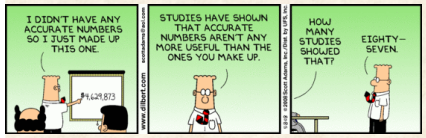
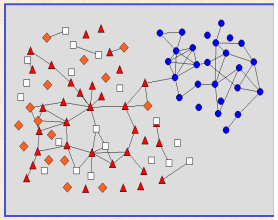
UNIVERSITY of WISCONSIN-MADISON
Computer Sciences Department

CS 202: Introduction to Computation — Professor Andrea Arpaci-Dusseau

How can computation… use data to solve problems?





Congrats to Game Winners…

Homework 4:
Kyle Junker
Jamie Jacobson
Quinn Else
Karl Foss
Brian Nelson

Will continue voting for Homework 5
- (Available Fri)

What topics have we covered in CS 202?

Part 1: Completed!

How do computers…?	Answer
Interact with humans?	Artificial intelligence
Solve problems?	Algorithms
Know what to do?	Programming languages
Make art?	Control flow: Sequential and Repeat
Show animated stories?	Flowcharts and Abstraction
Make decisions?	Decision Trees and If statements
Remember what has happened?	Variables
Avoid race conditions?	Critical sections
Educational software?	More variables and abstraction
Understand humans?	Natural language processing
Interact with humans?	Social robots
Guess what may happen?	Probability trials
Win games against you?	Game trees

Part 2: Where we are now…

How do computers…?	Answer
Solve societal problems?	Lots of data
Visualize data?	Lists
Find goal?	Optimization
Find stuff?	Search
Find stuff faster?	Binary search
Teach the world?	Digital StudyHall
Analyze text?	Histograms
Find web pages?	Search engines (Google)
Sort data?	Selection and insertion sort
Sort data faster?	Merge and quick sort
Predict the future?	Simulation
Share secrets?	Cryptography
Reach their limits?	P vs. NP

Part 3: End of Semester

How do computers…?	Answer
Work???	
Represent information and integers?	Bits and binary
Represent words, pictures, sound?	Encode and compress
Act logically?	Boolean logic, gates, and truth tables
Calculate?	Combinational circuits
Remember?	Memory
Execute instructions?	CPUs
Run multiple applications?	Virtualization (Operating systems)
Share memory well?	Caching
Communicate with others?	Networking
Use other languages?	Compilers
Tolerate faulty computers?	Logic with random information

Today's Exercise

What societal problem would you like to solve?

- What type of data would you need to collect?
- What would you look for in that data?

Example

- Concerned about car safety – reduce fatal accidents
- Data: Information about every accident
  - Location, time of day, fatalities, speed, cause
  - Mine data for commonalities (e.g., common locations)
- Set policy based on data (e.g., lower speed limit at those locations, more lights)

Brainstorm in groups of 4-5

New Data Sources

Scientific
- Sloan Digital Sky Survey
- Earth satellites
- Sensors: Temperature, earthquake
- GenBank Sequences

Medical
- Heartrate, blood pressure, temperature, attention?

Population statistics
- Surveys: Income, children, location

Entertainment and Sports

Business
- Stock trades
- Wal-Mart transactions

Social media
- Facebook, twitter, google searches
- Cell phone locations

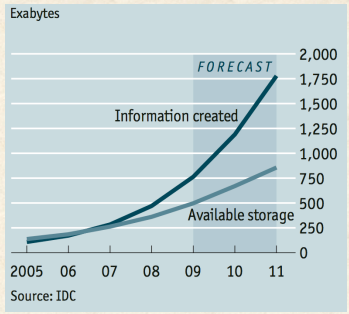
Digital media
- Digitized books, photos, texts in multiple languages

How much data?

What is an Exabyte?

- bit (b): 0/1
- byte (B): 8 bits
- Kilobyte (KB): 1000 bytes
- Megabyte (MB): 1000 KB
- Gigabyte (GB): 1000 MB
- Terabyte (TB): 1000 GB
  - 10 TB: Library of Congress
- Petabyte (PB): 1000 TB
  - 500 billion pages of text
- Exabtye (EB): 1000 PB
  - 5 EB: all words ever spoken
- Zettabyte (ZB): 1000 EB

Data produced at phenomenal rates



T. Cukier (2010). 'Data, data, everywhere', *The Economist*.

For data to solve problems, what do we need to do?

Capture
- Collect or obtain

Store
- Where? what if it gets lost?

Share
- Who should have access? Anonymize for privacy

Organize
- Sort it? put in categories?

Process it
- Search through it, analyze it, mine it for correlations

Visualize it
- What is useful for humans to look at?

Use results to inform our decision

Today: Two Examples

What can we learn from digitized books?

