CS464
Introduction to Machine Learning

CS464

Bilkent University

CS464

- Undergraduate-level introductory course aims at giving a broad overview of many concepts and algorithms in machine learning.

- The goal is to provide students with a deep understanding of the subject matter and skills to apply these concepts to real world problems.

Prerequisites

- Basic algorithms, data structures
- Basic probability and statistic
- Basic linear algebra
- Good programming skills

Data-rich Era

- The information **volume** grows exponentially
  - *Most data will never be seen by humans*
- Information **complexity** is also increasing greatly
  - *Most data cannot be comprehended by humans directly*
- In the computationally enabled, data-rich era **machine learning** is an essential component of the new scientific toolkit

Learning from Data

- Many applications require gaining insights from massive, noisy data sets
- **Commercial applications**
  - Speech recognition
    - https://youtu.be/2U1ImKsecB8
  - Consumer data (online advertising, viral marketing,…)
  - Health records (emergency rehospitalization,…)
- **Security related applications**
  - Spam filtering, fraud detections…
  - Surveillance …
- **Science**
  - Biology, astronomy, neuroscience, geology, …
  - Social science, economics

Spam or Ham

Osman Khan to Carlos show details Jan 7 (6 days ago) Reply

sounds good
+ok

Carlos Guestrin wrote:
Let's try to chat on Friday a little to coordinate and more on Sunday in person?

Carlos

Welcome to New Media Installation: Art that Learns

Carlos Guestrin to 10615-announce, Osman, Miche show details 3:15 PM (8 hours ago) Reply

Hi everyone,

Welcome to New Media Installation:Art that Learns

The class will start tomorrow.
Make sure you attend the first class, even if you are on the Wait List.
The classes are held in Doherty Hall C316, and will be Tue, Thu 01:30-4:20 PM.

By now, you should be subscribed to our course mailing list: 10615-announce@cs.cmu.edu.
You can contact the instructors by emailing: 10615-instructors@cs.cmu.edu

Ham

Natural _LoseWeight SuperFood Endorsed by Oprah Winfrey, Free Trial 1 bottle, pay only $5.95 for shipping mfw rlk Spam | x

Jaquelyn Halley to nherrlein, bcc: thehomey, bcc: ang show details 9:52 PM (1 hour ago) Reply

=== Natural WeightL0SS Solution ===

Vital Acai is a natural WeightL0SS product that Enables people to lose wieght and cleansing their bodies faster than most other products on the market.

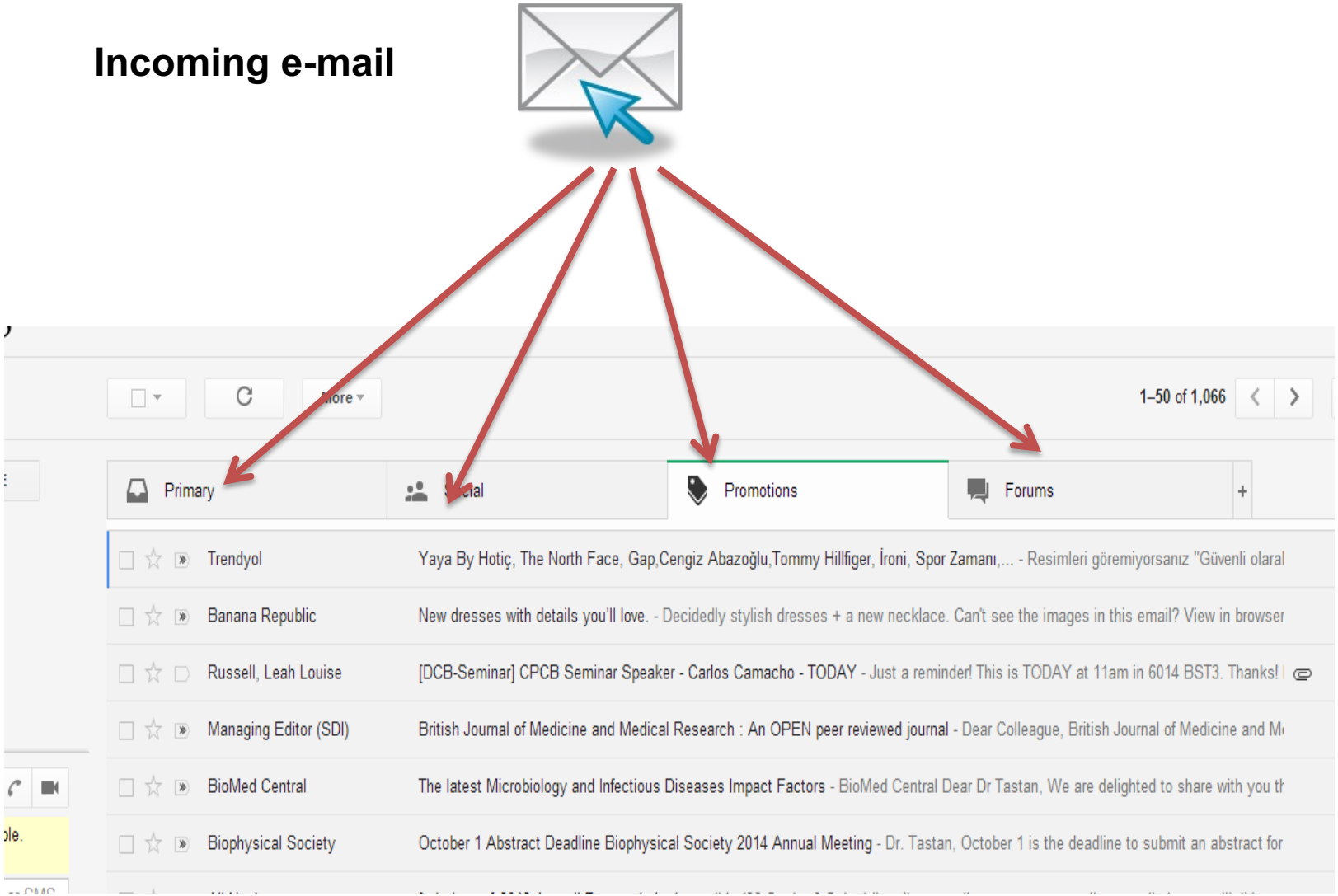
Here are some of the benefits of Vital Acai that You might not be aware of. These benefits have helped people who have been using Vital Acai daily to Achieve goals and reach new heights in there dieting that they never thought they could.

* Rapid WeightL0SS
* Increased metabolism - BurnFat & calories easily!
* Better Mood and Attitude
* More Self Confidence
* Cleanse and Detoxify Your Body
* Much More Energy
* BetterSexLife
* A Natural Colon Cleanse

Spam

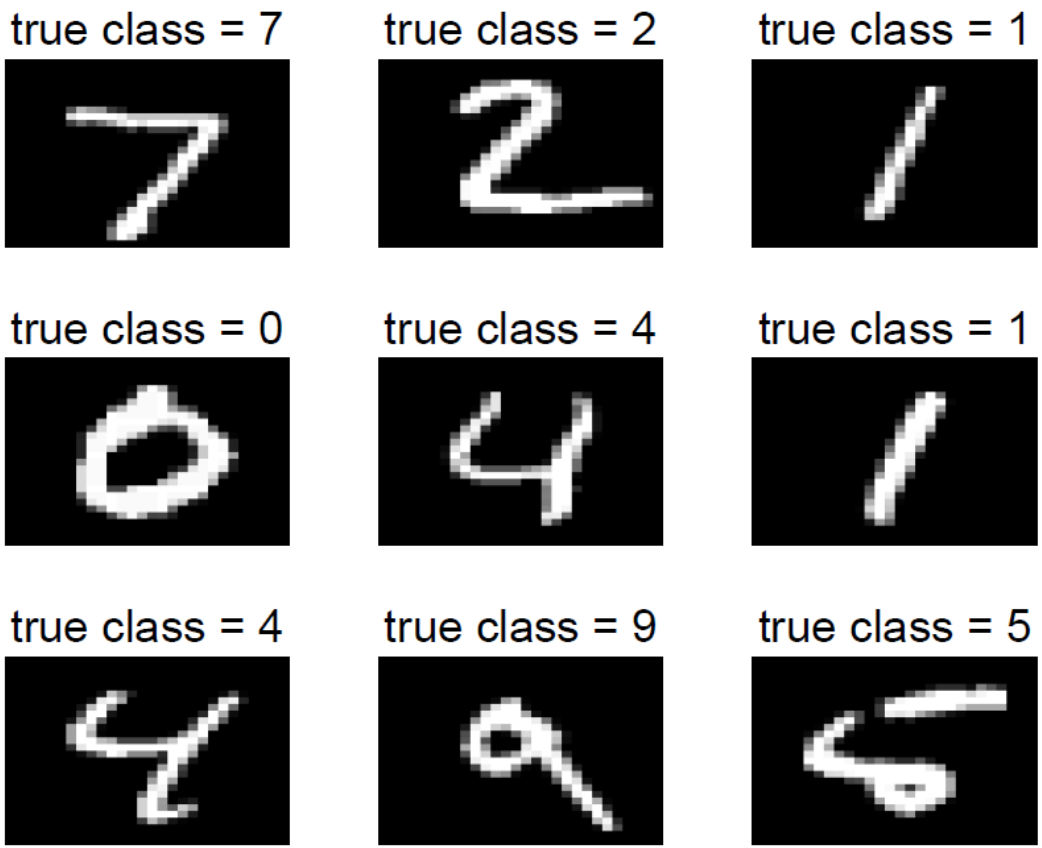
Text Classification

Incoming e-mail



Reading Zip Codes

- Classifying hand written digits from binary images



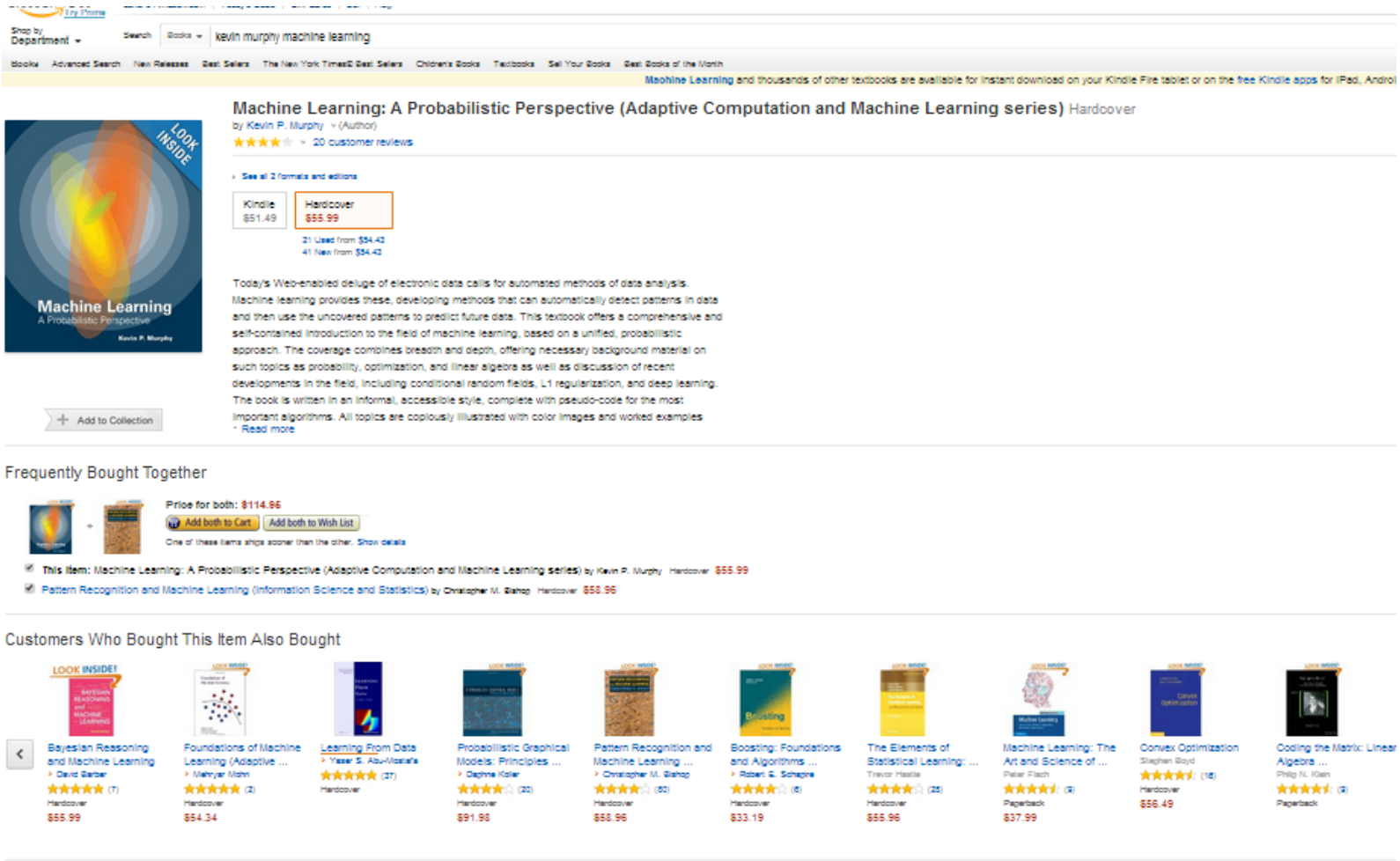
Recommendation Systems

users →

movies ↕

1		?	3	5	?
?	1				2
	4		4	5	?

