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

WPI

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 accesible 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?

WPI

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!

OAK RIDGE

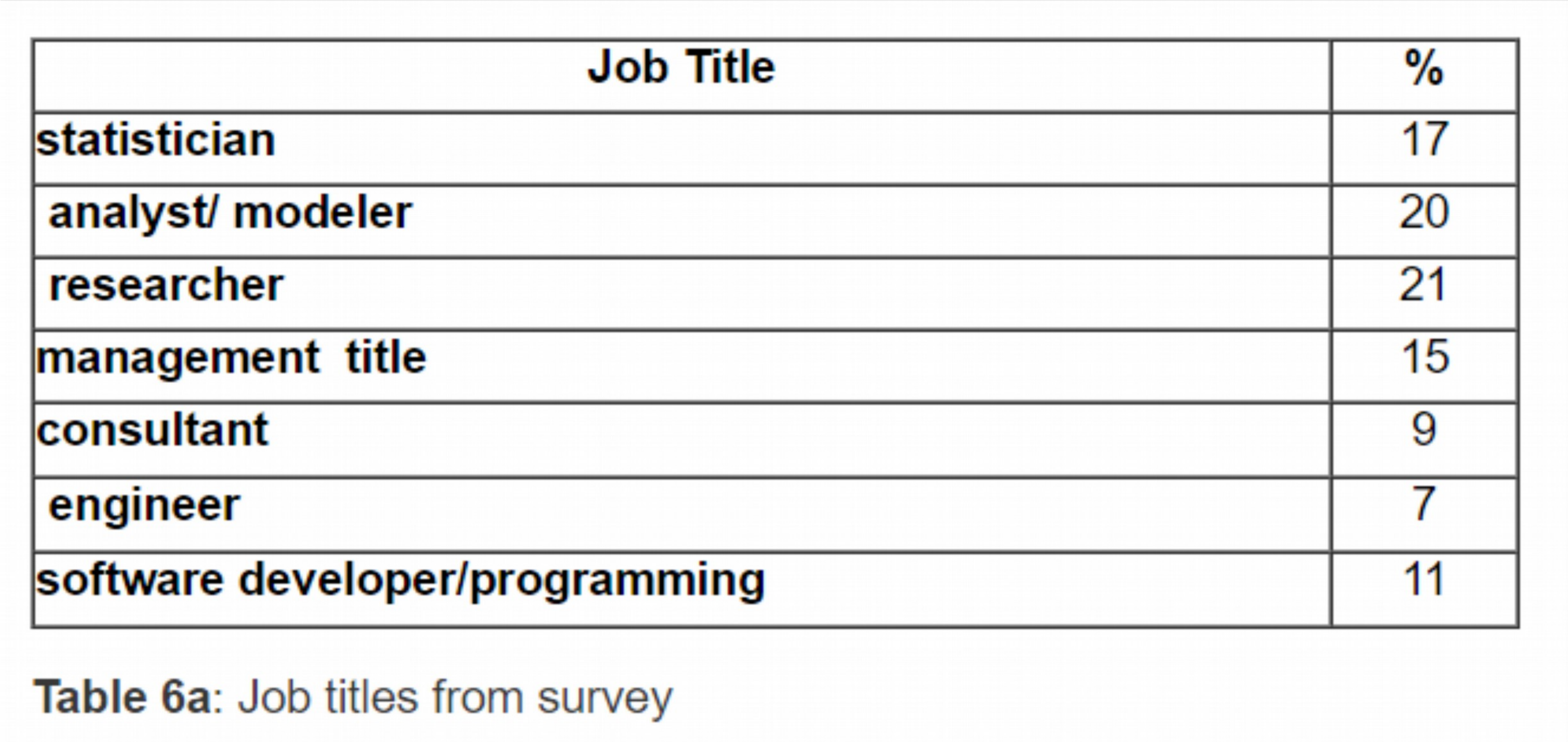
National Laboratory

MITRE

Google

Job titles can be deceiving...

http://www.everett-mead.co.uk/insurance-its-a-confusing-old-game/



Job Title	%
statistician	17
analyst/ modeler	20
researcher	21
management title	15
consultant	9
engineer	7
software developer/programming	11

Table 6a: Job titles from survey

A current really hot area
Data Science

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/

Data Science

Vs

Data Literacy

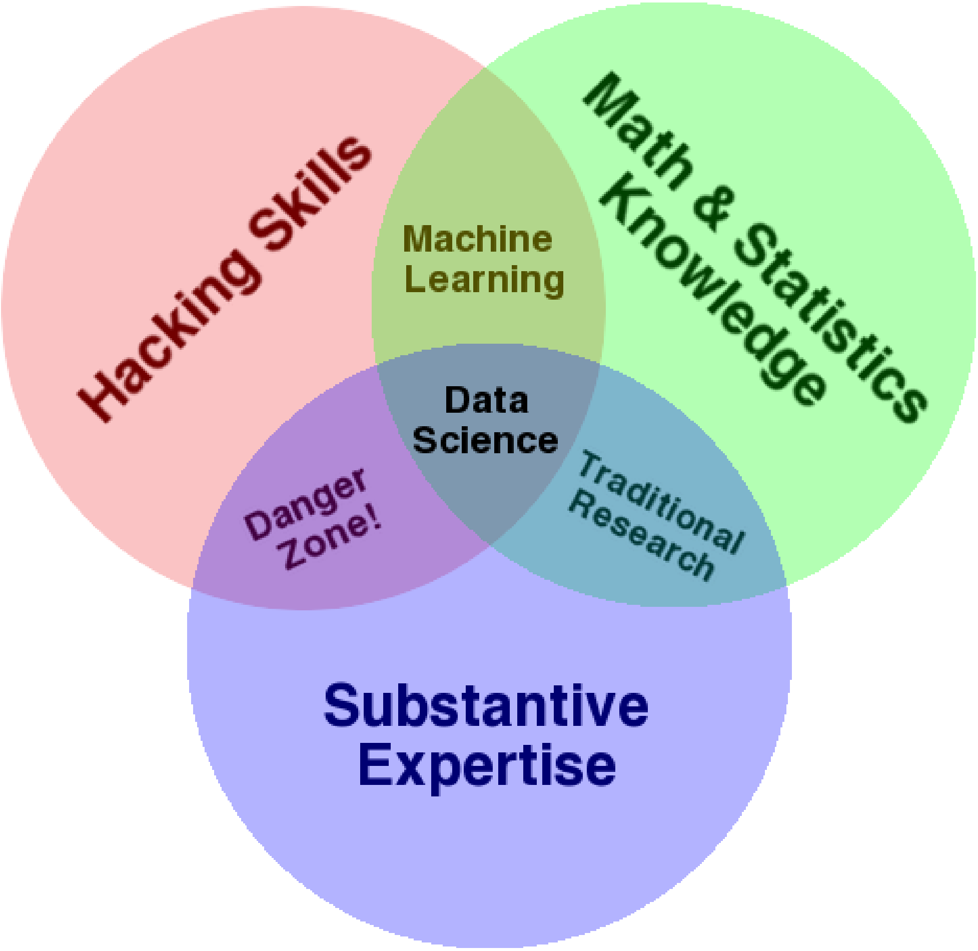
Vs

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?

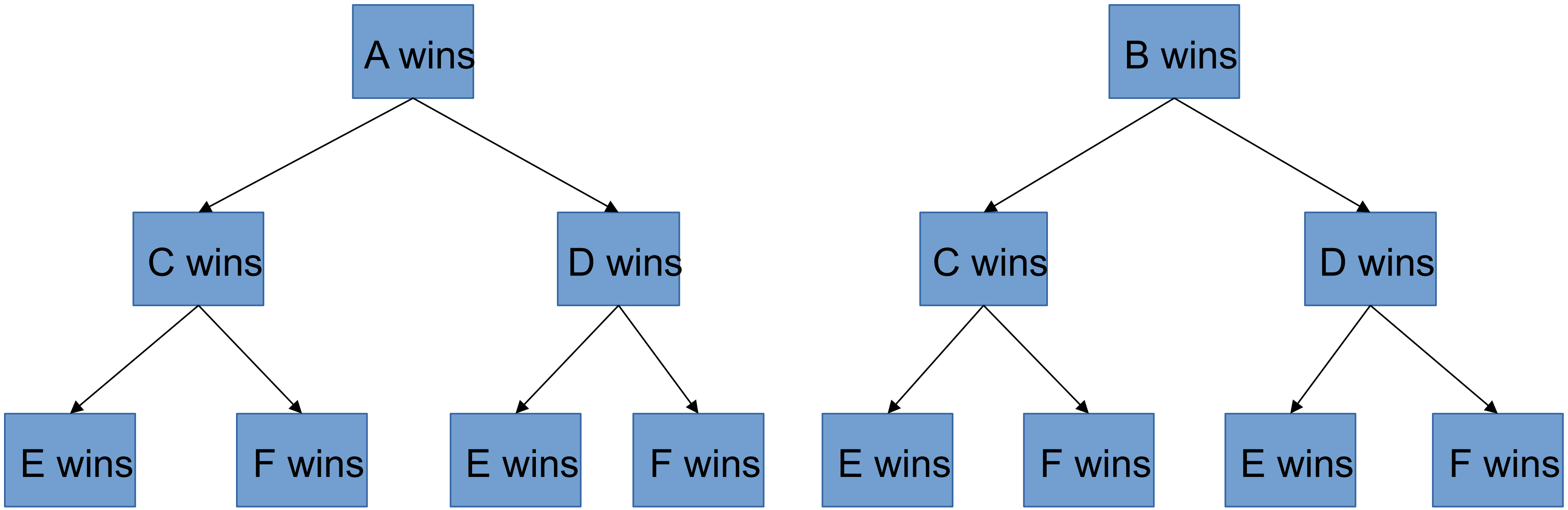


- Based upon Drew Conway's Data Science Venn Diagram
  - http://en.wikipedia.org/wiki/Data_science
  - http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Data Literacy – Your turn to work
Warm up problem

