# Lessons Learned in Transitioning from *Intro to Statistics* to *Reasoning with Data* (Converting a Class to a Community)

Rebecca Nugent

Carnegie Mellon University Statistics

August 1, 2017

# Carnegie Mellon University

- Private university in Pittsburgh, PA
- R1 research university designation
- $\approx$ 6500 undergrads, 6500 grads
- Six undergraduate colleges (admission is college-specific)
  - College of Fine Arts
  - Dietrich College of Humanities & Social Sciences (including Information Systems)
  - College of Engineering
  - Mellon College of Science
  - School of Computer Science
  - Tepper School of Business

# Carnegie Mellon Statistics

- Dietrich College of Humanities & Social Sciences (1200)
  - Economics (joint in Tepper)
  - English
  - History
  - Information Systems
  - Institute for Politics and Strategy
  - Modern Languages
  - Philosophy
  - Psychology
  - Social and Decision Sciences
  - Statistics
- Undergraduate Statistics: around 450 primary/add'l majors
  - Statistics (Concentration: Open, Math, Neuroscience)
  - Economics-Statistics
  - Statistics and Machine Learning

# Dietrich General Education Curriculum

*Current:*

- Computing @CMU
- Introduction to Statistics
- Writing/Interpretation
- History survey course
- Freshmen seminar
- Remaining courses are selected from categories: Communicating, Creating, Deciding, Modeling, Reflecting

Students largely feel like the courses aren't well-connected to each other; hard to get into some of them; often end up just looking for one to take or trying to get substitutions

Courses often aren't designed for general education

# Dietrich General Education Curriculum

In midst of multiple-year revamp/design of new requirements

College-wide committees looking for some general consensus (you can imagine how well that's going)

*Proposed for Fall 2018:*

- Reasoning with Data
- Writing minis: choose 2 of 3 themes (e.g.Writing with Data)
- Freshmen and Junior seminars
- More interdisciplinary courses across departments
- Experiential Learning and Service
- Portfolio and Self-Reflection
- Broader focus on courses that are truly general education

So how do we develop an updated gened intro course in statistics/data science for liberal arts and *some* STEM that

1. the students don't hate
2. the students find useful
3. doesn't really require any kind of computing (background)
4. illustrates the richness/diversity of types of data and interdisciplinary problems in data science
5. is reproducible and can grow/build on itself

# Where we're starting

*36-201: Statistical Reasoning and Practice*

- Between 120-240 each semester; required for all Dietrich
- Standard class structure: lecture, HWs, labs, exams
- Lab: hands-on practice of concepts
  - using Minitab as the software package
  - very results-oriented (calculate, basic interpretation)
  - staffed by master's/undergraduates (primarily) in Statistics
- Concepts: EDA, experimental design, elem. probability, discrete distributions, normal distribution, sampling distributions, confidence intervals, significance tests, one-way ANOVA, contingency tables, linear regression
- Emphasis is on learning steps of analysis, correct formulas to use for which scenario ("cookbook" approach), basics

# Where we're starting

*36-201: Statistical Reasoning and Practice Challenges:*

- Can be viewed as "dry"
  (common complaint about intro stat courses in general)
- Formula-driven, learning a set of steps
- Hard to have students see the whole picture
- Primarily focused on continuous and categorical data
- Students use different software packages in different disciplines
- Intro examples tend to be more STEM-driven, social science applications, very little humanities
- Contains almost no modern material or applications

# What we're doing now

Interviews w/ different Dietrich depts about issues/needs

- Students don't know the concepts
- Get tied to the specific software steps
- Can't see the big picture
- Class isn't really for them

*Our goals:*

- Modernize the course
- Emphasize concepts; tell stories with data
- More student-driven inquiry
- More adaptive material
- learn how different students interact with data
- **Convert the class into a community**

# What we're doing now

*Course Vision*

- Concepts, concepts, concepts
- Interpretation, interpretation, interpretation
- Remove computing cognitive load completely
- Interact/engage with problems
- Data: Continuous, Categorical, Text, Images, Networks, Time Series (neuroscience, economics, etc)
- Use atypical data types to illustrate intro level concepts
- Include overview of modern statistics methodology taught at conceptual level (classifiers, clustering, "big data"...)

