Syllabus - Intro To Data

- **Instructor:** Andy Choens, MSW
- **Course Name - Number:** Intro to Data - PBH210 & PBH211
- **Course Location - Time:** Thursday 18:30 - ~21:30
- **Office Hours and Location:** Available on request - Slack/Email is easiest
- **Contact Information:**
  - Phone: (518) 275 - 5984
  - Email: Andrew.Choens@acphs.edu
- **Slack Workspace:** acphsintrotodata

Mathematics is the language in which God has written the universe.

– Galileo Galilei

Schedule

Description	Date	Weekday
First Day of Spring Semester	Jan 14, 2019	Monday
Week 01	Jan 17, 2019	Thursday
Week 02	Jan 24, 2019	Thursday
Week 03	Jan 31, 2019	Thursday
Last day to drop a class	Feb 01, 2019	Friday
Week 04	Feb 07, 2019	Thursday
Week 05	Feb 14, 2019	Thursday
President's Day (College Closed)	Feb 18, 2019	Monday
Week 06	Feb 21, 2019	Thursday
Week 07	Feb 28, 2019	Thursday
Week 08	Mar 07, 2019	Thursday
Spring Break Begins	Mar 11, 2019	Monday
Spring Break (No Class)	Mar 14, 2019	Thursday
Spring Break Ends	Mar 15, 2019	Friday
Week 09	Mar 21, 2019	Thursday
Last day to withdraw from a course	Mar 22, 2019	Friday
Week 10	Mar 28, 2019	Thursday
Week 11	Apr 04, 2019	Thursday
Week 12	Apr 11, 2019	Thursday
Week 13	Apr 18, 2019	Thursday
Week 14	Apr 25, 2019	Thursday
Last day of Spring Semester classes	Apr 30, 2019	Tuesday
Reading Day	May 01, 2019	Wednesday
First day of Final Exams (Our Final Exam)	May 02, 2019	Thursday
Last day of Final Exams	May 08, 2019	Wednesday

Description	Date	Weekday
Emergency Make Up Day	May 09, 2019	Thursday

Attendance Policy: I expect you to attend/participate in class and lab. By doing so, you will learn more. With that said, you are an adult and I intend to treat you as such. If for some reason you are unable to attend class notify me as soon as possible and we will work out how you can complete the necessary assignments.

Office Hours: As an adjunct, I dont have an office other than our classroom and I spend most of my day in Troy, NY at my day job. If there is something you would like to discuss, contact me via Slack and we will arrange a time to meet. You are welcome to come out to my office in Troy where we can meet in a conference room. Alternatively, I will usually be in the classroom by ~6:00 PM on Thursday nights. If you tell me ahead of time - I can guarantee I will be in the classroom by 6:00. I am here to support you. Let me know how I can best do that.

Inclement Weather: If the school is closed we will not have class. If the school is open we will have class. With that said, our class is at night and I suspect at least some of you live off campus. If you feel unsafe about traveling before/after our class due to weather, let me know and we will work something out.

Course Description / Objectives

> Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, both structured and unstructured,[1][2] similar to data mining.
>
> Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data.[3] It employs techniques and theories drawn from many fields within the context of mathematics, statistics, information science, and computer science.
>
> Turing award winner Jim Gray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the data deluge.[4][5]
>
> Data science

Topics touched upon in this interdisciplinary course include:

Computer Science	Statistics	Research Fundamentals
Import data	Sampling theory	Develop research questions
Review/clean data	Confidence intervals	Answer research questions
Manage data	Monte Carlo simulation	Research ethics
Generate summary statistics	Machine Learning	
Literate programming		
Relational data		

Prerequisites

This course has a single prerequisite - curiosity. No analyst or researcher can succeed with out curiosity. Bring it in abundance.

Textbooks & Software

Out-of-pocket costs for this course are zero. All materials needed to complete this course are available online, for free. That said, these materials *are* very important. Because classroom time is limited, there will be important material covered in the reading that will not be covered in-class.

Textbooks

1. Hands-On Programming with R (HOPR) by *Garrett Grolemund*
2. R For Data Science (R4DS) by *Hadley Wickham and Garrett Grolemund*

There is a dead-tree edition of each text. These cost money, but the authors have more than earned it. I personally own a copy of each. However, the online version is identical to the dead-tree edition and is entirely sufficient for the course.

Recommended Reading

Specific readings will be assigned from these additional resources:

- R-bloggers
- Andrew Gelman's blog: Statistical Modeling, Causal Inference, and Social Science
- Frank Harrell's blog: Statistical Thinking

Software

The tools used in this class are Free and Open Source Software. We'll discuss what that means in class, but for now it can simply mean you don't have to pay anything to use them. The first lab includes time to install and configure these tools.

