CSCI 4550/6550 Interactive Visualization

https://www.cs.rpi.edu/~cutler/classes/visualization/S24/

Lecture 12: Ethics & Privacy

Today

- Ethics for Journalism and Visualization
- Readings for Today
  - "Ethical Dimensions of Visualization Research"
  - “How Deceptive are Deceptive Visualizations?”
  - "Guess Me If You Can: A Visual Uncertainty Model for Transparent Evaluation of Disclosure Risks in Privacy-Preserving Data Visualization"
- Related Papers
  - "Adaptive Privacy-Preserving Visualization Using Parallel Coordinates”
  - "Agile Ethics for Massified Research and Visualization”
- Informed Consent, FERPA, HIPPA, & Plagiarism Research
- Readings for Tuesday

Society of Professional Journalists' Code of Ethics

- Fact check, cite sources, disclose conflicts of interest
- Sources: question motives, anonymous if appropriate, avoid reimbursement for news
- Avoid stereotypes, misrepresentation, distortion, re-enactments, surreptitious/undercover work, pressure from advertisers & special interests
- Be compassionate, sensitive, invite dialog, admit mistakes
- *Today's Reality: Pressure of 24 hour cable news competition, accusations of "Fake News", clickbait, ...*

http://visual.ly/about/ethics

As an organization that both practices and recognizes quality data-journalism, Visual.ly subscribes to the code of ethics of the Society of Professional Journalists and agrees to abide by all of its principles.

We also agree to the following principles to support data analysis and visualization:

Data will be accurate and verifiable - Visual.ly will not "lie with statistics."

Proper Sourcing & Attribution - Visual.ly will always give credit where due and will do its own reporting.

Best Practices in Visual Representation - Visual.ly will not exploit idiosyncrasies of the human visual system to exaggerate or misrepresent data.

Most succinctly stated, Visual.ly's policy is one that encompasses accuracy, honesty, and transparency.

While Visual.ly will do our best to promote these standards, the policy applies only to the visualizations we create ourselves and those we feature as staff picks, not to those uploaded by members of the community.

Visual.ly's Code of Ethics for Data Visualization Professionals

- Data analysis is important
- Clearly state assumptions
- Too narrow focus or omission can lead to bias
- Incorrect analysis must be avoided
- Be open to criticism, learn from past work
- Be aware of colorblindness, cultural meaning
- Be open to criticism

A Hippocratic Oath

Jason Moore also suggested a hippocratic oath for visualization. This version is slightly edited from his original, and I guess some more work could be done on it. But I think it's a great start.

> *I shall not use visualization to intentionally hide or confuse the truth which it is intended to portray. I will respect the great power visualization has in garnering wisdom and misleading the uninformed. I accept this responsibility willfully and without reservation, and promise to defend this oath against all enemies, both domestic and foreign.*

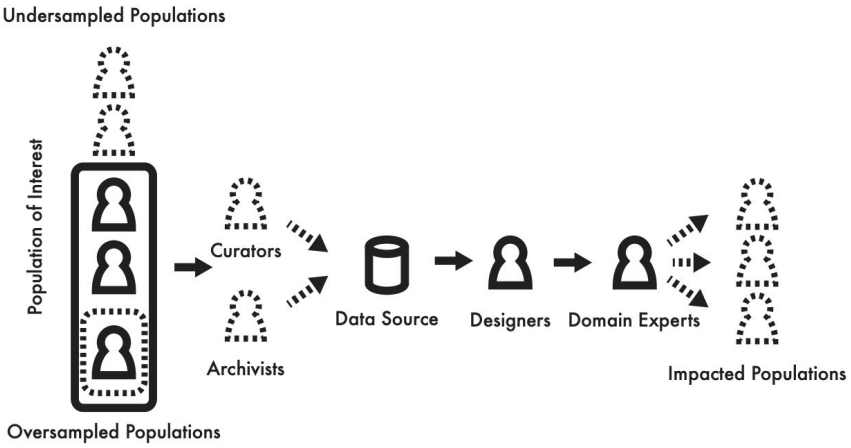
https://eagereyes.org/blog/2011/visualization-is-growing-up

