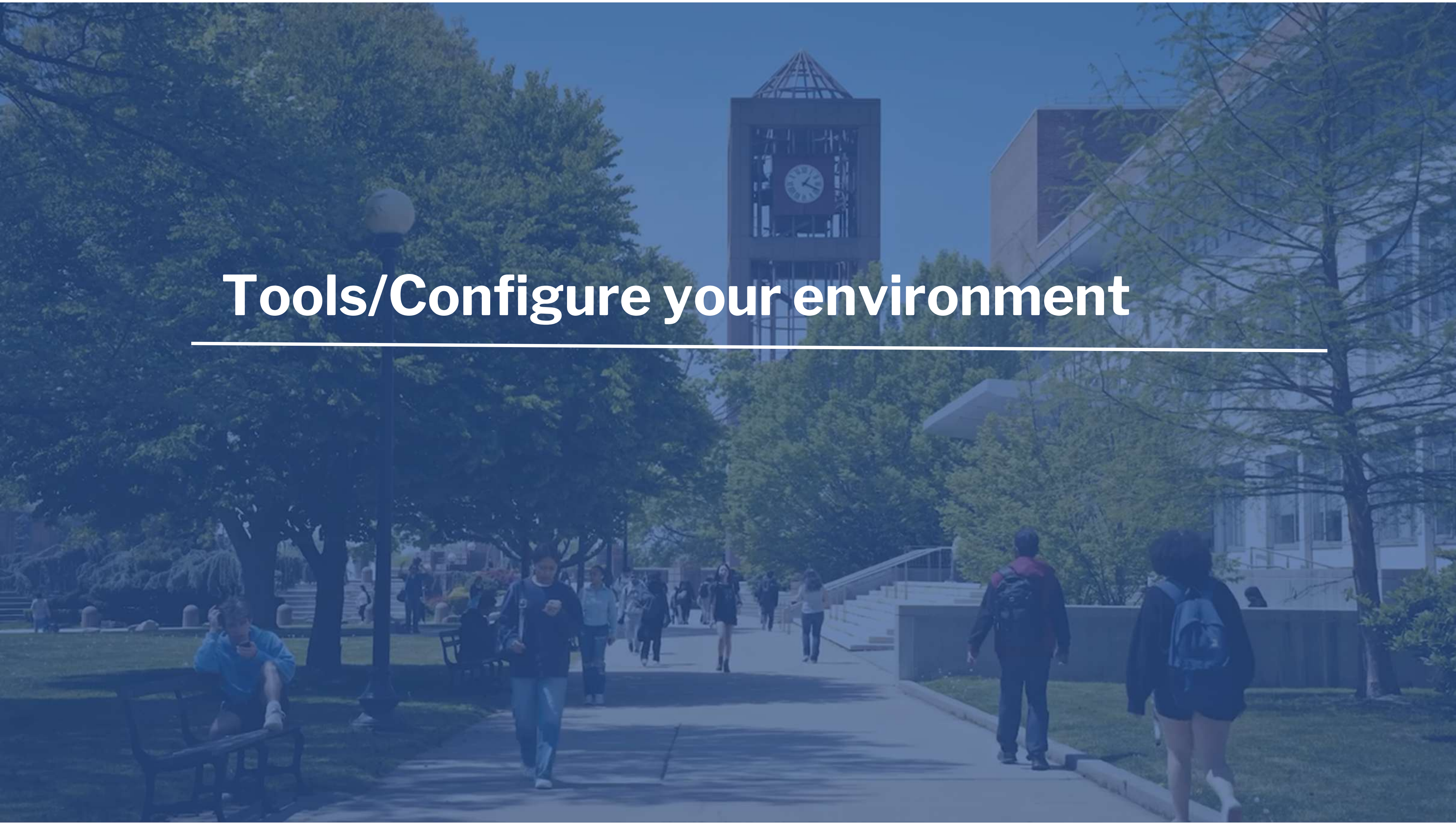
Introduction to Machine Learning (GAI 601)

WEEK 2

CUNY SPS

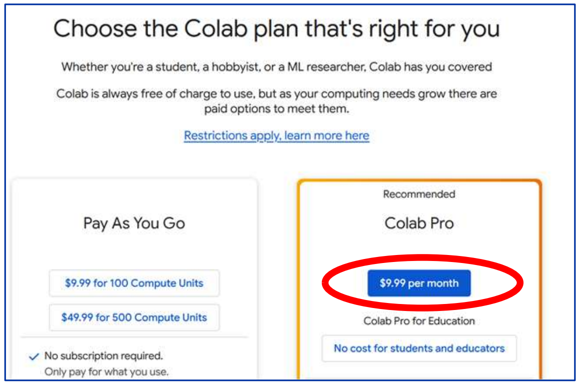


Tools/Configure your environment

Register for free tools

1. Colab

- Go to https://colab.research.google.com/signup
- Create a Google account (if you don't have one)
- Sign up for **Colab Pro for Education** (free for students)



Choose the Colab plan that's right for you

Whether you're a student, a hobbyist, or a ML researcher, Colab has you covered

Colab is always free of charge to use, but as your computing needs grow there are paid options to meet them.

Restrictions apply, learn more here

✓ No subscription required.
Only pay for what you use.

Colab Pro for Education

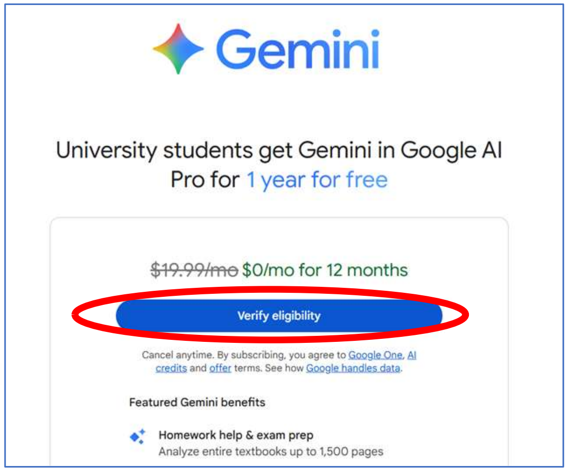
Pay As You Go
$9.99 for 100 Compute Units
$49.99 for 500 Compute Units
Recommended
Colab Pro
$9.99 per month
No cost for students and educators

2. GitHub Copilot

- Go to https://github.com/education
- Click on **Join GitHub Education**
- Create a GitHub account (if you don't have one)
- Click on Start an Application to get access

3. Google Gemini

- Go to https://gemini.google/us/students
- Fill out the verification form



University students get Gemini in Google AI Pro for 1 year for free

~~$19.99/mo~~ $0/mo for 12 months

Cancel anytime. By subscribing, you agree to Google One, AI credits and offer terms. See how Google handles data.

