IBM **Analytics**

The Data Science Skills Competency Model

A blueprint for the growing data scientist profession

Contents

Although data science as a field has existed for several decades, the rapid growth of artificial intelligence (AI) in business in the last five years has generated a demand for data scientists that far surpasses the availability of trained professionals. Today, 63% of executives cite a lack of talent as a prime barrier to adopting AI technology.[1]

This talent gap is an opportunity for aspiring professionals, and a challenge for companies striving for a competitive advantage in the market. For prospective data scientists and organizations building a data science team, gaining the necessary skills required can be a formidable obstacle. Formalizing necessary skills helps academic institutions, data scientists, hiring managers and resource talent development teams deliver on the promise of data science, machine learning (ML) and AI.

Lack of clarity in a growing field

In the 2018 August LinkedIn Workforce Report: Data Science Skills are in High Demand Across Industries report, LinkedIn reported that there were more than 151,000 unfilled data scientist jobs across the US, with “acute” shortages in New York City, San Francisco and Los Angeles. Combined with a 15% discrepancy between job postings and job searches on Indeed, demand for data scientists clearly outstrips supply.[2]

To make matters worse, there has been a lack of consistency on the skills required from candidates to fill a job. In some cases, job ads are too qualification-intensive, making it difficult to match skills to the job. In other cases, the candidates without the right qualifications are applying and being recruited as data scientists.

This situation has left companies at risk of losing valuable time and opportunity and disappointing results from poor implementations. And, in worst cases, when data science, ML and AI techniques are used incorrectly, they’re at risk of legal exposure. For aspiring data scientists looking for better jobs, this lack of consistency has left them unsure about which skills to develop to be successful in their careers.

What skills does a data scientist need to be successful?

Data science is a cross-disciplinary set of skills found at the intersection of statistics, computer programming and domain expertise. It comprises three distinct and overlapping areas:

1 **Statistics**, to model and summarize data sets
2 **Computer science**, to design and use algorithms to store, process and visualize data
3 **Domain expertise**, necessary to formulate the right questions and to put the answers in context

Other skills often missed are:

– Leadership
– Teamwork
– Communication

The data scientist engages in or leads the AI enterprise workflow, as shown in Figure 1. They must:

– Understand the business opportunity
– Work with data engineers and the IT department to find the right data sources
– Prepare data and build ML and specialized AI models
– Assist in the deployment of models into the operations of the organization
– Measure success and communicate that back to the business

