Applied and Industrial Mathematics Institute for Secondary Teachers

Randy Paffenroth
Associate Professor of Mathematical Sciences, and Associate Professor of Computer Science
Associate Professor of Data Science
Worcester Polytechnic Institute

7-19-2017
Outline

• A stranger in a strange land.
  • Working as a mathematician in industry
• A current area of interest at the intersection of academics and industry
  • Data Science
• How can you get involved!?
  • There are many beautiful Data Science problems that are accessible to high school students
• Most importantly, please ask questions!
If we are going to talk Data Science then we need some data!

A pop quiz...

• How many people know what Data Science is? (I didn't for a long time :-)

• How many people have used "data" when reaching a class?

• How many people have heard of the programming language Python?
Who am I?

- I am currently a professor at Worcester Polytechnic Institute
- I came to WPI three years ago as one of the two hires to kick off the WPI Data Science Program
- Before coming to WPI I was program director at a small company (50 people), and before that I worked at the California Institute of Technology and another small company (3 people!).
An example of a place where mathematicians work

http://www.numerica.us/

What did I do here?
Many others!
Job titles can be deceiving...

Table 6a: Job titles from survey

<table>
<thead>
<tr>
<th>Job Title</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>statistician</td>
<td>17</td>
</tr>
<tr>
<td>analyst/modeler</td>
<td>20</td>
</tr>
<tr>
<td>researcher</td>
<td>21</td>
</tr>
<tr>
<td>management title</td>
<td>15</td>
</tr>
<tr>
<td>consultant</td>
<td>9</td>
</tr>
<tr>
<td>engineer</td>
<td>7</td>
</tr>
<tr>
<td>software developer/programming</td>
<td>11</td>
</tr>
</tbody>
</table>
A current really hot area

Data Science

Data Science
Vs
Data Literacy
Vs
Data Analytics
What is Data Science?

- What do you think it is?
- What do you need to know to do it?
- How is it used?
What is Data Science?

- Based upon Drew Conway's Data Science Venn Diagram
Data Literacy – Your turn to work
Warm up problem

Data Literacy – Warm up problem

A wins

C wins

E wins

F wins

D wins

E wins

F wins

B wins

C wins

E wins

F wins

D wins

E wins

F wins
It can be easy to fool yourself!

Human beings are really good at pattern detection...
It can be easy to fool yourself!

Human beings are really good at pattern detection...

Perhaps a bit too good!

It can be easy to fool yourself!

Since statistical rigor is so important, let’s do a little fun math!

One of the things that opens students eyes to why combining math with data is important.
The Base Rate Fallacy is a very common error that people make when interpreting data.

- It is quite easy to describe (and hopefully understand).
- It does not require very much mathematical background.
- It demonstrates that our intuition can lead us astray.

https://en.wikipedia.org/wiki/Base_rate_fallacy
Suppose you have taken a test for a deadly disease.

The doctor tells you that the test is quite accurate, in that, if you have the disease then the test will correctly tell you that you have the disease 100% of the time.

However, if you don't have the disease, the test will very occasionally (say 1 time in 10) mistakenly tell you that you have it.

The test comes back positive (it says you have the disease)! Are you worried!?

In particular, can you estimate the probability that you actually have the disease given that the test came back positive?
Base Rate Fallacy

• What is your estimate?
  A) 99% probability I have the disease
  B) 90% probability I have the disease
  C) 50% probability I have the disease
  D) 10% probability I have the disease
  E) I don't know and I am mad at you for asking me!
The importance of asking the right question.
I was told the *probability* that I failed the test *given* that I have the disease.

\[ Pr(\text{I fail the test}|\text{I have the disease}) \]

I was told the *probability* that I failed the test *given* that I don’t have the disease.

\[ Pr(\text{I fail the test}|\text{I don’t have the disease}) \]

I want to know the *probability* that I have the disease *given* that I failed the test.

\[ Pr(\text{I have the disease}|\text{I fail the test}) \]
Base Rate Fallacy
Base Rate Fallacy
Base Rate Fallacy
Base Rate Fallacy
Base Rate Fallacy
Base Rate Fallacy
Base Rate Fallacy
The importance of asking the right question.
I want to know the probability that I have the disease given that I failed the test.

\[ Pr(\text{I have the disease}|\text{I fail the test}) \]

I do need to know the probability that I failed the test given that I have the disease.

