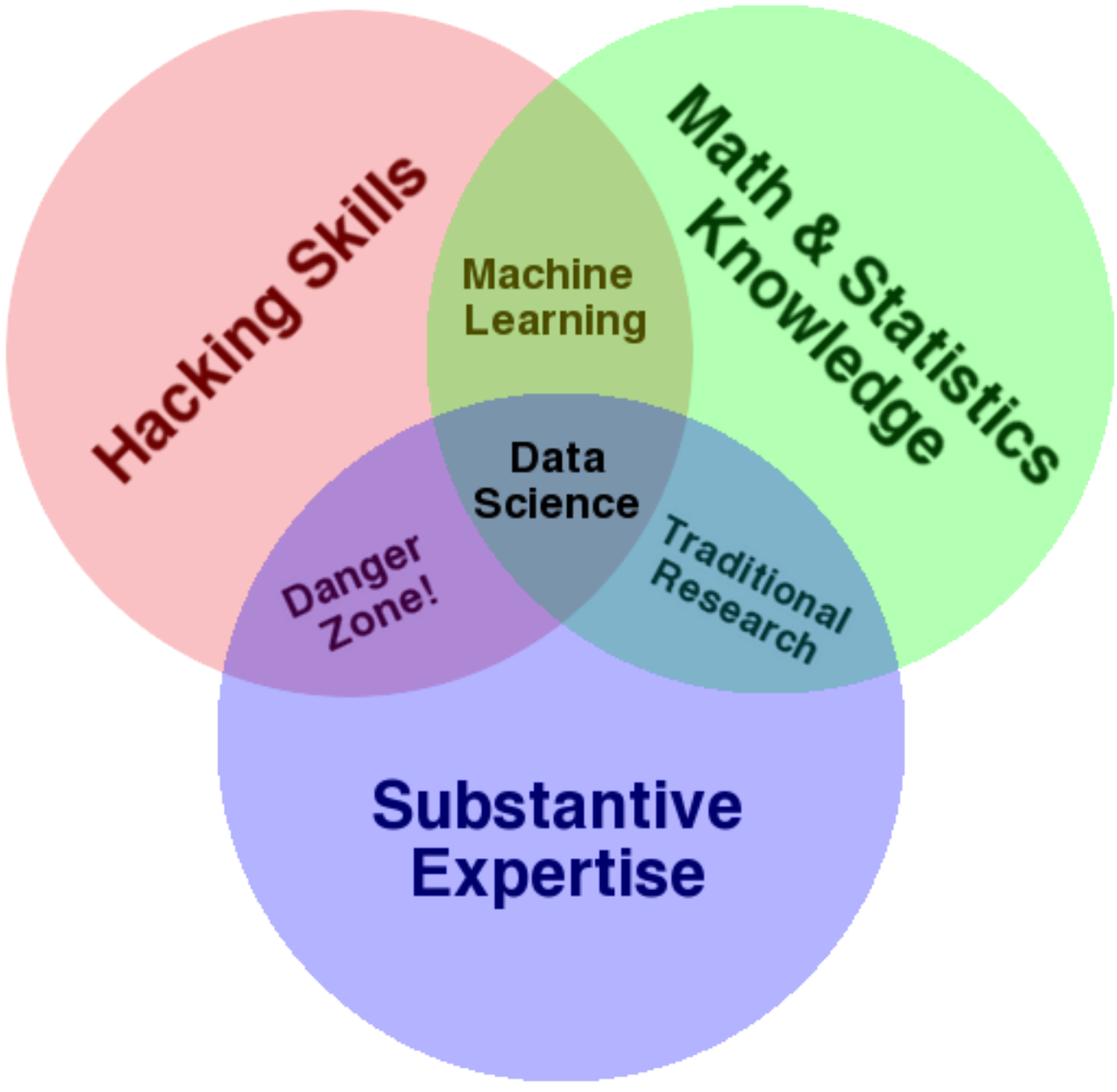
The Buzz Word Talk

Open Science, Big Data, Data Science and Bioinformatics

Bioinformatics and Data Science



Hacking Skills
(aka Programming)

- Programming is useful in dealing with “Big Data”

- We have a file of SNPs that is 736 Gb. That probably wouldn’t fit on any computer you have at home.