Students will use the following software:

Name	Recommended Download Link	License	Cost
R	Download R - CRAN	GPL v2	$0.00
R Studio	Download RStudio Desktop	AGPL v3	$0.00
Excel/LibreOffice	Download LibreOffice	MPL v2	$0.00

Students may use other R IDEs such as Emacs ESS, but classroom examples will focus on RStudio. Students may use either Microsoft Excel or LibreOffice Calc for spreadsheet assignments/projects (or other compatible software).

Assignments & Grading

The lab portion of the class will be graded separately from the class.

Table 4: Intro to Data Assignments

Assignments	Assign Date	Due Date	% Final Grade
Class Participation	2019-01-17	2019-05-09	5%
Assignment 01	2019-01-24	2019-01-31	5%
Assignment 02	2019-02-07	2019-02-21	10%
Midterm Exam		2019-03-07	25%
Assignment 03	2019-03-21	2019-03-28	10%
Assignment 04	2019-04-11	2019-04-18	10%
Assignment 05	2019-04-18	2019-04-25	10%
Final Exam		2019-05-02	25%
Total			100%

There will be one lab per week, except Week 8 (2019-03-07) and on the day of the final exam (2019-05-02). Each lab will be equally weighted.

Table 5: Lab Grade Criteria

Grade Letter	Numeric Score	Criteria
A	95	Your lab notebook contains complete, correct, and throughtful answers for all questions. It is evident you are completing the reading and learning the material.
B	85	Your lab notebook contains a mostly complete, correct, and relevant answers for most questions. It is apparent you are completing the reading and learning the material.
C	75	Your lab notebook is incomplete or contains consistent errors. It is not apparent you are completing the reading and learning the material.
D	65	Your lab notebook is incomplete or contains consistent errors. This is like a C but "more so". Are you completing the assigned reading? We should talk.
F	55	Your lab notebook demonstrates that you are not completing the readings and/or understanding the material. We should talk, soon.
NA	00	You failed to complete and/or submit the lab. We should talk about your future in this class immediately.

Tentative Lecture Plan

`PBH210` is a class. `PBH211` is a lab. Programming and working with data are skills best learned by asking lots of questions and engaging in the practice. There is a lecture component of this class. You will get more out of it if you are engaged. If you don't understand - tell me.

The subject matter see in the lab will built upon what we learned during the lecture portion of our evening. With the exception of Week 08 (Midterm) students are permitted to discuss their labs with others in the class, ask for help on Slack, etc. That said - what you submit should be YOUR work.

Students are expected to complete all class readings. I cannot stress this enough. Most of you, I assume, are not programmers. There is going to be a learning curve. If class is the first time you encounter a concept - you won't learn this material.

Week 01 - 2019-01-17

- **Topics:** Course Introduction
  - Who are you? Please complete your name poster!

  - Who are you? Who am I?
  - What is "Data Science"?
  - Class Survey
  - **Review Syllabus**
  - Our Tools (R, RStudio, Spreadsheets)
    * **Note:** Students may access these tools via the ACPHS virtual desktops.
  - Setup!
- **Homework:**
  - HOPR: Project 1 - Weighted Dice
  - HOPR: The Very Basics
  - HOPR: Packages and Help Pages

Week 02 - 2019-01-24

- **Topics:** Titanic Introduction To R (Vectors)
  - Creating
  - Atomic vector types
  - Indexing/Filtering
  - Basic Plotting
  - Compare R to Excel
  - Sampling
- **Homework:**
  - Assignment 01
  - HOPR: Project 2 - Playing Cards
  - HOPR: R Objects
  - HOPR: R Notation
  - HOPR: Modifying Values

Week 03 - 2019-01-31

- **Topics:** Titanic Introduction To R (Data Frames)
  - Relationship with vectors
  - Indexing/Filtering
  - Creating new columns
  - dplyr as an alternative to base R (filter)
  - More basic plotting
  - Simple intro to Tidy Data
  - Importing Excel/CSV
  - Compare R to Excel
- **Due:** Assignment 01
- **Homework:**
  - R For Data Science (R4DS): Welcome
  - R4DS: Introduction (1)

  - R4DS Explore-Intro
  - R4DS Data Visualization with GGPLOT
  - R4DS: R Markdown

Week 04 - 2019-02-07

- **Topics:** Visualize This
  - dplyr group_by, summarize, mutate
  - Visualization (GGPLOT)
  - Workflow Basics
  - Markdown
  - Working Directory
  - Compare R to Excel
- **Homework:**
  - Assignment 02
  - R4DS: Workflow Basics
  - R4DS: Data Transformation
  - Tidy Data by Hadley Wickham
  - Wikipedia: Literate Programming

Week 05 - 2019-02-14

- **Topics:** Logic & Flow
  - Review Boolean Logic and Testing
  - if(), for()
  - Why write a script?
  - More Data Import
  - Data Management Goal: Tidy Data!
  - Data Management Process
  - Review dplyr (group_by, filter, select, mutate, etc.)
  - Strengths of Excel/R
- **Homework:**
  - R4DS: Workflow: Scripts
  - R4DS: Exploratory Data Analysis
  - Why I Love R Notebooks
  - R4DS: Workflow: Projects

Week 06 - 2019-02-21

- **Topics:** Exploratory Data Analysis
- **Due:** Assignment 02
- **Homework:**
  - R4DS: Relational data

    * Read this one a couple of times. It isn't easy.