\[ Pr(\text{I fail the test}|\text{I have the disease}) \]

I also need to know the probability that I have the disease.

\[ Pr(\text{I have the disease}) \]

I also need to know the probability that I failed the test.

\[ Pr(\text{I fail the test}) \]
Bayes Theorem

\[ Pr(\text{I have the disease}|\text{I fail the test}) = \frac{Pr(\text{I fail the test}|\text{I have the disease}) \cdot Pr(\text{I have the disease})}{Pr(\text{I fail the test})} \]

Even T-shirts!

https://www.google.com/search?site=&tbm=isch&source=hp&biw=1241&bih=518&q=bayes+theorem+t-shirt&oq=bayes+theorem+t-shirt&gs_l=img.3...371.4856.0.4955.21.7.0.9.9.0.231.555.0j1j2.3.0....0...1ac.1.64.img..15.6.580.yrdHV_w79w
Base Rate Fallacy

• Many interesting examples
  – Psychology
  – Criminal justice
  – Etc.

• It can be explained and understood without a large amount of background
Many other problems that make great classroom examples

Problems that come up all the time.

- Biased sampling
- Misleading comparisons
- Etc.

As well as many excellent resources


With a classic being

- How to Lie with Statistics by Huff and Geis
  - http://www.amazon.com/dp/0393310728/?tag=mh0b-20&hvadid=3482023245&hvqmt=p&hvbnmt=bp&hvdev=c&ref=pd_sl_7f6adld6g8_p
What really gets me annoyed...

Lying with data visualizations

The numbers of men present are represented by the widths of the colored zones at a rate of one millimeter for every ten-thousand men; they are further written across the zones. The red [now brown] designates the men who enter into Russia, the black those who leave it.

Or anything by Edward Tufte
Practical advice

I hear and I forget.
I see and I remember.
I do and I understand.
- Confucius
The three keys

• In my experience there are three things that make for a great Data Science learning experience for students
The three keys

- In my experience there are three things that make for a great Data Science learning experience for students
- Working on a project actually using data!
The three keys

- In my experience there are three things that make for a great Data Science learning experience for students
  - Working on a project actually using data!
  - Working on a project actually using data!
The three keys

- In my experience there are three things that make for a great Data Science learning experience for students
  - Working on a project actually using data!
  - Working on a project actually using data!
  - Working on a project actually using data!
- And not listening to the pony-tailed professor blather on :-)

WPI
But, how do you actually do it?
You have access to an amazing wealth of information

- One of, if not the, most important single slides in the talk!
  - Many data sets on many subjects
  - Where I got the data sets for the workshop
- https://www.kaggle.com/
  - The best site on the Internet for Data Science competitions
- http://www.data.gov/
  - The U.S. government data repository
  - Many data sets of social importance from around the work
- And MANY, MANY more
A great example to work with, especially for a short project.

- http://www.gapminder.org
Python!
Oh, sorry… I meant Python!

https://upload.wikimedia.org/wikipedia/commons/4/4d/Ball_python_lucy.jpg
Python history

• First release in 1994
• More widespread use since version 1.5 in 1997.
• I been using Python since close to the start.
  • The code for my thesis was mostly in Python and that was 1999!

http://www.bennorthrop.com/Essays/2016/old-programmer.jpg
Why Python?

- Great programming model
  - Easy to learn
- Access to other languages
  - Can call C, C++, JAVA, etc.
- Lots of libraries
  - Numpy, scipy, pandas, pycuda, mpi4py, etc.
- Lots of great online resources!
- It is free!
You have to be careful though, the Internet is a dangerous place...

- https://www.youtube.com/watch?v=EUEHOYI0mRg
Batteries included...

- The default Python installation includes a vast library of functionality
  - [https://docs.python.org/3/library/](https://docs.python.org/3/library/)

![Batteries included!](https://upload.wikimedia.org/wikipedia/commons/6/68/Python_batteries_included.jpg)
Easy to install

• Anaconda!
  • Yes, there are far too many snake jokes…

• https://www.continuum.io/downloads
  • Or just search for “anaconda python” in Google.
Three main ways to run Python

- The default Python prompt
  - Either at the command line or running scripts
- The IPython interpreter
  - Either at the command line or running scripts
- The Jupyter notebook
  - A beautiful integrated development environment (IDE) for Python
- There are also many others
  - Such as Spyder, which is included with Anaconda
Quick example
Questions!?