- We can't just use Excel to analyze our data

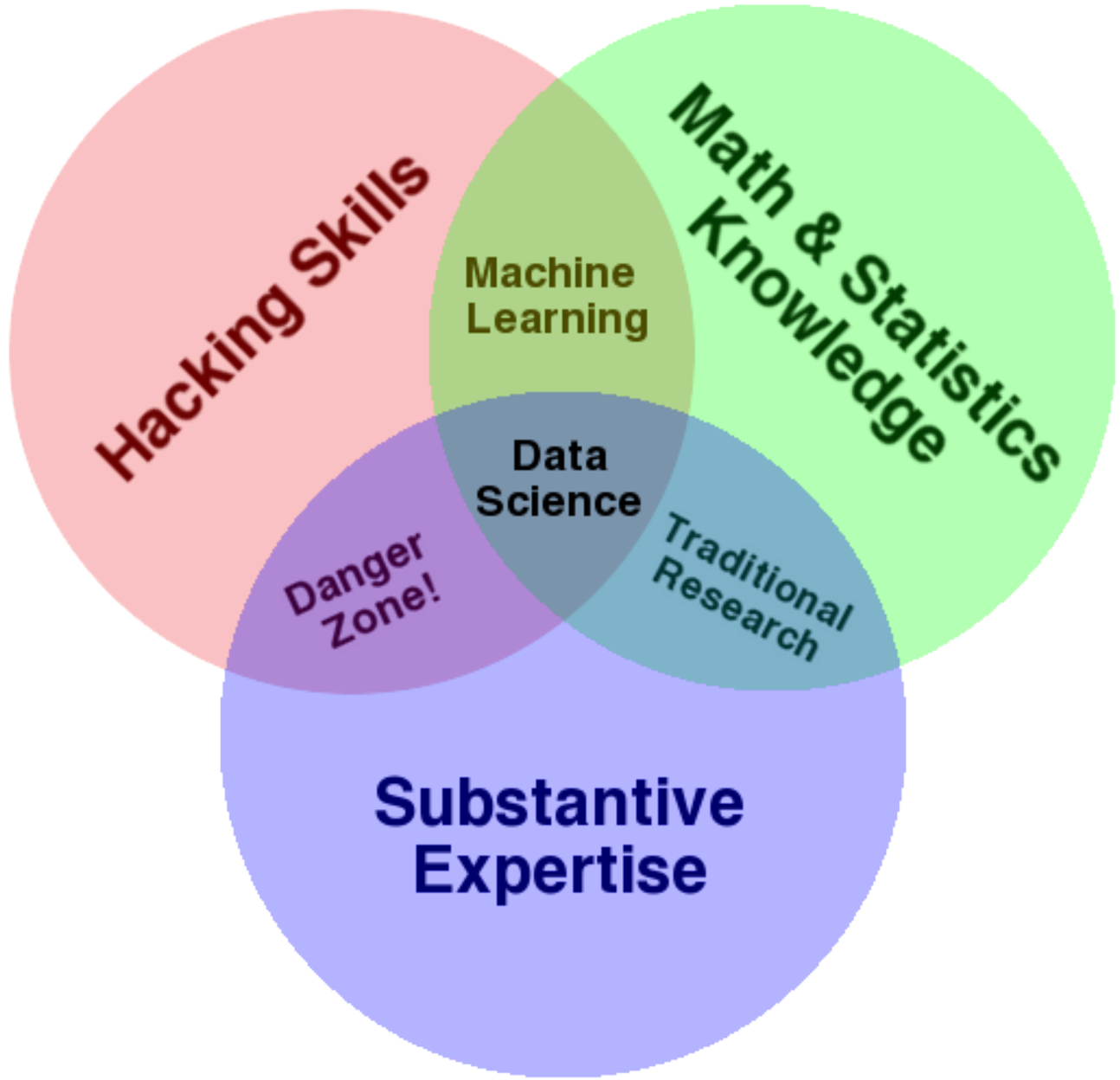

BIG DATA
dreamstime.com

Its Just Big

- Twitter - 100 Tb (?)
- Facebook - 30 Pb
- Google - 15,000 Pb
- Sanger Institute - 22 Pb (just DNA sequencing)

Its Really Big

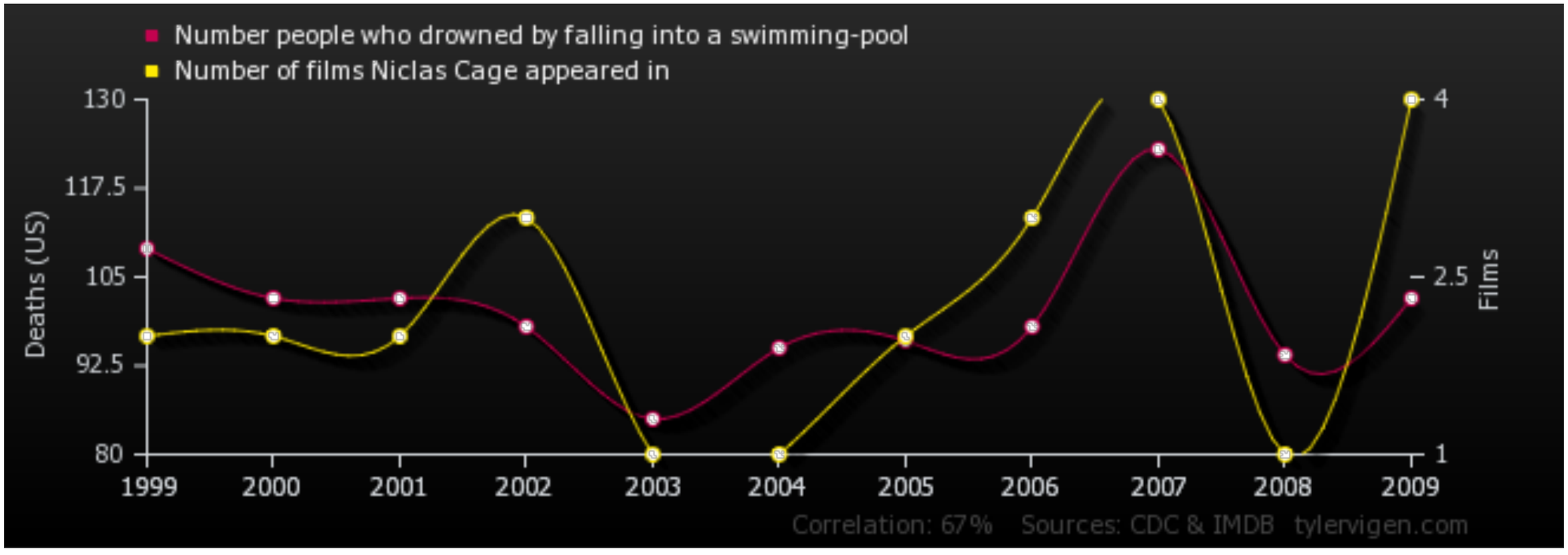
- Twitter - 100x larger than a home computer
- Facebook - 30,000x larger than a home computer
- Google - 15,000,000x larger than a home computer
- Sanger Institute - 22,000 larger than a home computer

