Big Data?
Big Promise, Big Problems

Jeanna Matthews (Clarkson University)
jnm@clarkson.edu

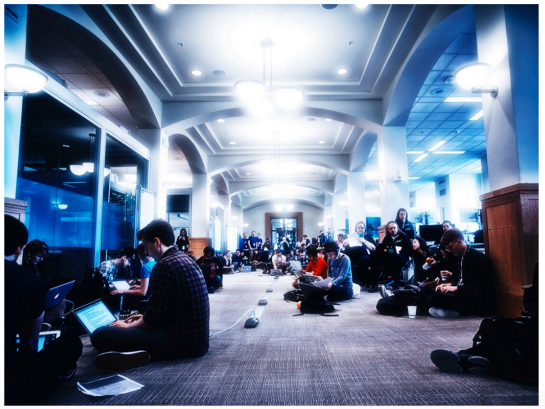
CNY Hackathon, Utica College
April 21 2017

Power of Big Data

- So many sources of data
- Understand/optimize ourselves and the world
- Connect with each other

MEGAN MOLTENI SCIENCE 02.13.17 5:35 PM

DIEHARD CODERS JUST RESCUED NASA'S EARTH SCIENCE DATA



Human Face of Big Data

Big Data


EACH OF US NOW LEAVES A TRAIL OF DIGITAL EXHAUST, AN INFINITE STREAM OF PHONE RECORDS, TEXTS, BROWSER HISTORIES, GPS DATA, AND OTHER INFORMATION, THAT WILL LIVE ON FOREVER.
Instead of "find my iPhone," some auto insurance companies are offering a service that may enable parents to "find my teenager." Progressive Insurance, for example, offers the Snapshot, a tracking device that reports on a car's location, acceleration, braking, and distance traveled. Owners who install the device can get a 10 to 15 percent discount on their policy. Privacy activists, however, fear the technology is ripe for abuse.

Our Digital Exhaust

- Emails, texts
- Social media
- Web browsing history, web site use and cross site correlations
- Cell phone location
- Purchase history, credit cards, wish lists, products viewed/ reviewed, frequent buyer cards,
- Cameras (yours, others, on street, accidental, aware/ unaware, facial recognition) , GPS tags in pictures
- Fitbits, microphones, Google glass,
- License plate readers, passport use, radio-frequency identification (RFID) readers, satellite imagery
- E-readers, streaming video use, MOOCs,

- We are voluntarily and enthusiastically making ourselves supremely trackable!
  - Often involve data collection through tracking devices or surveillance
  - Often voluntarily revealed info that takes advantage of our desire to connect with others
- Perfect storm
  - Cheap disk space
  - Fast and readily available network access
  - New types of tracking devices
  - Willingness of people to carry/use tracking devices
  - Strong human desire to connect and share
  - Disinterest in privacy

Wearable devices

- Fitbit, Fuelband, Jawbone Up
- Activity patterns
- Sleep patterns
- Some heart rate
- Some location

- Often must put data in cloud to even view it

- Promise? Example: More data about sleep patterns than ever before in history
- Problem? Two people in same place with high heart rate?

MOOCs

- Massively Open Online Courses
- How many times did you take a test or try an exercise?
- How long did you spend reading? Did you read every page? Where did you linger?

- Promise? Targeted experiments about best way to teach on a massive scale – more data than ever before!
- Problem? Sell to employers? Who is good at this task? Reputation sharing whether or not you participate

Spread of this data

- Correlation of different pools of data
  - Shared with identifying data
    - Sold? Shared with trusted partners? Businesses/ government
  - Shared "anonymously" but possible to deanomymize
  - Stolen by hackers
  - Insider misuse
  - ….

Data sharing/release

- Data you know you are revealing to support a specific goal
  - Orders placed with Amazon, credit card charges, emails sent, photos you posted
- Data you do not know you are revealing or don't believe is being recorded
  - GPS tagging on photos, last page accessed, pattern of reading in ereader, private conversations on a subway, picture that happened to catch you, facial recognition
- Recognize when they see you again
  - RFID, web browsing fingerprint, recognize text sample A is from same person who wrote sample B
- Predictions from this data
  - You didn't say something specifically but I deduce it.. Am I right?
  - Like Facebook page for a disease -> you have that disease
  - Buy prenatal vitamins or heart rate up on Fitbit -> you are pregnant

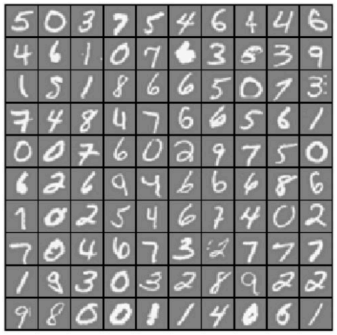
Used increasingly for big decisions

- We've all seen this used for advertising
  - Search for something on Google and see ads on Facebook and Amazon
  - Targeted advertising of housing? Jobs?
- Fed into machine learning algorithms
  - Predictions – credit decisions, bail, college admissions, housing decisions, allocation of public resources….
  - Look at potentially discriminatory attributes or infer them?
- What else is it used for?

