Bradley Voytek, Ph.D.

UC San Diego

Department of Cognitive Science
Halıcıoğlu Data Science Institute
Neurosciences Graduate Program

bvoytek@ucsd.edu
@bradleyvoytek

Project proposal (10%)

Project Proposals are Completed in Assigned Groups of 4-6.

These are due Sunday (11:59 PM) after **Week 4**.

These include a proposal of what question you are setting out to answer for your final project and what data you plan to use to complete this project.

I'm in an a cappella group!

I'm in an a cappella group!

JESSICA (BOLGER) VOYTEK

Mezzo

Jessica is an enthusiastic alumnus, originally from Los Alamitos California. Like many of the VoCals, Jess has been involved in music for as long as she can remember. From piano, violin, trumpet, and bass in elementary school, to choir at Los Alamitos High school, and finally to four wonderful years with the VoCals, her love has always been singing. Currently Jess is the Information Services Manager for the Girl Scouts of San Francisco Bay Area.

I love pets!

I love pets!

I really like to read, please recommend any books you like

I really like to read, please recommend any books you like

I play d&d

I play d&d

no! :D

“How much <technical skill> do I need?”

$$x = 3$$
$$x = x+1$$

COGS 108
Data Science in Practice

Ethics and Privacy (cont.)

UC San Diego

5. PRIVACY

- Can you guarantee privacy?
- What is the level of risk of your data, and how will you mitigate the risks? Are all subjects equally vulnerable?
- Anonymization: the process of removing personally identifiable information from datasets (PII)
- Use secure data storage, with appropriate access rights

HIPAA

Health Insurance Portability and Accountability Act of 1996 (HIPAA)

The HIPAA Privacy Rule

The HIPAA Privacy Rule establishes national standards to protect individuals' medical records and other personal health information and applies to health plans, health care clearinghouses, and those health care providers that conduct certain health care transactions electronically. The Rule requires appropriate safeguards to protect the privacy of personal health information, and sets limits and conditions on the uses and disclosures that may be made of such information without patient authorization. The Rule also gives patients rights over their health information, including rights to examine and obtain a copy of their health records, and to request corrections.

Source: US Department of Health and Human Services

HIPAA

- MUNI story

De-identification

Figure 1. Two methods to achieve de-identification in accordance with the HIPAA Privacy Rule.

Source: US Department of Health and Human Services

UC San Diego

De-identification

The first is the "Expert Determination" method:

(b) *Implementation specifications: requirements for de-identification of protected health information.* A covered entity may determine that health information is not individually identifiable health information only if:

(1) A person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable:

(i) Applying such principles and methods, determines that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information; and

(ii) Documents the methods and results of the analysis that justify such determination; or

Source: US Department of Health and Human Services

De-identification

The second is the "Safe Harbor" method:

(2)(i) The following identifiers of the individual or of relatives, employers, or household members of the individual, are removed:

(A) Names
(B) All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes, except for the initial three digits of the ZIP code if, according to the current publicly available data from the Bureau of the Census: (1) The geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people; and (2) The initial three digits of a ZIP code for all such geographic units containing 20,000 or fewer people is changed to 000

Source: US Department of Health and Human Services

UC San Diego

De-identification

(C) All elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older	
(D) Telephone numbers	(L) Vehicle identifiers and serial numbers, including license plate numbers
(E) Fax numbers	(M) Device identifiers and serial numbers
(F) Email addresses	(N) Web Universal Resource Locators (URLs)
(G) Social security numbers	(O) Internet Protocol (IP) addresses
(H) Medical record numbers	(P) Biometric identifiers, including finger and voice prints
(I) Health plan beneficiary numbers	(Q) Full-face photographs and any comparable images
(J) Account numbers	(R) Any other unique identifying number, characteristic, or code, except as permitted by paragraph (c) of this section [Paragraph (c) is presented below in the section "Re-identification"]; and
(K) Certificate/license numbers	

Source: US Department of Health and Human Services

UC San Diego

Case Study: Running Data

Strava, a company who made an app that released running data, geotagged from around the world [link]

Fitness tracking app Strava gives away location of secret US army bases

Data about exercise routes shared online by soldiers can be used to pinpoint overseas facilities

Latest: Strava suggests military users 'opt out' of heatmap as row deepens

▲ A military base in Helmand Province, Afghanistan with route taken by joggers highlighted by Strava. Photograph: Strava Heatmap

6. EVALUATION

- How will you evaluate the project?
  - Do you have a verifiable metric of success?
- Goodhart's Law: when a measure becomes a target, it ceases to be a good measure.

