# CompTIA DataX Certification Exam Objectives

## EXAM NUMBER: DY0-001

CompTIA

# About the Exam

The CompTIA DataX certification exam will certify the successful candidate has the knowledge and skills required to:

- Understand and implement data science operations and processes.
- Apply mathematical and statistical methods appropriately and understand the importance of data processing and cleaning, statistical modeling, linear algebra, and calculus concepts.
- Apply machine-learning models and understand deep-learning concepts.
- Utilize appropriate analysis and modeling methods and make justified model recommendations.
- Demonstrate understanding of industry trends and specialized data science applications.

## EXAM DEVELOPMENT

CompTIA exams result from subject matter expert workshops and industry-wide survey results regarding the skills and knowledge required of an IT professional.

## CompTIA AUTHORIZED MATERIALS USE POLICY

CompTIA Certifications, LLC is not affiliated with and does not authorize, endorse, or condone utilizing any content provided by unauthorized third-party training sites (aka "brain dumps"). Individuals who utilize such materials in preparation for any CompTIA examination will have their certifications revoked and be suspended from future testing in accordance with the CompTIA Candidate Agreement. In an effort to more clearly communicate CompTIA's exam policies on use of unauthorized study materials, CompTIA directs all certification candidates to the CompTIA Certification Exam Policies. Please review all CompTIA policies before beginning the study process for any CompTIA exam. Candidates will be required to abide by the CompTIA Candidate Agreement. If a candidate has a question as to whether study materials are considered unauthorized (aka "brain dumps"), they should contact CompTIA at examsecurity@comptia.org to confirm.

## PLEASE NOTE

The lists of examples provided in bulleted format are not exhaustive lists. Other examples of technologies, processes, or tasks pertaining to each objective may also be included on the exam, although not listed or covered in this objectives document. CompTIA is constantly reviewing the content of our exams and updating test questions to be sure our exams are current, and the security of the questions is protected. When necessary, we will publish updated exams based on existing exam objectives. Please know that all related exam preparation materials will still be valid.

## TEST DETAILS

| | |
|---|---|
| Required exam | DY0-001 |
| Number of questions | Maximum of 90 |
| Types of questions | Multiple-choice and performance-based |
| Length of test | 165 minutes |
| Recommended experience | A minimum of 5 years of hands-on experience as a data scientist |
| Passing score | Pass/fail only; no scaled score |

## EXAM OBJECTIVES (DOMAINS)

The table below lists the domains measured by this examination and the extent to which they are represented.

| DOMAIN | | PERCENTAGE OF EXAMINATION |
|---|---|---|
| 1.0 | Mathematics and Statistics | 17% |
| 2.0 | Modeling, Analysis, and Outcomes | 24% |
| 3.0 | Machine Learning | 24% |
| 4.0 | Operations and Processes | 22% |
| 5.0 | Specialized Applications of Data Science | 13% |
| **Total** | | **100%** |

# 1.0 Mathematics and Statistics

## 1.1 Given a scenario, apply the appropriate statistical method or concept.

- ***t*-tests**
- **Chi-squared test**
- **Analysis of variance (ANOVA)**
- **Hypothesis testing**
- **Confidence intervals**
- **Regression performance metrics**
  - $R^2$
  - Adjusted $R^2$
  - Root mean square error (RMSE)
  - F statistic
- **Gini index**
- **Entropy**
- **Information gain**
- ***p* value**
- **Type I and Type II errors**
- **Receiver operating characteristic/ area under the curve (ROC/AUC)**
- **Akaike information criterion/ Bayesian information criterion (AIC/BIC)**
- **Correlation coefficients**
  - Pearson correlation
  - Spearman correlation
- **Confusion matrix**
  - Classifier performance metrics
    - Accuracy
    - Recall
    - Precision
    - F1 score
    - Matthews Correlation Coefficient (MCC)
- Central limit theorem
- Law of large numbers

## 1.2 Explain probability and synthetic modeling concepts and their uses.

- **Distributions**
  - Normal
  - Uniform
  - Poisson
  - *t*
  - Binomial
  - Power law
- **Skewness**
- **Kurtosis**
- **Heteroskedasticity vs. homoskedasticity**
- **Probability density function (PDF)**
- **Probability mass function (PMF)**
- **Cumulative distribution function (CDF)**
- **Probability**
  - Monte Carlo simulation
  - Bootstrapping
  - Bayes' rule
  - Expected value
- **Types of missingness**
  - Missing at random
  - Missing completely at random
  - Not missing at random
- **Oversampling**
- **Stratification**

