CME 250:
Introduction to Machine Learning

Lecture 1:
Overview of Machine Learning

Sherrie Wang
sherwang@stanford.edu

Agenda

Slides are online at cme250.stanford.edu

- Definition of machine learning, applications
- Course logistics
- Machine learning overview
- K-nearest neighbors (KNN)

What is machine learning?

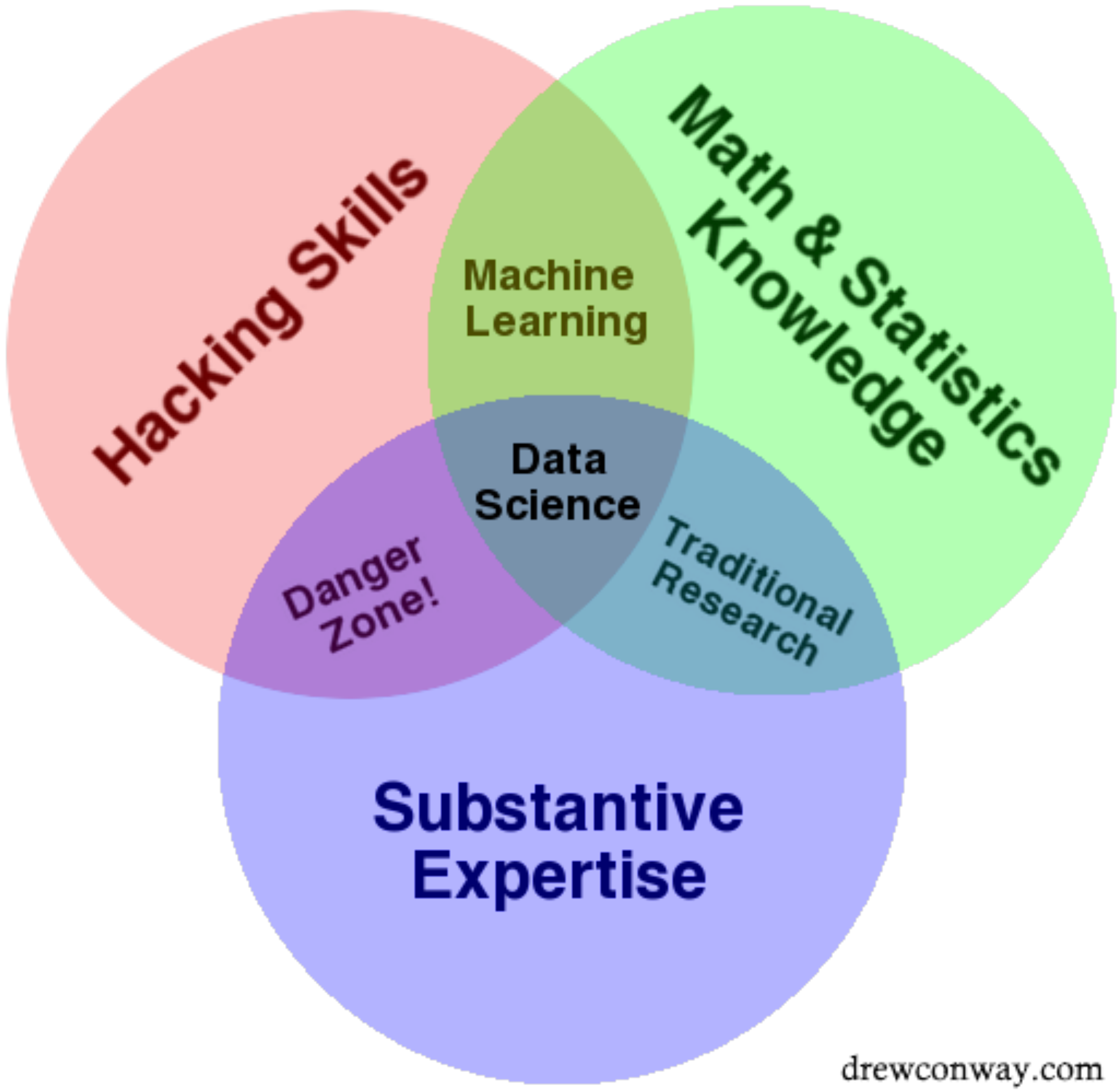
“[A] field of study that gives computers the ability to learn without being explicitly programmed.”

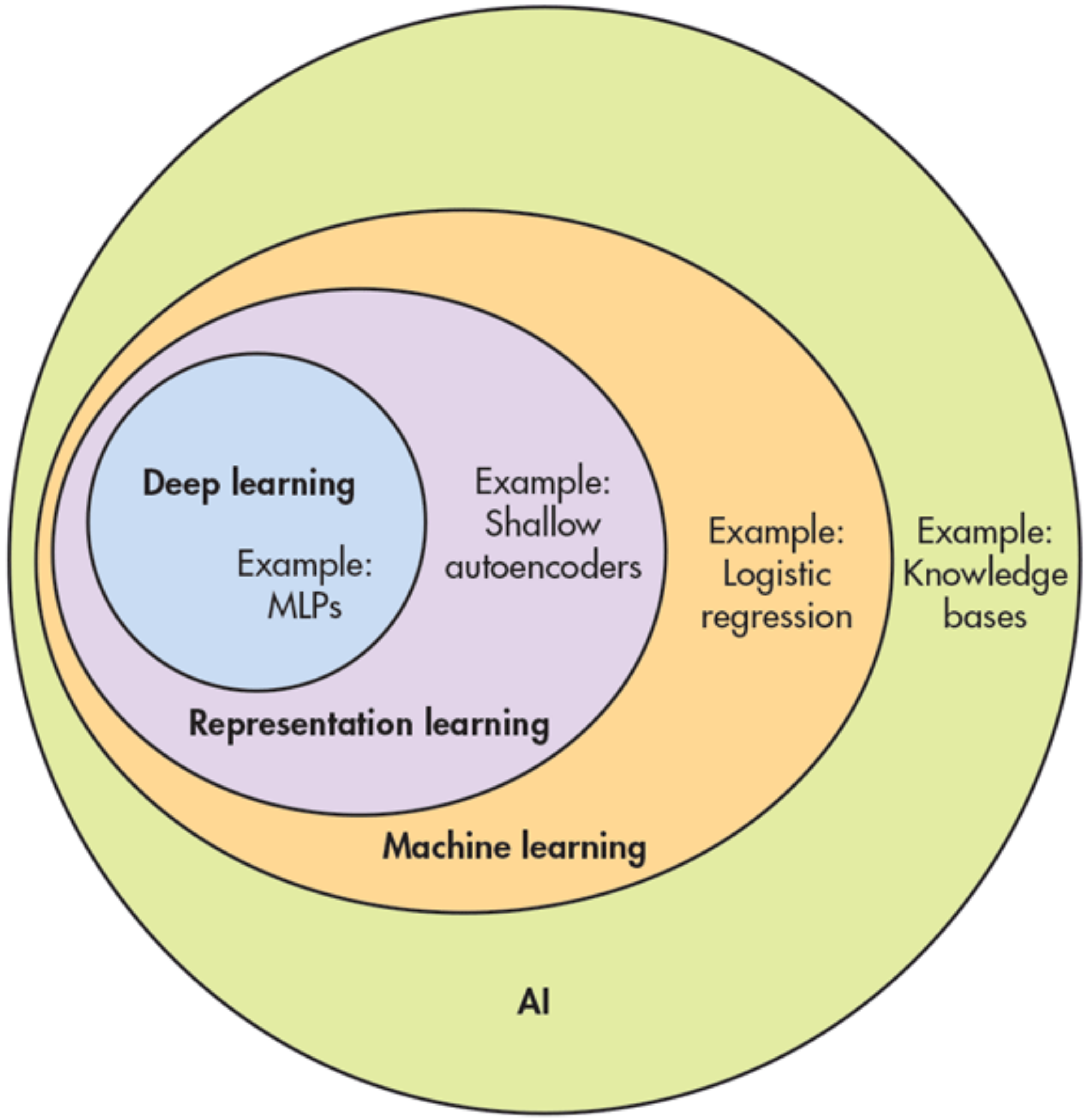
- Arthur Samuel (1959)

“A computer program is said to learn from experience E with some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E.”

- Tom M. Mitchell (1997)

What is machine learning?





Timeline of Machine Learning

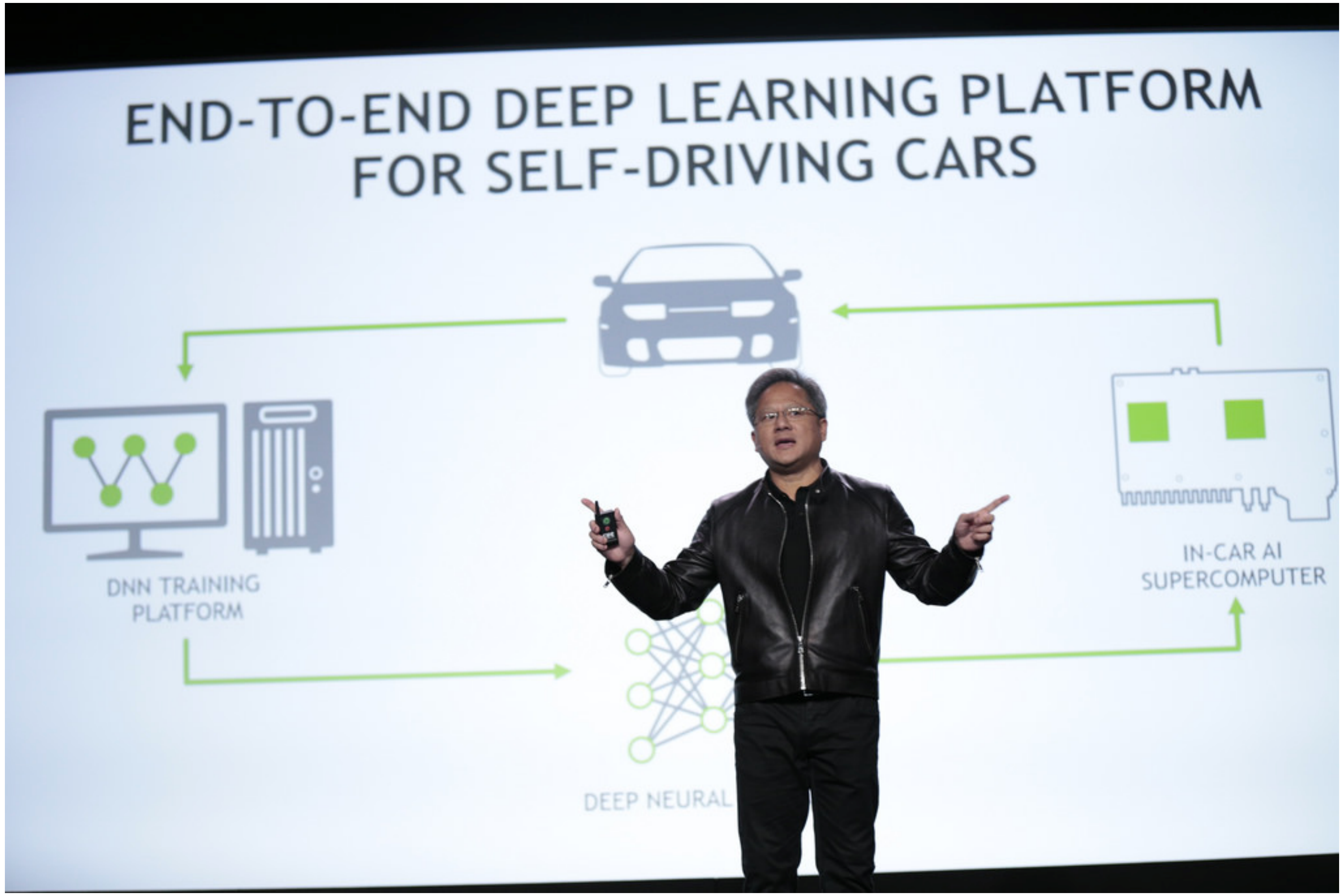
- **1805:** Legendre discovers the least squares method, the earliest form of linear regression.
- **1936:** Fisher proposes linear discriminant analysis.
- **1940s:** Various authors propose logistic regression.
- **1951:** Minsky and Edmonds build the first neural network machine, the SNARC.
- **1957:** Rosenblatt invents the perceptron, a binary classifier.
- **1967:** The nearest neighbor algorithm is created.
- **1970s:** AI winter caused by pessimism about machine learning effectiveness.
- **1980s:** Breiman, Friedman, Olshen, and Stone introduce CARTs. Backpropagation is rediscovered, causing a resurgence in machine learning research.
- **1995:** Ho describes random forests; Cortes and Vapnik introduce SVMs.
- **1997:** IBM's Deep Blue beats Kasparov, the world champion at chess.
- **2009:** ImageNet is created in Fei-Fei Li's group at Stanford. A catalyst for the current AI boom.
- **2016:** Google's AlphaGo defeats an unhandicapped human professional at Go.

Machine Learning Applications

Intelligent personal assistants (Alexa, Google Assistant, Siri)

