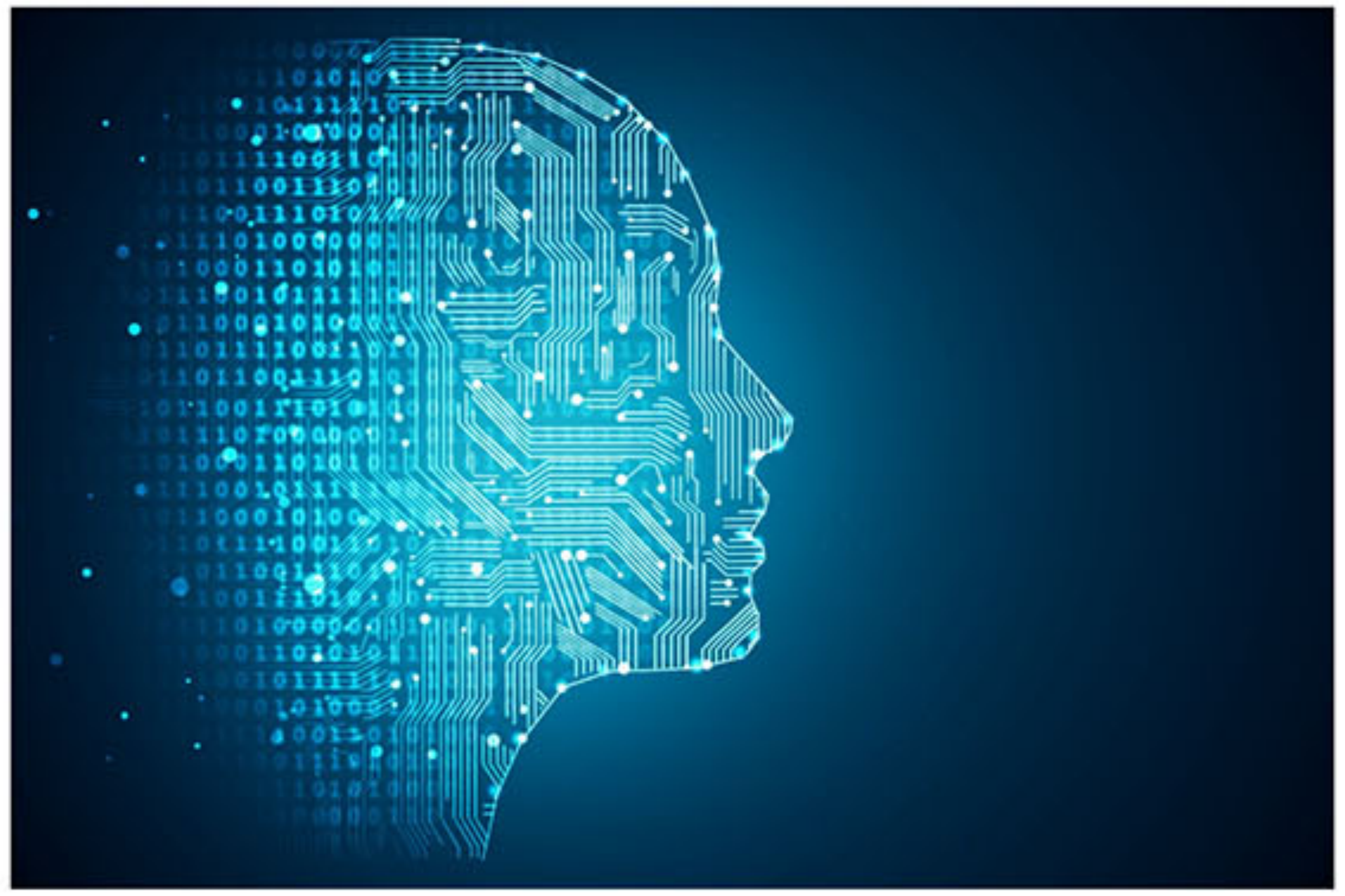
PRINCETON UNIVERSITY

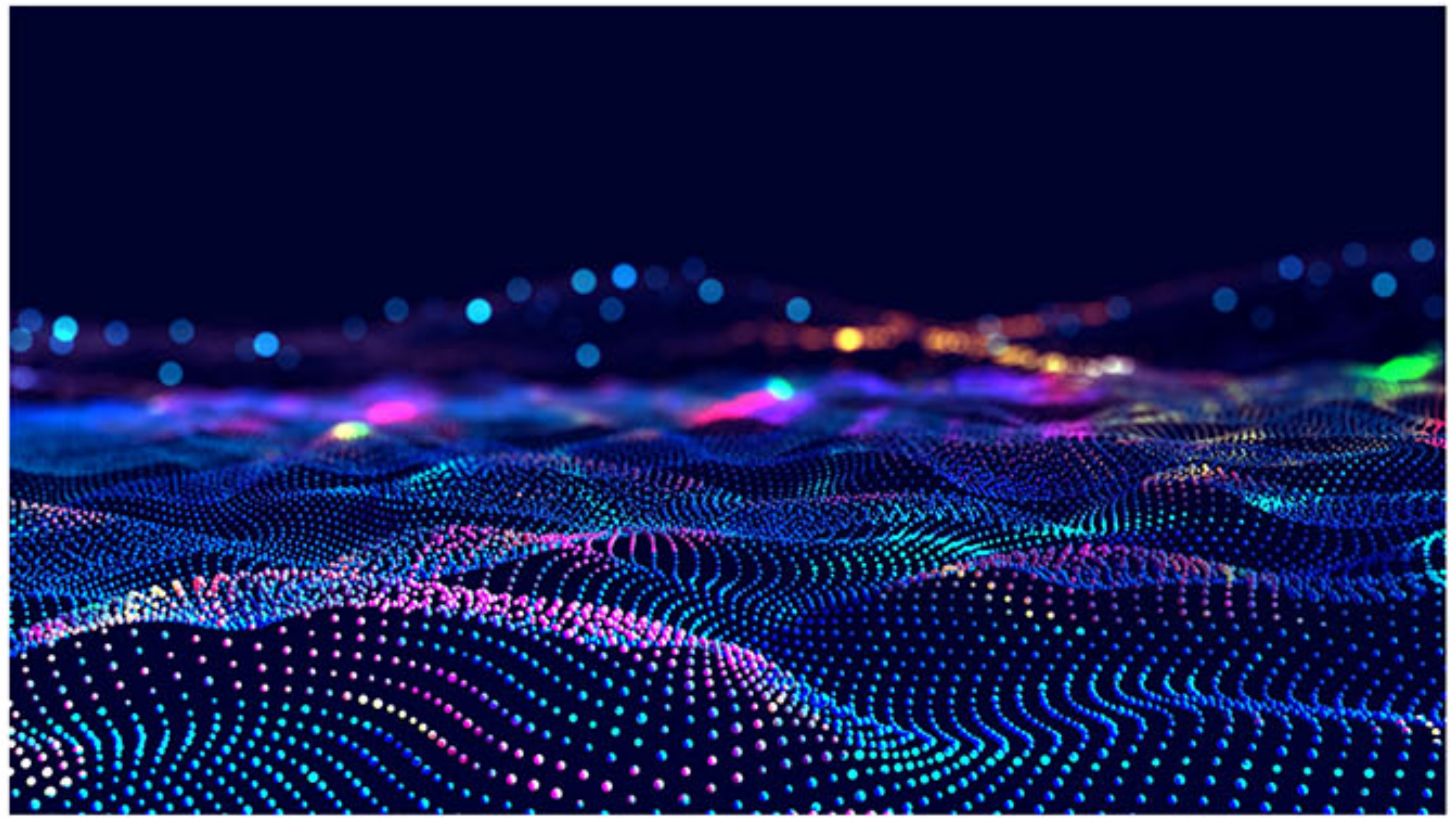
CENTER FOR STATISTICS AND MACHINE LEARNING

A Hands–On Introduction to Machine Learning



Wintersession 2025
January 15–17,21

Julian Gold
Gage DeZoort



With materials from:

Brian Arnold, Gage DeZoort, Julian Gold, Jonathan Halverson, Christina Peters, Savannah Thias, Amy Winecoff

DataX
Accelerating Scientific Discovery at Princeton

Mini-Course Outline

Date	Topic	Instructor
Wed. 1/15	Machine Learning Overview and Simple Models	Julian Gold
Thu. 1/16	Model Evaluation and Improving Performance	Julian Gold
Fri. 1/17	Introduction to Neural Networks	Gage DeZoort
Tue. 1/21	Survey of Neural Network Architectures	Gage DeZoort
Wed.+Thu. 1/22-1/23	Getting Started with LLMs with PLI	Simon Park, Abhishek Panigrahi
Wed. 1/22	Graph Neural Networks for Your Research	Gage DeZoort
Wed. 1/22	Machine Learning for the Physical Sciences	C. Jespersen, R. Pastrana, Q. Gallagher, H. Johnson

Agenda

- K-nearest Neighbors
  - Regression and classification
- Clustering with K-means
- Evaluation paradigms and improving performance

Intro to *K*-nearest neighbors (KNN)

- simple but powerful
- can be used for classification *or* regression!
- algorithm

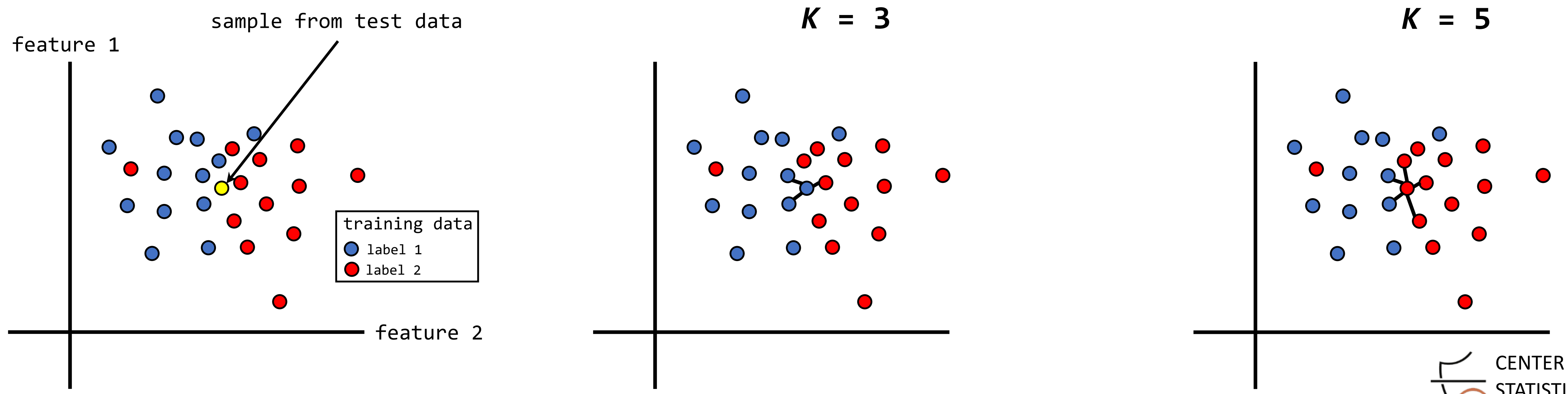
  1. for a given test sample (yellow dot), find the *K* nearest training samples in feature space

  2a. for **classification**, assign label by majority vote

  2b. for **regression**, assign value by mean of neighbors

K is a tunable parameter!