Recommendation Systems



Recommendation Systems

Recognizing Faces

Figure taken from: https://github.com/deepinsight/insightface

Semantic Segmentation

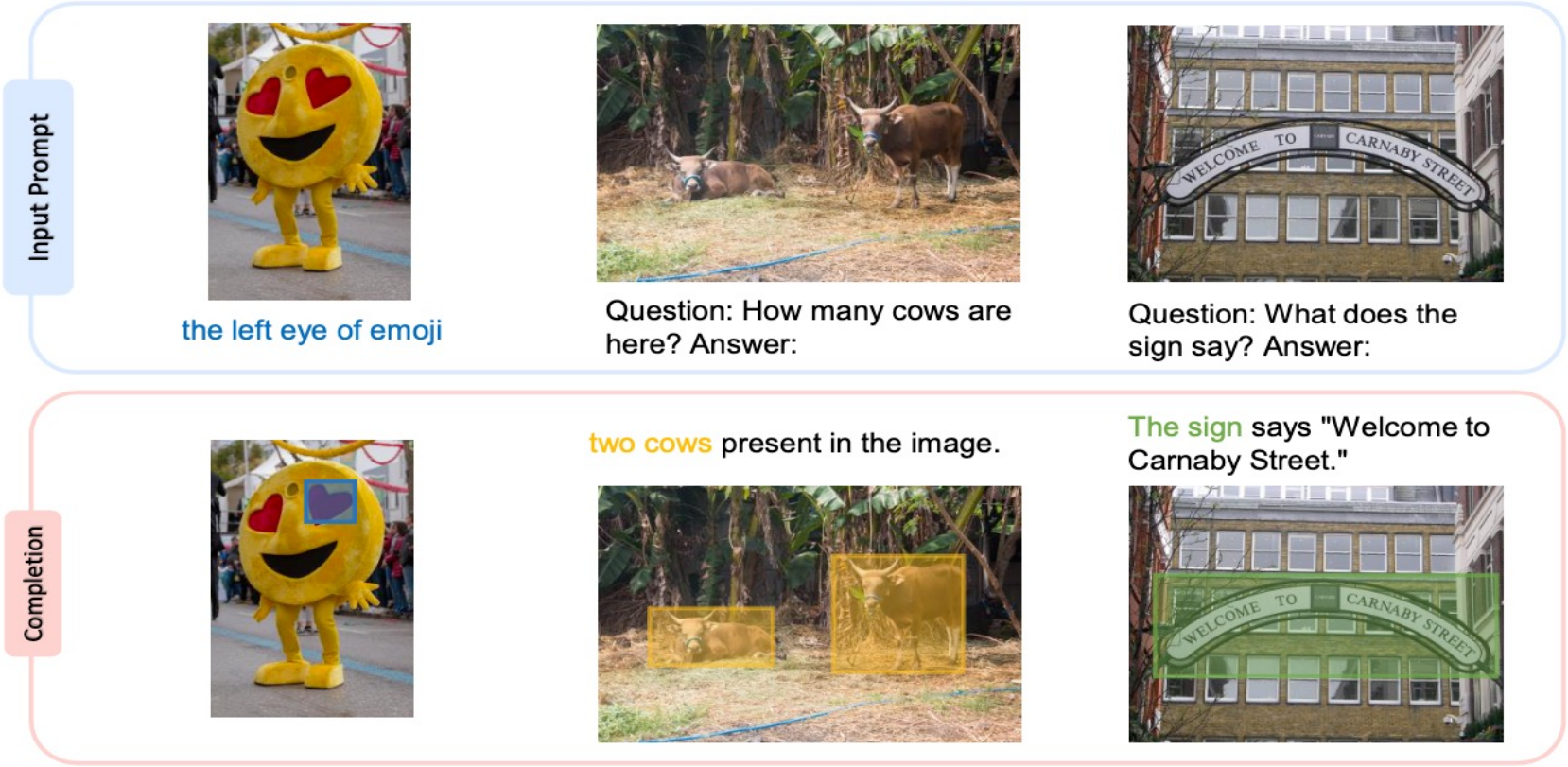
Figure taken from: https://github.com/facebookresearch/detectron2

Semantic Segmentation 2023

- **USER:** <IMAGE> What is the food that makes people feel spicy or hot in this image? Please output segmentation mask.
- **ASSISTANT:** <SEG>.

LISA: REASONING SEGMENTATION VIA LARGE LANGUAGE MODEL
Xin Lai1∗ Zhuotao Tian2∗ Yukang Chen1 Yanwei Li1 Yuhui Yuan3 Shu Liu2 Jiaya Jia
In submission August 2023

Semantic Segmentation 2023



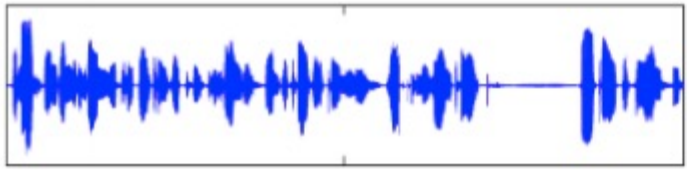
KOSMOS-2: Grounding Multimodal Large Language Models to the World
Zhiliang Peng∗, Wenhui Wang∗, Li Dong∗, Yaru Hao, Shaohan Huang, Shuming Ma, Furu Wei† Microsoft Research

In submission July 2023

Autopilot

Figure taken from: https://www.tesla.com/autopilot

Speech Recognition



https://youtu.be/D5VN56jQMWM

Generating Images

A highly detailed digital painting of a portal in a mystic forest with many beautiful trees. A person is standing in front of the portal.

eDiff-I: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers, November 2022

Editing Images

StyleRes: Transforming the Residuals for Real Image Editing with StyleGAN (CVPR 2023)
Hamza Pehlivan, Yusuf Dalva, Aysegul Dundar

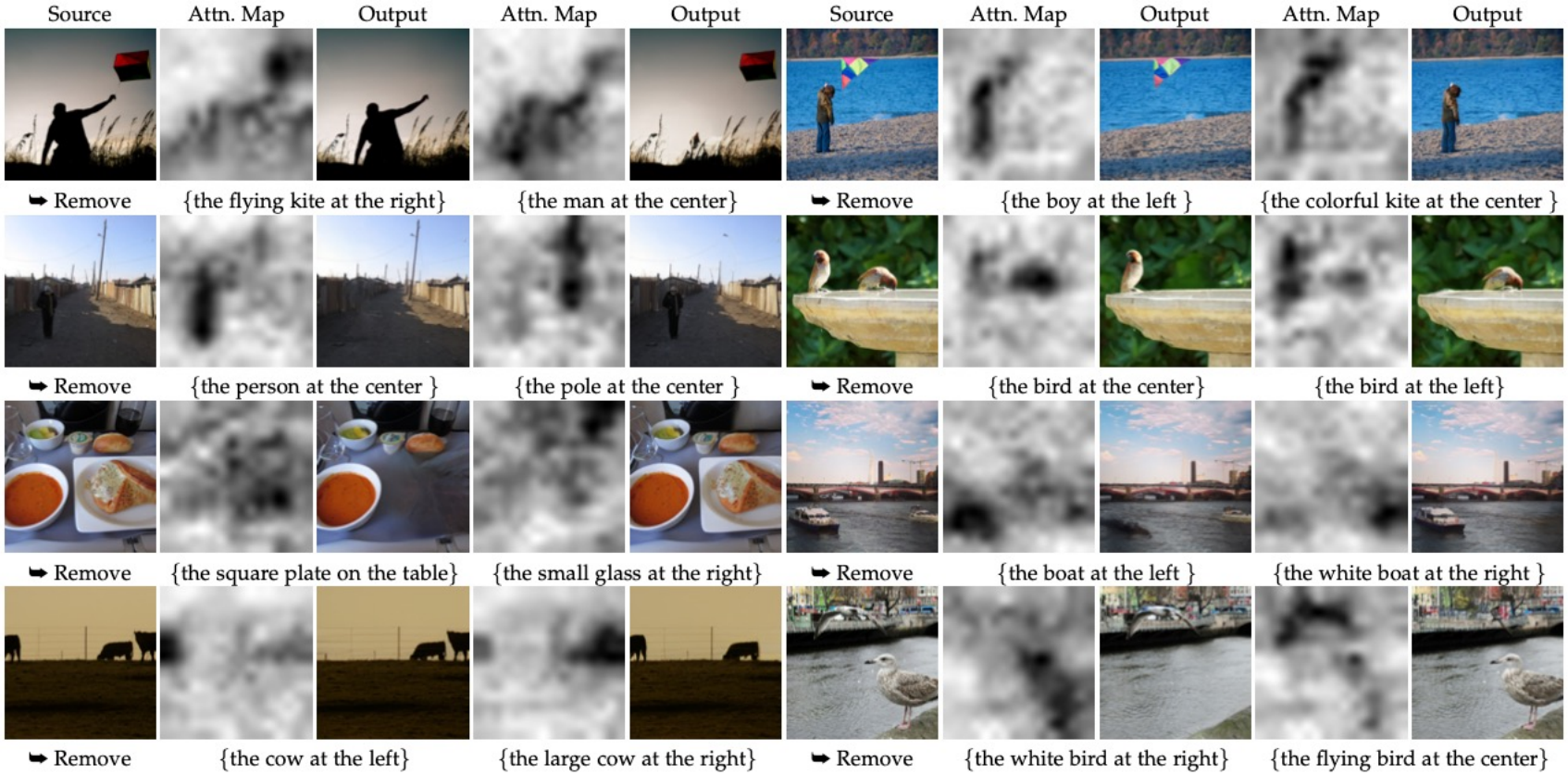
Inpainting and Editing Images

Diverse Inpainting and Editing with GAN inversion (ICCV 2023)
Ahmet Burak Yildirim, Hamza Pehlivan, Bahri Batuhan Bilecem, Aysegul Dundar

Inpainting with Instructions

Inst-Inpaint: Instructing to Remove Objects with Diffusion Models
(in submission, 2023)
Ahmet Burak Yildirim, Vedat Baday, Erkut Erdem, Aykut Erdem, Aysegul Dundar

Inpainting with Instructions



Inst-Inpaint: Instructing to Remove Objects with Diffusion Models
(in submission, 2023)
Ahmet Burak Yildirim, Vedat Baday, Erkut Erdem, Aykut Erdem, Aysegul Dundar

AlphaGo

- AlphaGO beats team of five leading Go players

Health

The Perils of Special Counsels / This Is Your ISIS on Drugs
Newsweek
06.02.2017
THE DOCTOR WILL SEE YOU NOW
HOW AI IS GOING TO CURE OUR SICK HEALTH CARE SYSTEM

November 19th, 2017
ROBOT OF THE YEAR
TIME

Many Other Applications

- Information Retrieval
- Online advertising
- Finance
- Fraud detection
- Economics
- Social Sciences

Term Projects

- **“Photobomb Detector”**
- **“Facial Expression Recognition”**
- **“Predicting the Stack Exchange Questions’ Category”**
- Some ideas from Stanford ML course:
  - http://cs229.stanford.edu/proj2019aut/index.html

TO DO

- Enroll on the Moodle:
- Read the syllabus:
  - https://stars.bilkent.edu.tr/syllabus/view/CS/464/
- Read information:
  - https://www.cs.bilkent.edu.tr/~adundar/teaching/cs464/

- Start talking to your classmates about project ideas and start establishing teams:
  - Groups of 5 people from the same section.

Growth of Machine Learning

- Machine learning is preferred approach to
  - Speech recognition, natural language processing
  - Computer vision
  - Medical outcomes analysis
  - Robot control
  - Computational biology
  - Sensor networks
- This trend is accelerating
  - Big data
  - Improved machine learning algorithms
  - Faster computers
  - Good open-source software

Why “Learn” ?

- Learning is useful when:
  - Human expertise does not exist (navigating on Mars),
  - Humans are unable to explain their expertise (speech recognition, object recognition)
  - Solution needs to be adapted to particular cases (personalization)

Experience = data

- Humans learn to improve their skills through practice

- The machine learning goal is to write algorithms that will improve with **experience**
  - Experience is encoded as **examples, or data**
  - Autonomous learning must gain from experience

- Where does training data come from?
  - Human annotation
  - Automatic collection

Objectives of Machine Learning

- **Algorithms:**
  - design of efficient, accurate, and general learning algorithms to deal with large-scale problems
  - make accurate predictions (unseen examples)
  - handle a variety of different learning problems

- **Theoretical questions:**
  - What can be learned? Under what conditions?
  - What learning guarantees can be given?
  - What can we say about the inherent ease or difficulty of learning problems?

