



UC Berkeley EECS  
Sr Lecturer SOE  
Dan Garcia

# The Beauty and Joy of Computing

Data



**Bendable  
Displays!!!**



<http://abcnews.go.com/Technology/lgs-flexible-screens-rolling-off-factory-lines/story?id=20498107>



# Data and 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.





# Ubiquitous data

---

...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' '1'
  - 1 Byte = 8 bit
  - 1 KB = 1024 Bytes, 1MB = 1024kB, 1GB = 1024MB, 1TB=1024GB, 1PB=1024TB, 1EB=1024PB, ...





# How much is?

---

- **1 KB?**
  - Paragraph of text
- **1 MB?**
  - 4 Mega pixel JPEG (compressed) image
- **1 GB?**
  - One hour of SD TV or 7 minutes of HDTV
- **1 TB?**
  - 2,000 hours of audio (uncompressed), 17,000 hours of MP3s
- **1 PB?**
  - Enough data to store the DNA of the entire population of the US – three times!





# The "biggest" data?

What do you think is the biggest data overall?

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





# 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 *biggest* data!





# Challenges

## ■ Storage

- No single hard disk/memory unit can store the data
- Need to parallelize haddisks
- 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





# Techniques that Help: Lossless Compression

- Entropy compression reduces data volume by removing **redundant** information
- This compression **is reversible** but has mathematically proven limits.
- Example:

AAAAAABBBBBCCC -> 6A5B3C







# Techniques that Help: Lossy Compression

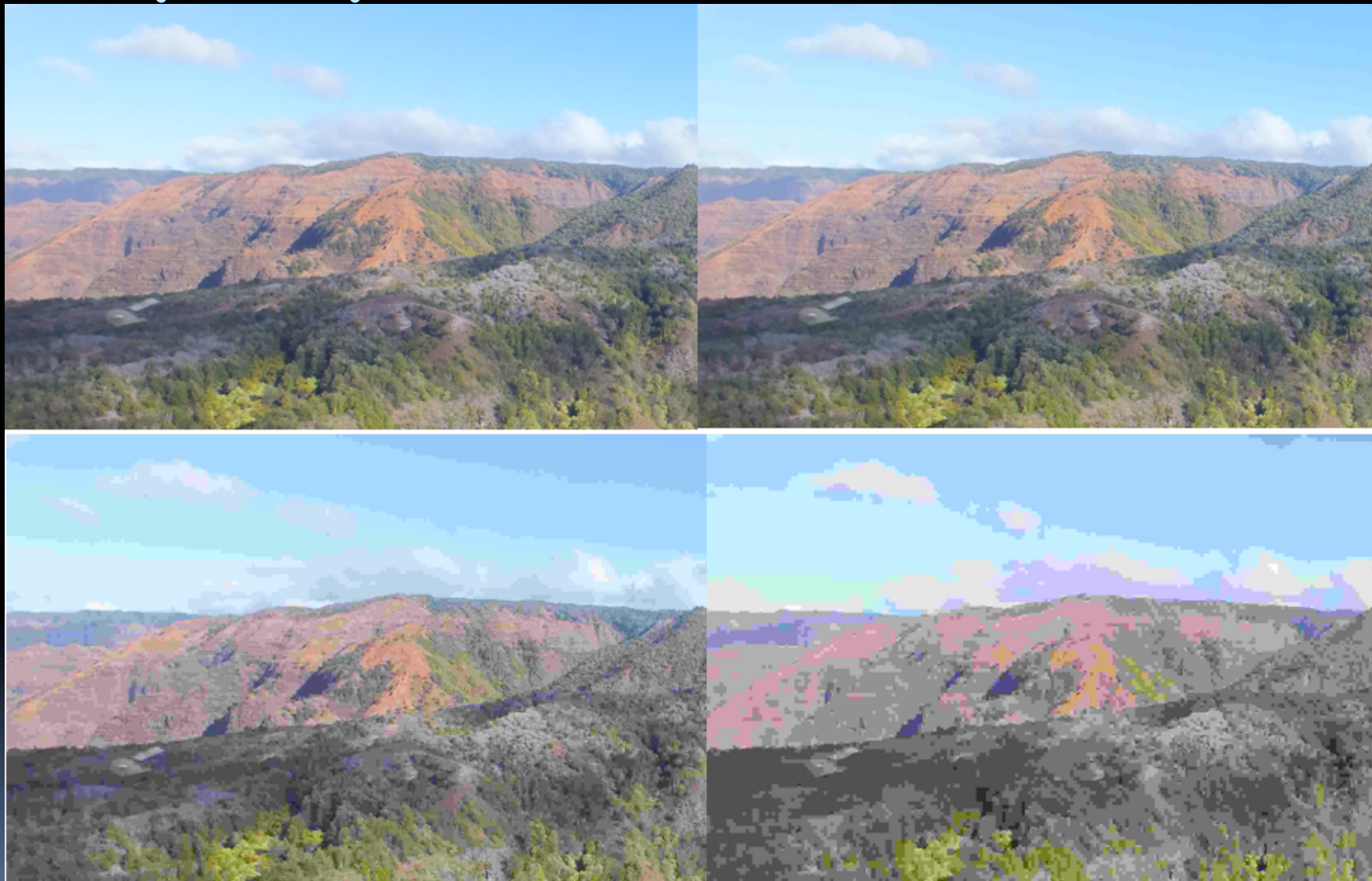
---

- Lossy compression reduces data volume by removing **irrelevant** information
- This compression is **not fully reversible** but only has perceptual limits.
  
- Compression needs an agreement on decompression = “format”





# Lossy Compression: JPEG





# Techniques that help: Metadata

- **Metadata: Data about data. Helps processing of data, e.g. search**
- **Example:**

The screenshot shows the OPANDA IEXIF 2 application window. The interface includes a menu bar with options like Summary, EXIF, GPS, and IPTC. A toolbar contains icons for Open, Save, Edit, Options, and Quick Mode. On the left, there is a thumbnail of the image and a text description. The main area displays a table of EXIF metadata.

Entry	Value	Tag	Type
Image			
Image Description	Door to the Soul	010E	A
Make	<a href="#">Nikon</a>	010F	A
Model	<a href="#">Nikon F5</a>	0110	A
Software	Opanda PowerExif	0131	A
Artist	Kenneth Garrett	013B	A
Copyright	Kenneth Garrett	8298	A
Exif IFD Pointer	Offset: 182	8769	L
Camera			
Exposure Time	1/30"	829A	R
F Number	F8	829D	R
Exif Version	Version 2.21	9000	U
Date Time Original	2005-04-07 11:46:28	9003	A
Date Time Digitized	2005-04-07 11:46:28	9004	A
Metering Mode	Average	9207	S
Light Source	Daylight	9208	S
User Comment	To house Egyptian pharaohs, pri...	9286	U
Focal Length In 35...	20mm	A405	S
Expand Lens	Nikkor 20-35mm f/2.8 zoom	AFC1	A
Expand Film	Kodak E100SW	AFC2	A

Copyright(C) 2003-2005 OPANDA Studio, All rights reserved.





# Two Main Reason for Digital Data

- Digital data can be copied without loss.
- Digital data can be processed by a computer, e.g. for search
- Problems:
  - Privacy
  - Security

The screenshot displays a map interface with several blue Twitter bird icons scattered across the area, representing digital data points. An inset window shows a photograph of the San Jose Civic Auditorium with the text "San Jose Civic Auditorium Address is approximate". To the right of the map is a calendar view for the week of July 22-28, 2013, with a header "READY OR NOT?". Below the calendar, a list of tweets is visible, including one from @stevewoz dated Saturday, July 27, 2013, at 7:05 PM, mentioning a concert at the San Jose Civic.





# One Main Reason for Big 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/](http://www.google.org/flutrends/us/)





en.wikipedia.org/wiki/Correlation\_does\_not\_imply\_causation!

