DATA SCIENCE PRIMER

Hi. I'm Mark Trovinger.

 github.com/PyDataFtWayne/datascienceprimer

 @MarkTrovinger

PyData

...is a gathering of users and developers of data analysis tools in Python.

NUMFOCUS

OPEN CODE = BETTER SCIENCE

WHAT IS

DATA SCIENCE?

USE OF:

STATISTICS

MACHINE LEARNING

SOFTWARE ENGINEERING

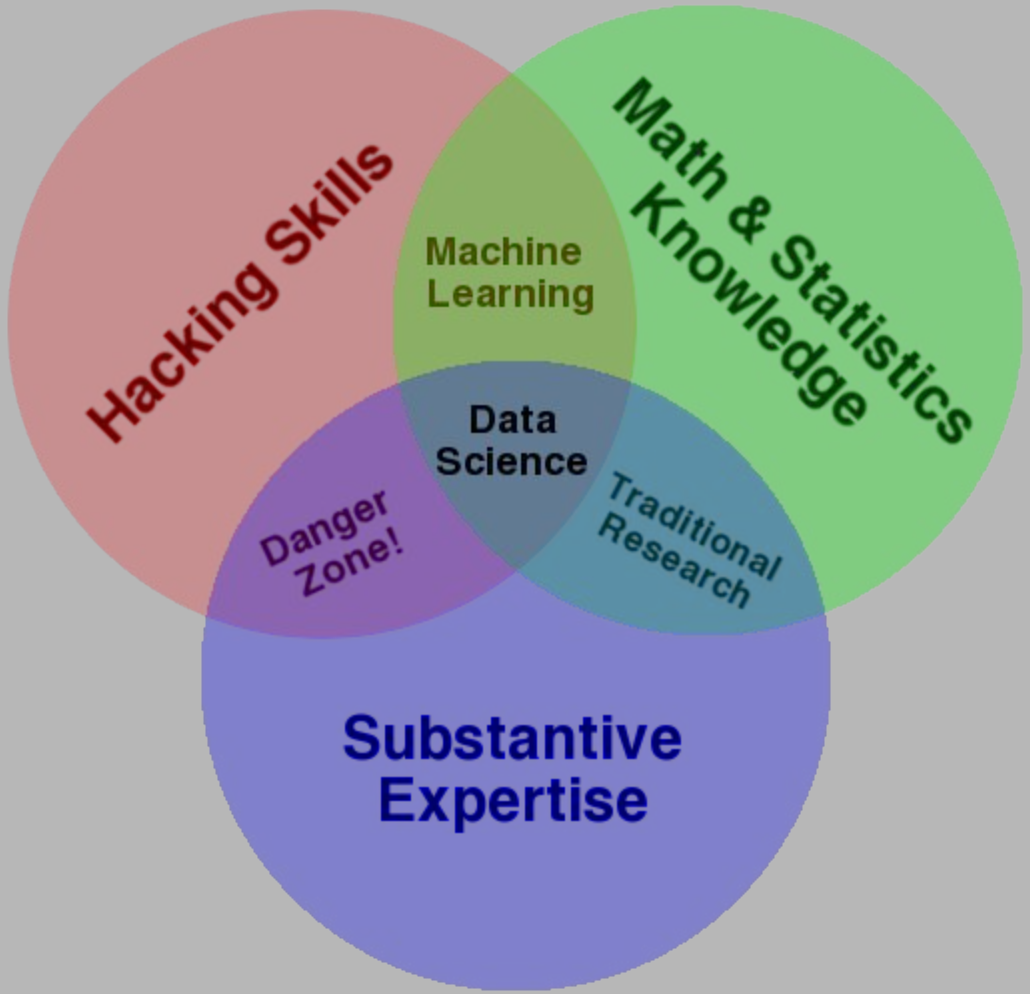
DOMAIN KNOWLEDGE

TO DO:

EXTRACT INFORMATION FROM DATA

MAKE PREDICTIONS FROM THAT INFORMATION

USE PREDICTIONS IN AN APPLIED SETTING


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

THE DATA SCIENCE PROCESS

1. FRAME THE PROBLEM
2. COLLECT RAW DATA
3. PROCESS DATA FOR ANALYSIS
4. EXPLORE THE DATA
5. PERFORM IN-DEPTH ANALYSIS
6. COMMUNICATE RESULTS

