DATA SCIENCE DEVELOPMENT TOOLS

PART 1: PYTHON, NUMPY

ANNOUNCEMENTS

Class #	Week #	Month	Date	Topic	Reading	Labs	Due date
1	1	Sep	4	Welcome, Introduction, Course Objectives, DS Lifecycle	Chapter 1 Intro DS	Lab 1: IDE Set Up, GitHub	Sept 6, 2025
2	2	Sep	9	Python setup, Google colab, Github	Chapter 2.1-2.5 Python		
3	2	Sep	11	NumPy, Vectorization		Lab 2: Vectorization	Sept 13, 2025
4	3	Sep	16	Pandas, Matplotlib, Seaborn	Chapter 2.6-2.8 Python		
5	3	Sep	18	Data Cleaning and Preparation		Lab 3: NumPy, Pandas	Sept 20, 2025
6	4	Sep	23	Data Acquisition, ETL, Populations, Sampling	Chapter 3 Data		
7	4	Sep	25	Descriptive Statistics	Chapter 4 Prob &		
8	5	Sep	30	Exploratory Data Analysis (EDA)	Chapter 5 EDA		
9	5	Oct	2	Principles of Data Visualization			
10	6	Oct	7	Business Intelligence with Tableau	Chapter 6 BI & T		
11	6	Oct	9	Tableau Dashboards, Stories			
12	7	Oct	14	Data management - databases, SQL queries			
13	7	Oct	16	More SQL Features, Joins	Chapter 7 DB an		
14	8	Oct	21	SQL continued, NoSQL, Connect to Tableau			
15	8	Oct	23	MIDTERM REVIEW			
16	9	Oct	28	Overview of AI, ML, DL, GenAI Topics for remainder of semester	Chapter 8 Unsup		
17	9	Oct	30	Unsupervised Learning- Kmeans			
18	10	Nov	4	Unsupervised Learning- Hierarchical, DBSCAN	Chapter 9 Supervised Learn		
19	10	Nov	6	Supervised Learning: Part 1	Chapter 10 Decision Trees	Lab 8: Cluster Analysis	Nov 08, 2025
20	11	Nov	11	Supervised Learning: Part 2	Chapter 11 Regression (optional)		
21	11	Nov	13	Evaluation of models, comparing performance	Chapter 12 Eval	Lab 9: ML Classification/Regression	Nov 15, 2025
22	12	Nov	18	ANN, Multi-Layer Perceptron, Backpropagation	Chapter 13 ANN		
23	12	Nov	20	Deep Learning		Lab 10: MLP and Backpropagation	Nov 22, 2025
24	13	Nov	25	GenAI - Introduction	Chapter 14 GenAI		
-	13	Nov	27	No Class. Thanksgiving Holiday.			
25	14	Dec	2	GenAI - Applications	Chapter 15 AI Ethics		
26	14	Dec	4	Ethics / Data Privacy / Business and Government Policy		**Lab 11: GenAI Applications**	December 06, 2
27	15	Dec	9	Review and Wrap Up, The Future of DS/ML/AI			
28	16	Dec	17	Finals Week. Final 12/12 2:30-5:30pm 1102 JKB		Final	

You are here

BYU

TOOLS FOR DATA MINING

- Start with Pickaxe and Shovel

- Move to jack hammer and other automated tools

- Python, SQL, R most common languages

- NumPy, Pandas, Matplotlib, Seaborn, Scikit-Learn

- Tableau, PowerBI

- AWS SageMaker, Dataiku, DataRobot, DataBricks

COMMON DATA SCIENCE PACKAGES FOR PYTHON

CS 180

Import name	Common alias	Description
`numpy`	`np`	NumPy includes functions and classes that aid in numerical computation. `NumPy` is used in many other data science packages.
`pandas`	`pd`	pandas provides methods and classes for tabular and time-series data.
`sklearn`	`sk`	scikit-learn provides implementations of many machine learning algorithms with a uniform syntax for preprocessing data, specifying models, fitting models with cross-validation, and assessing models.
`matplotlib.pyplot`	`plt`	matplotlib allows the creation of data visualizations in Python. The functions mostly expect `NumPy` arrays.
`seaborn`	`sns`	seaborn also allows the creation of data visualizations but works better with pandas DataFrame.
`scipy.stats`	`sp.stats`	SciPy provides algorithms and functions for computing problems that arise in science, engineering and statistics. `scipy.stats` provides the functions for statistics.
`statsmodels`	`sm`	statsmodels adds functionality to Python to estimate many different kinds of statistical models, make inferences from those models, and explore data.

BYU

NUMPY

- **Spelled: NumPy, Pronounced "num-pie"**

- **What is NumPy?** NumPy (Numerical Python) is the fundamental package for scientific computing with Python. It provides a high-performance multidimensional array object and tools for working with these arrays.

- **Why NumPy?**

  - **Speed:** NumPy arrays are more efficient and faster than Python lists for numerical operations, as they are implemented in C.

  - **Functionality:** It provides a rich set of functions for linear algebra, Fourier transforms, and random number generation.

  - **Foundation:** Many other data science libraries like Pandas, SciPy, and Scikit-learn are built on top of NumPy.

BYU

NUMPY ARRAY FUNCTIONS

Function	Parameters	Description

`array()`	`object`
	`dtype=None`
	`ndim=0`
		Returns an array constructed from `object`. `object` must be a scalar or an ordered container, such as tuple or list. The array element type is inferred from `object` unless a `dtype` is specified. `ndim` is the minimum number of array dimensions.

`delete()`	`arr`
	`obj`
	`axis=None`
		Deletes a slice of input array `arr`. `axis` is the axis along which to remove a slice. `obj` is the index of the slice along the axis.

`full()`	`shape`
	`fill_value`
	`dtype=None`
		Returns an array filled with `fill_value`. The `shape` tuple specifies array shape. `dtype` specifies the array type. If `dtype=None`, the type is inferred from `fill_value`.

`insert()`	`arr`
	`obj`
	`values`
	`axis=None`
		Inserts array `values` to input array `arr`. `axis` is the axis along which to insert. `obj` is the index before which `values` is inserted.

`zeros()`	`shape` `dtype=float`	Returns an array filled with zeros. The `shape` tuple specifies array shape. `dtype` specifies the array type.
`ones()`	`shape` `dtype=None`	Returns an array filled with ones. The `shape` tuple specifies array shape. `dtype` specifies the array type. If `dtype=None`, the type is float64.
`sort()`	`a` `axis=-1`	Sorts array `a` along `axis`. The default `axis=-1` sorts along the last axis in `a`. `axis=None` flattens `a` before sorting.

BYU

UPCOMING ASSIGNMENTS

- Reading: 2.1 – 2.5

- DS Lab 2: Vectorization due Sept 13th 11:59 pm

BYU