Featured Gemini benefits

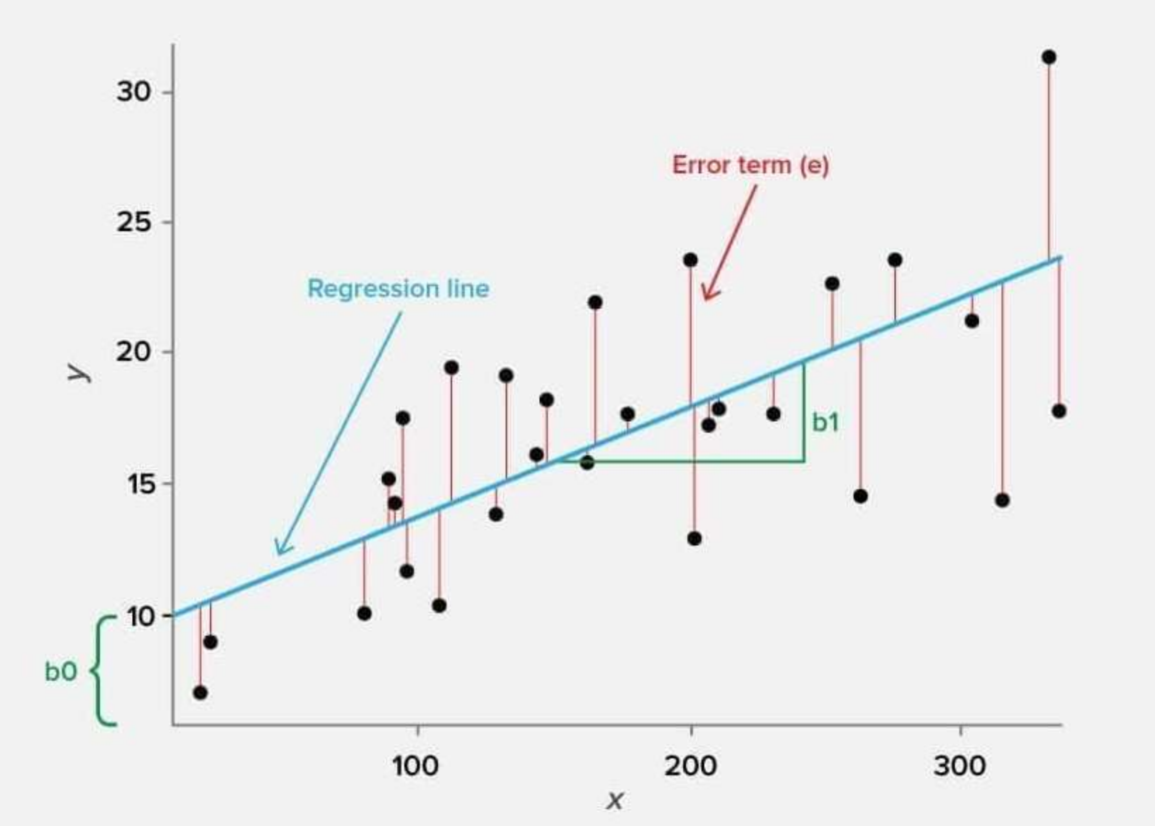
Homework help & exam prep
Analyze entire textbooks up to 1,500 pages

4. Microsoft Copilot 365

- Free with you Office 365 access (via CUNY)

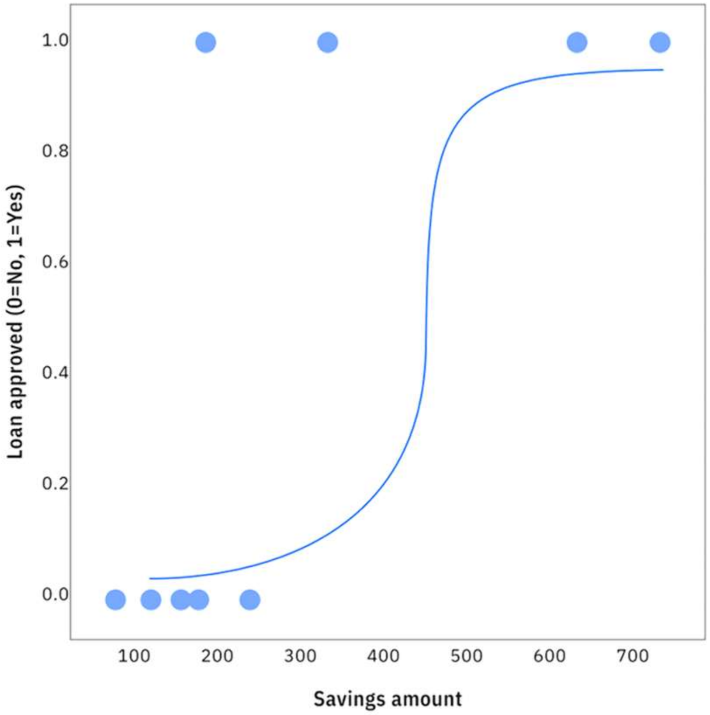
Linear & Logistic Regression

Linear Regression



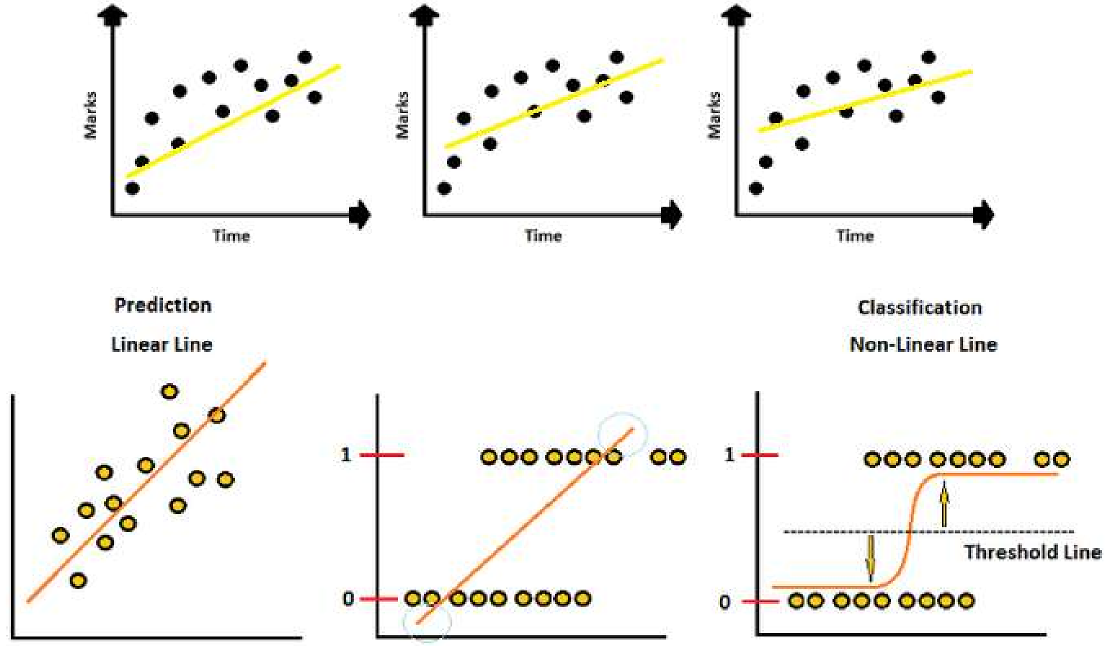
Source: Supervised Learning Algorithms: A Comparison, November 2020Kristu Jayanti Journal of Computational Sciences (KJCS) - DOI:10.59176/kjcs.v1i1.1259. Free access

