How much data (bytes) did we produce in 2020?
2020: 64.2 zetabytes

~2x increase every 2 years

IDC 2021
But *what* is in all this data?

Physical Sensors
Scott Snibbe, Cabspotting
Abortion

From Wikipedia, the free encyclopedia

For other uses, see Abortion (disambiguation).

Abortion is the ending of a pregnancy by removal or expulsion of an embryo or fetus before it can survive outside the uterus.[1] An abortion that occurs without intervention is known as a miscarriage or spontaneous abortion. When deliberate steps are taken to end a pregnancy, it is called an induced abortion, or less frequently "induced miscarriage". The unmodified word abortion generally refers to an induced abortion.[2] A similar procedure after the fetus has potential to survive outside the womb is known as a "late termination of pregnancy" or less accurately as a "late term abortion".[3]

When properly done, abortion is one of the safest procedures in medicine,[4] but unsafe abortion is a major cause of maternal death, especially in the developing world.[5] Making safe abortion legal and accessible reduces maternal deaths.[6] It is safer than childbirth, which has a 14 times higher risk of death in the United States.[7] Modern methods use medication or surgery for abortions.[8] The drug mifepristone in combination with prostaglandins appears to be as safe and effective as surgery during the first and second trimester of pregnancy.[9][10][11] The most common surgical technique involves dilating the cervix and using a suction device.[12] Birth control, such as the pill or intrauterine devices, can be used immediately following abortion.[11] When performed legally and safely on a woman who desires it, induced abortions do not increase the risk of long-term mental or physical problems.[11] In contrast, unsafe abortions (those performed by unskilled individuals, with hazardous equipment, or in unsanitary facilities) cause 47,000 deaths and 5 million hospital admissions each year.[11][13] The World Health Organization recommends safe and legal abortions be available to all women.[11]

Around 56 million abortions are performed each year in the world,[14] with about 45% done unsafely.[15] Abortion rates changed little between 2003 and 2008,[16] before which they decreased for at least two decades as access to family planning and birth control increased.[18] As of 2008, 40% of the world's 2008 pregnancies were unintended.[19] Countries that permit abortions have different limits on how late in pregnancy abortion is permissible, without limits as to reason.[20] Counties that permit abortions have different limits on how late in pregnancy abortion is permissible, without limits as to reason.[20] Abortion laws and...
“What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention, and a need to allocate that attention efficiently among the overabundance of information sources that might consume it.”

Herb Simon
as quoted by Hal Varian
Scientific American
September 1995

“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
January 2009
WHAT IS VISUALIZATION?

Ages of first-time mothers in 1980

New York Times 2018
Ages of first-time mothers in **2016**

New York Times 2018

Gray's Anatomy 1858
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]
### Summary Statistics

<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>

### Linear Regression

- **Set A**:
  - $u_X = 9.0$
  - $\sigma_X = 3.317$
  - $u_Y = 7.5$
  - $\sigma_Y = 2.03$
  - $Y = 3 + 0.5X$
  - $R^2 = 0.67$

Anscombe 73
WHY DO WE CREATE VISUALIZATIONS?

Why do we create visualizations?

• Answer questions (or discover them)
• Make decisions
• See data in context
• Expand memory
• Support graphical calculation
• Find patterns
• Present argument
• Tell a story
• Inspire
THE PURPOSE OF VISUALIZATION

**Record information**
Photographs, blueprints, ...

**Support reasoning about information (analyze)**
Process and calculate
Reason about data
Expand memory

**Communicate, inform, inspire (present)**
Share and persuade
Emphasize important aspects of data
Answer question

Gallop, Bay Horse “Daisy” [Muybridge 1884-86]

Answer question

Muybridge's setup [1884-86]
Photograph:
Phases of the moon

Yaorusheng / Getty Images

Drawing:
Phases of the moon

Galileo's drawings of the phases of the moon from 1616
http://galileo.rice.edu/sci/observations/moon.html
SUPPORT REASONING

Associated Press
2 of 13 pages of material faxed to NASA by Morton Thiokol [from Tufte 1997]
Make a decision: Challenger

Visualizations drawn by Tufte show how low temperatures damage O-rings [Tufte 97]
Visualizations drawn by Tufte show how low temperatures damage O-rings [Tufte 97]

Data in context

In 1854 John Snow plotted the position of each cholera case on a map. [from Tufte 83]
Data in context

Used map to support hypothesis
Broad St. pump was the cause.
[from Tufte 83]

Expand memory: Multiplication

Class exercise
Expand memory: Multiplication

\[
74 \times 48 = 3552
\]
Most powerful brain?