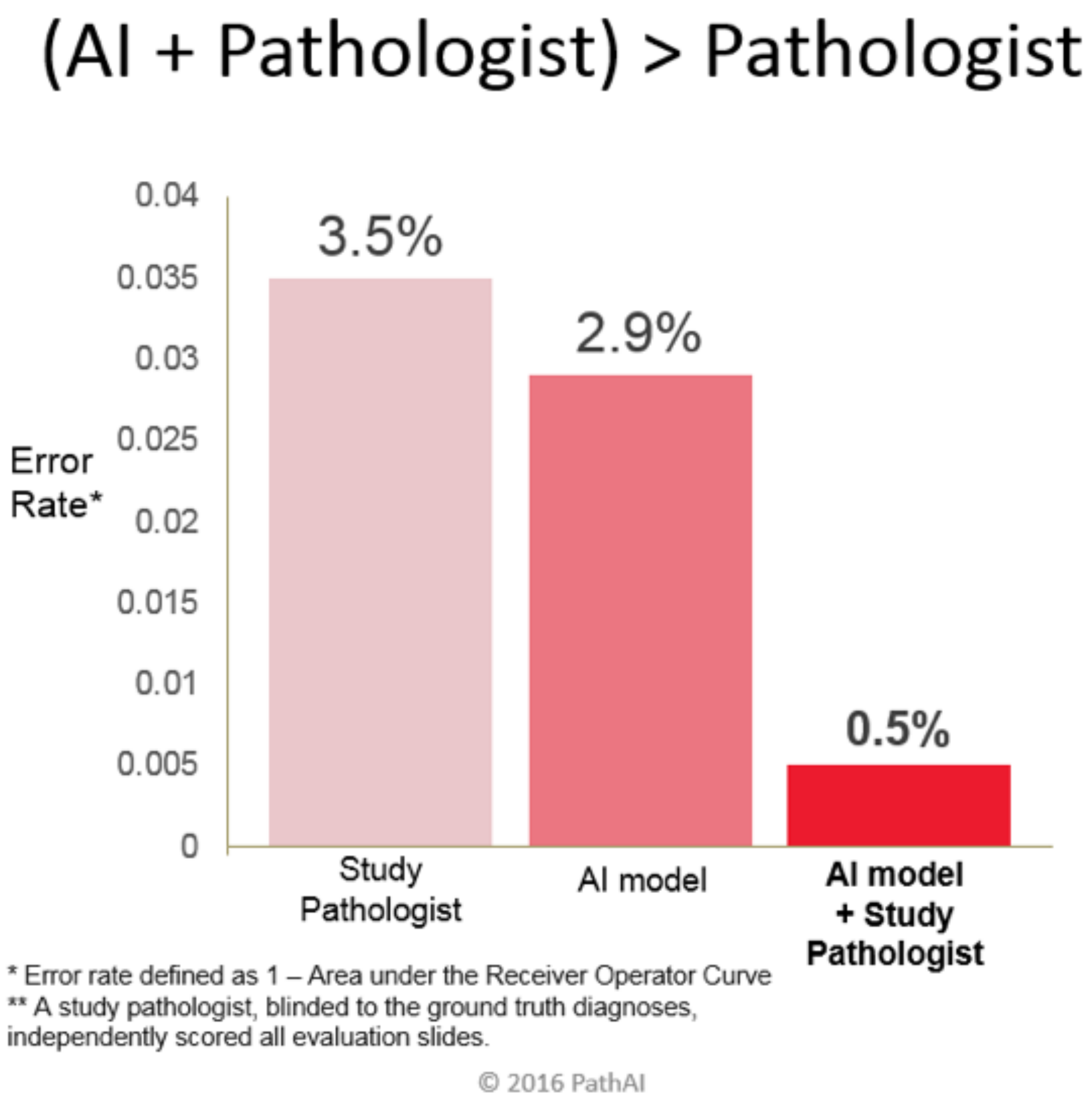
Machine Learning Applications

Self-driving cars



Machine Learning Applications

Deep learning for oncology



Machine Learning Applications

Recommender systems

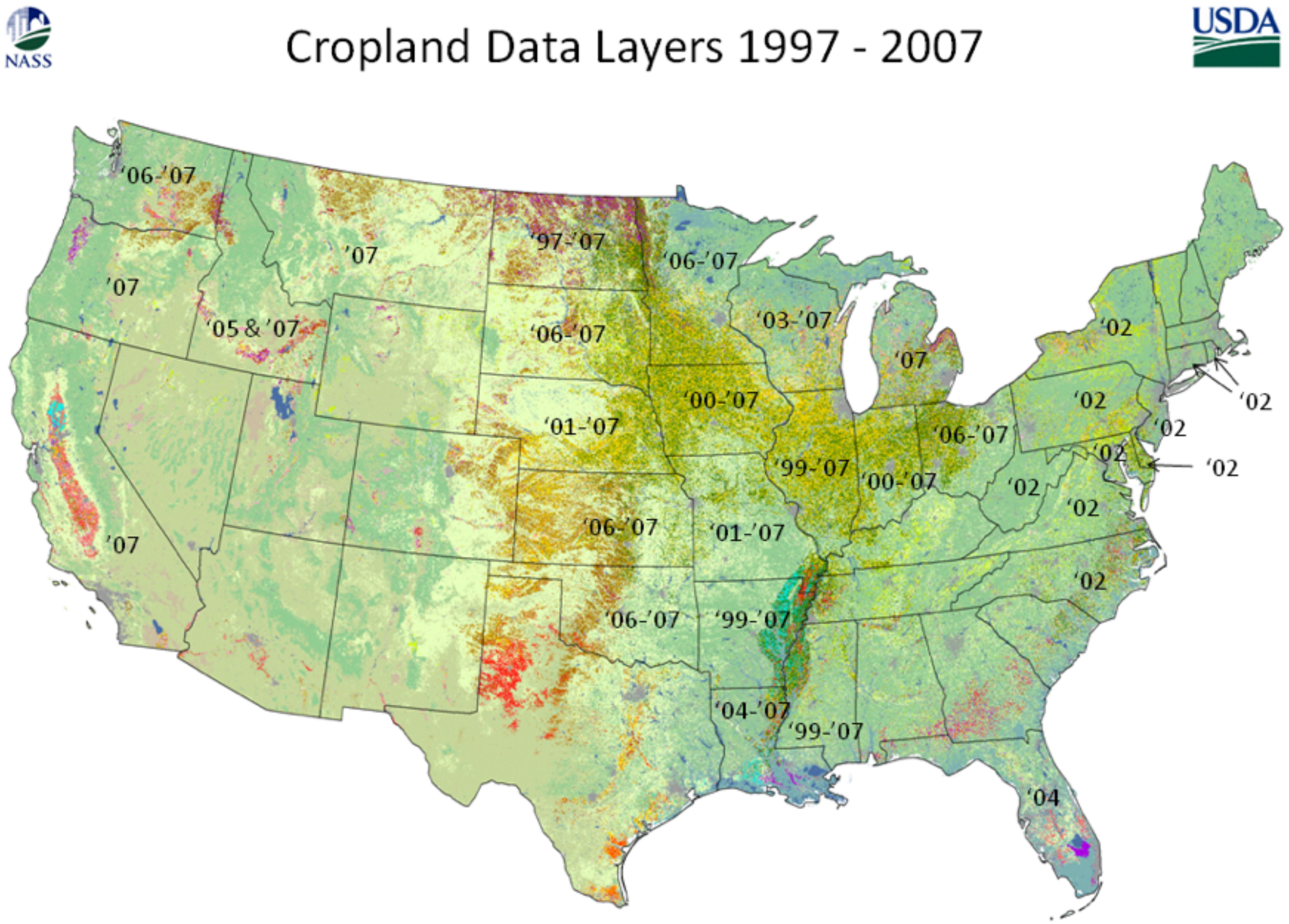
Machine Learning Applications

Legal contract review

LawGeex

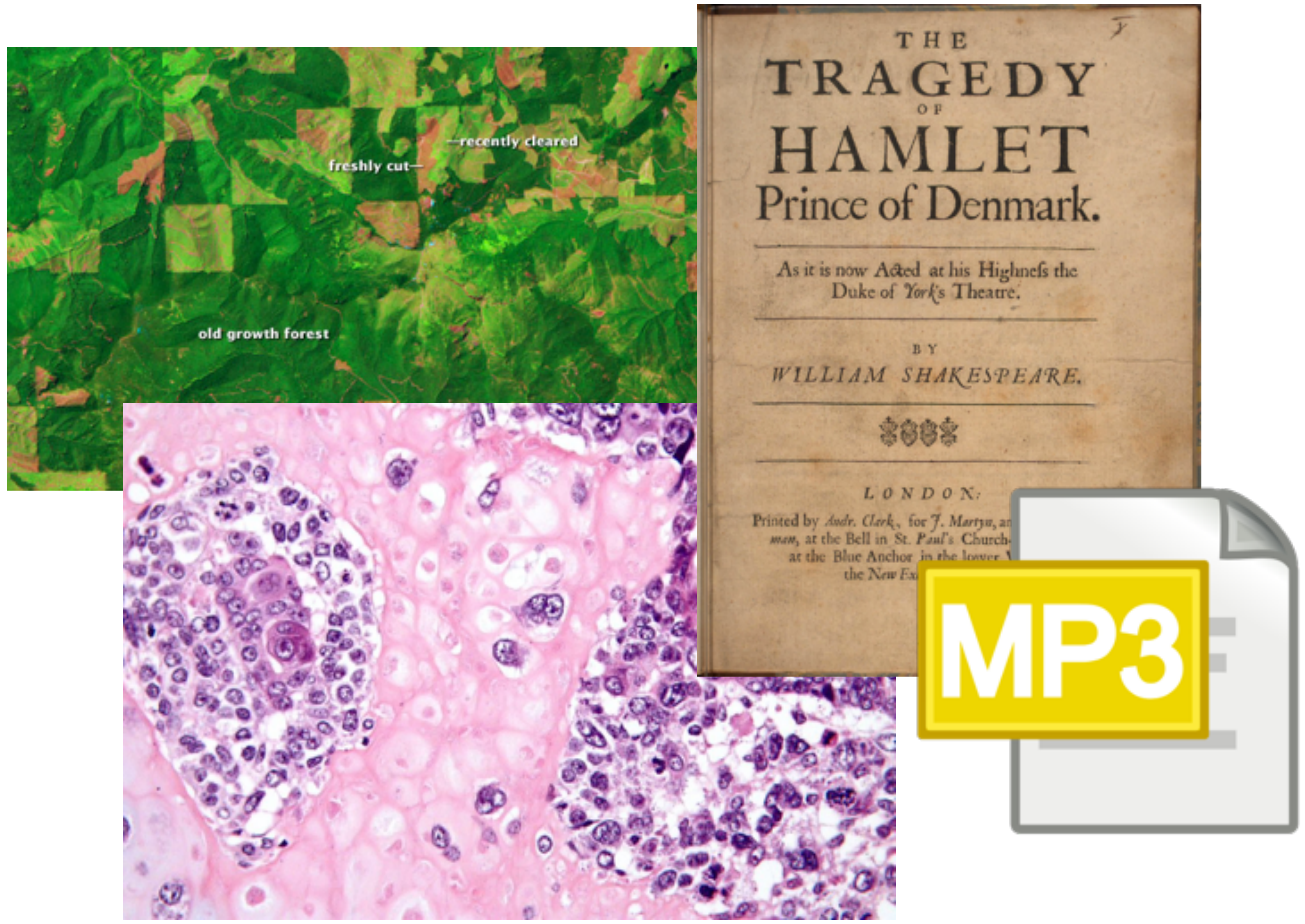
Machine Learning Applications

Mapping agriculture from satellite imagery

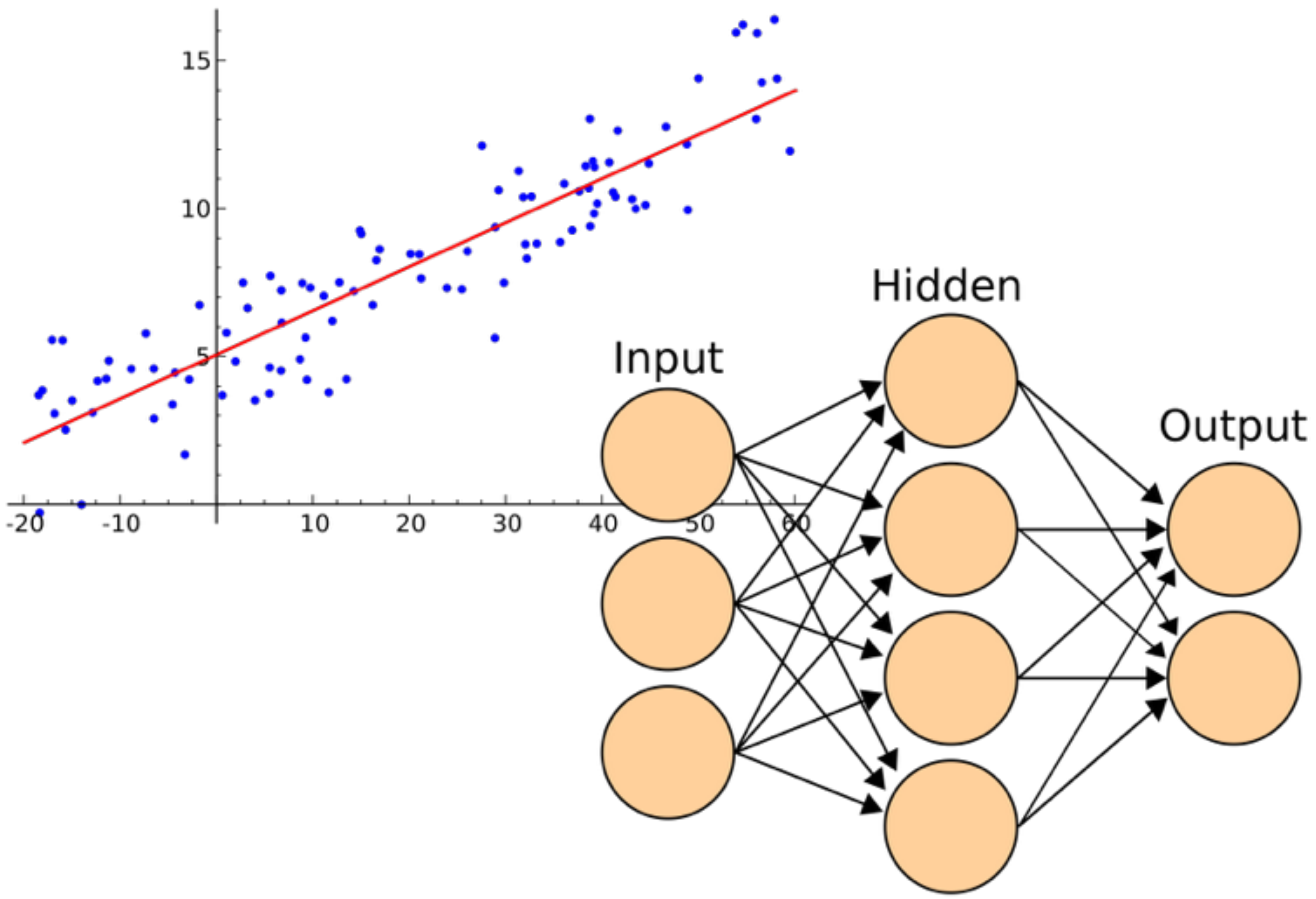


Machine Learning Applications

Data



Model



Course Logistics

Course Overview

Goals:

- High-level overview of well-known machine learning techniques
- Learn how to choose between methods
- Practical tips and best practices
- Familiarity with terminology

Intended Audience

- All disciplines welcome
- No background in machine learning is necessary
  - Course covers a subset of CS 229, STATS 315
- Prerequisites
  - Undergraduate-level statistics and linear algebra
  - Basic programming experience (Python, R, MATLAB)

Machine Learning Courses

Introduction

CME 250: Introduction to Machine Learning

CS 229A: Applied Machine Learning

Foundations

CS 229: Machine Learning

CS 221: Artificial Intelligence

CS 230: Deep Learning

STATS 315A/B: Modern Applied Statistics

CS 234: Reinforcement Learning

Theory

CS 229T: Statistical Learning Theory

Applications

CS 224N: Natural Language Processing with Deep Learning

CS 325B: Data for Sustainable Development

CS 231N: Convolutional Neural Networks for Visual Recognition

CS 273B: Deep Learning in Genomics and Biomedicine

CS 246: Mining Massive Data Sets

…and much more

Course Schedule

Week	Mon	Tue	Wed	Thu	Fri
1	First day of quarter				
2	Lecture 1		Lecture 2		
3	No lecture (MLK)		Lecture 3		HW 1 due by 5pm
4	No lecture		No lecture		HW 2 due by 5pm
5	Lecture 4		Lecture 5		
6	Lecture 6		Lecture 7		HW 3 due by 5pm
7	No lecture (President's)		Lecture 8		HW 4 due by 5pm

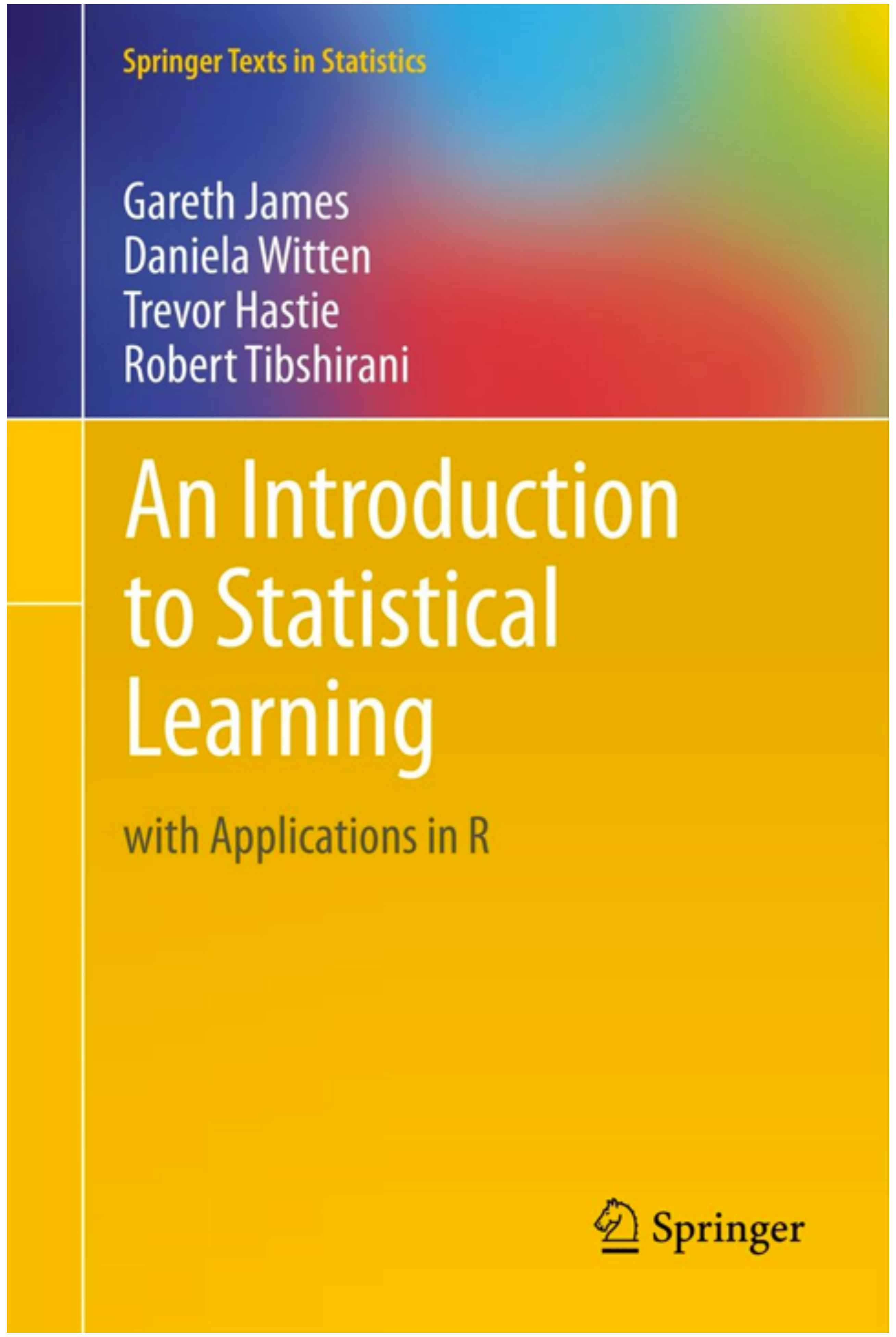
Course Schedule

Lecture 1	Lecture 2	Lecture 3	Lecture 4
Overview of Machine Learning	Linear and Logistic Regression	Regularization and Sparsity	Cross-validation and Imputation
Lecture 5	**Lecture 6**	**Lecture 7**	**Lecture 8**
Support Vector Machines	Classification and Regression Trees (CART)	Unsupervised Methods	Neural Networks

