UVA CS 4501: Machine Learning

Lecture 10: Supervised Classification

Dr. Yanjun Qi

University of Virginia

Department of
Computer Science

Where are we ? ➔ Five major sections of this course

- ❑ ~~Regression (supervised)~~
- ❑ Classification (supervised)
- ❑ Unsupervised models
- ❑ Learning theory
- ❑ Graphical models

e.g. SUPERVISED LEARNING

- Find function to map input space X to output space Y

$$f : X \longrightarrow Y$$

- So that the difference between *y* and *f(x)* of each example ***x*** is small.

e.g.

x

I believe that this book is not at all helpful since it does not explain thoroughly the material . it just provides the reader with tables and calculations that sometimes are not easily understood …

Input X : e.g. a piece of English text

$\longrightarrow$

y -1

Output Y: {1 / Yes , -1 / No }
e.g. Is this a positive product review

X_1	X_2	X_3	Y

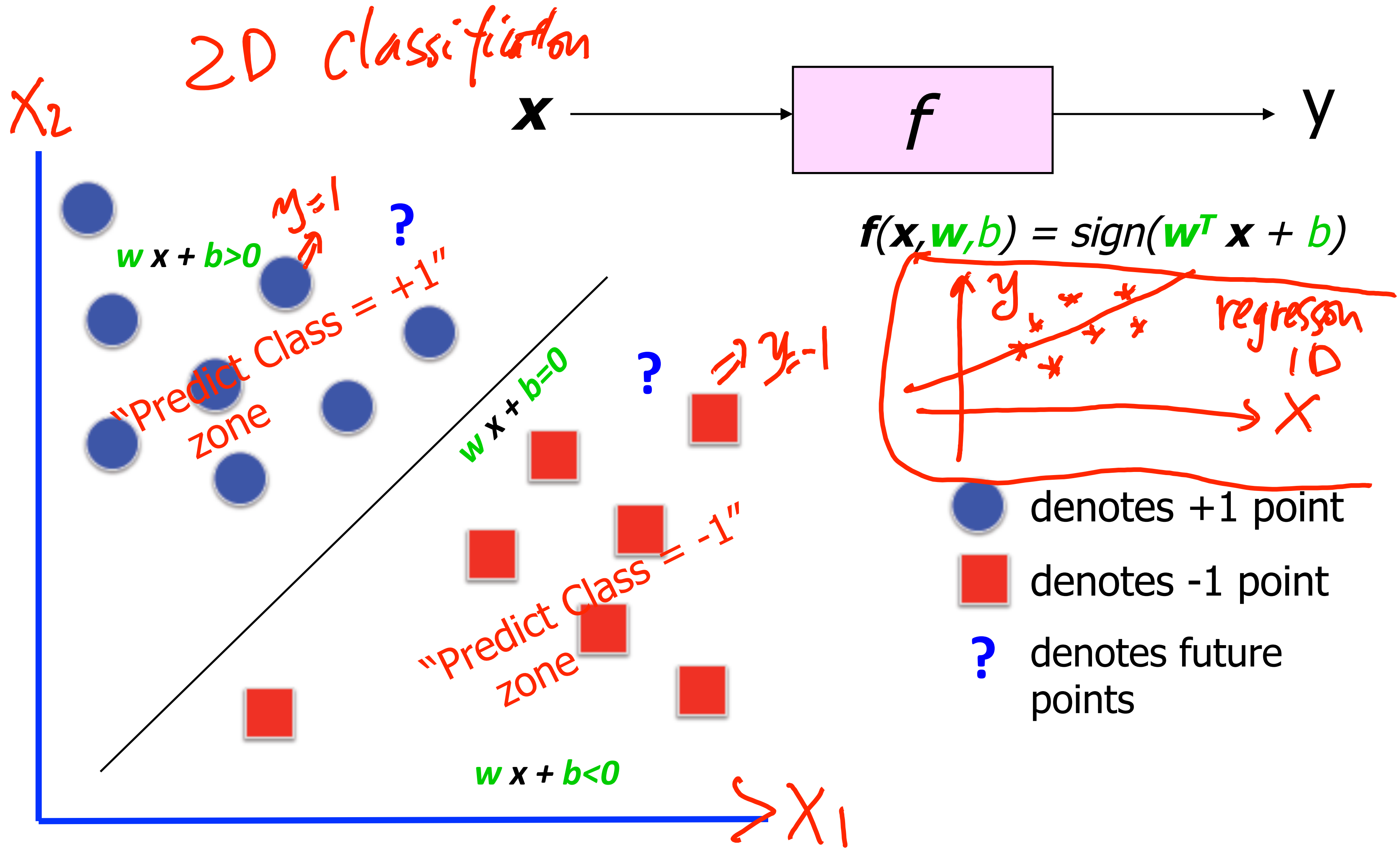
A Dataset for classification

$$f : \boxed{X} \longrightarrow \boxed{Y}$$

Output Class: categorical variable

- **Data**/*points/instances/examples/samples/records*: [rows]
- **Features**/*attributes/dimensions/independent variables/covariates/predictors/regressors*: [columns, except the last]
- **Target**/*outcome/response/label/dependent variable*: special column to be predicted [last column]

e.g. SUPERVISED Linear Binary Classifier

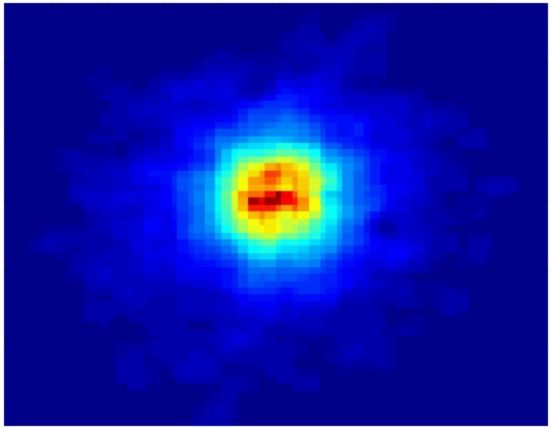


Courtesy slide from Prof. Andrew Moore's tutorial

Application 1: Classifying Galaxies

Courtesy: http://aps.umn.edu

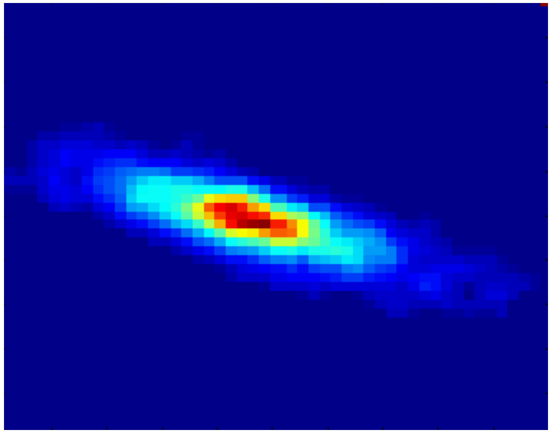
Early



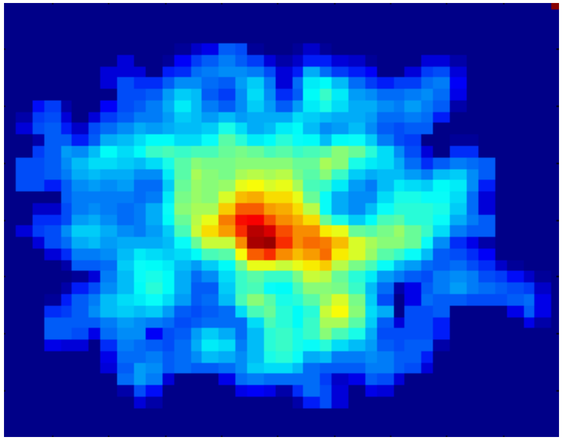
Class:
- Stages of Formation

Attributes:
- Image features,
- Characteristics of light waves received, etc.

