

# **Making Data Work: When is Data Size a Problem?**

# History

- The oldest big data problems: censuses
  - Han dynasty
    - First census in 2 A.D to determine tax revenue and military strength
    - Counted 57,671,000 people
  - Roman Census
    - The word “census” comes from the latin “censere” meaning to estimate
    - A census was carried out every five years in the Roman empire to determine taxes
  - Domesday book
    - In 1086 AD, King William I of England wanted to know how much land was in his kingdom, how it was occupied, and who was using it
    - The Domesday book compiled all this information about England

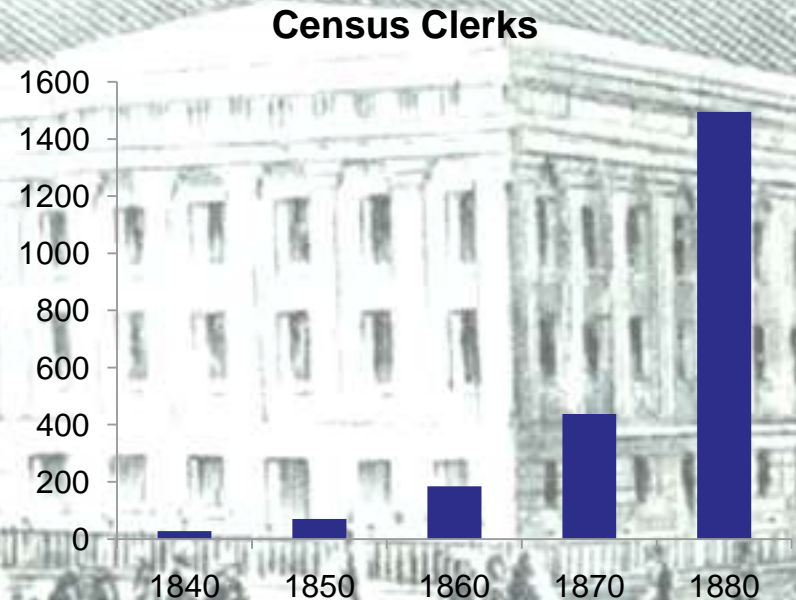
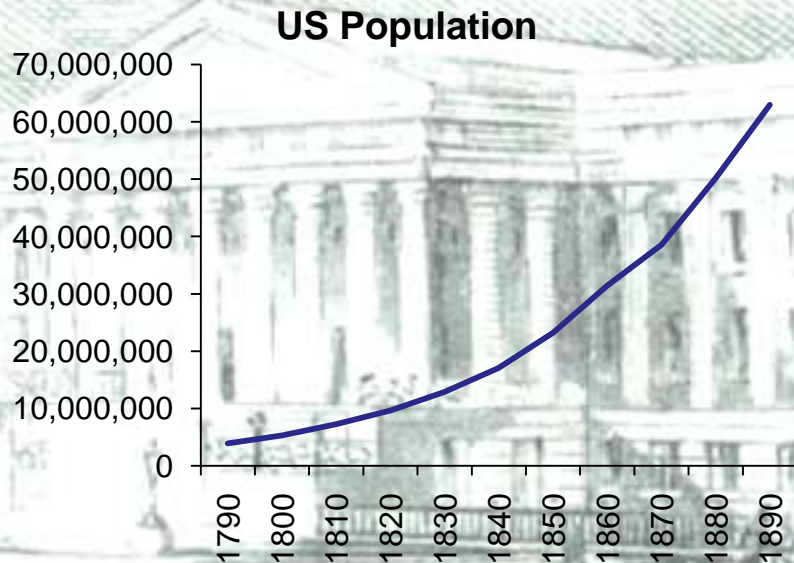
# History

- The US Constitution requires a census every 10 years:

Representatives and direct Taxes shall be apportioned among the several States which may be included within this Union, according to their respective Numbers... **The actual Enumeration shall be made within three Years after the first Meeting of the Congress of the United States, and within every subsequent Term of ten Years, in such Manner as they shall by Law direct.**

# History

- Between 1790 and 1890, the US population grew from 3 million to 60 million
- By 1880, tabulating data was a significant problem. It took 7 years to complete the 1880 census (whose results filled 3,473 pages).
- The census bureau was afraid that they would not finish tabulating results from the 1890 census until after they started collecting data for the 1900 census



