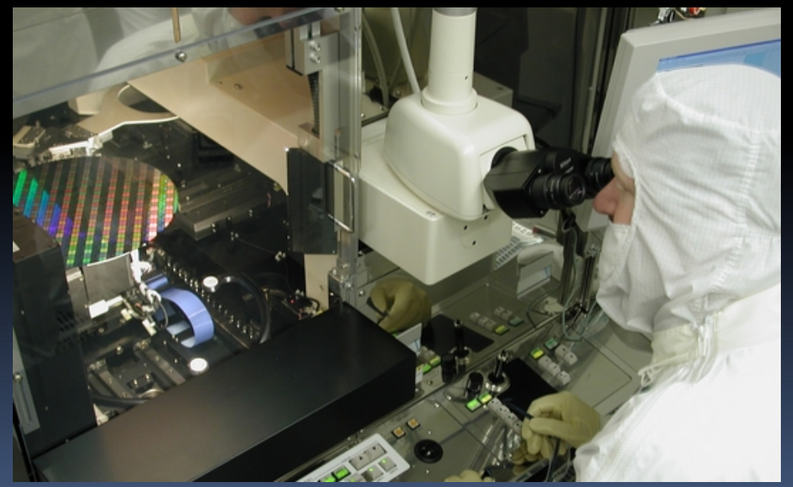
The Beauty and Joy of Computing

Lecture #16
Data and Information

"The Lost Language of Privacy"- Brooks

David Brooks waxes poetic about privacy in a NY Times piece about the pros/cons of police body cameras. He talks about why privacy is important to the development of full individuals, of families and friendships, and for communities.
Read it!



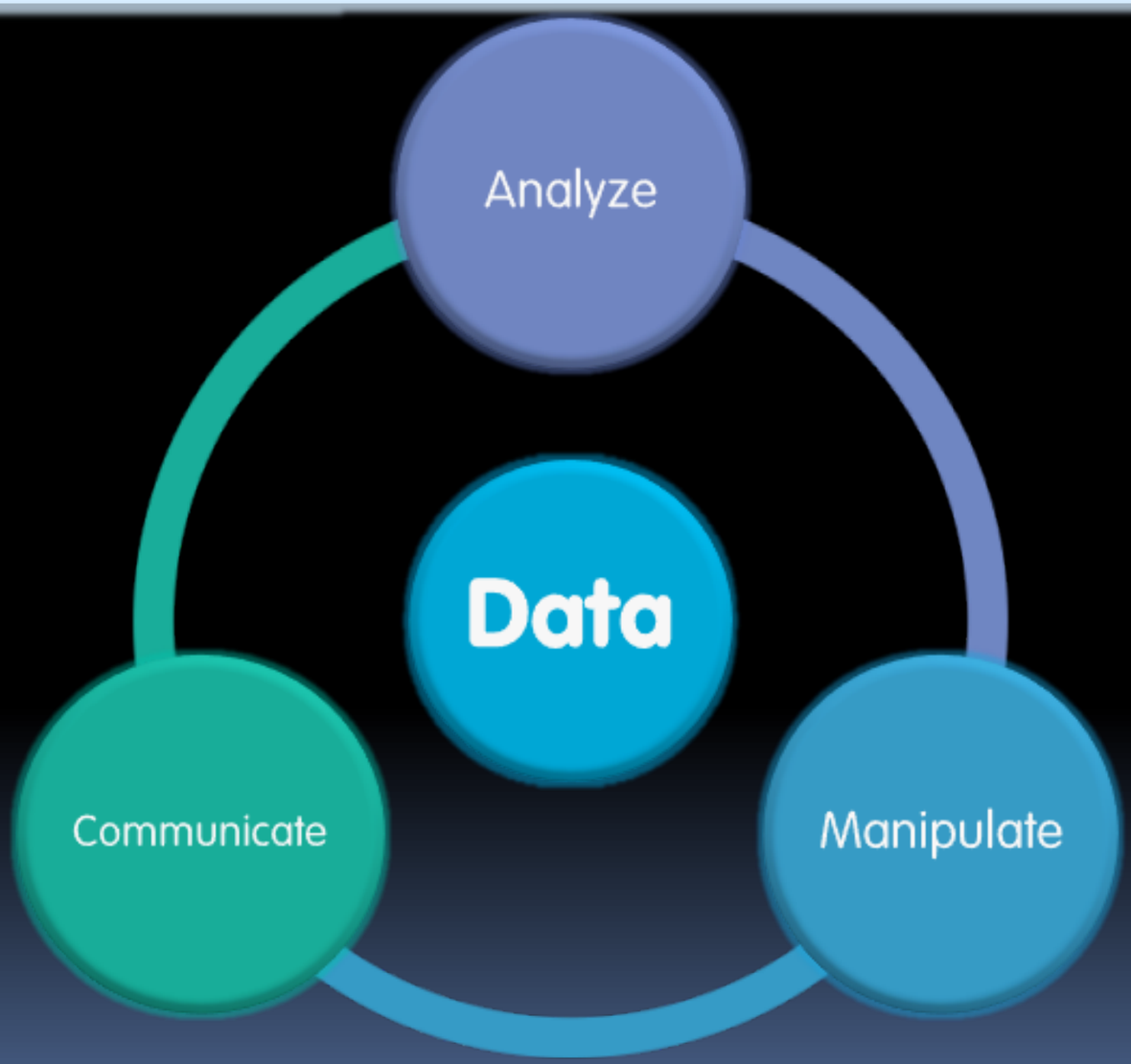
www.nytimes.com/2015/04/14/opinion/david-brooks-the-lost-language-of-privacy.html