# What we're doing now

*Class Structure*

- Lectures, weekly HW
- "Discussion Labs"
  - small (15-20)
  - staffed by TAs from Stat and other Dietrich units
  - presented as a case study/story
  - combination of questions and interpretation discussion
    What do these results mean? What might we do next?
- Quizzes/Final Exam
- Data Story project(s)

# What we're doing now

*Interactive Software Platform (ISLE)*

- No coding. More direct interaction with the concepts
- Easy to adapt new hypotheses, problems during class/lab
- Can collect answers from students; propagate through labs
- "Data Set Explorer": upload (formatted) data, variables
- Students can save graphs and work to editors that create websites/documents for a portfolio
- We can collect information on clicks, decisions, times, etc - how do students explore data?
- Combining tools like Java Script with R
- Built in modular form; can "mix-and-match"

Philipp Burckhardt, Tues 8/1, 11:05am, Session #344, CC-309

# What we're doing now

*Let's look at some examples:*

Interactive Venn Diagram
Conditional Probability with Text
Collecting Data
Forest Fires

Everything they do, everywhere they go - we can track them.
I promise it's not creepy.

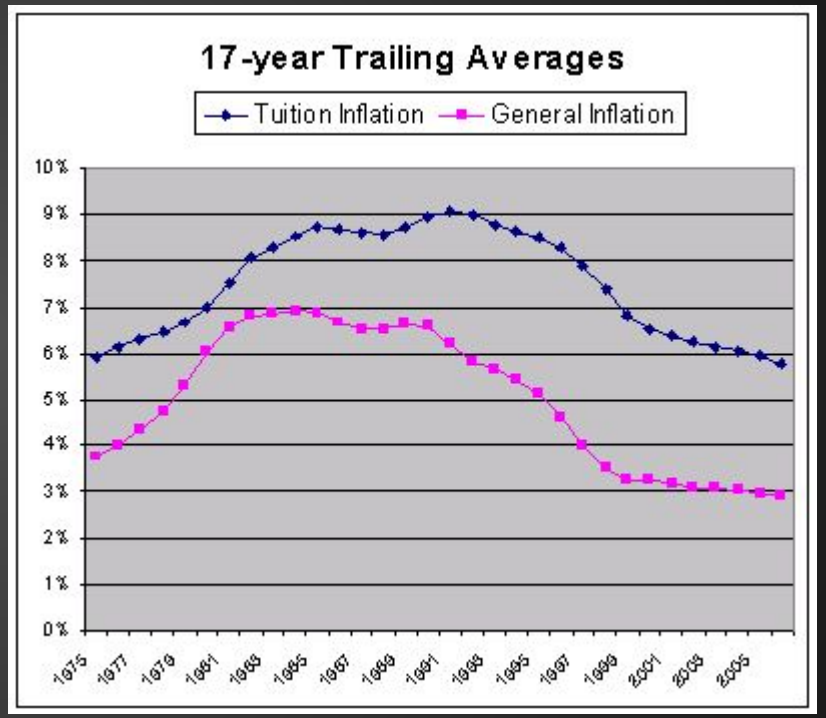
# What we're doing now

*Telling Stories with Data:*

- Tweaked version of case study approach
- Start with a general problem description and large multivariate data set (in the explorer)
- Build tools and skills in service of a layered data analysis/story
- Have students generate/motivate their own hypotheses
- Students can upload/contribute data sets to the explorer ("normal", text, images, etc)
- Projects can be suggested or self-defined
- Building a repository/community that grows with instructor supervision

# Increasing Tuition Rates

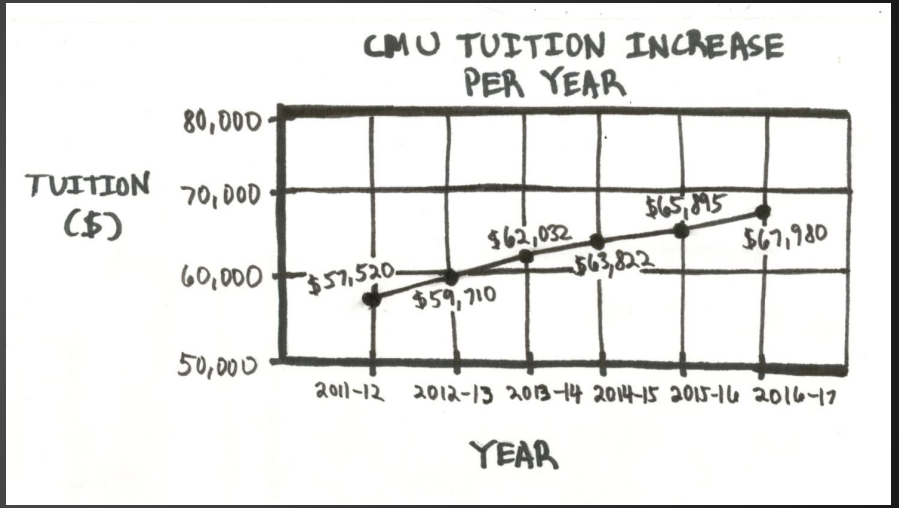
- College tuition rates have been increasing each year - on average between 6-9%
- Tuition rates increase at about 2x the general inflation rate
- Particularly bad period of growth that happened between 1958 and 2001
- Students rely heavily on federal aid, scholarships, grants, student loans, and more to attend college.