## 1.3 Explain the importance of linear algebra and basic calculus concepts.

- **Linear algebra**
  - Rank
  - Span
  - Trace
  - Eigenvalues/eigenvectors
  - Basis vector
  - Identity matrix
  - Matrix and vector operations
    - Matrix multiplication
    - Matrix transposition
    - Matrix inversion
    - Matrix decomposition
  - Distance metrics
    - Euclidean
    - Radial
    - Manhattan
    - Cosine
- **Calculus**
  - Partial derivatives
  - Chain rule
  - Exponentials
  - Logarithms

## 1.4 Compare and contrast various types of temporal models.

- **Time series**
  - Autoregressive (AR)
  - Moving average (MA)
  - Autoregressive integrated moving average (ARIMA)
- **Longitudinal studies**
- **Survival analysis**
  - Parametric
  - Non-parametric
- **Causal inference**
  - Directed acyclic graphs (DAGs)
  - Difference-in-differences
  - A/B testing of treatment effects
  - Randomized controlled trials

# 2.0 Modeling, Analysis, and Outcomes

## 2.1 Given a scenario, use the appropriate exploratory data analysis (EDA) method or process.

- **Univariate analysis**
- **Multivariate analysis**
- **Identification of object behaviors and attributes**
- **Charts and graphs**
  - Bar plot
  - Scatter plot
  - Box and whisker plot
  - Line plot
  - Violin plot
  - Heat map
  - Correlation plot
  - Histogram
  - Sankey diagram
  - Quartile-Quartile (Q-Q) plot
  - Density plot
  - Scatter plot matrix
- **Feature type identification**
  - Categorical variables
  - Discrete variables
  - Continuous variables
  - Ordinal variables
  - Nominal variables
  - Binary variables

## 2.2 Given a scenario, analyze common issues with data.

- **Common issues**
  - Sparse data
    - Sparse matrix
    - Sparse vectors
  - Non-linearity
  - Non-stationarity
  - Lagged observations
  - Difference observations
  - Multicollinearity
  - Seasonality
  - Granularity misalignment
  - Insufficient features
  - Multivariate outliers

## 2.3 Given a scenario, apply data enrichment and augmentation techniques.

- **Feature engineering**
- **Data transformation**
  - One-hot encoding
  - Label encoding
  - Cross-terms
  - Linearization
    - Logarithmic
    - Exponential
  - Box-Cox transformation
  - Normalization
  - Binning
  - Ratios
  - Pivoting
- **Geocoding**
- **Scaling**
- **Standardization**
- **Additional data sources**
  - Data augmentation
  - Data sets
  - Synthetic data

## 2.4 Given a scenario, conduct a model design iteration process.

- **Design and specifications**
  - Constraints
    - Time
    - Resource
    - Physical hardware
    - Cost
- **Performance evaluation**
  - Statistical metrics
  - Training time and cost
  - Inference performance over time
  - Model diagnostic plots
    - Residual vs. fitted values
- **Model selection**
  - Literature review
  - Hyperparameter tuning
  - Experiment tracking
  - Model architecture iteration
- **Requirements validation**

## 2.5 Given a scenario, analyze results of experiments and testing to justify final model recommendations and selection.

- **Benchmark against the baseline**
- **Benchmark against the conventional processes**
- **Specification testing results**
- **Final performance measures**
- **Satisfy business requirements**
  - Differentiate between business needs vs. business wants vs. reality

## 2.6 Given a scenario, translate results and communicate via appropriate methods and mediums.

- **Types of visualizations and reports**
- **Data selection for reports**
- **Effective communication and report considerations for peers and stakeholders**
  - Types of business executive stakeholders
  - Types of business domain stakeholders
  - Types of peers/professional stakeholders
- **Consider data types, dimensions, and levels of aggregation to produce appropriate visualizations/reports**
- **Avoid unintentionally deceptive charting and reporting**
- **Chart accessibility**
  - Font choice and size
  - Color choice
  - Content tagging
  - Effectiveness for accessibility
  - Government regulatory implications
- **Data and model documentation**
  - Code documentation
  - Data dictionary
  - Metadata
  - Change descriptions

