The Big Data Analyst Blueprint

Data analysis is a multi-disciplinary pursuit that consists of a whole lot of different skills. The Dev League Big Data Analyst Track is designed to help you build the skills that will be most useful to you leveraging the skills and experiences you bring to the course. The blueprint below outlines commonly needed skills in the field of data analysis. Your first job will be to customize this blueprint into your personal "Skills Backlog". This skills backlog will describe the skill you want to develop during each sprint of the course. Critically, it will also indicate why this skill will be useful to your career goals and, if at all possible, how it incorporates your incoming knowledge and projects.

Sub Specialties	We have identified 4 "tracks" or sub-specialities within the Data Analysis field that can help you focus your career and skill building objectives. An analyst usually specializes in one of these tracks. Data Scientists often have all of these skills with deep expertise in one or more. The core analytical objective of each track is the same: create value and insight from data. The difference is in the focus or approach. We've set up Data Journalists to focus most on the investigative aspects of data analysis. Our Data Engineer track focuses on technology, implementation, and programming of data systems. Our Statistical Modeler track is the most theoretically focused of the tracks. And the Business Analyst is focused on applying analyses to decision making.
Modules and Sprints	The 40 week course is divided into ten (10) 4-week modules, with each module consisting of two 2-week sprints. In all, the course consists of twenty (20) 2-weeks "Sprints". The purpose of each sprint is to build a specific skill and demonstrate it with a project that you add to your portfolio. Before the class starts, with the help of your coaching team, you will outline the skills you want to develop in each sprint. During the sprint, you will work with the coaching team and your cohort to select and develop a project that both helps you learn a new skill and demonstrate it to future employers / teammates / clients in your public portfolio.
User Stories	Each skill you select for your "Skills Backlog" will be expressed as a "User Story". Articulating features as user stories is a technique used in software project management to keep development focused on end-users and application to real-world problems. They're a useful way to keep your own development focused on your ultimate career objectives. Your Skills Backlog user stories keep you focused but also lets your coaching staff and cohort know how they can support you. A user story is expressed in the form: "As a ______ I need _______ so that ________". There is a suggested skill for each track in each sprint expressed as a user story in the blueprint below. You can select any one that seems relevant to your objectives or customize a new one.

		as a Data Journalist	**as a Data Engineer**	**as a Statistical Modeler**	**as a Business Analyst**
Acquiring Data and Data Formats	"Getting" the data you want to analyze means downloading it from websites, connecting to and querying databases, extracting it from HTML webpages, interfacing with APIs (application programmer interfaces) importing and exporting files, and converting back and forth between data formats. Programming languages, databases, command line-based applications and graphical applications each have something to offer. **Tools: Excel, Python, Command Line | Pre-reqs: git, Jupyter Notebooks**				
	Sprint 1 (weeks 1-2) Data Formats and Terminology	...I need to identify data formats to successfully load it into tools and investigate it	...I need to be able to be able to programmatically read, write, edit, and convert data files so that my tools can work with data sources	...I need to understand the basic terminology and structure of data so that I can apply statistical analyses	...I need to understand the story of where the data came from so I know how it is relevant to my action or recommendation
	Sprint 2 (weeks 3-4) Connecting to Data Sources	… I need to construct datasets from web resources, so that I can investigate issues where data is not readily available	… I need to understand the full technology operating system stack and ecosystem, so that I can interact with tools	...I need to understand basic scripting so that I can save repeatable analyses and work	… I need to be able to connect to my corporate databases, APIs, and data warehouses so that I can make use of available data resources
Basic Data Maniuplation	Before you engage in structured analysis, you often just want to "see" the data. This can mean pre-viewing a subset of it, summarizing the columns/attributes/features, sorting or reorganizing it and otherwise finding ways to immerse yourself in your data. Again, each data tool has something to offer, and our objective is to develop a good sense of the utilities available to you. **Tools: R, Python, Command Line | Pre-reqs: git, Jupyter Notebooks, some programming**				
	Sprint 3 (weeks 5-6) Data Operations	… I need to filter, search, and remove features from my data set so that I can conduct targeted investigations	… I need to understand the programming basics of automation so that I can develop tools that work efficiently (also big o notation, computability and complexity, sorting and searching algorithms)	…I need to understand the programming basics so that I can implement algorithms	…I need to transform data or derive additional statistics from input data, so that I can highlight more telling indicators