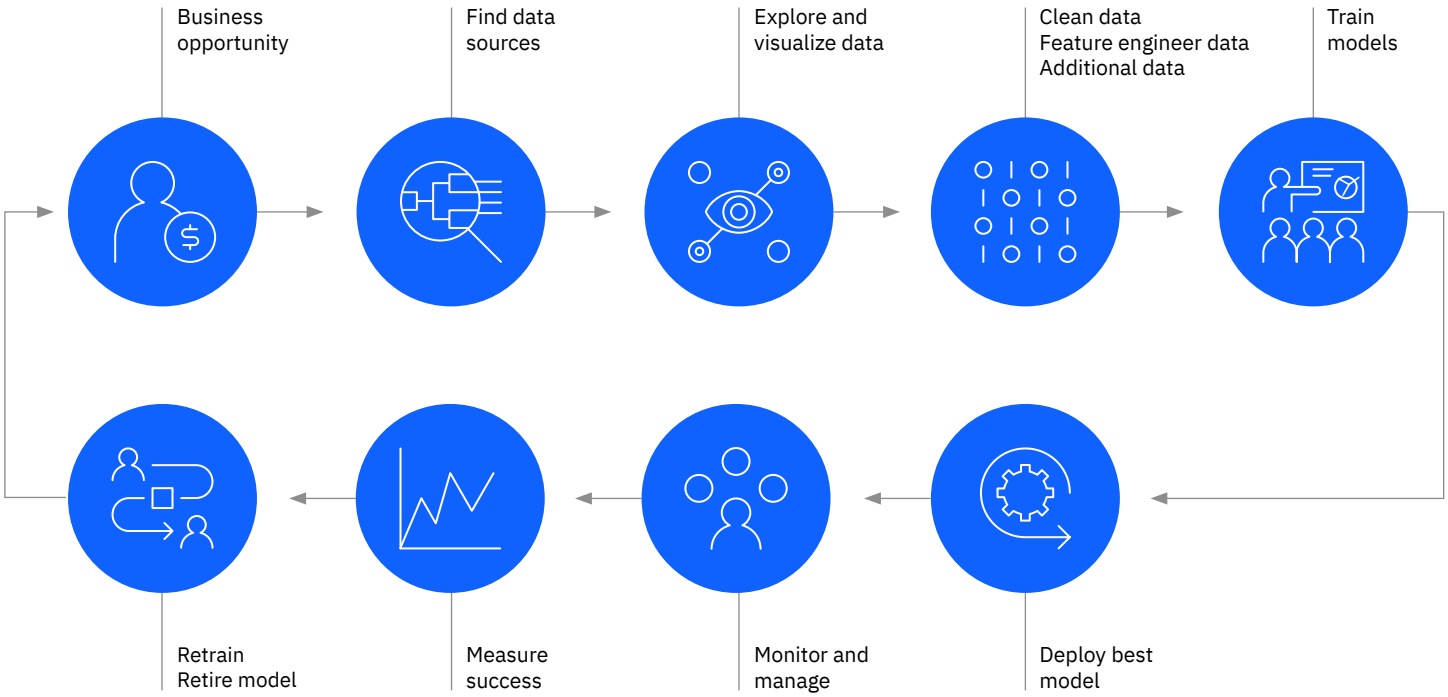

Figure 1. AI enterprise workflow

The Data Science Skills Competency Model

In 2018, IBM built the first Data Science Apprenticeship program in the United States. As part of the program, Ana Echeverri, Director and principal leader for AI Skills Learning and Certification at IBM, developed a transparent blueprint defining the skills competencies required for a data scientist. This blueprint can be used for:

- Recruitment
- Skills development
- Job expectations

This model, approved by the U.S. Department of Labor, can be found in the *Competencies and performance criteria* section of this paper. It's organized into the following seven areas, combining foundational skills with enterprise data science workflow skills:

1. **Statistics and programming foundation.** The competencies in this area are focused on the knowledge of key statistics concepts and methods essential to finding structure in data and making predictions. Further, job candidates must have Python programming skills—or other statistical programming skills—and the ability to visualize data, extract insights and communicate the insights in a clear and concise manner.

2. **Data science foundation.** A data scientist must be able to:
   - Characterize a business problem
   - Formulate a hypothesis
   - Demonstrate the use of methodologies in the analytics cycle
   - Plan for the execution

   Understanding the data science workflow and recognizing the importance of each element of the process is critical for successful implementations.

3. **Data preparation.** To ensure the data scientist can construct usable data sets, the key competencies required are:
   - Identifying and collecting the data required
   - Manipulating, transforming and cleaning the data

   A data scientist must also demonstrate the ability to deal with data anomalies such as missing values, outliers, unbalanced data and data normalization.

4. **Model building.** This stage is the core of the data science execution, where different algorithms are used to train the data and the best algorithm is selected. A data scientist should know:
   - Multiple modelling techniques
   - Model validation and selection techniques

   What differentiates a data scientist is understanding the use of different methodologies to get insight from the data and translating the insight into business value.

5. **Model deployment.** An ML model is valuable when it's integrated into an existing production environment and used to make business decisions. Deploying a validated model and monitoring it to maintain the accuracy of the results is a key skill.

6. **Big data foundation.** Organizations deal with large volume of structured and unstructured data. A data scientist must demonstrate understanding of how big data is used, the big data ecosystem and its major components. The data scientist must also demonstrate expertise with big data platforms, such as Hadoop and Spark.

7. **Leadership and professional development.** Data scientists must be good problem solvers. They must understand the opportunity before implementing the solution, work in a rigorous and complete manner, and explain their findings. A data scientist needs to understand the concepts of analyzing business risk, making improvements in processes and how systems engineering works.