Today

- Ethics for Journalism and Visualization
- Readings for Today
  - "Ethical Dimensions of Visualization Research"
  - “How Deceptive are Deceptive Visualizations?”
  - "Guess Me If You Can: A Visual Uncertainty Model for Transparent Evaluation of Disclosure Risks in Privacy-Preserving Data Visualization"
- Related Papers
  - "Adaptive Privacy-Preserving Visualization Using Parallel Coordinates”
  - "Agile Ethics for Massified Research and Visualization”
- Informed Consent, FERPA, HIPPA, & Plagiarism Research
- Readings for Tuesday

Reading for Today

- "Ethical Dimensions of Visualization Research", Michael Correll, ACM CHI 2019



- It is naive to claim that “data is neutral”
- Collecting data has political/ethical consequences
- Refraining from collecting data has political/ethical consequences
- Biases introduced by who collects and processes the data
- Bad if viewer assumes visualization is truth and does not critically think - Implying that you cannot disagree/argue with a visualization
- Lack of tools, lack of attempt to express data uncertainty
- Visualization can’t express empathy or acknowledge suffering
- Power imbalance, responsibility, conflict, propaganda
- Data feminism - recognition that data is biased, unequal power balance
- P-hacking - misuse of statistics to find patterns that appear significant but are not. From the P value test of statistical significance - how likely is it that the observed difference due to chance? Statistics is complicated, systems that automate analysis can be dangerous if they are promoting noise as unjustified conclusion
- Obligation to reveal decision making by machine learning (AI explainability) - provide transparency to otherwise opaque algorithmic decision-making - to empower those impacted by AI decision making
- Garden of forking paths
- Visualization of provenance (source) of data

- Visualize hidden labor (acknowledge all contributions, especially marginalized communities)
- Visualize hidden uncertainty (allow viewers to accurately assess risk of hurricane, etc.)
- Visualize hidden impacts (document and discuss the potential for harm or misuse of the new technology in an academic paper)

- Encourage “small data” – bigger data is not always better – scooping up more and more data with the promise of better service increases risk of privacy breaches
- Anthropomorphize data - include glyphs or images of people in visualization
- Obfuscate data to protect privacy - aggregating, fuzzing, restructuring

- Challenge structures of power - support data “due process”
- Right to explanation, right to privacy, right to be forgotten
- Be data advocates
- Instead of always using same boring standard dataset to demonstrate new visualization technique, contribute your expertise to help curate a new dataset and visualize something unseen but important - help combat injustices
- Don’t facilitate unethical analytical behavior - speak up about flaws, refuse to participate in questionable projects

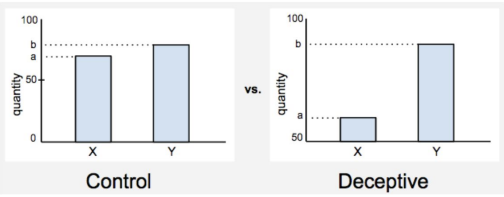
- Paper has LOTS of citations - difficult to read as first-time exposure to topic