Course Texts

An Introduction to Statistical Learning with Applications in R
by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani

The Elements of Statistical Learning
by Trevor Hastie, Robert Tibshirani, and Jerome Friedman



* Some of the figures in this presentation are taken from *"An Introduction to Statistical Learning, With Applications in R"* (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani

Course Requirements

- Short course, 1 unit
- Grading: Satisfactory / No Credit
- To receive credit, complete 4 homeworks at passing level (70+%)

Homework

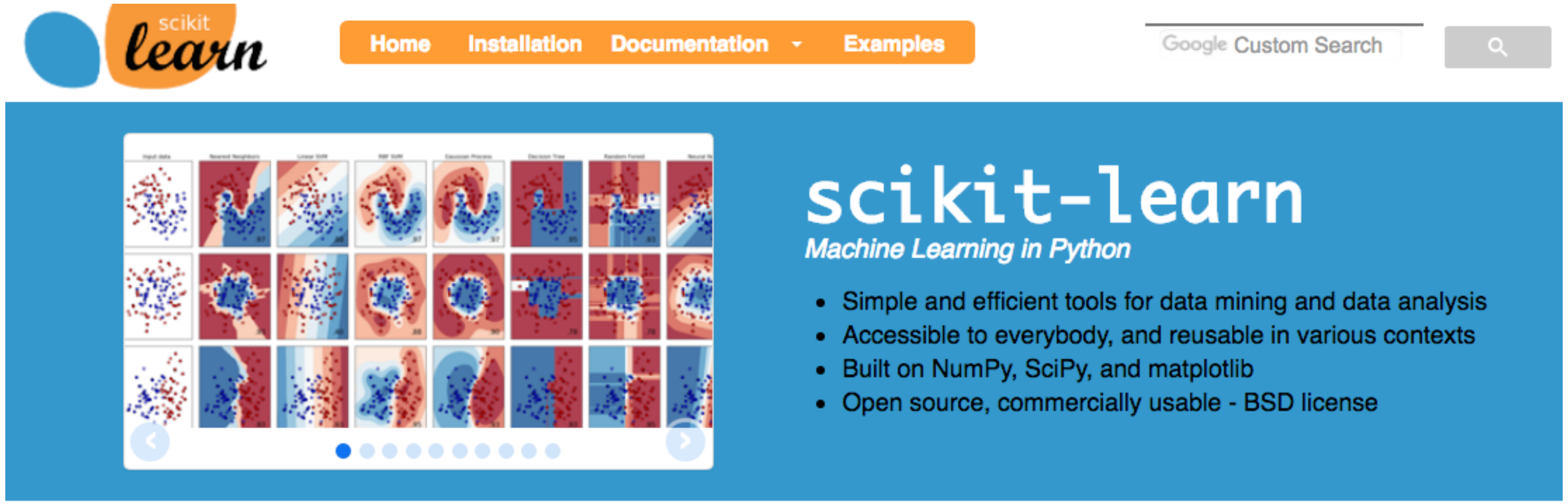
- **Part 1:** Conceptual multiple choice questions (10 pts)
- **Part 2:** Short applied exercises (5-10 pts)
- **Part 3:** Apply method covered in lecture to real dataset (5-10 pts)
- Each homework will be worth 25 points
- (Satisfactory grade = 70/100 or better)

Programming

- Course is not programming intensive, but does require students to write some code
- Students may use language of their choice for any programming exercises (e.g. Python, MATLAB, R)
- ISL textbook gives examples in R, a programming language with existing libraries for statistical analysis and machine learning
- Lecture programming examples will be given in Python, which has well-supported data science / machine learning libraries

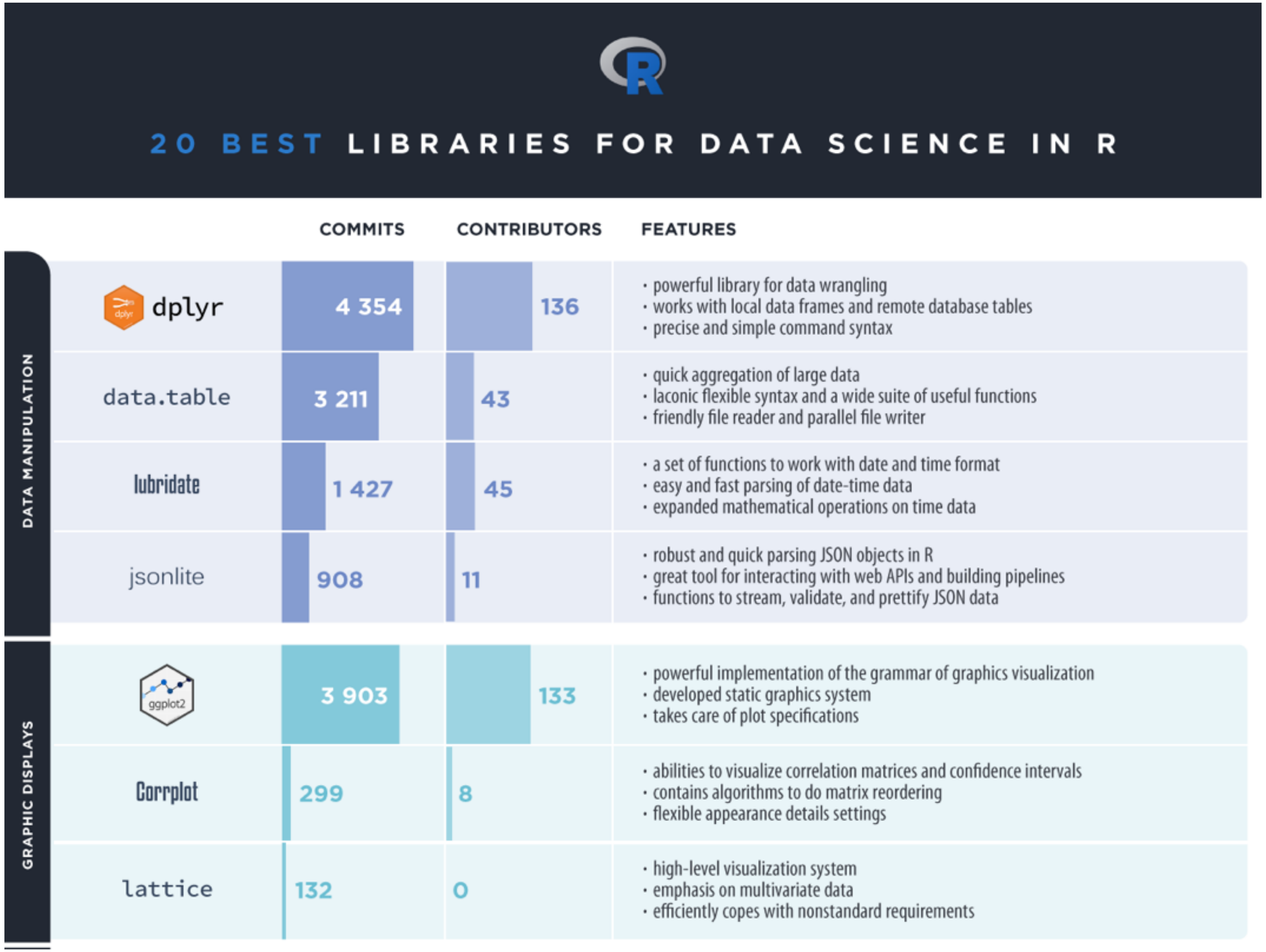
Python

- Python: www.python.org
- NumPy: http://www.numpy.org
- Pandas: https://pandas.pydata.org
- Scikit-learn: http://scikit-learn.org



R

- R: www.r-project.org
- A variety of packages for machine learning

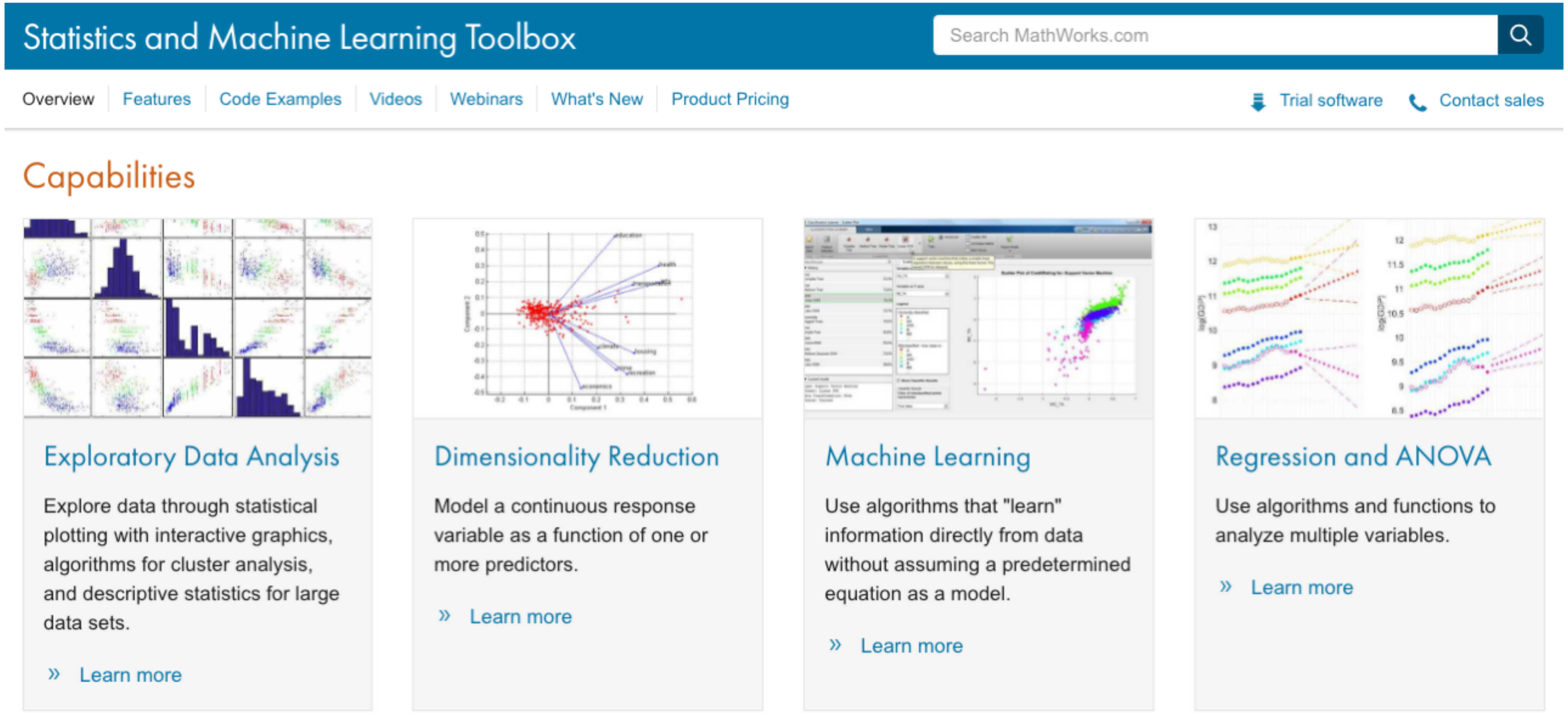


20 BEST LIBRARIES FOR DATA SCIENCE IN R

		COMMITS	CONTRIBUTORS	FEATURES
DATA MANIPULATION	dplyr	4 354	136	• powerful library for data wrangling • works with local data frames and remote database tables • precise and simple command syntax
	data.table	3 211	43	• quick aggregation of large data • laconic flexible syntax and a wide suite of useful functions • friendly file reader and parallel file writer
	lubridate	1 427	45	• a set of functions to work with date and time format • easy and fast parsing of date-time data • expanded mathematical operations on time data
	jsonlite	908	11	• robust and quick parsing JSON objects in R • great tool for interacting with web APIs and building pipelines • functions to stream, validate, and prettify JSON data
GRAPHIC DISPLAYS	ggplot2	3 903	133	• powerful implementation of the grammar of graphics visualization • developed static graphics system • takes care of plot specifications
	Corrplot	299	8	• abilities to visualize correlation matrices and confidence intervals • contains algorithms to do matrix reordering • flexible appearance details settings
	lattice	132	0	• high-level visualization system • emphasis on multivariate data • efficiently copes with nonstandard requirements

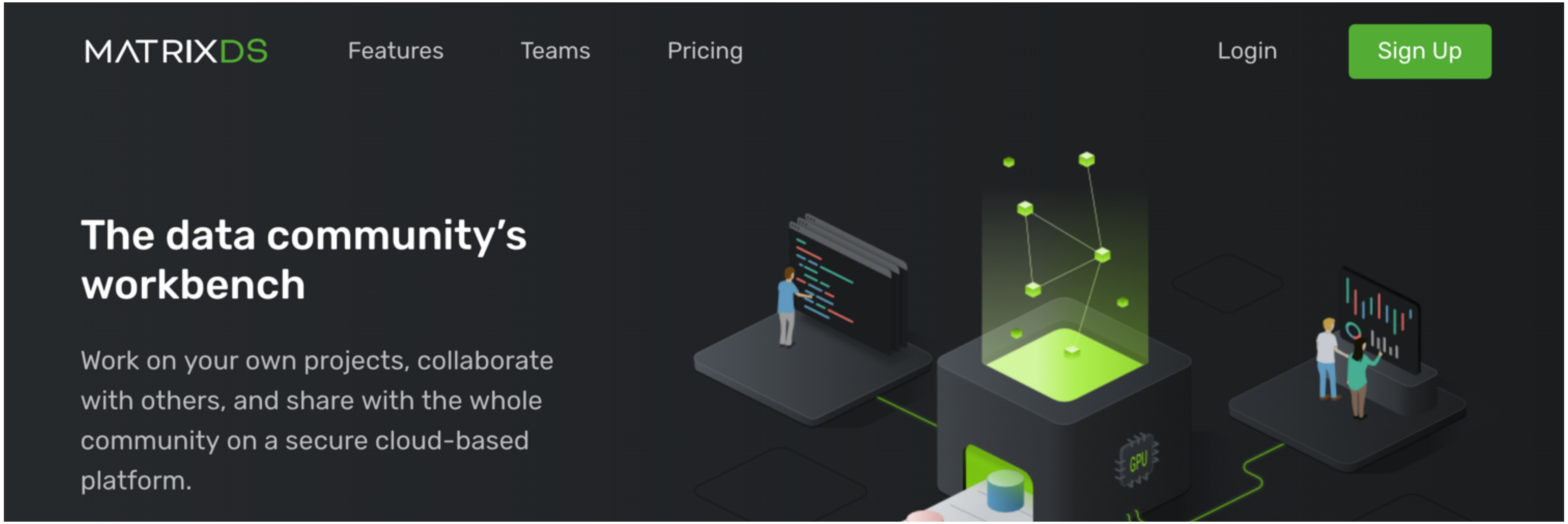
MATLAB

- MATLAB: www.mathworks.com/products/matlab.html



MatrixDS

- Cloud-based workbench that integrates data project needs in one location: www.matrixds.com



Online Resources

Website:

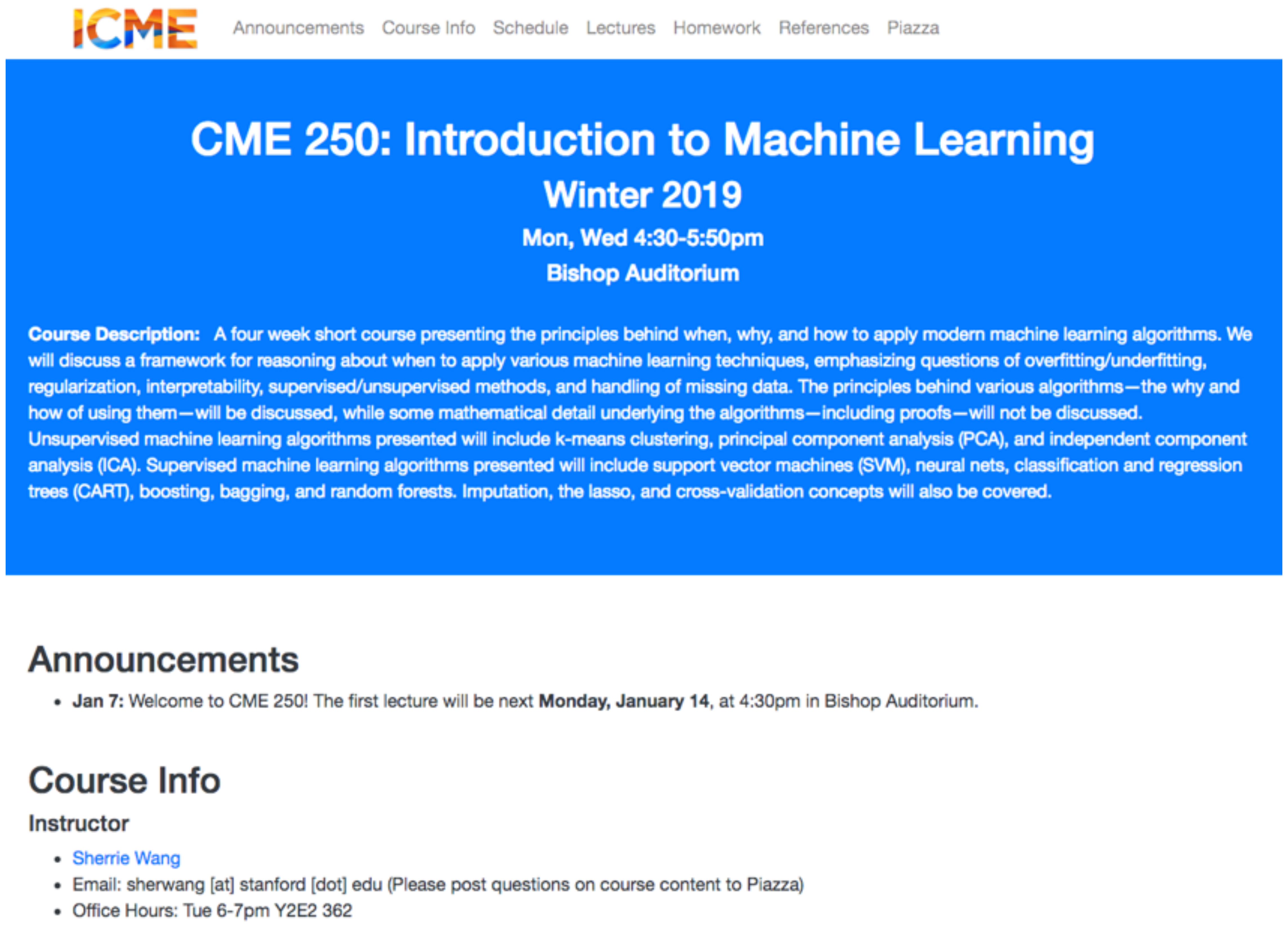
https://cme250.stanford.edu

Piazza:

https://piazza.com/stanford/winter2019/cme250

Gradescope:

https://www.gradescope.com/courses/33828



ICME Announcements Course Info Schedule Lectures Homework References Piazza

CME 250: Introduction to Machine Learning

Winter 2019

Mon, Wed 4:30-5:50pm

Bishop Auditorium

Course Description: A four week short course presenting the principles behind when, why, and how to apply modern machine learning algorithms. We will discuss a framework for reasoning about when to apply various machine learning techniques, emphasizing questions of overfitting/underfitting, regularization, interpretability, supervised/unsupervised methods, and handling of missing data. The principles behind various algorithms—the why and how of using them—will be discussed, while some mathematical detail underlying the algorithms—including proofs—will not be discussed. Unsupervised machine learning algorithms presented will include k-means clustering, principal component analysis (PCA), and independent component analysis (ICA). Supervised machine learning algorithms presented will include support vector machines (SVM), neural nets, classification and regression trees (CART), boosting, bagging, and random forests. Imputation, the lasso, and cross-validation concepts will also be covered.

Announcements

- **Jan 7:** Welcome to CME 250! The first lecture will be next **Monday, January 14**, at 4:30pm in Bishop Auditorium.

Course Info

Instructor

- Sherrie Wang
- Email: sherwang [at] stanford [dot] edu (Please post questions on course content to Piazza)
- Office Hours: Tue 6-7pm Y2E2 362

Machine Learning Overview

Problem Set-up

X: input variables (predictors, independent variables, features)

Y: output variable (response, dependent variable)

Machine learning: estimate a function f that describes the relationship between predictors and response

$$\mathrm{Y} = f(\mathrm{X}) + \varepsilon$$

How to find f ?

Training dataset containing n samples $\mathrm{i} = 1, 2, \ldots, n$

Input, output pairs: $(\mathrm{X}^{(1)}, \mathrm{Y}^{(1)}), (\mathrm{X}^{(2)}, \mathrm{Y}^{(2)}), \ldots (\mathrm{X}^{(n)}, \mathrm{Y}^{(n)})$

We will use these observations to build our model f.

$$\mathrm{Y} = f(\mathrm{X}) + \varepsilon$$

Algorithms vary in how they use this data and in the assumptions they place on f.

Prediction vs. Inference

Prediction:

- Predict response $\mathbf{Y}$ given inputs $\mathbf{X}$.

Inference:

- Understand the relationship between $\mathbf{Y}$ and individual predictors $\mathbf{X}_i$.
Do not want a black-box model.

Choosing an ML Algorithm

Which algorithm you use for a task will depend on:

- The type of problem you are trying to solve
- The type of data you have access to

Note that it's possible to have data ill-suited for the problem of interest. In this case, algorithms won't save you.

Machine Learning
Unsupervised Learning
Dimension Reduction
Big Data Visualisation
Structure Discovery
Feature Extraction
boats
mugs
Generative Networks
2D to 3D Modelling
Music Generation
Pattern Modelling
Image Generation
Clustering
Recommender Systems
Targeted Marketing
Customer Segmentation
Semi-Supervised Learning
Text Classification
Lane Finding
Supervised Learning
Classification
Fraud Detection
Gender Detection
Image Classification
Classification
Classification + Localization
Object Detection
Instance Segmentation
Medical Diagnosis
Regression
Weather Forecasting
Advertising Reach Prediction
Sales Growth Prediction
Estimating Life Expectancy
Market Forecasting
Reinforcement Learning
Real-time Decisions
Game AI
Robot Navigation
Learning Tasks
Skill Acquisition
By Benjamin Timmermans. http://btimmermans.com

Two Categories of Learning

Supervised learning:

Unsupervised learning:

Two Categories of Learning

Supervised learning:

- Builds a statistical model to predict an output from inputs

Unsupervised learning:

Two Categories of Learning

Supervised learning:

- Builds a statistical model to predict an output from inputs

Unsupervised learning:

- Learns structure from data without supervising output

Supervised Learning

Training data contains both the input variables and the associated response

➡ Mathematically, $\mathbf{X}^{(i)}$ and associated $\mathbf{Y}^{(i)}$ are available to learning algorithm for training

Goal: *generalize* to new data

Unsupervised Learning

Training data contains measurements for each observation, but no associated response of interest

➡ Mathematically, $\mathbf{X}^{(i)}$ are available but $\mathbf{Y}^{(i)}$ are not

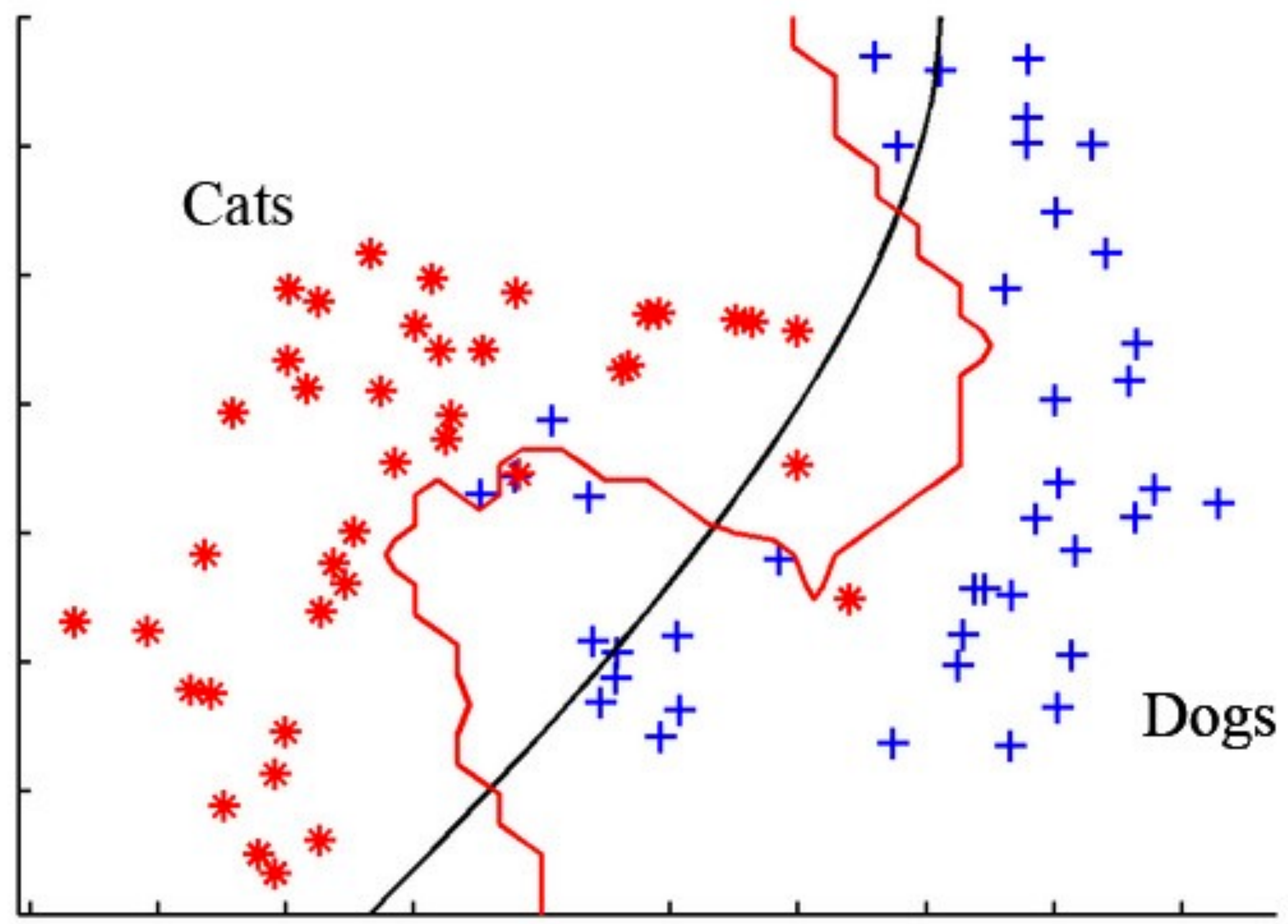
Goal: *understand relationships* between variables or among observations

Two Types of Supervised Learning

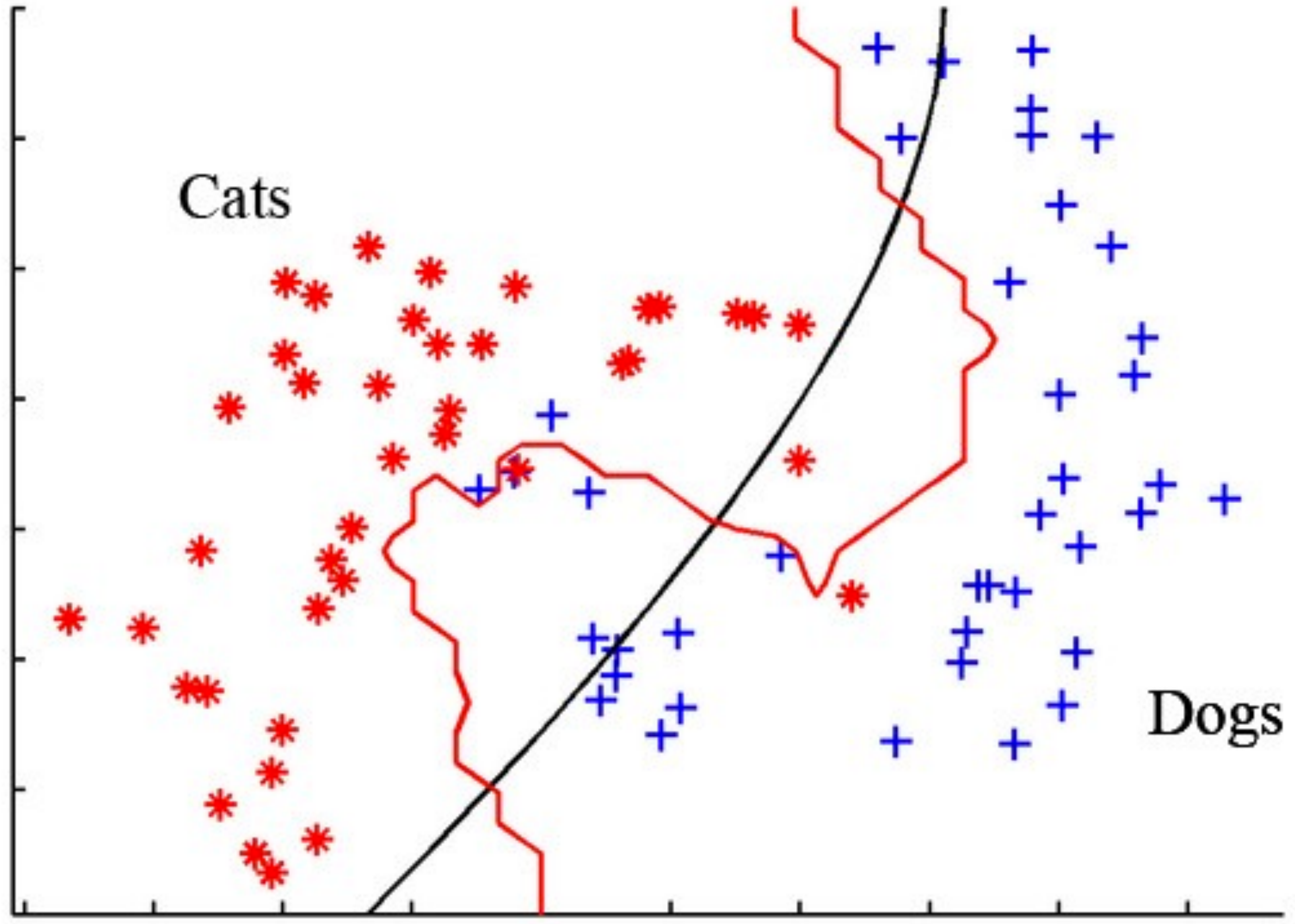
Two Types of Supervised Learning

Classification

- Output is qualitative (categorical)
- E.g. predict whether a credit card transaction is fraudulent

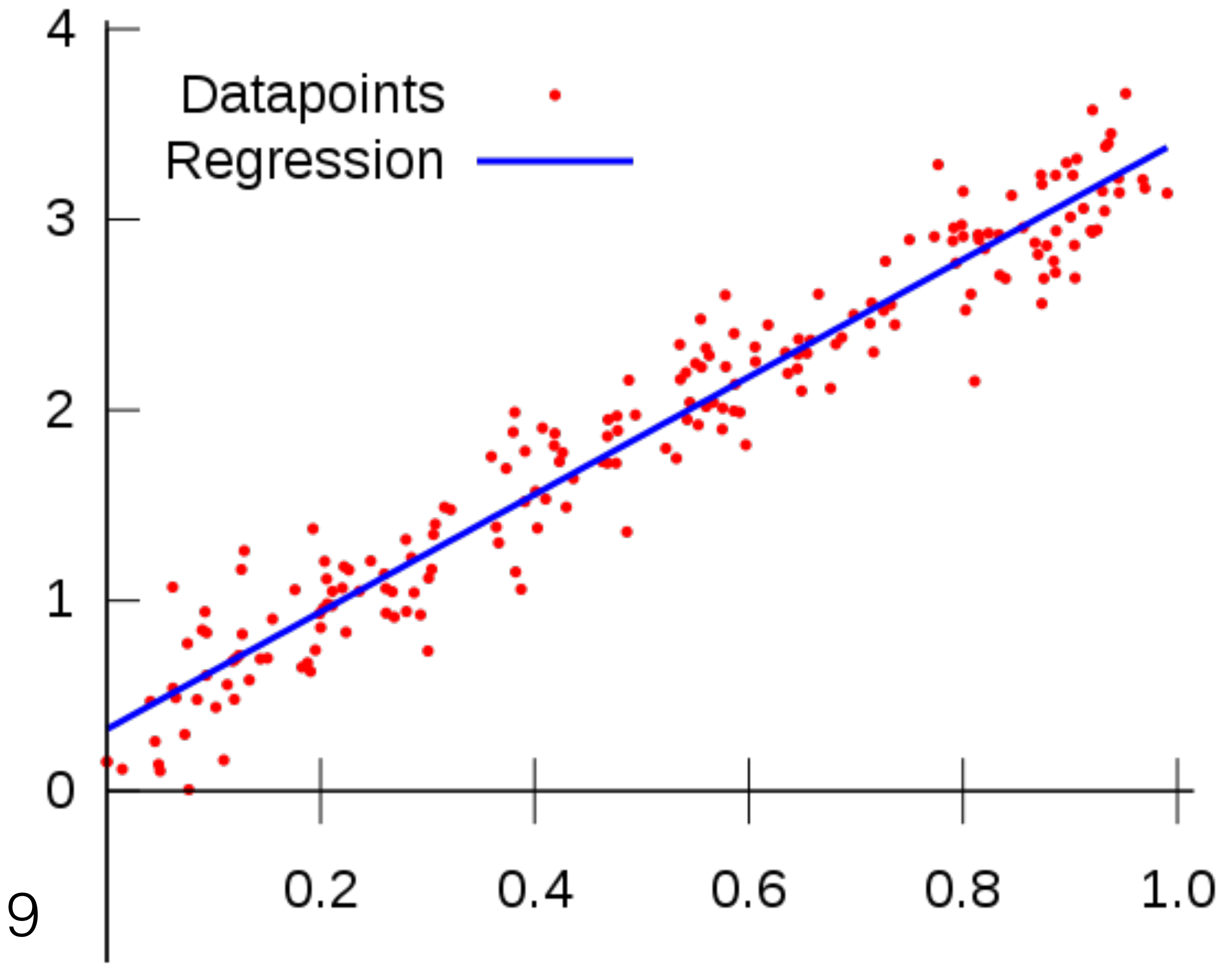


Two Types of Supervised Learning



Classification

- Output is qualitative (categorical)
- E.g. predict whether a credit card transaction is fraudulent