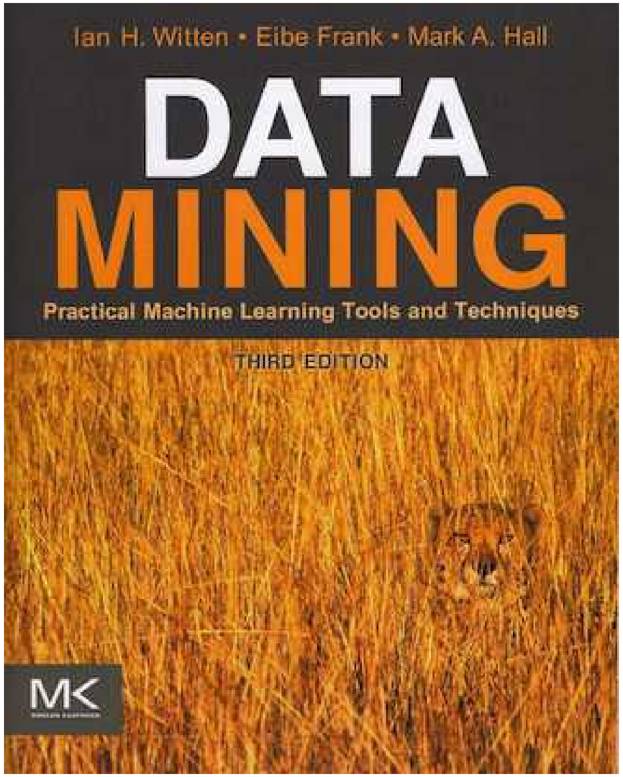
"2013 Alabama A-Day spring football game" by Patriarca12 - Own work. Licensed under CC BY 3.0 via Wikimedia Commons - https://commons.wikimedia.org/wiki/File:2013_Alabama_A-Day_spring_football_game.jpg#/media/File:2013_Alabama_A-Day_spring_football_game.jpg

Data Literacy – Warm up problem



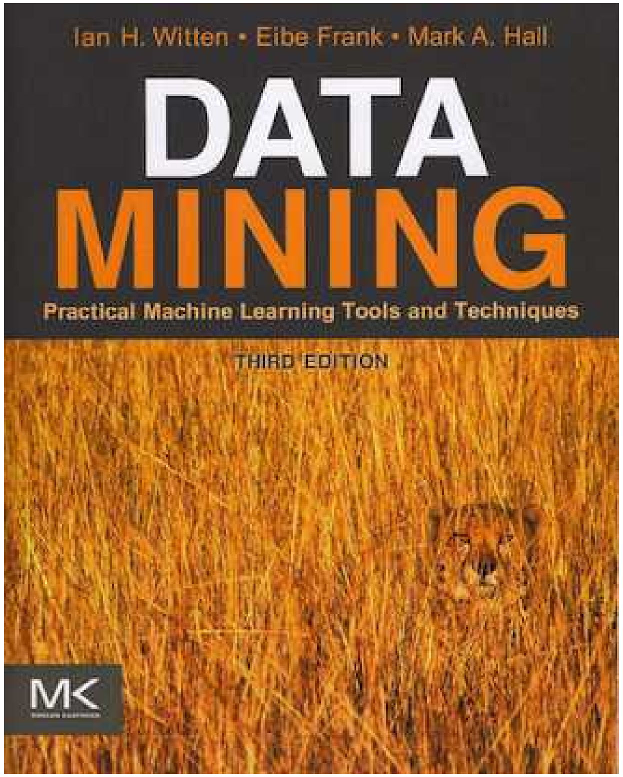
It can be easy to fool yourself!

Human beings are really good at pattern detection...

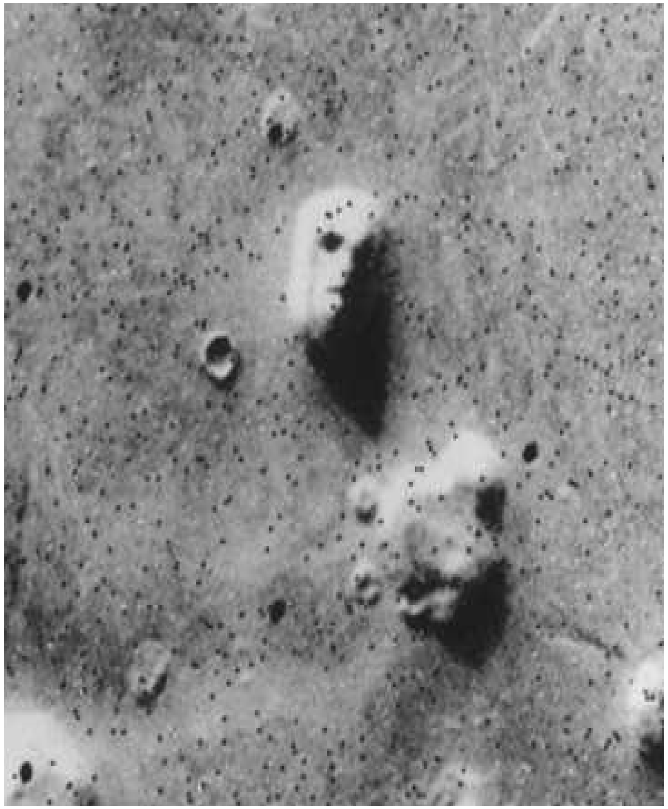

Ian H. Witten • Eibe Frank • Mark A. Hall
DATA MINING
Practical Machine Learning Tools and Techniques
THIRD EDITION
MK

It can be easy to fool yourself!

Human beings are really good at pattern detection...

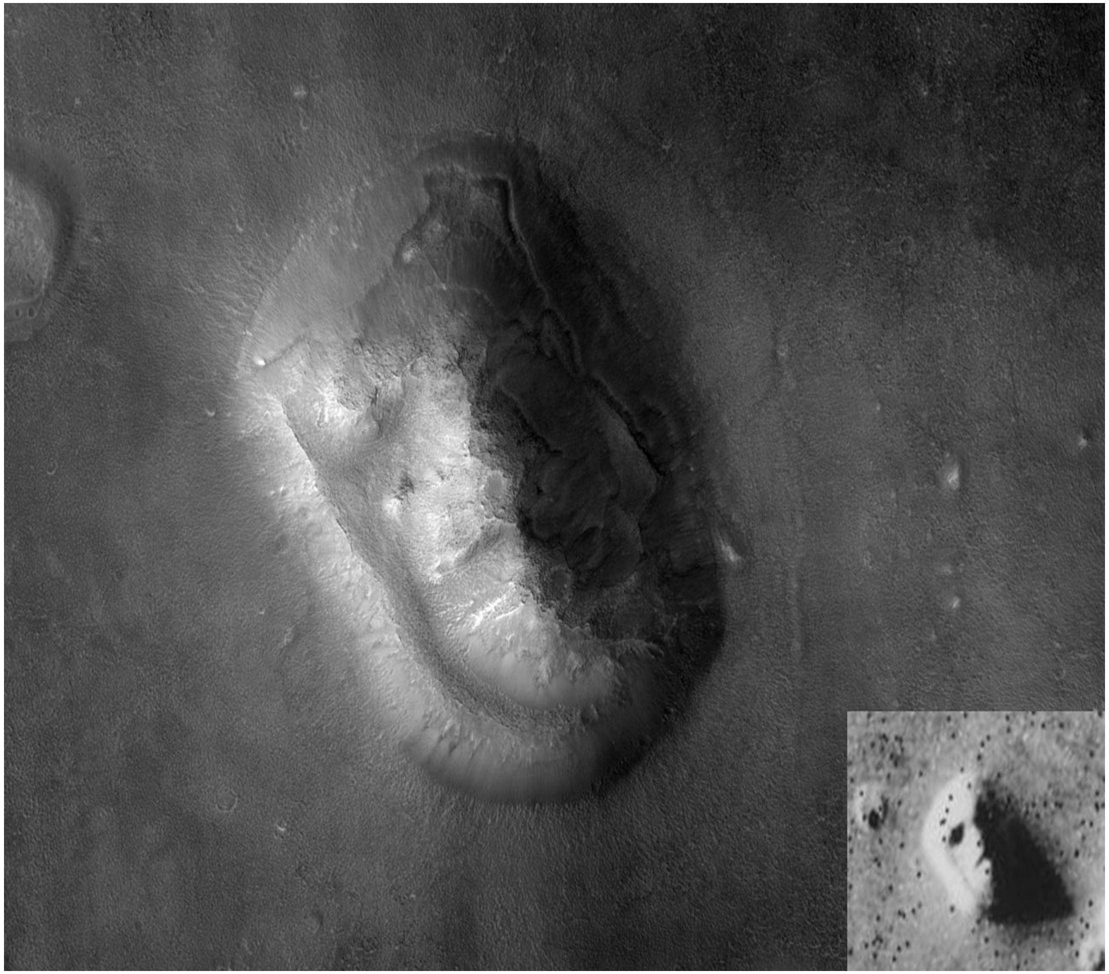

Ian H. Witten • Eibe Frank • Mark A. Hall
DATA MINING
Practical Machine Learning Tools and Techniques
THIRD EDITION
MK

Perhaps a bit too good!



http://en.wikipedia.org/wiki/Cydonia_(region_of_Mars)

It can be easy to fool yourself!



http://en.wikipedia.org/wiki/Cydonia_(region_of_Mars)

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.