Intermediate



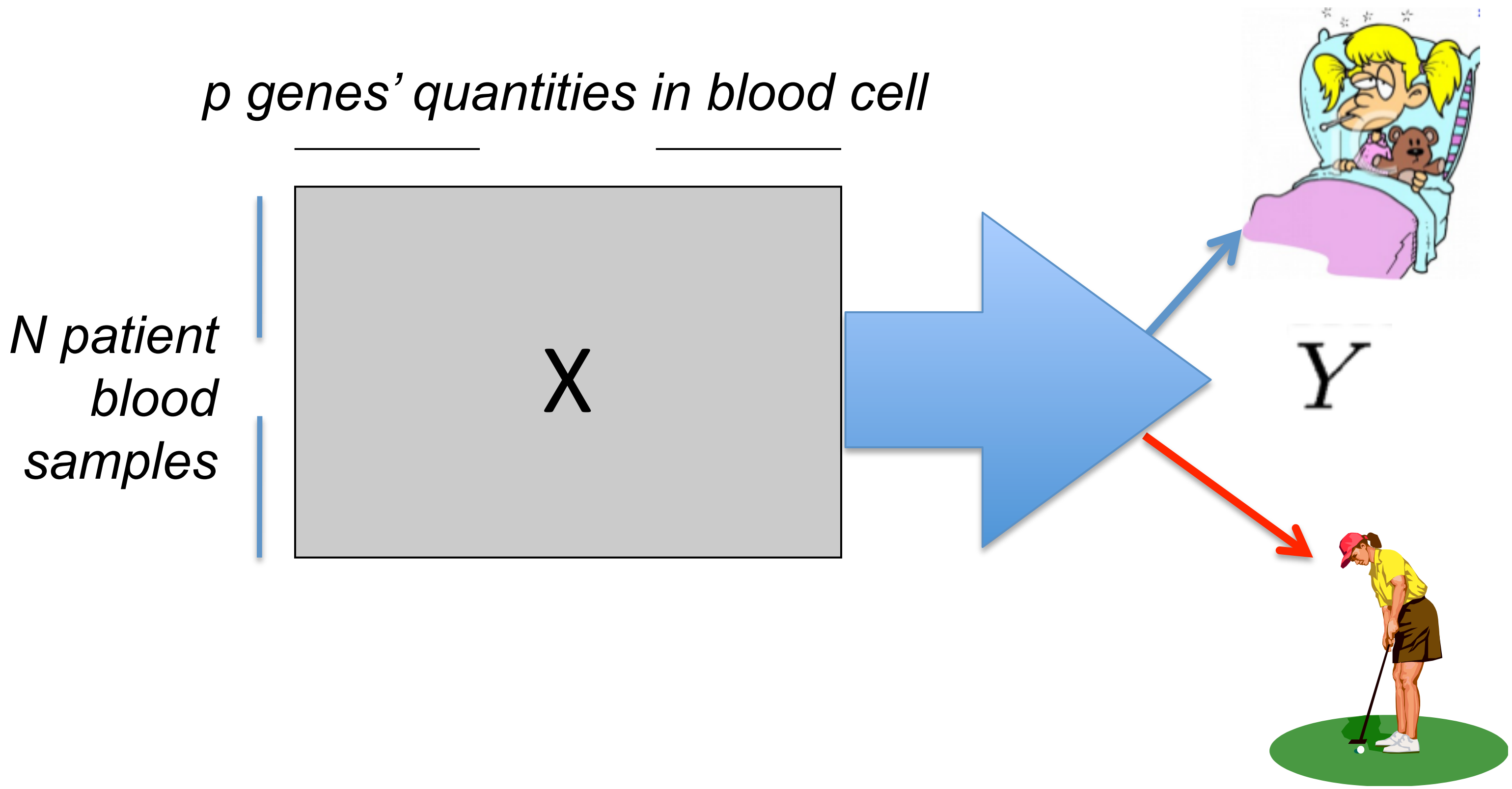
Late



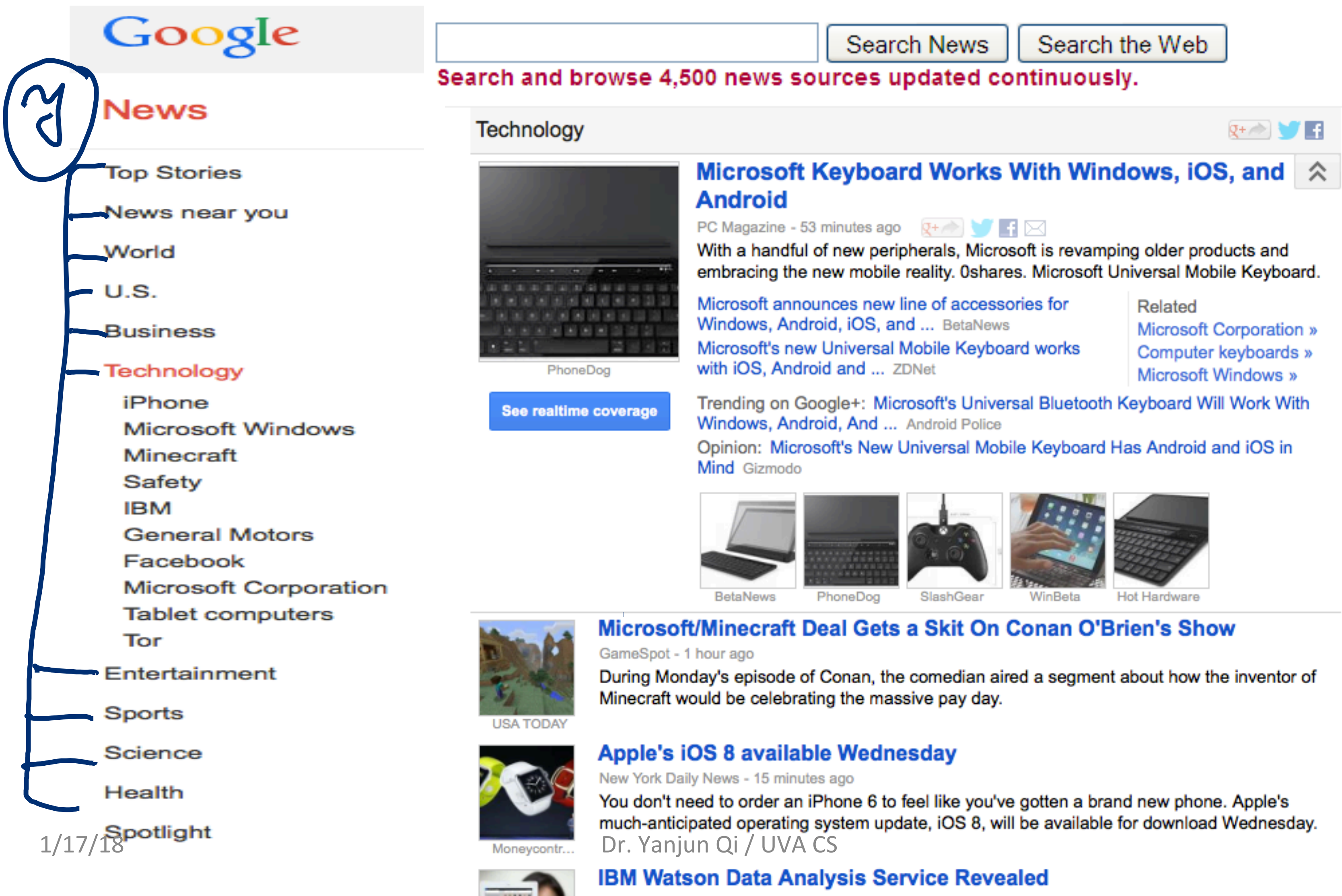
Data Size:
- 72 million stars, 20 million galaxies
- Object Catalog: 9 GB
- Image Database: 150 GB

From [Berry & Linoff] Data Mining Techniques, 1997

Application 2: **Cancer Classification using gene expression**



Application 3: –
Text Documents, e.g. *Google News*



Text Document Representation

- Each document becomes a `term' vector, [word]
  - each term is an (attribute) of the vector,
  - the value of each describes the number of times the corresponding term occurs in the document.

Bag of 'words'

