosely follow readings. Readings should be read by start of class.
Textbook

Interactive Data Visualization for the Web, 2nd Edition

For learning D3!
Book available online.
Code / examples on GitHub.

We will be using D3 v7.
https://d3js.org
Interactive Vega-Lite Notebooks

Hands-on engagement with course concepts and tools using Observable (JavaScript) notebooks.
Assignments

**CP**  Class Participation (10%)

**A1**  Expository Visualization (10%) - Due 1/12

**A2**  Deceptive Visualization (15%) - Due 1/24
  Peer Review (5%) - Due 1/30

**A3**  Interactive Prototype (20%) - Due 2/12
  Peer Review (5%) - Due 2/21

**FP**  Final Project (35%)

  Proposal - Due 2/16

  Demonstration Video - Due 3/6

  Final Prototype - Due 3/12
Grading Philosophy

A great submission gets a great grade (A- to A, 3.6 – 3.8), but an exceptional grade (A+, 3.9 – 4.0) requires exceptional effort.

Example: Typical A1 grades (out of 10 points). Everyone starts with a high score (9/10). We then deduct points for errors and also add points for going above and beyond the assignment requirements. The median score for A1 is typically 8.5 out of 10, which maps to an A-.
Course Participation

Lecture attendance and engagement
In-class exercises: team submissions
Online quizzes, submitted on Canvas

*Note*: You may miss up to 2 lectures and up to 2 in-class exercise days without penalty.
Online Practice Quizzes

Weeks 2-8 have online quizzes to emphasize important concepts.

Quizzes are due each Friday by 11:59pm, starting next week. They can be retaken as many times as needed to get full points.

Quiz completion counts towards course participation. Please note that quiz points will not be added directly to your course grade.
Final Project

Produce an explorable visual explanation

Initial prototype and design review

Final deliverables and video presentation

Submit and publish online (GitLab)

Projects from previous classes have been:

• Published as research papers
• Shared widely (some in the New York Times!)
• Released as successful open source projects
Why outbreaks like coronavirus spread exponentially, and how to “flatten the curve”

Harry Stevens, Washington Post 2020
Locations of each train on the red, blue, and orange lines at 5:13 am. Hover over the diagram to the right to display trains at a different time.

Trains are on the right side of the track relative to the direction they are moving.

See the morning rush-hour, midday lull, afternoon rush-hour, and the evening lull.

Service starts at 5AM on Monday morning. Each line represents the path of one train. Time continues downward, so steeper lines indicate slower trains.

Since the red line splits, we show the Ashmont branch first then the Braintree branch. Trains on the Braintree branch "jump over" the Ashmont branch.

Train frequency increases around 6:30AM as morning rush hour begins.
KEYBOARD WALKING

Passwords with a “keyboard walking” pattern start at an arbitrary key, then move in a direction (usually right or down) while continuing to hit keys. Sometimes this is combined with holding down the_shift key, so that some characters are uppercase or symbols to improve complexity.

While the generated password may seem to be random and unhackable, password crackers check for these keyboard patterns and guess them early on.

Many passwords in the leaked passwords dataset have a spatial pattern. Other than the numeric passwords like 123456, common keyboard walking offenders include qwerty and 1qaz@wsx.

Semantic Passwords
Vishal Devireddy  (CSE 512, Spring '21)
A1: Expository Visualization

Design a static visualization for a data set.

The climate of a place can have a tremendous impact on people's lived experience. You will examine average monthly climate measurements for six major U.S. cities, roughly covering the edges of the continental United States.

You must choose the message you want to convey. What question(s) do you want to answer? What insight do you want to communicate?
A1: Expository Visualization

Pick a **guiding question**, use it to title your vis. Design a **static visualization** for that question. You are free to **use any tools** (inc. pen & paper).

**Deliverables** (via Gradescope; see A1 page)
- Image of your visualization (PNG or JPG format)
- Short description + design rationale ($\leq$ 4 paragraphs)

Due by **11:59 pm, Fri Jan 12**.
Observable + Data Tutorial

Friday Jan. 5, 4:00-5:30pm on Zoom

Introduction to Observable notebooks, JavaScript basics, and data management and transformation.

Led by Shengqi & Zening.

Zoom link is available on Canvas.
The tutorial will be recorded.
Instructors

Instructor

Jeffrey Heer
OH: Tue 11:30-12:30pm

Teaching Assistants

Lisa Elkin
OH: Mon 1-2pm

Shengqi Hang
OH: Wed 5-6pm

Andrew Zhang
OH: Thu 1-2pm

Zening Qu
OH: Fri 3:30-4:30pm

Raymond Fok
OH: Online / By Appt.

Ron Pechuk
OH: Online / By Appt.

Yilun Sheng
OH: Online / By Appt.

cse442@cs
Lisa Elkin (She/Her)

Academic Background
- BMath, C&O and Pure Math, University of Waterloo, 2012
- MMath, Computer Science, University of Waterloo, 2018
- PhD, UW CSE, 2018 - ???

TA Experience
- HCI, Data Viz, Linear Algebra, Calculus, Intro CS, CS for non-majors

Industry Internships
- MSR 2018, Apple 2021, Meta 2022, Meta 2023
Winston Elkin

Academic Background

Doggy School Level 1*
PhD, UW CSE, 2019 - ???

Dissertation: Evaluating the Impact of Tree Species on a Dog’s Desire to Pee on it: an Autoethnography

Industry Internships

Meta 2022, Meta 2023. Even has the badge to prove it.

* Indicates candidate did not complete all requirements but passed due to being very cute.
Raymond Fok (he/him)
rayfok@cs.washington.edu

5th year PhD student, advised by Dan Weld

Research interests
• HCI + NLP
• Systems for scholarly communication

Hobbies
• Board games
• Traveling
• Growing vegetables in my backyard
Shengqi Hang  (He/him)

sqhang@cs.washington.edu

- First Year CSE Ph.D. student
- From: Jiangsu, China
- Research Interests
  - Computational biology
  - AI for medicine
- Hobbies
  - Movies
  - Soccer
  - Pingpong
Ron Pechuk ft. his cats
rpechuk@cs.washington.edu

- Senior in Computer Science
- Prior TA Experience: CSE 442, EE 393
- Interests: Front-end Development, Data Science
- Specialties: Web publishing, D3
- Hobbies: Basketball, Hiking, Board Games, Bowling
Hi, I am Zening Qu
I work on tools for better and easier visualization creations

1. Aunt Lily Can Say Her Visualizations

- Between views 1, 2, 3 and 4, the same field GDP Per Capita has different axis start and end values: [0,24000], [0,13000].

Change all GDP Per Capita axes start and end values to [0, 24000]

2. Dashboard Integrity Guard
Yilun Sheng (or Simon)

Second-year CSE Ph.D. student

Research Interests: Computational Biology, ML

Hobbies:
- Soccer (watching > playing), Premier League, Chelsea!
- Pokémon
- Bridge (the card game)

Contact: ylsheng@cs.washington.edu
Andrew Zhang (he/him)
azhang26@cs.washington.edu

- 4th year CSE BS/MS
- Previously TAed: CSE 14x, 12x, 373, and 312
- Interests: AI/ML, HCI, Data Visualization and Accessibility
- Hobbies: Soccer, tennis, boxing, hiking, snowboarding
Questions?