Course Syllabus

Defining Data Science and What Data Scientists Do

* Defining Data Science
  + Process of using data to understand different things, translating data to make decisions
* What is Data Science?
* Fundamentals of Data Science
  + Can help orginazations
    - Understand their environments
    - Analyze existing issues
    - Reveal previously hidden opportunities
  + Adds value to org by investigating data and exploring the best way to apply it to the business
  + Make sure you have the correct questions
  + What data do we need and where will it come from
    - Structured and unstructured dada
  + When the data info reveals itself, the data scientist becomes a story teller
* The Many Paths to Data Science
  + Most Data scientists don’t come from specific background, though strong math helps
* Advice for New Data Scientists
  + Curious, extremely argumentative, judgmental
  + Familiarity with software
  + Ability to tell a story
  + Figure out where your expertise starts
* Data Science: The Sexiest Job in the 21st Century

What Do Data Scientists Do?

* A day in the Life of a Data Scientist
* Old problems, new problems, Data Science solutions
  + Organizations can leverage the almost unlimited amount of data now available to them in a growing number of ways
    - To find optimum solutions to existing problems
* Data Science Topics and Algorithms
  + Regression
    - Statistical relationships
  + Data visualization
  + Artificial neural networks
  + Complex isn’t always better
* What is the cloud?
  + Can be used to store data or use computing power
* What Makes Someone a Data Scientist?

Data Science Topics

* Foundations of Big Data
* “Big Data refers to the dynamic, large and disparate volumes of data being created by people, tools, and machines. It requires new, innovative, and scalable technology to collect, host, and analytically process the vast amount of data gathered in order to derive real-time business insights that relate to consumers, risk, profit, performance, productivity management, and enhanced shareholder value.”
* here is no one definition of Big Data, but there are certain elements that are common across the different definitions, such as velocity, volume, variety, veracity, and value. These are the V's of Big Data
  + Volume is the scale of the data, or the increase in the amount of data stored.
    - Divers are increases in data sources
    - Higher resolution sensors
    - Scalable infrastructure
  + Variety is the diversity of the data
    - Structured data fits neatly into rows and colums
      * Relational database
    - Unstructured data is nor organized in a pre-defined way
      * Ex tweets, blog posts, pictures
    - Also reflects data that comes from different sources, machines, people and processes both interal and exter to the org
    - Divers
      * Mobile tech
      * Social media
      * Wearable tech
      * Geo tech
      * Video
      * Ect ect
  + Veracity is the quality and origin of data/ its ability to conform to facts and accuracy
    - Xompleteness
    - Integrity
    - Ambiguity
    - Drivers
      * Cost
      * Need for traceablitity
  + Value, monetary or social benefits, or customer satisfactions
* How Big Data is Driving Digital Transformation
  + Digital transformation affects business operation updating existing processes and operations and creating new ones to harness the benefits of new tech
* What is Hadoop?
* Data Science Skills & Big Data
* Data Scientists at New York University
* Data Mining
* Quiz: Data Mining

Deep Learning and Machine Learning

* What's the difference?
* Neural Networks and Deep Learning
* Applications of Machine Learning
* Regression
* Quiz: Regression

Data Science in Business

* Applications of Data Science
* How Data Science is Saving Lives
* How Should Companies Get Started in Data Science?
* Applications of Data Science
* The Final Deliverable
* Quiz: The Final Deliverable

Careers and Recruiting in Data Science

* How Can Someone Become a Data Scientist?
* Recruiting for Data Science
* Careers in Data Science
* High School Students and Data Science Careers

The Report Structure

* The Report Structure
* Quiz: The Report Structure
* Final Assignment

Course 2.

Tools for Data Science

Visual programming & modeling

Phython

R

SQL

Data management is the process of persisting and retrieving data

Data integration and transformation, often referred to as Extract, Transform, and Load, or “ETL”, is the process of retriening data from remote data management statems

Transfromting data and loading it into local data management systems is also a part of data intergration and transformation.

Data visualization is an ipart of an initial data exploration process as well as being part of a final deliverable

Model building is the process of creating a machine learning or deep learing model using an appropriate algorithm with a lot of data

Model monitoring- accurate, fairness, and adversarial robustness

Code asset managemtn

Data asset manament – data government/ data lineage

Integrated Development environments or IDEs

ETL – extract, transform, load (data refining, data cleaning)

Jupyter kernels

Jupyter lab

Apache zeppelin

R Studio

Spyder

Apache Spark, cluster computing, linear scalability, batch processing

Apache flink, stream processing

Knime (r and phython)

Orange

Commercial Tools for Data Science

Oracle database, Microsoft SQL server, IBM DB2

Informatica and IMB Infospheres DataStage Watson Studio Desktop

Tableau, IBM Cognos Analytics, Microsoft Power BI, Watson Studio Desktop

Data mining- SPSS modeler, SAS MInner, Watson studio desktop

Data asset management - Informatica, IBM Infosphere Information Governance Catalog

Cloud based tools for Data Science

Microshopt Azure machine learning

H2O driverless ai

SaaS- Software as a service (maintaince done by the cloud owner)

Amazon DyamoDB (noSQL)

Coudant is basically CouchDB, but with maintenance

Data integration and tranformations, ELT tools, informatica, IBM data refinery

Model building – google cloud AI platform training , IBM machine learning

Model Deployment

Libraries for Data science

* Python Libraries
  + Scientific Computing Libraries in Python
    - Usually contain built inmodules prociding different fuctionalities you can use directly
      * Sometimes called framewords
        + Pandas offers data structer and tools for data

