DLI Accelerated Data Science Teaching Kit

Lecture 1.1 - Teaching Kit Modules Overview

Teaching Kit Module Goals

- Teach fundamental and advanced topics in data collection and preprocessing, accelerated data science with RAPIDS, scalable and distributed computing, GPU-accelerated machine learning, data visualization and graph analytics
- Content also covers culturally responsive topics such as fairness and data bias, as well as challenges and important individuals from underrepresented groups
- This 3rd release of the Accelerated Data Science Teaching Kit now includes focused modules covering:
  - Graph Analytics
  - Streaming Data
  - Genomics
  - Text Analytics
  - CPU vs. GPU-accelerated Data Science
  - Working in Data Science Teams
  - Code Backup and Version Control
  - Team Project (Fake News Detection)

People Involved in Content Development

- Polo Chau, **Georgia Institute of Technology**
- Xishuang Dong, **Prairie View A&M University**
- Joe Bungo, **NVIDIA**
- Taurean Dyer, **NVIDIA**
- Aiswarya Bhagavatula, **Georgia Institute of Technology**
- Haekyu Park, **Georgia Institute of Technology**
- Zijie (Jay) Wang, **Georgia Institute of Technology**
- Scott Freitas, **Georgia Institute of Technology**
- Kevin Li, **Georgia Institute of Technology**
- Jon Saad-Falcon, **Georgia Institute of Technology**
- Zhiyan (Frank) Zhou, **Georgia Institute of Technology**

Teaching Kit Modules Overview

Module 1: Introduction to Data Science	1.1: Teaching Kit Modules Overview 1.2: What is Data Science? 1.3: Why is Data Science Important? 1.4: Learning Goals and Expectations 1.5: Analytics Building Blocks 1.6: Example Data Science Project 1: Apolo Graph Exploration 1.7: Example Data Science Project 2: NetProbe Auction Fraud Detection 1.8: Data Science Buzzwords, Hype Cycle, General vs Narrow AI 1.9: Career Paths and Challenges 1.10: Diversity Gaps in Science and Engineering 1.11: Hidden Figures in Data Science from Underrepresented Groups **Lab: Introduction to RAPIDS and cuDF**

Teaching Kit Modules Overview

Module 2: Data Collection	2.1: Collecting Data 2.2: Scraping Data 2.3: Popular Scraping Libraries 2.4: Data Annotation and Data Quality 2.5: SQLite as Simple, Effective Storage 2.6: SQL Refresher 2.7: Beware of Missing Indexes **Lab: Data Collection via API** **Lab: Data Annotation in Active Learning** **Lab: GPU-accelerated SQL with BlazingSQL** **DLI Online Course Section: Accelerating End-toEnd Data Science Workflows, Section 1: GPU-accelerated Data Manipulation**

Teaching Kit Modules Overview

Module 3: Data Pre-processing (ETL)	3.1: Introduction to Data Pre-processing 3.2: Data Cleaning and Statistical Preprocessing 3.3: Data Cleaners: OpenRefine and Wrangler 3.4: Feature Selection: Introduction to Filter Methods 3.5: Feature Selection: Introduction to Model-based Methods 3.6: Feature Reduction: PCA **Lab: Data Wrangling with OpenRefine** **Lab: Outlier Detection with IQR** **Lab: Feature Reduction with PCA**
Module 4: Data Ethics and Reducing Bias in Data Sets	4.1: Sources of Bias and Fairness Measures 4.2: Tools for Discovering and Interpreting Bias in Models 4.3: Challenges Faced by Underrepresented Groups Relating to Data Ethics and Bias **Lab: Classifier Audit with FairVis**
Module 5: Data Integration	5.1: Knowledge Graph 5.2: Data De-duplication

Teaching Kit Modules Overview

Module 6: Data Analytics, Concepts and Tasks	6.1: Break Complex Problems into Simpler Ones: Part 1 6.2: Break Complex Problems into simpler Ones: Part 2
Module 7: Visualization 101	7.1: What is Info Vis and Why it is Important 7.2: Human Perception 7.3: Gestalt Psychology 7.4: Chart Basics 7.5: Colors 7.6: Visual Exploratory Data Analytics with cuXFilter **Lab: Creating Visualizations**
Module 8: Fixing Common Visualization Issues	8.1: Fixing Bar Charts, Line Charts, Tables and More 8.2: Applying What You've Learned 8.3: Crown Jewel, Self-contained Figures and More Tips

Teaching Kit Modules Overview

Module 9: Data Visualization for Web (D3)	9.1: Why Learn D3? 9.2: Prerequisites: Javascript and SVG 9.3: D3 Overview 9.4: Enter-Update-Exit 9.5: Attributes, Styles, Classes and Text 9.6: Scales and Axes 9.7: Dynamic Data and Interaction **Lab: Web-based Visualization (D3)** **Lab: Server and Client-side Visualizations (Datashader, Plotly, Plotly Dash)**
Module 10: Scalable Computing (Hadoop, Hive)	10.1: Big Data is Common. How to Store It? 10.2: Why Hadoop? 10.3: MapReduce Overview 10.4: Example MapReduce Program 10.5: How to Try Hadoop 10.6: Pig and Hive **Lab: Hadoop**
Module 11: Scalable Computing (Spark)	11.1: Spark Overview 11.2: Example Spark Programs 11.3: Spark SQL and Other Spark Libraries 11.4: RAPIDS and Spark **Lab: Accelerated Spark with RAPIDS on AWS**