Regression

- Output is quantitative (continuous or ordered)
- E.g. predict the value of a stock tomorrow

Classification and Regression

Classification can often be formulated as a regression problem

- For a two-class (binary) problem: “What is the probability that observation belongs to class 1?” Probability lies in [0,1]

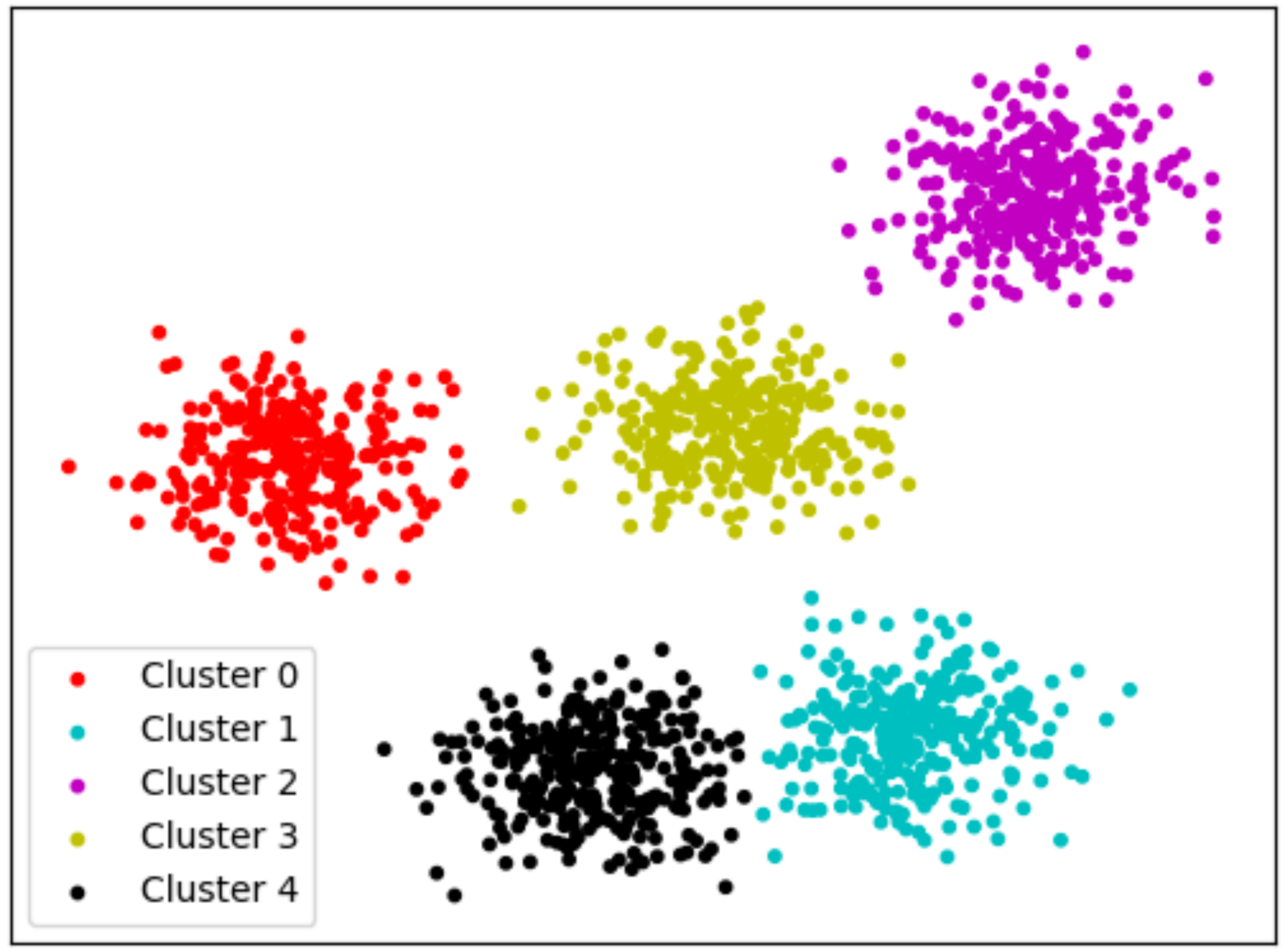
- Some methods work well on both types of problems (e.g. neural networks)

