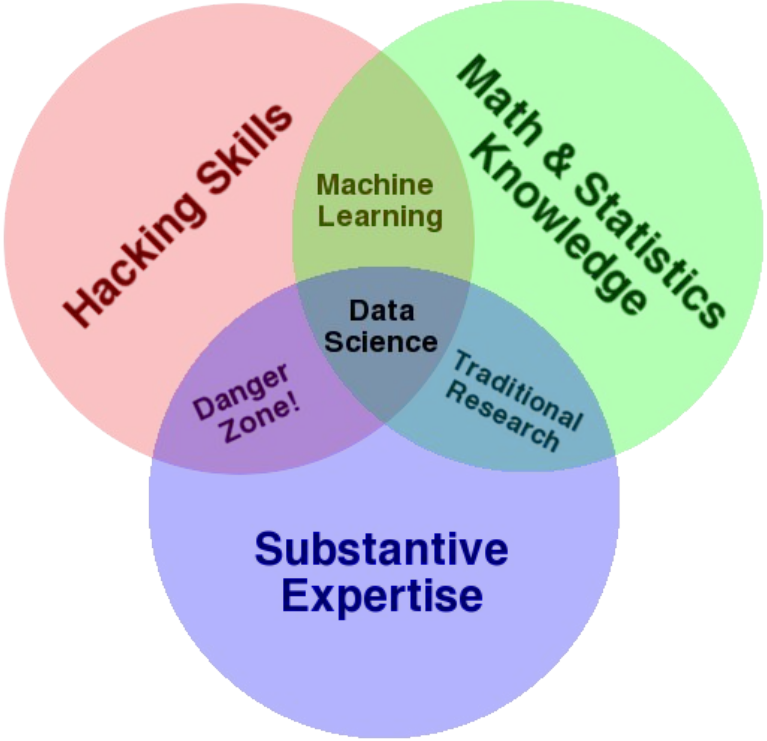
COSC/MATH 4931: Introduction to Data Science

Shion Guha
Mathematics, Statistics and Computer Science
January 17, 2017

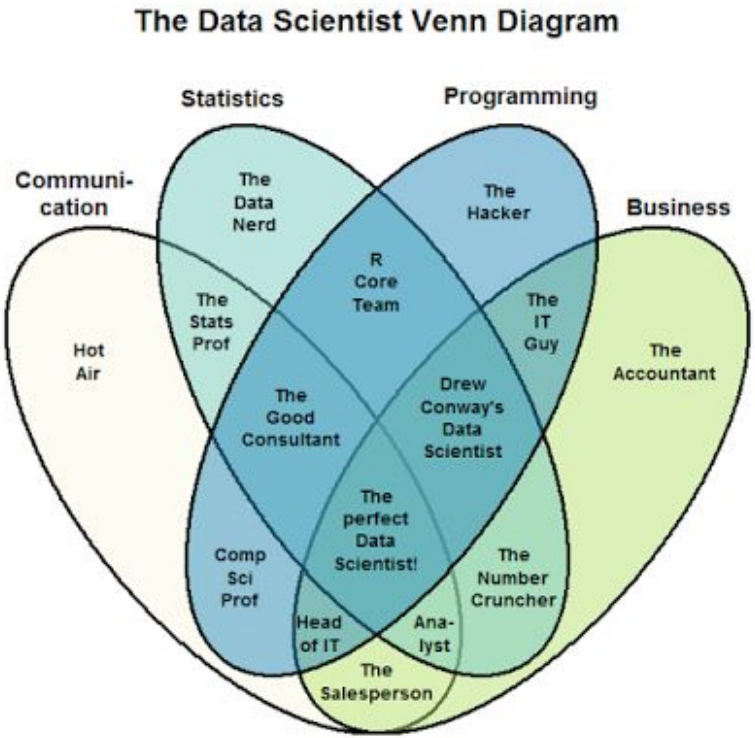
Course logistics

- Brief introduction to **modern** data science
- **Tools:** Jupyter, R-python stack
- Publishing to github regularly and writing reports.
- Understanding that data science is research - CSCW poster
- Topic presentation (20%), mid term project submission (30%), CSCW-style project paper (20%), final project (30%)

