THE DATA INCUBATOR REPLY

COURSE CONTENT

This document contains comprehensive information regarding the Data Incubator Reply's (henceforth DIR) 8-week fellowship and training program, from program logistics to graduation and offboarding.

COURSE CONTENT

DIR currently offers in-depth training in seven different topic areas. Each area is designed as a self-contained learning module which can be deployed individually or as a group. Below is a description of each module.

WEEK	TITEL	DESCRIPTION
1	Data Wrangling for Data Science	The first step of data science is mastering the computational foundations on which data science is built. We cover thefundamental topics of programming relevant for data science, including pandas, numpy, scipy, matplotlib, regular expressions, sql, json, xml, checkpointing, and web scraping that form the core libraries around handling structured and unstructured data in Python. Trainees gain practical experience manipulating messy, realworld data using these libraries. They also walk away with a firm understanding of tools like pip, git, ipython, jupyter notebooks, pdb, and unit testing that leverage existing open source packages to accelerate data exploration, development, debugging, and collaboration.
2	Introduction to Machine Learning	In world with abundant data, leveraging machines to learn valuable patterns from structured data can be extremely powerful. We explore the basics of machine learning, discussing concepts like regression, classification, model evaluation metrics, overfitting, variance versus bias, linear regression, ensemble methods, model selection, and hyperparameter optimization. Trainees come away with a strong understanding of the core concepts in machine learning and the ability to efficiently train and benchmark accurate predictive models. They gain hands-on practice with powerful packages like scikitlearn, building complex ETL pipelines to handle data in a variety of formats and techniques, developing models with tools like feature unions and pipelines that allow them to reuse existing models and reduce duplicate work, and practicing tricks like parallelization to speed up prototyping and development.

3	Introduction to Distributed Computation and Hadoop	With the advent of big data, diskspace, memory, and computational resources of a single computer are no longer sufficient. We introduce the basic concepts around distributed computing spreading out workloads over multiple computers. Topics covered include hadoop, HIVE, partitioning, fault tolerance, hadoop streaming, and mrjob. Trainees will walk away with a firm understanding of distributed computing paradigms, how to efficiently break up a workload across multiple nodes, and how to select between competing distributed computing paradigms. Trainees gain direct hands-on experience building, debugging, and deploying MapReduce jobs to run on large, real world data in the cloud (AWS EMR).
4	Exploratory and Explanatory Data Visualization	Data science is about helping humans understand the story behind the data and visualizations provide a powerful tool for telling that story. We delve into the use of visualizations as a tool for data exploration, using Python plotting libraries like matplotlib, seaborn, pandas, and bokeh to highlight the process of dissecting a large dataset to pull out meaningful relationships. We discuss the biases and limitations of both visual and statistical analysis to promote a more holistic approach. Trainees are also exposed to explanatory data visualization tools built on d3.js, and come away with a solid understanding of visualization design best practices, interactivity, scalability, as well as the engineering of browserbased delivery platforms and constructing dashboards. They apply that knowledge to creating a beautiful, functional web app on Heroku that incorporates realtime queries and user interaction.
5	Capston Project	The Capstone Project is a valuable component of the work trainees do at the Incubator program. Trainees will apply the techniques they have learned during the program to a real-world problem and will develop the necessary softskills to effectively communicate their results.
6	Advanced Machine Learning and Unstructured Data	While machine learning on structured data lays an important foundation, handling unstructured data and advanced machine learning techniques open up a larger world of analytical opportunities. We explore more advanced machine learning techniques such as support vector machines, decision trees, random forests, neural nets, clustering, KMeans, expectation maximization, word2vec, and handling unstructured data. Trainees come away with intuition about the suitability of different techniques for different problems. In addition to handling structured data, trainees directly apply these techniques to large volumes of real-world unstructured data, solving problems in natural language processing, bag of words, feature hashing, and topic modelling.
7	Distributed Computing with Spark and Scalding	Scala, Spark, and Scalding are technologies at the forefront of distributed computing that offer more abstract but more powerful APIs. The module focuses on the basics of Scala like map, flatmap, for comprehension, data structures, and core concepts of Spark like resilient distributed datastores, memory caching, actions, transformations, and distributed machine learning. Trainees come away with a solid understanding of the basics of Scala and Spark as well as critical tooling around Spark (sbt, jvm) to make them more productive. They apply that knowledge to directly developing, building, and deploying Spark jobs
8	Data Science for Business	Sometimes the most important question to ask in data science comes from thinking beyond the data itself. Important topics include data fidelity, relevance, and the value of additional data. Bias is a major theme, and trainees think about how their conclusions are influenced by data collection, external factors, internal structuring, procedural artifacts, and more. Trainees gain a broader understanding of how to balance trade-offs to suit the business problem, such as when to favor accuracy over interpretability and vice versa. The goal is to apply this knowledge to case studies that simulate what they would be expected to contribute as part of a real-world team faced with a business problem.