The Dragons of Eden [Carl Sagan]
Most powerful brain?

The Elements of Graphing Data [Cleveland]
Tell a story

Beautiful Evidence [Tufte]

Present an argument

“to affect thro’ the eyes what we fail to convey to the public through their word-proof ears”
Inform

Free-for-all

Attempted quarantine

Moderate distancing

Extensive distancing

https://www.washingtonpost.com/graphics/2020/world/corona-simulator/
GOALS OF VISUALIZATION RESEARCH

1. **Understand** how visualizations convey information
   - What do people perceive/comprehend?
   - How do visualizations correspond with mental models of data?

2. **Develop principles and techniques** for creating effective visualizations and supporting analysis
   - Leverage perception and cognition
   - Strengthen connection between visualization and mental models
**Mon: Exploratory Data Analysis**

**Wed: Using Space Effectively**

**Week 3**

**Mon: Interaction**

**Wed: Introduction to D3**

**Week 4**
Week 5

Mon: D3 Tutorial

Wed: Perception

Week 6

Mon: Visual Explainers

Wed: Color
Week 7

Mon: Animation

Wed: Network Layout

Week 8

Mon: Network Analysis

Wed: Deconstructing Visualizations
Final Project Review and Feedback

Week 9
**LEARNING GOALS**

- An understanding of key visualization techniques and theory, including data models, graphical perception and methods for visual encoding and interaction.

- Exposure to several common data domains and corresponding analysis tasks, including exploratory data analysis and network analysis.

- Practical experience building and evaluating visualization systems using Vega-Lite and D3.js.

- The ability to read and discuss research papers from the visualization literature.
YOU SHOULD EXPECT TO

• *Design, evaluate and critique* visualizations

• *Explore* data using existing visualization tools

• *Implement* interactive data visualizations

• *Develop* a substantial visualization project

COURSE MECHANICS

https://magrawala.github.io/cs448b-fa23/
Instructor: Maneesh Agrawala

Course Assistant: Jasmine Shih

Office Hours: 3:30-4:30pm Fridays
Location: Huang Basement
**Course Assistant:** Yifan Shen

**Office Hours:** 10-11am Wednesdays

**Location:** Huang Basement

---

**Course Assistant:** Anthony Xie

**Office Hours:** 4:30-5:30pm Mondays

**Location:** Bytes Cafe
Office Hours

Maneesh: 11-noon Thu, Coupa Café Y2E2
Jasmine: 3:30-4:30pm Fri, Huang Basement
Yifan: 10-11:00am Wed, Huang Basement
Anthony: 4:30-5:30pm Mon, Bytes Café

Happy to schedule other OH by appointment
Outside of OH use Slack to connect with us

https://canvas.stanford.edu/courses/180122/external_tools/11232

Textbooks

See also: www.edwardtufte.com
Interactive Notebooks

Interactivity
In addition to basic plotting and view composition, one of Vega-Lite's more exciting features is its support for interaction.

Starting with a scatter plot, we can add a basic (yet valuable!) form of interactivity – tooltips upon mouse hover – by including a tooltip encoding channel:

```javascript
vl.markPoint().encode(vl.x('Year'), vl.y('Miles_per_Dollar'), vl.color('Origin'), vl.tooltip(['Name', 'Origin'])); // show the Name and Origin fields in a tooltip
```

Hands-on engagement with course concepts and modern visualization tools (Vega-Lite / D3), in JavaScript (Observable)

Extra: Mon 10/2 4:30-5:30pm
Anthony will provide an intro to JavaScript/Observable on Zoom.
More details on course website.