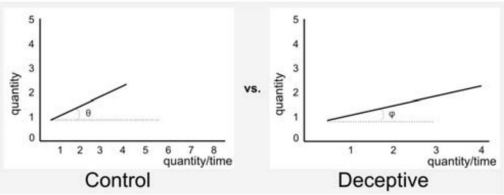
Today

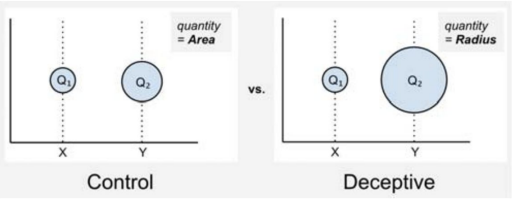
- Ethics for Journalism and Visualization
- Readings for Today
  - "Ethical Dimensions of Visualization Research"
  - “How Deceptive are Deceptive Visualizations?”
  - "Guess Me If You Can: A Visual Uncertainty Model for Transparent Evaluation of Disclosure Risks in Privacy-Preserving Data Visualization"
- Related Papers
  - "Adaptive Privacy-Preserving Visualization Using Parallel Coordinates”
  - "Agile Ethics for Massified Research and Visualization”
- Informed Consent, FERPA, HIPPA, & Plagiarism Research
- Readings for Tuesday

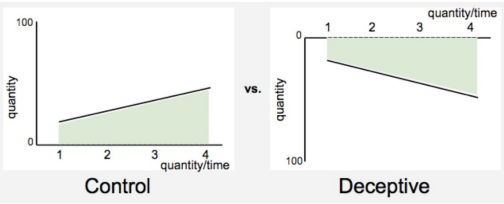
Reading for Today

- “How Deceptive are Deceptive Visualizations?”, Pandey, Rall, Satterthwaite, Nov, & Bertini, ACM CHI 2015









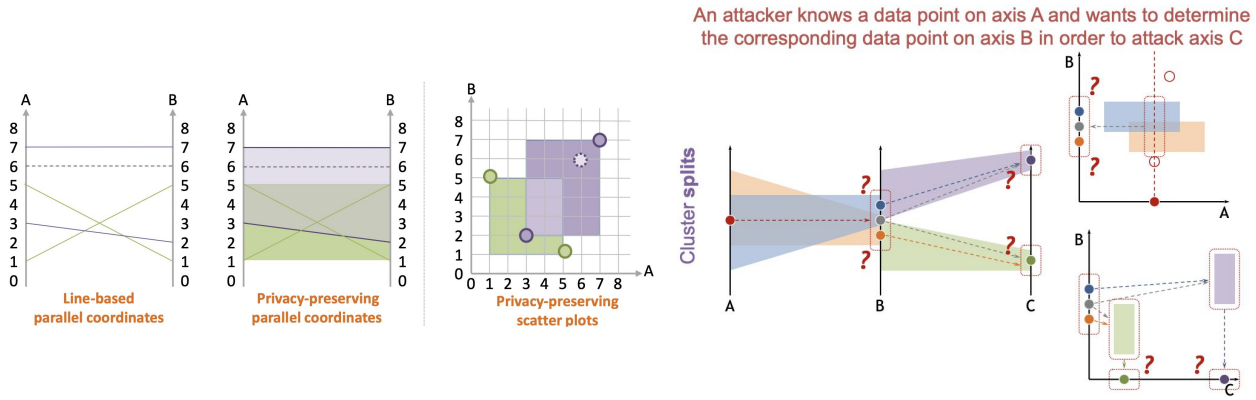
- Misrepresentations/distortions might come from lack of expertise, time constraints, or perhaps a malicious intention.
- Deceptive visualization can be: message reversal or message exaggeration/understatement
- Deception can be chart level (inability to read chart correctly - visual encoding flaws) or message level (incorrect interpretation - false beliefs)
- Enumerate common forms of deception/distortion in graphs (also done in prior work)
  - truncated axis, area as quantity, aspect ratio, and inverted axis
- Designed a user study to measure “message level” effectiveness/accuracy of visualization
- Does “deception” require intent?
  Is lack of knowledge of tools and conventions sufficient excuse?
  This paper does not explore or study the visualization creators (lack of) intent to deceive.
- Amazon Mechanical Turk - diverse subject pool, 330 participants total
  - Attention check (filter out random clickers)
- Probably need a test for “graphical literacy”

