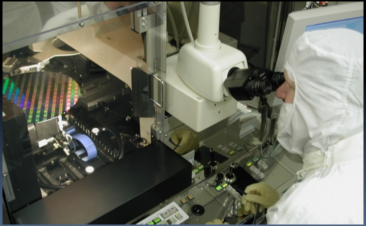
UC Berkeley EECS
Teaching Professor
Dan Garcia

The Beauty and Joy of Computing

Lecture #21
Data and Information

"The Lost Language of Privacy"- Brooks

David Brooks waxes poetic about why privacy is important in a NY Times piece about the pros and 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

(Cal) Admin Notes

- Schedule (see website)

Data and Information

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.

Garcia

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 bits
  - 1 KiB = 1024 Bytes, 1MiB = 1024 KiB, 1GiB = 1024 MiB, 1TiB=1024 GiB, 1PiB = 1024 TiB, …

How much is?

- 1 KiB?
  - Paragraph of text
- 1 MiB?
  - 4 Mega pixel JPEG (compressed) image
- 1 GiB?
  - One hour of SD TV or 7 minutes of HDTV
- 1 TiB?
  - 2,000 hours of audio (uncompressed), 17,000 hours of MP3s
- 1 PiB?
  - Enough data to store the DNA of the entire population of the US – three times!

☑

Garcia

(Cal) Clicker Question

What do you think is the biggest data overall?

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

Garcia

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!

Garcia

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
- This compression is reversible but has mathematically proven limits.
- Example:

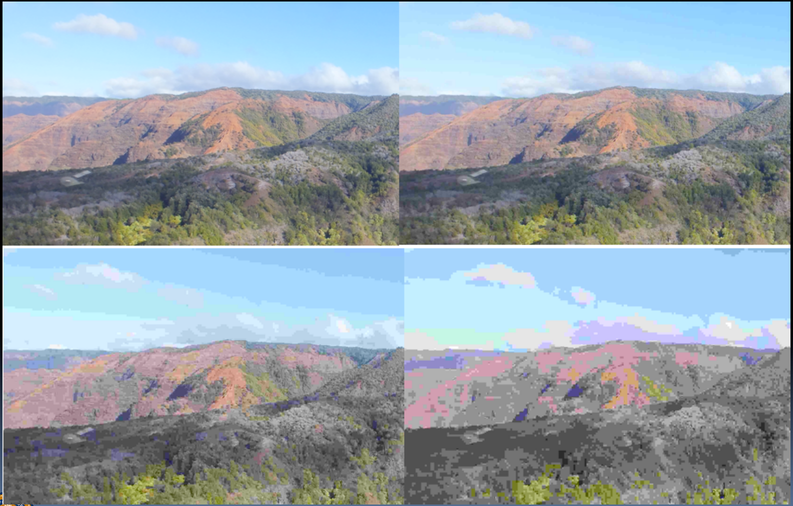
AAAAAABBBBBCCC -> 6A5B3C

Garcia

Helpful Techniques: 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"

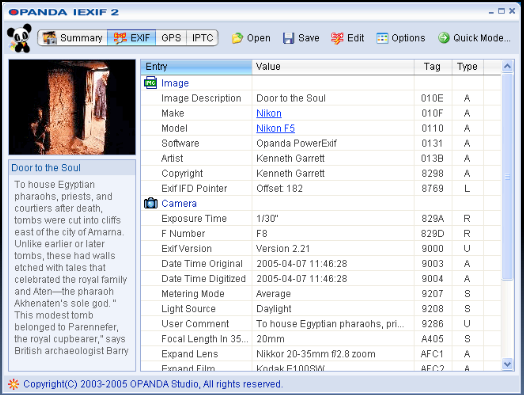
Lossy Compression Example: JPEG



Garcia

Techniques that help: Metadata

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



Door to the Soul

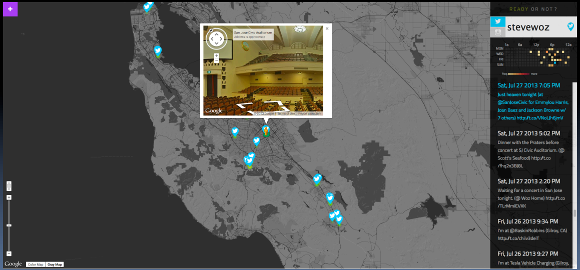
To house Egyptian pharaohs, priests, and courtiers after death, tombs were cut into cliffs east of the city of Amarna. Unlike earlier or later tombs, these had walls etched with tales that celebrated the royal family and Aten—the pharaoh Akhenaten's sole god." This modest tomb belonged to Parennefer, the royal cupbearer," says British archaeologist Barry