Machine Learning?

- **Role of Probability and Statistics:** Inference from a sample

- **Role of Computer Science**: Efficient algorithms to
  - solve the optimization problem
  - representing and evaluating the model for inference

Definitions and Terminology

- **Example**:
  - an object or instance in data used
- **Features (denoted by X):**
  - the set of attributes, often represented as a vector, associated to an example, e.g., height and weight for gender prediction
- **Labels (Y):**
  - in classification, category associated to an object
  - in regression, real-valued numbers

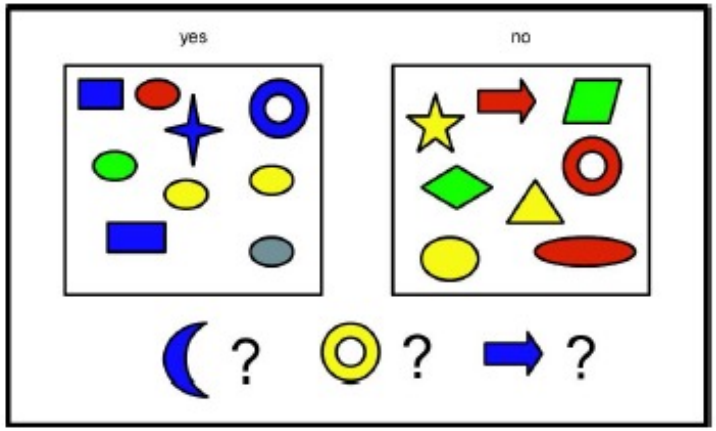
Definitions and Terminology

- **Training set:** Data used for training the learning algorithm

- **Test data:** Data used for evaluating the learned model

Feature Extraction and Mapping

- How you represent your data as a set of features is very! important.



(a)

D features (attributes)

N cases

Color	Shape	Size (cm)
Blue	Square	10
Red	Ellipse	2.4
Red	Ellipse	20.7

Label
1
1
0

(b)

Example: Recognizing Flowers

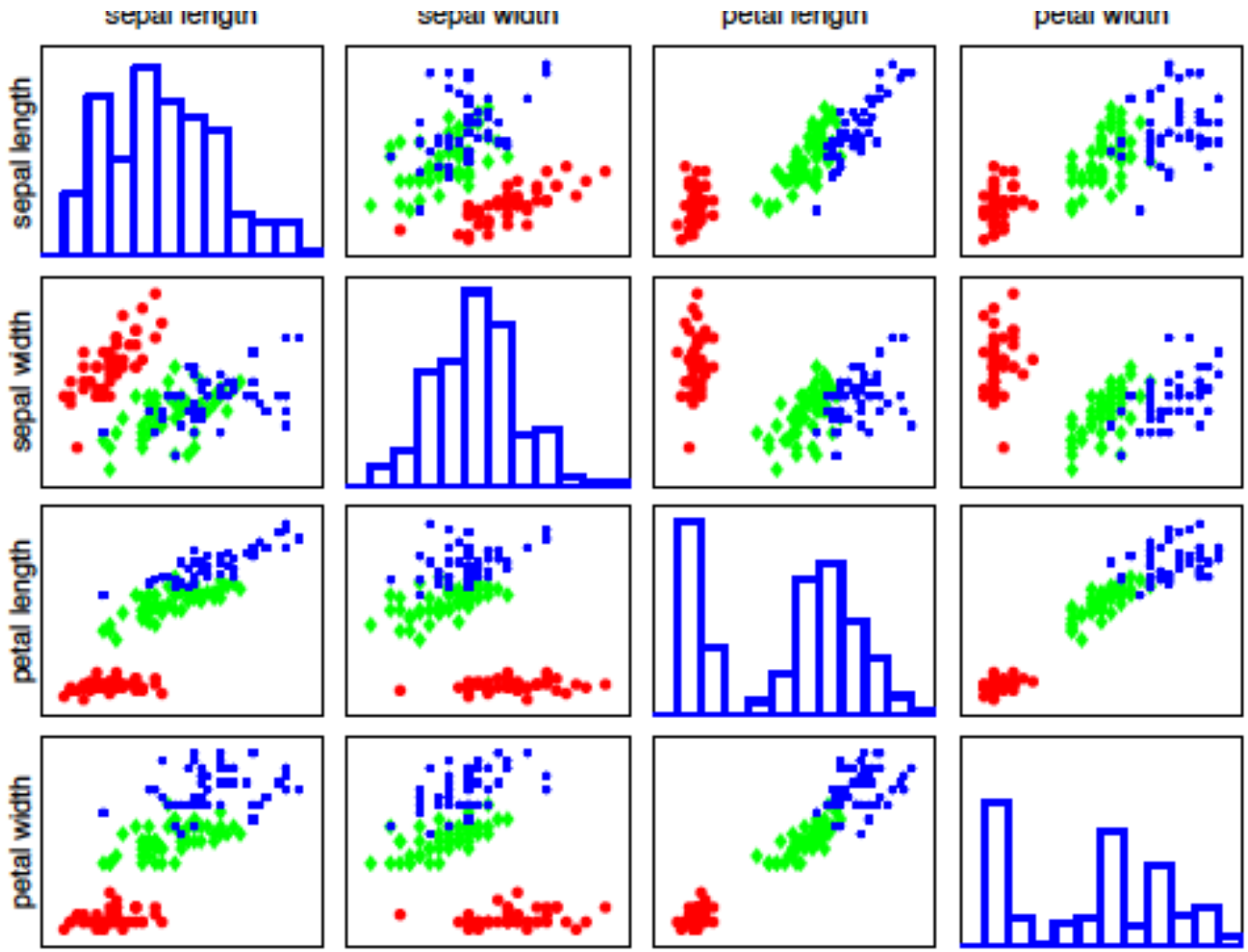
Three types of Iris Flowers

Setosa

Versicolor

Virginica

Flower Recognition: Features



Setosa

Versicolor

Virginica

Example: A Digit Recognizer

- Imagine you are asked to write a program to recognize hand written digits

- Now let’s use the terminology to define the problem

Terminology Example

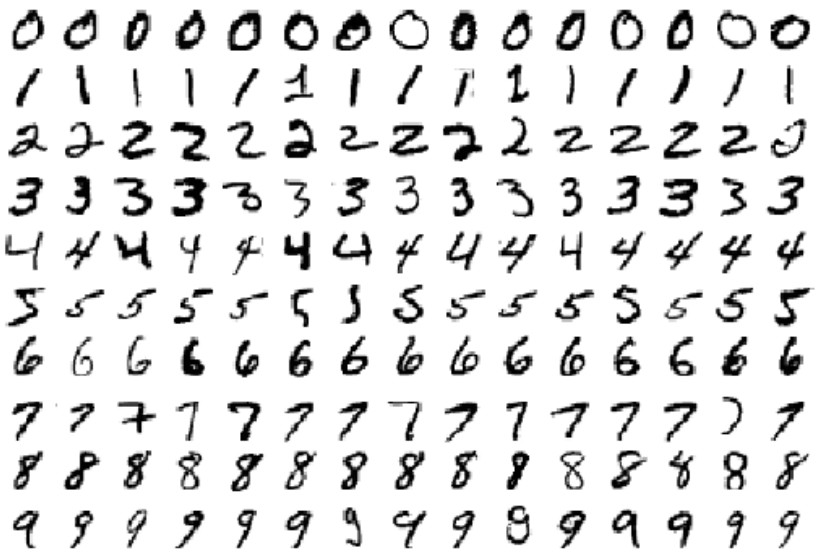
Classification Task:

Given a hand written digit classify it into {0,1,..,9}

→ 3

Digit Recognizer

- Each hand written digit image is an **example** (instance, or object) in your data

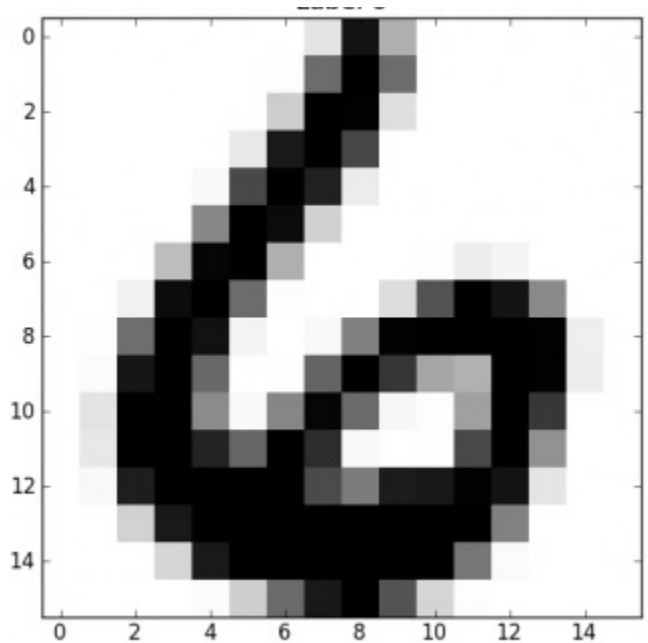


Labels (Space of outputs)

- Labels are all digits:

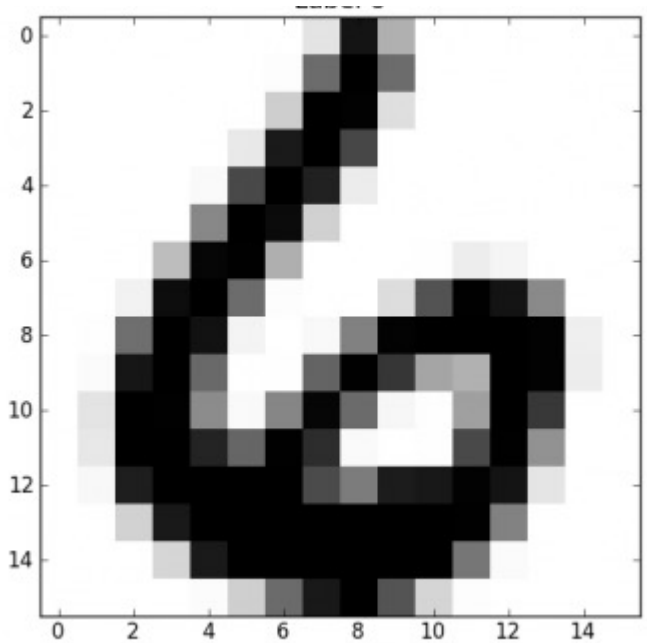
$$Y \in \{0, 1, \ldots, 9\}$$

Label of this example: 6



Features (Space of Inputs)

We could divide the image in 16x16 pixels and encode each pixel intensity as a numeric number.



16 x16 grid

Map to features →

Feature vector for the example

$$\begin{pmatrix} 0 \\ . \\ . \\ . \\ 1.2 \\ 3.2 \\ 4.5 \\ 1.2 \\ . \\ . \\ 0 \end{pmatrix}$$

256 x 1

Feature Mapping

- How you represent your data as features is very! important.

