How much data (bytes) did we produce in 2010?

2010: 1,200 exabytes
10x increase over 5 years

Gantz et al, 2008, 2010
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*

---

**A Brief History**

What was the **first** data visualization?

---

- Acquisition
- Cleaning
- Integration
- Visualization
- Analysis
- Presentation
- Dissemination

0 BC
The Rate of Water Evaporation, Lambert 1765

The Golden Age of Data Visualization

The Commercial and Political Atlas, William Playfair 1786
Statistical Breviary, William Playfair 1801

1822 Price of Wheat and Wages of Labor, William Playfair

1786 1826 Illiteracy in France, Pierre Charles Dupin

1856 “Coxcomb” of Crimean War Deaths, Florence Nightingale

"to affect thro' the Eyes what we fail to convey to the public through their word-proof ears"
The Rise of Statistics

Rise of **formal methods** in statistics and social science — Fisher, Pearson, ...

**Little innovation** in graphical methods

A period of **application and popularization**
Graphical methods enter textbooks, curricula, and **mainstream use**

The last few decades have seen the rise of formal theories of statistics, “legitimizing” variation by confining it by assumption to random sampling, often assumed to involve tightly specified distributions, and restoring the appearance of security by emphasizing narrowly optimized techniques and claiming to make statements with “known” probabilities of error.
While some of the influences of statistical theory on data analysis have been helpful, others have not.

Exposure, the effective laying open of the data to display the unanticipated, is to us a major portion of data analysis. Formal statistics has given almost no guidance to exposure; indeed, it is not clear how the informality and flexibility appropriate to the exploratory character of exposure can be fitted into any of the structures of formal statistics so far proposed.

<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

\[
\begin{align*}
\bar{X}_A &= 9.0, \quad \bar{X}_B = 3.97 \\
\bar{X}_C &= 7.5, \quad \bar{X}_D = 2.03
\end{align*}
\]

Linear Regression

\[
\begin{align*}
\hat{Y}_A &= 3 + 0.5X \\
R^2_A &= 0.67
\end{align*}
\]

Anscombe 1973
As a discipline, data analysis is a very difficult field. Data analysis must adapt itself to what people can and need to do with data. In the sense that biology is more complex than physics, and the behavioral sciences are more complex than either, it is likely that the general problems of data analysis are more complex than those of all three.
Course Goals

To explore how a broad class of data analysts can more effectively analyze data using novel interactive tools.

The class will be interdisciplinary in nature, with a goal of identifying and pursuing new research opportunities.

Course Topics

W1: Analytical Thinking
W2: Data Collection
W3: Data Cleaning & Assessment
W4: Data Integration
W5: Visual Analysis
Course Topics
W6: Managing Big Data
W7: Analysis Practices
W8: Social Network Analysis
W9: Text Analysis
W10: Dissemination / Collaboration

Course Expectations
You should expect to gain:
An overview of the research landscape in interactive data analysis (HCI, Vis, DB).
Engaged discussions with leading researchers and practitioners in the area of “data science.”
Hands-on experience working with scalable data analysis frameworks.
Insights from a quarter-long research project prototyping a new interactive analysis tool.

Course Staff - cs448g@cs
Instructor
Jeffrey Heer
Asst Professor, Computer Science
OH Tu 1030a-12p, Gates 375

Course Assistant
Diana MacLean
PhD Student, Computer Science
OH W 1-3p, Gates 372
... and you!

DataWrangler

<table>
<thead>
<tr>
<th>Transform Script</th>
<th>Impact Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Split data repeatedly on newline into rows.</td>
<td></td>
</tr>
<tr>
<td>Split split repeatedly on , into columns.</td>
<td></td>
</tr>
<tr>
<td>Promote row 0 to header.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year</th>
<th>Reported crime in Alabama</th>
<th>Property crime rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1980</td>
<td>14820</td>
<td>36.3</td>
</tr>
<tr>
<td>1985</td>
<td>15500</td>
<td>38.4</td>
</tr>
<tr>
<td>1990</td>
<td>16200</td>
<td>40.6</td>
</tr>
<tr>
<td>1995</td>
<td>16900</td>
<td>42.8</td>
</tr>
<tr>
<td>2000</td>
<td>17600</td>
<td>45.1</td>
</tr>
<tr>
<td>2005</td>
<td>18300</td>
<td>47.3</td>
</tr>
</tbody>
</table>