COURSE MODULES / COURSE NOTES

For each module, DIR provides interactive course notes, including introductions, tutorials, and sample code that trainees can play with both in and out of lecture, to begin grasping each module's material. Additionally, these course notes serve as a skeleton from which instructors structure their lectures, and as a comprehensive reference of each topic while trainees complete hands-on work via the mini-projects (see below).

MINI-PROJECTS

DIR takes an applied learning approach towards data science education. Lectures are important in introducing new concepts but do not cement practical hands-on data science skills. Our course is centered around mini-projects. These are fun, interactive projects that require trainees to accomplish a significant amount of work similar to what they might be asked to do on the job. Each mini-project reinforces the concepts taught in lecture and teaches trainees how to apply the tools they've learned. All mini-projects make use of real-world, open-source data and provide trainees with a place to practice, develop, debug, and deploy their newfound knowledge.

MODULE	MINI-PROJECT	DESCRIPTION
1	Graph	Trainees will begin to familiarize themselves with the Python analytics toolkit by graphing the social network of New York City's rich and famous.
1	SQL	Trainees will work with a large database of New York restaurants' inspections, and will build, manipulate and operate tables in order to extract relevant information.
2	ML	Trainees will take Yelp reviews and, based on various characteristics, build predictive models for a given restaurant's Yelp score.
3	MR	Using the English language Wikipedia, trainees will familiarize themselves with Hadoop and MapReduce technques in order to compute a variety of interesting statistics, facts, and relationships within the wikipedia dataset.
4	DataViz	Trainees will learn core visualization techniques, technologies, patterns, and best practices through webbased technologies such as D3.js, leaflet.js, and Bokeh. Trainees will visualize historical NYC Bus time, location, and schedule data.
6	TS	Trainees use historical weather data and timeseries analysis to build a weather model that predicts future temperatures.
6	NLP	Trainees revisit the Yelp review dataset and use its unstructured text data to build sophisticated, powerful predictive models of restaurant scores and uncover interesting relationships between words and phrases in the corpus.
6	Music	Trainees will develop models that are able to recognize the genre of a musical piece, first from pre-computed features and then working from the raw waveform. This is a typical example of a classification problem on time series data.
7	Spark	By examining the StackOverflow dataset, trainees familiarize themselves with Spark's computational workflow and analytic capabilities. This miniproject also ties together distributed computing techniques learned in MR and analytic techniques learned during the ML weeks.
8	DSB	Trainees will apply probability and statistics to be able to say something meaningful about Yelp ratings for businesses primarily in AZ and NV.

DAY-TO-DAY AND COURSE MODULE

The foundation of our Data Science training is our daily lectures. They set the stage for the skills that will be learned and demonstrated through hands-on application later in the day. DIR holds these lectures at the beginning of the morning and afternoon sessions, and the rest of the day is mainly spent in collaboratively work on the mini-projects. In the course notes, each module contains enough material for 5 days of lecture, divided so that each "notebook" contains enough material for one day of lecture. In some cases, the instructor will take extra time on some notebooks, and less time on others. Here is a sample daily schedule for the 8-week program:

+ 50' Coding challenge
+ 20' Discussion of coding challenge's solution
+ 60' Morning Lecture
+ 90' People working in mini-projects / capstone project
+ 60' Lunch break
+ 60' Afternoon Lecture
+ 120' People working in mini-projects / capstone project
+ 60' Partner panels / Soft-skill lecture / Miniproject discussion

REPLY DATA

Data Incubator Reply (DIR)
Data Reply
DE: www.reply.com/de/data-incubator
EN: www.reply.com/en/data-incubator
IT: www.reply.com/it/data-incubator

DIR Germany
Luise-Ullrich-Straße 10
80636 Munich - Germany
T. +49 89 411 142-600
dataincubator@reply.de

DIR Italy
Via Robert Koch, 1/4
20152 - Milano - Italy
T. +39 02 535761
dataincubator@reply.it