Cleaing

Manipulation

Annalasis

Two dimentional table consistion og columns and rows

Called dataFrame

Designed for easy indexing

* + - * + NumPy

Based on arrays

Enables mathamaticals to be applied

* + Visualization Libraries in Python
    - Great way to convey the meaningfulness of results
      * Matplotlib
        + Most well known library for data visualization

Great for graphs and plots

* + - * Seaborn
        + Based off of Matplotlib

Great for heatmaps, time series, violin plots

* + High Level-Machine learing and Deep Learning
    - Scikit-learn
      * Machine learning: regression, classification clustering, orhter
        + Built on NumPy, SciPy, and Marplotlib
    - Keras
      * Deep Learning Neural Networks,
  + Deap learning libraries in Physton
    - TensorFlow
      * Deep Learning Production and Deployment
    - PyTorch
      * Deep learing: regression classification
  + Apache Spark
    - Processes data through a network of computer
      * Similar to pandas NUmpy SciKit-lear
        + Can be used with Phython, RScala, or SQL
  + Other libraries in other languesas
    - Vegas
    - BigDL (deap learing
    - R Libraries
      * Ggplot2
      * Keras
      * Tensor Flow
      * R is being phased out by phython

Application programming interfaces

* What is an API
  + Lets two pieces of software talk to one another
* API Libraries
* Rest
  + Allow you to comnuicate using the internet
  + RE -Representational
  + S- State
  + T- Transfer
  + Request
    - Usually http message
    - JSON file
    - Sent to Service
  + Response

Data sets – Powering Data Science

* Data set
  + Structured collection of information
    - Tavbular data
      * Collection of rows and colloms
        + CSV File is a popular format

Comma Seperated Values

Each line represents a row

Colloms sepedrated by commas

* + - MNIST Dataset
      * It contains images of hanbdwritten digits and is commonly used to train image processing systems
  + Private data
    - Confidential
    - Private or personal information
    - Commercially sensitive
  + Open data
    - Scientific institutions
    - Governments
    - Organizations
    - Companies
    - Publicly available
  + Where to find open data
    - <http://datacaatalogs.org/>
    - Governmental, intergovernmental, and organization websites
    - Kaggle <https://www.kaggle.com/datasets>
    - Google <https://datasetsearch.research.google.com/>
  + Community Data License Agreement
    - CDLA
      * Sharing
        + Use and modify data

If you publish you must under the same terms of the original data

* + - * Permissive
        + Use and modify

Not required to share changes to the data

Sharing Enterprise Data

* Data Asset eXchange
  + DAX
    - Helps find open data sets for enterprises (commercial or research)
  + Model Asset eXchange

Machine learning Models

* Machine Learning and Deep learning models
  + What is a model
    - Data can contain a wealth of information
    - Machine learning uses model (algorithms) to identify patterns in data
    - A model must be trained on data before it can be used to make predictions
      * Supervised learning,
        + Data is labeled and model trained to make correct predictions
        + Humans provide input data and the correct outputs

Model then tries to identify relationships and dependencies between them

* + - * + Used to solve regression

Predict real numerical values

Home sales prices, stock market prices

* + - * + classification problems

classify things into categories

email spam filters, fraud detectors, image classification

* + - * unsupervised learning
        + data is not labeled by human
        + model tries to identify patterns without external help
        + Clustering

Are used to divide each record of a data set into one of a small number of similar groups.

EX Providing shopping recommendations

* + - * + Anomaly detection

Identifies outliers in the data set

EX fraudulent credit caard transactions or suspicious online log-in attempts

* + - * reinforcement learning
        + Conceptually similar to human learning process
        + Deep learning

Tries to loosely emulate how the human brain works

Applications

Natural language processing

Image audio video analysis

Time series forecasting

Much more

* + - * + Needs very large labeled data
        + Compute intensive
        + Can be built from scratch or downloaded off the internet

Public model repositories

Most frameworks provides a “Model Zoo”

Popular model zoos are provided by

Tensorflow

Pytorch

Keras

ONNX model Zoo

* + - * + Built using frameworks such as:

Tensorflow

Pytorch

Keras

The Model Asset Exchange (MAX

* Free Open-Souce Deep learning Microservices
  + Configured to use pre-trained or custom trained
    - Data+model+compute resources +domain expersites
    - Research, evaluate, test, train, and Validate model
  + Fully tested, deploy in minutes
* Available for variety of domains
  + Object detection
  + Image audio text classification
  + Named entity recognition
  + Image to text translation
  + Human pose detection

Overview Git/GetHub

* Preforming Version Controls
  + Keep track of documents
* Collaborating with others
* Free and open source
* Distributed version control system
* Accessible anywhere in the world
* One of the most common version control systems availed
* Can be used with
  + Code
  + Images
  + Documents
  + Any number of file types
* Glossary of Terms
  + SSH Protocol
    - Method for secure remote login from one computer to another
  + Repository
    - The folders of your project that are set up for version control
  + Fork
    - A copy of a repository
  + Pull request
    - The process you use to request that someone reviews and approves your changes before they become final
  + Working Directory
    - A directory on your file system, including its files and subdirectors, that is associated with a git repository
* Basic Git Commands
  + Init
    - Create?
  + Add
    - Moves changes from the working directory to the staging area
  + Status
    - Allows you to see the state of your working directory and the staged snapshop of your changes
  + Commit
    - Takes your staged snapshot of changes and commits them to the project
  + Reset
    - Undoes changes that you’ve made to the files in your working directory
  + Log
    - Enables you to browse previous changes to a project
  + Branch
    - Lets you create an isolated environment within your repository to make changes
  + Checkout
    - Lets you see and change existing branches
  + Merge
    - Lets you put everything back together again
* More resources <https://try.github.io/>