Base Rate Fallacy

- 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

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?

Licensed under Public Domain via Wikimedia Commons - https://commons.wikimedia.org/wiki/File:US_Navy_060105-N-8154G-010_A_hospital_corpsman_with_the_Blood_Donor_Team_from_Portsmouth_Naval_Hospital_takes_samples_of_blood_from_a_donor_for_testing.jpg#/media/File:US_Navy_060105-N-8154G-010_A_hospital_corpsman_with_the_Blood_Donor_Team_from_Portsmouth_Naval_Hospital_takes_samples_of_blood_from_a_donor_for_testing.jpg

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.

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

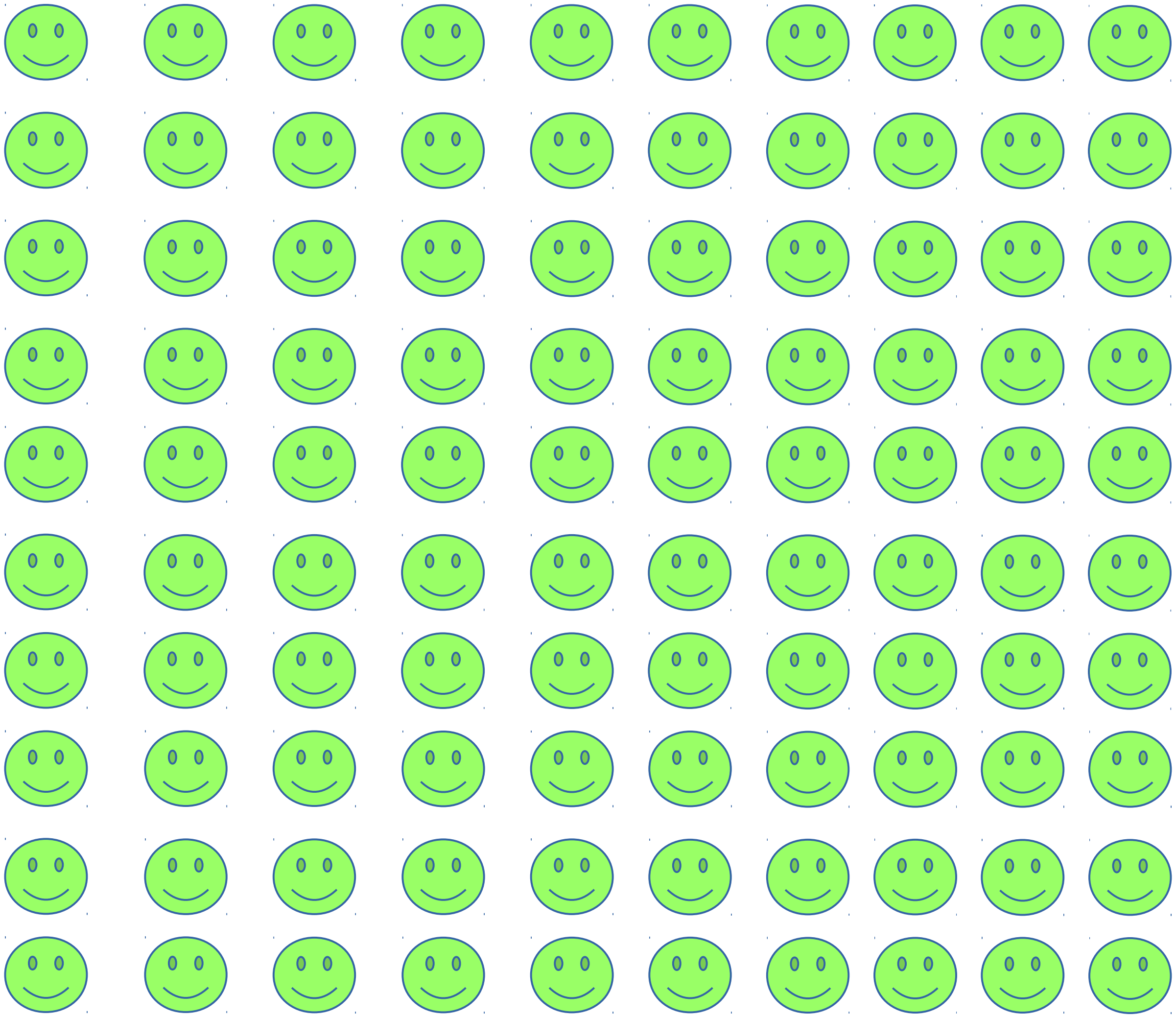
I was told the *probability* that I failed the test *given* that I don't have the disease.

$$Pr(\text{I fail the test}|\text{I don't have the disease})$$

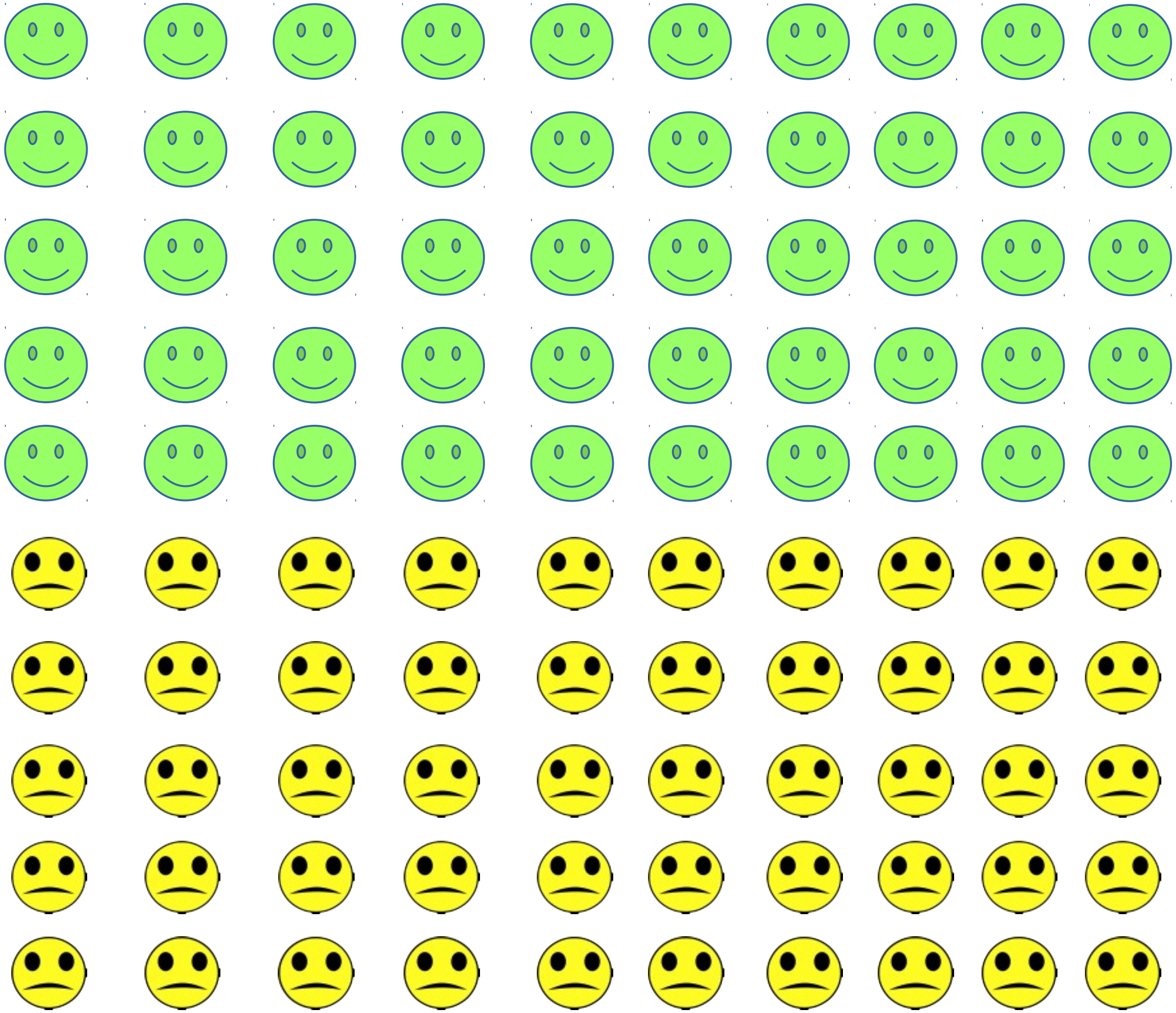
I want to know the *probability* that I have the disease *given* that I failed the test.