Today

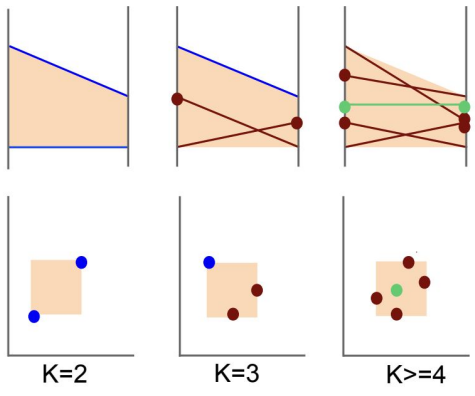
- Ethics for Journalism and Visualization
- Readings for Today
  - "Ethical Dimensions of Visualization Research"
  - “How Deceptive are Deceptive Visualizations?”
  - "Guess Me If You Can: A Visual Uncertainty Model for Transparent Evaluation of Disclosure Risks in Privacy-Preserving Data Visualization"
- Related Papers
  - "Adaptive Privacy-Preserving Visualization Using Parallel Coordinates”
  - "Agile Ethics for Massified Research and Visualization”
- Informed Consent, FERPA, HIPPA, & Plagiarism Research
- Readings for Tuesday

Reading for Today

- "Guess Me If You Can: A Visual Uncertainty Model for Transparent Evaluation of Disclosure Risks in Privacy-Preserving Data Visualization", Dasgupta, Kosara, & Chen, IEEE VizSec 2019



- Open data - risk of unintended disclosure of private information
- Post-anonymization risk of re-identification
  - either of specific info (attribute disclosure) or
  - ID of individual (identity disclosure)
- Prosecutor re-identification: knows an individual is in the database, wants to find out specific attributes
- Journalistic re-identification: wants to uncover the real ID of any individual – to prove breach is possible
- K anonymity: Each person's privacy/ID is protected. There are at least k-1 other people that have the same data attributes. Cluster is at least k people.
- Small clusters more vulnerable to privacy breach attack - fewer geometric configurations, possible/likely to be corner/edge/diagonal
- K-anonymity is necessary but not sufficient condition for privacy preservation
- Adversaries will re-order parallel coordinate axes to try to breach privacy



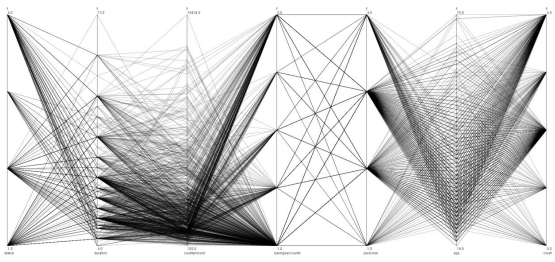
Today

- Ethics for Journalism and Visualization
- Readings for Today
  - "Ethical Dimensions of Visualization Research"
  - “How Deceptive are Deceptive Visualizations?”
  - "Guess Me If You Can: A Visual Uncertainty Model for Transparent Evaluation of Disclosure Risks in Privacy-Preserving Data Visualization"
- Related Papers
  - "Adaptive Privacy-Preserving Visualization Using Parallel Coordinates”
  - "Agile Ethics for Massified Research and Visualization”
- Informed Consent, FERPA, HIPPA, & Plagiarism Research
- Readings for Tuesday

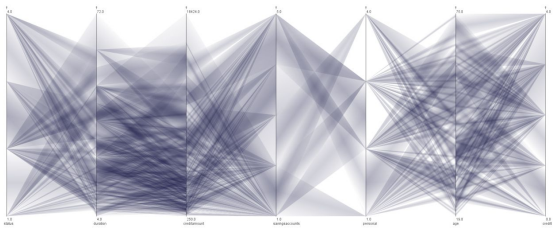
"Adaptive Privacy-Preserving Visualization Using Parallel Coordinates”, Dasgupta & Kosara, TVCG 2011