In addition to detailing the competencies, the model provides foundational performance criteria. Candidates and programs—such as data science undergraduate or graduate programs, or a set of courses in a massively open online course (MOOC)—can be evaluated using a common measurement system.

The IBM Data Science Apprenticeship program

In February 2019, IBM announced the Data Science Apprenticeship program. This program is part of IBM's "New Collar" jobs initiative for job candidates who may not have a college degree. It consists of three main components:

- Education
- Mentorship
- Practical experience

Over the course of the apprenticeship, employees work to meet the requirements to reach Level 1—Certified Data Scientist—of the Open Group Professional Certification Program for the Data Scientist Profession. This new program is recognized by the U.S. Department of Labor as a professional apprenticeship.

The competencies also help build the right work experiences, projects and mentoring relationships so aspiring data scientists can meet the skills requirements for future jobs.

Conclusion

The rapid growth of AI in business in the last five years presents an opportunity for professionals, and a challenge for companies striving for a competitive advantage in the market. Understanding the skills required to be successful is a common obstacle for aspiring data scientists and the organizations seeking to build a data science team.

The Data Science Competency Model identifies and defines the skills required by a data scientist to be successful within the enterprise data science workflow. Organizations will have a model to guide the selection or development processes for data scientists for today's competitive environment.

Competencies and performance criteria
Foundational competencies

Statistics and programming foundation

1.0 Understand sampling, probability theory, and probability distributions
2.0 Demonstrate knowledge of descriptive statistical concepts
3.0 Demonstrate knowledge of inferential statistics
4.0 Demonstrate knowledge of Python programming skills
5.0 Implement descriptive and inferential statistics using Python
6.0 Demonstrate ability to visualize data and extract insights
7.0 Demonstrate through a project the ability to analyze a dataset and communicate insights

Data science foundation

8.0 Demonstrate understanding of what is data science and what data scientists do
9.0 Demonstrate ability to characterize a business problem
10.0 Demonstrate ability to formulate a business problem as a hypothesis question
11.0 Demonstrate use of methodologies in the execution of the analytics cycle
12.0 Demonstrate through a project the ability to plan for the execution of a project

Data preparation

13.0 Demonstrate ability to identify and collect data – multiple formats
14.0 Demonstrate ability to manipulate, transform, and clean data
15.0 Demonstrate expertise with techniques to deal with missing values, outliers, unbalanced data, as well as data normalization
16.0 Demonstrate through a project the ability to construct usable data sets

Model building

17.0 Demonstrate understanding of Linear Algebra principles for machine learning
18.0 Demonstrate understanding of different modeling techniques
19.0 Demonstrate understanding of model validation and selection techniques
20.0 Communicate results translating insight into business value
21.0 Demonstrate through a project the ability to test different models on a dataset, validate and select the best model, and communicate results

Model deployment

22.0 Deploy and monitor a validated model in an operational environment
23.0 Demonstrate through a project the ability to deploy and use a deployed model

Big data foundation

24.0 Understand the concept of big data, and how big data is used at organizations
25.0 Understand with the big data ecosystem and its major components
26.0 Demonstrate through a project expertise with big data platforms (Hadoop, Spark)

Leadership and professional development skills

27.0 Participate as a data scientist on client engagements (internal or external)
28.0 Contribute to the profession by teaching or mentoring others

