Hand out anonymous survey about background with statistics, programming, spreadsheets, databases, anything else relevant + laptop info

Staff introductions

Student introductions

- Name, where from, which dorm
- One unusual thing that we will all remember

"Learning Goals"

- New course evaluation system emphasizes establishing learning goals, and it's the "year of learning"
- We have some in mind but would like to hear yours
- Four groups of four come up with 2-3 goals each

Components of Seminar

- In-class discussions of discoveries/fallacies/privacy, staff-led and student-led
- Apology that there will be some "lecture-style" sessions in first 3 weeks
    - Data analysis techniques - basic, no experience needed
    - Data visualization tools - basic, no experience needed
    - Data analysis tools - basic, no experience needed
- Recommended readings
- Outside guests, with and without case studies
- Student visualizations, individual
- Project #1: movie-rating predictions competition, individual or pair
- Project #2: individually-designed data analysis, individual or pair
- Field trip to Facebook or Google

Expectations

- Come to class and participate in discussions
- At least skim the recommended readings
- Prepare and present interesting data visualization
- Prepare and lead interesting in-class discussion
- Complete both projects by due-date, no lates
- (No exams)

Interlude: Facebook data-driven advertising video
What do you think?

Discovery #1: Beer and diapers

- One of the earliests much-discussed uses of Big Data was in retail: WalMart, Victoria's Secret
- *"Market-basket" data*

    B1: {milk,bread,eggs,beer,diapers}
    B2: {bread,beer}
    B3: {milk,beer,diapers}
    B4: {eggs,diapers}
    B5: {milk,bread,diapers}

B6: {milk,bread,eggs}

- *Frequent itemsets*
  A set of items is “frequent” if the items appear together in at least X% of baskets
  In real data X might be 1%. For X = 50%: {milk,bread}, {milk,diapers}, {beer,diapers}
- *Association rules*
  Set of items -> item: If *Set* is bought together then *item* is likely included too
  Set should be frequent (more than X% of baskets), and item should be in at least Y% of those baskets
  For X=Y=50%: {milk,eggs}->{bread}, {diapers}->{beer} (and others)
- First rule: French toast. *Why beer and diapers?*
  - Women send guys out for diapers and they pick up beer too
  - People with babies tend to drink at home
  - Underage buyers add diapers to make themselves seem older

Fallacy #1: False correlations

- *Correlation does not imply causation*
- Show a few of the spurious correlations
- Seat-belt cartoon: reverse causation
- Can also have “confounding variable” A where A causes both B and C therefore B and C appear to be correlated.
- Example: B = parental smoking causes C = delinquent children. Could be that C causes B, or that A = poverty causes both.

Fallacy #2: Football game result prediction scam

- You receive an email from “Prescient Polly” on Saturday predicting the winners of four of Sunday’s football games. She’s right.
- Same thing happens the following weekend, and then two more weekends. Four weekends of perfect predictions!
- On the fifth weekend, Polly offers to place bets for you on the next day’s games, for a modest fee.
- *Should you do it?* (breakout groups)
- *How many initial emails to have 100 possible takers on weekend five?*
  - 16 games = 2^16 = 65,536 possible outcomes. x100 = 6,553,600. Not that many!

Privacy: Google Maps traffic

- End of class food for thought
- Traffic in “old” days: sensors in roads
- Now: phones transmit location and speed
- Ethical? Pitfalls?