# 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., "predator-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).





# Is Data the Solution to Everything?

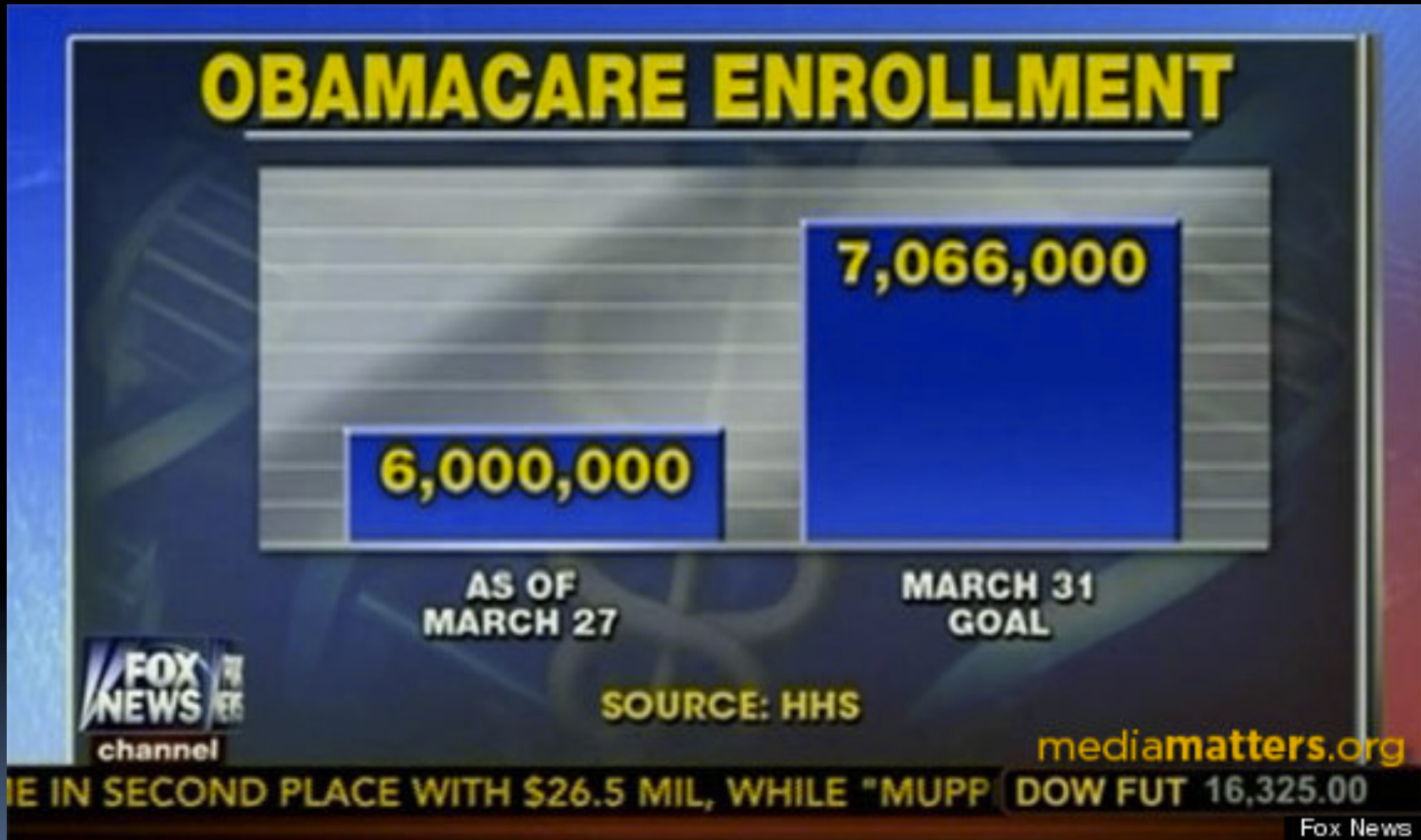
---

- “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?





