PYCON Notes

Cross stitch with Pillow

* Pillow is an image library
  + get\_pixel
    - results in rgb color tuples
  + imageStat
    - can use to find the ‘median’ color of an image
    - can make a custom palette with it
* scikit-image library
  + Can calculate distance between colors using a car manufacturer’s formula
* ih – experiment 626: the cross-stitch library

Tools to automate repetitive tasks

* tox
  + standardize python testing
  + automating Python packages
  + tests multiple versions
* nox
  + flexible test automation
  + similar to Flask
  + tests multiple versions
  + builds documentation
* sphinx
  + builds documentation
* invoke
  + task execution tool

Speeding up SQLAlchemy Operations

* SQL Expression Language (python DSL)
* Object Relational Mapper (ORM)
  + Bad SQL Access patterns mean that inefficiencies will go unnoticed at low volumes
    - Beware of implicit queries
    - 1 query per request
    - Don’t loop, calculate within the database
    - Coalesce a null into a zero

GraphQL

* Get data from multiple sources in query
* Can work with jinja
* Ariadne and Cannula libraries?

Pandas

* Save memory with inplace = True
* Options to normalize data while sorting it
  + You can convert NaN’s to 0’s in ‘read\_csv’
  + Concat + join
  + PivotTables
  + Cython- changes python into C and will make it run faster

Camelot and Excaliber

* Creates tables from pdf’s
* Tabula is another library that might work
* camelot.read\_pdf(similar to pandas—built on pandas)
  + Makes a pandas dataframe with a PDF File
  + Also has a parsing report to indicate accuracy
    - Lattice lines
    - Stream white spaces
* camelot-py
* Excaliber is the same deal, but it runs from a web server
  + excaliber-py

Numpy

* Has read file options

Papermill w/Jupyter

* Extension(?) or library to run multiple jupyter notebooks
* me\_dime
* github/bit-bucket
* review-md
* works with Spark clusters- notebook connects
* interact- might be easier to use
* can make a notebook to run a bunch of notebooks
* nb convert to strip code
* You can hide code cells with papermill
  + Hidden input/hidden output
* Jupyter labs- can grab a single document and split the view
* Dag execution systems

Instant Serverless APIs

* Serverless advantages and disadvantages
  + Scale to zero- only pay for the services you use
    - Google Run- Stateless—has no database
  + SQLite doesn’t scale well for writes
* Data Journalism
  + Analysts who work for newspapers
* Datasette.io
  + Tool for exploring and publishing data
  + It has instant mapping functions
  + CSVs to sqlite
* 538 has data files on github
* sqlite-utils
  + Utilities to convert csv to sqlite
* Glitch

Statistical Profiling

* How fast (or slow) a program runs
* Deterministic Profiling
* Stack Frame

Alternatives to Deep Learning

* Regression
  + Advantages:
    - Many kinds
    - Fast to fit
    - Works well with small data
    - Easy to interpret
  + Disadvantages
    - More prep work
    - Models require validation
  + Mixed effects regression
    - Finds patterns within a regression model
    - Statsmodel api
* Trees (Random Forests)
  + Advantages
    - Popular (winners on Kaggle)
    - Preform better than logistic regression
    - Less cleaning than regression
    - Packages available
  + Disadvantage
    - Can over fit
    - More sensitive to differences between datasets
    - Less interpretable than regression
    - Requires more compute and training points
  + xgboost
* Distance-based methods
  + Closest numbers form group
    - Advantages
      * Fast to train
      * Works well with small datasets
    - Disadvantages
      * Accuracy isn’t awesome
      * sklearn
* Deep learning
  + Advantages
    - Most accurate
  + Disadvantages
    - Requires most time, money, and compute
* Accuracy
  + 1. Deep learning
  + 2. Trees
  + 3. Regression
  + 4. Distance

Deep Learning

* Perceptrons as substitutes for neurons
  + Basically we build a brain for a computer
  + Feed Forward Network
  + Deep Feed Network
    - Hidden layers of perceptrons—aka, ‘thinking’
* Convulutional Neural Network
  + Takes in a lot of information and reduces it to make a decision
* Recurrent Neural Network
  + Only knows what’s happening at a given time
  + Has a feedback loop – ‘memory’
    - Can define how long stuff is remembered
* Libraries
  + NumPy
  + Keras
* Issues:
  + Memory
    - use mini batches
  + Write to disk
    - Joblib, h5py
  + Long running process
    - Tmux or screen
  + Not analyzing your errors
    - Figure out what’s wrong
    - Confusion matrix—figure out if you have false negatives and positives

Plug-ins

* Modularized code
  + Easier to read, maintain, test, and extend
* Decorator function
  + Wraps around another function
  + Data = decorator(function)
* Libraries
  + Pathlib and click
    - Func = getattr(readers, format)
    - Def read(plugin \*args, \*kwargs)
    - @register
  + Pyplugs
    - Simple plug-in packages

Testing

* Defensive programming
  + Raise assertion errors in code
    - Assert (condition)
* PyTest
  + Command line
    - Pytest(app name) -v
* Pandas
  + Notnull, isnull, duplicated
  + Util.testing(assert frame\_equal, check dtype, check like)
* Hypothesis
  + Automatically generates data for testing
  + Can use with pandas

Plotting tools

* Grammar-based visualization
* R- ggplot2
* Plotnine (Python)
* Javascript = VegaLite
* Altair (Python)