Training Set/Test Set

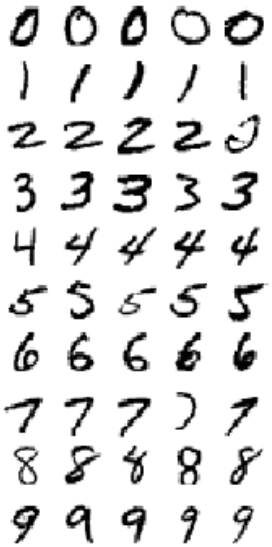
Training set

$$D_{train} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$$

+ their labels

Test set

$$D_{test} = \{(X_1, Y_1), \ldots, (X_m, Y_m)\}$$



+ their labels

Should not be overlapping

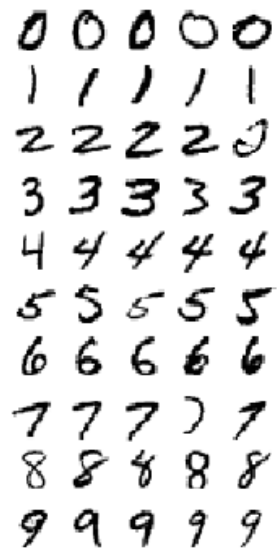
Terminology Example: Digit Recognizer

Training set

+ their labels

- **Built** the classifier using the features and the labels in the training set

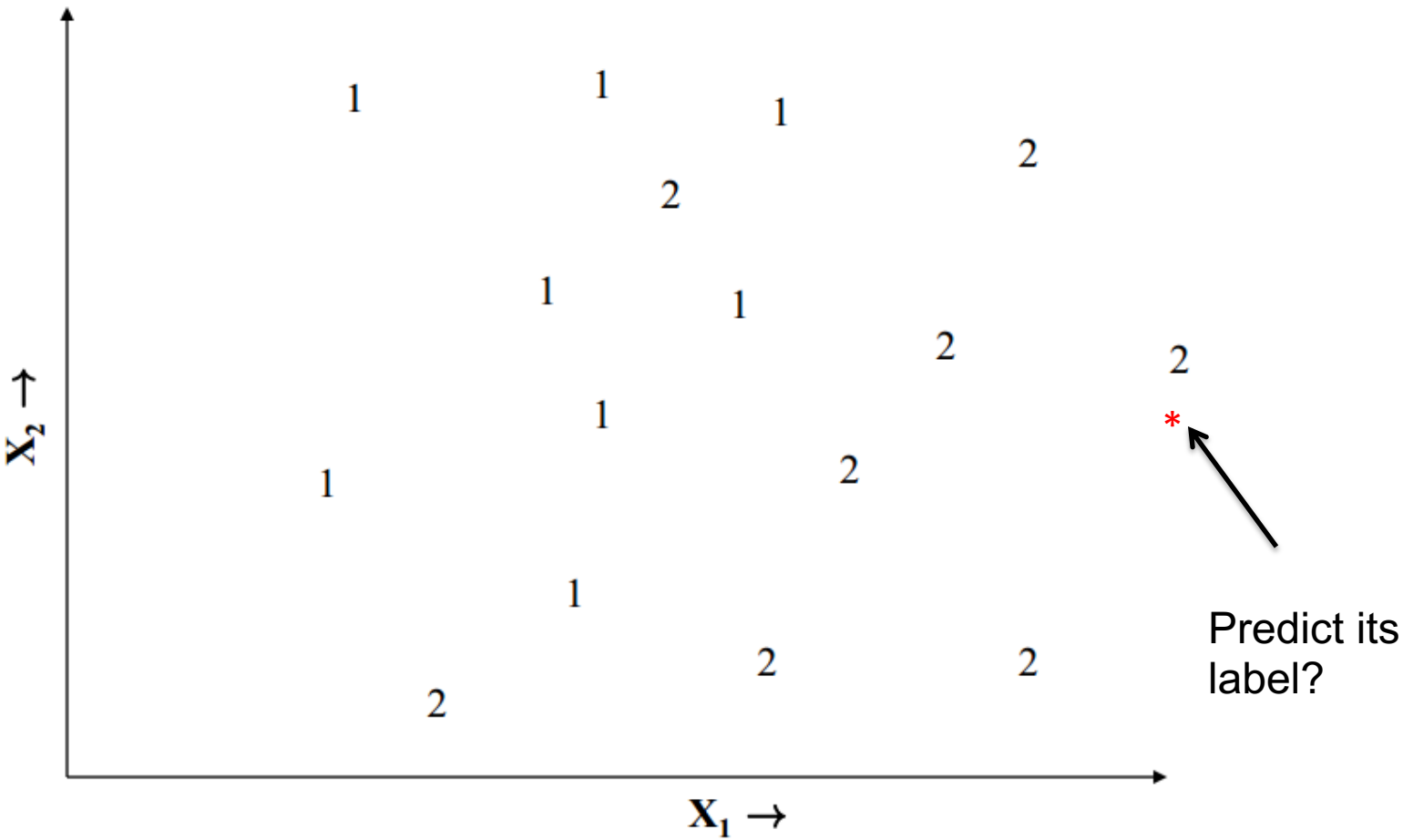
Test set



+ their labels

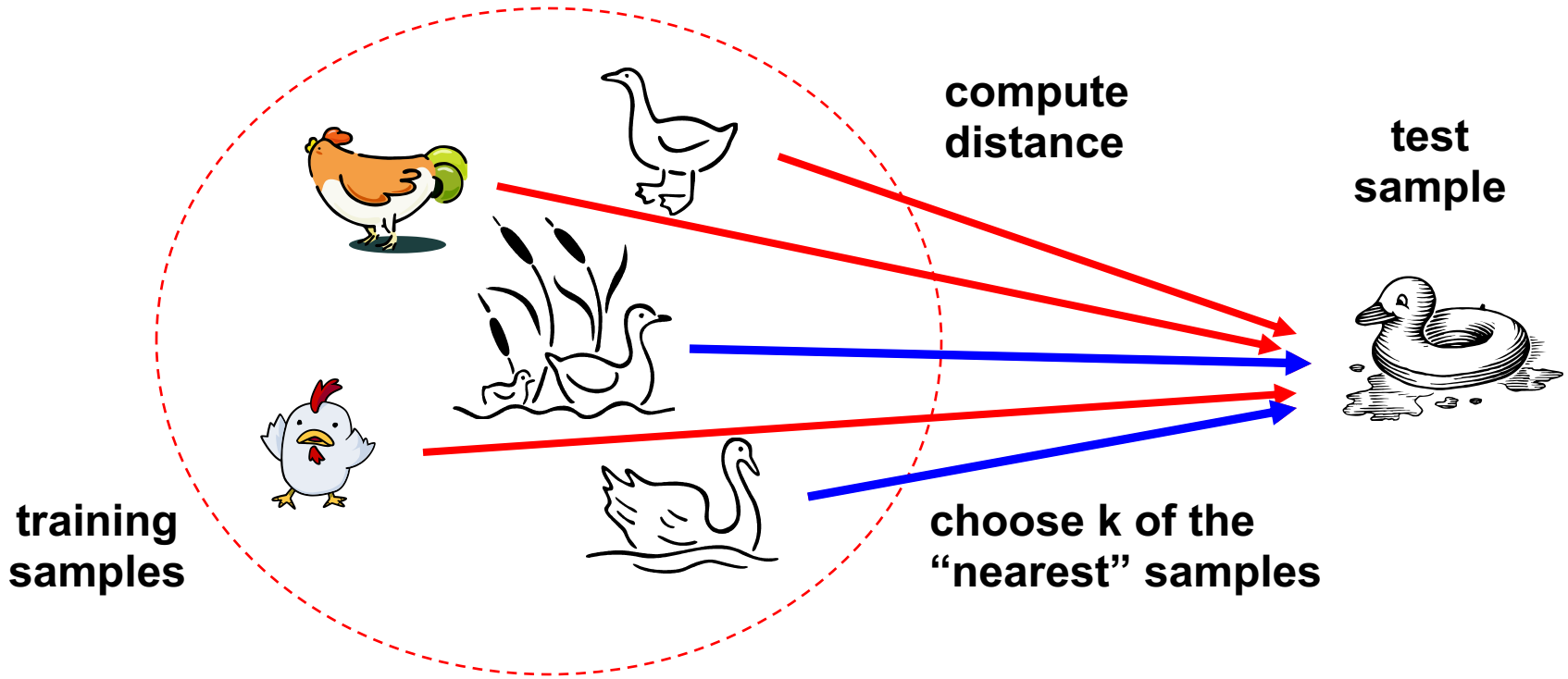
- **Predict** the label of **unseen** examples' features.
- **Compare** your predictions results with the actual labels.

How would you classify ?

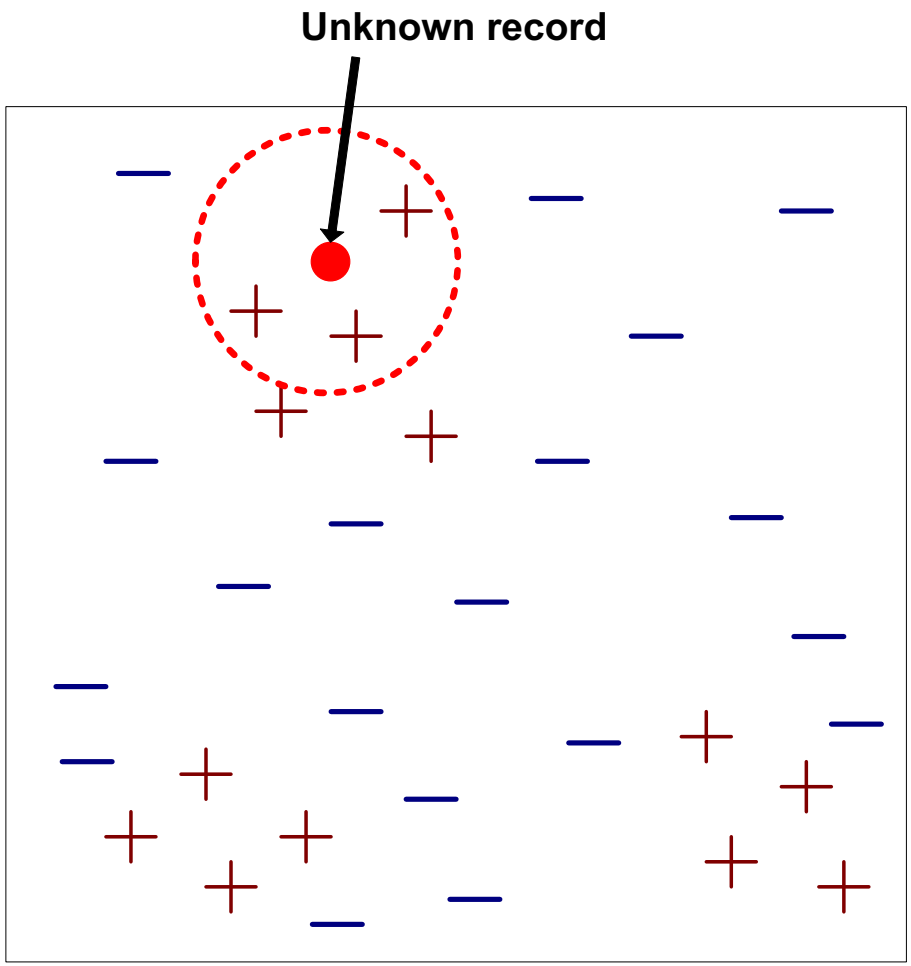


Nearest Neighbor Classifiers

- Basic idea:
  - If it walks like a duck, quacks like a duck, then it's probably a duck



Nearest Neighbor Classifiers



k-Nearest Neighbor (kNN)

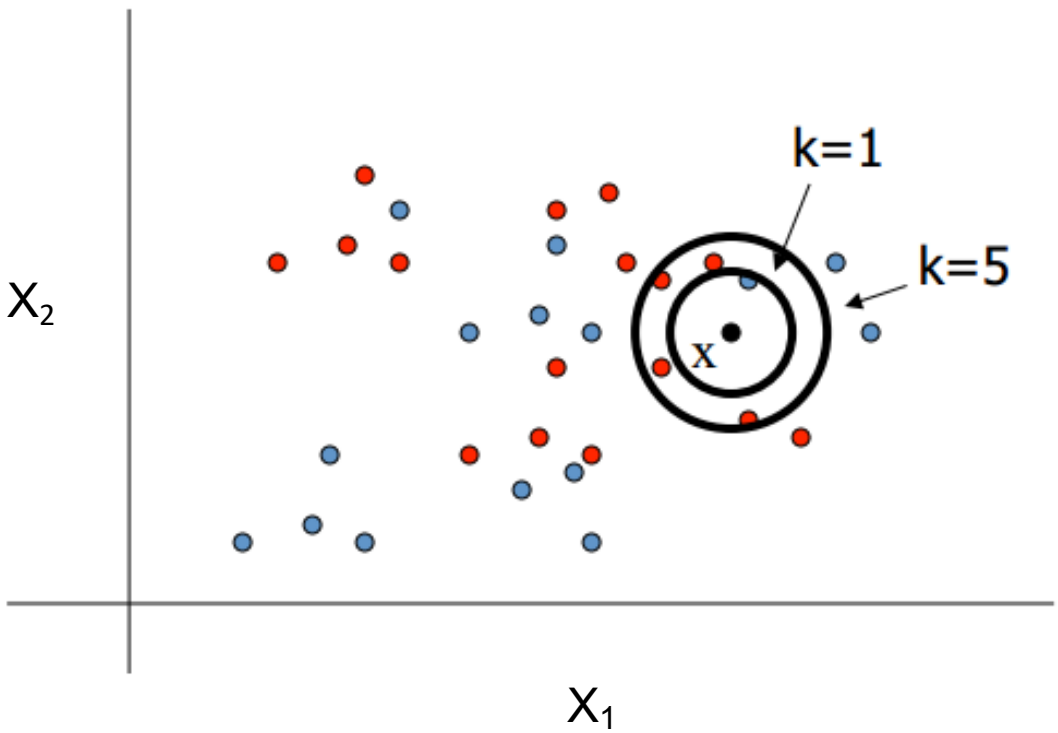
- Assumes all instances are points in n-dimensional space
- A distance measure is needed to determine the "closeness" of instances
- Classify an instance by finding its nearest neighbors and picking the most popular class among the neighbors