- Different purposes for “Visualization”:
  - Convey information or art? Is misleading necessarily bad when visualizing for art purposes? Context matters.
  - Find the appropriate balance between art (eye catching & memorable) and scientific accuracy & honest data presentation.
- Some visualization techniques are already quite lossy (e.g., pie chart) or just confusing (e.g., parallel coordinates?) and obfuscation may not be necessary
  - Making false connections
  - Limit visualization to 500 pixels to thwart attacks(!?)
  - Rounding pixel coordinates
- Impractical in the real world?
- So much effort to make a visualization *less useful.*
  (Why make the visualization at all? Who is the target audience of this visualization?)
- Need to investigate reliability of data before reporting it
- K-anonymity & l-diversity
- (Trusted) server does clustering & only sends clusters to (untrusted) client
- Writing - Diagrams & charts not well explained
  - Nice to read about the process and ideas for techniques that ultimately didn’t work
  - Would have been better written/easier to read if they had stated upfront what they wanted to accomplish

"Adaptive Privacy-Preserving Visualization Using Parallel Coordinates", Dasgupta & Kosara, TVCG 2011



(a) Default parallel coordinates view of the German credit dataset



(b) Clustered view with $k = 3$

- Could be further secured:
  - Don't place unclustered data on the public facing machine
  - Don't ever re-cluster the data (prevent clustering attacks)
- Client-server model: Computation & networking costs?
- How well does this work for small datasets?
- Hadn't considered visualization as a means of data breach, reassuring that people are researching the problem.
- How much accuracy in analysis do we lose?
  - Obscuring & blurring data (opposite of our usual focus on clarity & accuracy!)
- What about more sophisticated attacks? …?
  - Maybe this isn't 100% secure, but it's important to do something!
- Would like a user study comparing their Privacy-Preserving Parallel Coordinates to original full data Parallel Coordinates.
- Quasi-Identifiers (is an anonymous form really anonymous?)
- Privacy policies that declare how they can and will share your personal data…

Today

- Ethics for Journalism and Visualization
- Readings for Today
  - "Ethical Dimensions of Visualization Research"
  - “How Deceptive are Deceptive Visualizations?”
  - "Guess Me If You Can: A Visual Uncertainty Model for Transparent Evaluation of Disclosure Risks in Privacy-Preserving Data Visualization"
- Related Papers
  - "Adaptive Privacy-Preserving Visualization Using Parallel Coordinates”
  - "Agile Ethics for Massified Research and Visualization”
- Informed Consent, FERPA, HIPPA, & Plagiarism Research
- Readings for Tuesday

"Agile Ethics for Massified Research and Visualization”, Neuhaus & Webmoor, Information, Communication & Society 2012.

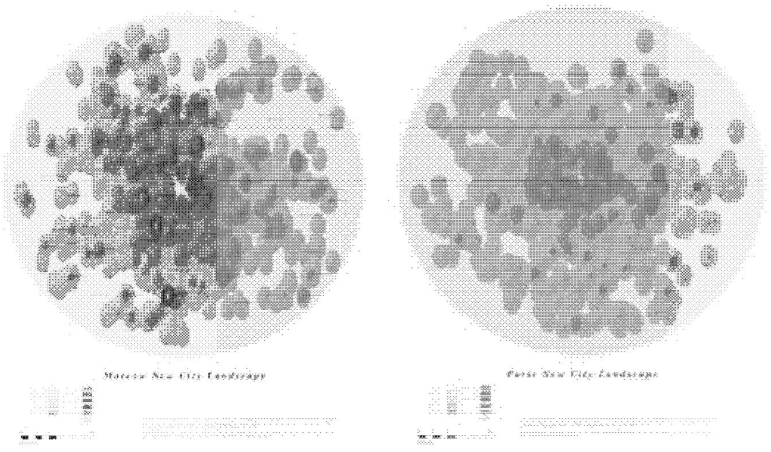


FIGURE 2 Two NCL maps as examples of aggregating potentially identifying information from Twitter. On the left is Moscow with a number of very active locations. Paris on the right exhibits a central main core of activity.

"Agile Ethics for Massified Research and Visualization", Neuhaus & Webmoor, Information, Communication & Society 2012.