Logistic Regression



Source: IBM https://www.ibm.com/think/topics/logistic-regression

Linear vs Logistic Regression



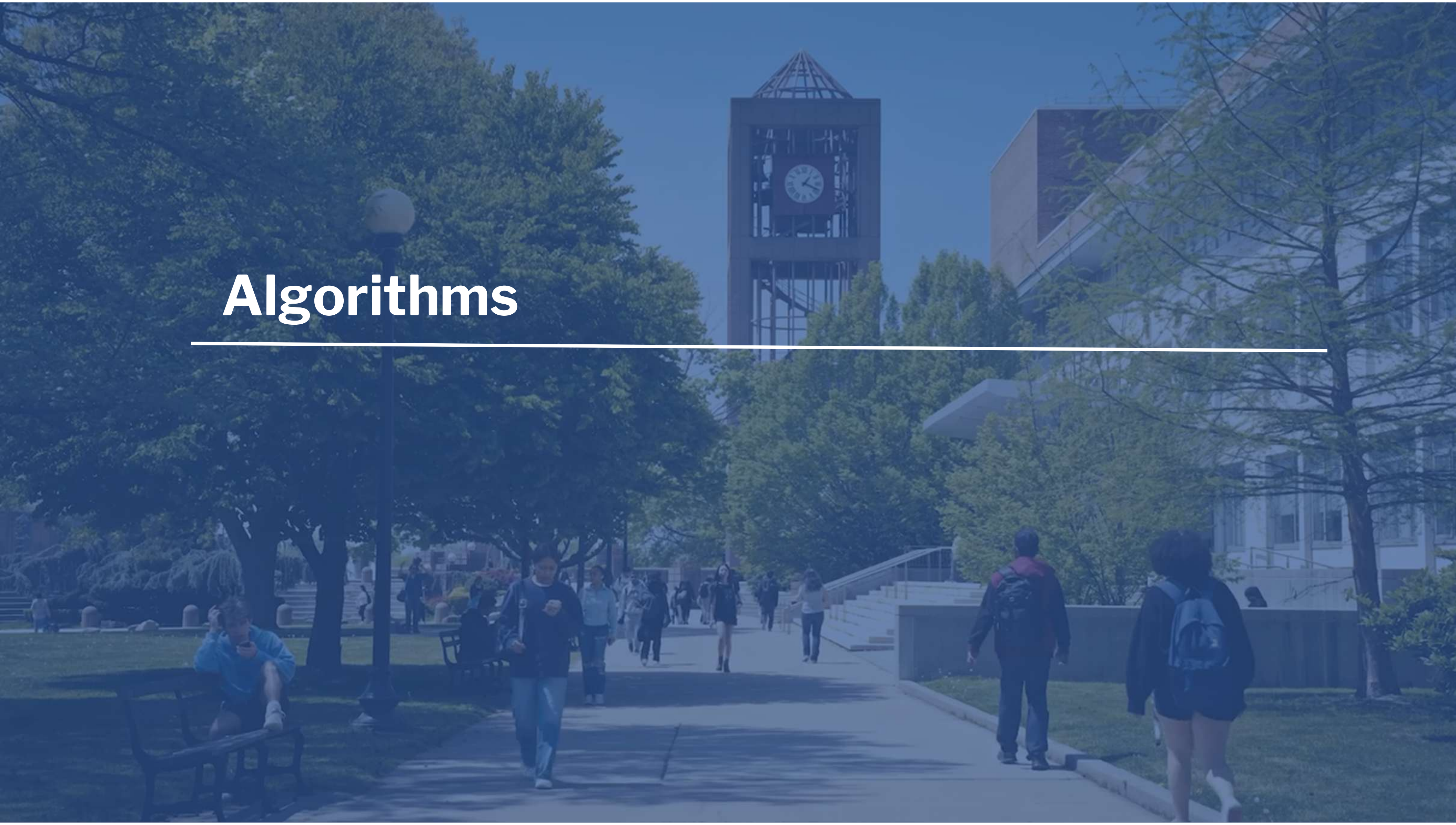
Linear vs Logistic Assumptions

Linear Regression Assumptions

1. Linearity
2. No Multicollinearity (predictor correlation)
3. Independence of Observations
4. Normal Distribution of Errors
5. Variance of errors is zero (homoskedasticity)
6. Errors are Independent (no autocorrelation)