Case Study: Teacher Rating

Washington, DC school district used an algorithm to rate teachers, based on test scores. Scores from this algorithm were used to fire 'low performers'

They had no independent measure of whether this measure improved teaching

7. ANALYSIS

- Do your analyses reflect spurious correlations?
  - Can you tease apart causation?
- What kind of covariates might you be tracking?
  - Are you inferring latent variables from proxies?

8. TRANSPARENCY & APPEAL

- Is your model a black box?
  - Is it interpretable as to how it came to any particular decision?
- Is there a way to appeal a model decision?
  - What kind of evidence would you need to refute a decision?

Case Study: Predictive Policing

- Predictive policing uses algorithms to predict crime, and recidivism
- Input data can be highly correlated [link] with race & SES, reflecting spurious correlations and leading to discriminatory decisions.
- These algorithms and decisions are often opaque and un-appealable.

9. CONTINUOUS MONITORING

- Healthy models maintain a back and forth with the thing(s) in the world they are trying to understand.
- Are you tracking for changes related to your data, assumptions, and evaluation metrics?
- Are you proactively looking for potential unintended side effects of your model itself or harmful outputs?
- Do you have a mechanism to fix and update your algorithm?

Case Study: NEWS SHARING

- Facebook is continuously making predictions about what you are going to do, which it uses to try to influence behaviour and then update its models based on the results
- Models optimize for engagement and sharing - can promote the spreading of misinformation

On systems & incentive structures

- Novel systems are not, *de facto*, equalizers. They will tend toward propagating existing inequalities.

- Companies working on these systems may have conflicts of interest with respect to the incentive structures imposed by the system and/or the business

On perpetuating inequality

- Data and algorithms can (will?) entrench disparities.
- The combination of damage, scale, and opacity can be destructive.
- They can introduce feedback in such a way as to enact self-fulfilling prophecies.

Putting it all together: The GOOD

- Well-posed question that you know something about.
- Have considered implications of work,
- Adequate data, covering population of interest, with known and manageable biases.
- Allowed to use the data.
- Have de-identified data, stored securely.
- Defined metrics for success, objectively measured.
- If suggesting causality, have actually established causality.
- Model is understandable, has procedure for appeal.
- Will monitor system for changes, have plans to update.

Putting it all together: The BAD

- Ill-posed question you know nothing about.
- Don't consider broader implications.
- Haphazardly collected, biased data.
- Didn't check, or are not allowed to use, data for this purpose.
- Un-anonymized, identifiable data, stored insecurely.
- No clear metric for success (meh, it 'seems to work').
- Present spurious correlations as meaningful.
- Model is a black box, no method for appeal in place.
- No monitoring, no way to identify biases or update model.

COGS 108
Data Science in Practice

Data Wrangling and Visualization

UC San Diego

Names

Looking at the absolute chance in percentages is interesting, but would not tell the full story. A change of, say 15% to 14% would be quite different and less drastic than a change from 2% to 1%, but the absolute change in percentage would measure those two things equally. Thus, I need a measure of the relative change in the percentages — that is, the percent change in percentages (confusing, I know). Fortunately the public health field has dealt with this problem for a long time, and has a measurement called the relative risk, where "risk" refers to the proportion of babies given a certain name. For example, let's say the percentage of babies named "Jane" is 1% of the population in 1990, and 1.2% of the population in 1991. The relative risk of being named "Jane" in 1991 versus 1990 is 1.2 (that is, it's (1.2/1)=1.2 times as probable, or (1.2-1)*100=20% more likely). In this case, however, I'm interested in instances where the percentage of children with a certain name decreases. The way to make the most sensible statistics in this case is to calculate the relative risk again, but in this case think of it as a decrease. That is, if "Jane" was at 1.5% in 1990 and 1.3% in 1991, then the relative risk of being named "Jane" in 1991 compared to 1990 is (1.3/1.5)=0.87. That is, it is (1-0.87)*100=13% less likely that a baby will be named "Jane" in 1991 compared to 1990.