What can we learn from cell phone data?

Example 1: Digitized Books

What if you had access to every book ever written?

What questions could you answer?

http://www.ted.com/talks/what_we_learned_from_5_million_books.html

Experiment!

http://books.google.com/ngrams

Example 2: Mobile Phone Data

What raw data can we get from phone calls?

Why could it be useful to gather?

Location

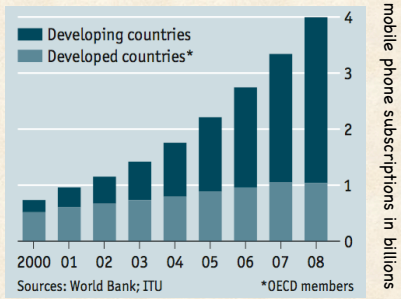
- Individual: Know where lost individuals are
- Big data: Look for anomalies (emergencies, riots); traffic planning?

Connections

- Infer relationships between people
- Individual: Insider-trading, exchanging secrets
- Big data: Contagious diseases

Big Data in the Developing World

Majority of 'behavioral data' comes from the developing world



T. Standage (2009).'Mobile Marvels', *The Economist*.

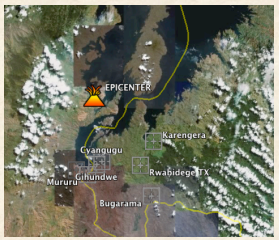
Challenge Question 1

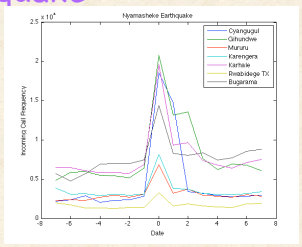
How could you use cell phone data to determine that something bad happened?

1) Earthquake
2) Epidemic

Assume don't have mechanism to collect data in other ways

Reactions to Regional Shocks: Earthquake



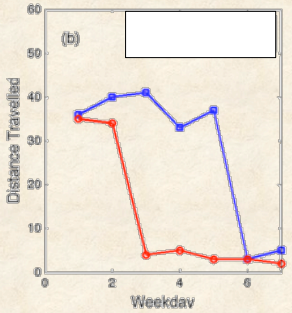


Call volume increases dramatically after event
Other (non-phone) data that indicates events?

A. Kapoor, N. Eagle and E. Horvitz (2010), "People, Quakes, and Communications: Inferences from Call Dynamics about a Seismic Event and its Influences on a Population", Proceedings of AAAI Artificial Intelligence for Development (AI-D'10).

Reactions to Regional Shocks: Epidemic

What happens to locations when epidemic is underway?



What is blue line?
What is red line?

Distances can be used for Disease Surveillance

Nathan Eagle, Leon Danon

Challenge Question 2

- How could you use cell phone data to predict how disease will spread?
  - RSV through community?
  - 50 Households in Kilfi, Kenya – Phones
- Can we link proximity patterns with infection?
- First goal: Can we use cell phone data to predict who are "friends"

Moses Kiti, Nathan Eagle, James Nokes.

Study Students Instead…

63 people over 9 months

Label at home or work

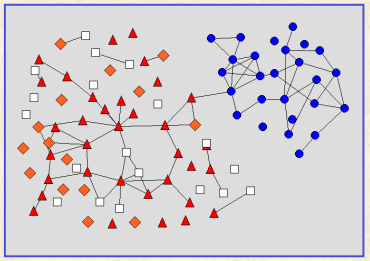
Eagle, N. "Machine Perception and Learning of Complex Social Systems", *PhD Thesis, Massachusetts Institute of Technology.* 2005.

Who are Friends?

What is most straight-forward way to find out if two people are friends or not?

Survey Them…

Grads, Undergrads



Self-Report Friendship

Might not be reciprocal, but ignore that!

Predict friendship from data

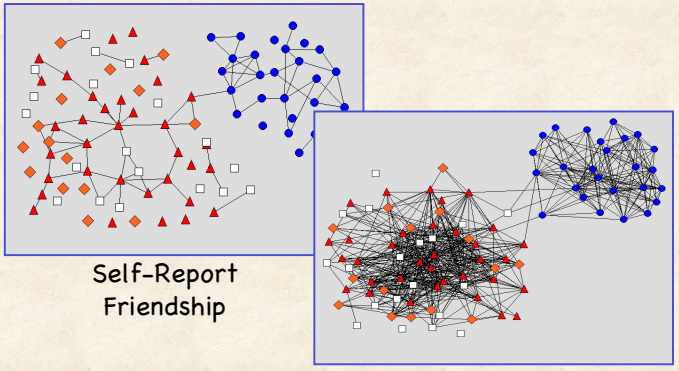
Can't always just ask people

- Tedious data collection
  - What happens when have millions of people???
- Incentives to lie

What other metrics could you use to predict friendship?