	Sprint 4 (weeks 7-8) Data manipulation Libraries and Tools	… I need to be able to convert published research and analysis from Excel / R / Python into a different tool so I can verify and audit the analysis	… I need to understand the basic data structures in Python so that I can diagnose and troubleshoot performance issues	…I need to understand the NumPy arrays and Pandas / R dataframes so I can supply data to algorithms, fit models, etc	… I need to understand how to export my advanced excel skills to R / Python so that I can build more powerful analyses on top of what I already know
Exploratory Data Analysis	Data analysis is built around questions, and exploratory data analysis helps you know what questions to ask. Descriptive statistics and basic visualizations that summarize features or suggest relationships inspire the generation of hypotheses to confirm with statistical tests or build into statistical models. **Tools: R, Python, Command Line | Pre-reqs: git, Jupyter Notebooks, some programming**				
	Sprint 5 (Weeks 9-10) Summarizing and Describing	… I need to summarize the data I have so that I can report basic findings	…I need to identify errors and inconsistencies in the data so that I can develop solutions to address them, and possibly their source	…I need to produce basic visual plots and summary statistics of the central tendencies and range of my data set so that I can develop an intuition for and a familiarity with my data set	… I need to construct inventories and quality assessment of the data available so that I can propose high value ways to use the data assets
	Sprint 6 (Weeks 11-12) Preliminary Findings and Hypotheses	…I need to identify interesting patterns so that I can direct further investigation	…I need to understand the volume and data types of data to understand their performance implications	…I need to produce statistical summaries that explain how variables in my data set relate to each other, so that I can develop hypotheses to guide my analysis	…I need to produce preliminary charts and dashboards so that I can communicate with other areas of the business about problems we need to solve with joint expertise and refine data collection based on feedback
Experimental Design and Research Methods	Experiments and the scientific method are at the heart of how we "know" what we know when it comes to data analysis. But how does it translate to the different situations we encounter in practice and what are some common pitfalls to be aware of? **Tools: R, Python, Command Line | Pre-reqs: git, Jupyter Notebooks, some programming**				
	Sprint 7 (Weeks 13-14) Sampling, Instruments and the Bias introduced by both	… I need to understand the different ways to study sample populations and the potential biases introduced so that I can assess the value of published research into my investigations	… I need to be able to implement valid sampling and collection procedures for data at all scales so that I can support analyses without inadvertently introducing bias	… I need to understand how sampling and instruments introduce bias so that I can design analyses that account for them	…I need to design effective data collection instruments so that I can answer critical questions for my business
	Sprint 8 (Weeks 15-16) Implementing Tests	… I need to undertstand how causation is established in scientific studies so that I can intrepret studies and focus my analyses	… I need to be able to implement experiment-driven algorithms such as A/B testing and Epsilon Greedy) so that I can provide a testing capability	… I need to understand how to isolate factors and design appropriate experiments so that I can answer a wide range of research questions	… I need to identify opportunities to test and optimize with techniques such as A/B Testing and Epsilon Greedy so that my organization can continuously improve
Probability Theory	Probability theory is the foundation for the inferential statistics we use to test hypotheses and critical to understanding the models and predictions we derive from data. An intuition for probability is an indispensable tool for effective analysis, and a rock-solid ability to explain it to non-statisticians is essential in virtually any real-world application. **Tools: R, Python, Command Line | Pre-reqs: git, Jupyter Notebooks, some programming**				
	Sprint 9 (Weeks 17-18) Random/Stochastic Prcesses and Variables	… I need to understand what kind of real world phenomena produce which probability distributions so I can recognize them as noteworthy patterns	… I need to be able to simulate data for common probability distributions so that I can synthesize data where needed	… I need to understand the probability structures and sequences that produce common probability distributions so that I can properly model phenomena in my analyses	… I need to understand which factors that drive my business are subject to random variation, and what drives the variation so that I can model them

