Certified Data Science Practitioner™ (CDSP) (Exam DSP-210) Bridge Document (From *Certified Data Science Practitioner™ (CDSP) (Exam DSP-110)*)

This bridge document is written for instructors who have used CertNexus' *Certified Data Science Practitioner™ (CDSP) (Exam DSP-110)* courseware and are looking to come up to speed on the *Certified Data Science Practitioner™ (CDSP) (Exam DSP-210)* courseware quickly and efficiently. Our instructional designers work to retain sequencing and activities wherever possible, while adding new content to stay up to date on the field of data science, align with the latest DSP-210 exam objectives, and provide an excellent class experience.

Exam Changes

The following table compares the exam domains that CertNexus publishes for the Exam DSP-110 certification and the Exam DSP-210 certification.

Domain	Exam DSP-110	% of Exam	Exam DSP-210	% of Exam
1.0	Defining the Question to Be Addressed Through the Application of Data Science	8%	Defining the Need to Be Addressed Through the Application of Data Science	8%
2.0	Extracting, Transforming, and Loading Data	21%	Extracting, Transforming, and Loading Data	21%
3.0	Performing Exploratory Data Analysis	31%	Performing Exploratory Data Analysis	31%
4.0	Building Models	24%	Building Models	23%
5.0	Testing Models	8%	Testing Models	5%
6.0	Communicating Findings	8%	Operationalizing the Pipeline	7%
7.0			Communicating Findings	5%

Overview of Changes

The *Certified Data Science Practitioner™ (CDSP) (Exam DSP-210)* course:

- Reflects changes and updates to the field of data science, including the push to democratize data in the organization.
- Reflects changes to the CertNexus exam blueprint and objectives, including the addition of the "Operationalizing the Pipeline" domain, which emphasizes an approach to deploying and delivering data-driven services to customers and other stakeholders in a business environment.
- Incorporates material about neural networks, particularly transformer architectures.
- Incorporates the latest CertNexus slide template, which has gone through a visual overhaul, and most notably, changed from standard 4:3 format to widescreen 16:9 format.
- Has an additional 2.5 hours of teaching time due to new material.

Lesson-Level and Topic-Level Structural Changes

The following table compares the lesson-level and topic-level outline of the *Certified Data Science Practitioner™ (CDSP) (Exam DSP-210)* course to the original *Certified Data Science Practitioner™ (CDSP) (Exam DSP-110)* course. Much of the outline is unchanged. However, there are a couple new topics, as noted in the following table.

Change color key:

- Topics with minor updates
- Topics with significant updates
- New topics

Certified Data Science Practitioner™ (CDSP)		
Lesson	**CNX0011: Certified Data Science Practitioner™ (CDSP) (Exam DSP-110)**	**CNX0020: Certified Data Science Practitioner™ (CDSP) (Exam DSP-210)**
1	Addressing Business Issues with Data Science A. Initiate a Data Science Project B. Formulate a Data Science Problem	Addressing Business Issues with Data Science A. Initiate a Data Science Project B. Democratize Data C. Formulate a Data Science Problem
2	Extracting, Transforming, and Loading Data A. Extract Data B. Transform Data C. Load Data	Extracting, Transforming, and Loading Data A. Extract Data B. Transform Data C. Load Data
3	Analyzing Data A. Examine Data B. Explore the Underlying Distribution of Data	Analyzing Data A. Examine Data B. Explore the Underlying Distribution of Data

Certified Data Science Practitioner™ (CDSP)		
Lesson	**CNX0011: Certified Data Science Practitioner™ (CDSP) (Exam DSP-110)**	**CNX0020: Certified Data Science Practitioner™ (CDSP) (Exam DSP-210)**
	C. Use Visualizations to Analyze Data D. Preprocess Data	C. Use Visualizations to Analyze Data D. Preprocess Data
4	Designing a Machine Learning Approach A. Identify Machine Learning Concepts B. Test a Hypothesis	Designing a Machine Learning Approach A. Identify Machine Learning Concepts B. Identify Transformer-Based Deep Learning Concepts C. Test a Hypothesis
5	Developing Classification Models A. Train and Tune Classification Models B. Evaluate Classification Models	Developing Classification Models A. Train and Tune Classification Models B. Evaluate Classification Models
6	Developing Regression Models A. Train and Tune Regression Models B. Evaluate Regression Models	Developing Regression Models A. Train and Tune Regression Models B. Evaluate Regression Models
7	Developing Clustering Models A. Train and Tune Clustering Models B. Evaluate Clustering Models	Developing Clustering Models A. Train and Tune Clustering Models B. Evaluate Clustering Models
8	Finalizing a Data Science Project A. Communicate Results to Stakeholders B. Demonstrate Models in a Web App C. Implement and Test Production Pipelines	Finalizing a Data Science Project A. Communicate Results to Stakeholders B. Demonstrate Models in a Web App C. Implement and Test Production Pipelines

Content-Level Changes

The following changes were made at the content level (organized by the structure in CDSP-210):