# History

- Example from the 1880 Census:  
The proportion of male deaths to female deaths from various causes

TABLE 5.—SHOWING FOR THE UNITED STATES AND FOR 50 CITIES THE PROPORTION OF MALE DEATHS TO 1000 FEMALE DEATHS OF CORRESPONDING AGES.

| Deaths from—                            | PROPORTION OF MALE TO 1000 FEMALE DEATHS. |                |            |                | Deaths from—                  | PROPORTION OF MALE TO 1000 FEMALE DEATHS. |                |            |                |
|---|---|----------------|------------|----------------|-------------------------------|---|----------------|------------|----------------|
|   | United States.                            |                | 50 cities. |                |                               | United States.                            |                | 50 cities. |                |
|   | All ages.                                 | Under 5 years. | All ages.  | Under 5 years. |                               | All ages.                                 | Under 5 years. | All ages.  | Under 5 years. |
| Alcoholism .....                        | 5267.7                                    |                | 2371.5     |                | Pleurisy .....                | 1078.5                                    | 1128.2         | 1280.0     | 1111.1         |
| Suicide .....                           | 4052.3                                    |                | 3006.0     |                | Enteric fever .....           | 1071.4                                    | 1046.3         | 1105.1     | 1220.0         |
| Accidents and injuries.....             | 2732.0                                    | 1225.1         | 2075.2     | 1444.0         | Bronchitis .....              | 1055.3                                    | 1150.6         | 1031.5     | 1097.6         |
| Diseases of the urinary organs.....     | 2234.0                                    | 1378.2         | 1301.8     | 1225.1         | Malarial fever .....          | 1020.5                                    | 1069.0         | 1110.0     | 1203.2         |
| Tetanus and trismus nascentium .....    | 1645.4                                    | 1361.2         | 1408.0     | 1393.9         | Scrofula and tabes .....      | 1008.0                                    | 1088.6         | 900.9      | 945.5          |
| Still-born .....                        | 1418.4                                    | 1418.4         | 1311.4     | 1311.4         | Infanticide .....             | 1000.0                                    | 1000.0         | 2000.0     | 2000.0         |
| Diseases of the bones and joints.....   | 1366.7                                    | 1292.1         | 1338.2     | 1109.4         | Heart disease and dropsy..... | 989.3                                     | 1228.6         | 1001.0     | 1208.9         |
| Pneumonia .....                         | 1287.8                                    | 1221.3         | 1183.8     | 1119.8         | Measles.....                  | 972.6                                     | 1070.8         | 952.8      | 1017.4         |
| Diseases of the respiratory system..... | 1219.2                                    | 1200.0         | 1155.0     | 1130.2         | Scarlet fever.....            | 966.8                                     | 1009.7         | 983.4      | 1040.1         |
| Croup .....                             | 1187.5                                    | 1202.4         | 1180.9     | 1167.8         | Diphtheria .....              | 962.9                                     | 1081.7         | 962.2      | 1040.5         |
| Diseases of the nervous system.....     | 1170.7                                    | 1200.3         | 1214.0     | 1224.3         | Hooping-cough .....           | 865.4                                     | 870.9          | 797.7      | 804.0          |
| Venereal diseases.....                  | 1165.4                                    | 1041.3         | 1293.0     | 1080.0         | Consumption .....             | 798.1                                     | 1094.3         | 1014.2     | 1112.6         |
| Diseases of the digestive system .....  | 1147.5                                    | 1213.1         | 1175.4     | 1208.5         | Peritonitis.....              | 719.0                                     | 1354.4         | 768.0      | 1293.3         |
| Diarrhoeal diseases .....               | 1109.8                                    | 1155.1         | 1120.1     | 1120.1         | Cancer.....                   | 595.0                                     | 901.5          | 526.3      | 1000.0         |
| Paralysis and apoplexy.....             | 1002.8                                    | 1128.3         | 1120.5     | 1311.0         |                               |   |                |            |                |