1- Nearest Neighbor Classifier Algorithm

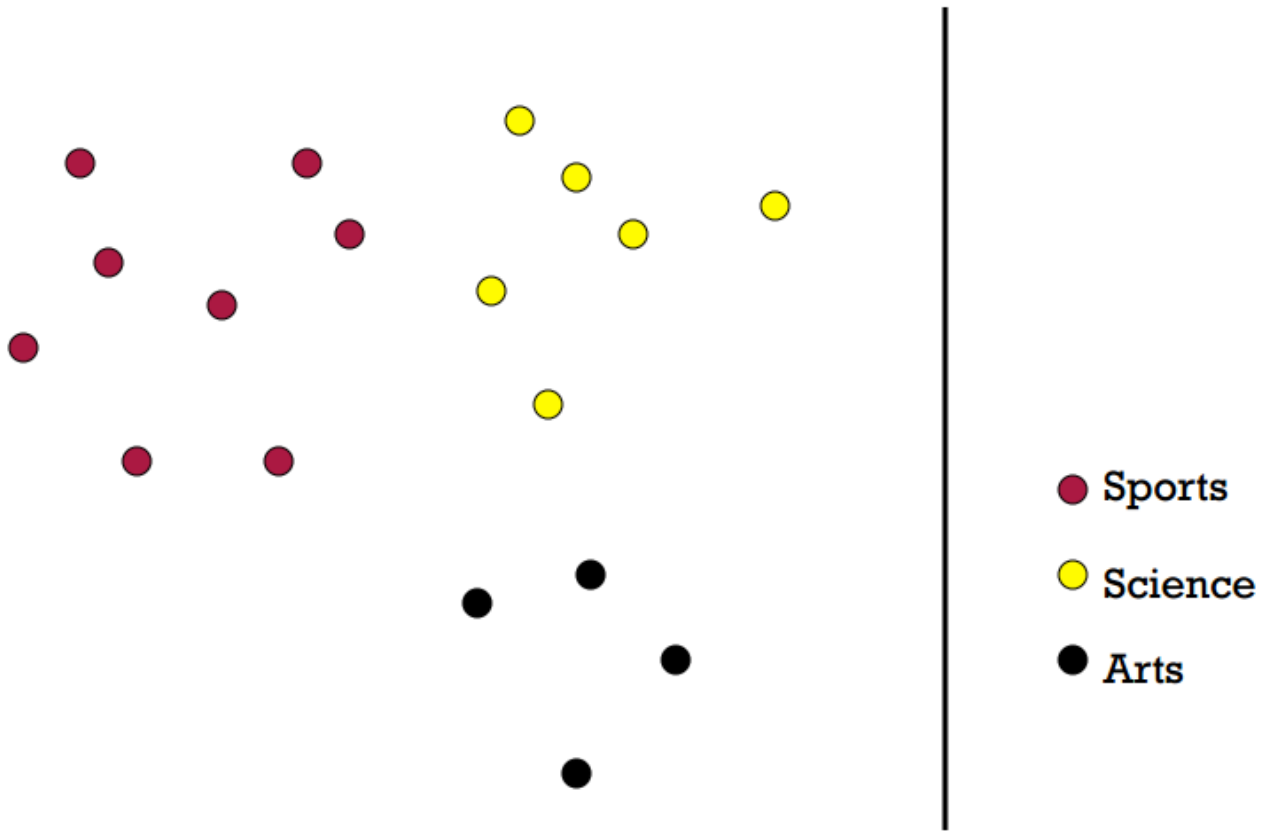
- The training data $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$
- Learning algorithm:
  - Store training examples
- Prediction algorithm:
  - To classify a new example x_{new} by finding the training example (x_i, y_i) that is nearest to x_{new}
  - Guess the class $y_{new} = y_i$

k-NN Classifier

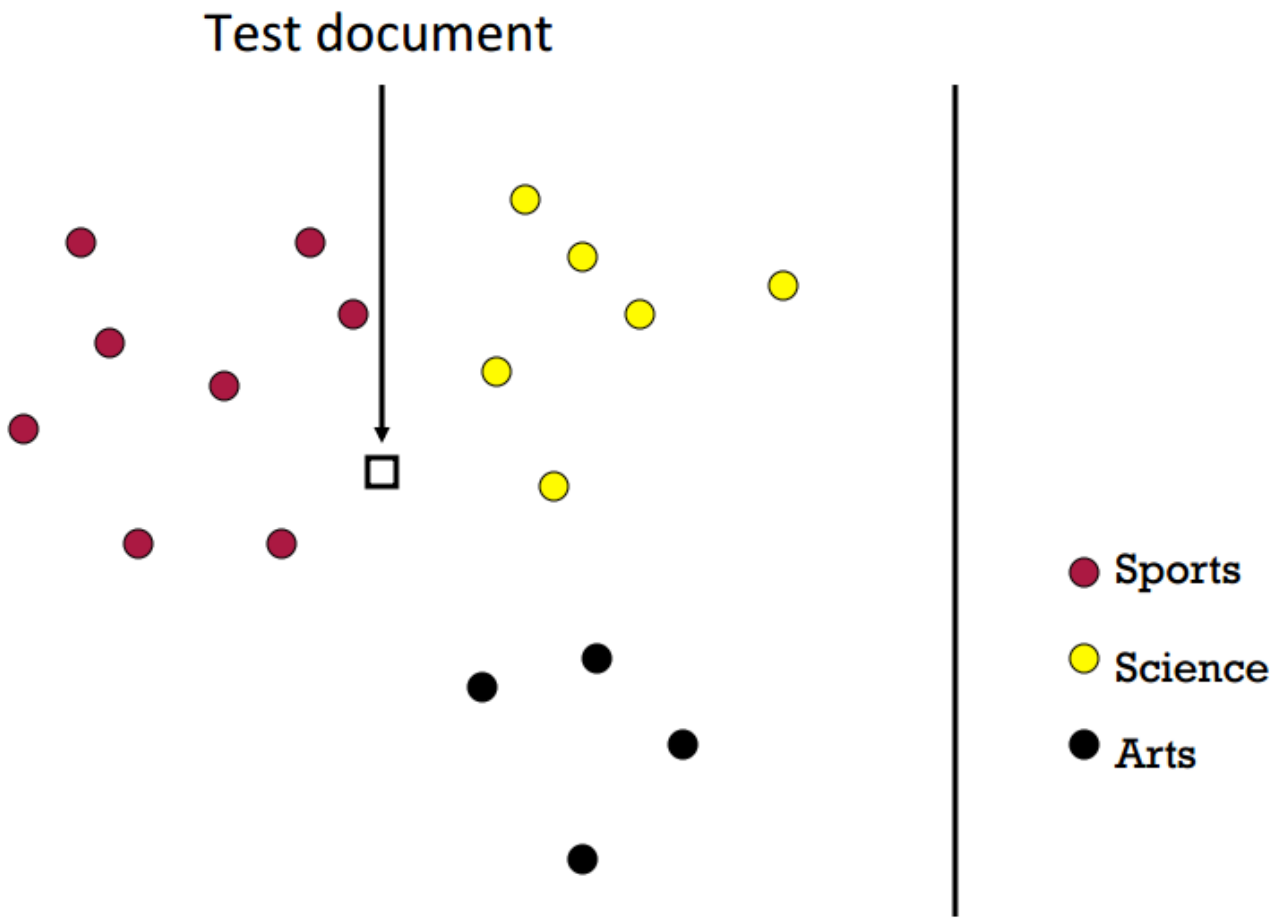
- To classify a new input vector x, examine the k-closest training data points to x and assign the object to the most frequently occurring class