Competencies and performance criteria

Foundational performance criteria

Competency outcomes	Assessment criteria	Evidence type
1.0 Understand sampling, probability theory, and probability distributions	1.1 Understand and apply different sampling techniques and ways to avoid bias 1.2 Understand the concepts of probability, conditional probability, and the Bayes' theorem 1.3 Demonstrate knowledge of distributions such as the normal distribution and binomial distribution	– Assignments, projects, case studies – Expert witness evidence – Mentor testimony – Observation – Questions and answers – Recognition of prior learning
2.0 Demonstrate knowledge of descriptive statistical concepts	2.1 Identify definitions of central tendency and dispersion (mean, median, mode, standard deviations) 2.2 Demonstrate knowledge about working with categorical data vs. numerical data 2.3 Recognize the difference between descriptive and inferential statistics	– Assignments, projects, case studies – Expert witness evidence – Mentor testimony – Observation – Questions and answers – Recognition of prior learning
3.0 Demonstrate knowledge of inferential statistics	3.1 Demonstrate understanding of the central limit theory and confidence intervals 3.2 Demonstrate the ability to develop and test hypothesis 3.3 Understand inference for comparing means (ANOVA) 3.4 Understand inference for comparing proportions 3.5 Articulate, and demonstrate knowledge of correlation and regression 3.6 Understand how to test and validate assumptions for regression models 3.7 Understand the impact of multicollinearity in regression 3.8 Use a regression model to predict numeric values	– Assignments, projects, case studies – Expert witness evidence – Learner products – Mentor testimony – Observation – Questions and answers – Recognition of prior learning
4.0 Demonstrate knowledge of Python programming skills	4.1 Demonstrate the ability to build Python code using variables, relational operators, logical operators, loops, and functions 4.2 Read and write data from csv and json files 4.3 Use data structures such as lists, tuples, sets, and dictionaries 4.4 Demonstrate knowledge of NumPy and SciPy libraries 4.5 Learn to use Git repositories 4.6 Demonstrate knowledge of Anaconda, and Jupyter notebooks	– Assignments, projects, case studies – Expert witness evidence – Learner products – Mentor testimony – Observation – Questions and answers – Recognition of prior learning
5.0 Implement descriptive and inferential statistics using Python	5.1 Understand use of histograms and box plots to understand and visualize data distributions 5.2 Master descriptive statistics Python code calculating mean, median, mode, standard deviation, and percentiles; and identifying outliers 5.3 Use Python code to test hypothesis, calculate correlations and to predict a continuous variable using regression 5.4 Validate regression assumptions	– Assignments, projects, case studies – Expert witness evidence – Learner products – Mentor testimony – Observation – Questions and answers – Recognition of prior learning
6.0 Demonstrate ability to visualize data and extract insights	6.1 Demonstrate expertise with Python visualization libraries 6.2 Demonstrate ability to visualize data for statistical analysis: histograms, box plots 6.3 Demonstrate ability to visualize data for insight sharing with nontechnical users	– Assignments, projects, case studies – Learner products – Mentor testimony – Observation – Questions and answers

Competency outcomes	Assessment criteria	Evidence type
7.0 Demonstrate through a project the ability to analyze a dataset and communicate insights	7.1 Demonstrate the ability to complete a project using all skills acquired up to this point: data exploration, descriptive and inferential statistics, and data visualizations 7.2 Build a report with findings 7.3 Deliver a presentation sharing insights 7.4 Demonstrate solid communication skills (written and verbal)	– Assignments, projects, case studies – Mentor testimony – Observation – Professional discussion – Questions and answers – Simulation
8.0 Demonstrate understanding of what is data science and what data scientists do	8.1 Articulate what are the benefits of using data science 8.2 Articulate what a data scientist does and the value of data scientists to an organization 8.3 Understand some of the tools and the technology behind data science (IBM DSX and others) 8.4 Articulate the value of data science in specific use cases	– Mentor testimony – Observation – Questions and answers
9.0 Demonstrate ability to characterize a business problem	9.1 Leverage business acumen to understand how to take a business problem and put it into quantifiable form 9.2 Collaborate with cross-functional stakeholders to identify quantifiable improvements 9.3 Define key business indicators and target improvement metrics	– Assignments, projects, case studies – Mentor testimony – Observation – Questions and answers
10.0 Demonstrate ability to formulate a business problem as a hypothesis question	10.1 Formulate business problem as a research question with associated hypotheses 10.2 Determine what data is needed to test the hypotheses 10.3 Ensure hypotheses to be tested are aligned with business value	– Assignments, projects, case studies – Mentor testimony – Observation – Questions and answers
11.0 Demonstrate use of methodologies in the execution of the analytics cycle	11.1 Demonstrate how to apply the scientific method to business problems 11.2 Demonstrate how to apply the CRISP-DM methodology 11.3 Demonstrate understanding of an experimentation approach to insight finding and solution building	– Assignments, projects, case studies – Mentor testimony – Observation – Questions and answers
12.0 Demonstrate through a project the ability to plan for the execution of a project	12.1 Demonstrate the ability to setup a new project and follow the application of the scientific method and the CRISP-DM methodology 12.2 Build a report explaining the project plan 12.3 Deliver a presentation sharing the project plan 12.4 Demonstrate solid communication skills (written and verbal)	– Assignments, projects, case studies – Mentor testimony – Observation – Professional discussion – Questions and answers
13.0 Demonstrate ability to identify and collect data – multiple formats	13.1 Demonstrate SQL skills for querying databases and joining tables 13.2 Demonstrate ability to work with data from multiple data sources: SQL Data bases, NoSQL Databases 13.3 Demonstrate ability to work with data in databases, csv and json files	– Assignments, projects, case studies – Expert witness evidence – Learner products – Mentor testimony – Observation – Questions and answers