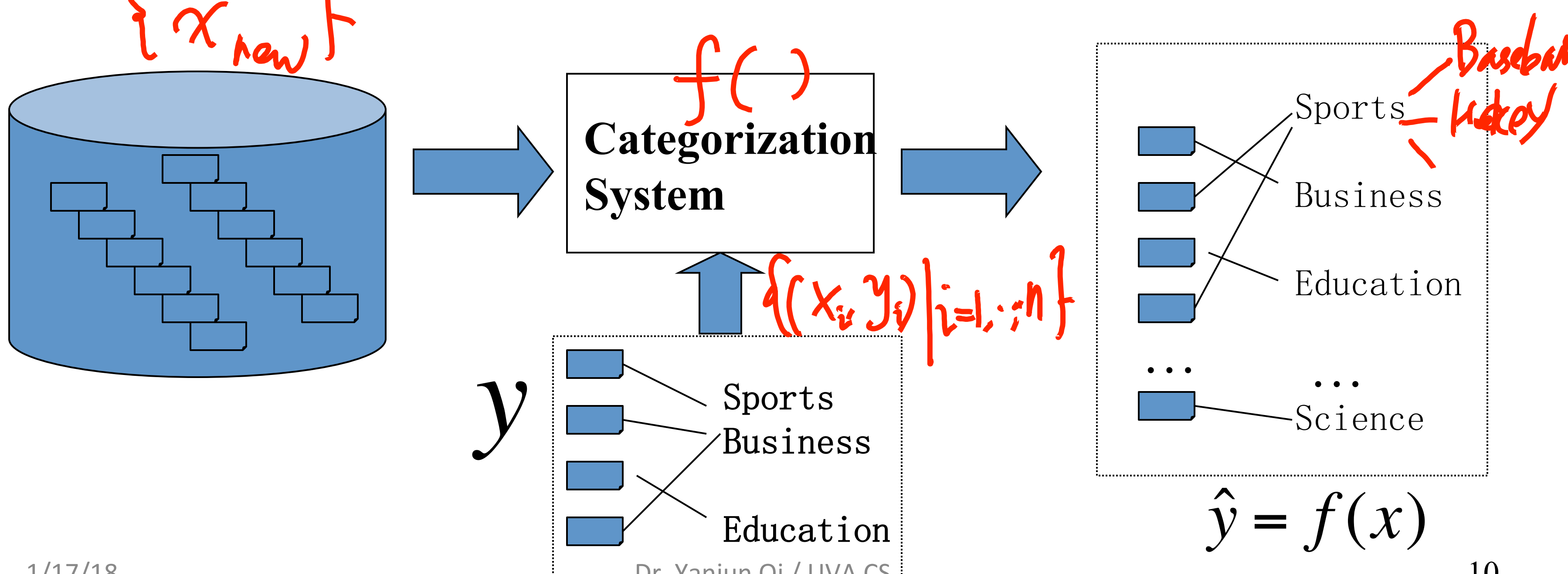
w_1 w_2 . . . . w_p

	team	coach	pla y	ball	score	game	wi n	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Text Categorization

$\{y_{tree}\}$