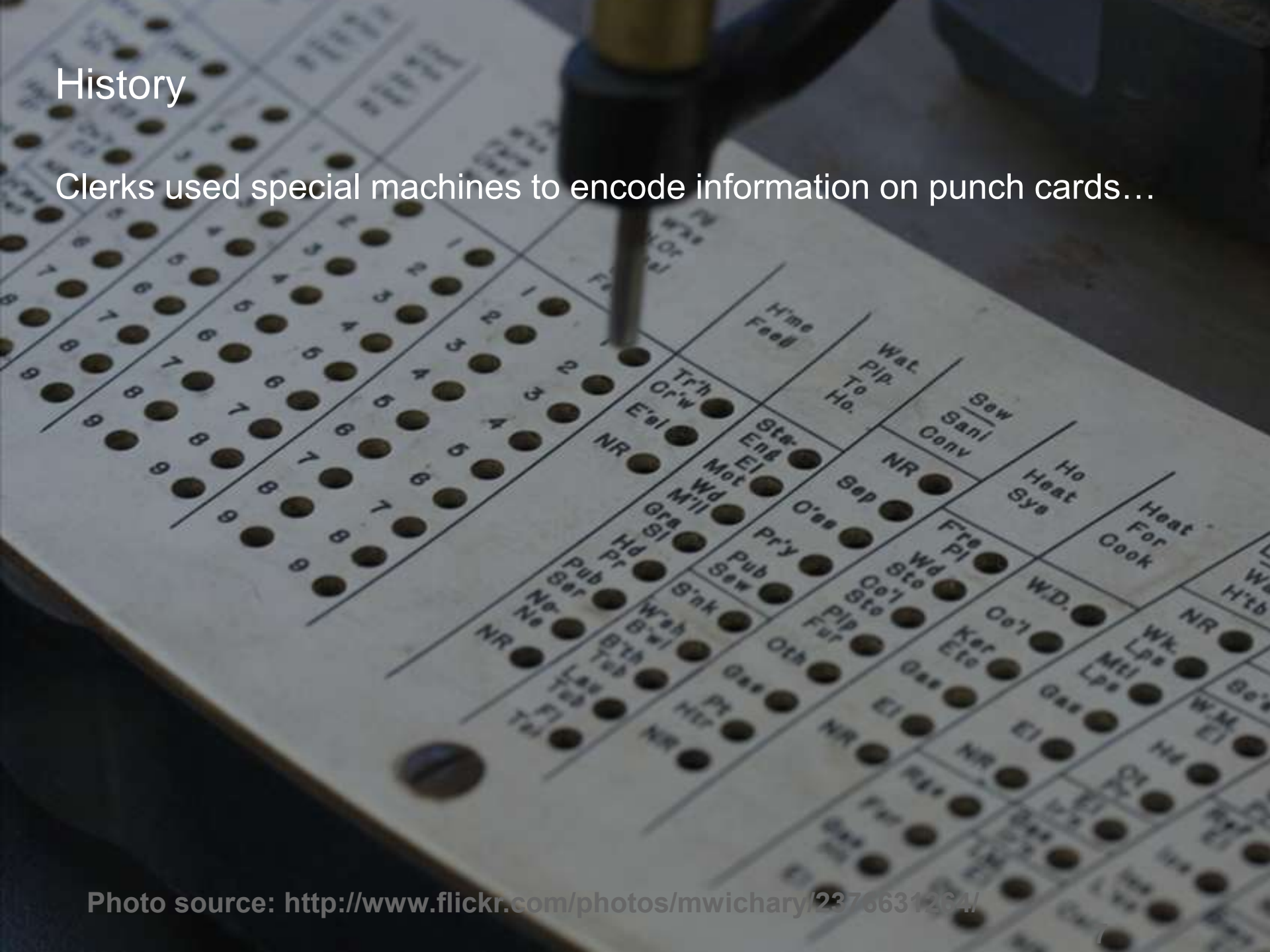
# History

- Herman Hollerith (February 29, 1860 – November 17, 1929)
- His advisor worked for the census bureau and hired Hollerith to be his assistant for the 1880 census, where Hollerith saw the pain of processing census results



# History

Clerks used special machines to encode information on punch cards...