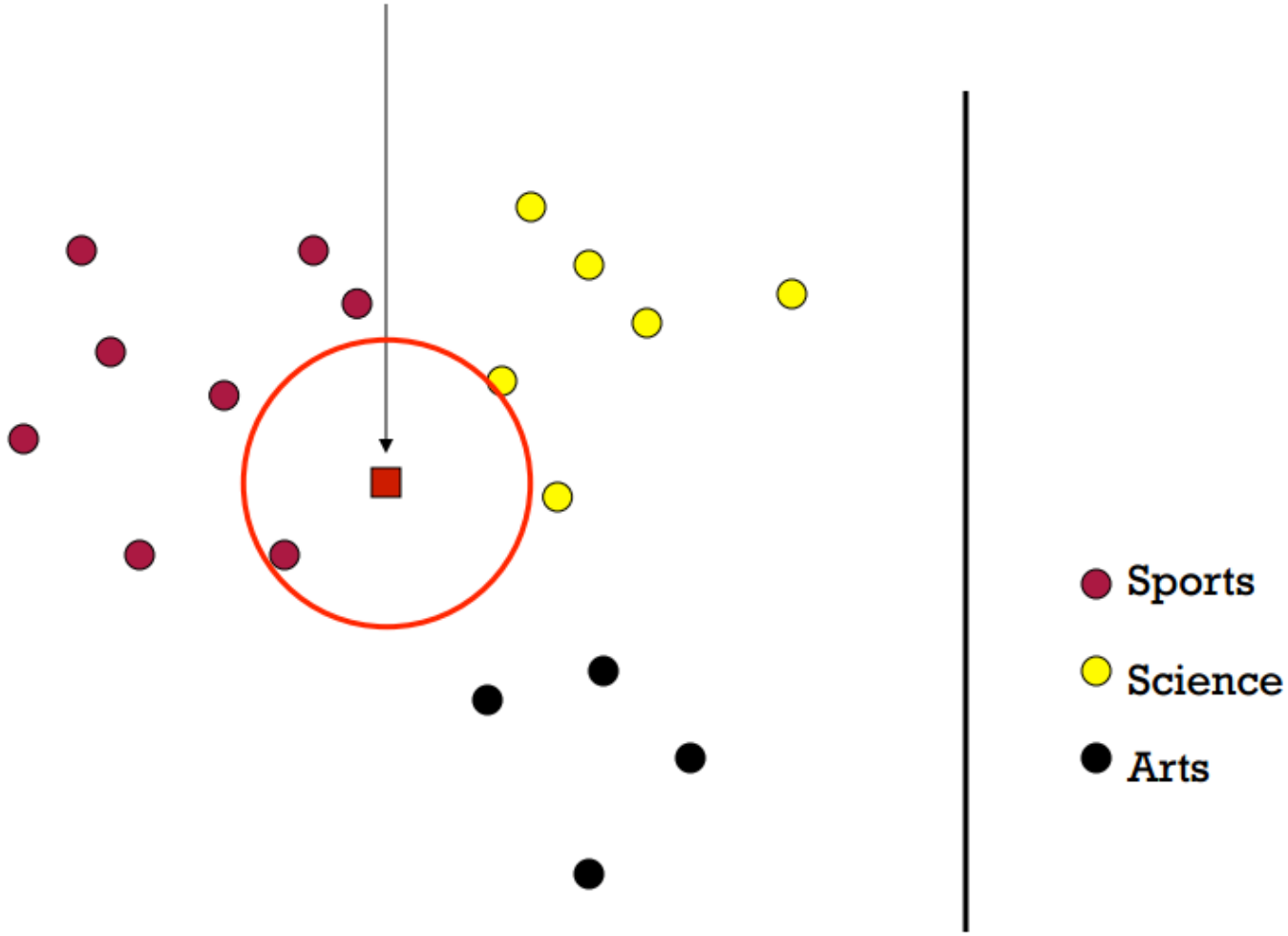
k-NN Classifier



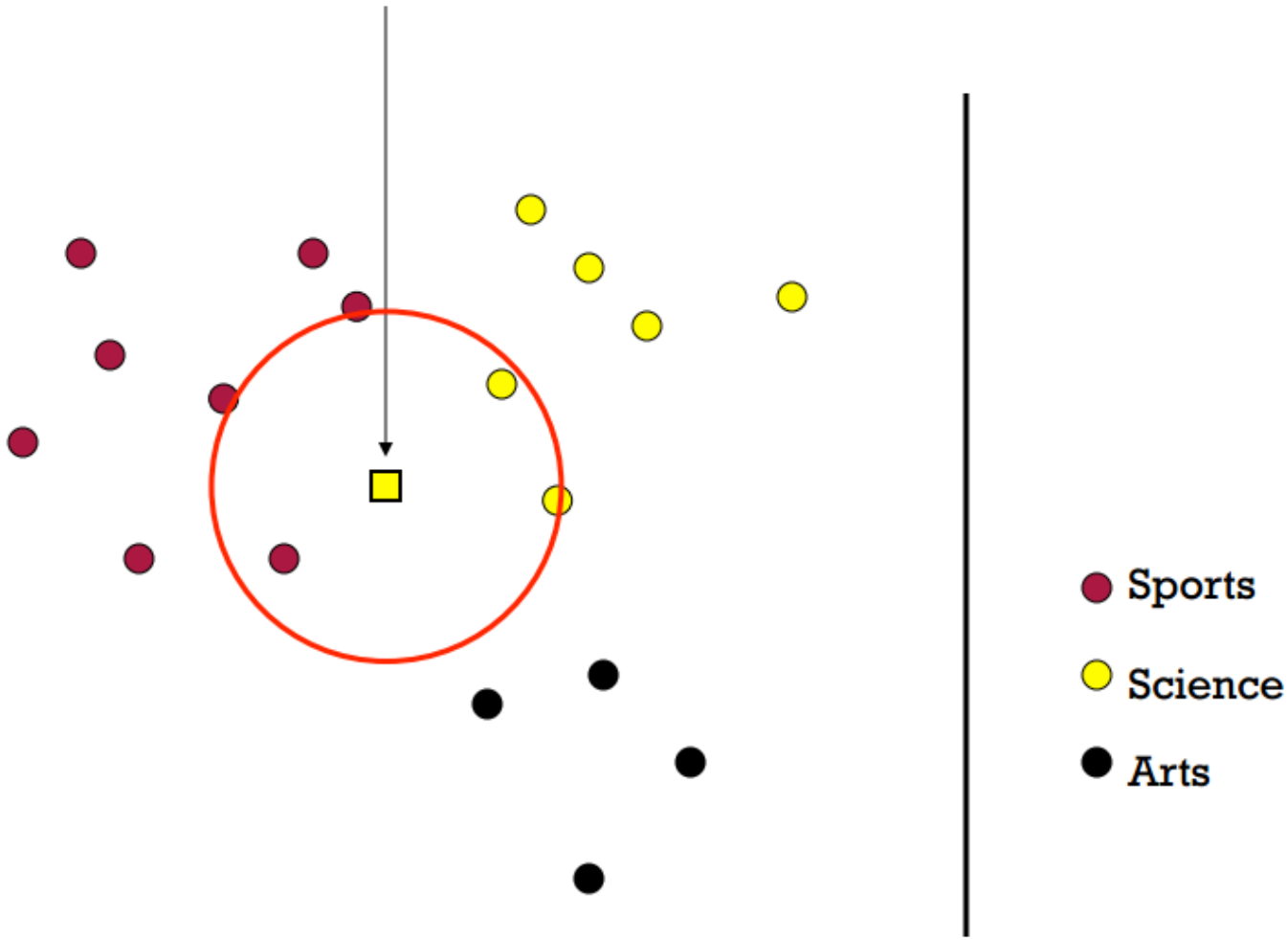
k-NN Classifier



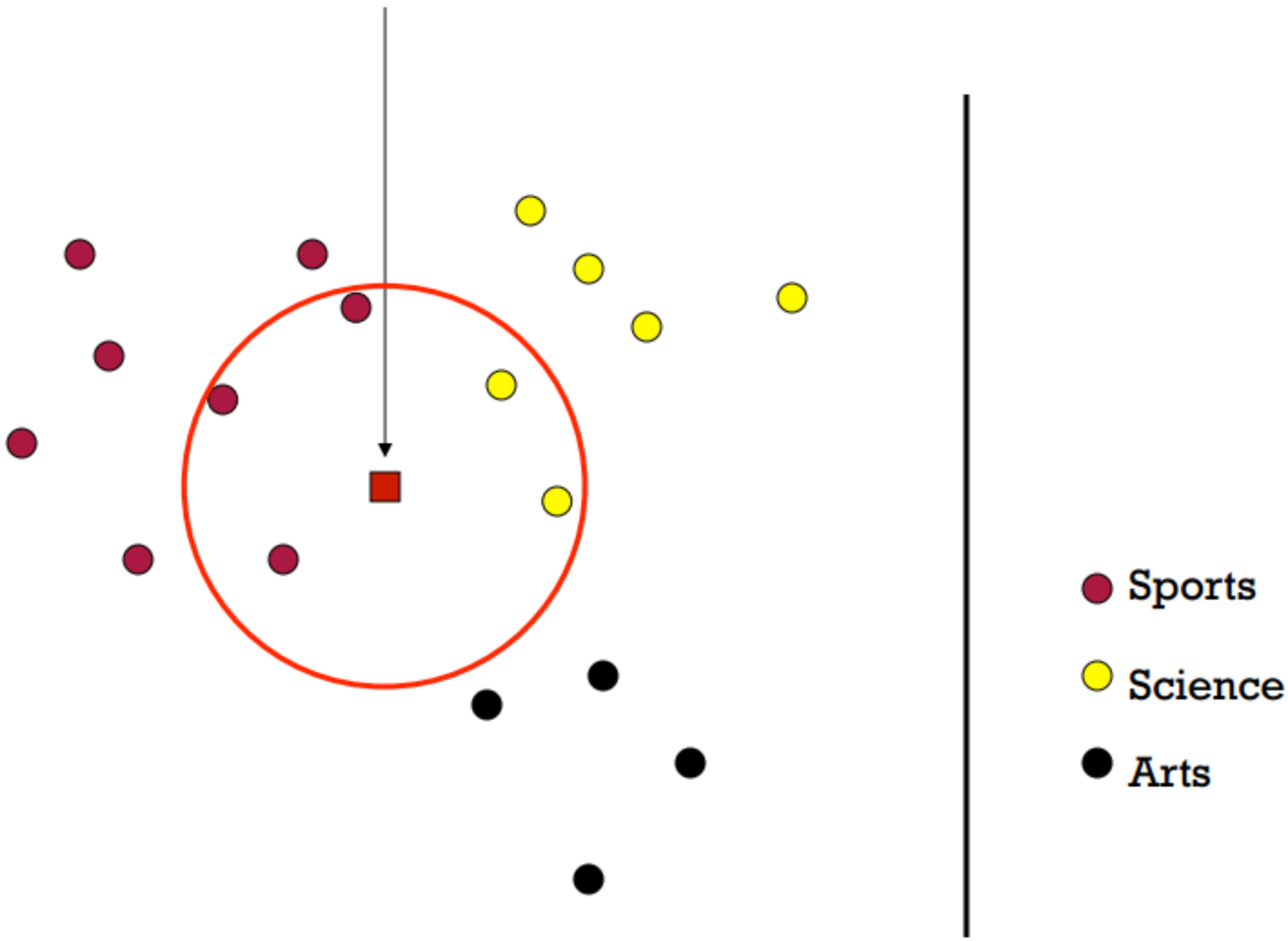
1-Nearest Neighbor Classifier (k = 1)



3-Nearest Neighbor Classifier (k=3)

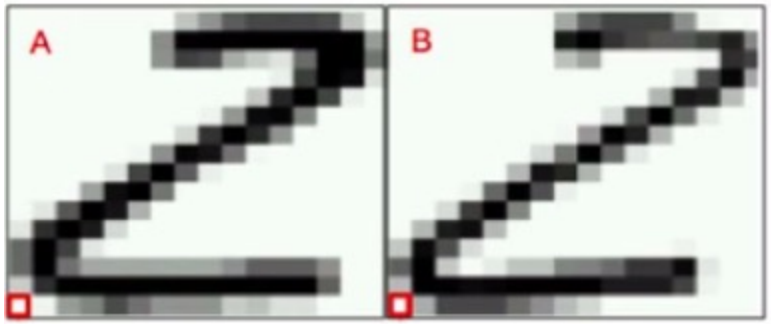


5-Nearest Neighbor Classifier (k=5)



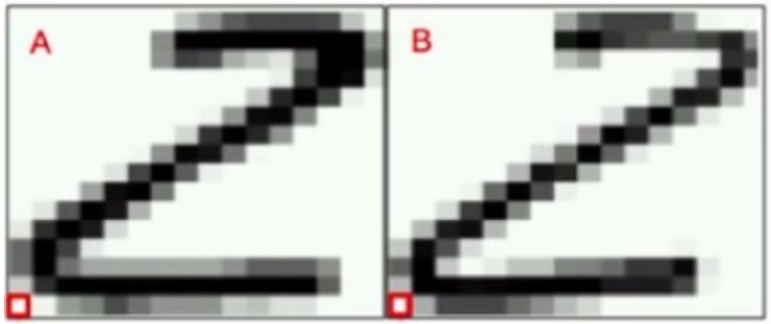
Example: Hand Written Digits

- 16x16 bitmaps
- 8-bit grayscale
- Euclidean distance over pixels

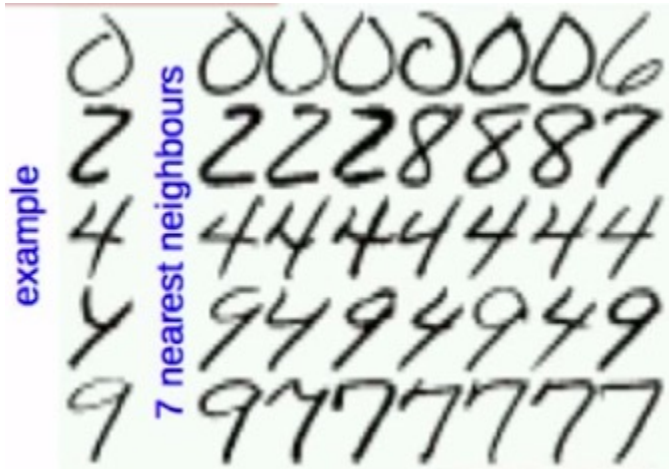


Example: Hand Written Digits

- 16x16 bitmaps
- 8-bit grayscale
- Euclidean distance over pixels

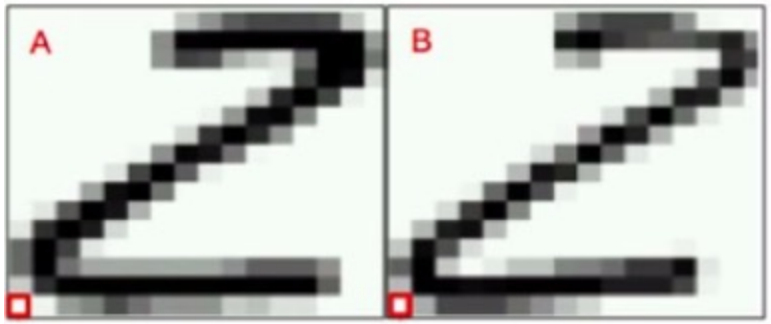


$$D(A,B) = \sqrt{\sum_r \sum_c (A_{r,c} - B_{r,c})^2}$$

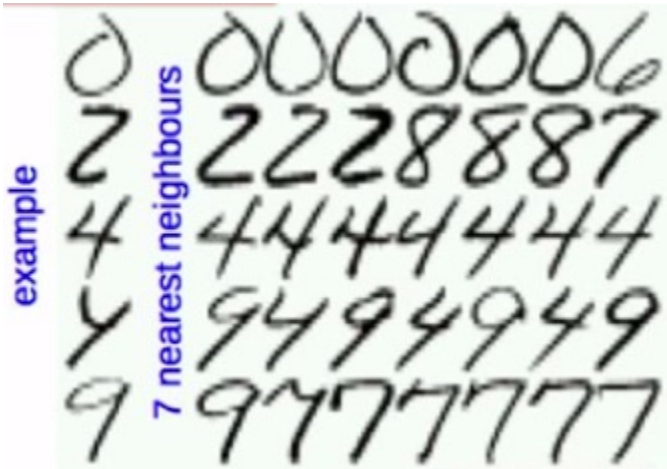


Example: Hand Written Digits

- 16x16 bitmaps
- 8-bit grayscale
- Euclidean distance over pixels



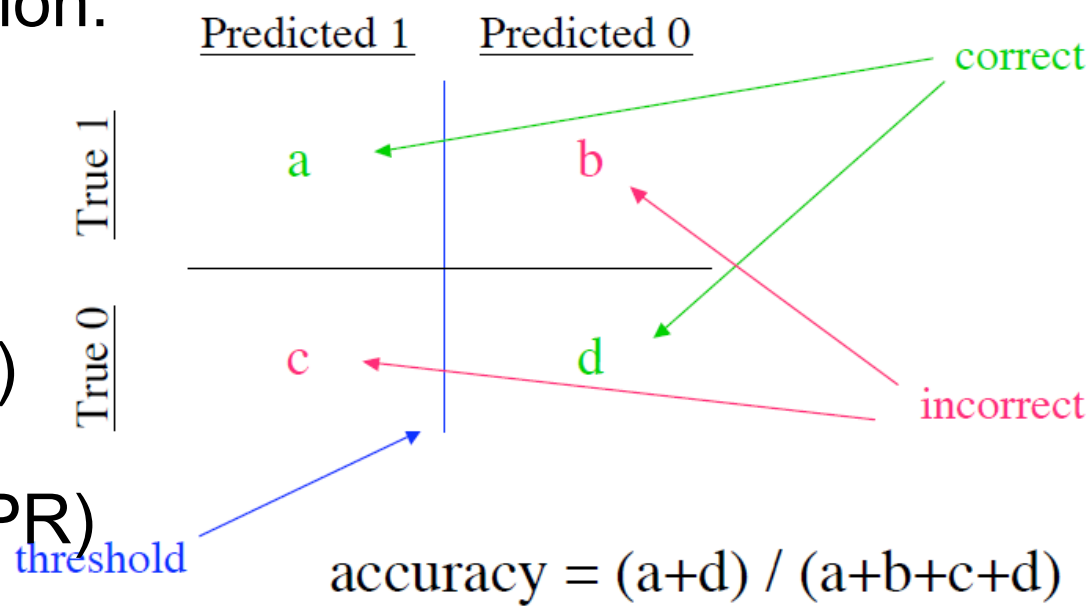
$$D(A,B) = \sqrt{\sum_r \sum_c (A_{r,c} - B_{r,c})^2}$$