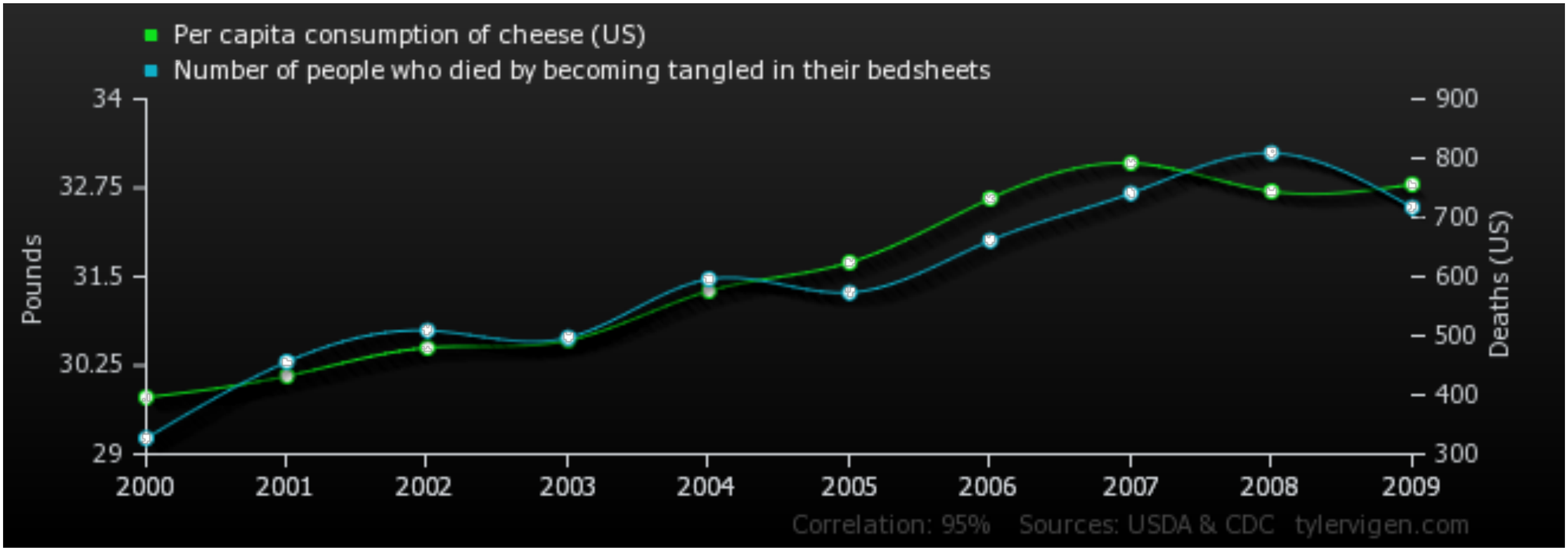

Hacking Skills
Math & Statistics Knowledge
Machine Learning
Data Science
Danger Zone!
Traditional Research
Substantive Expertise

Math and Statistics

- Without math and statistics it is easy to draw the wrong conclusions from data

- Perhaps the best example of this is “correlation does not mean causation”


Number people who drowned by falling into a swimming-pool
Number of films Niclas Cage appeared in
Deaths (US)
130
117.5
105
92.5
80
Films
4
2.5
1
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
Correlation: 67%
Sources: CDC & IMDB
tylervigen.com


Per capita consumption of cheese (US)
Number of people who died by becoming tangled in their bedsheets
Pounds
34
32.75
31.5
30.25
29
Deaths (US)
900
800
700
600
500
400
300
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
Correlation: 95%
Sources: USDA & CDC
tylervigen.com