# 3.0 Machine Learning

## 3.1 Given a scenario, apply foundational machine-learning concepts.

- **Loss function**
  - Variance minimization
- **Bias-variance tradeoff**
  - Overfitting
  - Underfitting
- **Variable/feature selection**
  - Feature importance
  - Multicollinearity
  - Correlation matrix
  - Variance inflation factor (VIF)
- **Class imbalance and mitigations**
  - Oversampling the minority class
  - Undersampling the majority class
  - Synthetic minority oversampling technique (SMOTE)
- **Regularization**
- **Cross-validation**
  - *k*-fold
- **The curse of dimensionality**
- **Occam's razor/law of parsimony**
- **In sample vs. out of sample**
- **Interpolation vs. extrapolation**
- **Ensemble models**
- **Hyperparameter tuning**
  - Grid search
  - Random search
- **Classifiers**
  - Binary classifiers
  - Multiclass (multinomial) classifiers
- **Recommender systems**
  - Collaborative filtering
  - Alternating least squares (ALS)
  - Similarity-based
- **Regressors**
- **Embeddings**
- **Post hoc model explainability**
  - Global explanations
  - Local explanations
- **Interpretable models**
- **Model drift causes**
  - Data drift
  - Concept drift
- **Data leakage**
  - Transfer learning
  - Cold start problem

## 3.2 Given a scenario, apply appropriate statistical supervised machine-learning concepts.

- **Linear regression models**
  - Ordinary least squares (OLS)
    - Assumptions
  - Weighted least squares
  - Ridge
  - Least Absolute Shrinkage and Selection Operator (LASSO)
  - Elastic net
- **Logistic regression models**
  - Probit
  - Logit
- **Linear discriminant analysis**
- **Quadratic discriminant analysis (QDA)**
- **Association rules**
  - Confidence
  - Lift
  - Reinforcement
  - Support
- **Naive Bayes**

## 3.3 Given a scenario, apply tree-based supervised machine-learning concepts.

- **Decision trees**
- **Random forest**
- **Boosting**
  - Gradient boosting
  - XGBoost
- **Bootstrap aggregation (bagging)**

## 3.4 Explain concepts related to deep learning.

- **Artificial neural network architecture**
  - Perceptron
  - Artificial neuron
  - Multilayer perceptron
  - Activation functions
    - Rectified linear unit (ReLU)
    - Sigmoid
    - Tanh
    - Softmax
  - Layer types
    - Input
    - Hidden
    - Pooling
    - Output
- **Dropout**
- **Batch normalization**
- **Early stopping**
- **Schedulers**
- **Back propagation**
- **One-shot learning**
- **Zero-shot learning**
- **Few-shot learning**
- **Deep-learning frameworks**
  - PyTorch
  - TensorFlow/Keras
  - AutoML
- **Optimizers**
  - Adam optimizer
  - Momentum
  - Root Mean Square Propagation (RMSprop)
  - Stochastic gradient descent
  - Mini-batch
- **Model types**
  - Convolutional neural network (CNN)
  - Recurrent neural network (RNN)
  - Long short-term memory (LSTM)
  - Generative adversarial networks (GANs)
  - Autoencoders
  - Transformers

## 3.5 Explain concepts related to unsupervised machine learning.

- **Clustering**
  - *k*-means
    - Silhouette score/elbow method
  - Hierarchical
  - Density-based spatial clustering analysis with noise (DBSCAN)
- **Dimensionality reduction**
  - Principal component analysis (PCA)
  - *t*-distributed stochastic neighbor embedding (*t*-SNE)
  - Uniform manifold approximation and projection (UMAP)
- ***k*-nearest neighbors (KNN)**
- **Singular value decomposition (SVD)**

# 4.0 Operations and Processes

## 4.1 Explain the role of data science in various business functions.

- **Compliance, security, and privacy**
  - Personally identifiable information (PII)
  - Proprietary
  - Anonymizing sensitive data
  - Data obfuscation
  - Data use regulations
- **Measures, metrics, and key performance indicators (KPIs)**
- **Requirements gathering**
  - Make recommendations based on cost-benefit analyses
  - Translate business need to the most appropriate solution
  - Relevant range of application

## 4.2 Explain the process of and purpose for obtaining different types of data.

- **Generated data**
  - Survey
  - Administrative
  - Sensor
  - Transactional
  - Experimental
  - Data-generating process
- **Synthetic data**
  - Costs and benefits
  - Creation process
  - Limitations
  - Sampling
  - Rationale
- **Commercial/public data**
  - Costs and benefits
  - Availability
  - Licensing
  - Restrictions

