



The Beauty and Joy of Computing

Lecture #10 Data



UC Berkeley EECS
Lecturer
Gerald Friedland

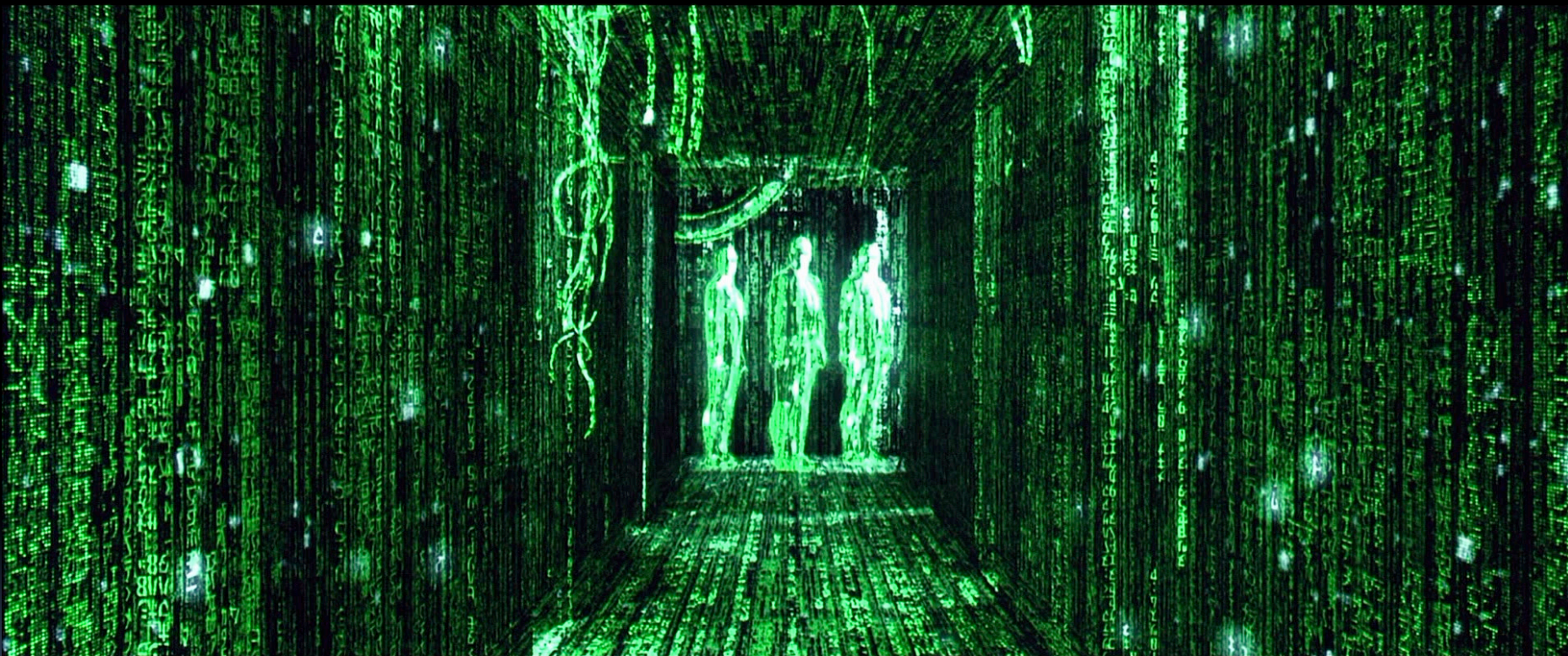
Bendable Displays!!!



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



Data: A Definition





We don't really know what it is...

...but 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?

- **1kB?**
 - Paragraph of text
- **1 MB?**
 - 4 Mega pixel JPEG (compressed) image
- **1GB?**
 - One hour of SD TV or 7 minutes of HDTV
- **1TB?**
 - 2,000 hours of audio (uncompressed), 17,000 hours of MP3s
- **1PB?**
 - 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





BIGDATA

- 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 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