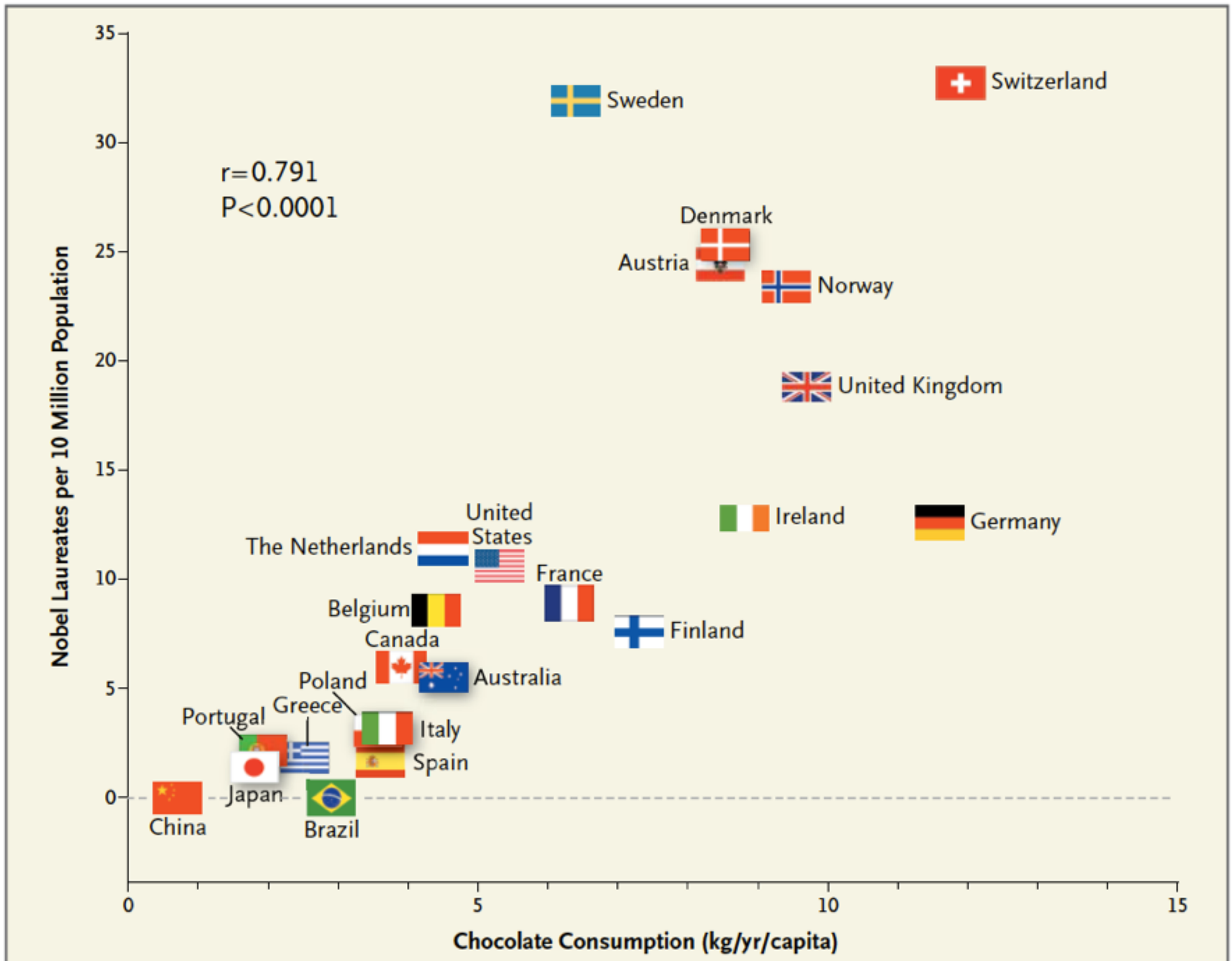
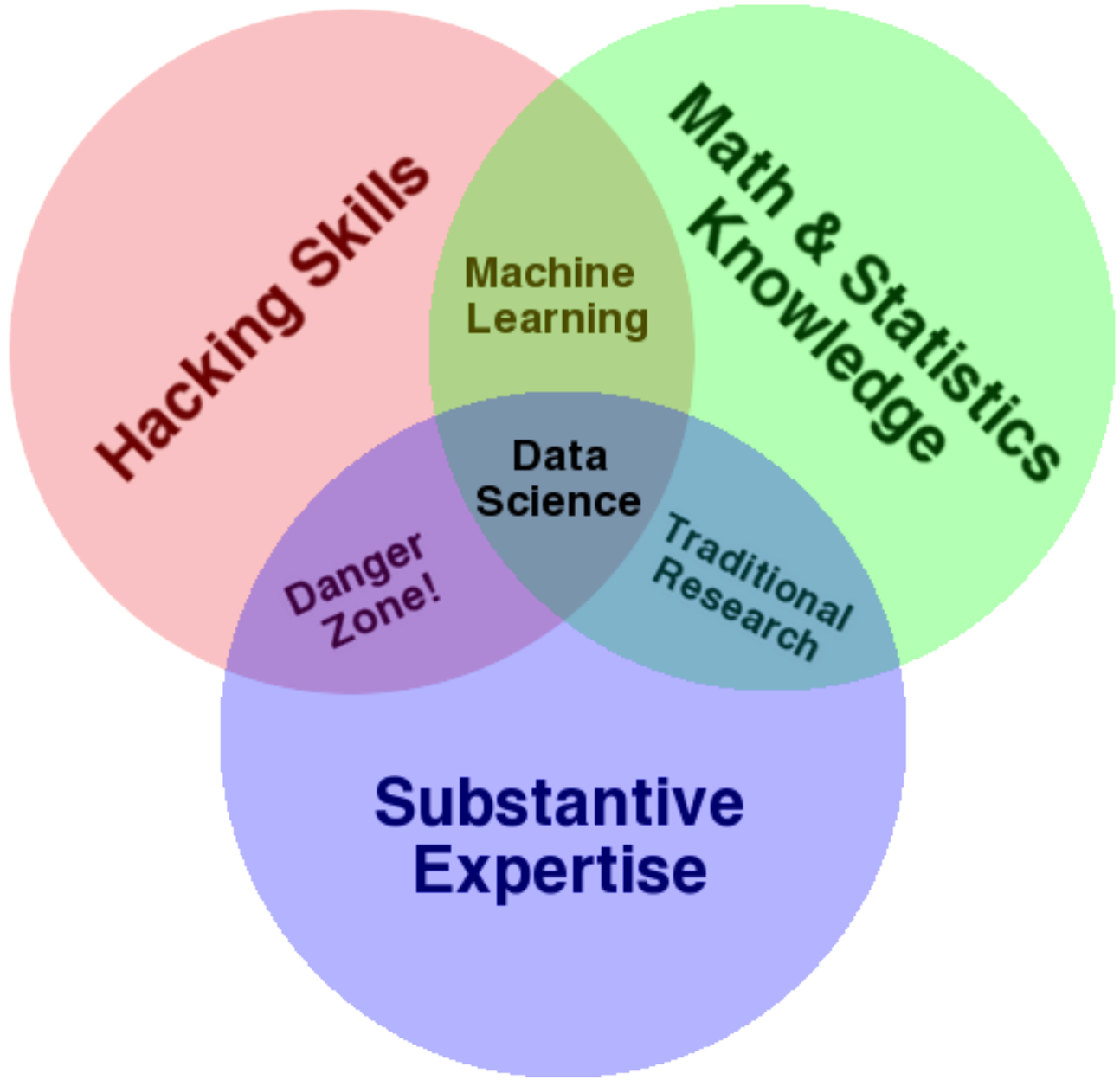