- Privacy, confidentiality, anonymity, and informed consent
  - Minimize risk to participants
  - Observational study does not require consent
  - (current) IRB process does not/cannot work at massive scale
  - A single person can more easily do what used to take a team of researchers much more time.
- Agile ethics: too high level to be enforceable?
- Even though individuals made this data available, it is researchers responsibility to not put them in danger
- Not showing individuals, only general trends
- If the research team's information is similarly public/vulnerable, its ok?
- Geometry for twitter
- Public-private greyscale
- "Now that you've read this… Just be more careful, ok?"
- Writing
  - Confusing paper organization
  - Low resolution images
  - Unconventional acronyms
  - Footnotes at end of paper (prefer at end of each page)

- Jargon-y
- Unnecessarily lengthy?
- Websurfing is dangerous
- Twitter is scary (amount of personal data available surprising)
  - Users can (now?) disable location tracking
  - Are Twitter "protected" accounts new?
  - Read the fine print before sharing your data!
- TimeRose visualization is new to me
- Scattered topics
- Fitness tracking applications offer to post your morning run on facebook are dangerous
- Proposed solutions are too idealistic?

- Internet is an ocean of data
- Research results poured back into ocean of data
- Surveillance: Shopping malls are private spaces, but made to feel like public spaces
- File/log planned data collections in advance (pre-planning required, data more precious)
- Researchers should make themselves equally public

Today

- Ethics for Journalism and Visualization
- Readings for Today
  - "Ethical Dimensions of Visualization Research"
  - “How Deceptive are Deceptive Visualizations?”
  - "Guess Me If You Can: A Visual Uncertainty Model for Transparent Evaluation of Disclosure Risks in Privacy-Preserving Data Visualization"
- Related Papers
  - "Adaptive Privacy-Preserving Visualization Using Parallel Coordinates”
  - "Agile Ethics for Massified Research and Visualization”
- Informed Consent, FERPA, HIPPA, & Plagiarism Research
- Readings for Tuesday

Informed Consent

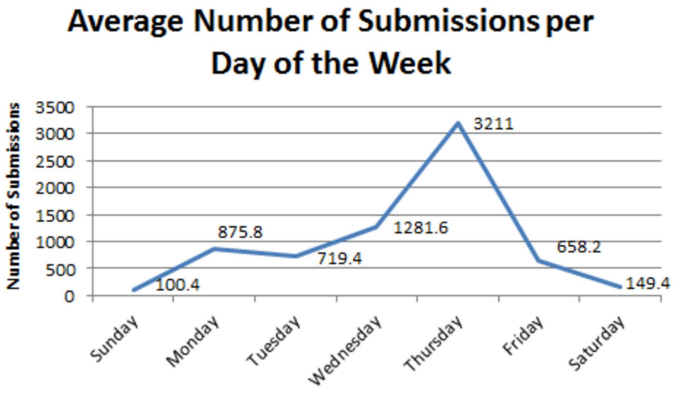
- Do you carefully read every document you sign?
  Every “agree to terms” button you click?

- Data can be taken out of context, used in ways other than intended

- Previously: required a team of researchers to gather data
- Now: a single person can do it alone -
  We’ve lost informal peer consultation of ethical concerns

Potential to Use Submitty Data for Research?

- Does an individual student's grade rise over time (repeated submissions)?
- Do students who start submitting earlier in the week have a higher final grade for that homework?
- Do students who submit more often get higher grades?



How to anonymize? Which are sensitive attributes?

