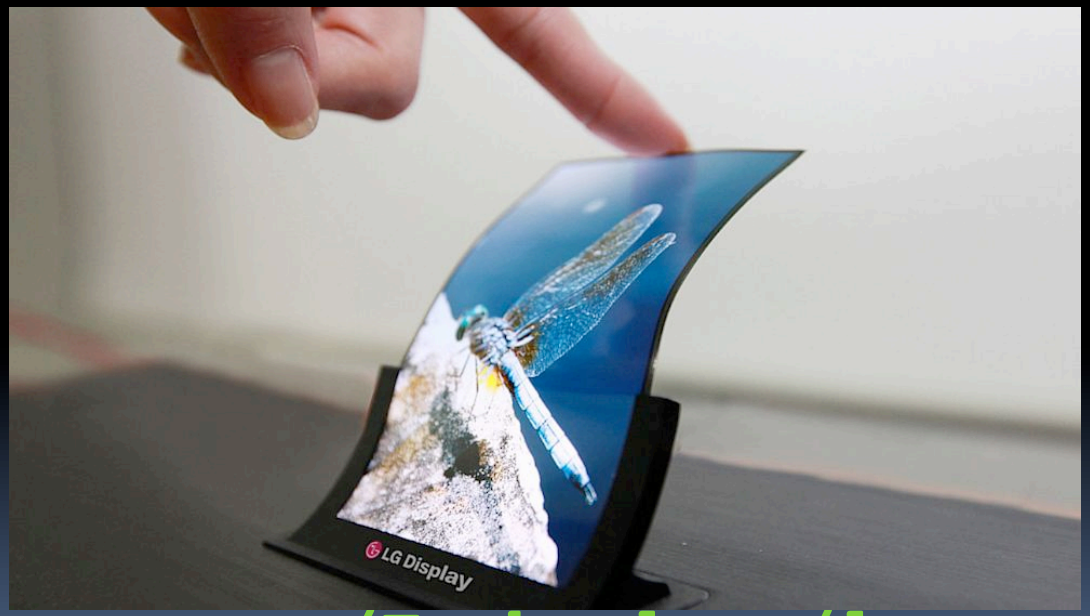
The Beauty and Joy of Computing

Lecture #10
Data

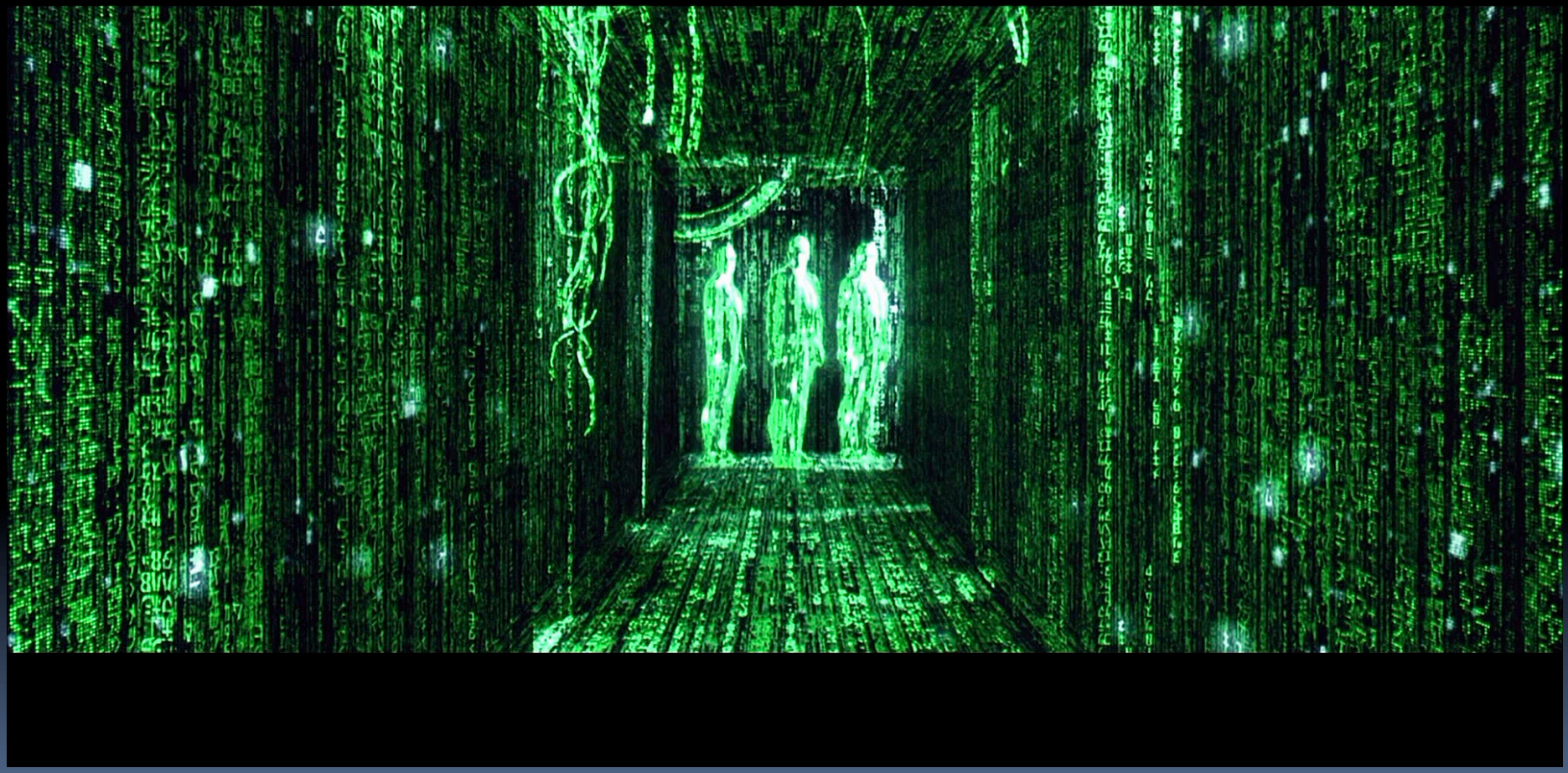
UC Berkeley EECS
Lecturer
Gerald Friedland