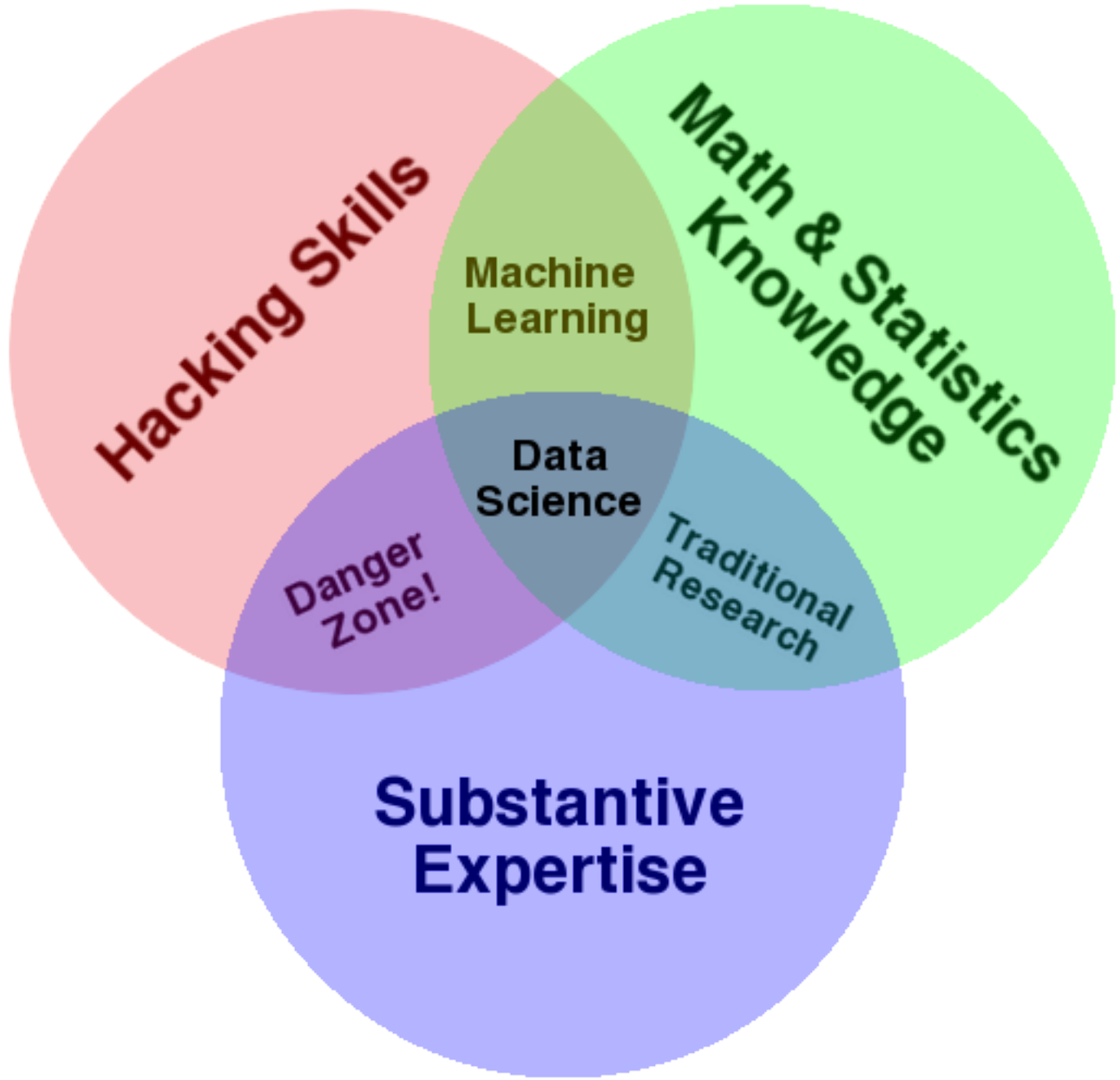