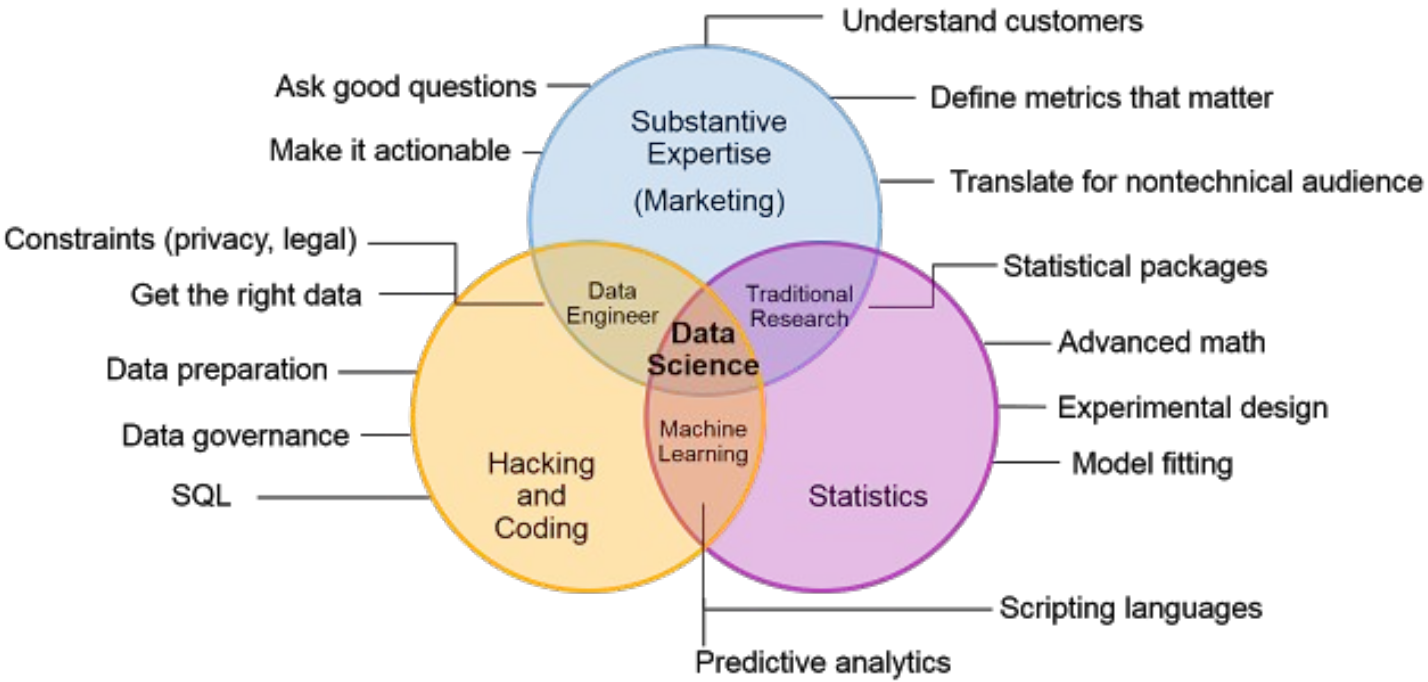
Data science: Conway, 2013



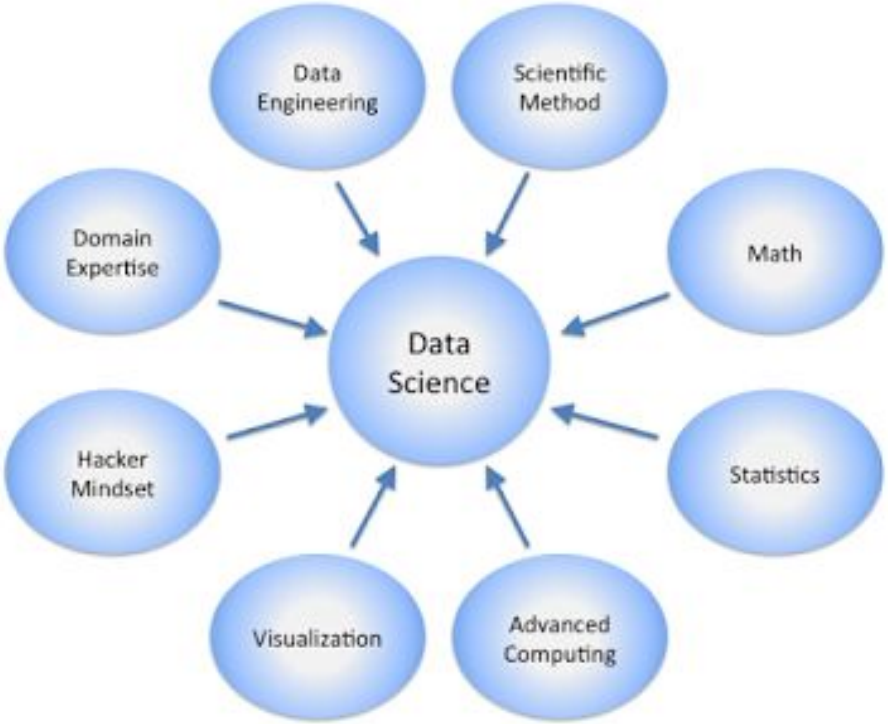
Data science: Kolassa, 2015



Data science: Gartner, 2016



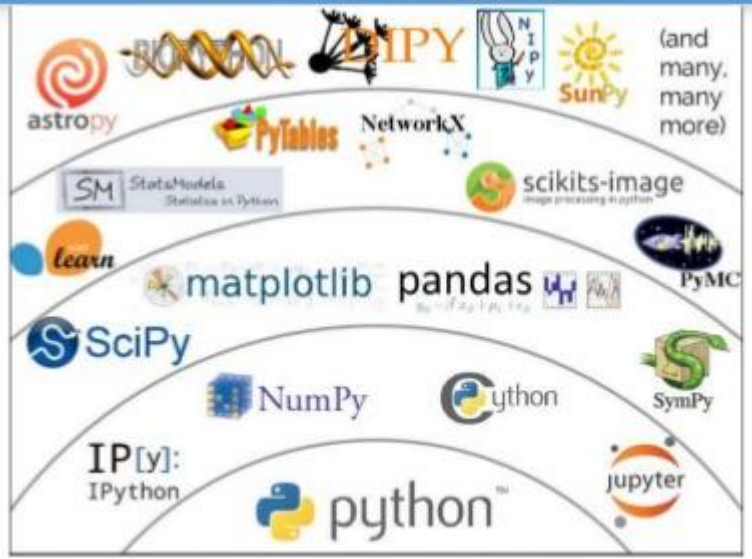
Data science: Wikipedia, 2016



Doing Data Science

Doing Data Science: Python Stack

Python Data Science stack



https://speakerdeck.com/jakevdp/the-state-of-the-stack-scipy-2015-keynote