Week 07 - 2019-02-28

- **Topics:** Relational Data 01
  - Primary Key
  - Foreign Key
  - Inner Join
  - Outer Join
- **Homework:**
  - What Has Happened Down Here Is The Winds Have Changed
  - R4DS: Strings
  - R4DS: Factors

Week 08 - 2019-03-07

- **Topics:** Relational Data 02
  - More relational data.
  - Let's discuss what didn't make sense last week.
  - Two Hats: Science v Technical
    * How to write a research question.
  - **Midterm Exam**
    * This is basically a lab - but you can't help one another.
    * Will NOT include any relational data questions/topics.
    * You *may not* help each other.
    * You *may* use your class notes and Google.
    * But you still need to prepare. This is a real test and it will *test* you.
- **Homework:**
  - R4DS: Iteration
  - R4DS: Functions
  - Reproducible Research With A Marmot
    * This is the most entertaining read assignment of the course.
  - Holy Coding Error, Batman
  - Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogof:
    * This is a long read. Just get the general flavor.
  - Reinhart Rogoff: Everyone makes coding mistakes we need to make it easy to find them graphing uncertainty

This is genuinely more reading than I would assign in a single week. Fortunately, you have two weeks to finish this and some of these readings are actually quite entertaining.

Spring Break - 2019-03-14

- Please use this time to review anything you did not understand in the first half of the course.
- Based on our progress during the first half of the course, I may change the readings/discussions in the second half.

Week 09 - 2019-03-21

- **Topics:** Speaking Truth To A Marmot
  - Literate Programming and Reproducible Research
  - Explain it to the marmot
  - How I Met Your ~~Mother~~ Package
  - Revisiting: The Working Directory
- **Homework:**
  - Assignment 03
  - Tidyverse Packages
    - I do *NOT* expect you to read everthing here. Just look over the page and get a sense of what is available.

Week 10 - 2019-03-28

- **Topics:** Tell Me Straight - Does It Fit?
- **Due:** Assignment 03
- **Homework:**
  - Is Most Published Research Wrong?
    - Possibly - Yes
  - Hurricanes vs. Himmicanes
  - Wikipedia: Linear regression
  - R4DS: Model Basics
  - R4DS: Model Building
  - R4DS: Many Models
- Reproducibility v Replicability
- Simple linear regression:
- Multivariate linear regression

Week 11 - 2019-04-04

- **Topics:** The Surprising Adventures of Baron Munchausen

- **Homework:**
  - Incrementalism - Provided by the instructor.
  - Abandon Statistical Significance
  - Pull Oneself Up By One's Bootstraps
  - Bootstrapping
  - Forget Excel: This Was Reinhart and Rogoff's Biggest Mistake
- Sampling Theory Primer
  - Standard Deviation
  - Confidence Intervals
- Bootstrap

Week 12 - 2019-04-11:

- **Topics:** Lasso Your Future :robot_face:
  - Ethical Analysis 1
  - More Bootstrapping: Risk Ratios
  - Lasso
- **Homework:**
  - Assignment 04
  - How Machines Learn
  - Wikipedia: Lasso
  - Do Scientists Feel Pressure To Produce Positive Results?

Week 13 - 2019-04-18:

- **Topics:** Can't See The Forest For All These Trees
  - Ethical Analysis 2
  - More About Machine Learning :robot_face:
  - Decision Trees
  - Random Forests
    - The relationship between the random Forest and bootstrapping
- **Due:** Assignment 04
- **Homework:**
  - Assignment 05
  - Learning To Fool Our Algorithmic Spies
  - How Open Source Can Fight Algorithmic Bias

Week 14 - 2019-04-25:

- **Topics:** From Soup To Nuts
  - *Use* the Analytical Template
  - Choose Your Own Adventure
- **Due:** Assignment 05

- **Homework:**
  - We Used Broadband Data We Shouldn't Have, Here's What Went Wrong
  - Can You Use This Data Set to Find Serial Killers?
  - Exploring the ChestXray14 dataset: problems

Final Exam - 2019-05-02

- The final exam includes complimentary pizza.
  - Please tell me if you have any allergies or dietary requirements.
- This is a choose your own adventure exam!
  - You will choose a "Research Project" from a set of options.
  - You will be told how to access the data.
  - You will be given a set of questions you will need to answer.
  - You will need to:
    * Import the data.
    * Answer the research question using one or more tables or visualizations.