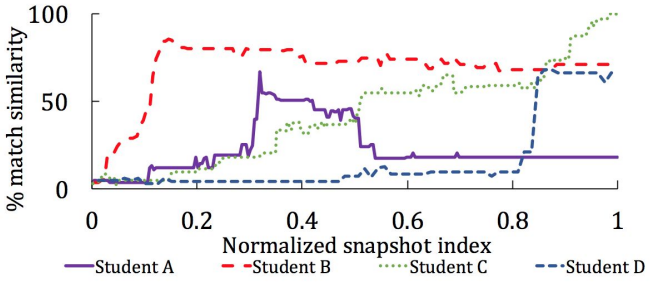
- Grades, plagiarism, sleep patterns, who took the class, time/# of attempts needed to complete assignment
- Remove explicit identifiers (username, RIN)
- Small datasets cause problems
- Quasi identifiers
- Sanitize data on the fly, constraining the interaction (don't allow inspection of complete historical details of single student -- even if the name has been removed)
- Assume data holder is trusted and aware of data sensitivity *(and appropriately concerned)*

FERPA - The Family Educational Rights and Privacy Act

- Students/parents can inspect & review information in their educational records
- Students/parents can request a correction to their record.
- Schools may disclose, without consent, "directory" information
  - @RPI: name, address, photographs, phone #, e-mail, date/location of birth, major field of study, academic load, participation in officially recognized activities and sports, weight and height of members of athletic teams, dates of attendance, degrees,
    honors and awards received, class year in school, and most recent previous educational institution attended
- *However, schools must allow students/parents to opt out of directory information disclosure*
- Students/parents must be regularly informed about their rights

- Submitty stores all of your submissions
  - It's your choice when & what to submit
- What if we asked you to install a plugin for your IDE that:
  - Captured your files after every save? every keystroke?
  - Watched what other programs were used simultaneously?
  - Saved your physical location & who you were with
  - Spied through your camera/microphone?

This is creepy, we have no intention of doing this!



"TMOSS: Using Intermediate Assignment Work to Understand Excessive Collaboration in Large Classes", Yan et al, SIGCSE 2018

Privacy & Visualization

- Most visualization computation *assumes* unrestricted access to data
- How do we do this computation with partial information?
- How do you design hardware/software system to ensure data security?

Privacy & Visualization

- Who would potentially benefit from access to this data? (Why is this a grey area?)
  - Scientific discovery
  - Improve healthcare
  - Improve education
- What data has privacy concerns?
  - Corporate secrets
  - Health records
  - Academic records
  - Personal finances
  - Personal location

Health Insurance Portability and Accountability Act (HIPAA)

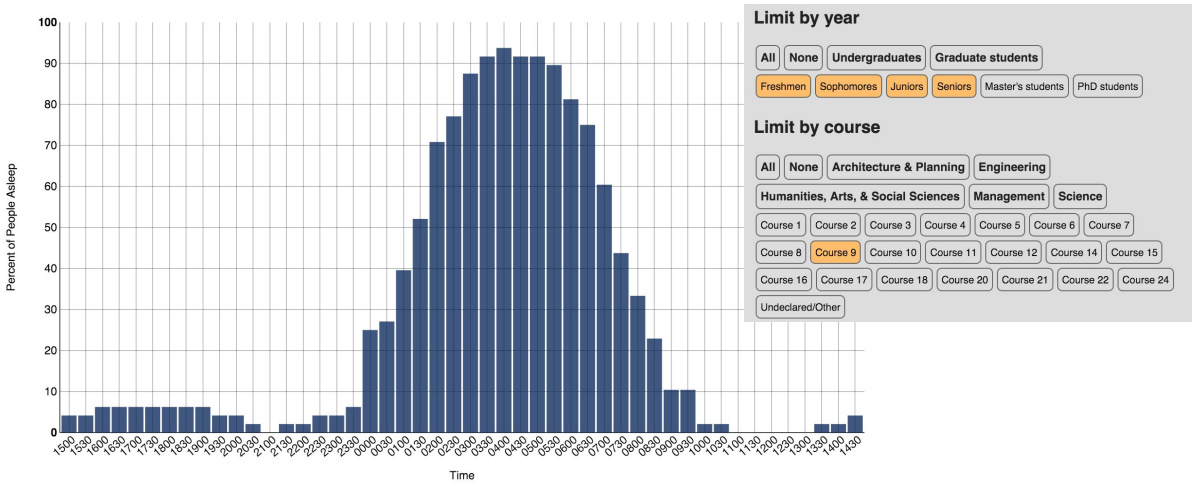
- Long Title: “An Act To amend the Internal Revenue Code of 1986 to improve portability and continuity of health insurance coverage in the group and individual markets, to combat waste, fraud, and abuse in health insurance and health care delivery, to promote the use of medical savings accounts, to improve access to long-term care services and coverage, to simplify the administration of health insurance, and for other purposes.”
- Unintended negative outcomes
  - Reduced retrospective chart-based research (responses dropped from 96% to 34% in one study on heart-attack follow up surveys)
  - Legalistic details on privacy preservation techniques has made informed consent forms even longer and less user-friendly
  - Stiff penalties for violations, lead doctors to withhold information (even sometimes from people who have rights to see it!)
  - Expensive to implement
  - Requires training healthcare providers