Admin Notes

- Midterm tomorrow. Good Luck!
  - No Lecture
- "In-Lab" Portion:
  - 90 minutes
  - Take the survey
- Don't Stress!
  - Clobber Policy
- Next Week: "Explore" Post, a bit lighter.

Data and Information

3 Components



Data & Information Facilitate Knowledge

- Computers are used in an iterative and interactive way when processing digital information to gain insight and knowledge.
- Digital information can be filtered and cleaned by using computers to process information.
- Combining data sources, clustering data, and data classification are part of the process of using computers to process information.
- Insight and knowledge can be obtained from translating and transforming digitally represented information.
- Patterns can emerge when data is transformed using computational tools.

Ball

A Human Component

- Collaboration, in working with data to gain insight and knowledge:
  - …is an important part!
  - …applies multiple perspectives, experiences, and skill sets.
  - …involves communication between participants.
  - …includes developing hypotheses and questions, and in testing hypotheses and answering questions.
  - …can benefit from face-to-face and using online collaborative tools
  - …when Investigating large data sets can lead to insight and knowledge not obtained when working alone.

A Computational Component

- Large data sets
  - provide opportunities and challenges for extracting information and knowledge.
  - provide opportunities for identifying trends, making connections in data, and solving problems.
  - Need computing tools to facilitate the discovery of connections
- For efficiently finding, recognizing patterns and finding trends in information…
  - Search tools are essential.
  - Information filtering systems are important tools
  - Software tools, including spreadsheets and databases can help

Big Data Concerns

- Large data sets include data such as transactions, measurements, texts, sounds, images, and videos.
- The storing, processing, and curating of large data sets is challenging or sometimes impossible.
- Structuring large data sets for analysis can be challenging.
- Maintaining privacy of large data sets containing personal information can be challenging.
- Scalability of systems is an important consideration when data sets are large.
- Analytical techniques to store, manage, transmit, and process data sets change as the size of data sets scale.
- **Technical problems affect how data set is used**.

xkcd.com/1273

Data is Ubiquitous

…we work with it all the time:

- Data is collected any moment of your life
- Data is stored, copied, transmitted, deleted, edited.
- Computers perform operations on data
- Data enters and exits through sensors
- We can measure it!
  - 1 bit = '0' or '1'
  - 1 Byte = 8 bits
  - 1 KiB = 1024 Bytes, 1MiB = 1024 KiB, 1GiB = 1024 MiB, 1TiB=1024 GiB, 1PiB = 1024 TiB, …

How much is…?

- 1 KiB?
  - Long Paragraph of text (500-1,000 characters)
- 1 MiB?
  - 4 megapixel JPEG (compressed) image
  - .33 MP image UNcompressed
- 1 GiB?
  - One hour of standard def TV or 7 minutes of HDTV
- 1 TiB?
  - 2,000 hours of audio (uncompressed), 17,000 hours of MP3s
- 1 PiB?
  - Store the DNA of the entire population of the US – 3x!