Two Types of Unsupervised Learning

Two Types of Unsupervised Learning

Clustering

- Partition data into subsets that share common characteristics



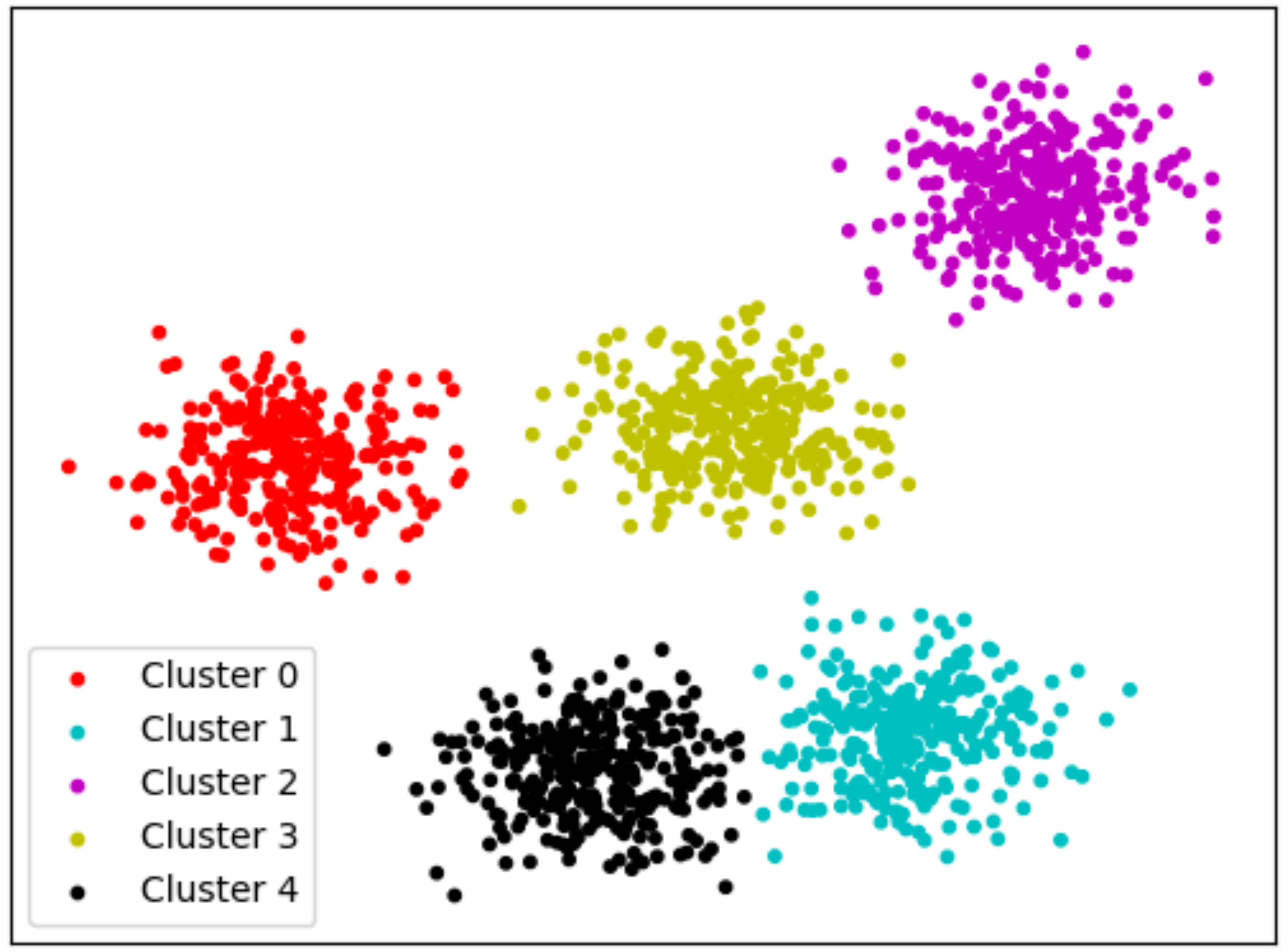
Two Types of Unsupervised Learning

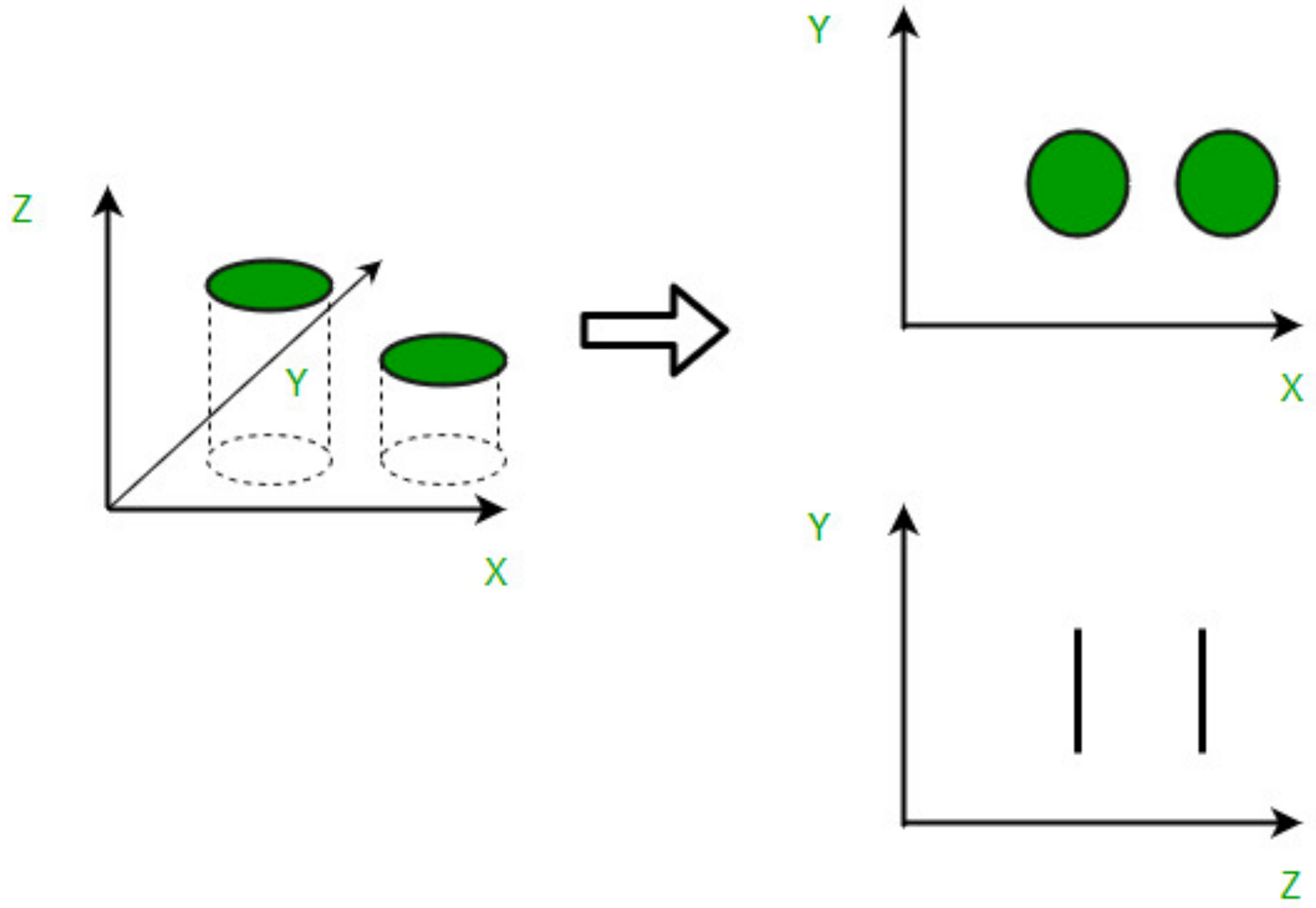
Clustering

- Partition data into subsets that share common characteristics

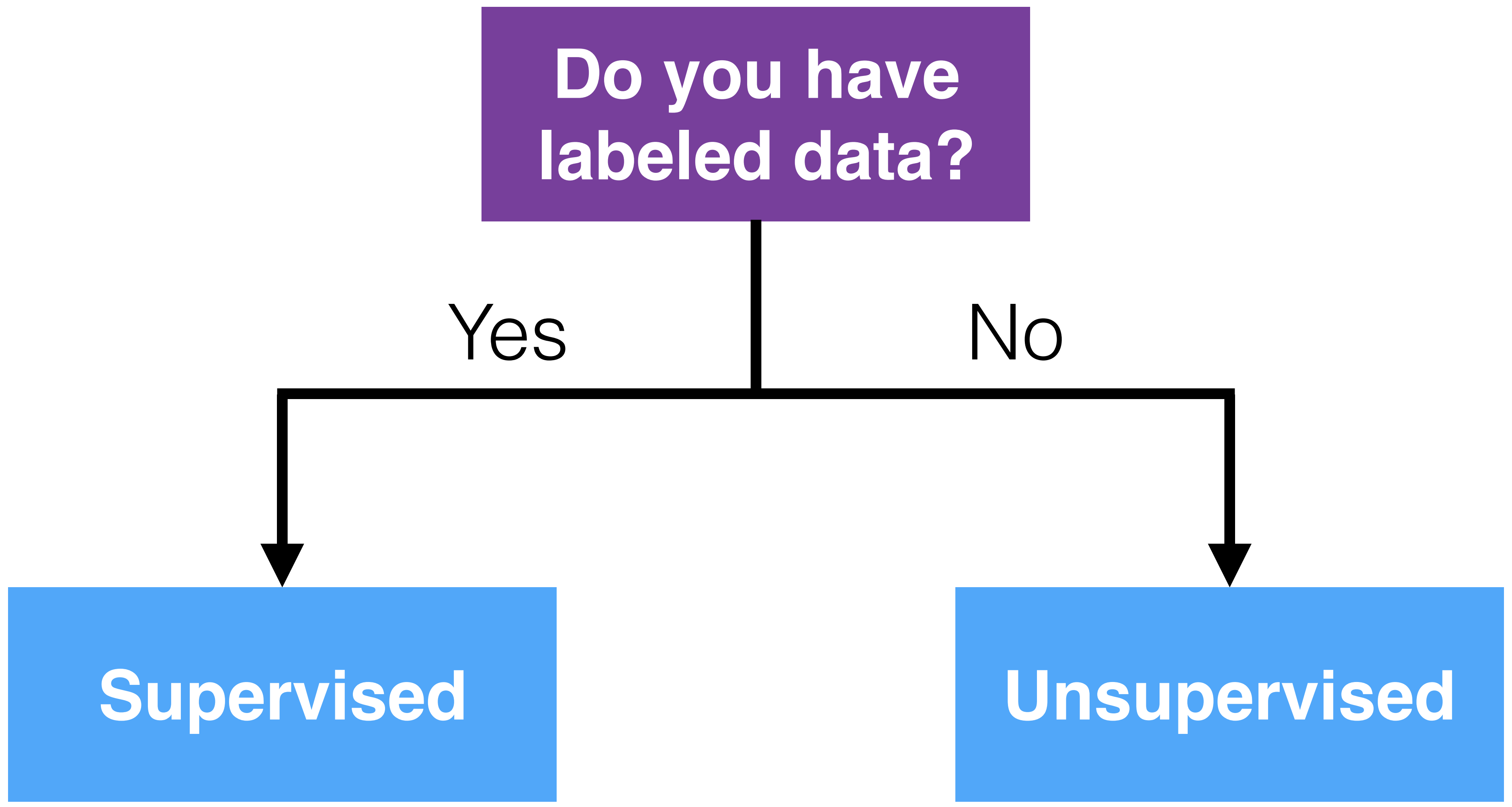
Dimensionality reduction

- Create new features from original inputs that retain important information

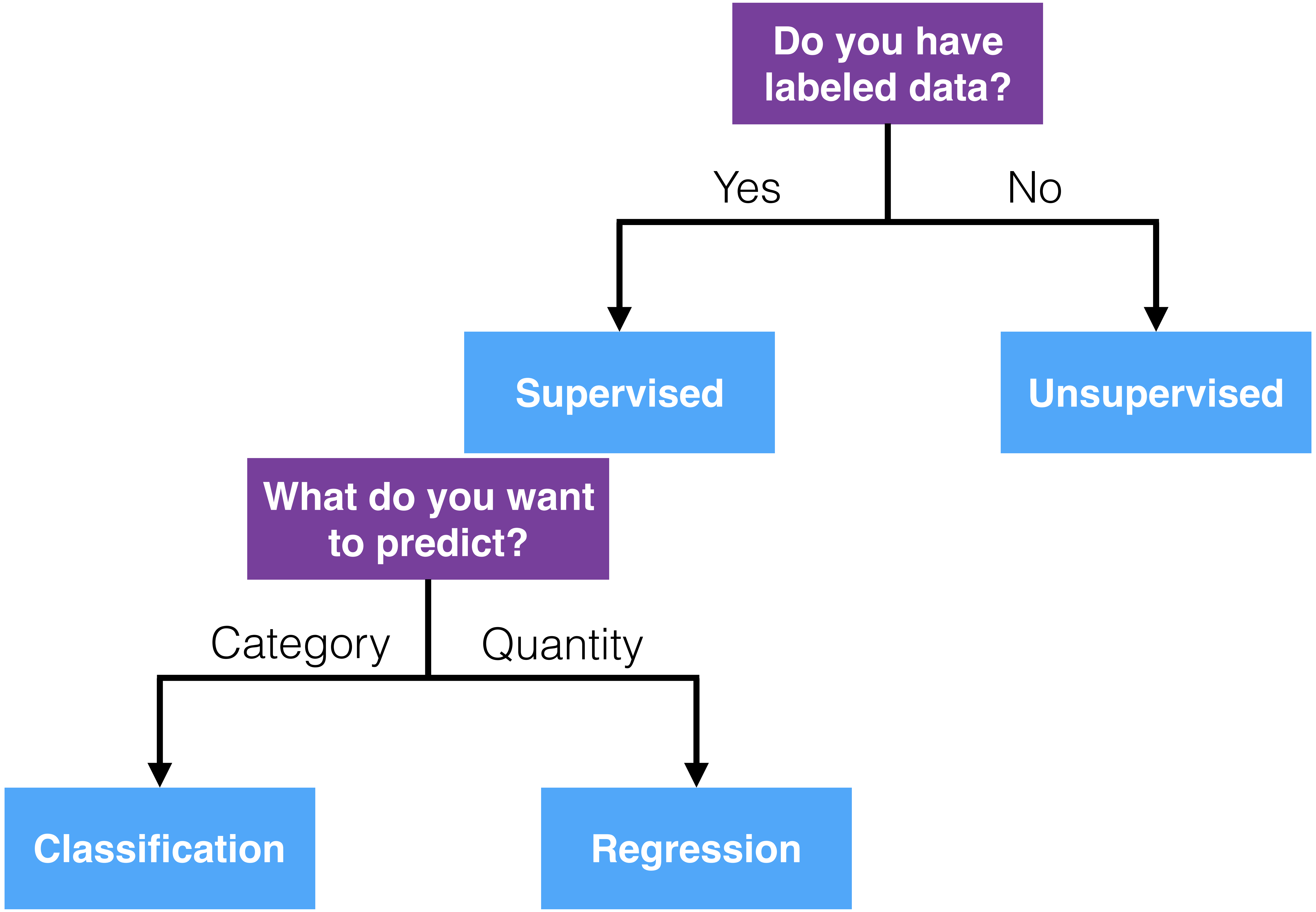




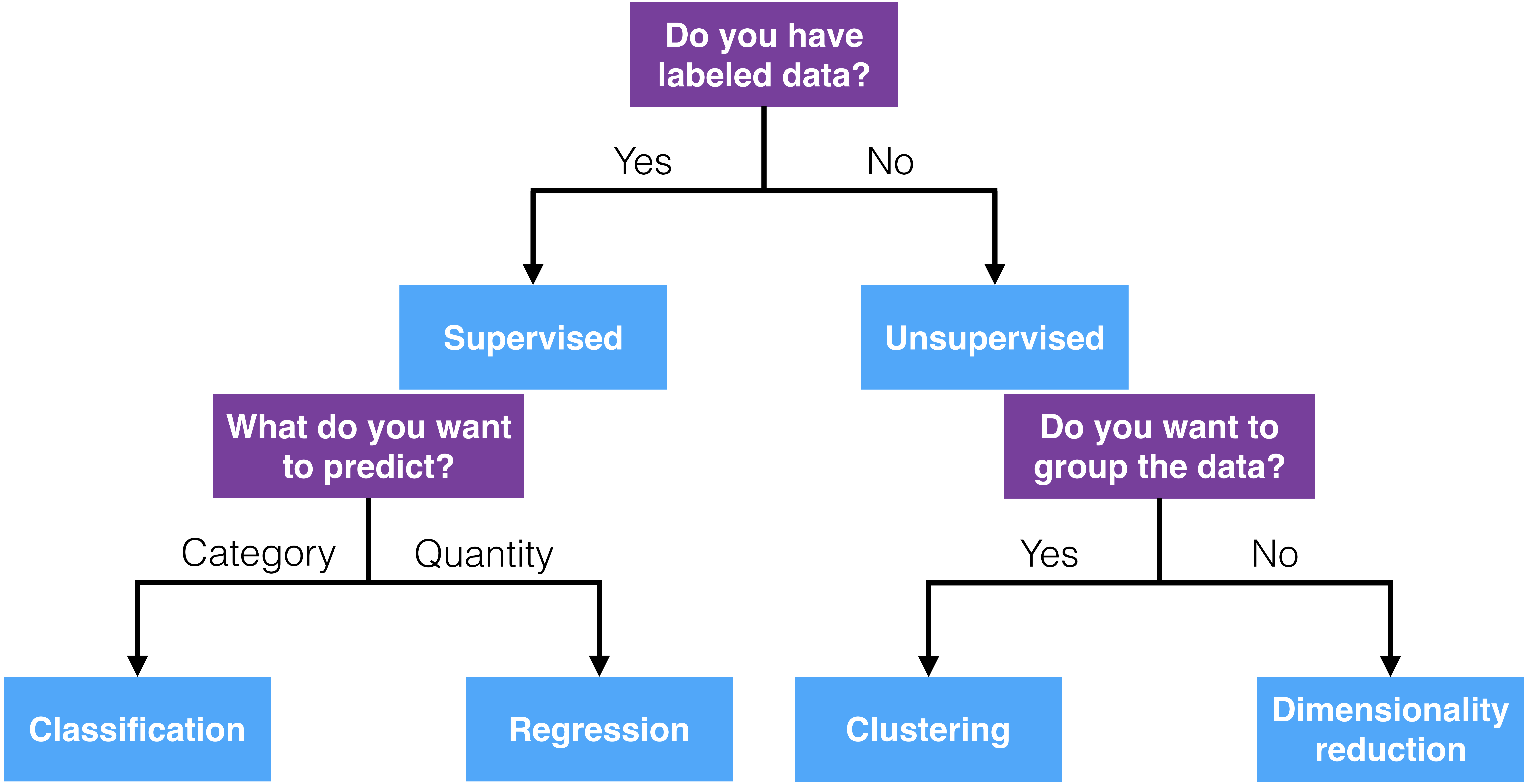
Choosing an ML Algorithm

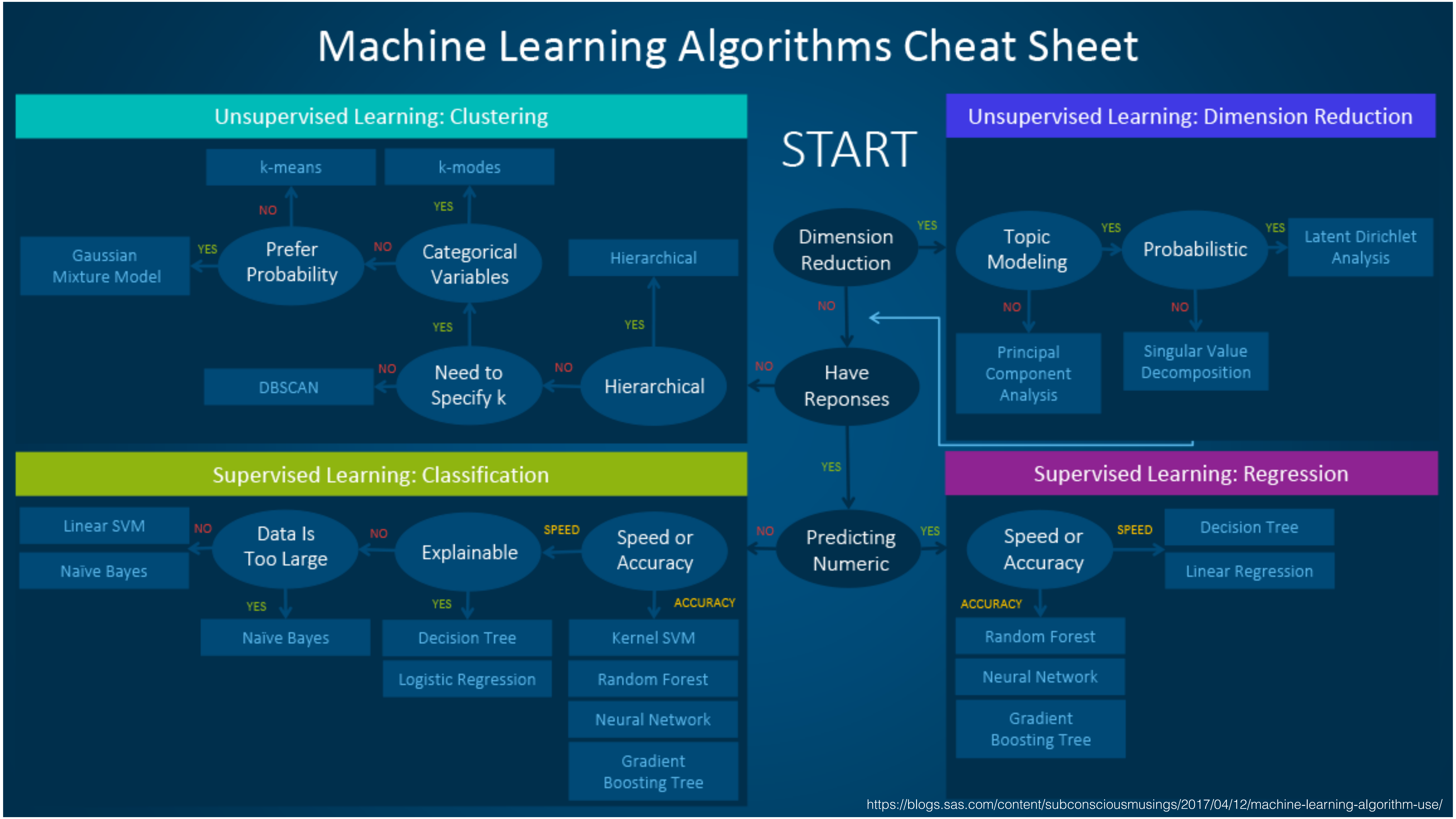


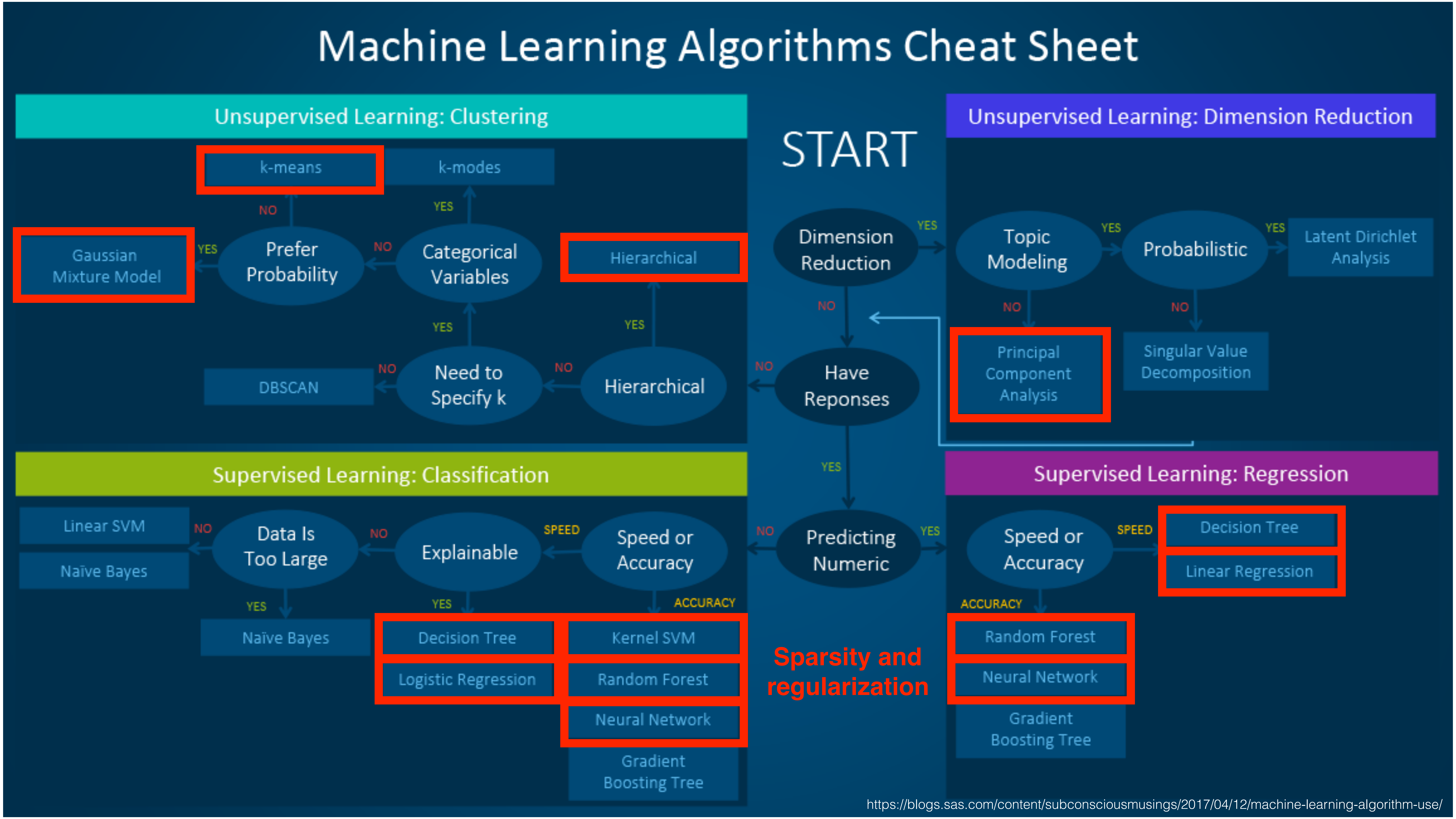
Choosing an ML Algorithm

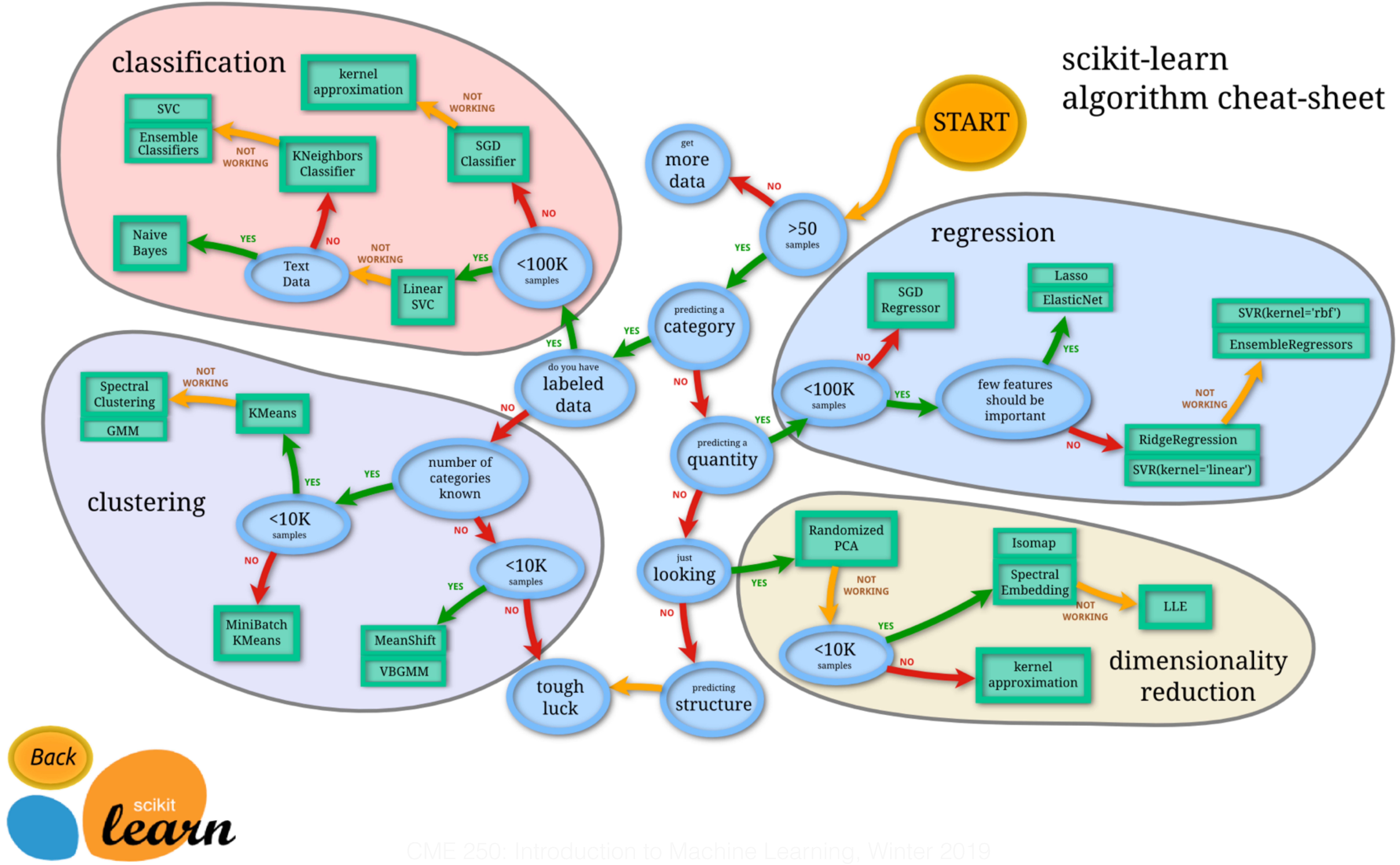


Choosing an ML Algorithm




Machine Learning Algorithms Cheat Sheet
START
Unsupervised Learning: Clustering
Unsupervised Learning: Dimension Reduction
Supervised Learning: Classification
Supervised Learning: Regression
Dimension Reduction
YES
NO
Topic Modeling
YES
NO
Probabilistic
YES
NO
Latent Dirichlet Analysis
Principal Component Analysis
Singular Value Decomposition
Have Reponses
NO
YES
Predicting Numeric
NO
YES
k-means
k-modes
Gaussian Mixture Model
Prefer Probability
Categorical Variables
Hierarchical
Need to Specify k
Hierarchical
DBSCAN
Linear SVM
Naïve Bayes
Data Is Too Large
Explainable
Speed or Accuracy
SPEED
ACCURACY
Naïve Bayes
Decision Tree
Logistic Regression
Kernel SVM
Random Forest
Neural Network
Gradient Boosting Tree
Speed or Accuracy
SPEED
ACCURACY
Decision Tree
Linear Regression
Random Forest
Neural Network
Gradient Boosting Tree
https://blogs.sas.com/content/subconsciousmusings/2017/04/12/machine-learning-algorithm-use/

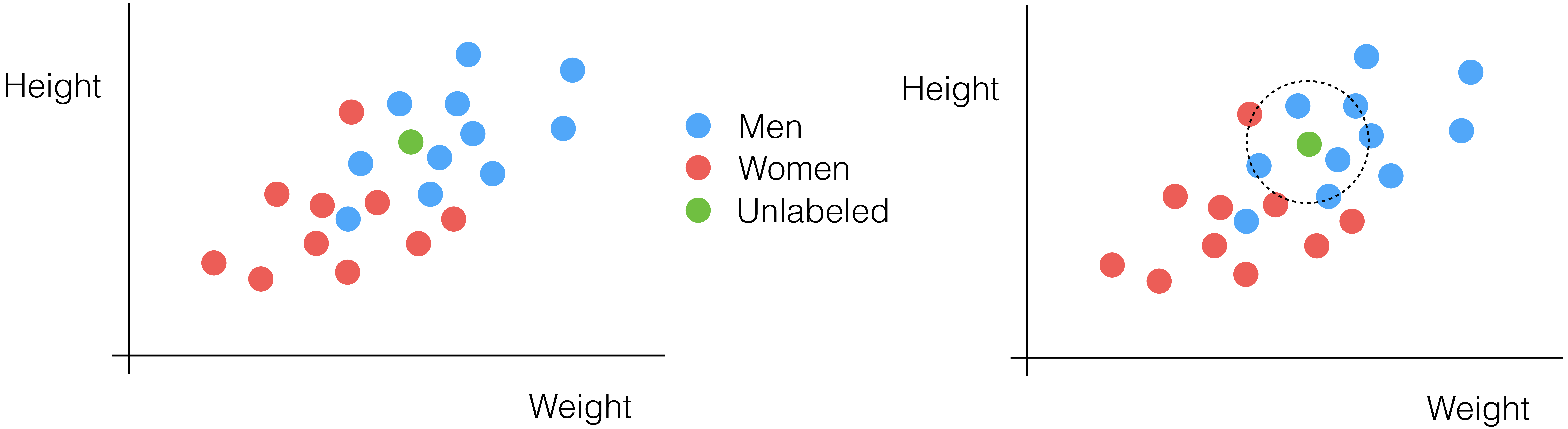

Machine Learning Algorithms Cheat Sheet
Unsupervised Learning: Clustering
k-means
k-modes
NO
YES
Gaussian Mixture Model
YES
Prefer Probability
NO
Categorical Variables
Hierarchical
YES
YES
DBSCAN
NO
Need to Specify k
NO
Hierarchical
START
Dimension Reduction
YES
NO
Have Reponses
NO
YES
Predicting Numeric
NO
YES
Unsupervised Learning: Dimension Reduction
Topic Modeling
YES
Probabilistic
YES
Latent Dirichlet Analysis
NO
NO
Principal Component Analysis
Singular Value Decomposition
Supervised Learning: Classification
Linear SVM