# History

... which were then tabulated using machines designed by Hollerith

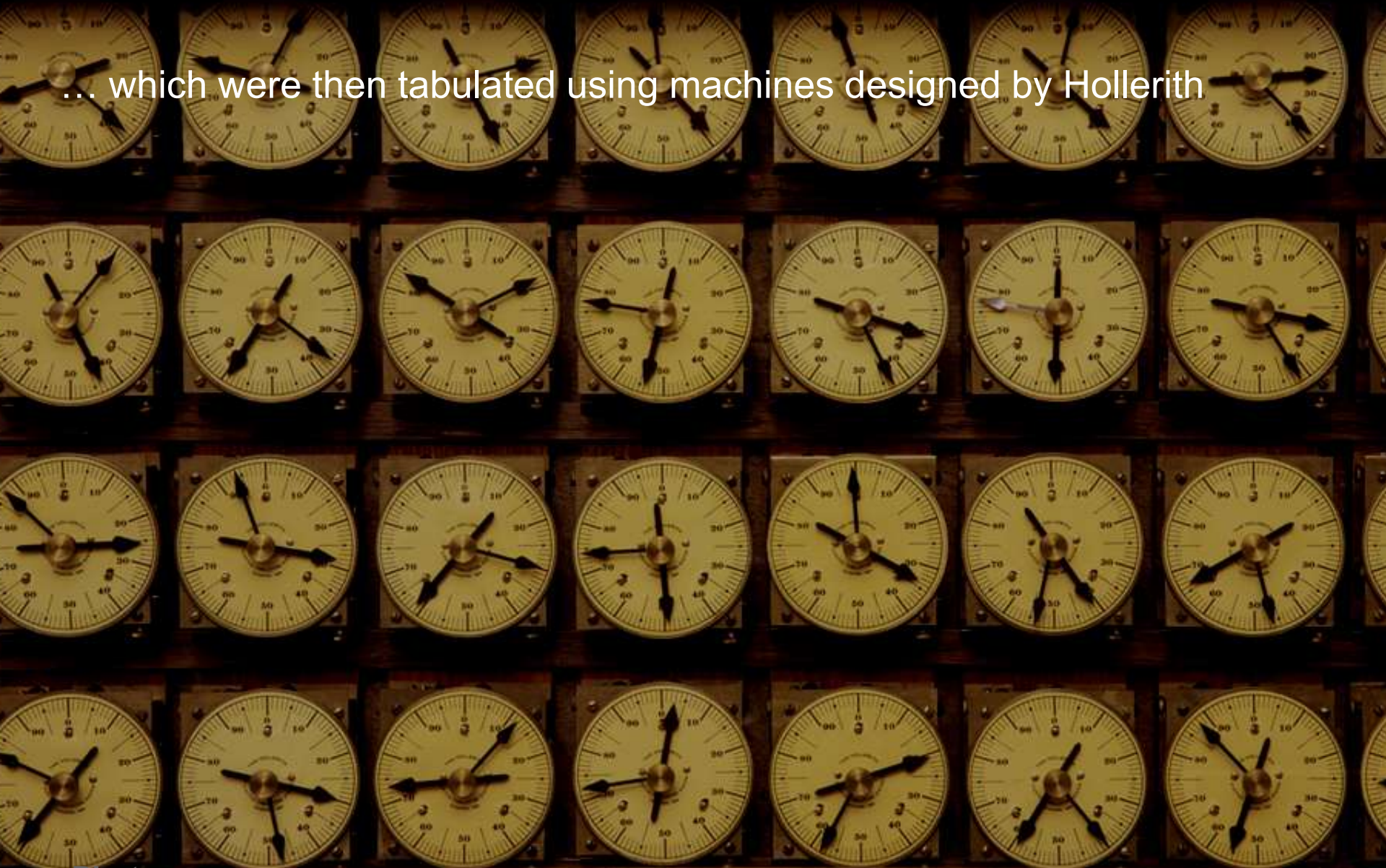


Photo source:

<http://www.flickr.com/photos/mwichary/4358926764/sizes/l/in/photostream>

# History

- 1890 Census Facts
  - 80 clerks operated the machines
  - 1000 cards per clerk per hour
  - 500,000 cards per day
  - Finished in 2.5 years
  - 24,408 pages of reports
  - Cost \$11.5 million
  - Machines saved approximately \$5 million
- In 1896 Hollerith incorporated his company as the “Tabulating Machine Company,” one of the companies that eventually became IBM

# History

- Herman Hollerith was the first data scientist
  - He understood his subject matter
  - He knew basic principles of math and science
  - He looked for a creative solution to his problem
- The 1890 Census was the first big data problem...