Clicker Question

What do you think is the biggest data overall?

a) Text
b) Images
c) DNA
d) Videos
e) Census Data

Big Data, Compression, Metadata

Big Data

- Netflix is said to use 1 PB to store the videos for streaming.
- World of Warcraft is stored on 1.3PB to maintain the game.
- Internet Archive: About 10PB
- AT&T transfers about 30PB of data through its networks each day.
- YouTube processes about 40PB of videos a day.
  - Multimedia data is the biggest data!
- Warning: may change! The internet is big…

Challenges

- Storage
  - No single hard disk/memory unit can store the data
  - Need to parallelize harddisks
  - All the problems of concurrent programming!
    - How to access the data?
    - What if a disk fails?
    - How fast is the access (read, write, delete)?
    - Physical limits: Energy cooling

Helpful Techniques: Lossless Compression

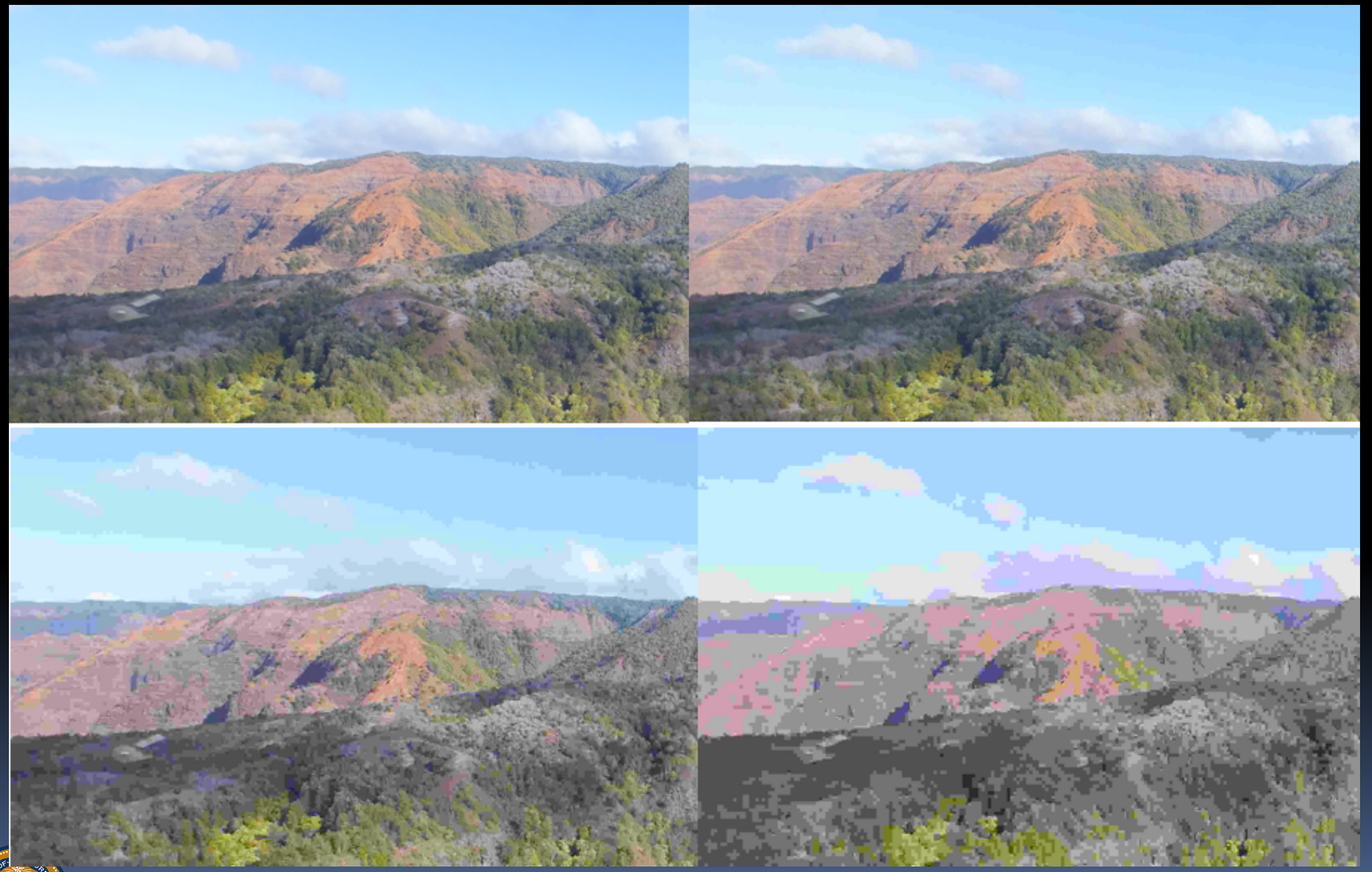
- Entropy compression reduces data volume by removing redundant information (so fewer bits are used to store or transmit it!)
- This compression is reversible but has mathematically proven limits.
- Example:

AAAAAABBBBBCCC -> 6A5B3C

Helpful Techniques: Lossy Compression

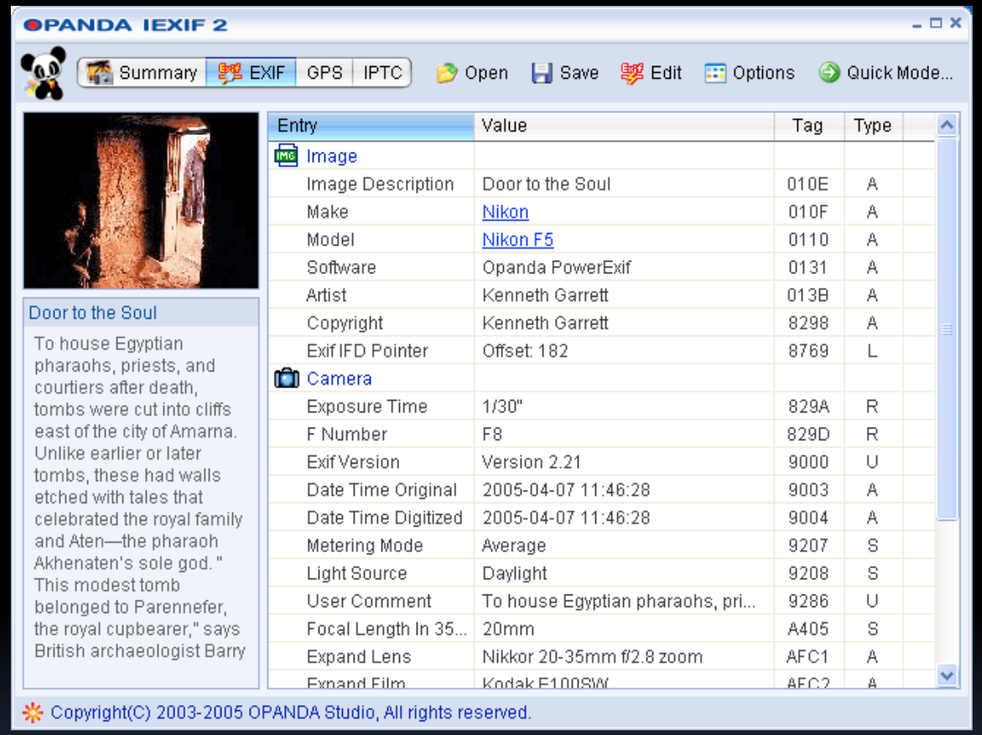
- Lossy compression reduces data volume by removing irrelevant information (and often can yield *smaller files* than lossless)
- This compression is not fully reversible but only has perceptual limits.
- There are trade-offs in using lossy and lossless compression techniques for storing and transmitting data.
- Compression needs an agreement on decompression = "format"

Lossy Compression Example: JPEG



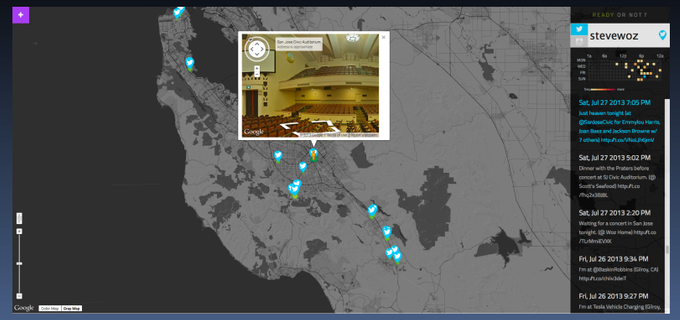
Techniques that help: Metadata

- Metadata: Data about data
  - can be descriptive data about an image, a Web page, or other complex objects.
  - can increase the effective use of data or data sets by providing additional information about various aspects of that data.
    - (e.g., search)