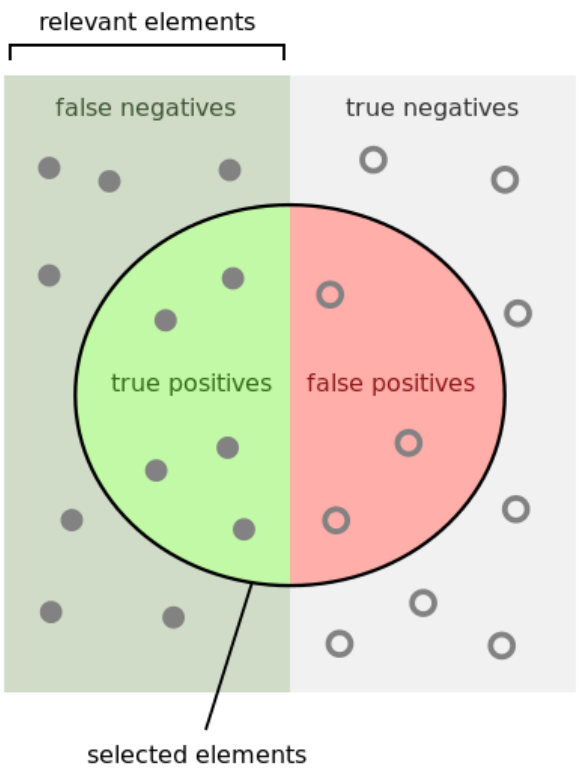
Accuracy: 7-NN ~95.2%
SVM ~95.8%
Humans ~97.5%

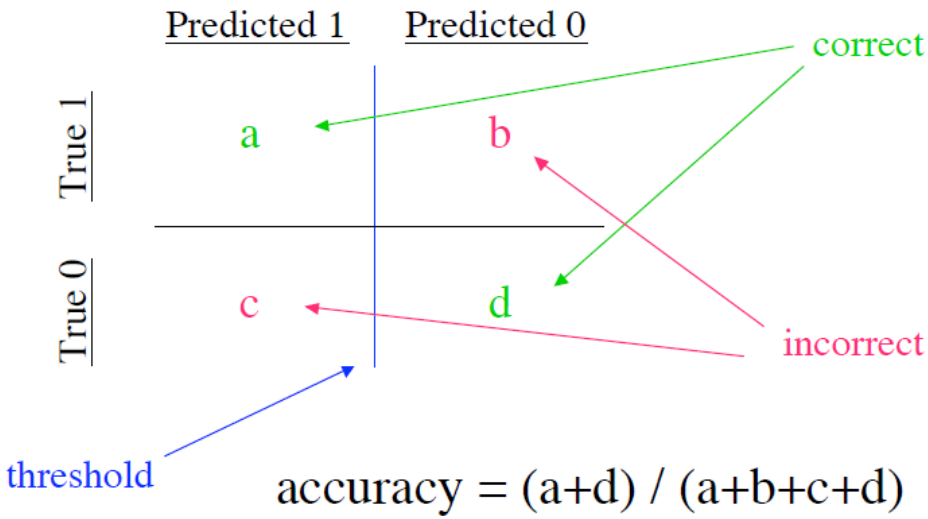
Performance Measures

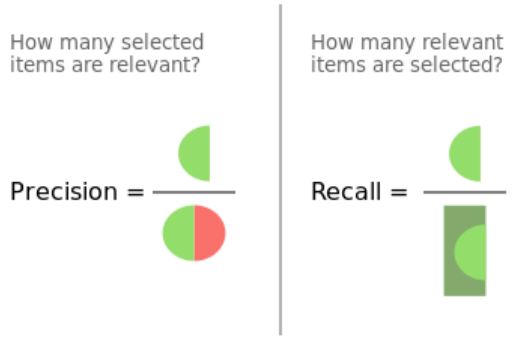
- For binary classification:
  - True positives
  - True negatives
  - False positives
  - False negatives
- Precision (selectivity)
  - a/(a+c)
- Recall (sensitivity, TPR)
  - a/(a+b)
- Specificity, FPR
  - d/(c+d)
- There is often a trade-off between precision and recall
  - Receiver Operating Characteristic (ROC) curves are used to accurately assess performance in relation to this trade-off



Performance Measures







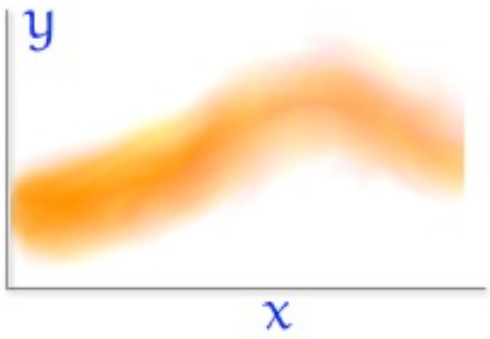
Source: wikipedia

Relationship to Modeling

Functional Representation

- Variable X is related to variable Y, and we would like to learn this relationship from the data
- **Example:** X = [weight, blood glucose, . . .] and Y is whether the person will have diabetes or not.
- We assume there is a relationship between X and Y:
  - it is less likely to see certain X co-occur with “low risk” and unlikely to see some other X co-occur with “high risk".

- This relationship is encapsulated by P(X, Y)



Formulation of Learning

- Data denoted by : $\mathcal{D} = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$, where n is the sample size
- The distribution $P(X, Y)$ is unknown to us (otherwise there is no learning to be done)
- Assumption: the data is sampled independently from $P(X, Y)$ (the i.i.d assumption)
- Upon observing the data: $\mathcal{D} = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$, the classifier is asked to summarize what it learned about the relationship X and Y
- The learner's summary takes the form of a function $\hat{f}_D \mathcal{X} \longmapsto \mathcal{Y}$

Multitude of Learning Frameworks

- Presence/absence of labeled data
  - Supervised learning: $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$
  - Unsupervised learning: $\{X_1, X_2, \ldots, X_n\}$
  - Semi-supervised learning: A mixed of above

Multitude of Learning Frameworks

- Type of Labels:
  - Binary Classification $Y \in \{0, 1\}$
  - Multi class classification $Y \in \{0, \ldots, K\}$
  - Regression $Y \in \mathbb{R}$
  - Structure prediction: Y is a complex object like a graph or a tree

Multitude of Learning Frameworks

- Problems also differ in the protocol for obtaining data
  - Passive Learning
  - Active Learning
- and in assumptions in data
  - Batch
  - Online

Examples of Regression Problems

- Predict tomorrow's stock market price given current market conditions and other possible side information.
- Predict the age of a viewer watching a given video on YouTube.
- Predict the location in 3D space of a robot arm end effector, given control signals (torques) sent to its various motors.
- Predict the amount of prostate specific antigen (PSA) in the body as a function of a number of different clinical measurements.
- Predict the temperature at any location inside a building using weather data, time, door sensors, etc.

Linear Regression

Residual error

$$y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + \epsilon = \sum_{j=1}^{D} w_j x_j + \epsilon$$

Outcome variable (real-valued labels)

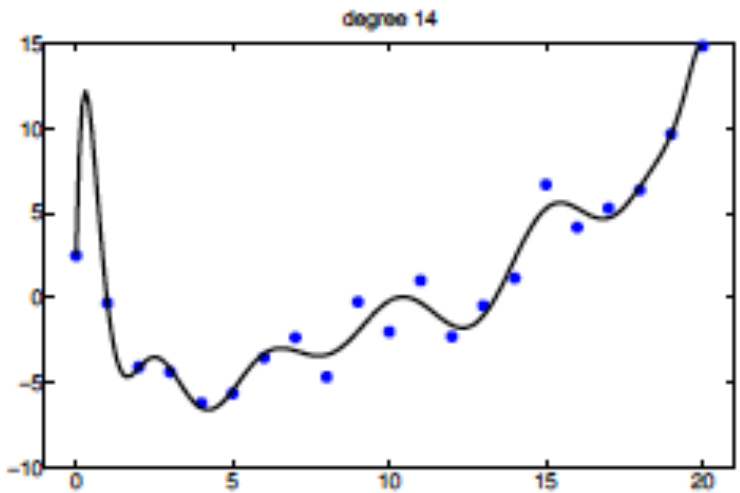
Coefficient vector

Feature vector

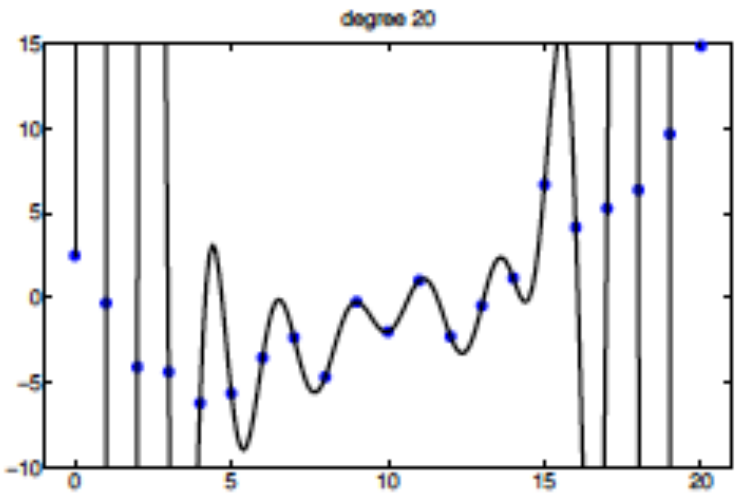
- **Optimization problem:** Given instances of *x* and *y*, compute the coefficient vector that will minimize the residual error
- Common assumption: Residual error has a Gaussian (normal) distribution

What Can Go Wrong?

Overfitting



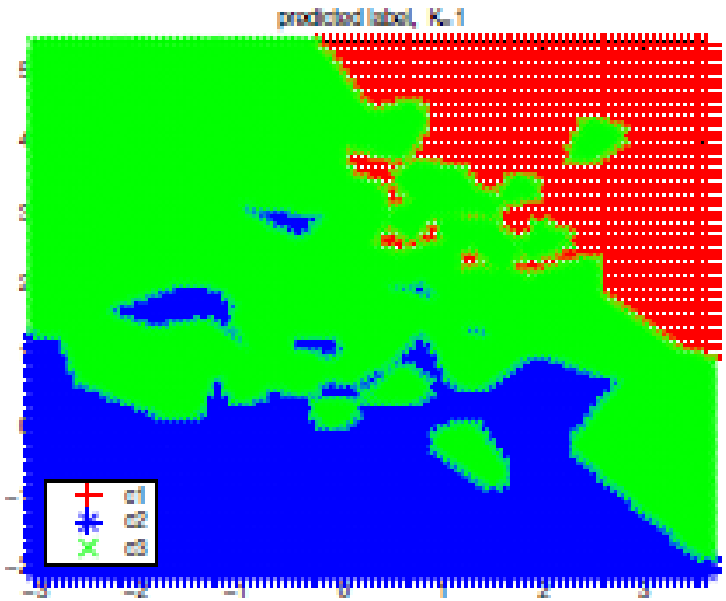
14-degree polynomial



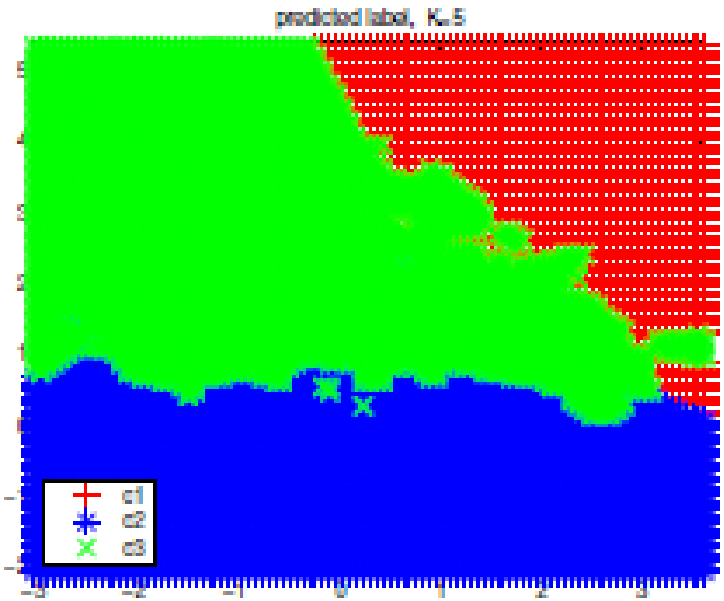
20-degree polynomial

- Which model is more likely to be generalizable?
  - If we try to fit the model to the training data too accurately, we will end up with a too complex model that considers every bit of variation in the data as information
- **Occam's Razor**