hulu

Doing Data Science: Jupyter Notebooks

Run some Python code!

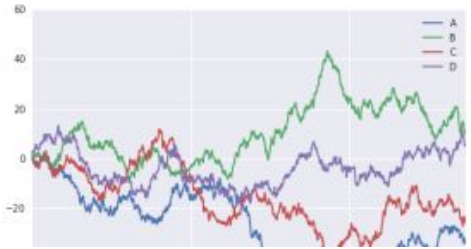
To run the code below:

1. Click on the cell to select it.
2. Press SHIFT+ENTER on your keyboard or press the play button (▶) in the toolbar above.

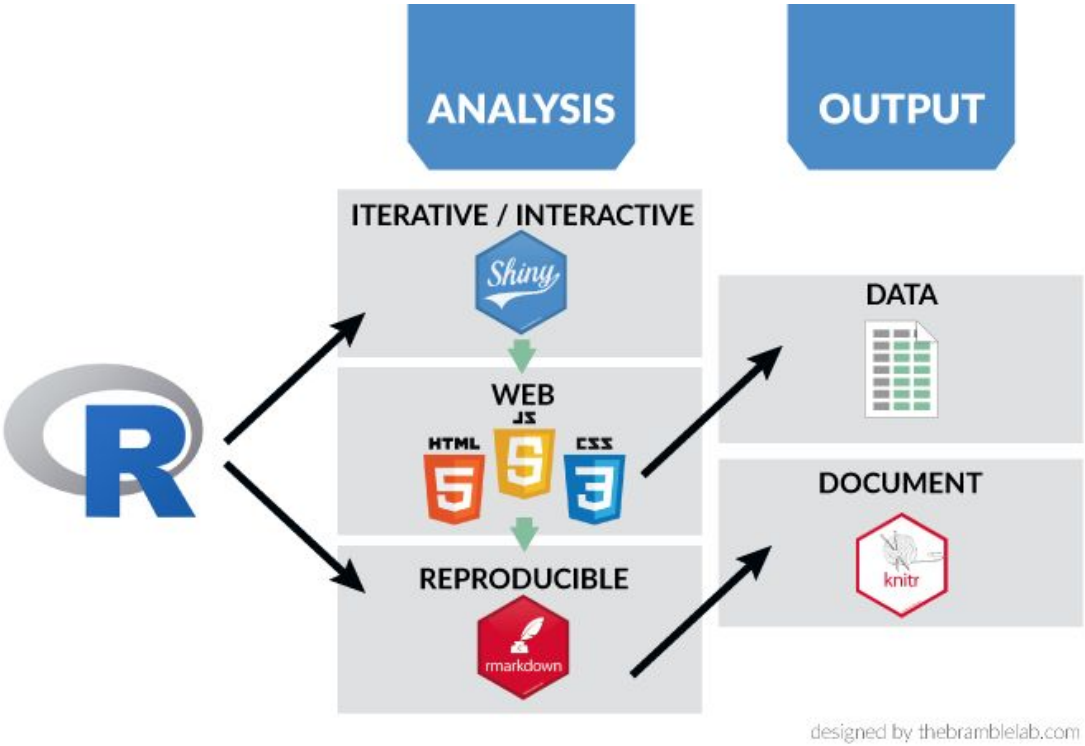
A full tutorial for using the notebook interface is available here.