- Lesson 1:
  - Topic A: Added content on voluntary disclosure and informed consent to address a new task in exam objective 1.2.
  - Topic B: Added a new topic on data democratization to address the new exam objective 7.2. The activity incorporates several discussion questions related to the overarching Greene City National Bank (GCNB) scenario.
  - Topic C: Expanded the discussion of learning modes to include more information on semi-supervised and reinforcement learning.
- Lesson 2:
  - Topic A: Added content on data sources and distinguished them from data repositories; and added content on data-sharing agreements. Also added content on generated data. These address new tasks in exam objective 2.1.
  - Topic B: Added content on data governance in the data-cleaning process to address a new subtask in objective 2.2. Also improved the content on continuous vs. discrete variables and expanded the content on word-embedding techniques.
- Lesson 3:
  - Topic A: Added content on statistical bias to address new subtasks in exam objective 2.2. Also, distinguished statistical bias from model bias here and later on to avoid confusion.
  - Topic B: Improved the explanation and visualization of range measures/quartiles.
  - Topic D: Added formulas for normalization and standardization.
  - **Note**: Some minor adjustments have been made to the code in the "Analyzing Data" notebook, several of which are incorporated into later notebooks where relevant. For example:
    - Some statistical functions/methods in pandas, like `corr()`, now include `numeric_only = True` as an argument since those functions/methods were changed in the pandas library to default to `False`.
    - The `datetime_is_numeric` argument was removed from the `describe()` method due to changes in the pandas library.
    - Datetime data is now excluded from summary stats/visualizations due to changes in default behavior for pandas functions/methods.
- Lesson 4:
  - Topic A: Added content on data leakage to address a changed task in exam objective 4.1.
  - Topic B: Added a new topic on transformer-based, neural-network architectures to address a changed task in exam objective 4.2. The change to the task wording in the exam blueprint is minor but necessitates a significant amount of explanatory material. The material is somewhat technical but still very high level and is not meant to be a thorough exploration of transformer architectures or large language models (LLMs). The main activity presents several discussion questions that relate to the overarching GCNB scenario. There is also an *optional* activity that takes students through the basic process of fine tuning a pre-trained language model using the PyTorch and Transformers

libraries. The code has been pre-written and pre-executed within a Jupyter® Notebook. The provided VM does not have adequate resources to run the notebook, so students simply examine the code and its outputs.
    - Topic C: Added content on statistical power.
- Lesson 5:
    - **Note**: Some minor adjustments have been made to the code in the "Developing Classification Models" notebook to address changes to Python® libraries. For example:
        - The pandas library has new classes for calculating ROC, PRC, confusion matrices, etc.
        - The prediction results for the naïve Bayes, decision-tree, random-forest, and gradient-boosting models are slightly different.
        - The randomized/grid-search results are also slightly different for some of the models.
        - The evaluation results for some of the models are also slightly different.
        - Despite the differences in results, the overall conclusions are the same.
- Lesson 6:
    - Topic A: Added content on different gradient-descent techniques to address a new subtask in exam objective 4.2.
    - **Note**: Some minor adjustments have been made to the code in the "Developing Regression Models" notebook to address changes to Python libraries. Most notably, the XGBoost model has slightly different prediction/search/evaluation results, though the overall conclusions are the same.
- Lesson 7:
    - Topic A: Added an animated slide for *k*-means clustering visualization and improved the hierarchical agglomerative clustering (HAC) visual. Also replaced the discussion on latent-class analysis (LCA) with one on density-based clustering to address the change in a subtask in exam objective 4.2.
    - Topic B: Added content on evaluating density-based clustering methods.
    - **Note**: The *k*-means clustering models in the "Developing Clustering Models" notebook now assign data examples to different clusters. The overall conclusions are the same, though the clustering numbers are different, and the distributions of some clusters have changed as well. Also, in the silhouette-analysis portion of the notebook, 20 clusters are used as the alternative amount rather than 15 in order to obtain a better score.
- Lesson 8:
    - Topic A: Added content on explainable AI (XAI), interpretability tools, model documentation, participatory design, and related subject matter to address new tasks in exam objective 7.1.
    - Topic B: The web app in the activity is unchanged, but it loads new versions of the classification and clustering models due to the aforementioned minor changes to those models.
    - Topic C: Added model deployment as a step in the data-pipeline process and clarified offline vs. online models. Also added content on pipeline integration, data fabrics, data meshes, pipeline security, and pipelining tools like dbt and Fivetran, to address new tasks in exam objective 6.1.

Additional Notes

- Due to issues getting new versions of Anaconda® to install on Debian, the Linux® distribution in the VM has been changed from Debian 10 to Lubuntu® 22.04. The overall experience is very similar, and the Jupyter Notebook interface has not changed.
- The course setup has not changed drastically. The Linux VM is still provided and includes an installation of Anaconda, as well as all written code files.
- Newer versions of Anaconda and various Python libraries are installed in the VM. As noted in the content-level changes, this has led to a few changes in the code and/or the outputs of the code. However, the activities and the notebooks still follow the same structure and lead to the same overall conclusions.
- Non-screenshot images (and image elements) in the course PDF are now in vector format and retain full quality when scaled. This includes process diagrams, charts, and formulas.