Today

- ... and the solution was very close to the Map-Reduce algorithms that we use today



# When is data size a problem?

- Example problem: Biomarker data

| <b>Technique</b> | <b>Commercially Viable</b> | <b>Data Size</b>                  |
|------------------|----------------------------|-----------------------------------|
| mRNA Microarrays | 1990s                      | Tens of thousands of variables    |
| SNP Microarrays  | 2000s                      | Hundreds of millions of variables |
| Gene Sequencing  | Today                      | Billions of variables             |

# When is data size a problem?

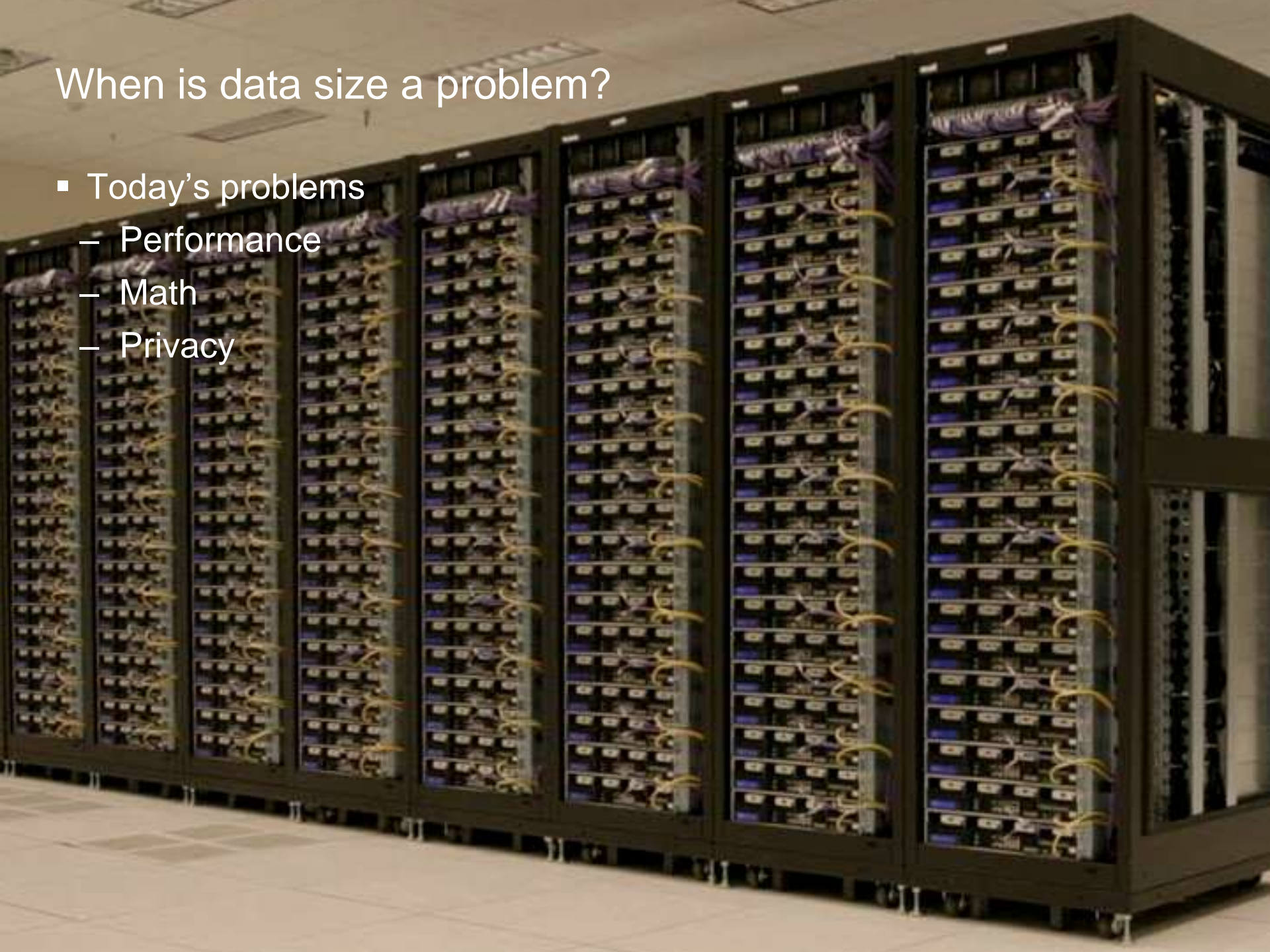
- Gene sequencing is becoming commercially feasible