In [1]:
```
%matplotlib inline

import pandas as pd
import numpy as np
import matplotlib

from matplotlib import pyplot as plt
import seaborn as sns

ts = pd.Series(np.random.randn(1000), index=pd.date_range('1/1/2000', periods=1000))
ts = ts.cumsum()

df = pd.DataFrame(np.random.randn(1000, 4), index=ts.index,
                  columns=['A', 'B', 'C', 'D'])
df = df.cumsum()
plt.figure(); df.plot(); plt.legend(loc='best')
```

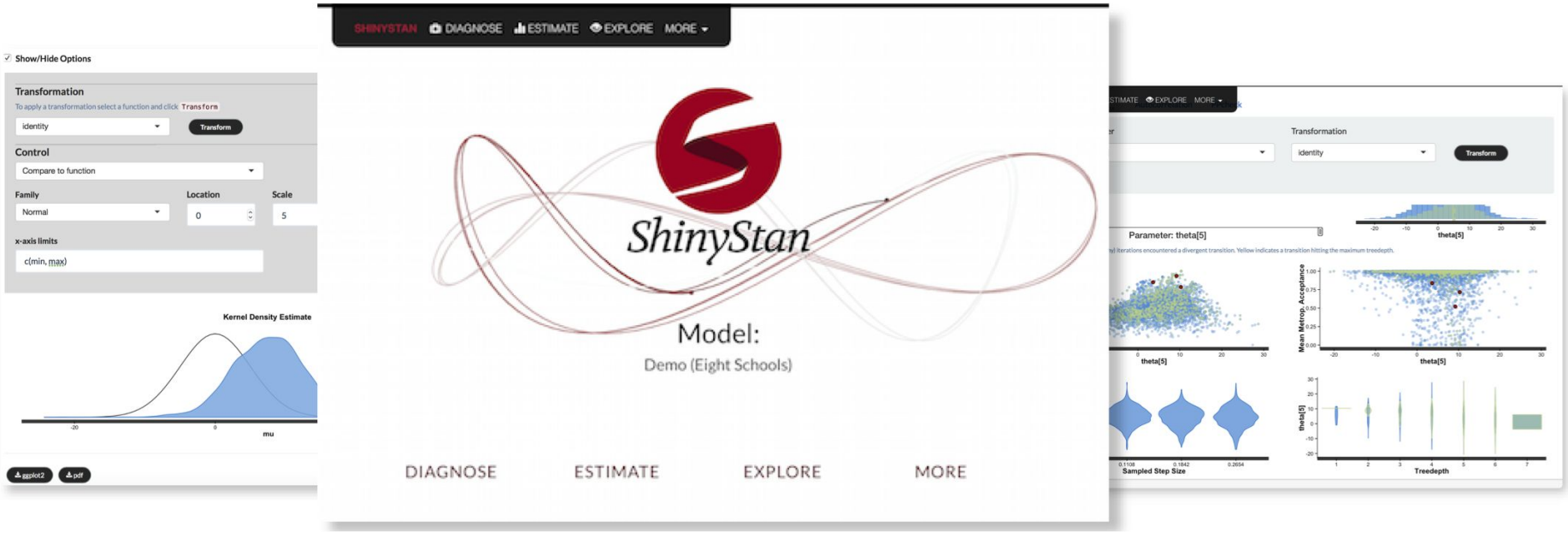
Out[1]:
```
<matplotlib.legend.Legend at 0x7fb27b72fcc0>

<matplotlib.figure.Figure at 0x7fb283672b70>
```



Doing Data Science: R Stack



Doing Data Science: R Stack



Doing Data Science: Github Publishing

frankcleary / data-science-bytes

Unwatch 2 | Star 0 | Fork 0

branch: master | data-science-bytes / content / bart-annotate.ipynb

Frank Cleary on Mar 21 fixed by removing bs4 usage in parsing

0 contributors

3013 lines (3012 sloc) | 257.024 kb | Raw | Blame | History

To extend on my post about plotting and reshaping data from the BART API, I worked a bit with the matplotlib annotion interface to add text and arrows to a plot. The meat of this post is in cell #4 below. Download notebook.

