DATA SCIENCE CHEATSHEET

1. Learn Python and Jupyter

Python is one of the most popular programming languages using the data science field. Its the easiest to pick up and will give you maximum bang for buck. Focus on learning:

- Data structures - Base Types, Lists, Dictionaries,
- Looping - for, in, while,
- Control Flow and Operators
- Creating Functions and Classes

Jupyter Notebook and Jupyter Lab are the most common Data Science development environments for Data Science with Python.

2. Identifying data science tasks

Common Data Science Tasks and the Algorithms behind them
Sales Forecasting - Regression
Churn Prediction - Binary Classification
Customer Propensity to Buy - Regression
Market Basket Analysis - Association Rules
Sign Language Recognition - Object Detection
Defect Analysis - Semantic Segmentation
Human Pose Modelling - Posenet

How to find data science tasks?
FIRST: Search for "Machine Learning Examples for <Industry You're Interested In> industry"
THEN: Look for examples of those tasks on GitHub or Kaggle to get an idea of how they're structured

3. Understand the types of data you'll encounter and how to work with them

You'll encounter lots of different types of data during your journey as a data scientist. Its useful to know how to work with each of them.
Structured - CSV, Excel, SQL Views
Unstructured - Images, Video, Text, Audio

4. Analysing and Visualising your datasets

The most popular methods for analysing and visualising **STRUCTURED** data.

- Viewing and Transforming: Microsoft Excel, Pandas (Python library for data analytics, think Excel but using Python), Numpy (Python library for basic array and mathematical functions)
- Visualisation: Matplotlib (Most popular Python plotting library), Seaborn (Easy to use and great looking visualisation library)

The most popular methods for analysing and visualising **IMAGES** and **VIDEOS** are:

- Viewing and Transforming: OpenCV (ridiculously powerful computer vision Python library)
- Visualisation: Matplotlib (handles viewing images in a Jupyter Notebook using plt.imshow() method)

The most popular methods for analysing and visualising **TEXT** and **NATURAL LANGUAGE** are:

- Viewing and Transforming: NLTK, TextBlob and Spacy
- Visualisation: Matplotlib

The most popular methods for analysing and visualising **AUDIO** are:

- Viewing and Transforming: Scipy
- Visualisation: Numpy

5. Preprocessing and transformation

The core goal of preprocessing and transformation is to get your data ready for modelling. Learn how to perform these tasks:

- **Structured**: handling missing values, normalizing and scaling data, splitting your data into dependent and independent values and creating a training and testing dataset
- **Images**: checking your images are valid, labelling images using labelme and labelImg, performing image augmentation using OpenCV
- **Text**: removing punctuation, stripping out stop words, applying lemmatization and tokenization
- **Audio:** conversion of .wav files to spectrograms

6. Modelling, Algorithms and Evaluation

Supervised

- Structured Regression - Random Forest Regressor, Gradient Boosting Regressor
- Structured Classification - Random Forest Classifier, Gradient Boosting Classifier
- Image Classification - Keras Sequential Neural Network
- Object Detection - Tensorflow Single Shot Detector
- Semantic Segmentation - Tensorflow Mark R-CNN
- Pose Estimation - PoseNet
- Reinforcement Learning - Stable Baselines
- Sentiment Analysis - Text Blob Sentiment

Unsupervised

- Clustering - K-Means
- Anomaly Detection - One Class SVM
- Dimensionality Reduction - Principal Component Analysis

7. Deployment and Integration

Being able to deploy your models to cloud services allows you to integrate your work with other parts of the business or startup.
Cloud Machine Learning Providers: Watson Machine Learning, AWS Sagemaker, Azure ML

You should also get an understanding of how to deploy your models using Open Source tools including:

- FastAPI
- Django
- Flask

8. Domain Expertise and Presentation Skills

How to learn about your industry?
Read blog posts, financial reports and industry white papers. Look for data science examples in that industry.

How to improve your presentation skills?

- Join a toastmasters club
- Practice presenting at a meetup
- Make a YouTube video describing a project your built!

NICHOLAS RENOTTE