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

Base Rate Fallacy

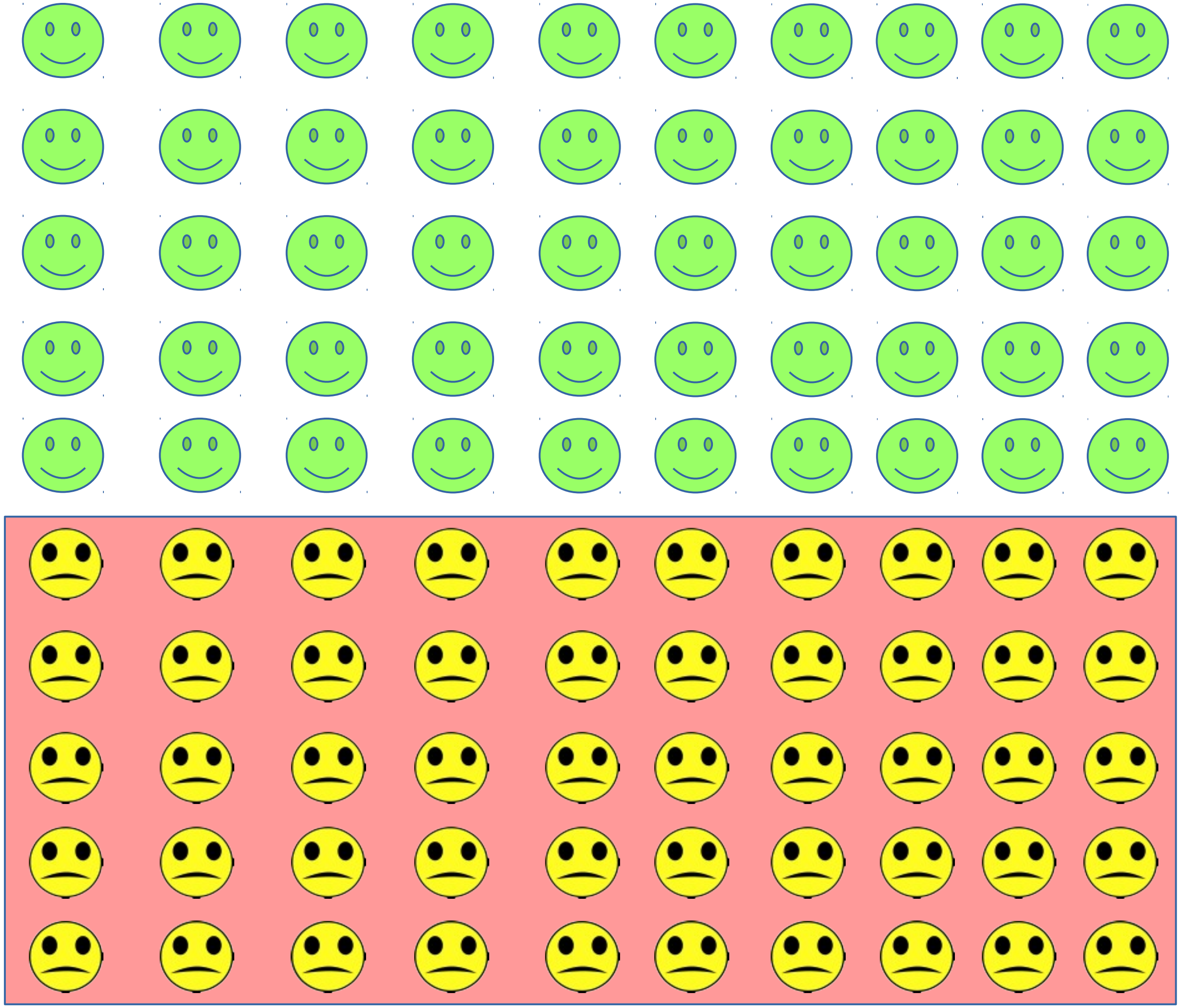


Base Rate Fallacy



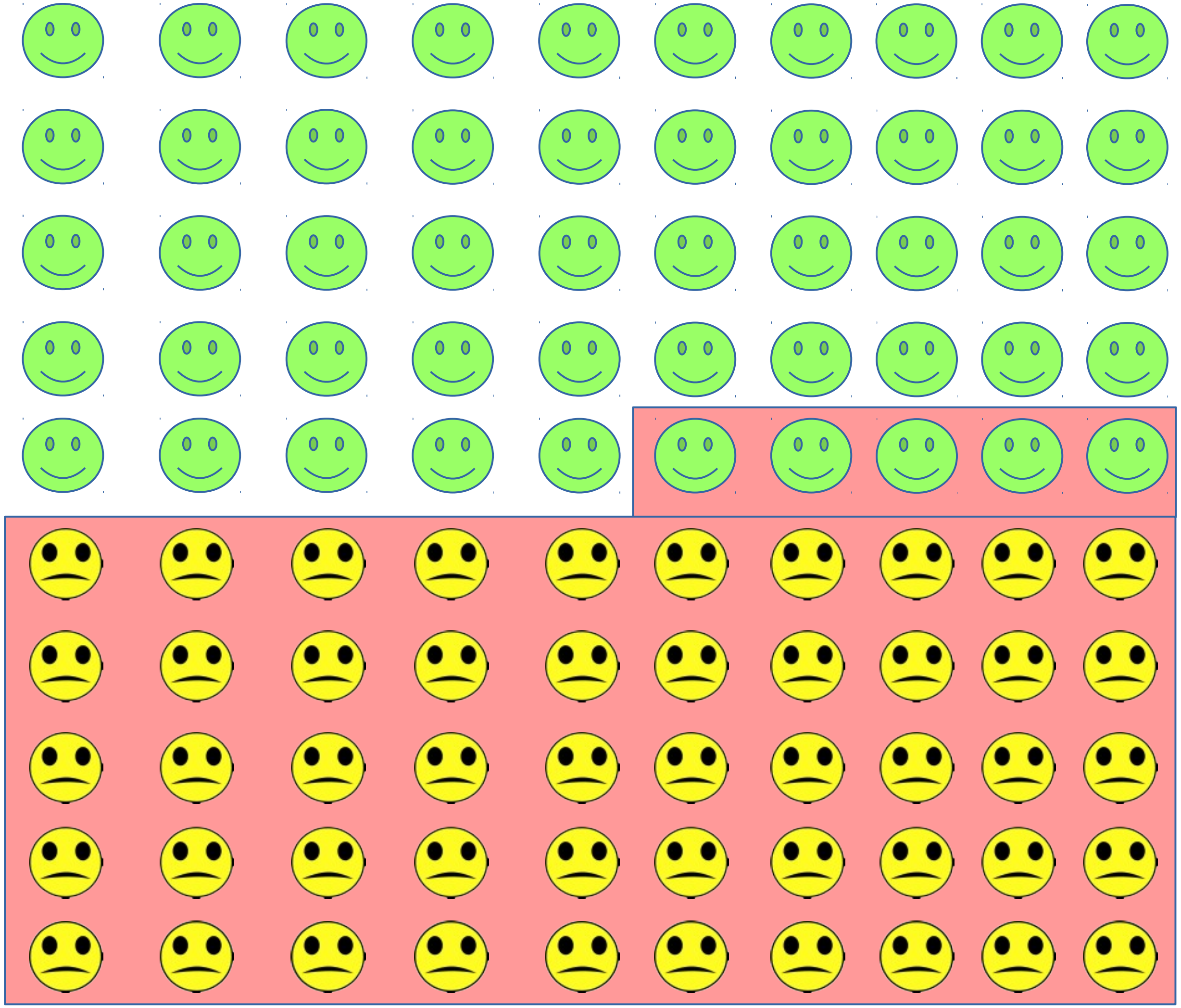
By Jgsho (Own work) [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons

Base Rate Fallacy



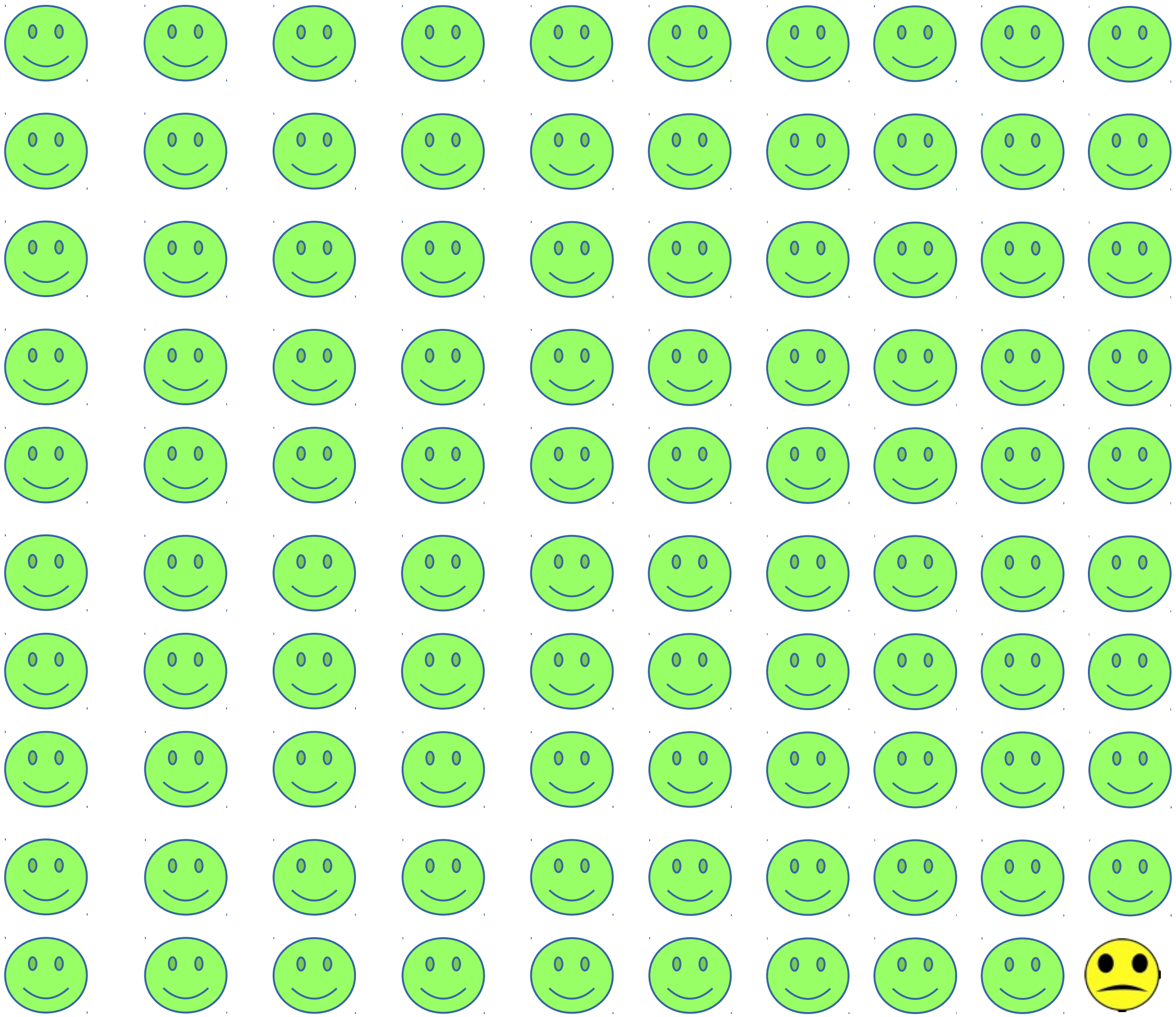
By Jgsho (Own work) [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons

Base Rate Fallacy



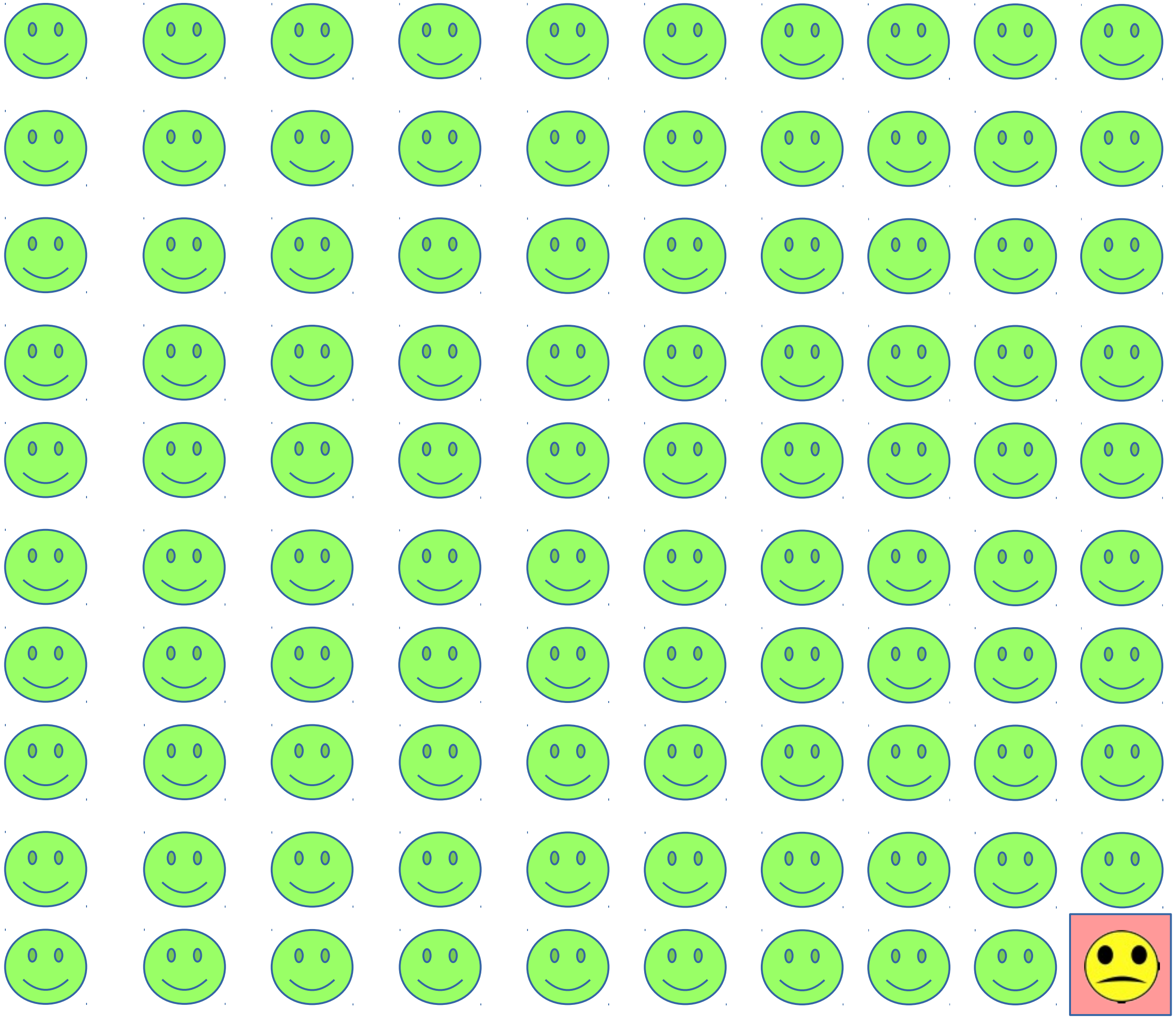
By Jgsho (Own work) [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons

Base Rate Fallacy

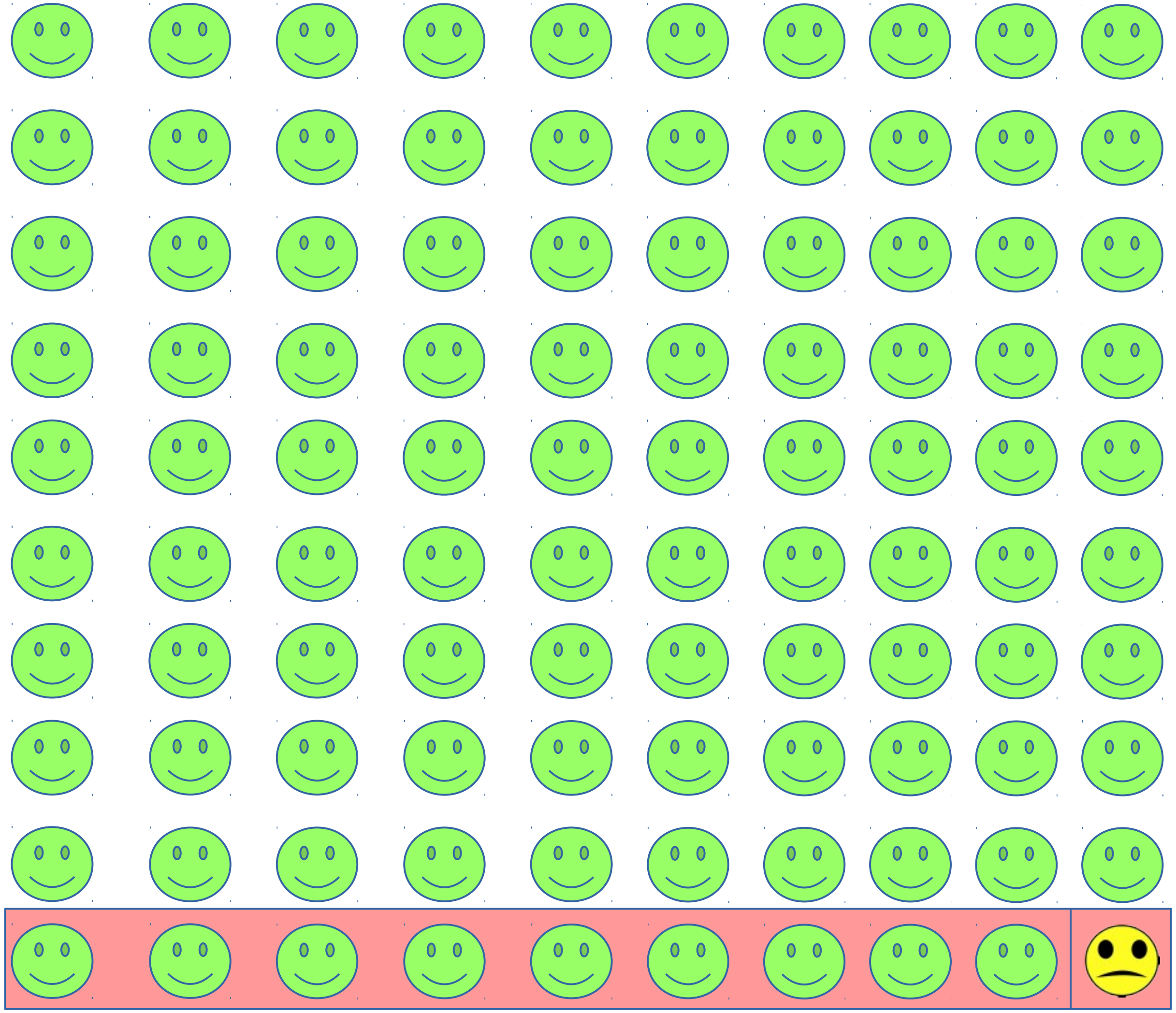


By Jgsho (Own work) [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons

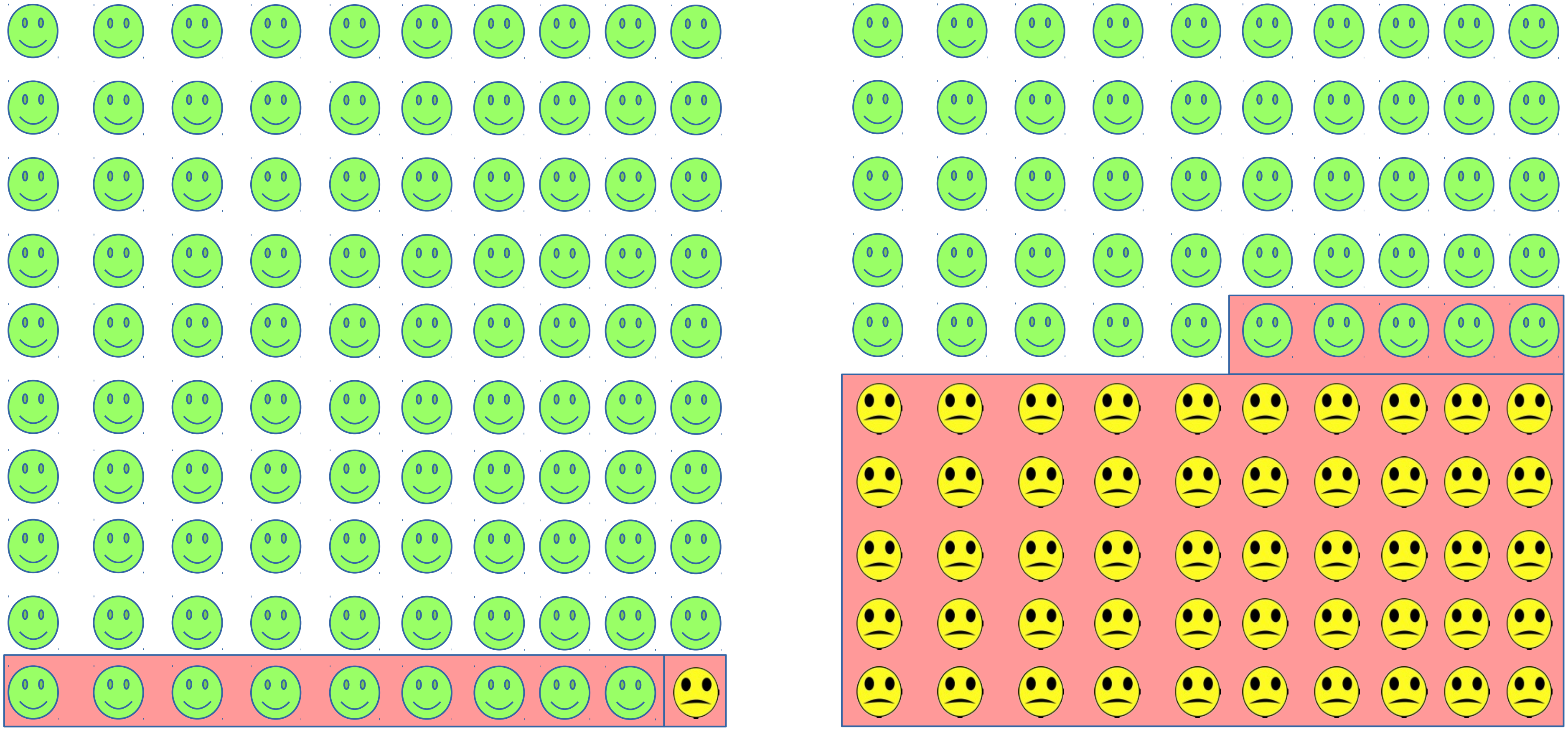
Base Rate Fallacy



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.

$$Pr(\text{I have the disease})$$

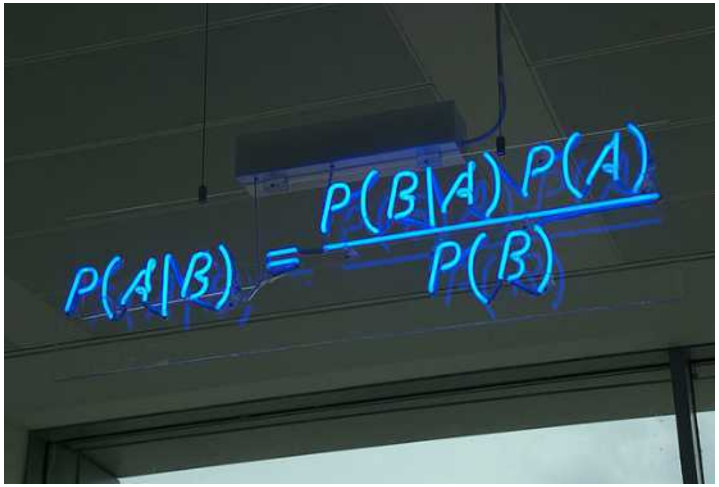
I also need to know the *probability* that I failed the test.

$$Pr(\text{I fail the test})$$

Bayes Theorem

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

"Thomas Bayes" by unknown - [2][3]. Licensed under Public Domain via Wikimedia Commons - https://commons.wikimedia.org/wiki/File:Thomas_Bayes.gif#/media/File:Thomas_Bayes.gif



"Bayes' Theorem MMB 01" by mattbuck (category) - Own work by mattbuck.. Licensed under CC BY-SA 3.0 via Wikimedia Commons - https://commons.wikimedia.org/wiki/File:Bayes%27_Theorem_MMB_01.jpg#/media/File:Bayes%27_Theorem_MMB_01.jpg

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.yrkdHV_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

- http://www.stat.columbia.edu/~gelman/bag-of-tricks/chap10.pdf
- http://cseweb.ucsd.edu/~ricko/CSE3/Lie_with_Statistics.pdf

With a classic being

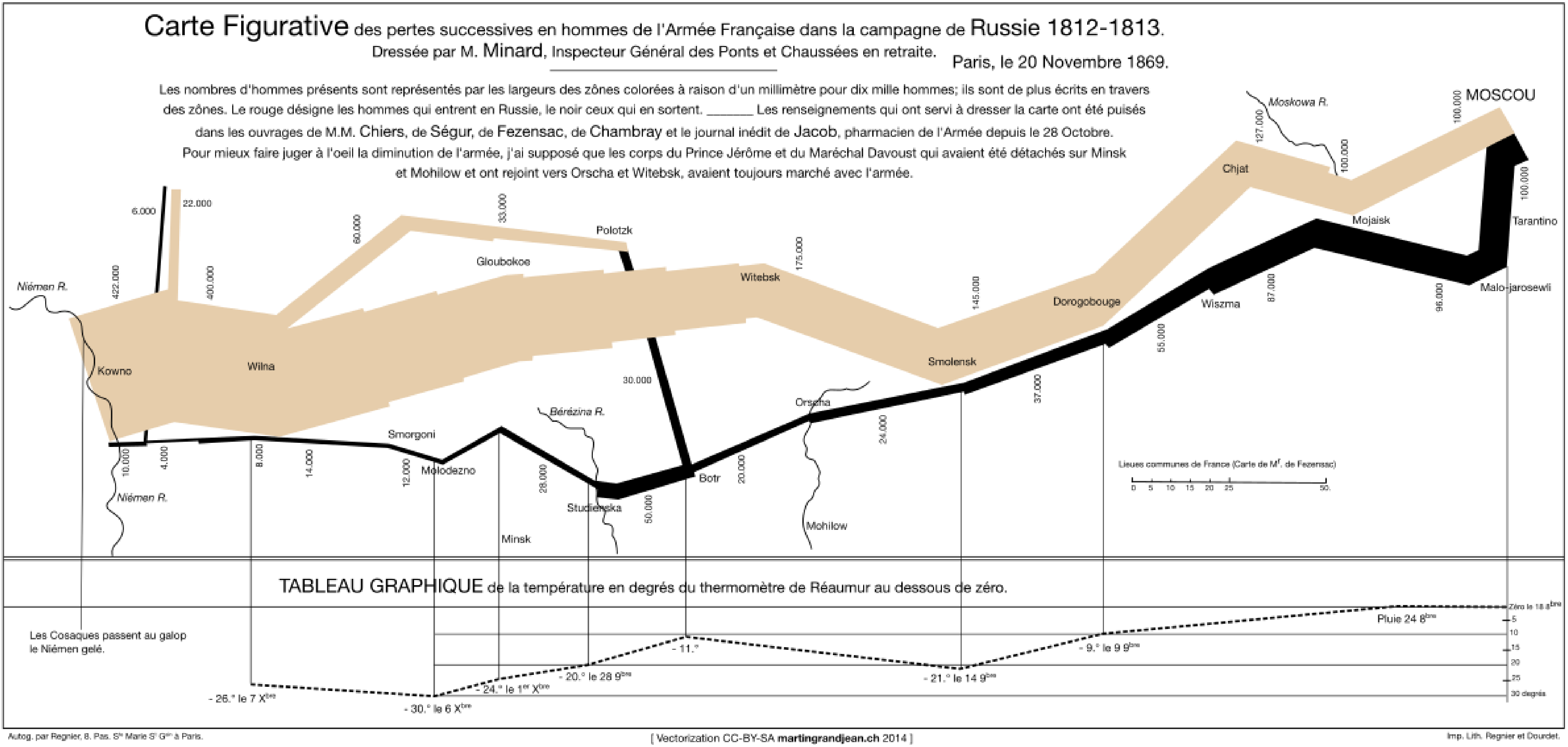
- How to Lie with Statistics by Huff and Geis
  - http://www.amazon.com/dp/0393310728/?tag=mh0b-20&hvadid=3482023245&hvqmt=p&hvbmt=bp&hvdev=c&ref=pd_sl_7f6adld6g8_p

What really gets me annoyed...