# Visualization ... Epic FAIL



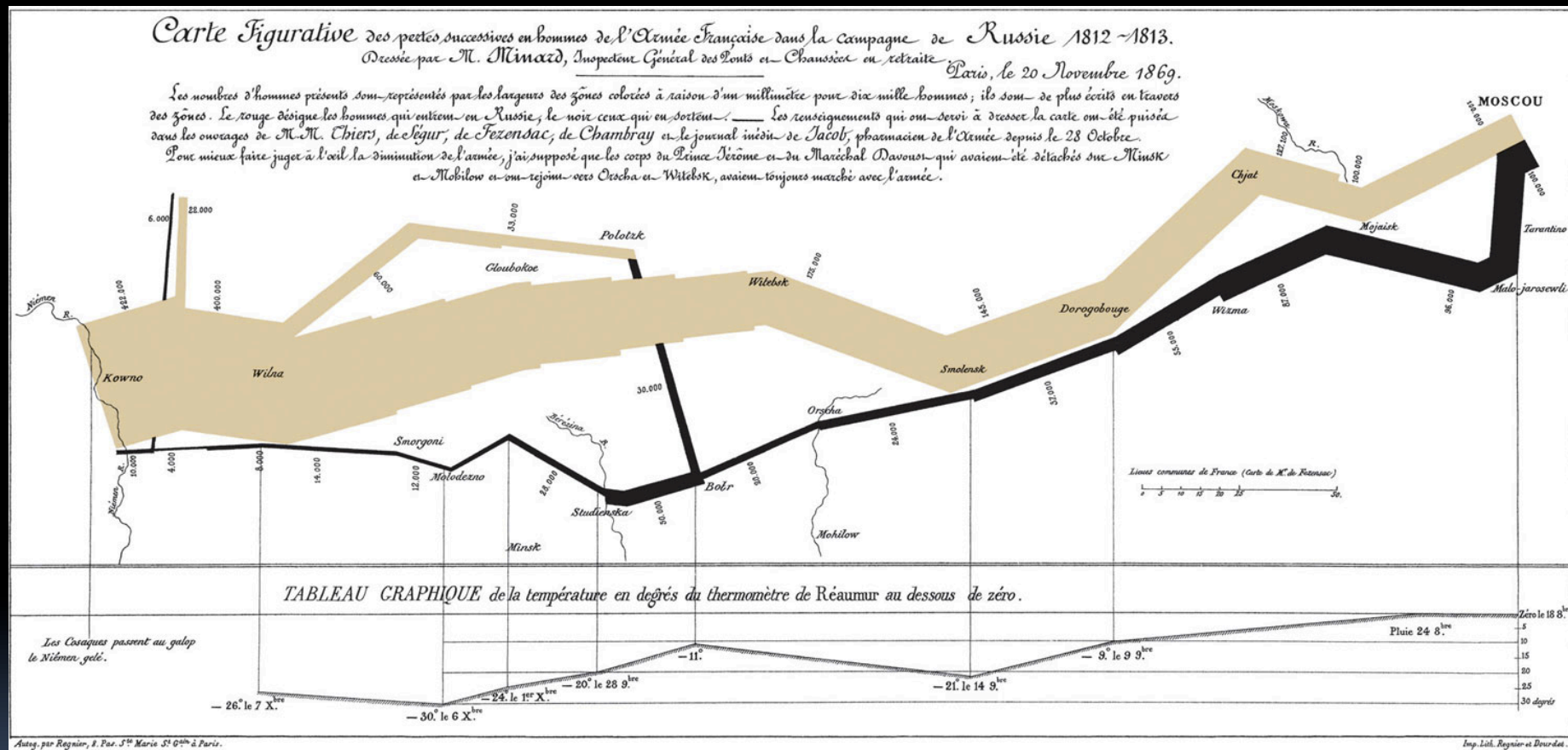
Fox News Graphic on the "Obamacare" Enrollment as of 2014-03-27







# Visualization ... Epic WIN



## Charles Joseph Minard, Napoleon's 1812 Russian Campaign