Figure 1. **Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**


Hacking Skills
Math & Statistics Knowledge
Machine Learning
Data Science
Danger Zone!
Traditional Research
Substantive Expertise

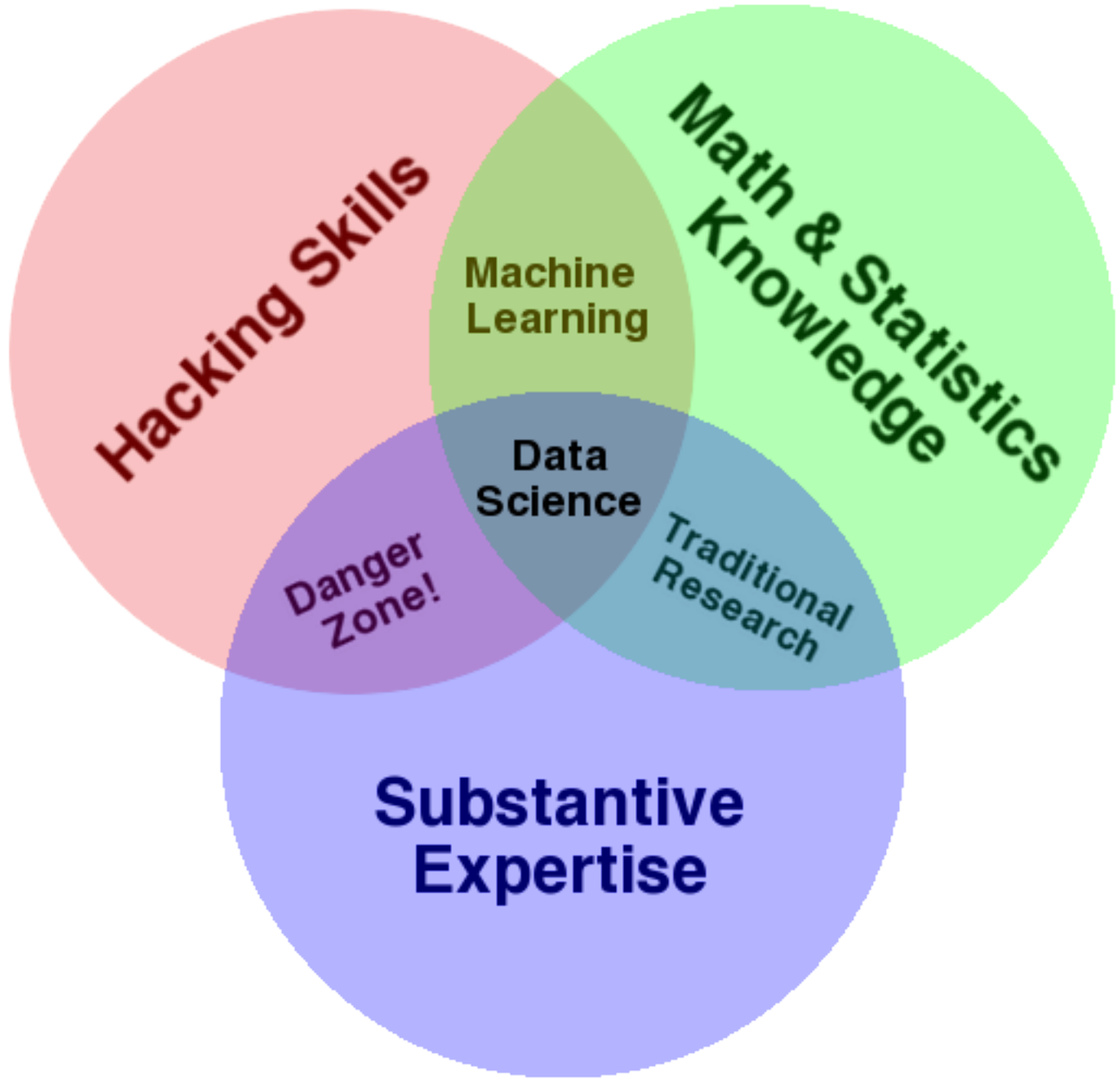
Substantive Expertise

- You need to know what the data is telling you
- Without a knowledge of biology all you really have is a fancy program and some numbers
- You need to be able to tell a story


Hacking Skills
Math & Statistics Knowledge
Machine Learning
Data Science
Danger Zone!
Traditional Research
Substantive Expertise