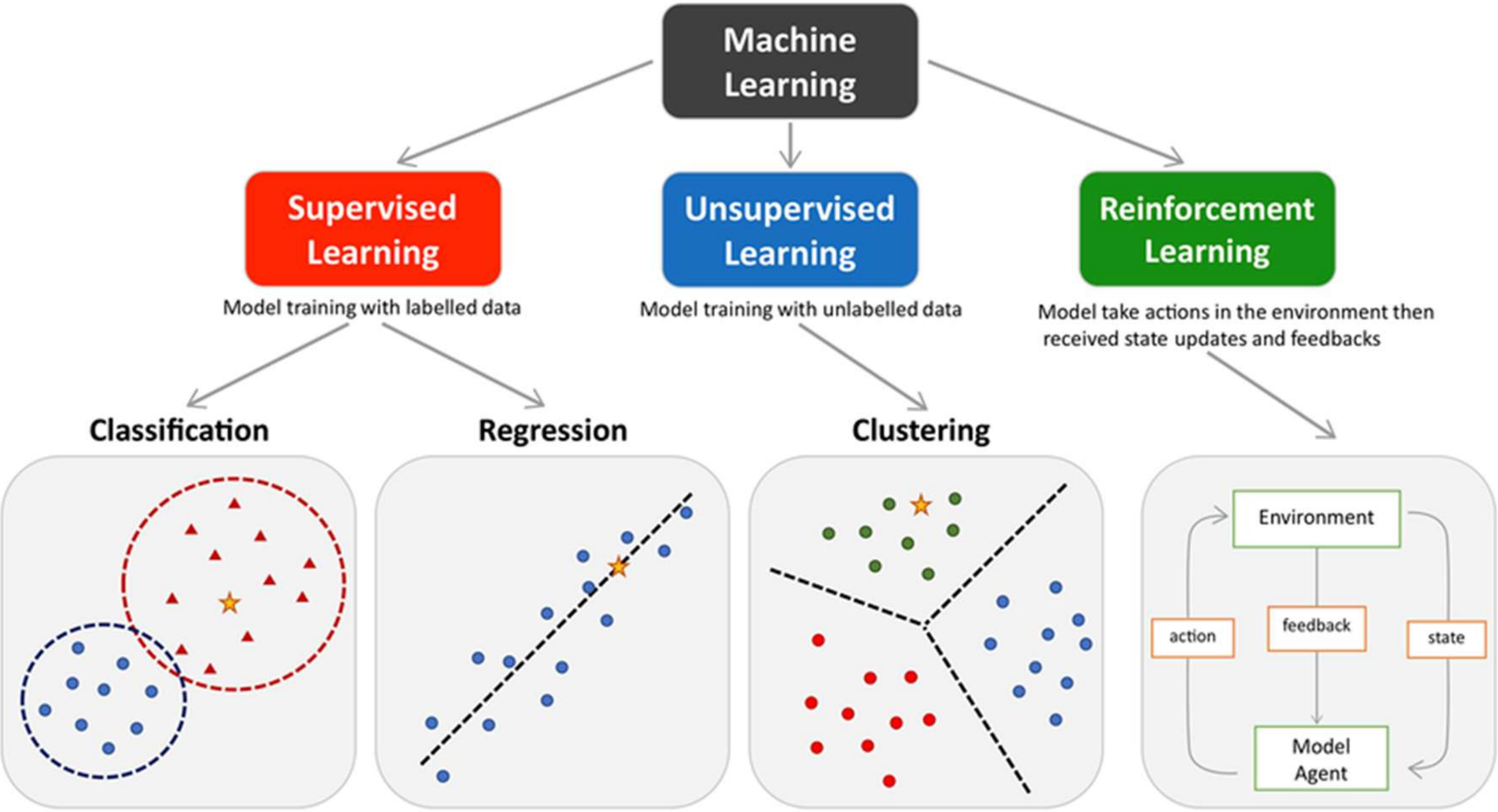
Logistic Regression Assumptions

1. Binary Outcome
2. Log-Linearity
3. Independence of Observations
4. No Multicollinearity
5. Large Sample Size
6. No Outliers



Algorithms

Types of Machine Learning



Machine Learning

Supervised Learning

Model training with labelled data

Unsupervised Learning

Model training with unlabelled data

Reinforcement Learning

Model take actions in the environment then received state updates and feedbacks

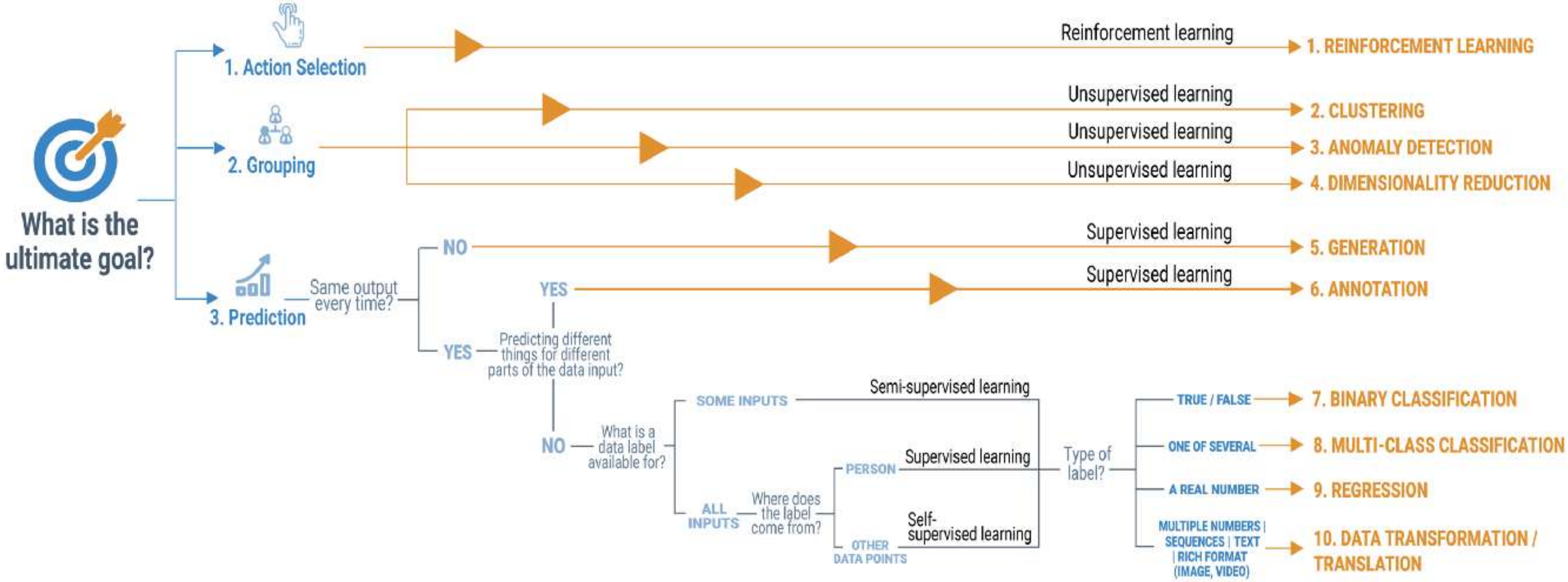
Classification

Regression

Clustering

Source: "Machine Learning Techniques for Personalised Medicine Approaches in Immune-Mediated Chronic Inflammatory Diseases: Applications and Challenges", Pend et al, Sep 2021

Types of Machine Learning



What is the ultimate goal?

1. Action Selection

2. Grouping

3. Prediction

Same output every time?

NO

YES

Predicting different things for different parts of the data input?

YES

NO

What is a data label available for?

SOME INPUTS

ALL INPUTS

Where does the label come from?

PERSON

OTHER DATA POINTS

Semi-supervised learning

Supervised learning

Self-supervised learning

Type of label?