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



Sat, Jul 27 2013 7:05 PM
Just heaven tonight (at @SanJoseCivic for Emmylou Harris, Joan Baez and Jackson Browne w/ 7 others) http://t.co/VNoLJh6jmV

Sat, Jul 27 2013 5:02 PM
Dinner with the Praters before concert at SJ Civic Auditorium. (@ Scott's Seafood) http://t.co/fhq2x3BJ8L

Sat, Jul 27 2013 2:20 PM
Waiting for a concert in San Jose tonight. (@ Woz Home) http://t.co/TLrMmiEVXK

Fri, Jul 26 2013 9:34 PM
I'm at @BaskinRobbins (Gilroy, CA) http://t.co/chiiv3deiT

Fri, Jul 26 2013 9:27 PM
I'm at Tesla Vehicle Charging (Gilroy,

Garcia

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/

☑

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?

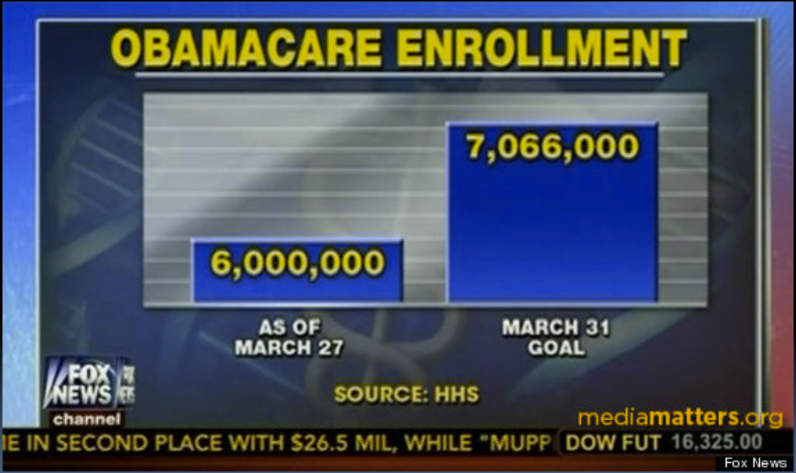
en.wikipedia.org/wiki/Correlation_does_not_imply_causation!

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., “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).

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

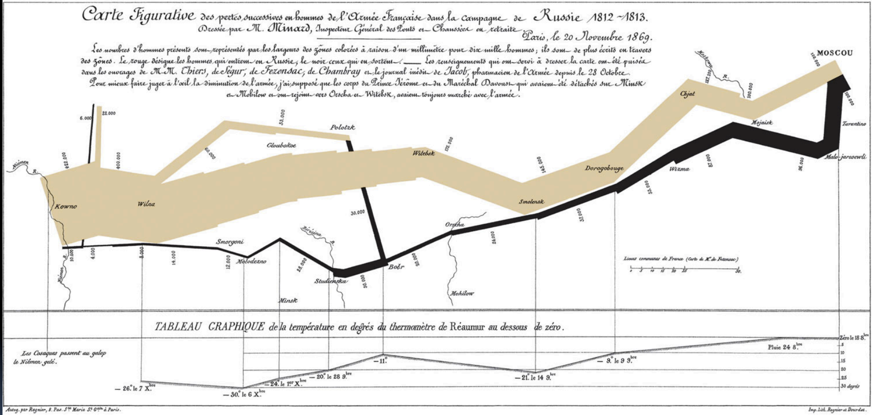
Visualization … Epic FAIL (2014)



Garcia

www.edwardtufte.com/tufte/posters

Visualization … Epic WIN (1869)



Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812–1813.

Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.

Paris, le 20 Novembre 1869.

Les nombres d'hommes présents sont représentés par les largeurs des zônes colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zônes. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. —— Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.

Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mobilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.

TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.

Les Cosaques passent au galop le Niémen gelé.

Charles Joseph Minard, Napoleon's 1812 Russian Campaign

Garcia

youtu.be/jbkSRLYSojo

Visualization … Epic WIN (2009)

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.