Main Reason for Digital Data

- Digital data can be copied without loss, and sent instantly, permanently to everyone.
  - Data is stored in many formats depending on its characteristics (e.g., size and intended use).
  - The choice of storage media affects both the methods and costs of manipulating the data it contains.
  - Reading data and updating data have different storage requirements.
- Problems with data that contains personal information:
  - Privacy
  - Security
- Trade-offs in storing, transmitting



Main Reasons for Big Digital Data

- Analyzing data at Internet-scale helps understand the world on never-before-seen scale.
- Useful for empirical sciences:
  - What are the economic trends based on Google searches?
  - Are there animals that dance to music without human training?
  - How is the flu progressing?
    - `www.google.org/flutrends/us/

Ball

Data: Conclusions and Visualization

Is Data the Solution to Everything?

- "Even" Internet data is biased
- It's easy to draw conclusions too quickly
- Sometimes finding the questions to ask is the hard part…
- E.g., NetFlix Prize
  - "Predict whether someone will enjoy a movie based on how much they liked or disliked other movies"
  - Dataset: users and movie ratings
  - What questions can we ask of this data set?

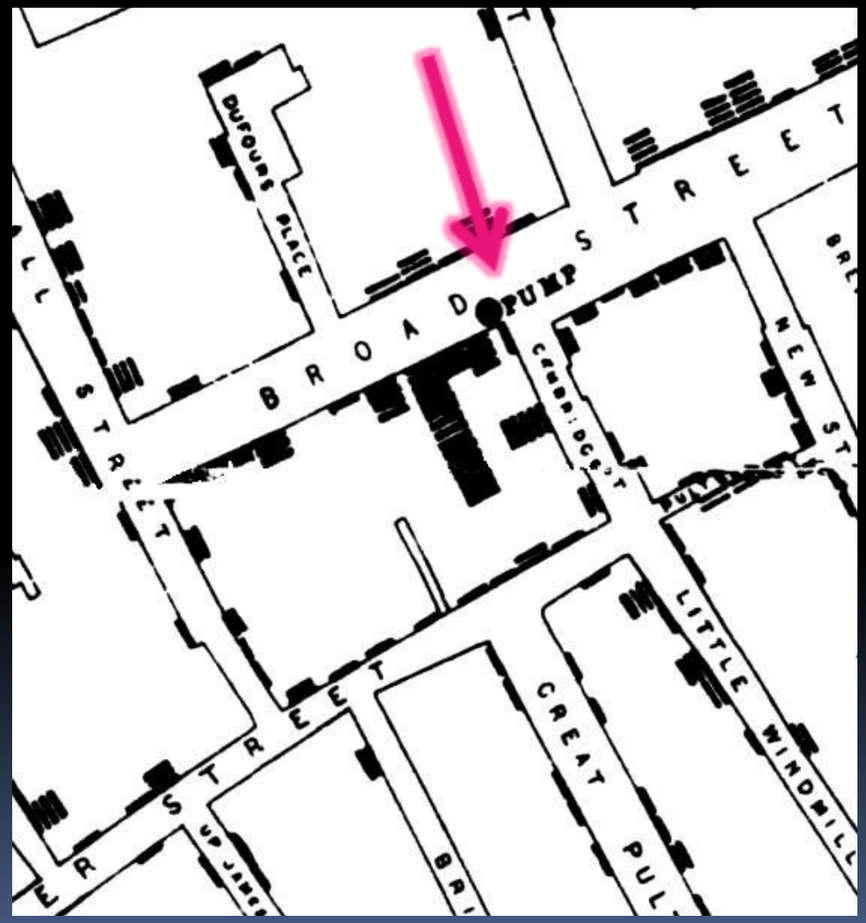
Correlation does not Imply Causality!

- cum hoc ergo propter hoc logical fallacy:
  - A occurs in correlation with B.
  - Therefore, A causes B.
- Just because A and B are correlated does not necessarily imply one causes the other! It could be that…
  - A may be the cause of B
  - B may be the cause of A
  - some unknown third factor C may actually be the cause of A and B.
  - A caused B AND B caused A. This is a self-reinforcing system.
    - E.g., "preditor-prey" relationships
  - the "relationship" is a coincidence or so complex or indirect that it is more effectively called a coincidence (i.e. two events occurring at the same time that have no direct relationship to each other besides the fact that they are occurring at the same time).