Lying with data visualizations

- http://gizmodo.com/how-to-lie-with-data-visualization-1563576606

But, nothing is more beautiful when you get it right!



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.

"Minard's Map (vectorized)" by Calvinius - Own work http://www.martingrandjean.ch/historical-data-visualization-minard-map/. Licensed under Creative Commons Attribution-Share Alike 3.0 via Wikimedia Commons - http://commons.wikimedia.org/wiki/File:Minard%27s_Map_(vectorized).svg#mediaviewer/File:Minard%27s_Map_(vectorized).svg

Or anything by Edward Tufte

- https://en.wikipedia.org/wiki/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 :-)

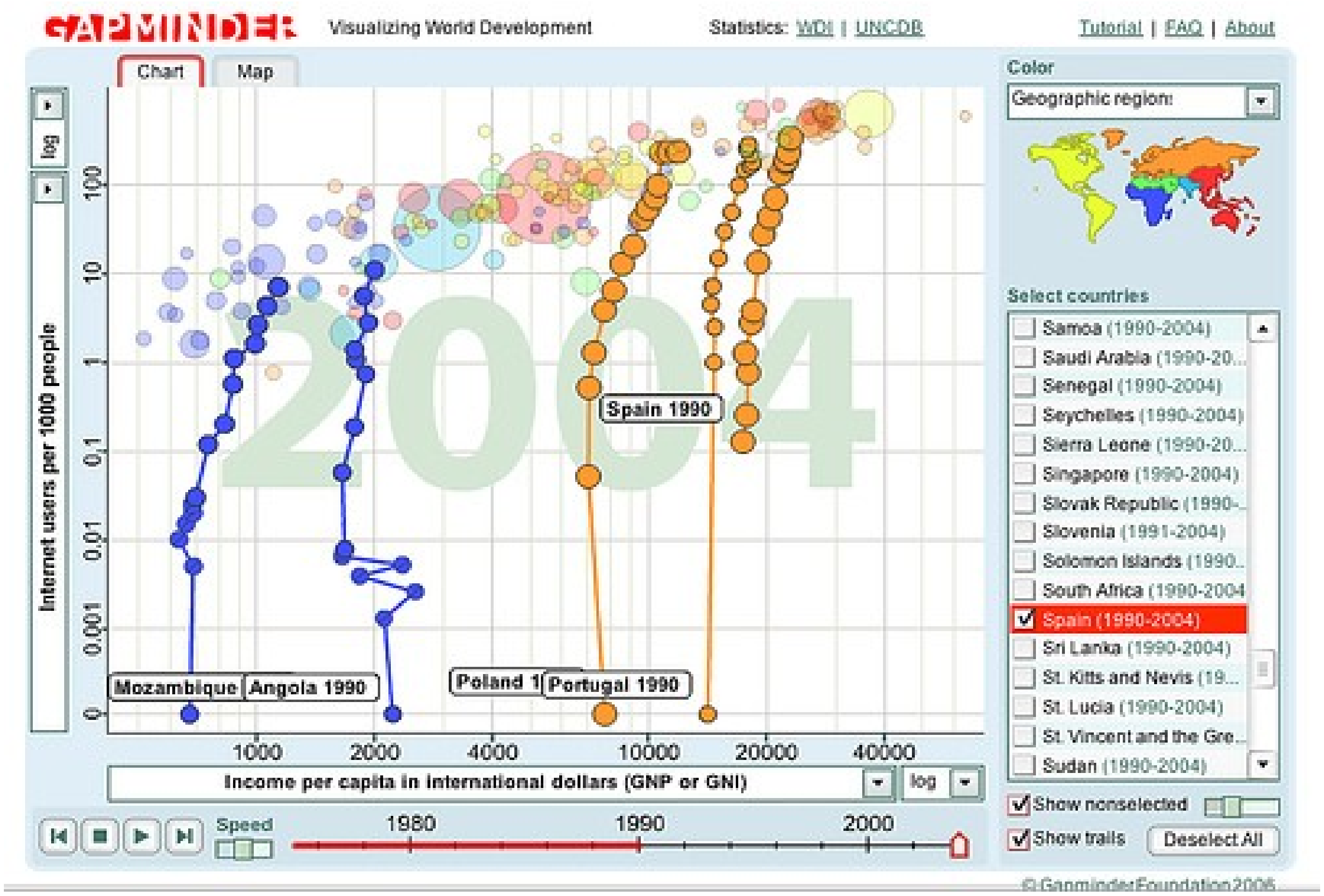
But, how do you actually do it?

You have access to an **amazing** wealth of information

- **One of, if not the, most important single slides in the talk!**
- http://archive.ics.uci.edu/ml/
  - 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
- http://databank.worldbank.org/data/home.aspx
  - Many data sets of social importance from around the work
- And MANY, MANY more

A great example to work with, especially fot a short project.

- http://www.gapminder.org



Color: Geographic region

Income per capita in international dollars (GNP or GNI) | log

- □ Samoa (1990-2004)
- □ Saudi Arabia (1990-20...
- □ Senegal (1990-2004)
- □ Seychelles (1990-2004)
- □ Sierra Leone (1990-20...
- □ Singapore (1990-2004)
- □ Slovak Republic (1990-...
- □ Slovenia (1991-2004)
- □ Solomon Islands (1990...
- □ South Africa (1990-2004
- ☑ Spain (1990-2004)
- □ Sri Lanka (1990-2004)
- □ St. Kitts and Nevis (19...
- □ St. Lucia (1990-2004)
- □ St. Vincent and the Gre...
- □ Sudan (1990-2004)

☑ Show nonselected

☑ Show trails

Python!

https://upload.wikimedia.org/wikipedia/commons/4/4d/Ball_python_lucy.JPG

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?

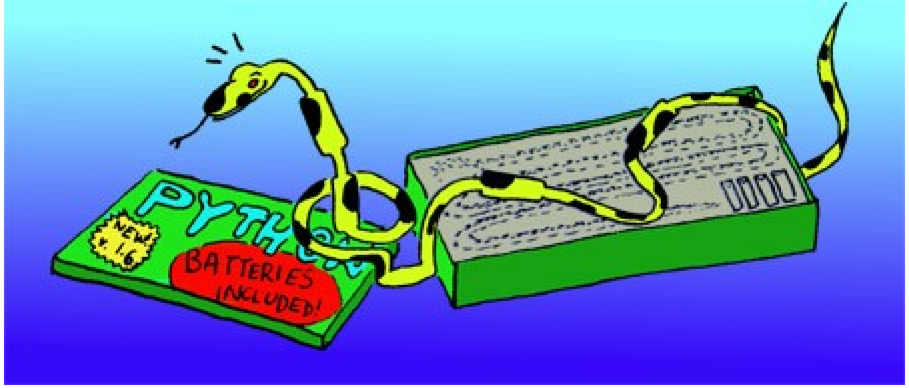
- Great programming model
  - Easy to learn
- Access to other languages
  - Can call C, C++, JAVA, etc.
- Lots of libraries
  - Numpy, scipy, pandas, pycuda, mpi4py, etc.
- Lots of great online resources!
- It is free!

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://upload.wikimedia.org/wikipedia/commons/6/68/Python_batteries_included.jpg