#reported crime in Alaska

<table>
<thead>
<tr>
<th>Year</th>
<th>Reported crime in Alaska</th>
<th>Property crime rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1980</td>
<td>1578.9</td>
<td>36.5</td>
</tr>
<tr>
<td>1985</td>
<td>1652.9</td>
<td>38.7</td>
</tr>
<tr>
<td>1990</td>
<td>1727.5</td>
<td>40.9</td>
</tr>
</tbody>
</table>

MedHelp > Lyme

<table>
<thead>
<tr>
<th>Year</th>
<th>Reported crime in Alabama</th>
<th>Property crime rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1980</td>
<td>14820</td>
<td>36.3</td>
</tr>
<tr>
<td>1985</td>
<td>15500</td>
<td>38.4</td>
</tr>
<tr>
<td>1990</td>
<td>16200</td>
<td>40.6</td>
</tr>
<tr>
<td>1995</td>
<td>16900</td>
<td>42.8</td>
</tr>
<tr>
<td>2000</td>
<td>17600</td>
<td>45.1</td>
</tr>
<tr>
<td>2005</td>
<td>18300</td>
<td>47.3</td>
</tr>
</tbody>
</table>

#reported crime in Alaska

<table>
<thead>
<tr>
<th>Year</th>
<th>Reported crime in Alaska</th>
<th>Property crime rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1980</td>
<td>1578.9</td>
<td>36.5</td>
</tr>
<tr>
<td>1985</td>
<td>1652.9</td>
<td>38.7</td>
</tr>
<tr>
<td>1990</td>
<td>1727.5</td>
<td>40.9</td>
</tr>
</tbody>
</table>

TopicFlow

Diana MacLean

Jason Chuang
Course Participation

Lecture attendance is mandatory.

We expect you to think critically before lecture
→ Come prepared for discussion!

Each week has a set of required readings
→ Start the readings over the weekend
**Reading Responses**
You are responsible for a **thoughtful written response** (2-3 para) for each week’s readings.
Choose **one** of the readings:
- What did you find most insightful?
- Are there notable flaws or shortcomings?
- What questions remain unaddressed?
- Did the reading spark any new research ideas?
Responses are due by **8am before lecture**, on days without a guest speaker, to cs448g@cs

**Discussants**
You are responsible for helping **lead in-class discussion** once during the quarter.
Prepare a **short presentation** on the readings (15 min maximum). Highlight the main results, implications, and research opportunities.
**Moderate discussion** of the topic. Use your own questions or input from reading responses to spur an **engaged and sustained debate**.

**Guest Lecture Discussion**
You must also help **prepare questions for a guest lecture** once during the quarter.
Formulate a set of **5 or more critical discussion questions** prior to a guest lecture.
Not sure what to ask? Use the week’s readings; look at guests’ web pages, bios & publications.
Questions are due **by 8am** before the guest lecture, send them to cs448g@cs

**Assignments**
- **A0** Class Participation (30%) – quarter long
- **A1** A Failure of Analysis (5%) – due 4/4
- **A2** Analyzing Big Data (15%) – due 4/11, 4/18
- **FP** Final Project – 4/18 until quarter’s end
  (1) Select a domain and data set(s)
  (2) Identify a research problem, write abstract
  (3) Iteratively develop a prototype system
Assignment 0: Scheduling

By 5pm on Wed 3/31, please send us:

Your preferred email address.

Your top 3 picks for discussant topics. There are 8 options: W2 through W9.

Your top 3 picks for guest lecture questions. There are 9 options: W2 through W10.

Extra credit: generate questions for Wed 3/30.

Week 1 Readings


Illuminating the Path: The Research and Development Agenda for Visual Analytics. (Chapter 2: The Science of Analytical Reasoning)

A1: A Failure of Analysis  due Mon 4/4

Find and document a failure of analysis.

Write a short report (no more than 1 page) that describes the situation, the information serving as input to the analysis process, the nature of the failure, and its consequences.

If applicable, describe the failure in terms of cognitive biases. Suggest ways that improved analysis tools might have prevented the failure.

Questions?

hci.stanford.edu/courses/cs448g