course_intro

March 22, 2018

In [4]:
```
%%html
<style>
.h1_cell, .just_text {
    box-sizing: border-box;
    padding-top:5px;
    padding-bottom:5px;
    font-family: "Times New Roman", Georgia, Serif;
    font-size: 125%;
    line-height: 22px; /* 5px +12px + 5px */
    text-indent: 25px;
    background-color: #fbfbea;
    padding: 10px;
}

hr {
    display: block;
    margin-top: 0.5em;
    margin-bottom: 0.5em;
    margin-left: auto;
    margin-right: auto;
    border-style: inset;
    border-width: 2px;
}
</style>
```

```
<IPython.core.display.HTML object>
```

Module 0: Topics in Data Science

This quarter, we will focus on the topic of text analytics. Professor Dou and I, along with an army of graduate students, have been working on a specific research question in the text analytics area: if we ask a college student to write a short summary of their weekly reading (the text part), can we (the computer) spot misconceptions/misunderstandings (the analytics part)? I can give you the punch line: it is not easy! We are now into our 4th year on this project.

There are two basic approaches to analyzing text. Alternative 1 I'll call knowledge-based. It attempts to use human knowledge of linguistics, sentence structure, word meaning to analyze a piece of text. This is a time-honored approach dating back to the very start of AI in the 1960s. And it is this approach we are using currently for our misconception analysis.

The second approach is the new kid on the block. It is statistical and even alien in some ways. There is no attempt to use the knowledge-based approach. Instead, we turn text (words) into data that reflect complicated relationships with each other. We end up with vectors in hundreds and more typically thousands of dimensions. And what analytics method is good at analyzing huge amounts of numeric data? Yep, machine learning and deep learning in particular. So we turn the study of linguistics, creative writing and human communication into big number crunching.

I hope this stirs you a bit to say "Yes, AI conquers all!" or "You've now gone too far." What I would like to do in this course is at least delve into some intro pieces of the statistical approach. You don't have to agree with the approach. But you should have knowledge of how it works. Given time, we might work in some modules on the knowledge-based approach.

Course caveats

This is a pilot course. And really a joint exploration of many of the topics: we will be exploring them together. I chose to look at NLP and deep learning because I want to know more about it. Perhaps it is a better method for the misconception problem we have been working on for 4 years, who knows. Given all this, I cannot promise a polished syllabus. Nor a smooth set of modules that each have just the right amount of content and homework. I'll do my best but it could be bumpy. I had to laugh when a student asked me if they could see my detailed syllabus for the class. I said "Uh, no. I barely was able to get the course blurb together!"

This course will be taught again next year and much more polished so that might be an option for you.

Also note that the course is heading for a purely web-based version in the future. To move in that direction, I will not be holding in-class lectures. Instead, I'll use our class time for office hours.

Jupyter notebooks

I love jupyter notebooks. I think they are simple but revolutionary in terms of doing data science. What you are reading is a jupyter notebook. I'll do a little coding to show you. I'll bring in a dataset from the Titanic passenger list and display the first 5 rows. I'll meet you again down past the code.

```
In [1]: import pandas as pd
        url = 'https://docs.google.com/spreadsheets/d/1z1ycUZjJpmMWB4gXbhwRQ9B_qa42CwzAQkf82mL
        titanic_table = pd.read_csv(url)

In [2]: #I am setting the option to see all the columns of our table as we build it, i.e., it
        pd.set_option('display.max_columns', None)

In [3]: titanic_table.head()
```

Out[3]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1
4	5	0	3				

```
4                            Allen, Mr. William Henry    male  35.0      0

   Parch            Ticket     Fare Cabin Embarked
0      0         A/5 21171   7.2500   NaN        S
1      0          PC 17599  71.2833   C85        C
2      0  STON/O2. 3101282   7.9250   NaN        S
3      0            113803  53.1000  C123        S
4      0            373450   8.0500   NaN        S
```

Cool, huh. I can mix text (markdown and html) with runnable code. There is an html cell at the very top of the notebook that allows me to tailor how things show up. You can change this to whatever you like. All of the course content will be presented as jupyter notebooks.

You are viewing this because github has a notebook viewer built-in. Thanks, github. But you cannot run my code from github. To do that, you can download this notebook to your own computer and make sure you keep the .ipynb suffix. Then I recommend downloading anaconda, which contains the jupyter notebook server along with lots of useful data science packages.

ANother caveat: I am using Python 2.7, somewhat reluctantly. I have yet to convince myself that all the packages we will be using are stable in 3.6. However, when you donwlaod anaconda, you may choose 2.7 but actually 3.6 will come along with it. You can follow directions on the web for setting up environments (activations) for both. And the cool thing is that I have my jupyter environment set up so I can switch between 2.7 and 3.6 for this notebook, right now, by going up to the kernel tab and changing the kernel. Pretty slick.