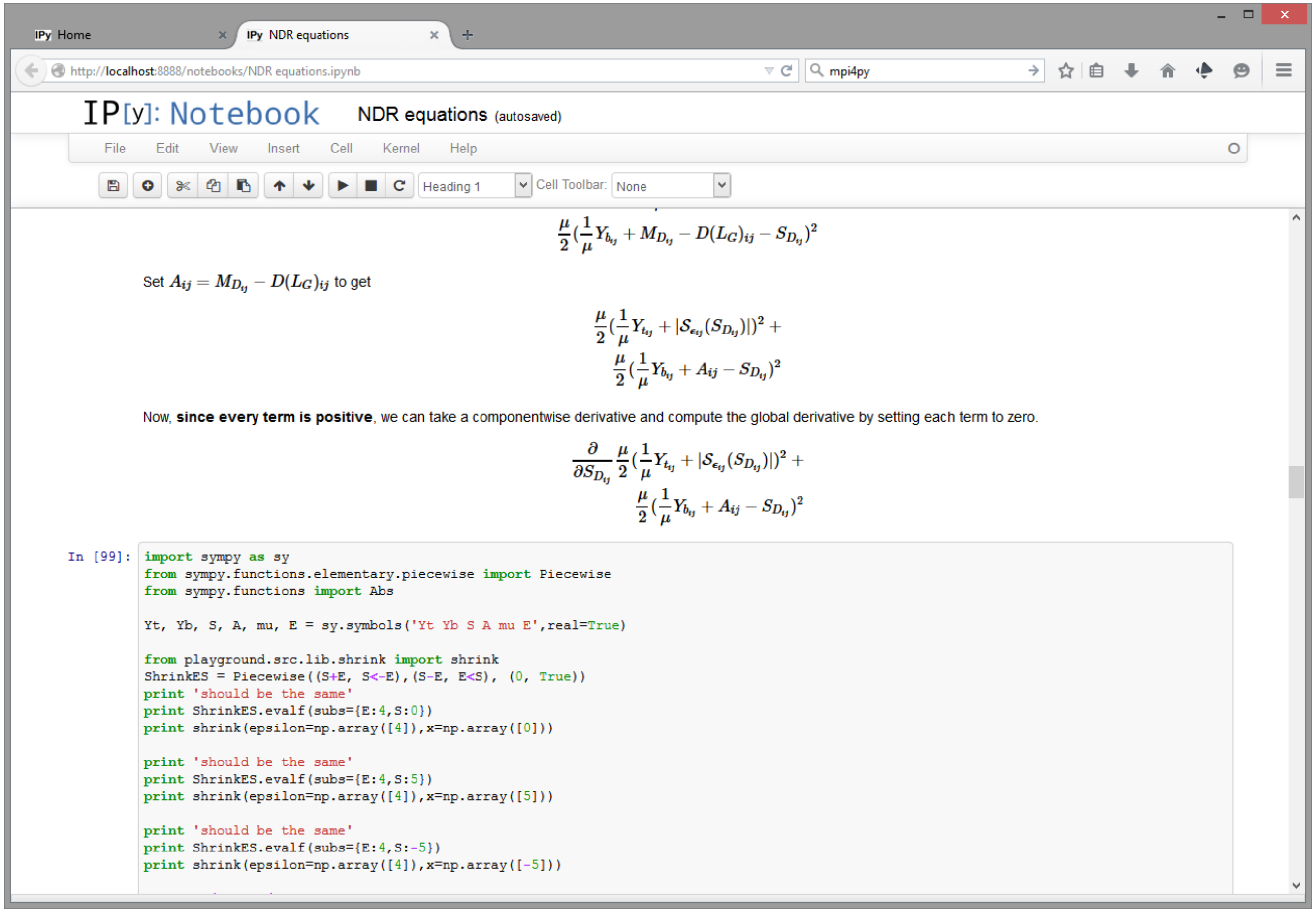
Easy to install

- 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 interpretor
  - 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 Anacoda

Quick example



$$\frac{\mu}{2}(\frac{1}{\mu}Y_{b_{ij}} + M_{D_{ij}} - D(L_G)_{ij} - S_{D_{ij}})^2$$

Set $A_{ij} = M_{D_{ij}} - D(L_G)_{ij}$ to get

$$\frac{\mu}{2}(\frac{1}{\mu}Y_{t_{ij}} + |\mathcal{S}_{\epsilon_{ij}}(S_{D_{ij}})|)^2 +$$
$$\frac{\mu}{2}(\frac{1}{\mu}Y_{b_{ij}} + A_{ij} - S_{D_{ij}})^2$$

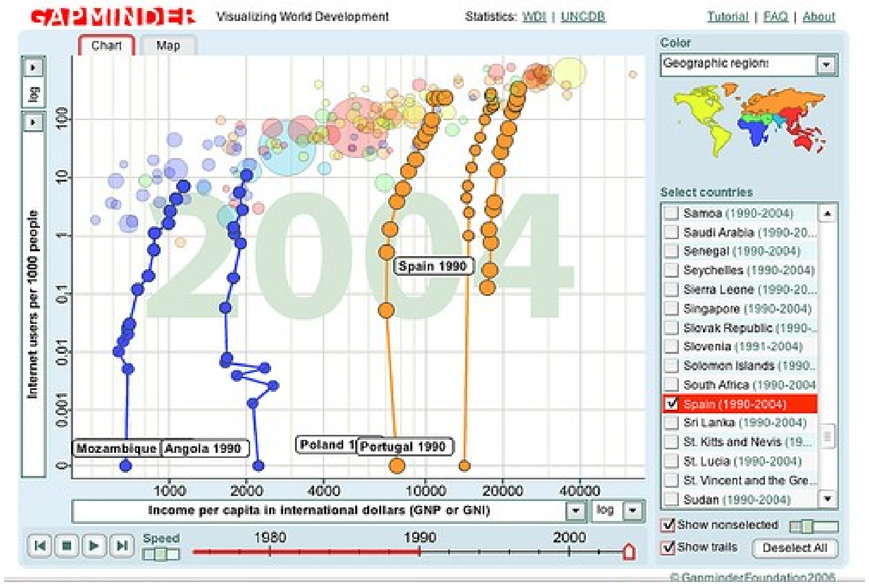
Now, **since every term is positive**, we can take a componentwise derivative and compute the global derivative by setting each term to zero.

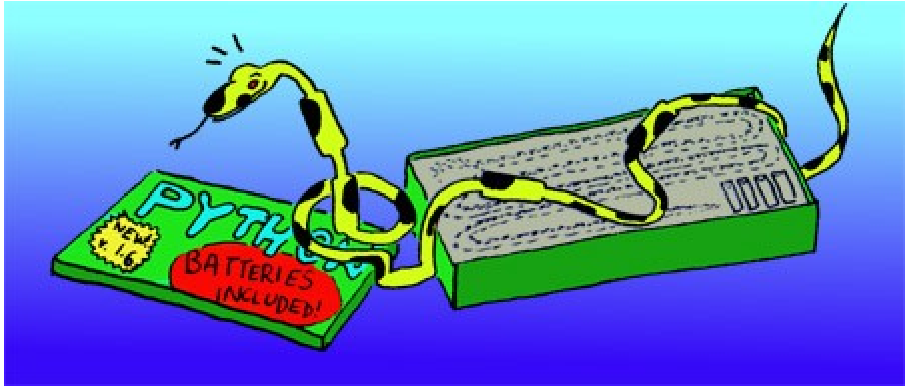
$$\frac{\partial}{\partial S_{D_{ij}}}\frac{\mu}{2}(\frac{1}{\mu}Y_{t_{ij}} + |\mathcal{S}_{\epsilon_{ij}}(S_{D_{ij}})|)^2 +$$
$$\frac{\mu}{2}(\frac{1}{\mu}Y_{b_{ij}} + A_{ij} - S_{D_{ij}})^2$$

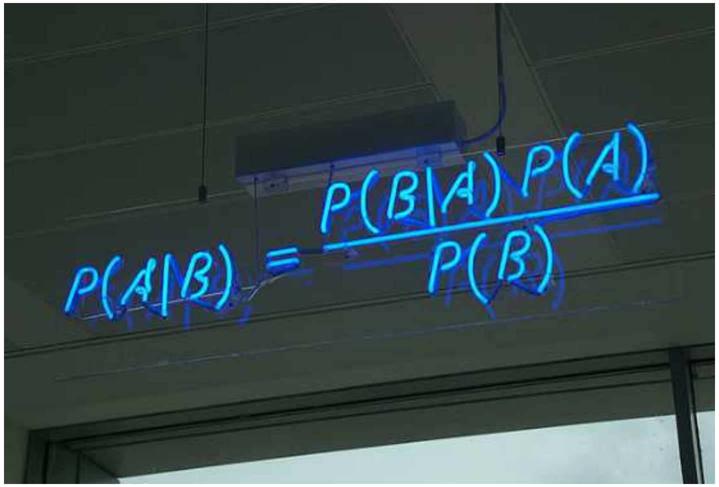
In [99]:

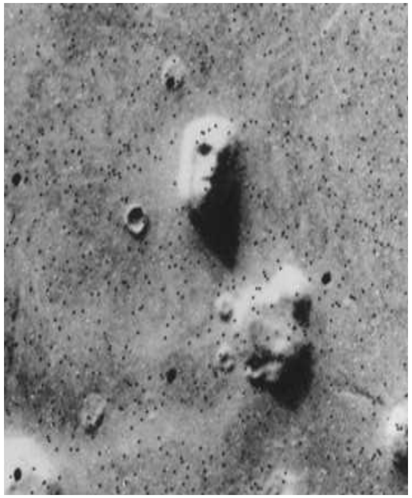
```
import sympy as sy
from sympy.functions.elementary.piecewise import Piecewise
from sympy.functions import Abs

Yt, Yb, S, A, mu, E = sy.symbols('Yt Yb S A mu E',real=True)

from playground.src.lib.shrink import shrink
ShrinkES = Piecewise((S+E, S<-E),(S-E, E<S), (0, True))
print 'should be the same'
print ShrinkES.evalf(subs={E:4,S:0})
print shrink(epsilon=np.array([4]),x=np.array([0]))

print 'should be the same'
print ShrinkES.evalf(subs={E:4,S:5})
print shrink(epsilon=np.array([4]),x=np.array([5]))

print 'should be the same'
print ShrinkES.evalf(subs={E:4,S:-5})
print shrink(epsilon=np.array([4]),x=np.array([-5]))
```

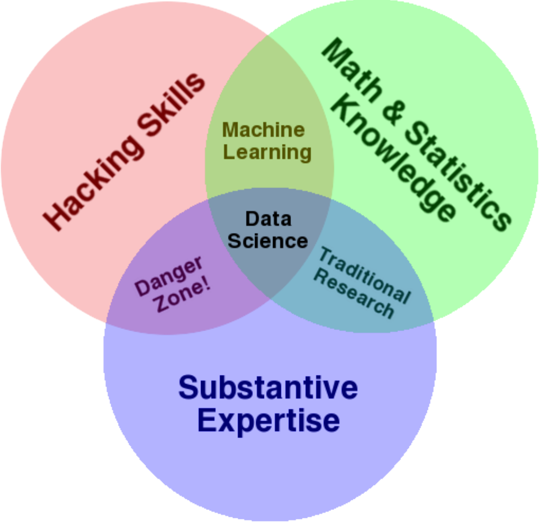
Questions!?










Hacking Skills

Math & Statistics Knowledge

Machine Learning

Data Science

Danger Zone!

Traditional Research

Substantive Expertise