## 4.3 Explain data ingestion and storage concepts.

- **Infrastructure requirements**
  - Resource sizing
  - Graphics processing unit (GPU)/ Tensor Processing Unit (TPU)
- **Data formats**
  - Common formats
    - Comma-separated values (CSV)
    - JavaScript Object Notation (JSON)
    - Parquet
  - Compressed format
  - Structured storage
  - Semi-structured storage
  - Unstructured storage
- **Streaming**
- **Batching**
- **Pipeline implementation**
- **Orchestration/automation**
- **Persistence**
- **Refresh cycles**
- **Archiving**
- **Data lineage**

## 4.4 Given a scenario, implement common data-wrangling techniques.

- **Merging/combining**
  - Defining keys
  - Data matching
    - Match rates
    - Fuzzy join
  - Observation tracking
  - Union
  - Intersection
  - Types of joins
- **Cleaning**
  - Date/time standardization
  - Regular expressions
  - Deduplication
  - Unit conversion/standardization
  - Missing codes
- **Data errors**
  - Idiosyncratic
  - Systematic
- **Outliers**
  - Identification
  - Winsorization/cut points
  - Error vs. valid data point
- **Data flattening**
  - Extensible Markup Language (XML)
  - JSON
- **Imputation types**
- **Ground truth labeling**

## 4.5 Given a scenario, implement best practices throughout the data science life cycle.

- **Data science workflow models**
  - Cross-Industry Standard Protocol for Data Mining (CRISP-DM)
  - Data Management Association (DAMA)
- **Version control**
  - Code
  - Data
  - Hyperparameters
  - Models
- **Integrated development environment (IDE)**
- **Dependency licensing**
- **Access via application programming interface (API)**
  - Data access and retrieval
  - Model endpoint/model services
- **Process documentation**
  - Markdown
  - Docstring
  - Appropriate code commenting
  - Reference data and documentation
- **Clean code methods**
- **Unit test writing**

## 4.6 Explain the importance of DevOps and MLOps principles in data science.

- **Data replication**
- **Continuous integration/ continuous deployment (CI/CD) pipelines**
- **Model deployment**
- **Container orchestration**
- **Virtualization**
- **Code isolation**
- **Model performance monitoring**
- **Model validation**
  - Online
  - Offline
  - Model A/B testing

## 4.7 Compare and contrast various deployment environments.

- **Containerization**
- **Cloud deployment**
- **Cluster deployment**
- **Hybrid deployment**
- **Edge deployment**
- **On-premises deployment**

# 5.0 Specialized Applications of Data Science

## 5.1 Compare and contrast optimization concepts.

- **Constrained optimization**
  - Network topology
    - Traveling salesman
  - Scheduling
  - Linear solvers
    - Simplex method
  - Non-linear solvers
  - Pricing
  - Resource allocation
  - Bundling
  - Boundary cases
- **Unconstrained optimization**
  - One-armed bandit
  - Multi-armed bandit
  - Finding local maxima or minima

## 5.2 Explain the use and importance of natural language processing (NLP) concepts.

- **Tokenization/bag of words**
- **Word embeddings**
  - *n*-grams
- **Term frequency-inverse document frequency (TF-IDF)**
- **Document term matrix**
- **Edit distance**
- **Large language models**
  - Word2vec
  - GloVe
- **Text preparation**
  - Lemmatization
  - Stop words
  - Augmenters
  - String indexing
  - Stemming
  - Part-of-speech (POS) tagging
- **Topic modeling**
  - Latent Dirichlet Allocation
- **Disambiguation**
- **NLP applications**
  - Sentiment analysis
  - Question-and-answer/dialogue
  - Named-entity recognition (NER)
  - Auto-tagging
  - Text generation
  - Matching models
  - Speech recognition and generation
  - Text summarization
  - Natural language understanding (NLU)
  - Natural language generation (NLG)

## 5.3 Explain the use and importance of computer vision concepts.

- **Optical character recognition**
- **Object/semantic segmentation**
- **Object detection**
- **Tracking**
- **Sensor fusion**
- **Data augmentation**
  - Filter application
  - Rotation
  - Occlusion
  - Spurious noise
  - Flipping
  - Scaling
  - Holes
  - Masking
  - Cropping

## 5.4 Explain the purpose of other specialized applications in data science.

- Graph analysis/graph theory
- Heuristics
- Greedy algorithms
- Reinforcement learning
- Event detection
- Fraud detection
- Anomaly detection
- Multimodal machine learning
- Optimization for edge computing
- Signal processing