Machine Learning

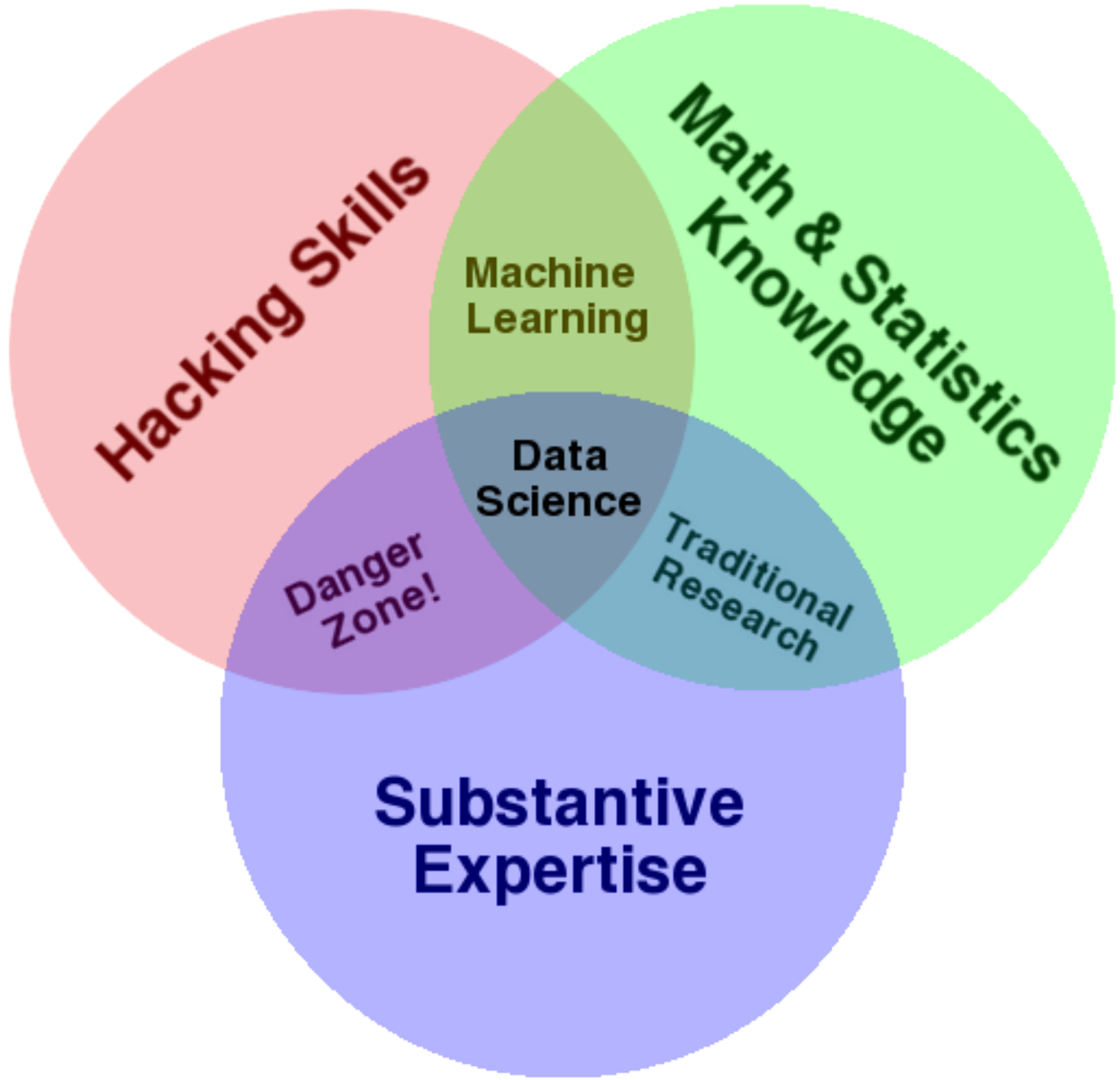
- Machine learning is the intersection of programming and statistics
- It is a collection of techniques that allow computers to explore data
- Facebook's face recognition algorithms use machine learning


Hacking Skills
Math & Statistics Knowledge
Machine Learning
Data Science
Danger Zone!
Traditional Research
Substantive Expertise

Traditional Research

- Bioinformatics is inherently interdisciplinary
- This makes it necessary to have a knowledge of all of the areas

- This is the difference between a bioinformatician and a biologist who spends their whole time studying one system


Hacking Skills
Math & Statistics Knowledge
Machine Learning
Data Science
Danger Zone!
Traditional Research
Substantive Expertise

Bioinformatics

- When you combine biology, math and statistics with programming skills, add in a little bit of coffee you get bioinformatics