(Note that I'm not doing any model fitting here because I'm not interested in any parameter estimates — I have my entire population! I'm just summarizing the data in a way that makes sense.)

Source: Hilary Parker, https://hilaryparker.com/2013/01/30/hilary-the-most-poisoned-baby-name-in-us-history/

Names

Looking at the absolute chance in percentages is interesting, but would not tell the full story. A change of, say 15% to 14% would be quite different and less drastic than a change from 2% to 1%, but the absolute change in percentage would measure those two things equally. Thus, I need a measure of the relative change in the percentages — that is, the percent change in percentages (confusing, I know). Fortunately the public health field has dealt with this problem for a long time, and has a measurement called the relative risk, where "risk" refers to the proportion of babies given a certain name. For example, let's say the percentage of babies named "Jane" is 1% of the population in 1990, and 1.2% of the population in 1991. The relative risk of being named "Jane" in 1991 versus 1990 is 1.2 (that is, it's (1.2/1)=1.2 times as probable, or (1.2-1)*100=20% more likely). In this case, however, I'm interested in instances where the percentage of children with a certain name decreases. The way to make the most sensible statistics in this case is to calculate the relative risk again, but in this case think of it as a decrease. That is, if "Jane" was at 1.5% in 1990 and 1.3% in 1991, then the relative risk of being named "Jane" in 1991 compared to 1990 is (1.3/1.5)=0.87. That is, it is (1-0.87)*100=13% less likely that a baby will be named "Jane" in 1991 compared to 1990.

(Note that I'm not doing any model fitting here because I'm not interested in any parameter estimates — I have my entire population! I'm just summarizing the data in a way that makes sense.)

These plots looked quite curious to me. While the names had very steep drop-offs, they also had very steep drop-ins as well.

This is where this project got deliriously fun. For each of the names that "dropped in" I did a little research on the name and the year. "Dewey" popped up in 1898 because of the Spanish-American War — people named their daughters after George Dewey. "Deneen" was one name of a duo with a one-hit wonder in 1968. "Katina" and "Catina" were wildly popular because in 1972 in the soap opera *Where the Heart Is* a character is born named

Source: Hilary Parker, https://hilaryparker.com/2013/01/30/hilary-the-most-poisoned-baby-name-in-us-history/

Names

The Most Trendy Names in US History

POSTED TO DATA UNDERLOAD | TAGS: NAMES | NATHAN YAU

Names are incredibly personal things. It's how we identity ourselves. We associate others, places, and points in our past with names. Maybe you recall a family member, a celebrity, or a significant other.

At the same time, it's not uncommon for two people with the same name to run into each other, and it's why gift shops can sell and profit from those mini license plates. Parents decide what they want to call their kid at some point. So as you walk through history, you end up with names that surge, some that die off, and some that come back again.

Hilary Parker already looked at the most poisoned name in US history (her own). Here we look at names from the other direction. The most trendy:

The Most Unisex Names in US History

BY NATHAN YAU / POSTED TO DATA UNDERLOAD / TAGS: NAMES

Source: Nathan Yau, http://flowingdata.com/2013/07/29/the-most-trendy-names-in-us-history/; /2013/09/25/the-most-unisex-names-in-us-history/

Names over time

Source: babynamewizard.com

Names over time

Source: babynamewizard.com

Names over time

Source: babynamewizard.com

UC San Diego

Names over time

Source: babynamewizard.com

UC San Diego

Names over time

Source: *Abacaba*

Names over time

Source: *Abacaba*