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 live
- 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!

Friedland

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!

Friedland

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:

AAAAAABBBBBCCC -> 6A5B3C

Friedland

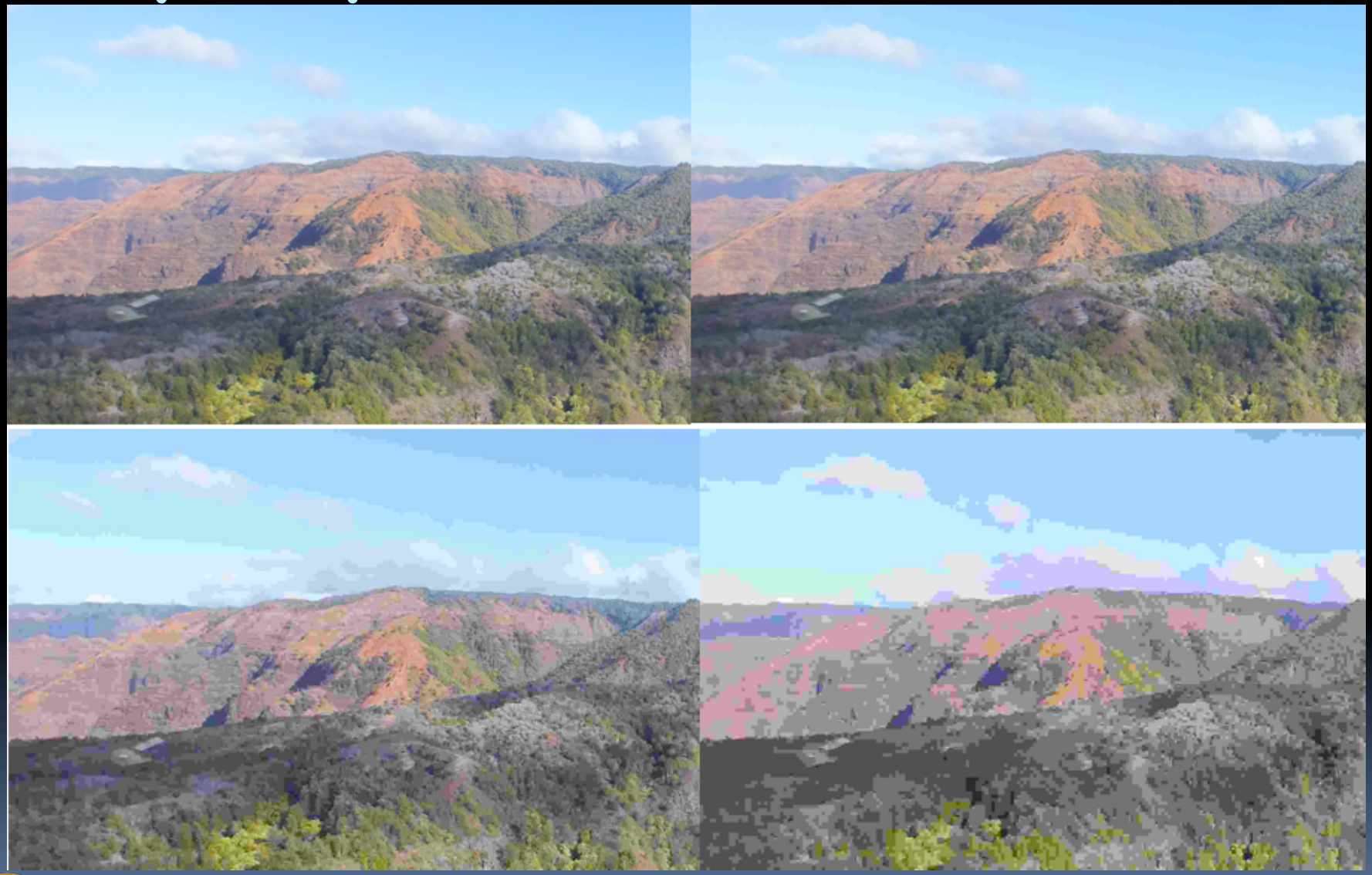
Techniques that Help: Compression

- 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"