	Sprint 10 (Weeks 19-20) Applying Probability Models	… I need to be able to apply Bayes Theorem when new evidence is gathered so that I can update my understanding in a given investigation	… I need to be able to implement the application of probability models such as regression, Monte Carlo simulations and Markov Chain Models so that modeled phenomena can be useful in practice	… I need to understand the different kinds of probability models and different techniques on how to apply each of them so that I have a target in mind when I create them	… I need to be able to interpret the output of probability models such as regression, markov chains, and Monte Carlo simulations so that I can understand scenarios, ranges of outcomes, and risk as it relates to my organization and its processes
Inferential Statistics	Inferential statistics allow us to infer something about an entire "population" by measuring only a sample as well as giving us the tools we need to test our hypotheses about differences, relationships, (focuses on superpopulation inference, how to work with samples, models that have interpretable parameters). **Tools: R, Python, Command Line | Pre-reqs: git, Jupyter Notebooks, some programming**				
	Sprint 11 (Weeks 21-22) Population Estimates and Hypotheses	… I need to understand how population characteristics are inferred from their samples so I can draw accurate conclusions about third party research as well as my own analysis	… I need to understand the computing and analytical performance tradeoffs between different levels of sampling so that I can optimize for different objectives	… I need to understand the kinds of statistical hypotheses I can make as well as the tests they apply to so that I can answer a variety of research questions	… I need to understand how to construct a testable hypothesis about the populations represented by my business data so that I can drive strategic decisions about novel scenarios
	Sprint 12 (Weeks 23-24) Linear Regression Models	… I need to create regression models of data I am investigating to describe the relationships between key factors in my investigation	… I need to know how to implement different regression estimation methods so that I understand their performance characteristics	… I need to understand the different methods for estimating regression models and their relative tradeoffs so that I can efficiently arrive at a model that works for my purposes	… I need to create regression models that describe the relationships between key factors in my business so that I can use that information to drive decision making
Machine Learning	Summarizing the analysis of data into a mathematical or algorithmic model – that explains relationships between different data features, or predicts some features given others – is the culmination of all of the preparatory analytical steps described above. The model serves as the basis for the data product that applies the newly gained insight to the real world. **Tools: R, Python, Command Line | Pre-reqs: git, Jupyter Notebooks, some programming**				
	Sprint 13 (Weeks 25-26) Machine Learning Capabilities	… I need to understand how to use natural language and text processing tools, particularly research summarizing and assistance to enhance my investigation capabilities	… I need to understand the basic implementation of all sci-kit learn algorithms so that I can understand their performance characteristics	… I need to understand the full range of supervised and unsupervised machine learning techniques so that I can apply them to a broad range of problems	...I need to understand the most common uses of machine learning in business so that I can identify opportunities to leverage data assets
	Sprint 14 (Weeks 27-28) Machine Learning Optimization	… I need to understand how to train my machine learning models with different data so they perform better	… I need to know how to set up and implement the testing of models for their accuracy and performance so I can support model optimization	… I need to understand how to tweak hyperparameters, elect appropriate accuracy / error measures and use other techniques so that I can generally optimize model performance	...I need to understand how to tweak the data my organization collects so that my models perform better
Data Governance	We are learning just how powerful data is, and like with any powerful tool, we must understand the dangers inherent in it use. Who is affected? What are the negative consequences of ungoverned data? What can we do to protect against those consequences? **Tools: R, Python, Command Line | Pre-reqs: git, Jupyter Notebooks, some programming**				
	Sprint 15 (Weeks 29-30) Sanitizing Data	… I need to know how to structure a data set that is sanitized so that my data requests are more likely to be supplied	...I need to know how to structure data systems that can produce aggregated or depersonalized data so that we can ensure the privacy of data subjects	… I need to know how personally identifiable data can be constructed from multiple non-identifying data sets so I can advise the team on how to sanitize.	… I need to know how to strip personally identifiable data out of data sets so I can safely share my working data and analyses with others
	Sprint 16 (Weeks 31-32) Securing Data	… I need to know how data should be secured so if I am sharing sensitive data received as part of an investigation I can protect information subjects	… I need to know how to store personal and sensitive data in a secure way so that I can build systems that protect the information subjects	… I need to know how to secure data so that I can share powerful data with collaborators while protecting the information subjects.	… I need to know how to secure data so I that I can protect the subjects of any research (market or otherwise) that I am conducting