**1975 - 2005: Tuition rates in response to general inflation**
http://www.finaid.org/savings/tuition-inflation.phtml

# Changes at CMU

- CMU raised its tuition by 3.38% (2016) and will increase it again by 3% (2017)
- *"...supports the expansion of undergraduate education initiatives, and the improvement of facilities and various services to help students succeed and thrive at [Carnegie Mellon]." - Provost Jahanian*
- Tuition rates continue to increase and financial aid packages don't change




**Graph showing the tuition increase at CMU over the years.**
https://thetartan.org/2017/5/1/news/tuition-increase

# Out of State Tuition vs. # Full-Time Undergrads

**Question #1:** Is there a correlation between **Out of State Tuition and the Number of Full-time Undergraduate Students?**

**Assumptions:** Tuition costs increase as more full-time students join the student body. Need to pay for more buildings, utilities, classrooms, etc. Have to hire more professors.

**Null:** No correlation between Out of State Tuition and NumberFTUGs.

**Alternative:** The correlation is greater than 0.

# Out of State Tuition vs. # Full-Time Undergrads

**Hypothesis test for correlation between OutStateTuition and NumberFTUGs:**

$$H_0 : \rho = 0 \; vs. \; H_1 : \rho > 0$$

```
t-test for Pearson correlation coefficient

Alternative hypothesis: True correlation coefficient is greater tl

        pValue: 1
        statistic: -6.3514
        95% confidence interval: [-0.3041,1]

Test Decision: Fail to reject null in favor of alternative at 5% s
```

Surprised that the correlation test came back with a P-Value of 1 and a statistic of -6.35. We fail to reject the null in favor of the alternative at a 5% significance level. The actual correlation coefficient is not greater than 0. The reported 95% confidence interval tells us that the correlation value could be anywhere between [-0.3041, 1]. The actual correlation between Out of State Tuition and Number of Full Time Undergraduates is -0.244.

# Out of State Tuition vs. # Full-Time Undergrads


Grouping by Type: Correlations become (+)!
Private: +0.260 | Public: +0.305

- Two very distinct groups. Private data is mostly scattered around the bottom in a linear fashion.
- *What's happening*? Doesn't make sense to look at them together.
- *Public: More undergrads, cheaper $$$.*
  - *1-30,000+ students, $5k-10k*
- *Private: More expensive, less # of UGs.*
  - *1-500 students, wide range for cost*

# What we're doing now

*Data@Dietrich*

- Video series from across Dietrich: *What Is Data?*
- Assigned to students as HW or used for case studies
- Example from Modern Languages

- Students get to see breadth/depth of problems
- Statistics majors see beyond their coursework
- Non-stat majors see potential future projects
- Faculty get advertising for all their projects and classes at minimal cost to them
- Everybody gets an active public-facing website repository

# What we're going to do this year

*How do liberal arts and Stat students interact with data?*

- Track them as they move through the platform
- Analyze the different paths
- In practice, can adapt as the lab is happening
- Different discussion labs for different groups?

# What we're going to do this year

*Discover Data@Dietrich*

- Currently have Tartan Data Science Cup "DataFest"; company sponsorship; big community event
- Intro level workshops for Dietrich college (or anyone)
- Joint between Statistics and Dietrich unit
- Present research problem and question
- Data, beginning analysis steps
- Workshop the problem, see what happens

# Takeaways

- Building Data Science courses is not enough
- Humanities/Liberal Arts students may need something completely different; need to figure out what that is
- Need software/platforms that allow for customization without requiring comp background (for students, teachers)
- More interaction with data, more interaction with each other
- Convert class to community
- Give “ownership” to stakeholders
- Community should grow, evolve with contributions
- Build easily-traversed bridges

# The Team/Upcoming

- Philipp Burckhardt: 11:05am, Session #344, CC-309
- Gordon Weinberg
- Christopher Peter Makris
- Kayla Frisoli
- Jamie McGovern (Happy Anniversary!)

- Electronic Conference on Teaching Statistics (eCOTS): May 21-25, 2018
- Carnegie Mellon Sports Analytics Conference/Tartan Data Science Cup: October 28-29, 2017

  `http://www.stat.cmu.edu/tartandatasciencecup`

  `http://www.cmusportsanalytics.com/conference`

rnugent@stat.cmu.edu, `http://www.stat.cmu.edu/~rnugent`