EXAMPLES:

AIRBNB: HELP RENTERS SET PRICES

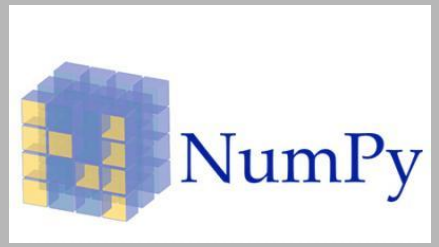
BAYES IMPACT: BETTER MATCHES BETWEEN ORGAN DONORS AND RECIPIENTS

PREDICT EMPLOYEE ATTRITION

PYTHON

DATA SCIENCE

STACK


IP[y]:
IPython

NumPy

ython

jupyter

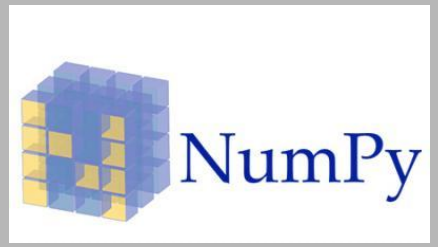


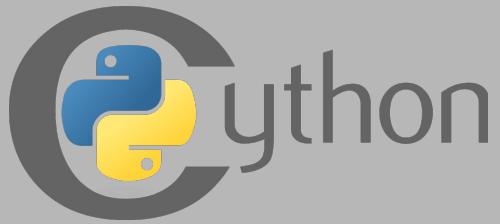
SciPy


matplotlib

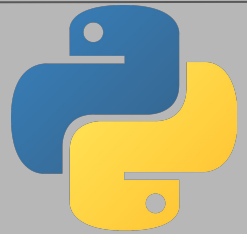

pandas


IP[y]:
IPython


NumPy


Cython


jupyter




scikit
learn


SM
StatsModels
Statistics in Python

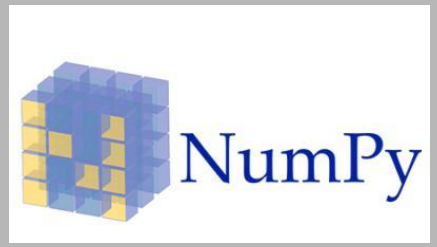

scikit-image
image processing in python

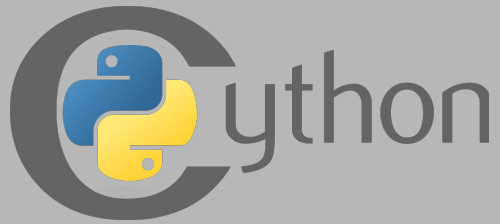

SciPy


matplotlib

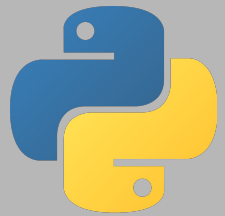

pandas


IP[y]:
IPython


NumPy


Cython


jupyter



AND MANY MORE:

PERFORMANCE:

NUMBA, THEANO, NUMEXPR

DATA STRUCTURES AND COMPUTATION:

DASK, BLAZE, DISTARRARY, PYSPARK, GRAPHLAB, XRAY

VISUALIZATION:

BOKEH, SEABORN, PLOTLY, GGPLOT, TOYPLOT, HOLOVIEWS

I'll just leave this right here...

If you haven't switched, it's time

TITANIC DEMO

FUTURE TOPICS:

BIG(ISH) DATA WITH DASK

BIG(GER) DATA WITH AWS AND SPARK

FAST PYTHON WITH NUMBA

DEEP LEARNING WITH TENSORFLOW

AUTOMATING WITH TPOT

DATA SCIENCE PRIMER 2: SON OF DATA SCIENCE PRIMER

REFERENCES:

STATE OF THE STACK, SCIPY 2015 JAKE VANDERPLAS

TREY CAUSEY, DATAQUEST.IO INTERVIEW

DATA SCIENCE VENN DIAGRAM, DREW CONWAY

KAGGLE TITANIC, ANDREW CONTI