In [1]:
```
%matplotlib inline
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
import datetime

def prettify_axis(ax, ylabel='', xlabel=''):
    label_format_dict = dict(fontsize=20, fontweight='bold')
    tick_format_dict = dict(labelsize=16, direction='out', top='off', right='off',
                            length=4, width=1)
```

Core:

- Introduction to Data Science (**this course! 3xxx level**)
- Advanced Data Science (coming soon! 4xxx level elective)
- Data at Scale (i.e. Parallel & Distributed Systems) (MSCS 6060)

Augment:

- Data Structures and Algorithms (COSC 2100)
- Data Mining (COSC 4610)
- Discrete Mathematics (MATH 2100)
- Linear Algebra (MATH 3100)
- Probability (MATH 4700)
- Statistical Methods (MATH 4720)

What it isn't:

- Not a programming course.
- Not a theory course.
- Not your usual lecture-homework-exam course model.
- Strict no spoonfeeding policy!

What it is:

- **Flipped classroom:** Demo-code-discuss model
- Simulates real world data science projects
- Learn how to do research and ask research questions.
- Brief introduction to commonly used data science methods.
- Critically read, write and present about data.

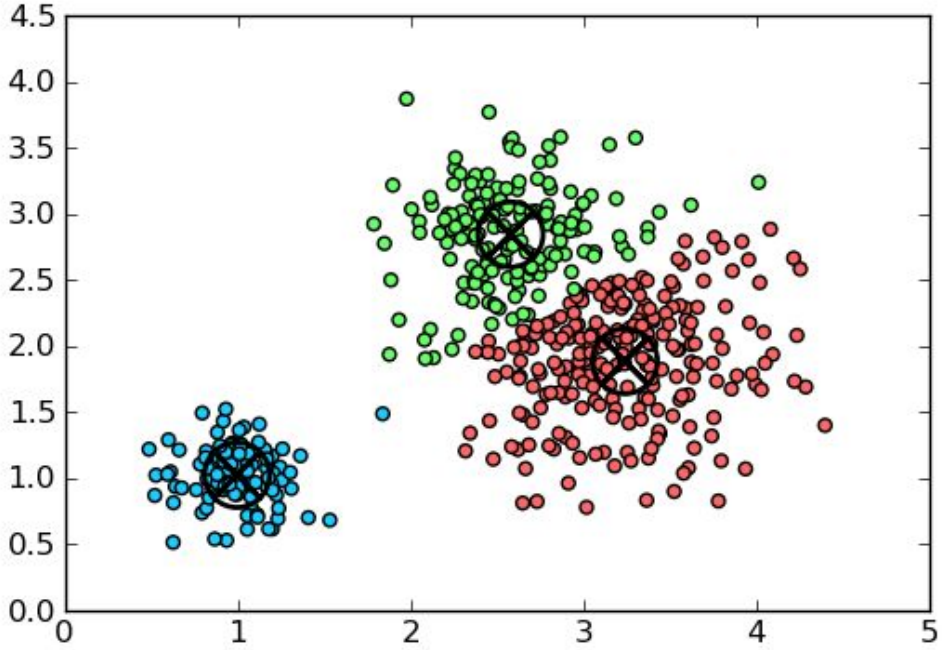
What it isn't:

- Not a programming course.
- Not a theory course.
- Not your usual lecture-homework-exam course model.
- Strict no spoonfeeding policy!

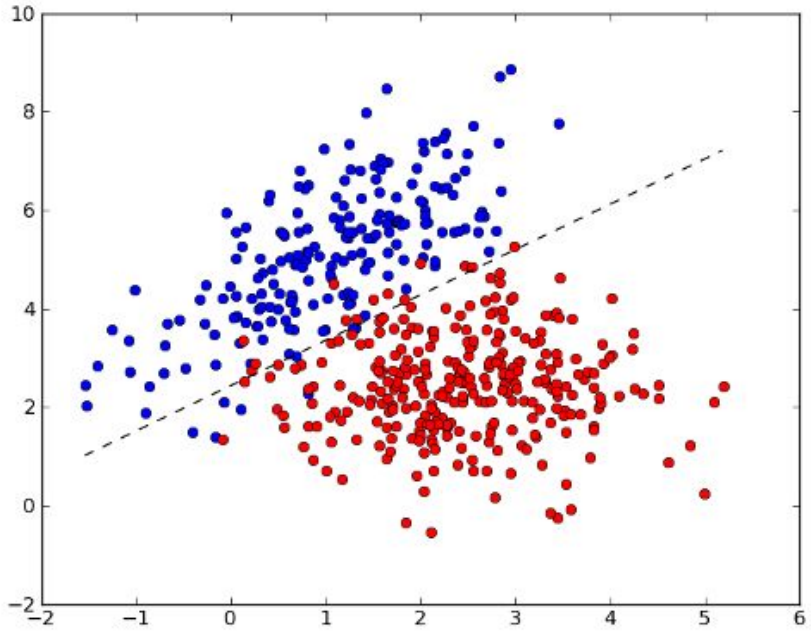
What it is:

- **Flipped classroom:** Demo-code-discuss model
- Simulates real world data science projects
- Learn how to do research and ask research questions.
- Brief introduction to commonly used data science methods.
- Critically read, write and present about data.