|                        | Year      | Technology            | Cost            | Time     |
|------------------------|-----------|-----------------------|-----------------|----------|
| Human genome project   | 1991-2003 | Sanger                | \$3,000,000,000 | 12 years |
| Watson gene sequencing | 2007      | Next-Gen (Academic)   | \$2,000,000     | 2 months |
| Complete Genomics      | 2009      | Next-Gen (Commercial) | \$5,000         | 2 weeks  |
| ?                      | 2011      | ?                     | \$100?          | 1 Day?   |

- Each person's genes take up approximately **375 MB**

# When is data size a problem?

- Today's problems
  - Performance
  - Math
  - Privacy



# Performance

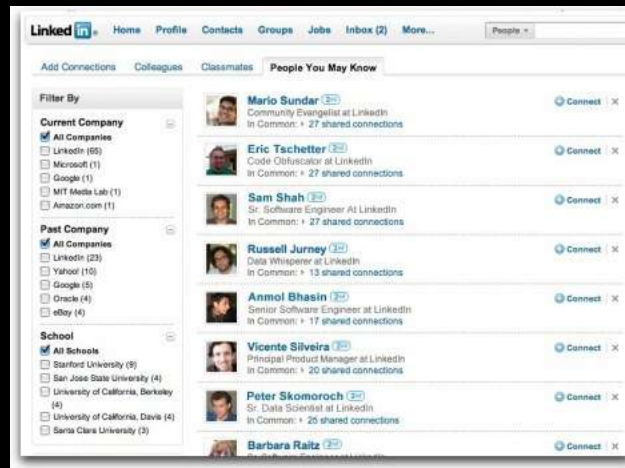
- Performance problems
  - On gigantic data sets, simple computations can be time consuming
    - Summary statistics (min, max, mean, etc)  $O(n)$
    - Sorting  $O(n \log n)$
    - Indexing  $O(n \log n)$

# Performance

- Many practical problems are harder
- Example: matrix decomposition and multiplication
  - Matrix operations occur everywhere
    - Scientific computing
    - Statistics
    - Social networks
  - Matrix multiplication
    - Theoretical bound: at least  $\Omega(n^2 \log n)$  steps
    - Best current algorithm:  $O(n^{2.37})$

# Performance

- Example:
  - At LinkedIn, we have built an analytical application called “People You May Know”
  - Making predictions for PYMK requires big matrix calculations
  - In the worst case, LinkedIn would need to consider whether every one of our 70,000,000 users knows every other user: that’s an  $O(n^2)$  problem