Benefits of Visualizations

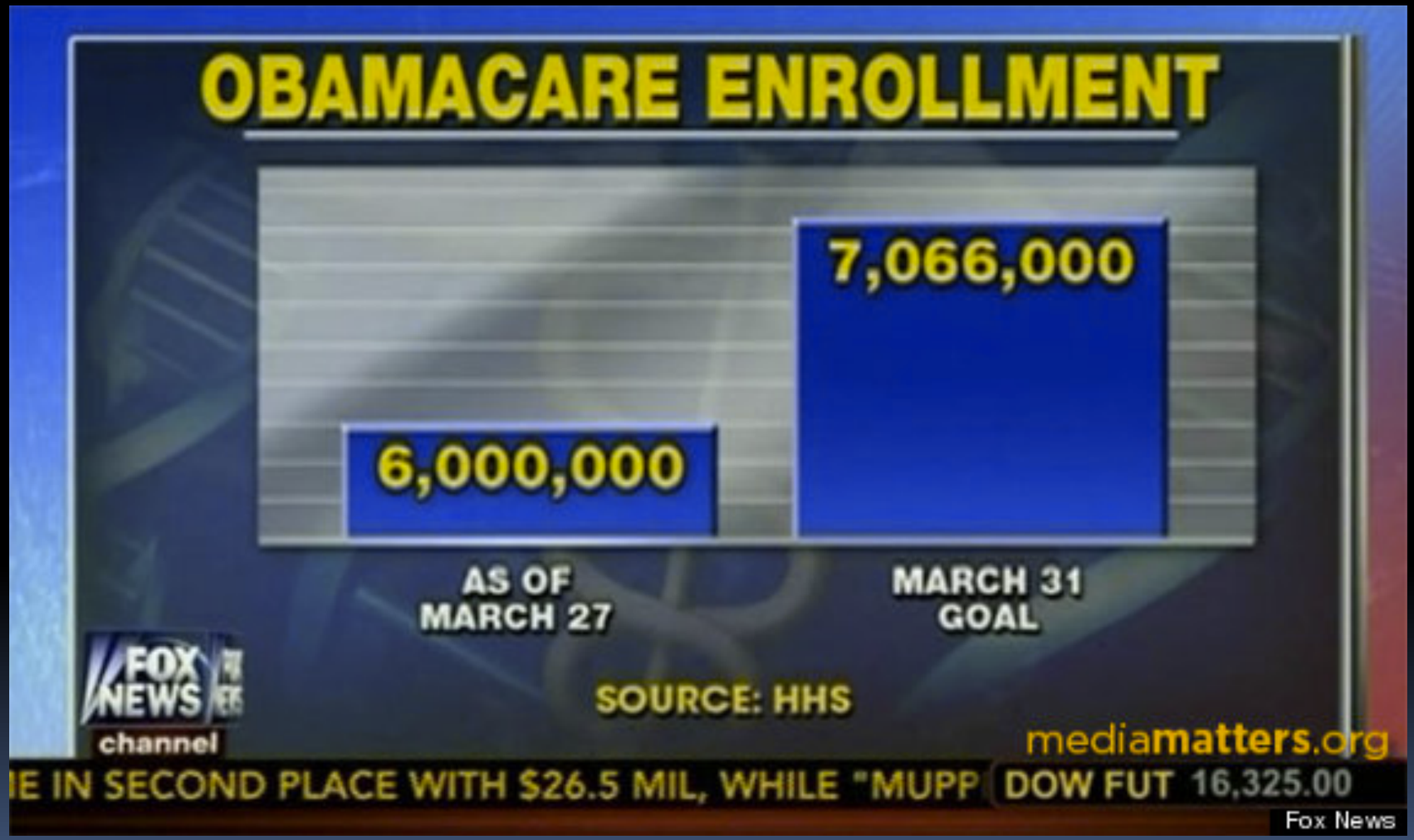
- Visualization tools and software can communicate information about data
- To communicate effectively the insights and knowledge gained from digitally represented information
  - Summarize the data analyzed computationally
  - Transform the information
  - Tables, diagrams, and textual displays
- Interactivity with data is an aspect of communicating