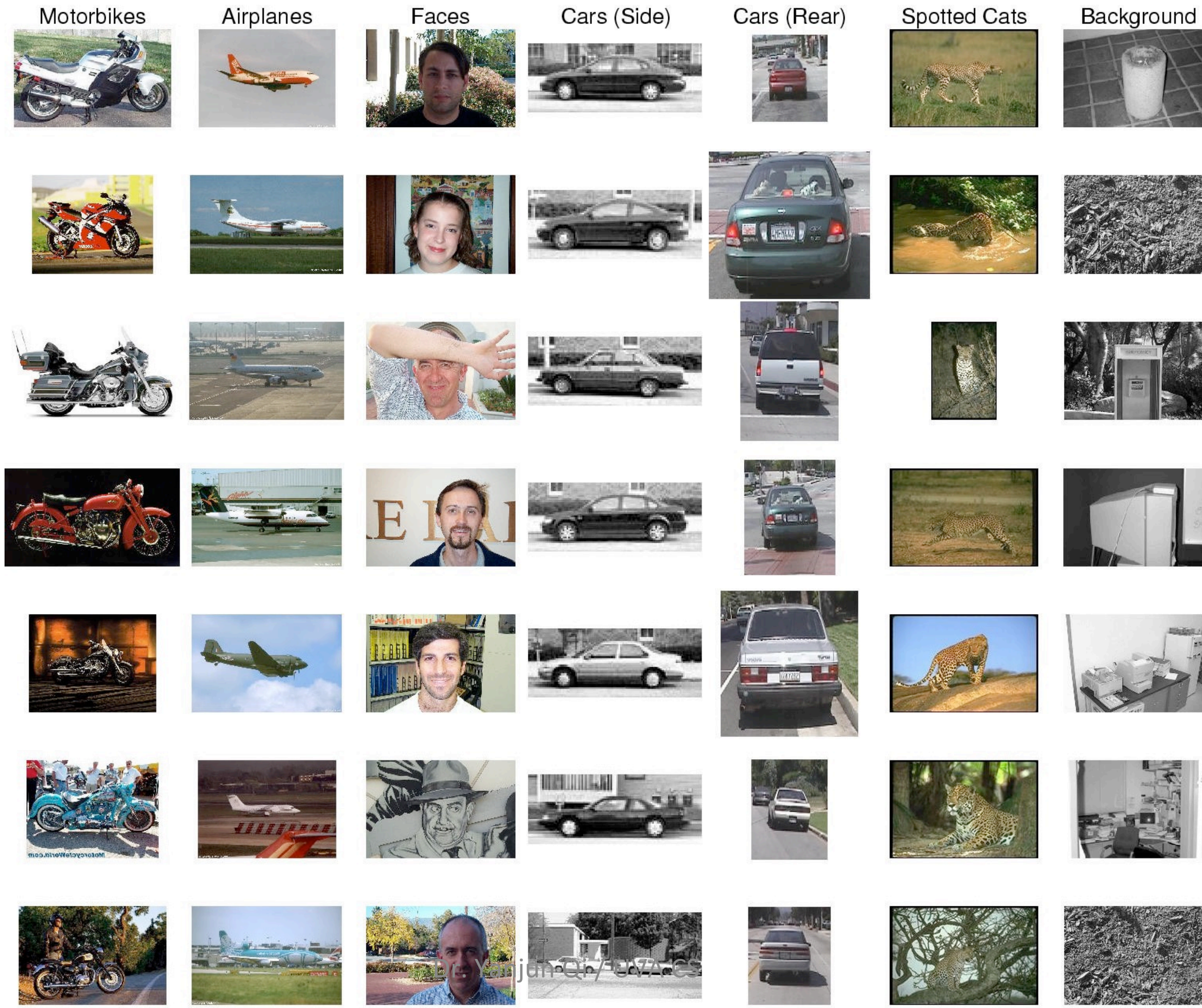
- Pre-given categories and labeled document examples (Categories may form hierarchy)
- Classify new documents
- A standard supervised learning problem