Friedland

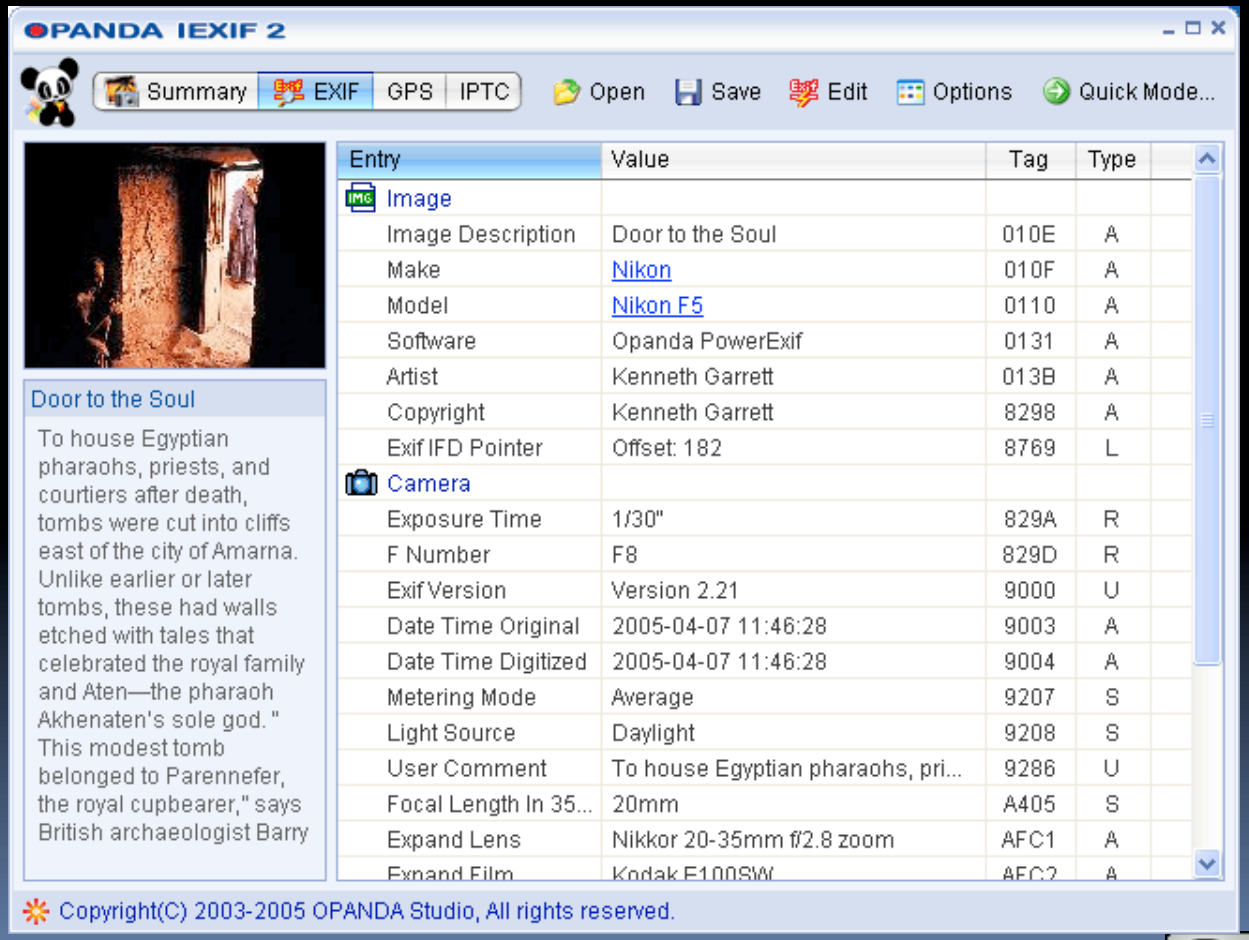
Lossy Compression: JPEG



Friedland

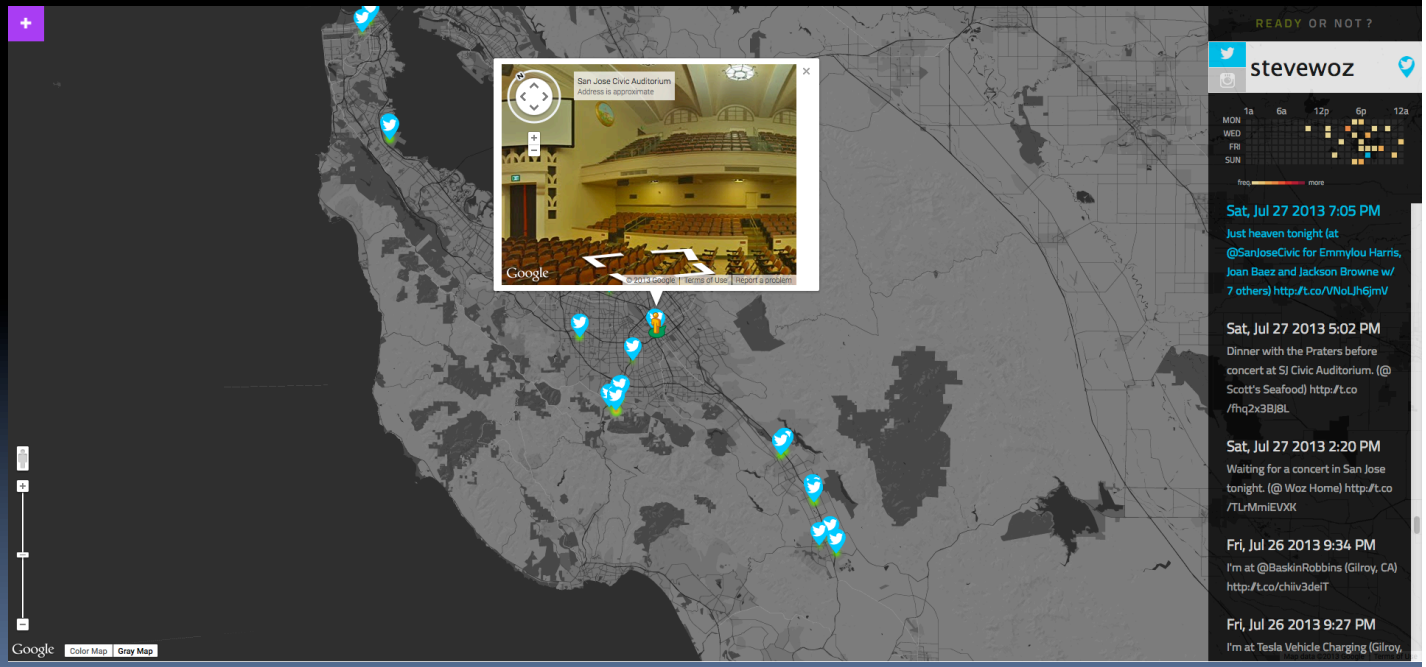
Techniques that help: Metadata

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



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