TRUE / FALSE

ONE OF SEVERAL

A REAL NUMBER

MULTIPLE NUMBERS | SEQUENCES | TEXT | RICH FORMAT (IMAGE, VIDEO)

Reinforcement learning → 1. REINFORCEMENT LEARNING

Unsupervised learning → 2. CLUSTERING

Unsupervised learning → 3. ANOMALY DETECTION

Unsupervised learning → 4. DIMENSIONALITY REDUCTION

Supervised learning → 5. GENERATION

Supervised learning → 6. ANNOTATION

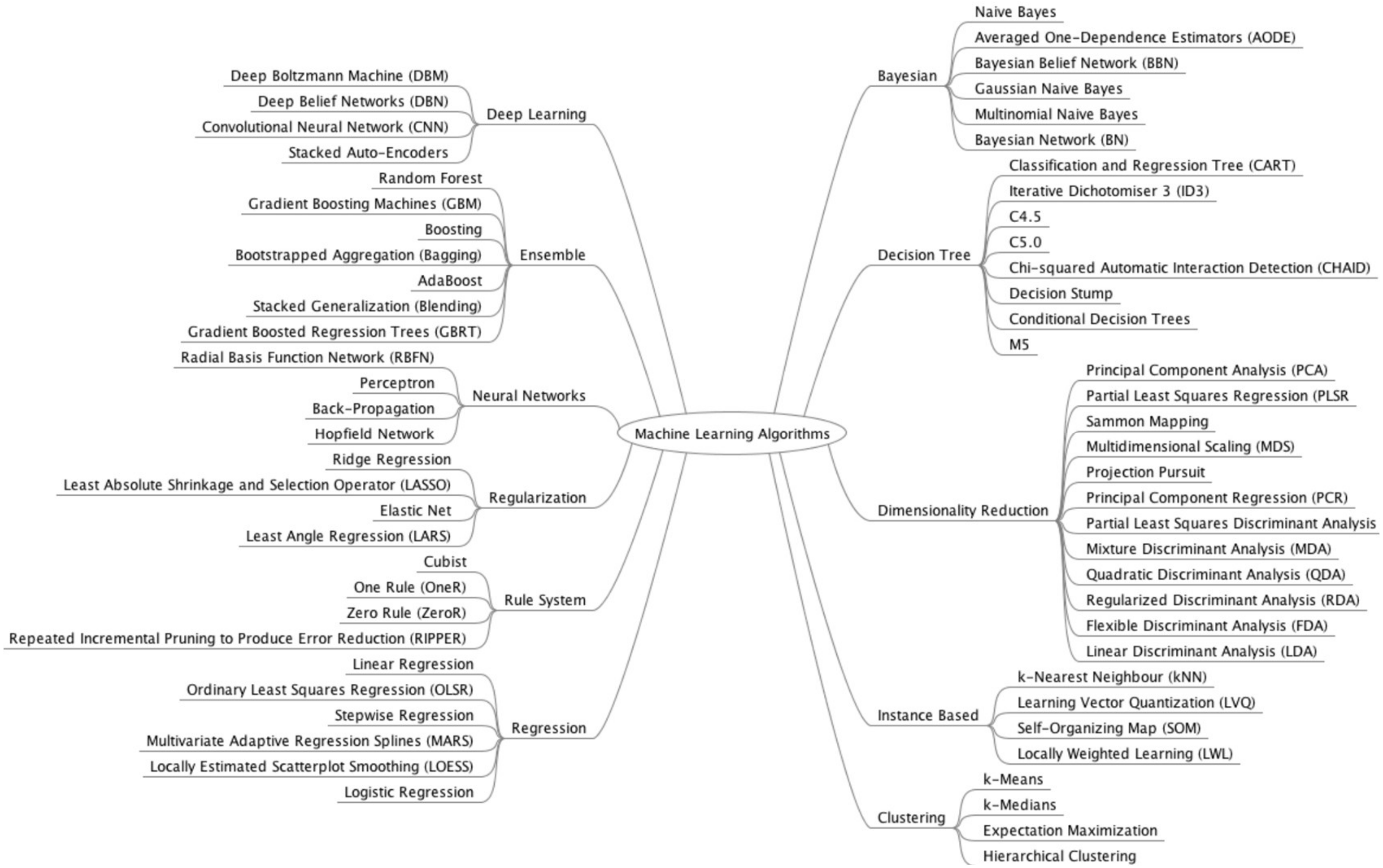
7. BINARY CLASSIFICATION

8. MULTI-CLASS CLASSIFICATION

9. REGRESSION

10. DATA TRANSFORMATION / TRANSLATION

Machine Learning Algorithms



Machine Learning Algorithms

Deep Learning
Deep Boltzmann Machine (DBM)
Deep Belief Networks (DBN)
Convolutional Neural Network (CNN)
Stacked Auto-Encoders

Ensemble
Random Forest
Gradient Boosting Machines (GBM)
Boosting
Bootstrapped Aggregation (Bagging)
AdaBoost
Stacked Generalization (Blending)
Gradient Boosted Regression Trees (GBRT)

Neural Networks
Radial Basis Function Network (RBFN)
Perceptron
Back-Propagation
Hopfield Network

Regularization
Ridge Regression
Least Absolute Shrinkage and Selection Operator (LASSO)
Elastic Net
Least Angle Regression (LARS)

Rule System
Cubist
One Rule (OneR)
Zero Rule (ZeroR)
Repeated Incremental Pruning to Produce Error Reduction (RIPPER)

Regression
Linear Regression
Ordinary Least Squares Regression (OLSR)
Stepwise Regression
Multivariate Adaptive Regression Splines (MARS)
Locally Estimated Scatterplot Smoothing (LOESS)
Logistic Regression

Bayesian
Naive Bayes
Averaged One-Dependence Estimators (AODE)
Bayesian Belief Network (BBN)
Gaussian Naive Bayes
Multinomial Naive Bayes
Bayesian Network (BN)

Decision Tree
Classification and Regression Tree (CART)
Iterative Dichotomiser 3 (ID3)
C4.5
C5.0
Chi-squared Automatic Interaction Detection (CHAID)
Decision Stump
Conditional Decision Trees
M5