Examples of Text Categorization

- News article classification
- Meta-data annotation
- Automatic Email sorting
- Web page classification

Application 4: – Objective recognition / Image Labeling (Label Images into predefined classes) {y} tree

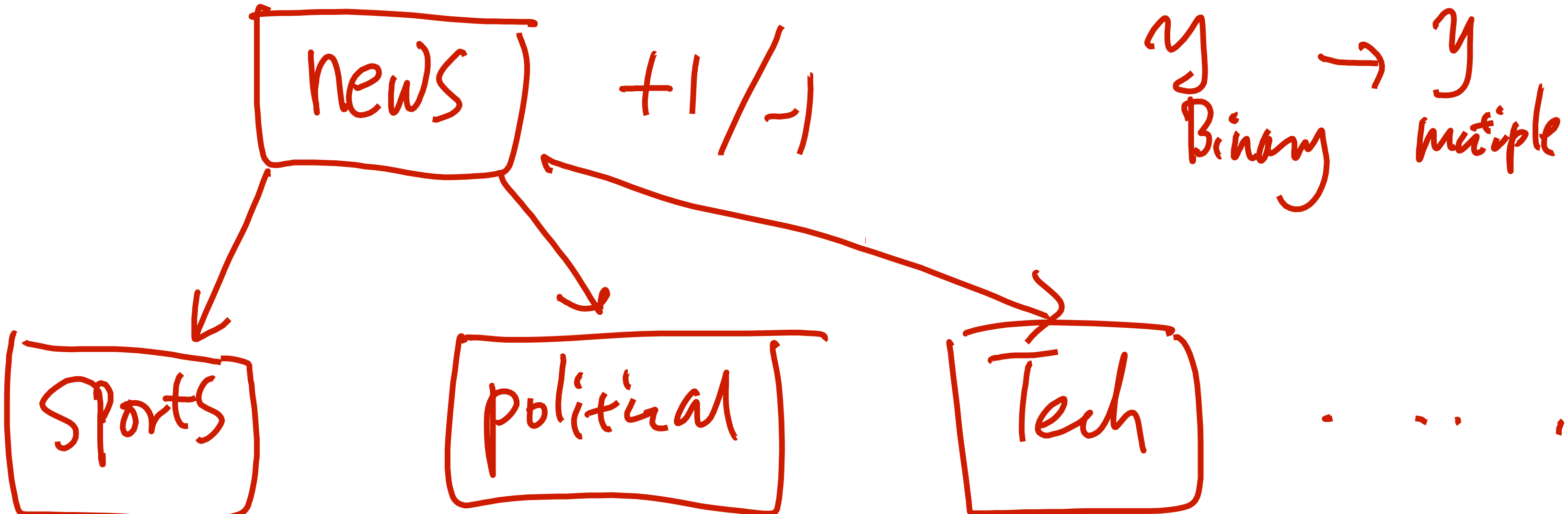


– X

$\{ x, \{ y_{tree} \} \}$

Hierachical supervised classification

$y_{Binary} \rightarrow y_{multiple} \rightarrow y_{tree}$



+1/-1

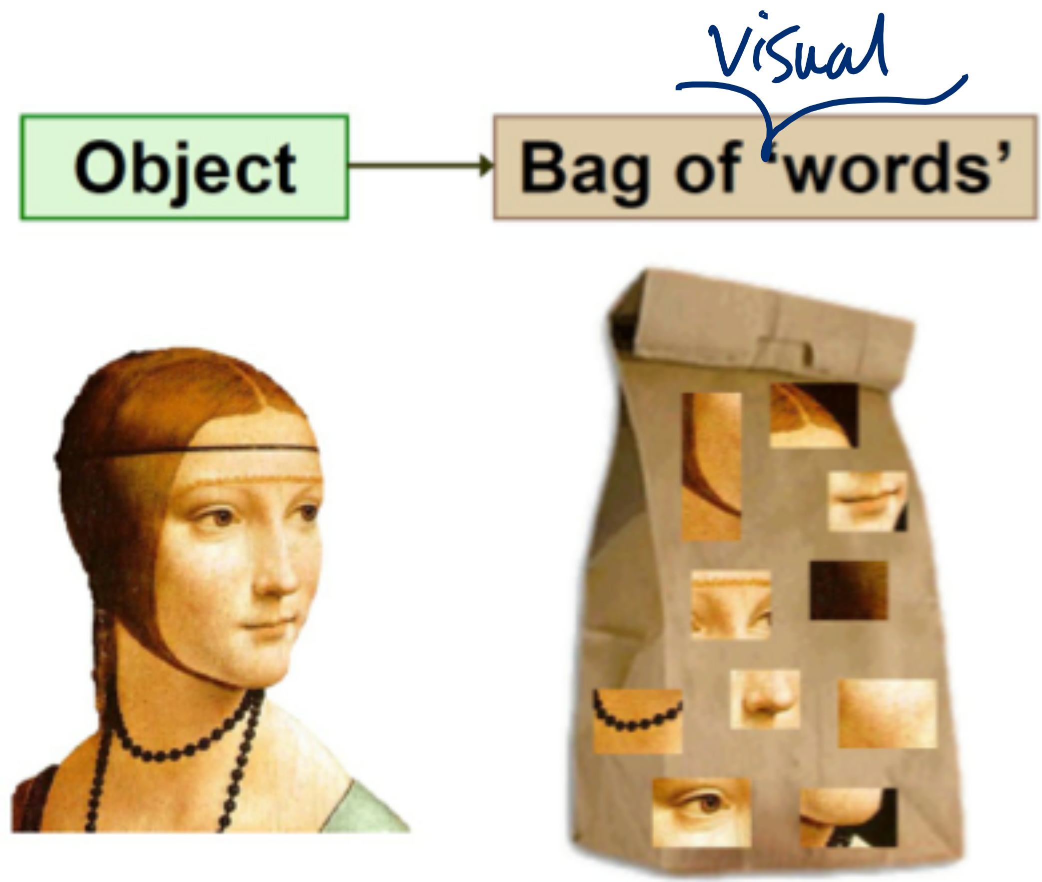
$f_1(\ )$ ① $X \rightarrow$ news $+1/-1$ $\rightarrow$ $f_2(\ )$ Tech $+/-$