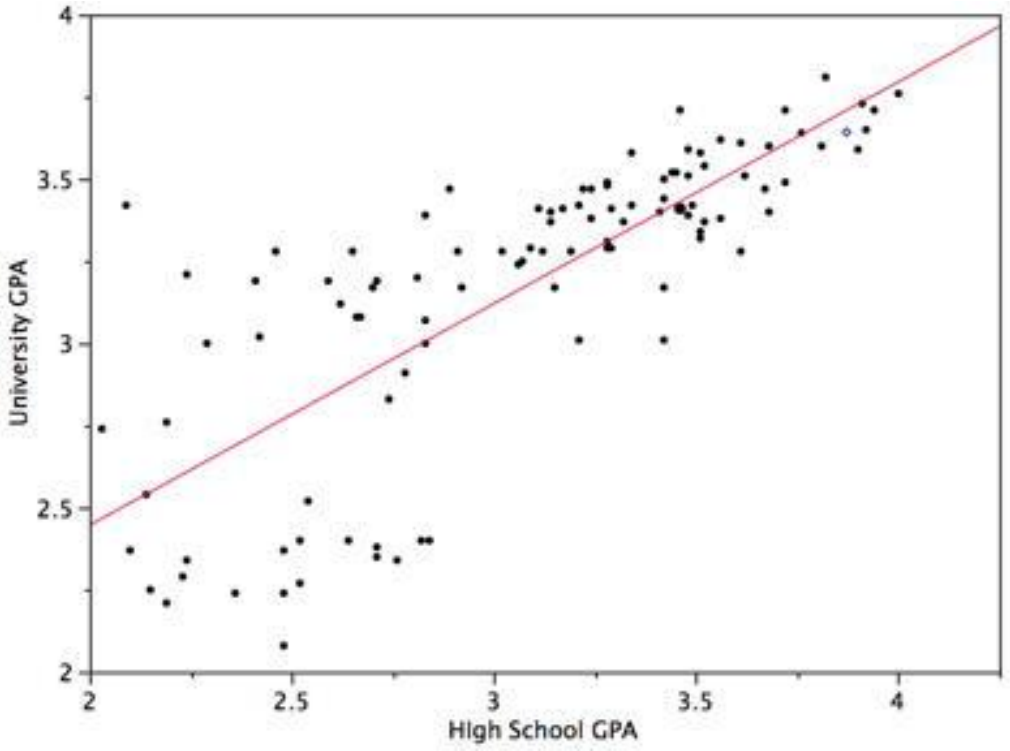
Data Science Methods: Clustering

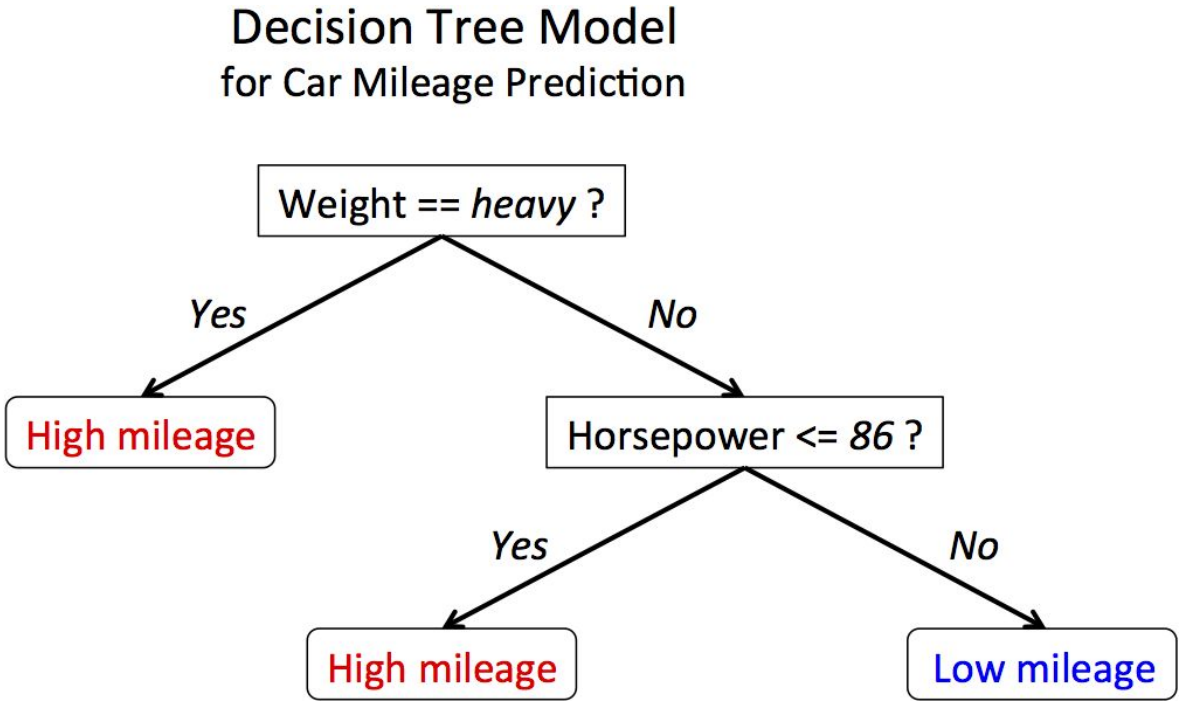


Data Science Methods: Classification

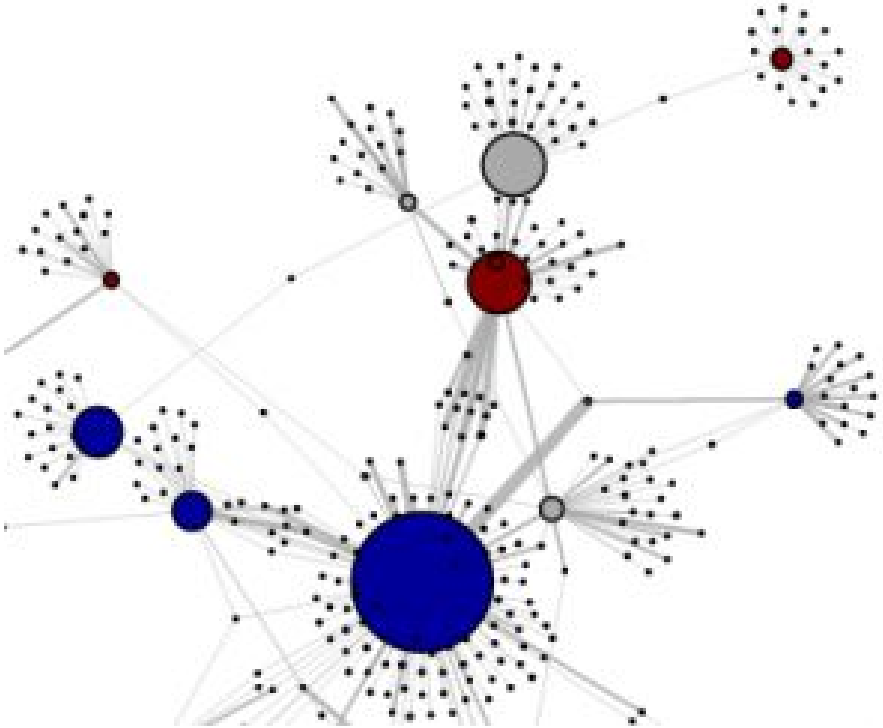


Data Science Methods: Regression

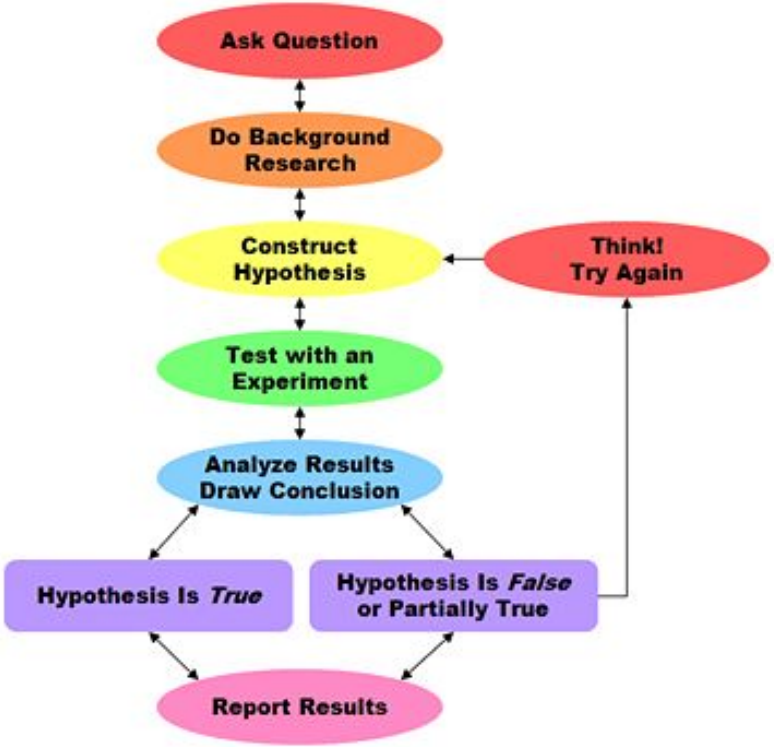


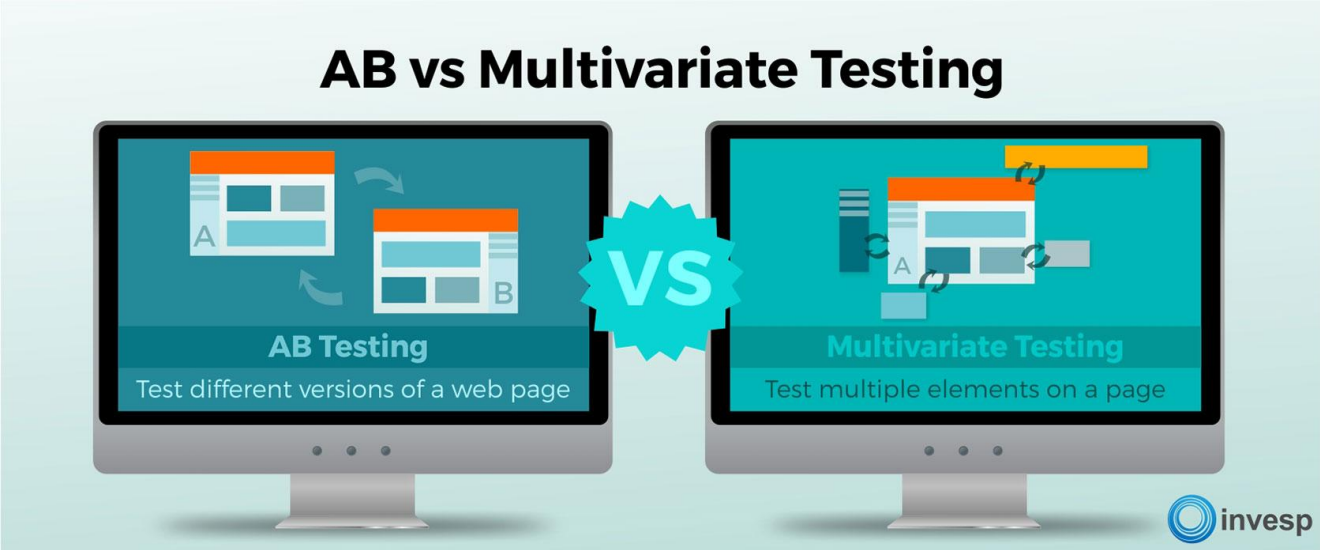

Decision Tree Model
for Car Mileage Prediction
Weight == heavy ?
Yes
No
High mileage
Horsepower <= 86 ?