Overfitting Example: k-NN

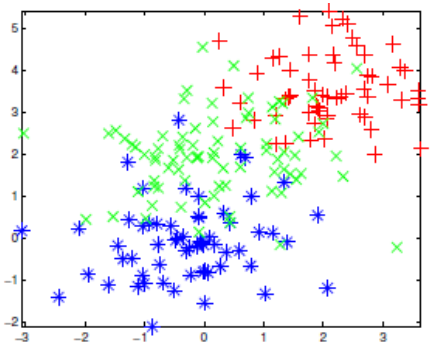


MAP Estimate for K=1



MAP Estimate for K=5

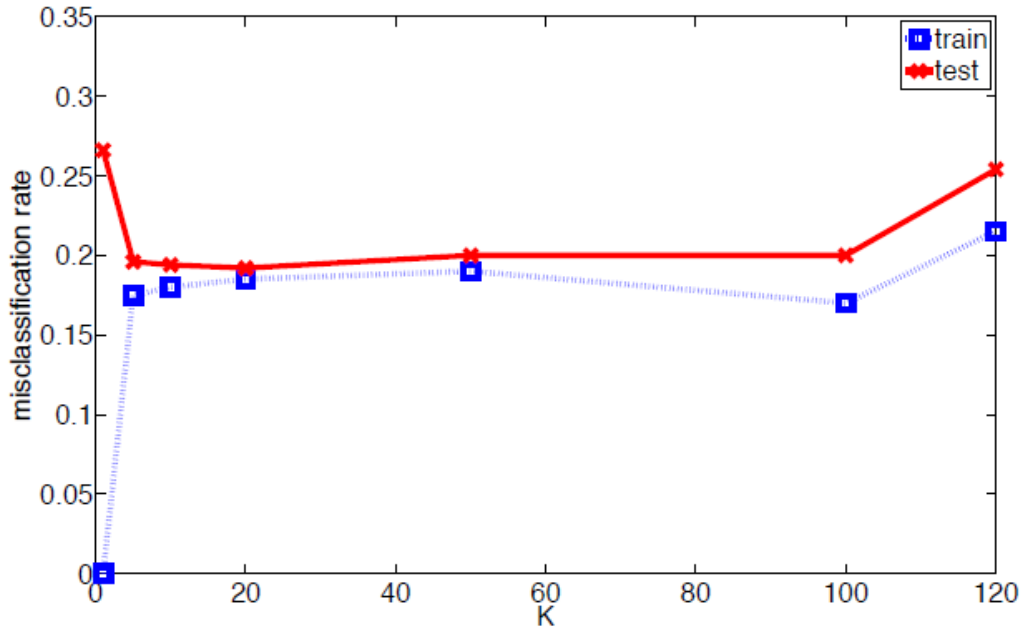
Training Data:

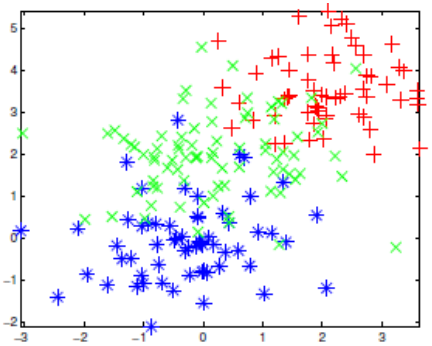


- What happens if we grow K further?
- How about when K=N (number of samples)?

Overfitting vs. Underfitting

Performance of K-NN on training and test samples as a function of K:

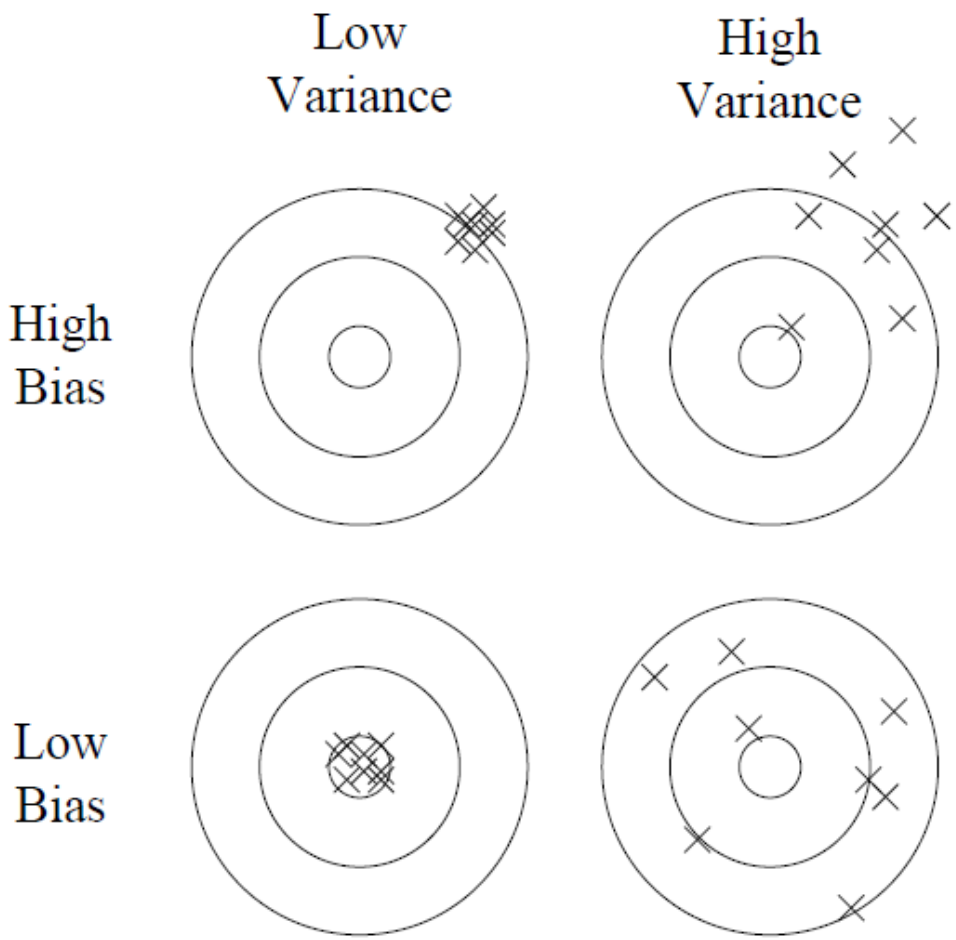




To assess overfitting, reserve a portion of these as test samples

Misclassification rate: $\mathrm{err}(f, \mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(f(\mathbf{x}_i) \neq y_i)$

Bias and Variance



Model Selection

Do not confuse this K with the K in K-NN!

- K-fold Cross-Validation
  - Randomly partition training samples into K groups
  - Repeat for i=1 to K:
    - Reserve the ith group as the test set
    - Train on the remaining (K-1)N samples
  - Results in one prediction per sample
    - Use a measure of classification accuracy to assess performance
  - Repeat the above procedure multiple times
    - Report the **mean** and **variance** of performance figures across these multiple runs

- K = 5 is usually a good choice
  - K=N => Leave-One-Out Cross Validation
- Why repeat multiple times?
- Why variance?

Curse of Dimensionality

- The space grows exponentially with the number of dimensions (features)
  - If the number of training samples stays constant as the number of features goes up, the data gets sparser
  - Distance/locality-based classifiers (e.g., k-NN) are particularly vulnerable to curse of dimensionality

Breaking The Curse

- Feature Selection
  - Explicitly aim at reducing the number of features by trying and evaluating models with different combinations of few features
    - Computationally difficult since the number of combinations is exponential
  - A parsimonious (less complex) model is always more desirable for interpretability and generalizability
- Dimensionality Reduction
  - Pre-process the data to obtain lower-dimensional representations by discovering latent patterns
    - Principal Component Analysis (PCA)/ Singular Value Decomposition (SVD)

Dimensionality Reduction via PCA

Mean

Principal Component 1

Principal Component 2

Interpretability vs. Accuracy

- **Data mining:** Find interpretable patterns
  - Emphasis on understanding
- **Machine learning:** Make accurate predictions
  - Emphasis on utility

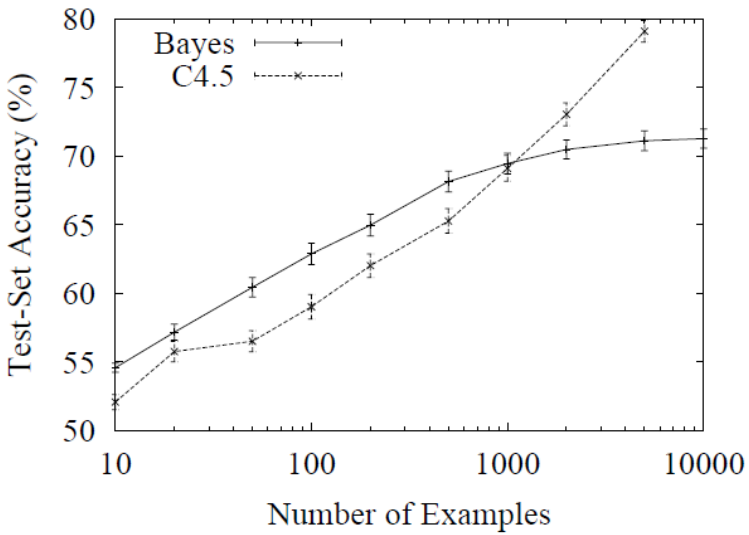


Figure 2: Naive Bayes can outperform a state-of-the-art rule learner (C4.5rules) even when the true classifier is a set of rules.

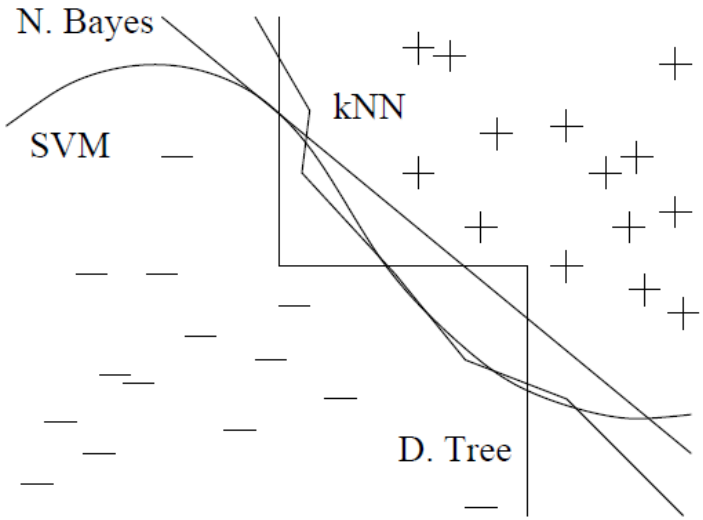


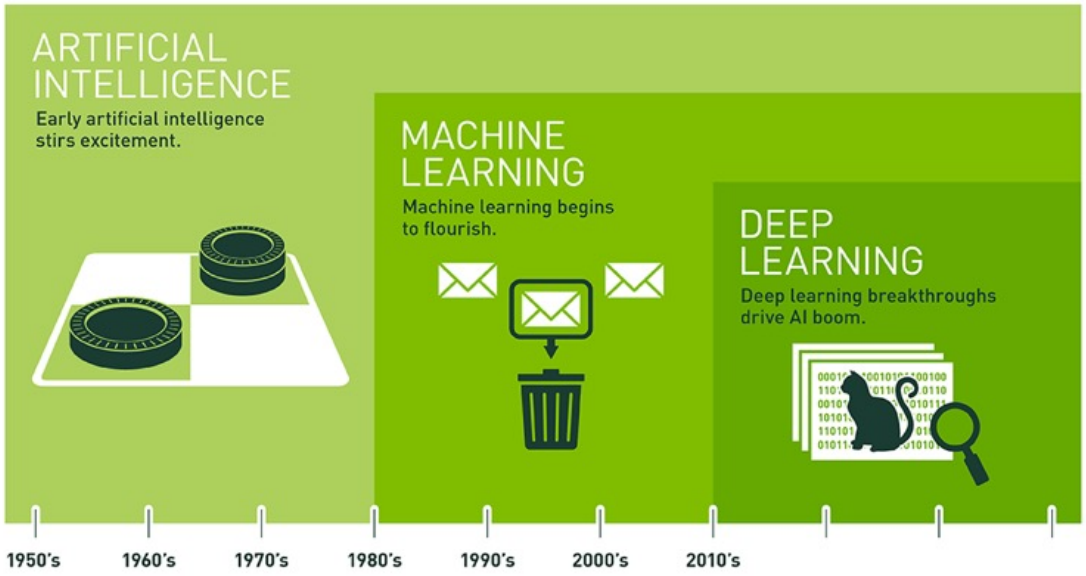
Figure 3: Very different frontiers can yield similar class predictions. (+ and − are training examples of two classes.)

More Data - Cleverer Algorithm

- Lack of sufficient number of examples is usually the biggest challenge in machine learning
  - Algorithmic sophistication can address this problem only to a certain extent, since increasing complexity can result in overfitting
- The classifier should get better as it sees more examples
  - If not, can we say it is a good algorithm?
- Extreme case: Deep Learning (Convolutional Neural Networks)
  - Works poorly with few samples
  - Works great if you have millions of samples

AI vs ML vs DL

- Nice blog post by NVIDIA:
- https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

What will be Covered?

- This class should give you the basic foundation for applying machine learning
- **Supervised learning**
  - Bayesian learning
  - Linear models for regression and classification
  - Instance-based learning
  - Support vector machines
  - Decision Trees
  - Ensemble models
  - Deep Learning
- **Unsupervised learning**
  - Clustering
  - Dimensionality reduction
- **Model selection and evaluation**
- **Sequential Models**
  - Hidden Markov Models
- **Additional topics (if time permits)**