A TOUR OF THE PYTHON DATA SCIENCE ECOSYSTEM

USING:

- PYTHON 3.5
- JUPYTER

W/ LIBRARIES:

- PANDAS
- NUMPY
- SCIKIT-LEARN

BONUS:

- TENSORFLOW
  - KERAS
- XGBOOST

I AM:

BEN KAMPHAUS

MACHINE LEARNING/SOFTWARE ENGINEER AT THINKTOPIC

I SPLIT TIME BETWEEN PYTHON AND CLOJURE.

STRUCTURE:

- SALIENT FEATURES OF PYTHON (10 MINUTES)
  - PANDAS MODULE (10 MINUTES)
  - NUMPY MODULE (10 MINUTES)
  - SKLEARN MODULE (10 MINUTES)
  - BONUS MATERIAL (10 MINUTES)

BUT FIRST, DEMO!!!!

BIRD'S EYE VIEW OF PYTHON FEATURES

DEPENDENY MANAGEMENT:

```
pip install sklearn
```

++ QUICK AND EASY AT THE COMMAND LINE
-- NATIVE DEPENDENCIES. SYSTEM GLOBAL

DOCKER. `virtualenv`. ETC.

MULTIPARADIGM

OBJECT-ORIENTED

```
class Point:
    def __init__(self, x, y):
        self.x = x
        self.y = y
```

FUNCTIONAL (-ISH)

```
map(lambda x: x**2, range(10))
```

```
[x**2 for x in range(10)]
```

TYPICALLY IMPERATIVE:

```
l = []
for i in range(10):
    if i % 2:
        l.append(i)
```

BUT STILL. HIGHER ORDER FUNCTIONS!

```
def add_to(x):
    def add(y):
        return x + y
    return add
```

AND COOL PYTHONIC THINGS LIKE GENERATORS:

```
def geometric_series(a, r):
    power = 0
    yield a
    while True:
        power += 1
        yield a * r**power
```

PANDAS

ELEVATOR PITCH:

FAST TABULAR DATA MANIPULATION

WHEN YOU USE IT:

- IO OR DATABASE ACCESS LAYER
- RELATIONAL ALGEBRA OPERATIONS
- BASIC STATISTICS
- SIMPLE VISUALIZATIONS
- INPUT TO SOME MACHINE LEARNIG APIS

PANDAS PHILOSOPHY

- FAVOR VIEWS FOR SELECT, FILTER, SLICE (VIA NUMPY)
- MUTATION PRODUCES A COPY (EXCEPT WITH `inplace=True`)
  - OPTIMIZED IN C OR CYTHON, SO FAST
  - INDEXES AND COLUMNS ARE LABELED

INTERACTIVE DEMO!!!

NUMPY

ELEVATOR PITCH:

NDARRAYS AND FAST LINEAR ALGEBRA ROUTINES

WHEN YOU USE IT:

- IMPLEMENTING MACHINE LEARNING OR DATA SCIENCE METHODS FROM PAPERS
- FEATURE ENGINEERING AND TRANSFORMATION
- DATA NORAMLIZATION
- INTERPOLATION AND NUMERICAL COMPUTING

YOU KNOW. MATLAB. IDL. APL. ETC.

TYPICAL NUMPY OPERATIONS:

- LINEAR ALGEBRA OPERATORS: DOT PRODUCTS, SCALING, ETC.
  - VECTORIZED ARRAY FUNCTION CALLS
  - RESHAPING, STACKING, FLATTENING
  - SLICING, MASKING
- MORE ADVANCED: DECOMPOSITIONS, FACTORIZATION, MATRIX INVERSION/PSEUDO-INVERSION

INTERACTIVE DEMO!!!

SCIKIT-LEARN (SKLEARN)

ELEVATOR PITCH:

YOUR ONE STOP MACHINE LEARNING SHOP.

WHEN YOU USE IT:

- TEST/TRAINING SPLITS AND CROSS-VALIDATION
- PREPROCESSING
- DIMENSIONALITY REDUCTION
- CLUSTERING

WHEN YOU USE IT (CONT):

- FEATURE ENGINEERING
  - LEARNING/FITTING
- PERFORMANCE EVALUATION
- HYPERPARAMETER SEARCH

TYPICAL SKLEARN WORKFLOW:

- INITIAL PREPROCESSING/IO/MUNGING FROM PANDAS + NUMPY
- SCALING AND/OR DIMENSIONALITY REDUCTION
- SPLIT BETWEEN TRAINING AND HOLD-OUT TEST DATASETS

TYPICAL SKLEARN WORKFLOW (CONT):

- INITIAL ASSESSMENT OF METHODS W/TRAINING DATA
- OUT-OF-FOLD ACCURACY ASSESSED W/CROSS-VALIDATION
- HYPERPARAMETER SEARCH FOR FINAL MODEL SELECTION

INTERACTIVE DEMO!!!

BONUS MATERIAL!

TENSORFLOW

TENSORFLOW BUILDS A DATA FLOW GRAPH OUT OF A MIX OF NUMPY IDIOMS AND NEURAL NETWORK SPECIFIC CONSTRUCTS.

KERAS

KERAS WRAPS TENSORFLOW (AND THEANO, WHICH IS SIMILAR) WITH AN SKLEARN LIKE INTERFACE.

XGBOOST

XGBOOST CAN BE CALLED IN A MANNER SIMILAR TO SKLEARN LEARNING ALGORITHMS.

OR DIRECTLY VIA A FULLY SKLEARN COMPATIBLE WRAPPER.

LET'S MAKE SENSE OF SOMETHING ON KAGGLE!