Yes
No
High mileage
Low mileage

Data Science Methods: Network Analysis

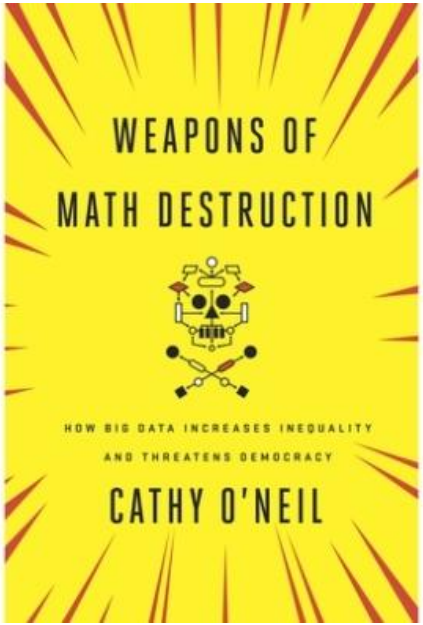


Scientific Method




AB vs Multivariate Testing
A
B
VS
AB Testing
Test different versions of a web page
A
Multivariate Testing
Test multiple elements on a page
invesp

Human Centered Data Science



Forbes · New Posts +26 posts this hour · Popular What 20-Year-Olds Do · Lists America's Top Colleg · Video Digital C

TECH | 2/16/2012 @ 11:02AM | 2,119,347 views

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

319 comments, 169 called-out + Comment Now + Follow Comments

Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. Target, for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.

TARGET

Target has got you in its aim

38.0k Share 15.3k Tweet 5.8k Share 362 Submit

MARQUETTE
UNIVERSITY

BE THE DIFFERENCE.