# Performance

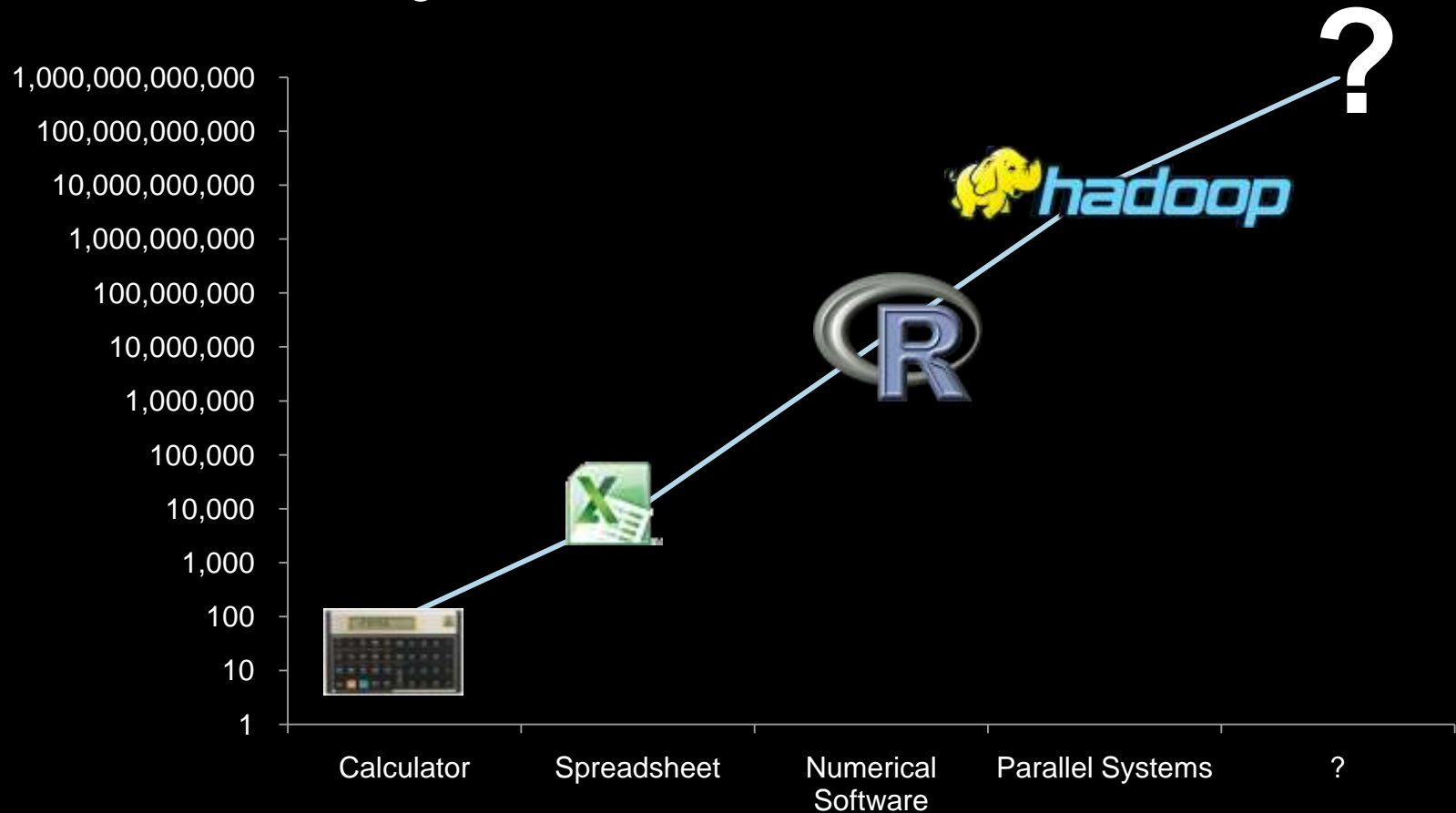
- Dealing with scale using your current tools
  - Compress your data
    - CPU speeds are much faster than I/O speeds
    - Compression helps keep data in memory, and not on disk
  - Eliminate data
    - Cut extraneous variables
    - Partition data (by time, demographics, etc)
    - Randomly sample

# Performance

- Relax your problem requirements
  - Solve an easier problem
  - Look for an approximate solution

# Performance

- Pick the right tool for the job
  - Learn something new!



# Math

- Math problem: “Wide” data
  - Wide means
    - More variables than observations
    - More columns than rows
  - Wide datasets come up naturally in many contexts
    - Biology data
    - Web data
    - Consumer finance data

# Math

- Learning from data
  - Calculating statistics
  - Looking for correlations
  - Modeling effects
    - Example: a linear model
$$y = c_0 + c_1x_1 + c_2x_2 + \dots + c_mx_m$$
- Wide data sets occur naturally in predictive models
  - One row per response
  - Many columns represent predictors

# Math

- It's easy to over-fit models with wide data
  - Analogy: picture a system of linear equations  $Ax=b$ :

$$\begin{aligned}a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n &= b_1 \\a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n &= b_2 \\a_{31}x_1 + a_{32}x_2 + \dots + a_{3n}x_n &= b_3 \\&\dots \\a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mn}x_n &= b_m\end{aligned}$$

- If  $m < n$  there are an infinite number of solutions
  - If  $m = n$  there is usually a unique solution
  - If  $m > n$ , there is usually no solution
- If you have more variables than observations, you may find false
    - Correlations between variables
    - Relationships between response variables and predictors