Competency outcomes	Assessment criteria	Evidence type
14.0 Demonstrate ability to manipulate, transform, and clean data	14.1 Demonstrate an understanding of when/why data transformations are necessary 14.2 Apply feature selection techniques 14.3 Demonstrate understanding of techniques to clean data 14.4 Demonstrate mastery of the pandas library for data transformation and manipulation 14.5 Demonstrate expertise with slicing, indexing, sub-setting, and merging and joining datasets	– Assignments, projects, case studies – Learner products – Mentor testimony – Observation – Questions and answers
15.0 Demonstrate expertise with techniques to deal with missing values, outliers, unbalanced data, as well as data normalization	15.1 Able to identify in which situations data may need to be scaled 15.2 Able to select the best way to handle missing values 15.3 Able to identify outliers and understand options to handle outliers 15.4 Able to understand the impact of working with unbalanced data 15.5 Able to construct a fully usable dataset	– Assignments, projects, case studies – Learner products – Mentor testimony – Observation – Questions and answers
16.0 Demonstrate through a project the ability to construct usable data sets	16.1 Demonstrate the ability to complete a data engineering project using all skills acquired up to this point: cleaning and transforming data and building a usable dataset 16.2 Build a report documenting decisions made on the data 16.3 Deliver a presentation sharing process and results 16.4 Demonstrate solid communication skills (written and verbal)	– Assignments, projects, case studies – Mentor testimony – Observation – Professional discussion – Questions and answers – Simulation
17.0 Demonstrate understanding of linear algebra principles for machine learning	17.1 Demonstrate understanding of working with vectors 17.2 Demonstrate understanding of working with matrices 17.3 Understand the application of eigenvectors and eigenvalues	– Assignments, projects, case studies – Expert witness evidence – Mentor testimony – Observation – Questions and answers – Recognition of prior learning
18.0 Demonstrate understanding of different modeling techniques	18.1 Learn how to build models using libraries such as scikit-learn, and algorithms such as regressions, logistic regressions, decision trees, boosting, random forest, Support Vector Machines, association rules, classification, clustering, neural networks, time series, survival analysis, etc. 18.2 Understand the process for experimentation and testing of different models on a dataset 18.3 Demonstrate expertise selecting potential models to test, based on the available data, data distributions, and the goal of the project: explaining relationships or prediction 18.4 Apply feature selection techniques 18.5 Demonstrate use of principal component analysis	– Assignments, projects, case studies – Learner products – Mentor testimony – Observation – Questions and answers