Techniques that Help: Compression

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

- **Example:**

AAAAAABBBBBBCCC -> 6A5B3C





Techniques that Help: Compression

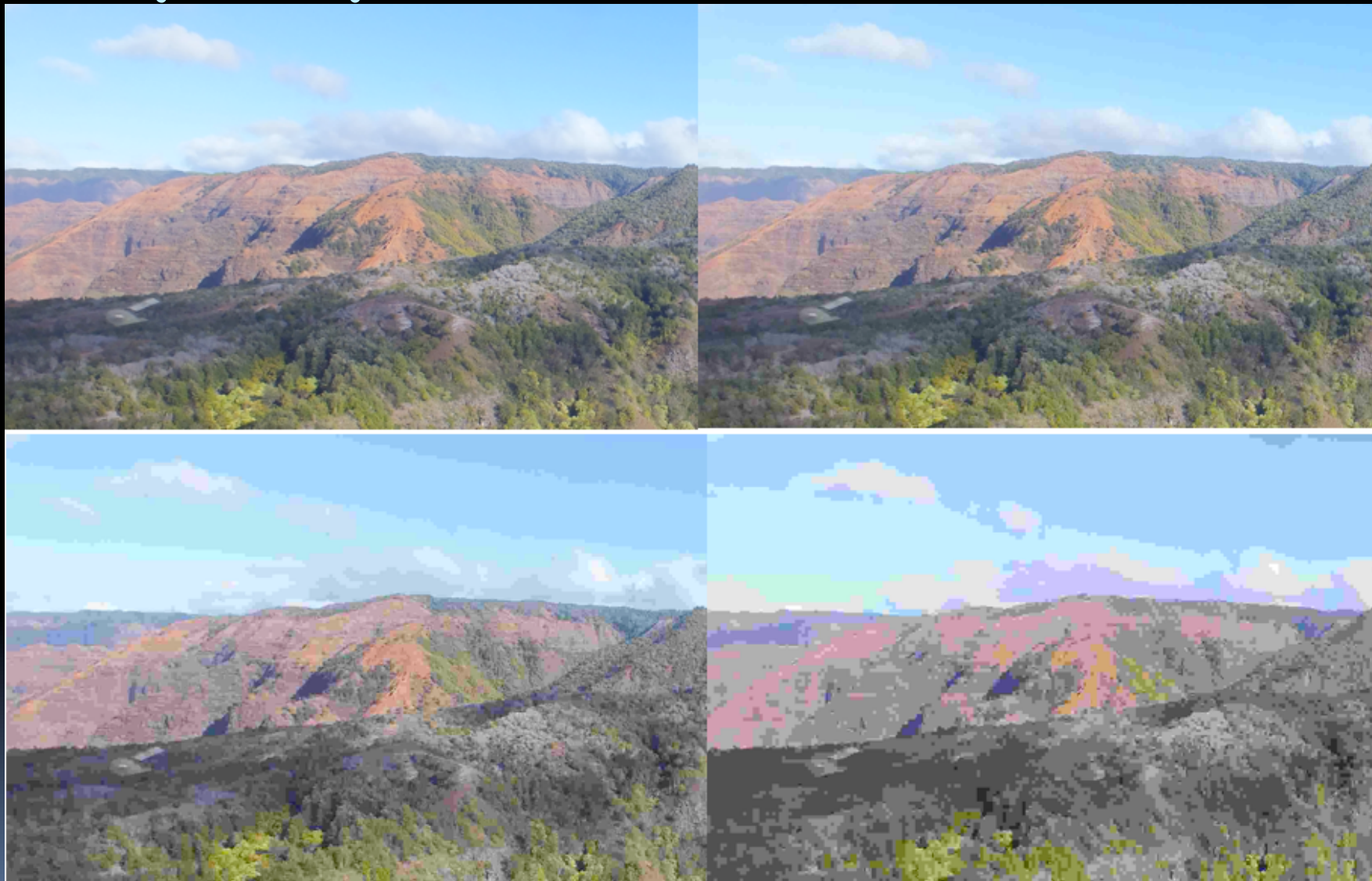
- Lossy compression reduces data volume by removing **irrelevant** information
- This compression is **not fully reversible** but only has perceptual limits.

- Compression needs an agreement on decompression = “format”





Lossy Compression: JPEG



Friedland





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	Nikon	010F	A
Model	Nikon F5	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 shows a map interface with several blue Twitter bird icons scattered across a geographic area. A central window displays a photograph of the interior of the San Jose Civic Auditorium, with the text "San Jose Civic Auditorium" and "Address is approximate" above it. To the right of the map is a vertical sidebar containing a calendar for the week of July 22-28, 2013, and a list of tweets from the user "stevewoz". The tweets include dates and times such as "Sat, Jul 27 2013 7:05 PM" and "Sat, Jul 27 2013 5:02 PM", along with text about concerts and dinners. At the bottom of the map area, there are "Google", "Color Map", and "Gray Map" options.





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?
- But privacy is a challenge: Future Lecture





Is Data the Solution to Everything?

- Careful: Correlation does not imply causation
- “Even” Internet data is biased
- It’s easy to draw conclusions too quickly

- The right questions need to be answered using proper data





Asking the Right Question

If you choose an answer to this question at random, what is the chance you will be correct?

- A) 25%
- B) 50%
- C) 60%
- D) 25%



e) None of the above



<http://flowingdata.com/2011/10/28/best-statistics-question-ever/>

UC Berkeley "The Beauty and Joy of Computing" : Data (15)

Friedland