John Snow's visualization of the outbreak of Cholera in 1854 London



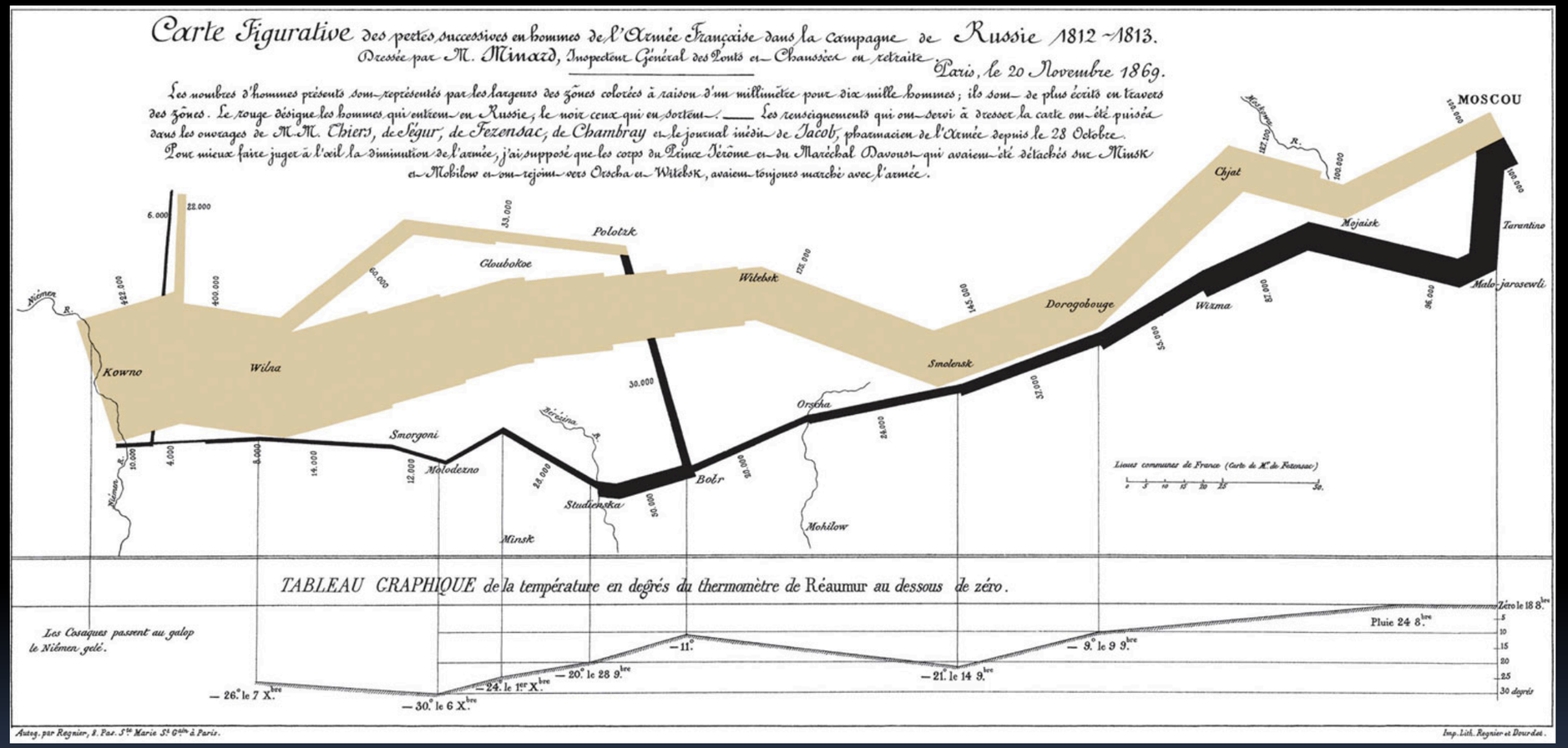
www.huffingtonpost.com/2014/03/31/fox-news-obamacare-graphic_n_5063582.html

Visualization … Epic FAIL (2014)



www.edwardtufte.com/tufte/posters

Visualization … Epic WIN (1869)