# CompTIA DataX DY0-001 Acronym List

The following is a list of acronyms that appear on the CompTIA DataX DY0-001 exam. Candidates are encouraged to review the complete list and attain a working knowledge of all listed acronyms as part of a comprehensive exam preparation program.

| Acronym | Spelled Out |
|---|---|
| AIC-BIC | Akaike Information Criterion - Bayesian Information Criterion |
| ALS | Alternating Least Squares |
| ANOVA | Analysis of Variance |
| API | Application Programming Interface |
| AR | Autoregressive |
| ARIMA | Autoregressive Integrated Moving Average |
| AUC | Area Under the Curve |
| CDF | Cumulative Distribution Function |
| CI/CD | Continuous Integration/Continuous Deployment |
| CNN | Convolutional Neural Network |
| CRISP-DM | Cross-industry Standard Process for Data Mining |
| CSV | Comma-separated Values |
| DAG | Directed Acyclic Graph |
| DAMA | Data Management Association |
| DBSCAN | Density-based Spatial Clustering Analysis with Noise |
| EDA | Exploratory Data Analysis |
| FFNN | Feed Forward Neural Network |
| GAN | Generative Adversarial Networks |
| GPU | Graphics Processing Unit |
| GUID | Globally Unique Identifier |
| HDBSCAN | Hierarchical Density-based Spatial Clustering Analysis with Noise |
| HPC | High-performance Computing |
| HTTP | Hypertext Transfer Protocol |
| IDE | Integrated Development Environment |
| IP | Internet Protocol |
| JSON | JavaScript Object Notation |
| KNN | *k*-Nearest Neighbors |
| KPI | Key Performance Indicator |
| LASSO | Least Absolute Shrinkage and Selection Operator |
| LSTM | Long Short-term Memory |
| MA | Moving Average |
| MAC | Media Access Control |
| MCC | Matthews Correlation Coefficient |
| ML | Machine Learning |
| NER | Named-entity Recongition |
| NLG | Natural Language Generation |
| NLP | Natural Language Processing |
| NLU | Natural Language Understanding |
| OLS | Ordinary Least Squares |
| OS | Operating System |
| PCA | Principal Component Analysis |
| PDF | Probability Density Function |
| PII | Personally Identifiable Information |
| PIP | Preferred Installer Program |
| POS | Part of Speech |
| QDA | Quadratic Discriminant Analysis |
| Q-Q | Quantile-Quantile |
| RegEX | Regular Expression |
| ReLU | Rectified Linear Unit |
| REST | Representational State Transfer |
| RPC | Remote Procedure Call |
| RMS | Root Mean Square |
| RMSE | Root Mean Square Error |
| RMSprop | Root Mean Square Propagation |
| RNN | Recurrent Neural Network |
| ROC-AUC | Receiver Operating Characteristic - Area Under the Curve |
| RPC | Remote Procedure Call |

| Acronym | Spelled |
|---|---|
| RSS | Residual Sum of Squares |
| SARIMA | Seasonal Auto-regressive Integrated Moving Average |
| SMOTE | Synthetic Minority Oversampling Technique |
| SOAP | Simple Object Access Protocol |
| SVD | Singular Value Decomposition |
| SVM | Support Vector Machines |
| SVN | Subversion |
| TF-IDF | Term Frequency Inverse Document Frequency |

| Acronym | Spelled Out |
|---|---|
| TPU | Tensor Processing Unit |
| *t*-SNE | *t*-distributed Stochastic Neighbor Embedding |
| UMAP | Uniform Manifold Approximation and Projection |
| VIF | Variance Inflation Factor |
| WSL | Windows Subsystem for Linux |
| XML | Extensible Markup Language |

# CompTIA DataX DY0-001 Hardware and Software List

CompTIA has included this sample list of hardware and software to assist candidates as they prepare for the DataX DY0-001 certification exam. This list may also be helpful for training companies that wish to create a lab component for their training offering. The bulleted lists below each topic are sample lists and are not exhaustive.

**Equipment**

- Workstations with CUDA-compatible GPU
- GPU on cloud providers

**Software**

- Linux kernel-based operating systems (preferred)
- Windows operating systems
  - Regional packs
  - Unicode
  - Windows Subsystem for Linux (WSL)
  - Docker desktop
- CoderPad
- Python or R
  - Relevant packages (visualization, modeling, cleaning, and machine learning)
- Notebook environment/tool set
- Visual Studio Code
- Git

**Other**

- Large data sets
- Small data sets
- Various types of data sets

CompTIA