Near each other on some day?
Call each other frequently?

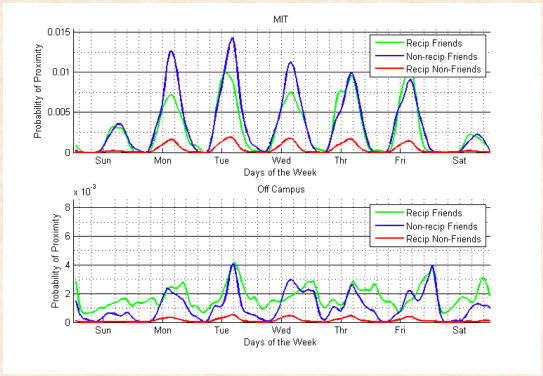
Is this a good predictor?



Self-Report Friendship

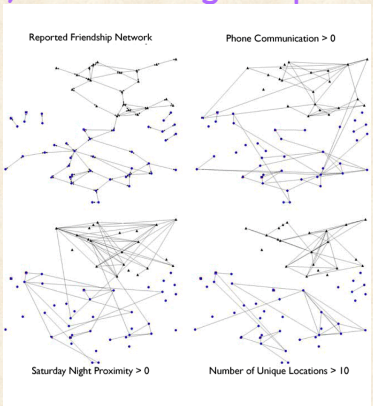
1-Day Proximity

Friendship Inference: Time Matters!



N. Eagle, A. Pentland, and D. Lazer. (2009) "Inferring Social Network Structure using Mobile Phone Data", *Proceedings of the National Academy of Sciences (PNAS)*, September 8, 2009 vol. 106 no. 36, 15274-15278.

Are any of these good predictors?



N. Eagle, A. Clauset, A. Pentland, and D. Lazer (2010), "Multi-Dimensional Edge Inference", *Proceedings of the National Academy of Sciences (PNAS)*, 107(9), pp. E31.

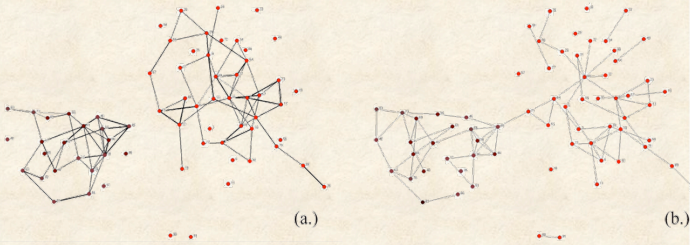
Are any of these good predictors?

	Friends		Not Friends	
	avg	std	avg	std
Total Proximity (minutes / day)	**72**	150	**9.5**	36
Saturday Night Proximity (minutes / week)	**7.3**	18	**.20**	1.7
Proximity with no Signal (minutes / day)	**12**	20	**2.9**	20
Total Number of Towers Together	**20**	36	**3.5**	4.4
Proximity at Home (minutes / day)	**3.7**	8.4	**.32**	2.2
Phone Calls / day	**.11**	.27	**.001**	.017

N. Eagle, A. Clauset, A. Pentland, and D. Lazer (2010), "Multi-Dimensional Edge Inference", *Proceedings of the National Academy of Sciences (PNAS)*, 107(9), pp. E31.

Inferred v. Reported Network

Combine metrics as well as we can; Can we predict who are friends?

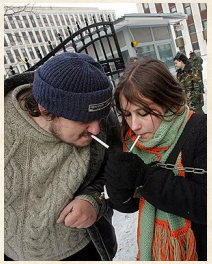


Somewhat!

N. Eagle, A. Pentland, and D. Lazer. (2009) "Inferring Social Network Structure using Mobile Phone Data", *Proceedings of the National Academy of Sciences (PNAS)*, September 8, 2009 vol. 106 no. 36, 15274-15278.

Big Data from Small Samples: 150 Smokers in New York City

- 150 Undergraduate Smokers / Recent Quitters
  - Location, Bluetooth, Communication, Context-Driven Surveys
- Question: Is there a behavioral signature associated with relapse?



Yuelin Li and Nathan Eagle.

Conclusions

Huge amounts of data being collected

- Technical challenges to storing it (3rd part of course)
- This part of course:
  - How to visualize it?
  - How to search through it?
  - How to organize it (sort it)?
  - How to find web pages?
  - How to keep information private?
  - Which algorithms best handle huge amounts of data?

Up to you all to decide how information should be used!

Announcements

HW 4 graded

- Send email if any problems

HW 5 available Friday (due 1 week from Fri)

Exam 1 Returned Friday