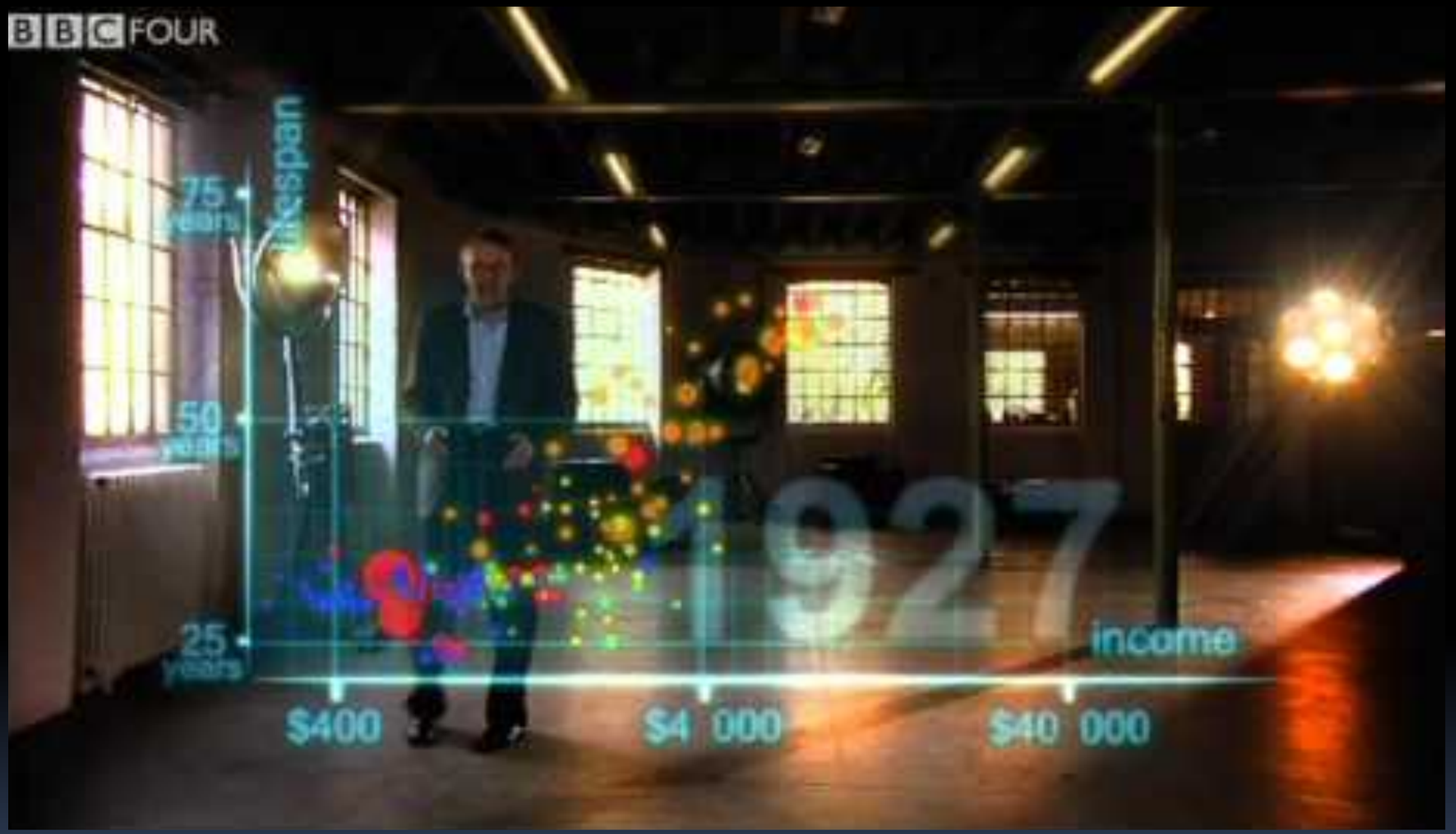


Charles Joseph Minard, Napoleon's 1812 Russian Campaign

youtu.be/jbkSRLYSojo

Visualization … Epic WIN (2009)

Additional Videos:
Short Visualizations:
https://www.youtube.com/watch?v=QpdyCJi3Ib4
Full TED Talk:
http://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen



Hans Rosling's 200 countries, 200 years, 4 minutes – the joy of stats

Summary

- The right questions need to be answered by the proper data.
- The rewards are high but handling data is an ongoing challenge to computer scientists as well as security specialists and privacy preservers.

Data & Information Facilitate Knowledge

- Computing enables and empowers new methods of information processing that have led to monumental change across disciplines, from art to business to science.
- Managing & interpreting an overwhelming amount of raw data is part of the foundation of our information society and economy.
- People use computers and computation to translate, process, and visualize raw data, and create information.
- Computation and computer science facilitate and enable a new understanding of data and information that contributes knowledge to the world.
- You will work with data using a variety of computational tools and techniques to better understand the many ways in which data is transformed into information and knowledge.