Teaching Kit Modules Overview

Module 12: Scalable Computing (Hbase)	12.1: HBase Overview 12.2: How HBase Scales Up Storage 12.3: How to Use HBase 12.4: Learn More About Hbase
Module 13: Scalable Computing (Dask and UCX)	13.1: Using Dask and UCX with RAPIDS and BlazingSQL

Teaching Kit Modules Overview

Module 14: Machine Learning (Classification)	14.1: Overview 14.2: Introduction to Supervised Learning 14.3: Linear Regression 14.4: RAPIDS Acceleration: Linear Regression 14.5: Overfitting and Cross Validation 14.6: Decision Tree 14.7: Visualizing Classification: ROC, AUC, Confusion Matrix 14.8: Bagging 14.9: Random Forests 14.10: RAPIDS Acceleration: Random Forest 14.11: Boosting 14.12: XGBoost with RAPIDS 14.13: k-NN with RAPIDS **Lab: Decision Tree Classification Clustering** **Lab: Classification (Random Forest)** **Lab: Image Classification with RAPIDS-based Random Forest** **DLI Online Course Section: Accelerating End-to-End Data Science Workflows, Section 2: GPU-accelerated Machine Learning**

Teaching Kit Modules Overview

Module 15: Machine Learning (Clustering and Dimensionality Reduction)	15.1: Introduction to Unsupervised Learning 15.2: KMeans and Hierarchical Clustering 15.3: RAPIDS Acceleration: KMeans 15.4: DBSCAN 15.5: t-SNE 15.6: UMAP 15.7: Visualizing Clusters 15.8: RAPIDS Acceleration: PCA, UMAP, DBSCAN **Lab: KMeans Clustering** **Lab: Dimensionality Reduction and Visualization**

Teaching Kit Modules Overview

Module 16: Neural Networks	16.1: Introduction to Artificial Neural Networks 16.2: Activation Function and Perceptron 16.3: Multilayer Perceptron 16.4: Advanced Deep Neural Networks **Lab: Binary Classification with Perceptron** **DLI Online Course: Getting Started with Deep Learning** **DLI Online Course: Deep Learning at Scale with Horovod** **DLI Online Course: Getting Started with Image Segmentation** **DLI Online Course: Modeling Time-Series Data with Recurrent Neural Networks in Keras** **DLI Online Course: Medical Image Classification Using the MedNIST Dataset** **DLI Online Course: Image Classification with TensorFlow: Radiomics — 1p19q Chromosome Status Classification**

Teaching Kit Modules Overview

Module 17: Graph Analytics	17.1: How to Represent and Store Graphs 17.2: Graph Power Laws 17.3: Centralities: Degree, Betweenness, Clustering Coefficient 17.4: PageRank and Personalized PageRank 17.5: Interactive Graph Exploration **Lab: Graph Analytics with cuGraph**
Module 18: Streaming Data	18.1: Machine Learning for Streaming Data Analysis 18.2: Data Preprocessing 18.3: Learning Process 18.4: Reasoning and Data Resource **Lab: Sales Forecasting via RAPIDS Linear Regression**
Module 19: Genomics	19.1: Introduction to Genomics 19.2: Data Preprocessing 19.3: Clustering and Validation 19.4: Statistical Analysis **Lab: Cancer Recognition on Genomics Data via Decision Tree Algorithm**

Teaching Kit Modules Overview

Module 20: Text Analytics	20.1: Basics: Preprocessing, Representation, Word Importance 20.2: Latent Semantic Indexing (Singular Value Decomposition) 20.3: SVD: Dimensionality Reduction, and Other Uses 20.4: Text Visualization **Lab: Latent Semantic Indexing for Text via Singular Value Decomposition (cuML)**
Module 21: CPU vs. GPU-accelerated Data Science	21.1: RAPIDS Benefits 21.2: Refactoring Workloads **Lab: Accelerating Workloads Using RAPIDS** **DLI Online Course Section: Accelerating End-to-End Data Science Workflows, Section 3: Data Analysis to Save the UK**

Teaching Kit Modules Overview

Module 22: Working in Data Science Teams	22.1: Forming Great Teams 22.2: Project Idea Checklist: Heilmeier Questions 22.3: Pay Attention to Software Licenses Early On
Module 23: Code Backup and Version Control	23.1: Git: Overview and Benefits 23.2: Warning! Keep Your Repository Private Initially 23.3: GitHub and Bitbucket
Module 24: Team Project (Fake News Detection)	24.1: Introduction to Project 24.2: Evaluation of Team Project **Team Project: Fake News Detection (cuML)**

DEEP
LEARNING
INSTITUTE

PRAIRIE VIEW
A&M UNIVERSITY

DLI Accelerated Data Science Teaching Kit

Thank You