**FUNDAMENTALS of DATA SCIENCE**

(From Booz Allen Fundamentals of Deep Learning)

Identifying the unseen

Advanced computing enables us to make connections like never before. We can

teach machines to detect cancer, identify cybersecurity threats before they happen,

and optimize business operations – the possibilities are endless. So how can

machines continue to learn, without the dedication of abundant resources?

Deep learning, a sub-mechanism of artificial intelligence (AI), is a computing process that enables

machines to find patterns in data. Deep learning sifts through data and makes connections, identifying

odd trends that would be unfindable by the human senses. Imagine a world in which we could identify

risks before they happen, capitalize on trends that could save lives, and rely on machines to make

conclusions based off of data and fact, instead of emotion and perception – that’s the world of deep

learning.

**Data Science** is the analysis of diverse data sets (using statistics and programming). Data analysis that includes all kinds of data (structured/unstructured), anything that you can use to get insight and actionable steps out of your data.

Rare qualities and in high demand.

The ability to take unstructured data and find, order, meaning and value in that data.

For businesses, data-driven insights is an area with the greatest opportunity for growth and differentiation.

**EXPLORATORY DATA ANALYSIS**

Sets the stage for modeling

**Visual Exploratoration**

Exploration is always a critical first step in any good data analysis

Precursor to numerical exploration

Review Data, Check Assumptions, Check Anomalies, Data Suggestions

Determine if your data appropriate to reach your conclusions?

**Exploratory Graphs/Charts**

Information Dense: they communicate so much more so much faster

Check for Shape, Gaps, Outliers

Start with single/univariate distributions (histogram)

Then evaluate joint/bivariate distributions (scatterplot)

Look for **unusual** cases (exceptional values)

Look for **errors** in the data (check for patterns)

Look for **missing** (biased) data (check for patterns, impute)

Bar Plot

Histograms

For quantitative variables

Shows the shape of then distribution

Box Plots

For quantitative variables

Good for identifying outliers ( a number which is less than Q1 or greater than Q3 by more than 1.5 times the interquartile range ( IQR=Q3−Q1 ).

Scatter Plot (Pair Plot)

Shows associations between quantitative variables

Multicollinearity: try using fewer variables/features or combining features (based on domain knowledge)

Conclusion:

Use a method that you find quick and easy

When you’re going through your charts you are determining:

Do you have what you need to answer your questions?

Are there gaps in the data?

Are there exceptional cases (outliers)?

Are there errors in the data?

**Numerical (Statistical) Exploration**

First Graphics, then numbers

You’re still exploring, you’re not modeling the data

It’s good to get multiple perspectives on the data

Will the students be O.K. with the class content being different from what was posted?

Topic Outline

Course Introduction

Installing Anaconda

Introduction to Jupyter Notebooks

\*Brief Overview of Data Science

\*The Differences Between BI, Data Analytics and Data Science

Introduction to Python

Python Exercises

Statistics Fundamentals – I will only cover Descriptive Statistics. I will briefly mention Inferential Statistics and how it will be covered in the 3-day program.

Introduction to Pandas - I will briefly mention NumPy as I lead into I will briefly demonstrate Scikit-learn toward the end to pull everything together).

Pandas Exercise

\*The Data Science Process

Data Cleaning

Exploratory Data Analysis - I will briefly demonstrate Matplotlib while discussing exploratory data analysis.

Introduction to Scikit-Learn – I will briefly mention test/train split and demo creating and testing a model.

**CONCLUSION:**

(B-storm: You can analyze you company’s Salesforce data, or you can monitor social media through sentiment analysis about your company or some issue affecting your company.)

(Slides – Opening 2 Slides)

(Slide – AGENDA)

Data Science Overview

Data Analytics vs. Data Science

The Data Science Process

Python Libraries for DS

Next Steps

(Slide) Logistics

(Slide) Stanford & Big Data Trunk

(Slide) Big Data Trunk

--- Introductions

(Slide?) Jupyter Notebook (Hands-on Lab)

(Slide) Python Libraries for Data Analysis (w/only graphics)

(Slide) Data Science w/ Venn Diagram

(Slide) Data Analytics vs. Data Science

(Slide) Python

Notebook - Python

(Slide - Python Exercises)

Descriptive Statistics (SLIDES)

Pandas

Notebook – pandas

(Slide - Pandas Exercises)

Data Science Process (DIAPERS Slide)

(Slide) Define the Problem

(Slide) Ingest the Data

(Slide) Data Sources

(Slide) Analyze/Explore the Data

(Slide) Matplotlib

(Notebook - Matplotlib)

(Slide) Prepare the Data

(SLIDES) Clean the Data

(Slide) Evaluate the Model

(Slide) Scikit-learn

(Slide) Scikit-learn Interface

.fit(), .score(), .predict()

Notebook – Scikit-learn

(Slide) Next Steps

(Slide) Upcoming Course

(Slide) Thank you