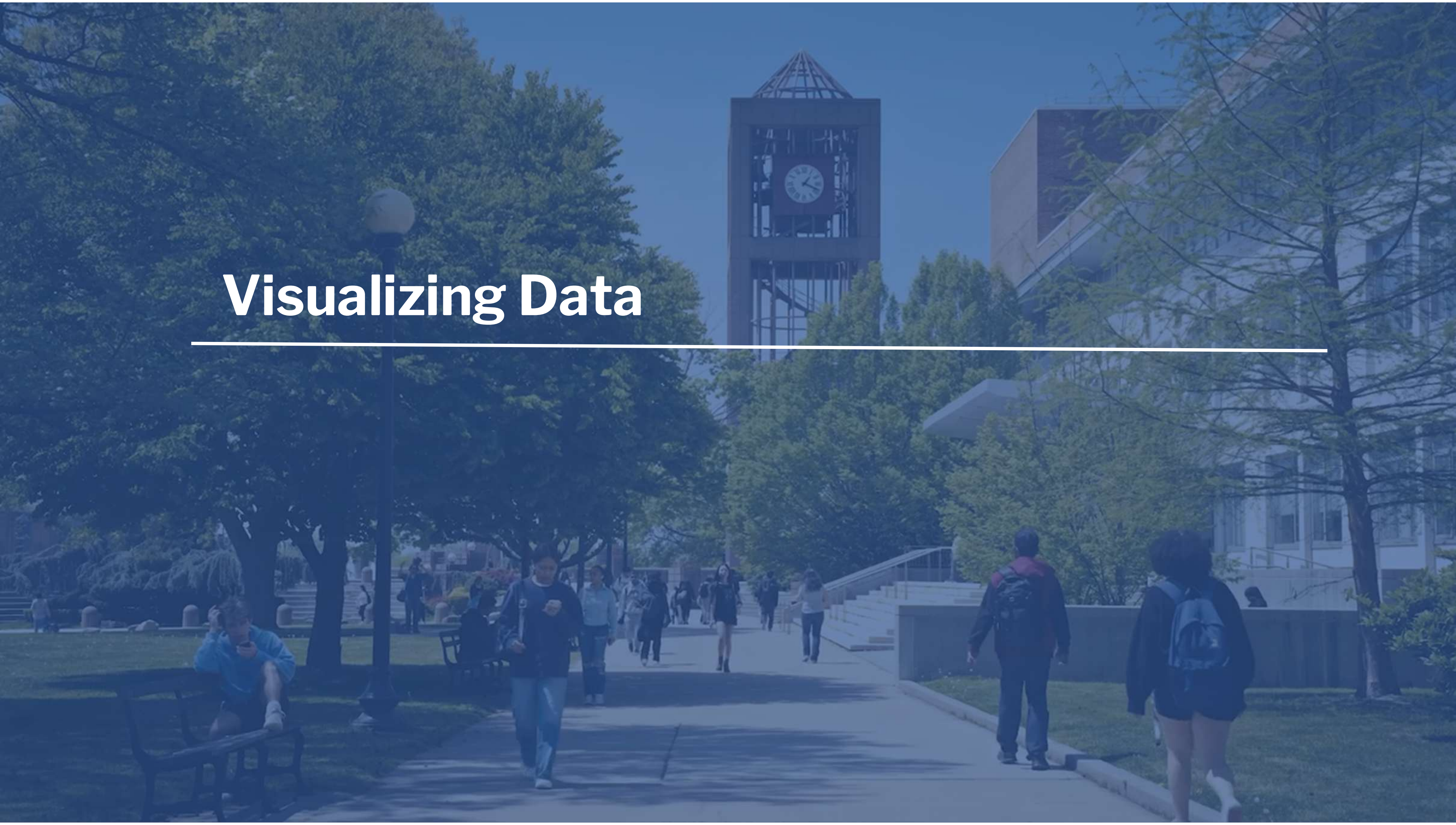
Dimensionality Reduction
Principal Component Analysis (PCA)
Partial Least Squares Regression (PLSR
Sammon Mapping
Multidimensional Scaling (MDS)
Projection Pursuit
Principal Component Regression (PCR)
Partial Least Squares Discriminant Analysis
Mixture Discriminant Analysis (MDA)
Quadratic Discriminant Analysis (QDA)
Regularized Discriminant Analysis (RDA)
Flexible Discriminant Analysis (FDA)
Linear Discriminant Analysis (LDA)

Instance Based
k-Nearest Neighbour (kNN)
Learning Vector Quantization (LVQ)
Self-Organizing Map (SOM)
Locally Weighted Learning (LWL)

Clustering
k-Means
k-Medians
Expectation Maximization
Hierarchical Clustering

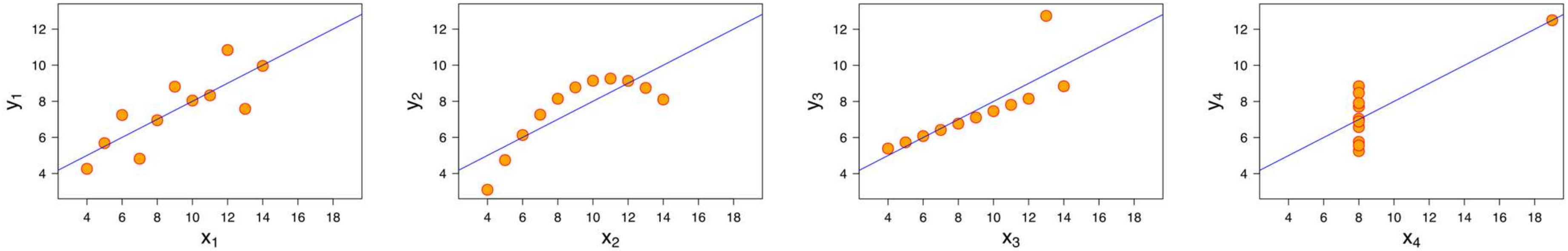
Source: Brownlee, J., 2016

Visualizing Data



Exploratory Data Analysis (EDA)

What do these data sets have in common?



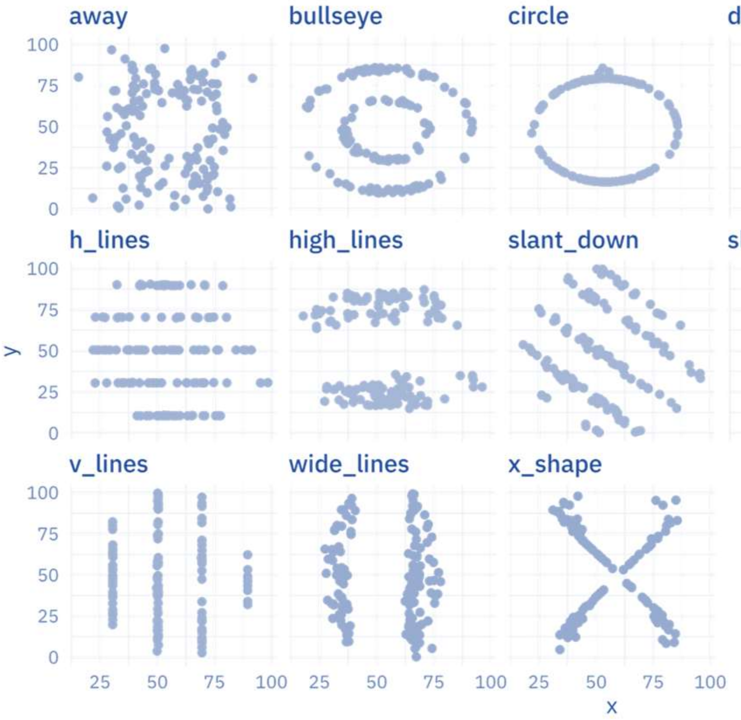
Not much. Except….

