COGS 9 – A05 Discussion

Deadlines

- Reading Quiz 3 – October 20th (Yesterday), Late submission till 22nd
  - 10 points
- Mid Way Team Evaluations E.C. – October 28th (Friday)
  - 2 bonus points if you fill the form
- Assignment 2 – October 28th (Friday)
  - 40 points
  - Read through and follow the instructions

- Assignment 1 grades will be released next week
  Some issues with gradescope

Announcement

- THE QUESTION BELOW HAS TWO CORRECT ANSWERS NOW EVEN IF YOU ALREADY SUBMITTED YOUR QUIZ WE WILL MAKE SURE YOU GET CREDIT IF YOU MARKED EITHER

  Look at figure Figure 2A from Broman & Woo's Data Organization …
  [ ] 06-21-2015
  [ ] 2015-06-21
  [ ] NA
  [ ] date

- EVERYONE WILL GET FULL CREDIT FOR THE QUESTION BELOW

  Which of the following are true of a tidy dataset?
  [ ] Column headers are variable names, not values.
  [ ]Multiple variables are stored in separate columns.
  [ ]Variables are stored in rows and columns.
  [ ] All of the above

Lecture 8 (**Programming for Data Science**)

- Why and How to Program?
  - Tips on how to learn to code and resources
- Different roles in the industry
  - Software Engineer, Software Developer, Computer Scientist
  - Data Scientist, Machine Learning Engineer, Data Analyst
- R (in Academia), Python, SQL
  - Imperative vs Declarative
- Version Control (Git) -> Github
  - GOLD MINE - https://education.github.com/pack/offers
- Python Data Stack
  - Anaconda, Miniconda, Numpy, Pandas, Matplotlib, Seaborn, Scipy, SKLearn, PyTorch

Guest Lecture (Bradley Voytek)

- Cool brain stuff!! Growing brains, video of pulsating brain, zombies!
- Why data science? Turn observations into numbers and learn the fundamental laws of the universe (Planets revolving around the Sun)
- What is Data Science? Different interpretations, definitions
- Why data context is really important? Understanding where data comes from (Citizen birth b/w Sep 3-13 1752, different formats (year)
- Ecological fallacy = Conclusions of individuals based on aggregates
- Prof Voytek's data science career – How he started and work at Uber

Reading 2

- Data's Day of Reckoning
  - Large amounts of data and usage has been unethical
  - Ethics and Security training must be at the heart of the curriculum of a course
  - Ethical principles are forgotten when you're in a hurry. Solution: Checklists!
  - Ethics and security should be a part of an organization's culture. Security is gradually becoming a part, but integrating ethics has been challenging. Some ideas like the andon cord, escalating issues without retaliation, ethical challenges during hiring, etc.
  - Regulation not easy -> Agile implies policies can't catch up with code
  - Concludes by saying that we need to incorporate ethics and talk about it!

Reading 2

- Myths and Fallacies of Personally Identifiable Information (PII)
  - Developing effective privacy protection technologies is critical
  - There's too much information about us on the internet, so called "PIIs"
  - Just like alchemists believed in a Philosopher's stone that could turn any metal to gold, people believe that data can be de-identified by removing PII
  - Legal and technological definitions of PII
  - Companies assume that removing PII magically makes the data publishable. Remember that data points are real people with real lives. Think of the effect
  - K-anonymity tries to anonymize identifying attributes to make joins with external datasets difficult, but how do you distinguish identifying attributes
  - Re-identifying using any information that distinguishes one from another. Too much data and sophisticated algorithms that PII has no meaning
  - De-identification provides weak form of privacy, same in health care. PII is meaningless in HIPAA Privacy Rule
  - Differential privacy (better) but no universal methodology. Query-based

Reading 2 Quiz Hard Questions

- Storing user data as de-identified is often not enough to guarantee that no re-identification can occur if the data is leaked/stolen, why?
  - This is incorrect, properly de-identified, e.g. HIPPA's Safe Harbor strategy, data does guarantee anonymity to users. Even if the data is leaked/stolen.
  - Algorithms currently exist to take such a dataset and figure out who the user is.
  - These de-identified records can often be joined with other leaked or public data sets to statistically re-identify users with a high degree of certainty.
  - De-identification practices can easily be reversed engineered with today's compute capability, e.g. cloud and distributed compute resources.

Reading 2 Quiz Hard Questions

- Fill in the blanks: Although many software startups tend to follow the "move fast and break things" approach. It is __**"x"**__ to "move fast and break things" while also considering ethical issues by using methods such as __**"y"**__.
  - x: Impossible, y: Waterfall
  - x: Possible, y: Scrum
  - x: Impossible, y: Agile
  - x: Possible, y: Lean

Reading 2 Quiz Hard Questions

- True or False: For organizations, a viable strategy includes adopting policy similar to the Institutional Review Board's practices in order to reduce ethical violations/issues.

- True or False: Personal identifiable information (PII) is a useful concept in the context of the Privacy Rule in HIPAA.