- Its not magic

Some of our tools

Python

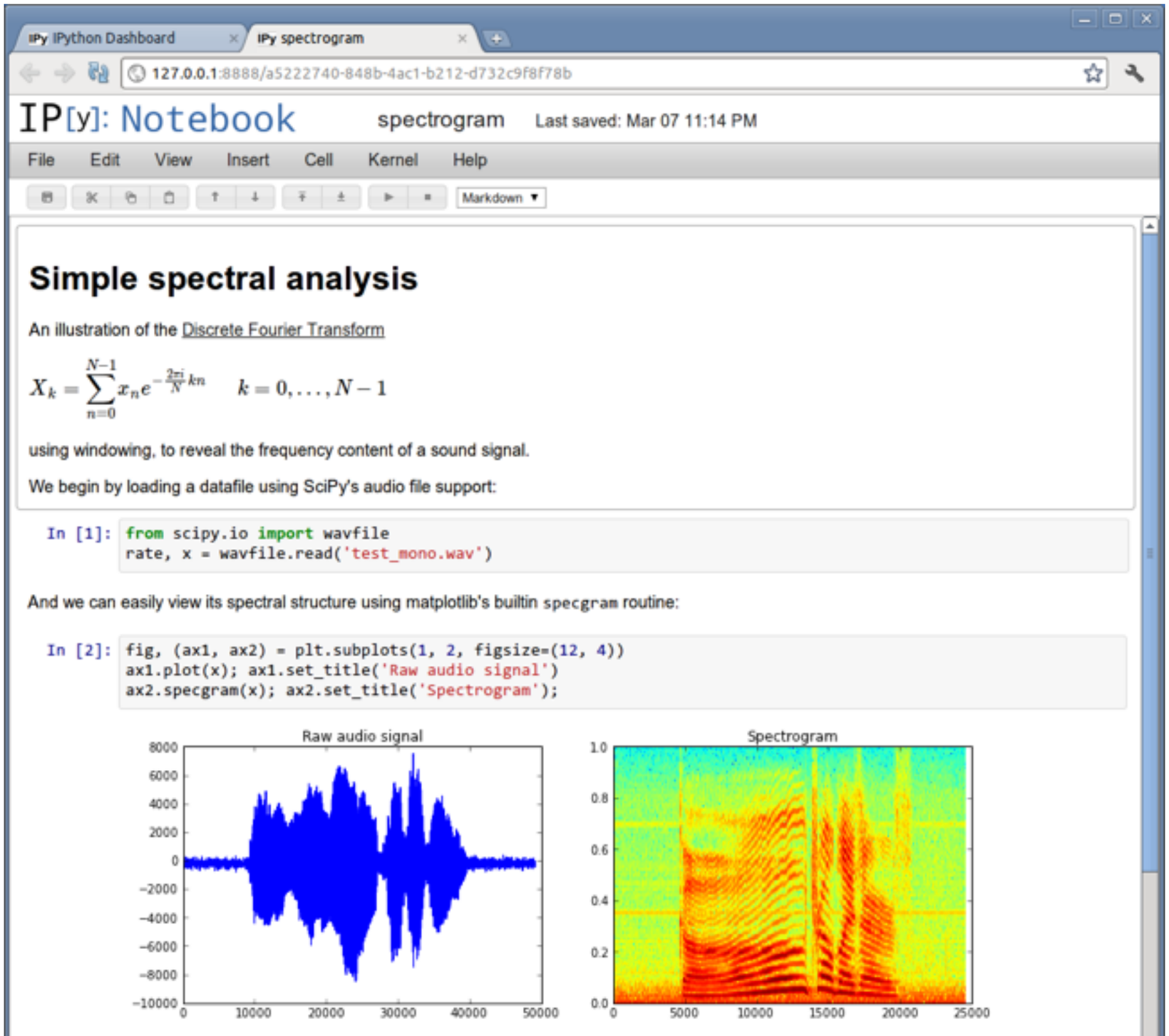
- Python is a multi-purpose programming language
- We use it for data manipulation, data processing and even analysis
- Its simple to read, quick to write, and powerful

Python

- name = raw_input('What is your name?\n’)
  print 'Hi, %s.' % name

- print 'Hello, world!'

IPython Notebook



NBViewer

Initialization

In [4]:
```
import pickle as pickle
import pandas as pd
import os as os

from Data.Containers import Run

from Data.Containers import get_run
from Data.Containers import Cancer
from Initialization.InitializeCN import initialize_cn
from Initialization.InitializeReal import initialize_real
from Initialization.InitializeMut import initialize_mut
from Initialization.PreprocessMethylation import process_meth
```

In [5]:
```
from IPython import utils
from IPython.display import HTML
```

In [6]:
```
css_file = 'profile_default/static/custom/custom.css'
base = utils.path.get_ipython_dir()
styles = "<style>\n%s\n</style>" % (open(os.path.join(base, css_file),'r').read())
display(HTML(styles))
```

In [7]:
```
!curl http://gdac.broadinstitute.org/runs/code/firehose_get_latest.zip -o fh_get.zip
!unzip fh_get.zip
```

```
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100  6542  100  6542    0     0   8343      0 --:--:-- --:--:-- --:--:-- 10620
Archive:  fh_get.zip
replace firehose_get? [y]es, [n]o, [A]ll, [N]one, [r]ename: ^C
```

GitHub

- GitHub is a place to keep code, and to share it with other people.

- A lot of bioinformatics software is now being published on GitHub

shilab

Filters ▾ | Find a repository…

+ New repository

wrapple

Python ★ 0 ⑂ 0

Updated 4 days ago

RPipeline PRIVATE

R ★ 0 ⑂ 0

RMarkdown Pipeline

Updated 11 days ago

seqParLASSOcode

Matlab ★ 0 ⑂ 0

seqParLASSOcode

Updated on Jan 19

DockerGuide

★ 0 ⑂ 0

People 8 ›

Invite someone

Teams 5 ›

Jump to a team

Owners
2 members · 14 repositories

DisorderedProteins
2 members · 1 repository

ovarian
2 members · 0 repositories

Create new team

Reproducible Research and Open Science

- Doing research that can be re-done by others is one of the cornerstones of science

- However there are currently many problems with reproducibility

- 47/53 cancer research papers were deemed unreproducible by Begley and Ellis

- A 2014 paper found less than 25% of computer science papers it analyzed were reproducible

Reproducibility and Open Science

- Having your code available (GitHub)
- Having your data available (FigShare, DataDryad)
- Having your workflow available (IPython Notebook, NBViewer, OSF)
- Open access journals or pre-prints (Arxiv, PeerJ Preprints, PLOS, etc)
- There are many problems, and many people are working on solutions

Why should you care?

- If you aren't already, you will pay for a lot of this research
- We as scientists have a responsibility to the public that our work is open and available to them