

**The Beauty and Joy of Computing**

**Data**



**Bendable Displays!!!**

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

UC Berkeley EECS  
Sr Lecturer SOE  
Dan Garcia

*bjc*

**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.

UC Berkeley "The Beauty and Joy of Computing" : Data (2)

**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, ...

UC Berkeley "The Beauty and Joy of Computing" : Data (3)

**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!

UC Berkeley "The Beauty and Joy of Computing" : Data (4)

**The "biggest" data?**

What do you think is the biggest data overall?

- Text
- Images
- DNA
- Videos
- Census Data




UC Berkeley "The Beauty and Joy of Computing" : Data (5)


**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!

UC Berkeley "The Beauty and Joy of Computing" : Data (6)

## 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



UC Berkeley "The Beauty and Joy of Computing" : Data (7)

## Techniques that Help: Lossless Compression

- Entropy compression reduces data volume by removing redundant information
- This compression is reversible but has mathematically proven limits.
- Example:
   
AAAAAABBBBBBCCC -> 6A5B3C

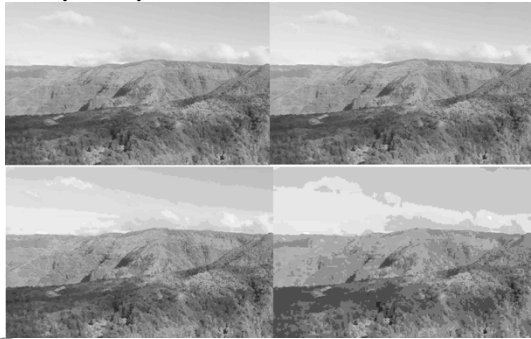
UC Berkeley "The Beauty and Joy of Computing" : Data (8)

## 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"

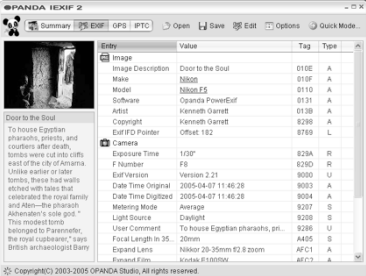
UC Berkeley "The Beauty and Joy of Computing" : Data (9)

## Lossy Compression: JPEG



UC Berkeley "The Beauty and Joy of Computing" : Data (10)


## Techniques that help: Metadata

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

UC Berkeley "The Beauty and Joy of Computing" : Data (11)

## 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



UC Berkeley "The Beauty and Joy of Computing" : Data (12)

**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/flu-trends/us/

UC Berkeley "The Beauty and Joy of Computing": Data (3)

**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).

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

UC Berkeley "The Beauty and Joy of Computing": Data (4)

**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?

UC Berkeley "The Beauty and Joy of Computing": Data (5)

**Visualization ... Epic FAIL**

www.huffingtonpost.com/2014/03/31/fox-news-obamacare-graphic\_n\_5063582.html

UC Berkeley "The Beauty and Joy of Computing": Data (6)

**Visualization ... Epic WIN**

www.edwardtufte.com/tufte/posters

UC Berkeley "The Beauty and Joy of Computing": Data (7)

**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.

UC Berkeley "The Beauty and Joy of Computing": Data (8)