# Math

- Math problem: “Very Long Data”
  - Long means
    - More observations than variables
    - Many more rows than columns
  - It’s easy to find statistically significant effects in giant data sets, many of which aren’t meaningful

# Math

- Example:
  - While I was writing *R in a Nutshell*, I analyzed data from the CDC describing every birth in the United States
  - 4,273,225 million rows
  - Let's look at the number of male and female births by day of week

# Math

- 10% sample (from the book):

```
> table(births2006.smpl$SEX, births2006.smpl$DOB_WK)
```

```
      1      2      3      4      5      6      7
F 19709 30614 34332 34189 34310 33328 22171
M 20565 32143 35443 36101 35854 35052 23512
```

```
> b10 <- table(births2006.smpl$SEX, births2006.smpl$DOB_WK)
```

```
> b10[1,] / b10[2,]
```

```
      1      2      3      4      5      6      7
0.9583759 0.9524313 0.9686539 0.9470375 0.9569365 0.9508159 0.9429653
```

```
> chisq.test(b10)
```

Pearson's Chi-squared test

```
data: table(b10)
```

```
X-squared = 7.1446, df = 6, p-value = 0.3077
```

# Math

- 100% sample

```
> table(births2006.raw$SEX, births2006.raw$DOB_WK)
```

|   | 1      | 2      | 3      | 4      | 5      | 6      | 7      |
|---|--------|--------|--------|--------|--------|--------|--------|
| F | 196062 | 307095 | 340904 | 342233 | 342649 | 334942 | 221072 |
| M | 206863 | 322151 | 356963 | 360017 | 357580 | 350115 | 234579 |

```
> b100 <- table(births2006.raw$SEX, births2006.raw$DOB_WK)
```

```
> b100[1,] / b100[2,]
```

|  | 1         | 2         | 3         | 4         | 5         | 6         | 7         |
|--|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
|  | 0.9477867 | 0.9532642 | 0.9550121 | 0.9506023 | 0.9582443 | 0.9566628 | 0.9424203 |

```
> chisq.test(b100)
```

Pearson's Chi-squared test

```
data: table(b100)
```

```
X-squared = 26.8505, df = 6, p-value = 0.0001544
```

# Math

- How do you deal with these math problems?
  - Understand your subject matter
    - Weed out relationships that don't make sense
    - Example: Biologists restrict biomarkers to specific pathways
  - Limit your sample size
    - Partition your data into smaller sets (by time, demographics, whatever)
    - Select a sample of observations
  - Use more robust statistics
    - Cross-validation / holdout samples
    - Penalized models

# Privacy

- As a data scientist, it's tempting to use as much data as possible
  - Think about the fascinating things you could learn if you had a database with...
    - Credit history
    - Health care data
    - Income information
    - Web surfing habits
    - Purchase history (online and offline)
    - Social web site usage

# Privacy

- DoubleClick Case Study
  - In 1999, I worked for DoubleClick
  - I worked on a product to combine
    - Search queries
    - Sites visited
    - Ads clicked on with
    - Names and addresses
    - Demographic data
    - Catalog and internet purchases
  - Why?
    - We wanted to show more relevant ads to end users and make publishers more profitable



# Privacy

- What happened?
  - Bad press
  - Customer anger
  - Lawsuits



## **Surfer beware: Advertiser's on your trail DoubleClick tracks online movements**

January 26, 2000

The Internet's largest advertising company has begun tracking Web users' online movements, not just by anonymous identifying numbers but also by their actual names, addresses and real-world purchasing habits, USA TODAY has learned.

Privacy experts say it's the first time a huge real-world database has been linked to movements from site to site, even where the user has never made a purchase or registered.

DoubleClick, which places banner ads on 11,500 Web sites, has 100 million files of individual, usually anonymous, online behavior, often collected without surfers' knowledge. It's combining that with detailed data, including the phone numbers and purchasing habits, of 90 million U.S. households kept by direct-marketing services company Abacus Direct, which DoubleClick bought in June.

# Privacy

- When you put lots of data in a database,
  - You may alienate your users
  - You may be breaking the law
  - You may be creating a valuable target for criminals

# Questions and Answers

- email: [baseballhacks@gmail.com](mailto:baseballhacks@gmail.com)
- Twitter: @j Adler
- LinkedIn: <http://www.linkedin.com/in/josephadler>