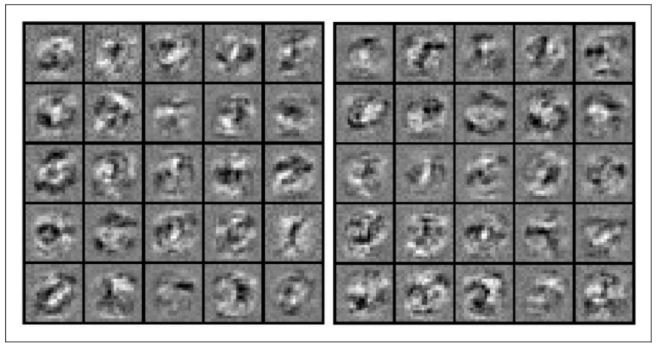
Algorithmic Transparency/ Accountability

- Big data used for big decisions
  - Are you offered a job?
  - Can you buy a house?
  - Will there be more police surveillance in your neighborhood?
- How are these decisions made?
  - Top 20 job candidates how chosen?
  - Right to understand biases built in by programmers or more likely biases learned from historical data

Black Box Decision Making



- Machine learning on digits
- Machine learning on dogs vs. wolves
- Machine learning on your digital exhaust?
  - Credit card charge for marriage counseling => raise interest rates, lower credit limit
  - Credit card charge? Facebook like? Colleague on Linked In? Happen to be near a demonstration site?



(a) Husky classified as wolf (b) Explanation

Figures from “How the Machine 'Thinks:' Understanding Opacity in Machine Learning Algorithms”, Burrell and ““Why Should I Trust You?": Explaining the Predictions of Any Classifier”, Ribeiro et al.

Anonymity in Big Data

- Once your data is collected, one thing is use by the collector directly
  - Or when the collector sells it or trades it with others in a non-anonymized form
- But also common is that it can be released in some "anonymized" way for research purposes
  - Many times with very good intent
- But does anonymization work?
- Lets look at some examples over time

AOL Search Data (2006)

- AOL released users' search queries without their permission
  - AOL username changed to a random ID number
  - Ability to analyze all searches by a single user often allows identification of the user
    - The data includes personal names, addresses, social security numbers and everything else someone might type into a search box.
    - Search for your name (vanity search), phone numbers, etc.
  - Researchers contacted people who confirmed that their search query set has been successfully deanonymized

Netflix Recommendation Data (2007)

- 10 million movie rankings by 500,000 customers published as part of a challenge for people to come up with better recommendation system
- Anonymized by removing personal details and replacing names with random numbers
- Researchers correlated with data from IMDB where people had published their rankings of movies
  - Rankings of enough movies publicly served as fingerprint into Netflix data – revealing recommendations they did *not* make public

Massachusetts Governor's Health Records (2002)

- Researcher Sweeney correlated medical data with voting records
  - Group Insurance Commission (GIC) medical data given to researchers and sold to industry
  - Voting records public
  - William Weld (governor of Massachusetts) uniquely identified by birth date, gender and 5-digit zip code
- Estimated that 87% of the US population uniquely identified by this same combination

Ok lets try that again

- Same researcher, Sweeney, proposes k-anonymity
  - Hide in a crowd of at least k for any answer
- Other researchers poke holes in k-anonymity, propose l-diversity
  - E.g. What if everyone in crowd has same sensitive attribute or most do
- Other researchers poke holes in l-diversity, propose t-closeness
- Each step does more to obscure the data and still data can be successfully deanonymized

When you feel anonymous, beware

- Anonymous blog post?
  - Other researchers claim they can identify many characteristics of writers (e.g. gender, place of birth etc) from anonymous text and even uniquely match writer based on stylistic fingerprints (stylometry)
- Contribution of open source code?
  - Same for stylistic fingerprints in source code
- Surfing the web?
  - Browser configuration revealed in web request often unique fingerprint (even without cookies)
  - Panopticlick.eff.org
- Carrying a passport or enhanced license?
  - RFID readable at a distance
- Checking out someone's public Facebook profile?

Need for privacy?

- Need space for dissent and discussion
- Need space to try and develop ideas
- Need space to grow and change
- Need space to allow discussion and consideration of the idea you hate the most
- May need to be able to keep private what others may use to prejudge you
- "The clash of ideas is the sound of freedom." Lady Bird Johnson
- Live in a world where everyone knows everything about everyone?

What else can you do?

- I don't recommend living paranoid!
- Instead
  - Turn off tracking when you don't really want or need it
  - Ask why/challenge data collection requests
  - Lobby for laws that give individual's rights to inspect/correct/delete their data
  - Use/support services with stronger privacy/anonymity properties
  - Request interfaces that allow you to view your data without contributing it to "cloud"
  - Hold governments and companies accountable for their data collection and use
  - Push for transparency /symmetry/explainability/accountability
  - Models for successful data sharing

Can we have the best of both worlds?

- Use detailed personal information to connect with others and help people make better choices without potential for discrimination and oppression?
  - Accountability and transparency in decision making?
- The ability to learn high level information without drilling down to individual's details?
  - Trust who is holding the data? No anonymous release?

Two recommend reads

Data and Goliath, Schneier

Weapons of Math Destruction, O'Neil