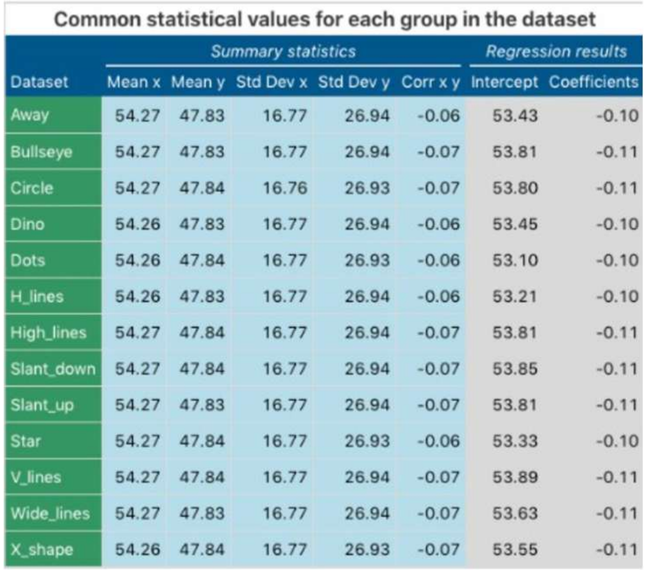
Identical means, variances, correlation, coefficients of determination, regressions

Property	Value
Mean of x	9
Sample variance of x: s_x^2	11
Mean of y	7.50
Sample variance of y: s_y^2	4.125
Correlation between x and y	0.816
Linear regression line	y = 3.00 + 0.500x
Coefficient of determination of the linear regression: R^2	0.67

Source: Anscombe's quartet from Wikipedia

Another Example





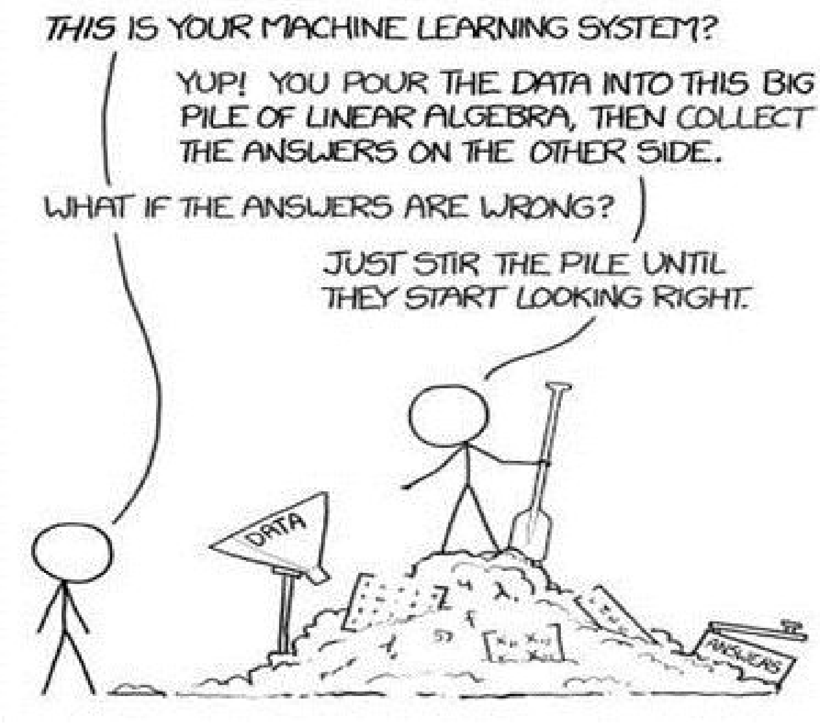
Common statistical values for each group in the dataset

	Summary statistics					*Regression results*	
Dataset	Mean x	Mean y	Std Dev x	Std Dev y	Corr x y	Intercept	Coefficients
Away	54.27	47.83	16.77	26.94	-0.06	53.43	-0.10
Bullseye	54.27	47.83	16.77	26.94	-0.07	53.81	-0.11
Circle	54.27	47.84	16.76	26.93	-0.07	53.80	-0.11
Dino	54.26	47.83	16.77	26.94	-0.06	53.45	-0.10
Dots	54.26	47.84	16.77	26.93	-0.06	53.10	-0.10
H_lines	54.26	47.83	16.77	26.94	-0.06	53.21	-0.10
High_lines	54.27	47.84	16.77	26.94	-0.07	53.81	-0.11
Slant_down	54.27	47.84	16.77	26.94	-0.07	53.85	-0.11
Slant_up	54.27	47.83	16.77	26.94	-0.07	53.81	-0.11
Star	54.27	47.84	16.77	26.93	-0.06	53.33	-0.10
V_lines	54.27	47.84	16.77	26.94	-0.07	53.89	-0.11
Wide_lines	54.27	47.83	16.77	26.94	-0.07	53.63	-0.11
X_shape	54.26	47.84	16.77	26.93	-0.07	53.55	-0.11