Production Development	The computation, memory, and storage demands of a prototype can be drastically different from an implementation of the same predictive model or analytical process out "in the wild". Building robust data software that can scale is a big part of taking advantage of big data. **Tools: R, Python, Command Line | Pre-reqs: git, Jupyter Notebooks, some programming**				
	Sprint 17 (Weeks 33-34) Storing and Computing in the Cloud	… I need to know how to store large datasets in the cloud so that I can investigate large datasets that do not fit on my personal computer	… I need to know how to set up production environments in the cloud so that my team can flexibly deploy data products.	… I need to understand how to access cloud-based storage and computational resources to be able to perform more resource intensive analyses	… I need to know how to access cloud-based storage and computing resources so that I am not constrained by my company's software and infrastructure
	Sprint 18 (Weeks 35-36) Performance	… I need to know how to publish data tables and visualizations to the cloud so that I can enhance the quality of my analysis communication	… I need to know how to break down and individually monitor the performance of the individual components of my data system so that I can optimize overall performance	… I need to understand the relative performance characteristics of different tools, techniques, and libraries so that I can optimize my analysis time and computing	… I need to understand the performance characteristics of and bottlenecks in the various elements of my corporate data systems so that I can design tools that are responsive and don't take up critical company resources
Data Products	Many of the most ubiquitous uses of data – Netflix content recommendations, Amazon purchase suggestions – don't appear to be data at all. To be successful with clients, managers, and customers that don't "speak data" it's essential to turn the output of your analysis into actionable insight or value-adding products. **Tools: R, Python, Command Line | Pre-reqs: git, Jupyter Notebooks, some programming**				
	Sprint 19 (Weeks 37-38) Interactive Tools & Dashboards	… I need to know how to publish data tables and visualizations to the cloud so that I can enhance the quality of my analysis communication	… I need to be able to create cloud-based services and APIs for data resources so that I can design and supply more flexible data systems	… I need to be able to share my research and analyses in interactive and reproducible formats online so that I can solicit feedback and input from a distributed network of experts	… I need to be able to create tools and dashboards from my analyses so that teams within my organization can work more effectively
	Sprint 20 (Weeks 39-40) Integrated Analytical Tools	… I need to create interactive tools that take user input so that I can contextualize the insight from investigations	… I need to know how to connect models developed and trained on local training data to cloud-based computing and storage resources to apply the analyses to a production scale system	… I need to be able to package my analytical work into libraries so that my work is accessible to other projects	… I need to create automated tools such as recommendation engines and classifiers so that analyses can be integrated into end-user products
		Data Journalist	**Data Engineer**	**Statistical Modeler**	**Business Analyst**
	What's Next	Advanced data and statistical communication involves a sharp command of statistical and probabilistic principles, a thorough knowledge of visual processing and user interface design, and an artistic and aesthetic sensibility that allows one to reveal hundreds or thousands or millions of simultaneous narratives embedded in data of all kinds. The Data Journalist can build on the core with projects that explore public data, interactive visualizations, data-driven essays and investigative reporting skills.	It's said that advances in data science are not more sophisticated models, but existing models that are powered by more data. Data Engineers make more data possible. Leveraging cloud computing resources, multi-core/processor/appliance architecture, and functional programming languages, data engineers put the "Big" in Big Data Analyst and build on the core skills with advanced programming and architectural magic.	The most productive frontiers of AI are not in traditional rules-based programming techniques, but machine learning. Training machines with data is likely to be the future of our most significant technological advances. Adding a portfolio of deep learning, reinforcement learning, and natural language processing will put you in position to help companies take advantage of that future.	Some of the best, most interesting data is collected by private companies who all need talented, analytical individuals and teams to help them turn their data into value for their products, operations, or strategy. Business Analysts build on the core with developing some literacy in accounting, the language of business and examining industry specific analytical challenges and needs.