Naïve Bayes
NO
Data Is Too Large
NO
Explainable
SPEED
Speed or Accuracy
YES
YES
ACCURACY
Naïve Bayes
Decision Tree
Logistic Regression
Kernel SVM
Random Forest
Neural Network
Gradient Boosting Tree
Sparsity and regularization
Supervised Learning: Regression
Speed or Accuracy
SPEED
Decision Tree
Linear Regression
ACCURACY
Random Forest
Neural Network
Gradient Boosting Tree
https://blogs.sas.com/content/subconsciousmusings/2017/04/12/machine-learning-algorithm-use/


scikit-learn
algorithm cheat-sheet
START
>50 samples
get more data
NO
YES
predicting a category
do you have labeled data
<100K samples
Linear SVC
Text Data
Naive Bayes
KNeighbors Classifier
SVC
Ensemble Classifiers
SGD Classifier
kernel approximation
NOT WORKING
classification
clustering
number of categories known
<10K samples
KMeans
Spectral Clustering
GMM
MiniBatch KMeans
MeanShift
VBGMM
tough luck
predicting a quantity
regression
SGD Regressor
few features should be important
Lasso
ElasticNet
RidgeRegression
SVR(kernel='linear')
SVR(kernel='rbf')
EnsembleRegressors
just looking
predicting structure
Randomized PCA
Isomap
Spectral Embedding
LLE
dimensionality reduction
Back
scikit learn

Supervised Algorithm #1:
K-Nearest Neighbors

K-Nearest Neighbor (KNN) Classifier

A **classification** algorithm that labels observations based on “nearby” examples with known labels.



K-Nearest Neighbor (KNN) Classifier

Many classifiers build a model of

$$\Pr(Y \mid X)$$

KNN classifier predicts that the class for observation X is the class most common among its k nearest neighbors in the training set

$$\Pr(X \text{ belongs to class } Y) \approx (\#\ k \text{ nearest neighbors of } X \text{ in class } Y) \div k$$

K-Nearest Neighbor (KNN) Classifier

Suppose $k = 3$.

To classify **x** in this example, we find its 3 nearest neighbors.

Two of them are blue, and one is yellow.

Therefore a KNN classifier with $k = 3$ assigns **x** to the blue class.

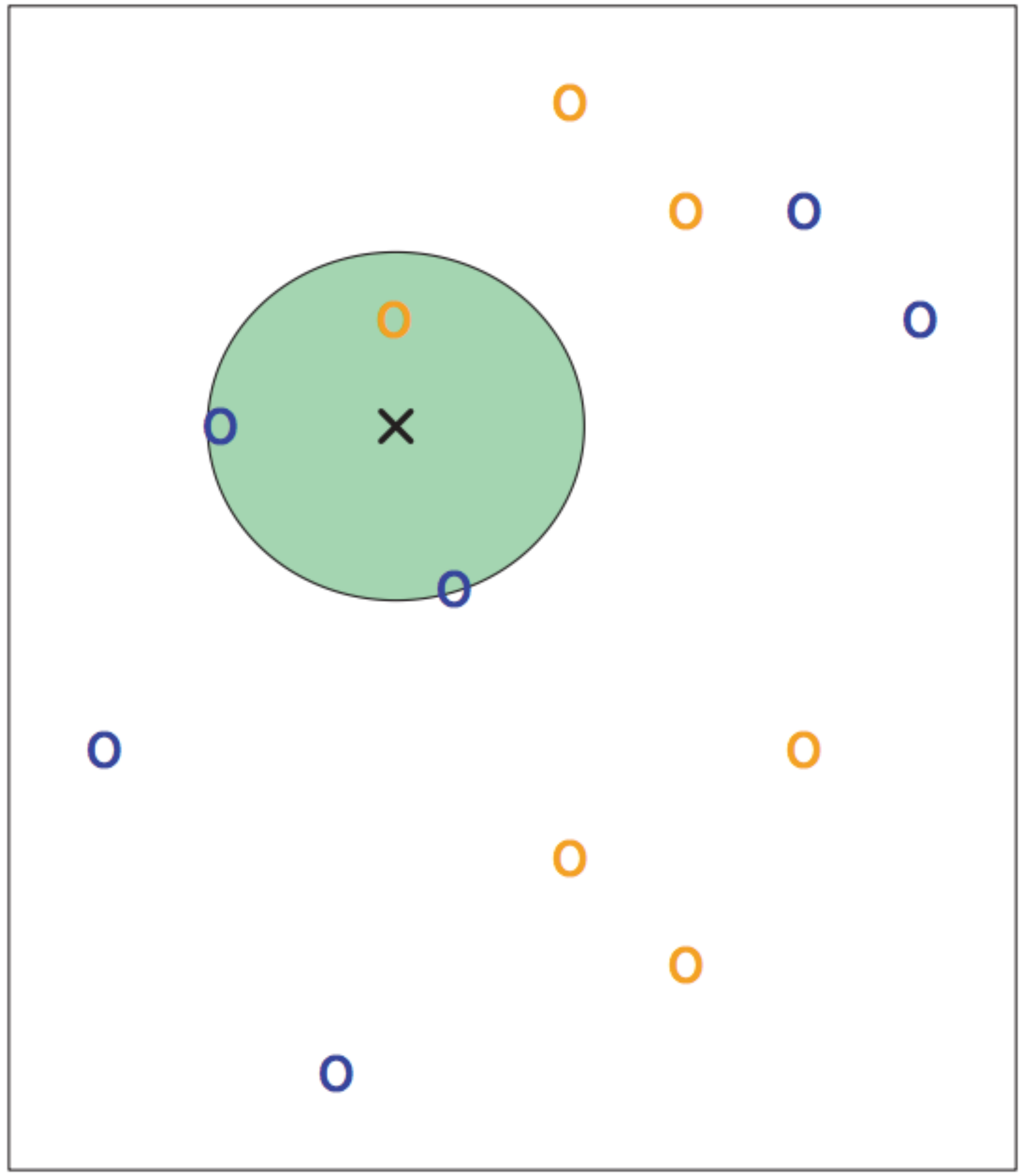


FIGURE 2.14, ISL (8th printing 2017)

K-Nearest Neighbor (KNN) Classifier

The classifier partitions the feature space into decision regions, each with a class label.

Decision boundaries separate decision regions.

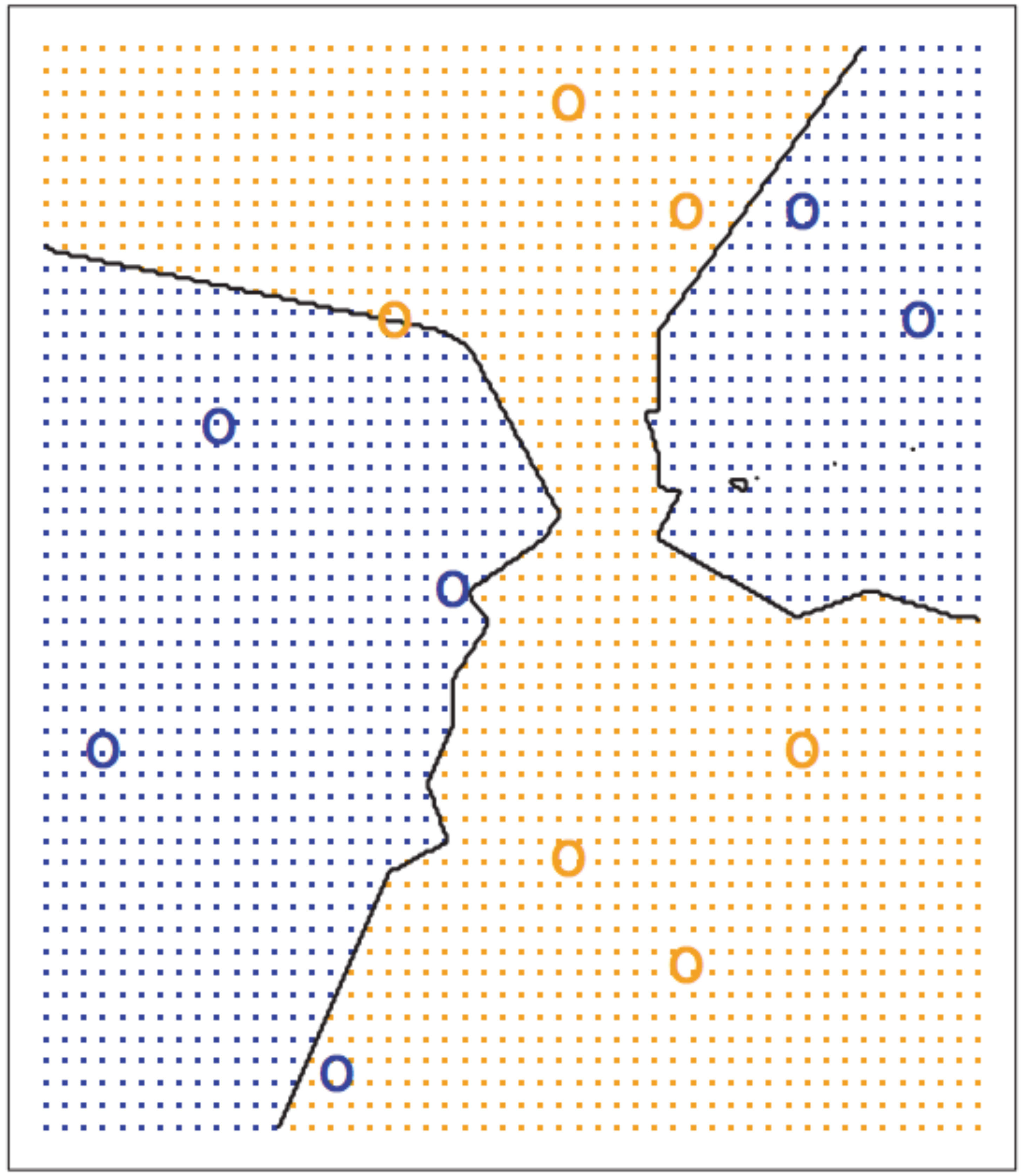


FIGURE 2.14, ISL (8th printing 2017)

How to choose k?