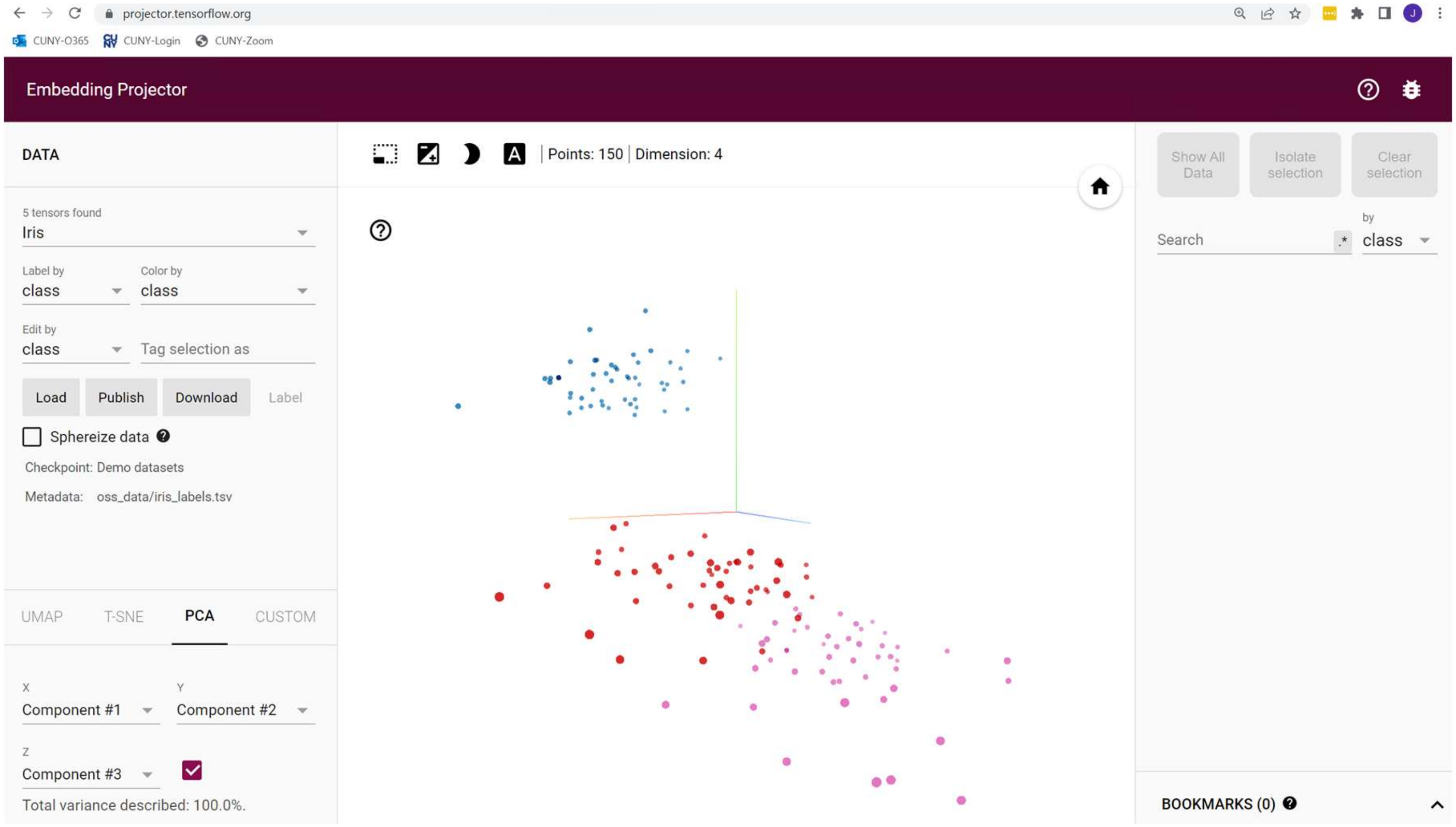
Descriptive statistics can be misleading. Data visualization helps.

Source: Datasaurus Dozen from Wikipedia

What Data Science isn't…


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

Visualizing Data



Points: 150 | Dimension: 4

5 tensors found
Iris

Label by
class

Color by
class

Edit by
class

Tag selection as

Checkpoint: Demo datasets

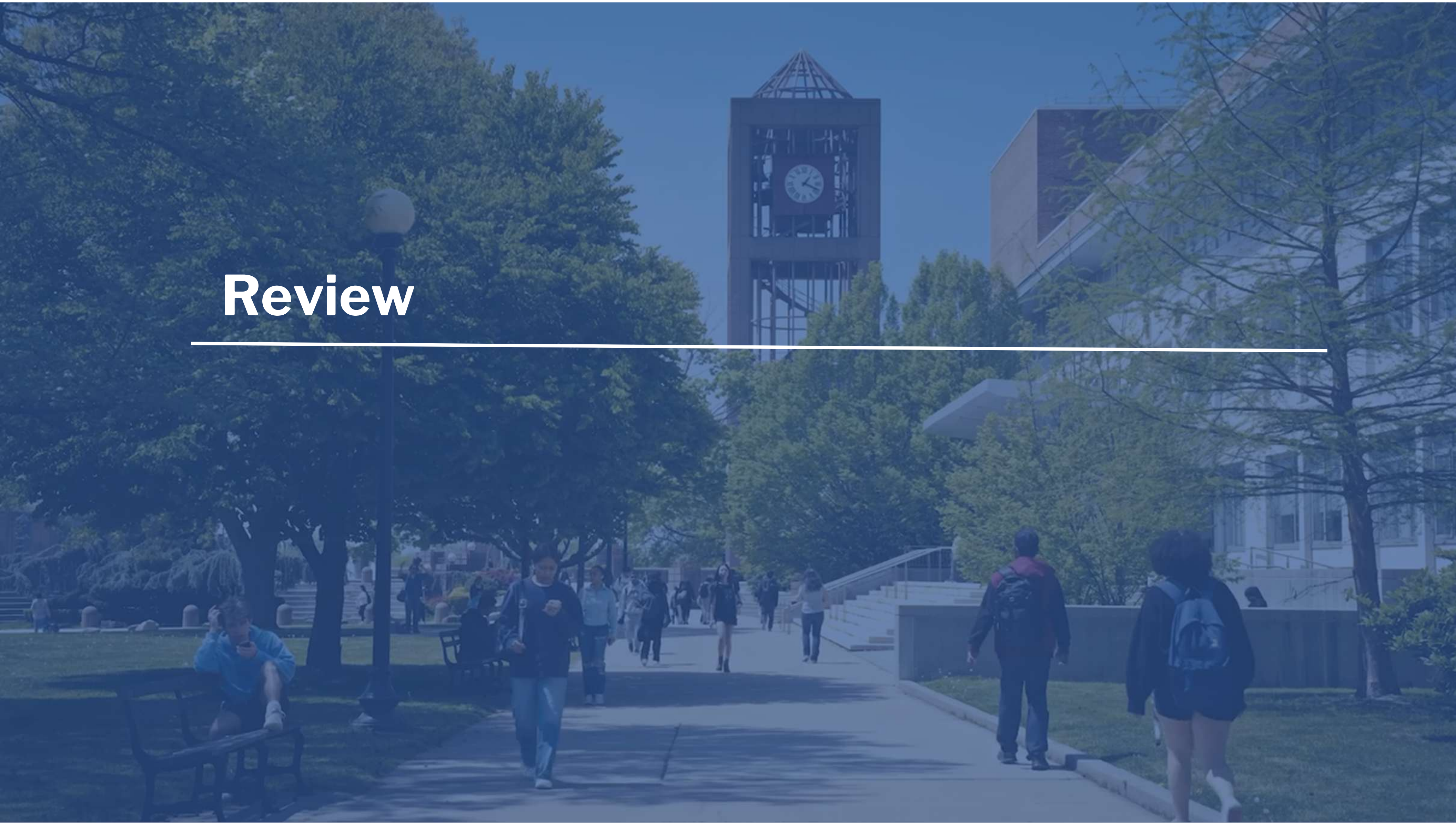
Metadata: oss_data/iris_labels.tsv

X
Component #1

Y
Component #2

Z
Component #3 ☑

Total variance described: 100.0%.



Review

This week we covered

Lesson Objectives/Topics

1. Fit and interpret linear regression models to identify relationships between variables
2. Apply logistic regression for binary classification tasks in business scenarios
3. Evaluate model performance using error metrics and classification accuracy
4. Explain the assumptions and limitations of regression techniques

CUNY
School of Professional Studies