Tools for Data Science notes

* Languages
  + Python
    - Clear syntax
    - General purpose
    - Large standard library
  + R
    - Free software
      * Different organization
      * Focused on values rather than business
    - Array-oriented syntax (math to code)
  + SQL (structured query language)
    - Limited to querying and managing data
    - Relational database
  + Other languages
* Categories of tools
  + Data management
  + Integration and transformation (retrieving, ETL)
  + Visualization – initial exploration
  + Model building – algorithms
  + Deployment – available to other apps
  + Monitoring and assessment
  + Asset management
  + Development environments
  + Execution environments
* Open source tools
  + Management
    - Relational databases
  + ETL, extract-transform-load (integration and transformation)
* Commercial tools
* Cloud based
* Libraries (Python)
  + Sci Comp
    - Pandas – data structure
    - NumPy – arrays
  + Visualization
    - Matplotlib – graphs/plots
    - Seaborn – heat maps, etc.
  + Machine Learning
    - Scikit-learn
    - Keras – deep learning
    - TensorFlow
    - PyTorch
  + Spark – cluster
    - Compliments:
      * Vegas
      * Big DL
  + R
    - Ggplot2
* API (application programming interfaces)
  + Let’s two softwares communicate
  + REST (rep state transfer)
    - Interact with web services
    - “resource” vs. “client”
* Data Sets
  + Def: structured collection of data
    - Tabular
      * CSV (comma separated)
    - Hierarchical
    - Network
    - Raw files
* Sharing enterprise
  + DAX
* Machine Learning Models
  + Predictions
  + Identifying patterns
  + Classes
    - Supervised
      * Labeled and trained
      * Regression
      * classification
    - Unsupervised
      * Data not labeled
      * Clustering/anomaly detection
    - Reinforcement
      * Based on human learning process
  + Deep Learning
    - Emulate human brain
    - Fraemworks:
      * Tensorflow
      * Pytorch
      * Keras
    - Repositories (model zoos)
      * ONNX
    - EX.
      * Prepare data
      * Build model
      * Train model
      * Deploy model
* Model Asset Exchange (MAX)
  + Reduces time to value
  + Model serving microservices