- choose value that gives better predictions on test data



***K* = 3**

***K* = 5**

Coding in Python!

https://github.com/PrincetonUniversity/intro_machine_learning/tree/main/day2
https://jdh4.github.io/intro-ml

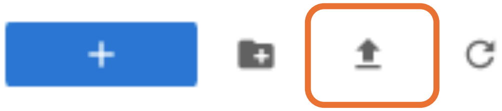
Google Colab

1. Open notebook (.ipynb):

2. Click “Open in Colab”:

JupyterLite

1. Download notebook (.ipynb) from left.
2. Open JupyterLite:
3. Upload the notebook and open:



DataX
Accelerating Scientific Discovery at Princeton

K-Means Clustering Visualization

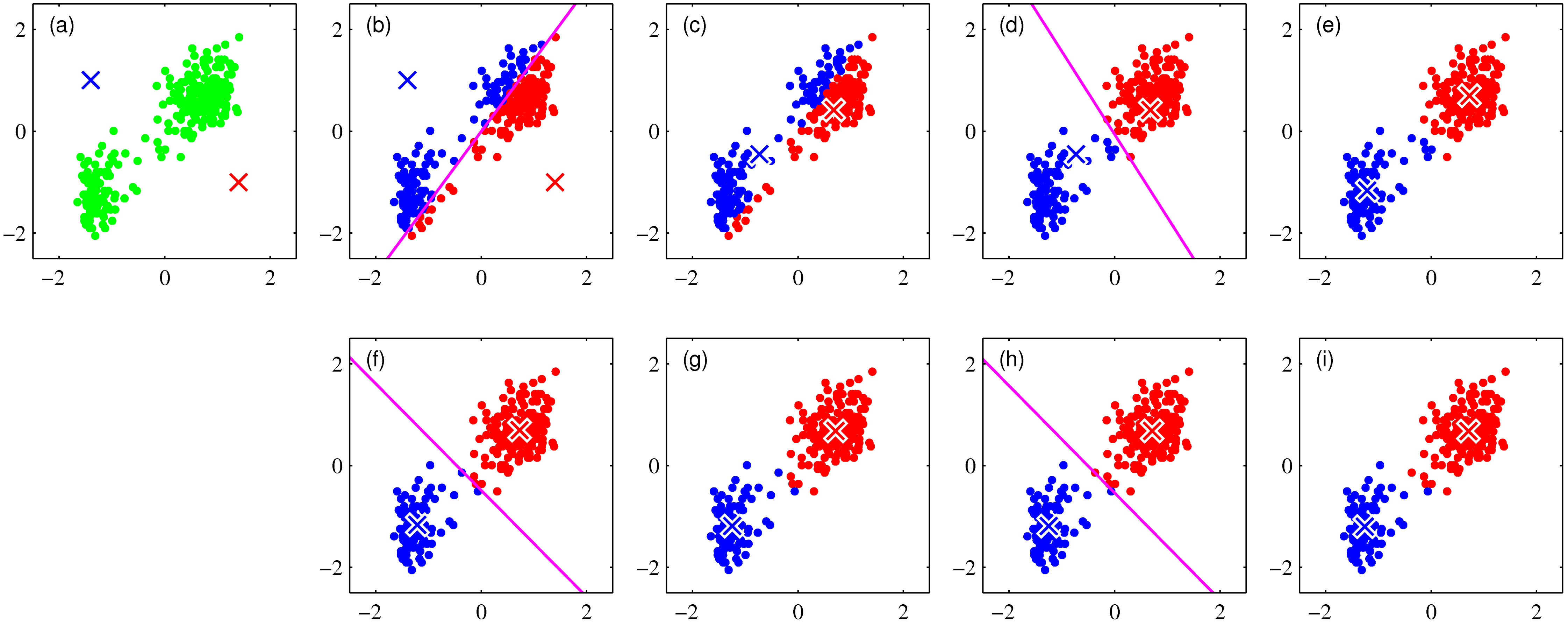


Figure credit: Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning*.

DataX
Accelerating Scientific Discovery at Princeton

Overview of Machine Learning Process

Datasets? Input features?
Targets? Evaluation metrics?

1. Define the problem to be solved.
2. Split the data into train / validation / test.
3. Run the validation loop:
   a. Choose a set of models.
   b. Train each model by optimizing its parameters on the training set.
   c. Evaluate the performance of each model on the validation set.
   d. Repeat until performance is satisfactory.
4. Evaluate final performance on the test set.