Risks to users/participants?

- Quasi-identifiers & Doxing/doxxing (document tracing):
  “Internet-based practice of researching and publishing personally identifiable information about an individual. The methods employed in pursuit of this information range from searching publicly available databases and social media websites like Facebook, to hacking, and social engineering. It is closely related to cyber-vigilantism, hacktivism and cyber-bullying.” (definition from Wikipedia)
- If you’re not interesting (now or ever in the future), you probably have privacy?

When are MIT students asleep?

- Leon Lin and Aaron Scheinberg, MIT Tech



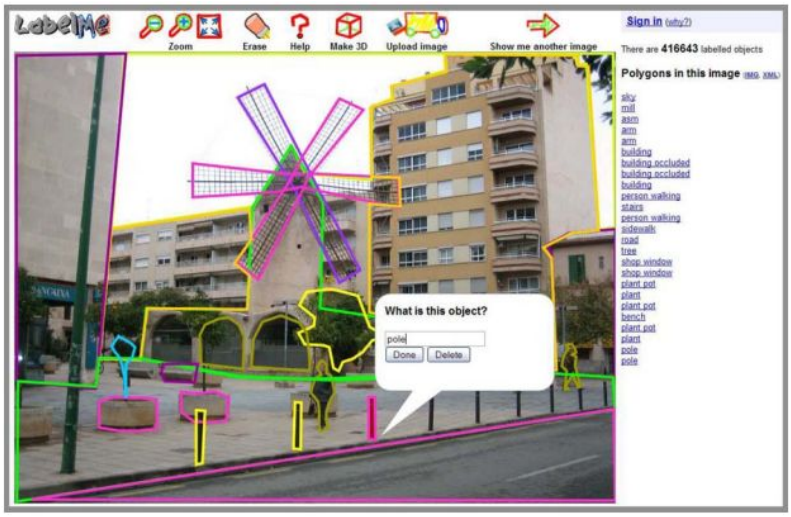
http://tech.mit.edu/V132/N59/pressure/sleepinghours/index.htm

Today

- Ethics for Journalism and Visualization
- Readings for Today
  - "Ethical Dimensions of Visualization Research"
  - “How Deceptive are Deceptive Visualizations?”
  - "Guess Me If You Can: A Visual Uncertainty Model for Transparent Evaluation of Disclosure Risks in Privacy-Preserving Data Visualization"
- Related Papers
  - "Adaptive Privacy-Preserving Visualization Using Parallel Coordinates”
  - "Agile Ethics for Massified Research and Visualization”
- Informed Consent, FERPA, HIPPA, & Plagiarism Research
- Readings for Tuesday

Reading for Tuesday *pick one*

- "LabelMe: Online image annotation and applications" Torralba, Russell, & Yuen, IEEE 2010



Reading for Tuesday *pick one*

- "QSplat: A Multiresolution Point Rendering System for Large Meshes", Rusinkiewicz & Levoy, SIGGRAPH 2000