Optional Textbook
For additional theory and depth
Readings

From books, notebooks and linked articles
Many open to public, some may require SUNetID/Password

Material in class will be loosely based on readings
Readings should be read by start of class

Post comment (on reading, notebooks or lecture) in Canvas Discussion
One comment per week through week 9
Must post by end of the week (Sun at 8pm)
You have 1 pass for the quarter

Class home page
https://magrawala.github.io/cs448b-fa23/

Reading Responses

Good responses typically exhibit one or more
Critiques of arguments made in the papers/lectures
Analysis of implications or future directions for ideas in readings/lectures
Insightful questions about the readings/lectures

Responses should not be summaries
Should be substantive (1-2 paragraphs is typical)
In-Person Discussion

Discussion and critique are essential for effective design and evaluation of visualizations

• In-person discussion is more effective than online and benefits all attendees
• Attendance is very strongly recommended
• Will be considered in grading for non-SCPD students

Assignments

Class participation (10%)

Assignment 1: Visualization Design (10%) due 10/2

Assignment 2: Exploratory Data Analysis (15%) due 10/16
Use Tableau or Vega-Lite

Assignment 3: Interactive Prototype (25%) due 10/30
Should be familiar with Javascript (start now if you are not)
Will cover basics of Vega-Lite and D3 in class

Final Project (40%) proposal due 11/6, design review 11/27, 11/29, final submission 12/10
Final project

Produce an interactive visual explainer
Initial prototype and design review
Final deliverables and video

Projects from previous classes have been
Published as research papers
Shared widely (e.g., gone viral on blogs)
Released as successful open source projects

Structure of Musicals

Lyrical themes in Hamilton

Lyrical themes in Hamilton [Townley-Smith, Sterman, Cook 2016]
Visualization of Narrative Structure

Character interactions and sentiment in The Hobbit [Bilenko, Miyakawa 2013]

ASSIGNMENT 1: VISUALIZATION DESIGN
Due 10/2 11:30 AM
Design a static visualization for a data set
You must choose the message you want to convey. What question(s) do you want to answer? What insight do you want to communicate?

Data: Stanford Undergraduate Majors
Stanford University publishes a variety of datasets through the Stanford Institutional Research & Decision Support website. They have published a data table containing information about the number of Stanford undergraduates obtaining a Bachelor's degree in 75 different fields of study from 2003 to 2022. We have filtered and wrangled this data to the top 10 fields of study by cumulative degrees conferred over the time period to produce a dataset with the following attributes:

- **Year**: Academic year between 2003 and 2022. (Academic years run July-June so Year=2003 covers July 2002 to June 2003.)
- **FieldOfStudy**: Field in which degree was obtained.
- **Count**: Number of students earning a Bachelor's degree.

The extracted dataset is available in csv format: TopFieldsStanfordBachelors.csv
Assignment 1: Visualization Design

Due: Monday Oct 3, 2022 by 11:59am

In this assignment, you will design a visualization for a small data set and provide rigorous rationale for your design choices. You should choose the best means to explore the correlation of each pair of the display. The see how to use any graphics or charting but please excluding drafting it by hand.

Data: Stanford Undergraduate Majors

Stanford University publishes a variety of datasets through the Stanford Institutional Research & Doctoral Support websites. They have published a data table containing information about the number of Stanford undergraduates obtaining a bachelor’s degree in 15 different fields of study from 2000 to 2020. We have selected and managed the data for the top 10 fields of study by graduation degrees centered over the years 2005 and 2010.

- Academic year: July 2004 to June 2005.
- Graduating year: July 2005 to June 2010.
- Undergraduates: Undergraduates with a bachelor’s degree.
- Graduating: Number of students earning a bachelor’s degree.
- The extended dataset is available on the web.

Assignment

Your task is to download this data and design a report (i.e., simple visualizations that you believe effectively communicate one aspect of the data) and provide a short write-up (no more than 4 paragraphs) describing your design choices. Start by choosing a question you’d like your visualization to answer. Design your visualizations to answer that question, and use the question as the title of your graphs.

While you may use the data set given, you are free to filter, transform and augment the data as you see fit. Such transformations may include (but are not limited to) log transformation, smoothing percentages or averages, grouping elements into new categories, and removing data that are not relevant to your driving question. Your data should include external data as you see fit. Your chart should be interpretable without recourse to your short write-up. Do not forget to include titles, and labels of legends as needed.

According to the question your visualization is designed to answer.

As different visualizations can emphasize different aspects of a data set, your write-up should ensure some aspects of the data that you are emphasizing in most effectively communicate. In short, what does all you trying to tell? Just as important, also note which aspects of the data might be obscured due to your visualization design.

In your write-up, you should provide a rigorous rationale for your design decisions and explain whenever possible your design decisions. Document the visual encoding you used and why they are appropriate for the data. These decisions include the choice of visualization types, axes, titles, labels, and legend. Show how these choices were made.

Please include a short list of the tools you used to create the visualization.

Grading