The decision regions depend on the value of k.

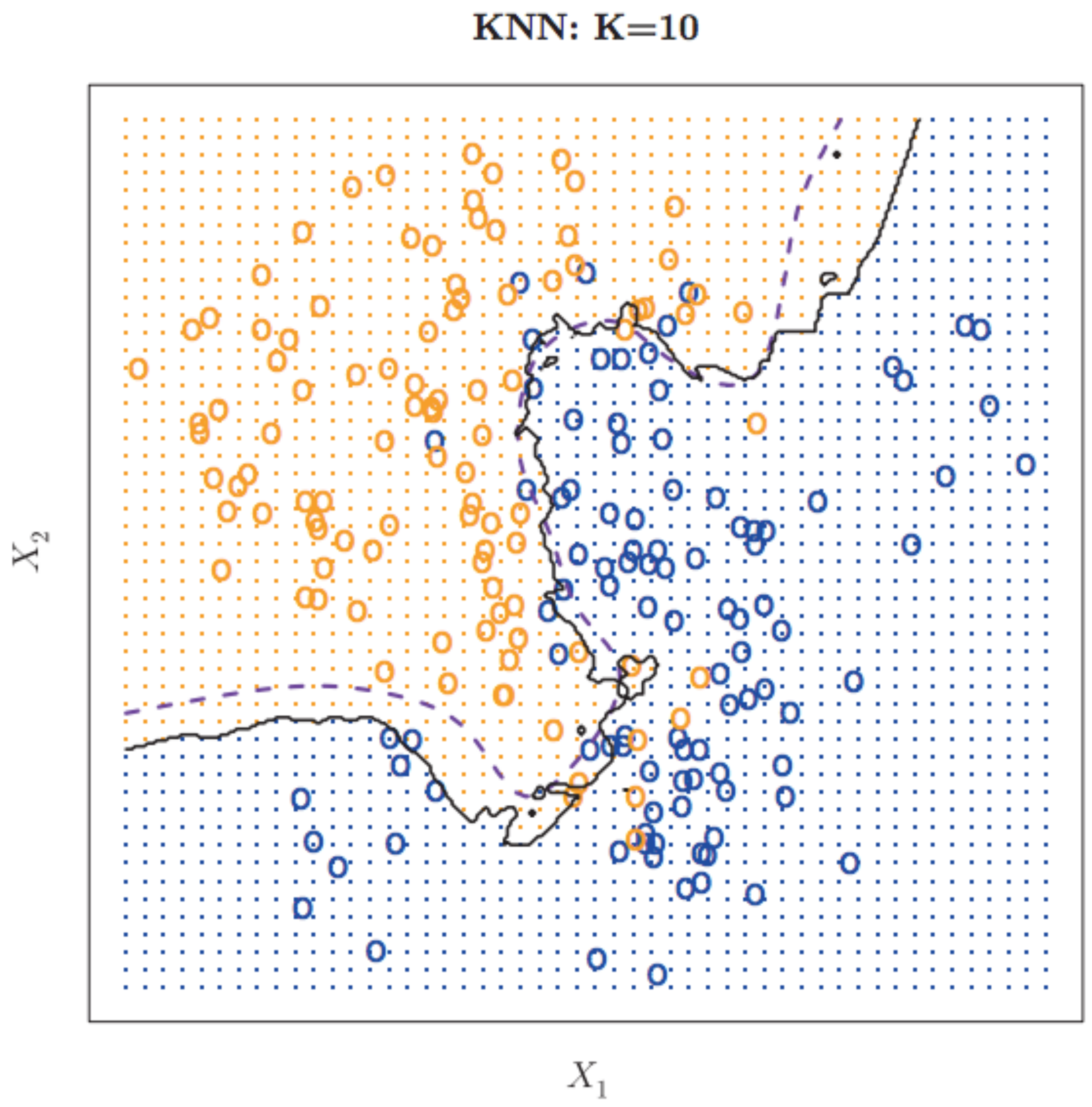


FIGURE 2.15, ISL (8th printing 2017)

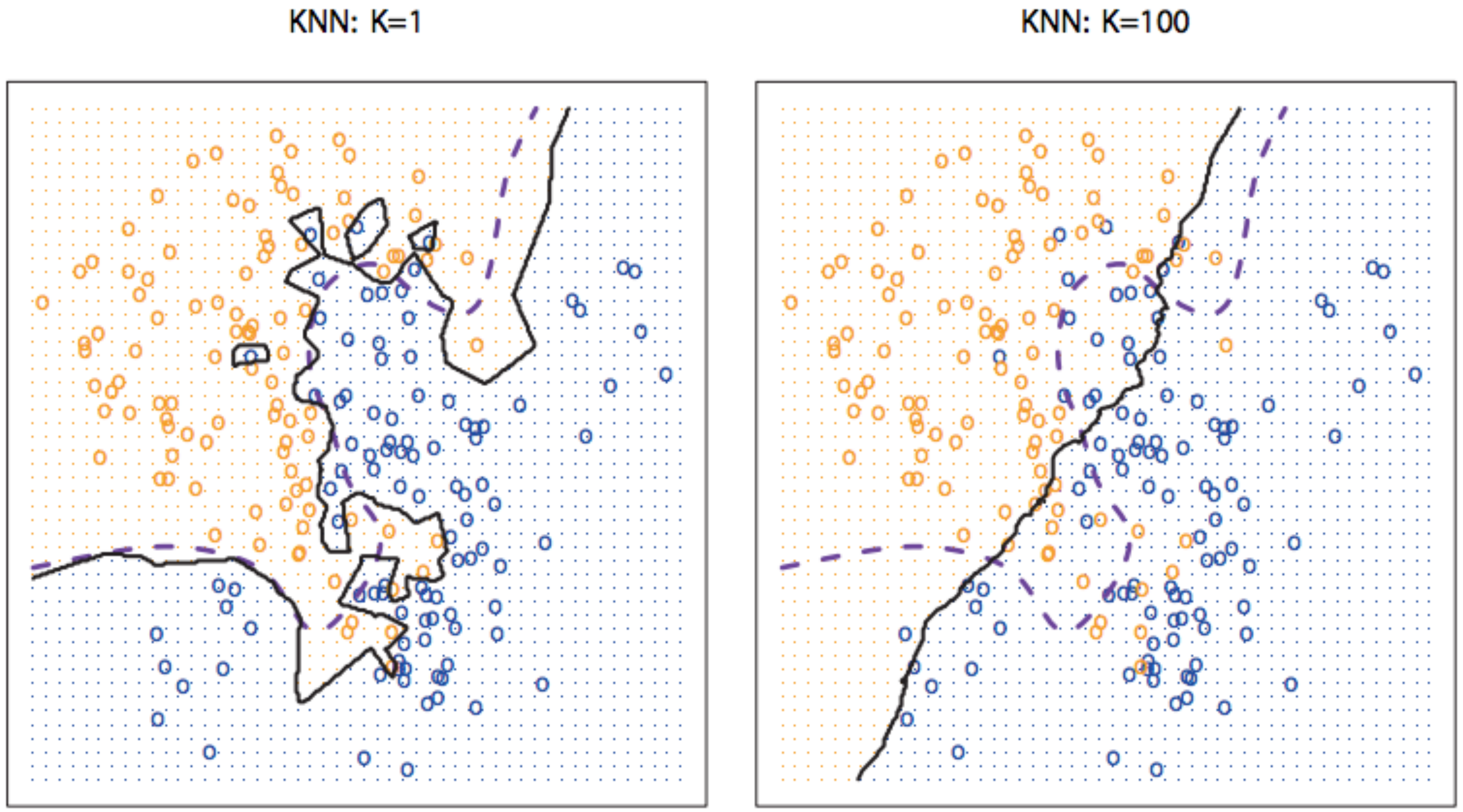


FIGURE 2.16, ISL (8th printing 2017)

How to choose k?

In machine learning terminology, k is a *hyperparameter*.

A *hyperparameter* is set before the learning process begins. We will learn how to tune hyperparameters in a later lecture.

How to choose k?

k small

- More flexible decision boundary, but more likely to *overfit*

k large

- Less flexible decision boundary, but less likely to *overfit*

Overfitting occurs when we learn random noise in training data rather than underlying trend

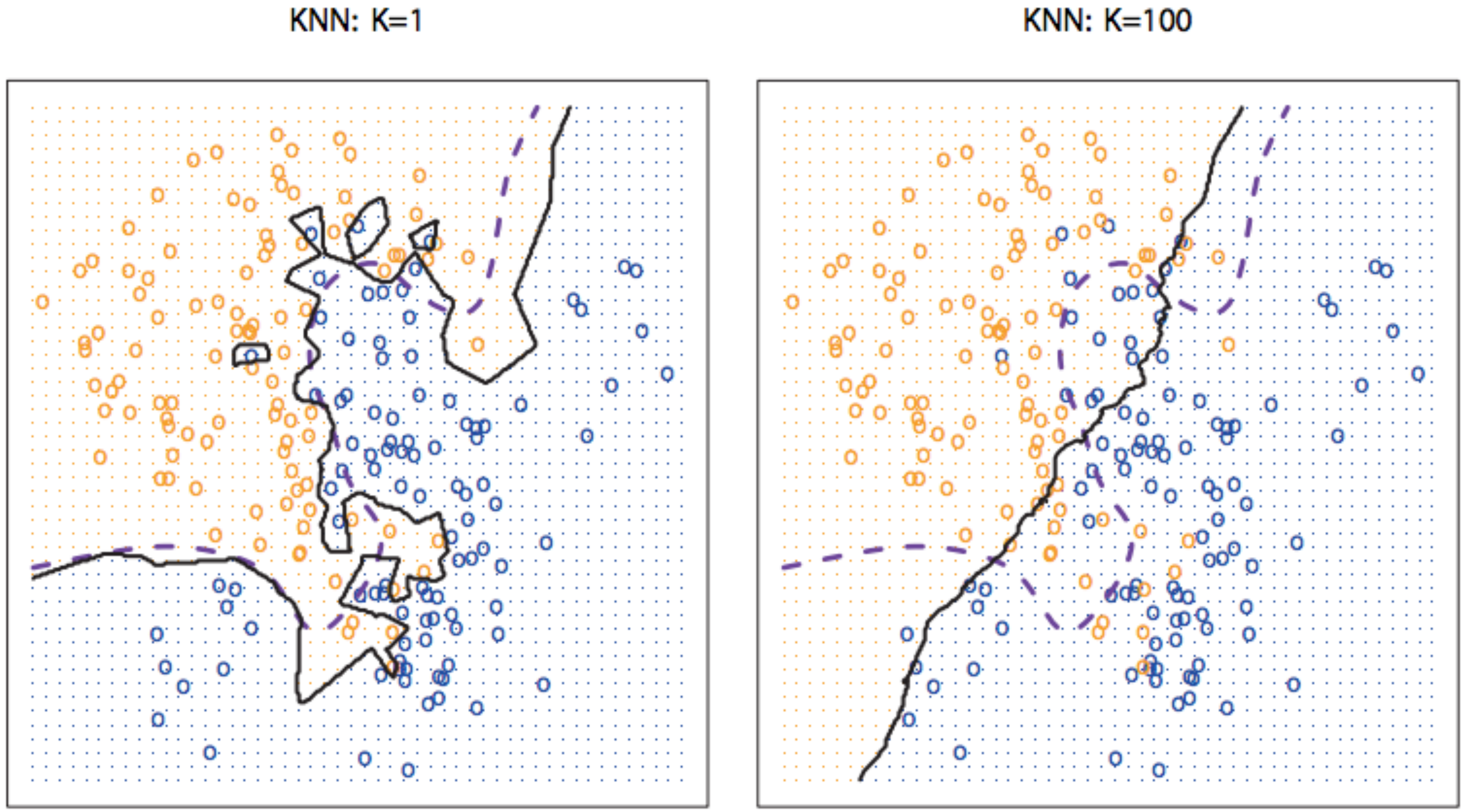


FIGURE 2.16, ISL (8th printing 2017)

How to choose k?

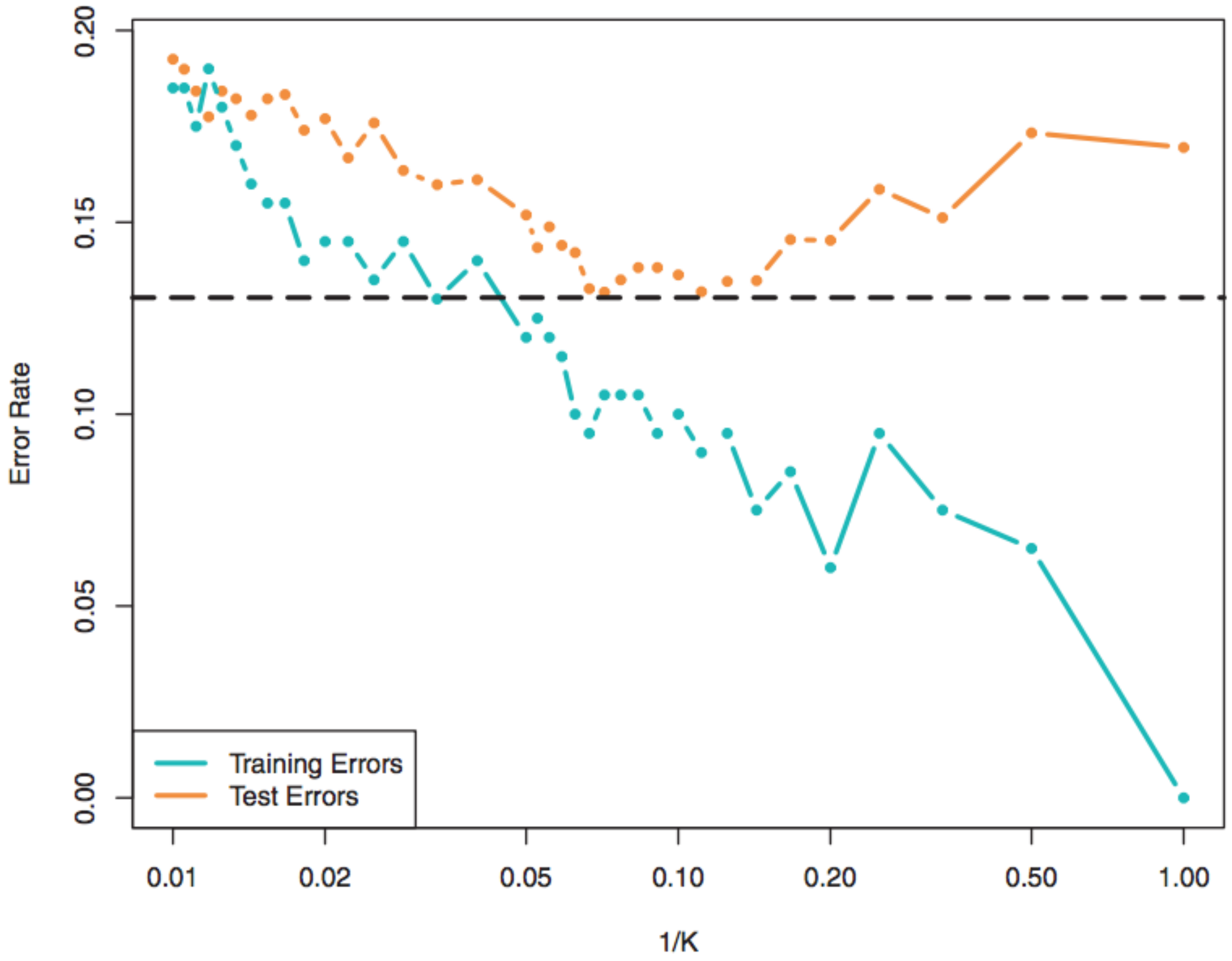


FIGURE 2.17, ISL (8th printing 2017)

K-Nearest Neighbor Summary

Advantages:

- Simple to implement
- Few tuning parameters (k, distance metric)
- Flexible, classes do not have to be linearly separable

Disadvantages:

- Computationally expensive ($O(nd)$ where d is input dimension)
- Sensitive to imbalanced datasets
- Sensitive to irrelevant inputs

“Best” Machine Learning Algorithm

Bad news: No algorithm is the best

- No machine learning algorithm will perform well on every task

Good news: All of them are the best

- Each machine learning algorithm will perform well on some task

“No free lunch” theorem

- Wolpert (1996): All algorithms perform equally when averaged over all possible problems

Trade-offs and Decisions

- Bias vs. variance
- Accuracy vs. interpretability
- Accuracy vs. scalability
- Domain-knowledge vs. data-driven
- More data vs. better algorithm
- Accuracy vs. fairness