One Main Reason for Big Data

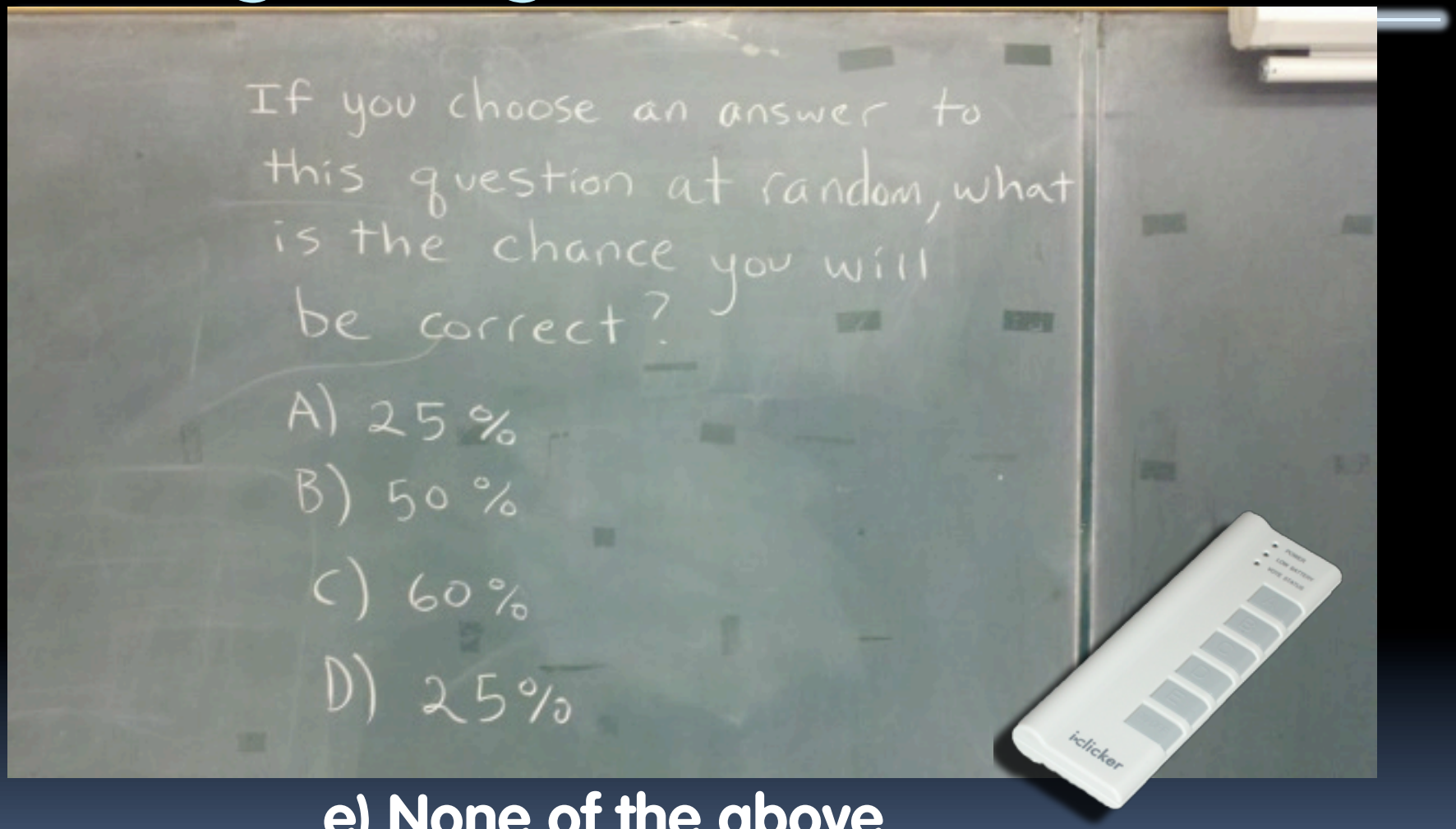
- **Analyzing data at Internet-scale helps understand the world on never before seen scale.**
- **Useful for empirical sciences:**
  - What are the economic trends based on Google searches?
  - Are there animals that dance to music without human training?
  - How is the flu progressing?
- **But privacy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



e) None of the above

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

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.