Competency outcomes	Assessment criteria	Evidence type
19.0 Demonstrate understanding of model validation and selection techniques	19.1 Demonstrate successful application of model validation and selection methods 19.2 Demonstrate use of cross-validation 19.3 Demonstrate use of model accuracy metrics such as Confusion Matrix, Gain and Lift Chart, Kolmogorov Smirnov Chart, AUC – ROC, Gini Coefficient, Concordant – Discordant Ratio, and Root Mean Squared Error	– Assignments, projects, case studies – Learner products – Mentor testimony – Observation – Questions and answers
20.0 Communicate results translating insight into business value	20.1 Demonstrate the ability to turn data insight into business value 20.2 Demonstrate the ability to adapt final deliverables and presentations based on the audience: data scientists, or business stakeholders	– Assignments, projects, case studies – Mentor testimony – Observation – Questions and answers
21.0 Demonstrate through a project the ability to test different models on a dataset, validate and select the best model, and communicate results	21.1 Demonstrate the ability to complete a project using all skills acquired up to this point: defining a business challenge as a hypothesis, selecting and evaluating different models on a date set and selecting a final “best” model 21.2 Build a report with findings and conclusions for a data science audience and for a business audience 21.3 Deliver a presentation sharing results for a data science audience and for a business audience 21.4 Demonstrate solid communication skills (written and verbal)	– Assignments, projects, case studies – Mentor testimony – Observation – Professional discussion – Questions and answers – Simulation
22.0 Deploy and monitor a validated model in an operational environment	22.1 Demonstrate how to deploy a model 22.2 Demonstrate the ability to monitor model performance and to define thresholds for model re-training 22.3 Demonstrate how to use a deployed model from a Python application	– Assignments, projects, case studies – Learner products – Mentor testimony – Observation – Questions and answers
23.0 Demonstrate through a project the ability to deploy and use a deployed model	23.1 Demonstrate the ability to complete a small project building a simple application that will use a machine learning deployed model to predict results	– Assignments, projects, case studies – Mentor testimony – Observation – Professional discussion – Questions and answers – Simulation
24.0 Understand the concept of big data, and how big data is used at organizations	24.1 Understand what big data is and how big data is used at organizations 24.2 Understand the concepts and major applications of distributed and cloud computing paradigm 24.3 Demonstrate knowledge of the big data ecosystems	– Assignments, projects, case studies – Mentor testimony – Questions and answers
25.0 Understand the big data ecosystem and its major components	25.1 Demonstrate knowledge of how each major component in the big data ecosystems works (HDFS, YARN, MapReduce, Spark, Pig, Hive, Flume, Flink, Kafka, etc.) 25.2 Demonstrate hands-on experience with HDFS, MapReduce, Spark, Pig, Hive	– Assignments, projects, case studies – Learner products – Mentor testimony – Observation – Questions and answers

Competency outcomes	Assessment criteria	Evidence type
26.0 Demonstrate through a project expertise with big data platforms (Hadoop, Spark)	26.1 Demonstrate the ability to complete a small project using the Hadoop and spark frameworks	– Assignments, projects, case studies – Mentor testimony – Observation – Professional discussion – Questions and answers – Simulation
27.0 Participate as a data scientist on client engagements (internal or external)	27.1 Participate as a data scientist in a minimum of 2 projects with clients (internal or external) 27.2 Demonstrate teamwork abilities, and the ability to manage project risks, and stakeholder conflict	– Assignments, projects, case studies – Mentor testimony – Observation – Professional discussion – Questions and answers – Simulation
28.0 Contribute to the profession by teaching or mentoring others	28.1 Demonstrate commitment to the profession by writing publications, and teaching and mentoring others 28.2 Demonstrate the ability to create reusable assets such as notebooks, libraries and documentation	– Mentor testimony – Observation – Professional discussion – Reflective accounts/personal statements

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the United States of America
January 2020

[1] Francesco Brenna, Giorgio Danesi, Glenn Finch, Brian Goehring and Manish Goyal. “Shifting toward Enterprise-grade AI: Resolving data and skills gaps to realize value.” *IBM Institute for Business Value*, September 2018. https://www.ibm.com/downloads/cas/QQ5KZLEL

[2] “August LinkedIn Workforce Report: Data Science Skills are in High Demand Across Industries.” *LinkedIn*, August 10, 2018. https://economicgraph.linkedin.com/resources/linkedin-workforce-report-august-2018

55029955-USEN-00