② $X \rightarrow$ Tech news $+/1$ $f_3(\ )$

LATER

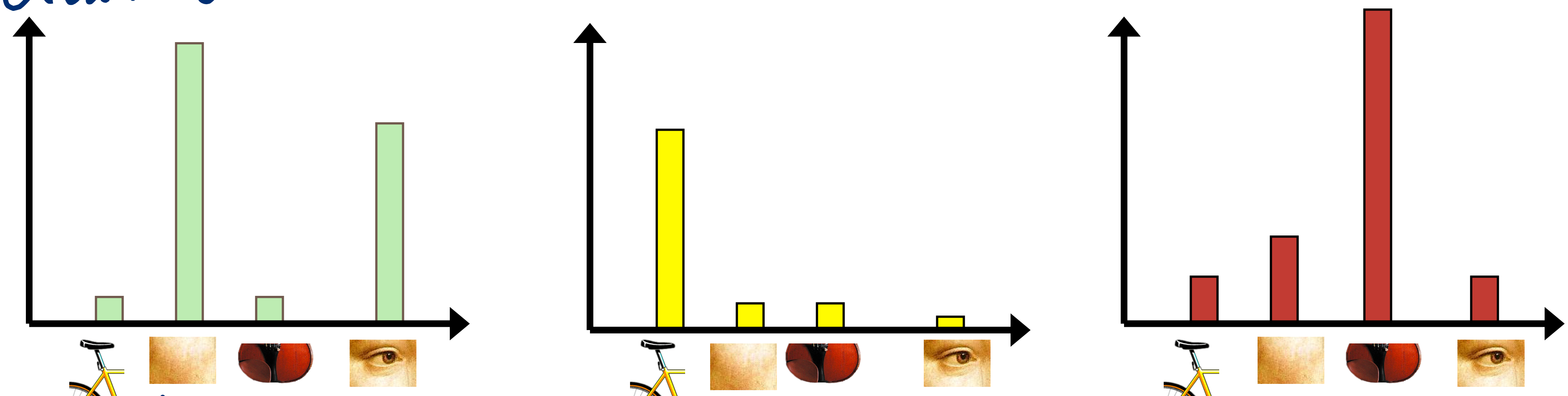
Image Representation for – Objective recognition

- Image representation → bag of "visual words"



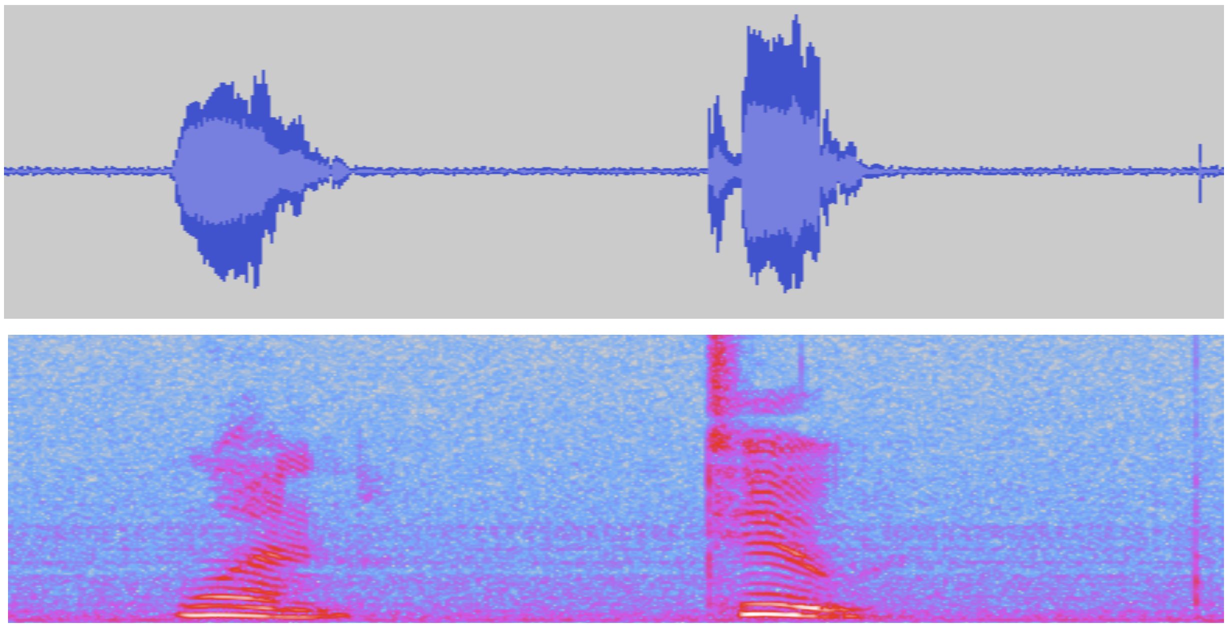
- An object image: histogram of visual vocabulary – a numerical vector of D dimensions.

Occurrence



w_1 w_2 w_3.. w_D

Application 5: – Audio Classification



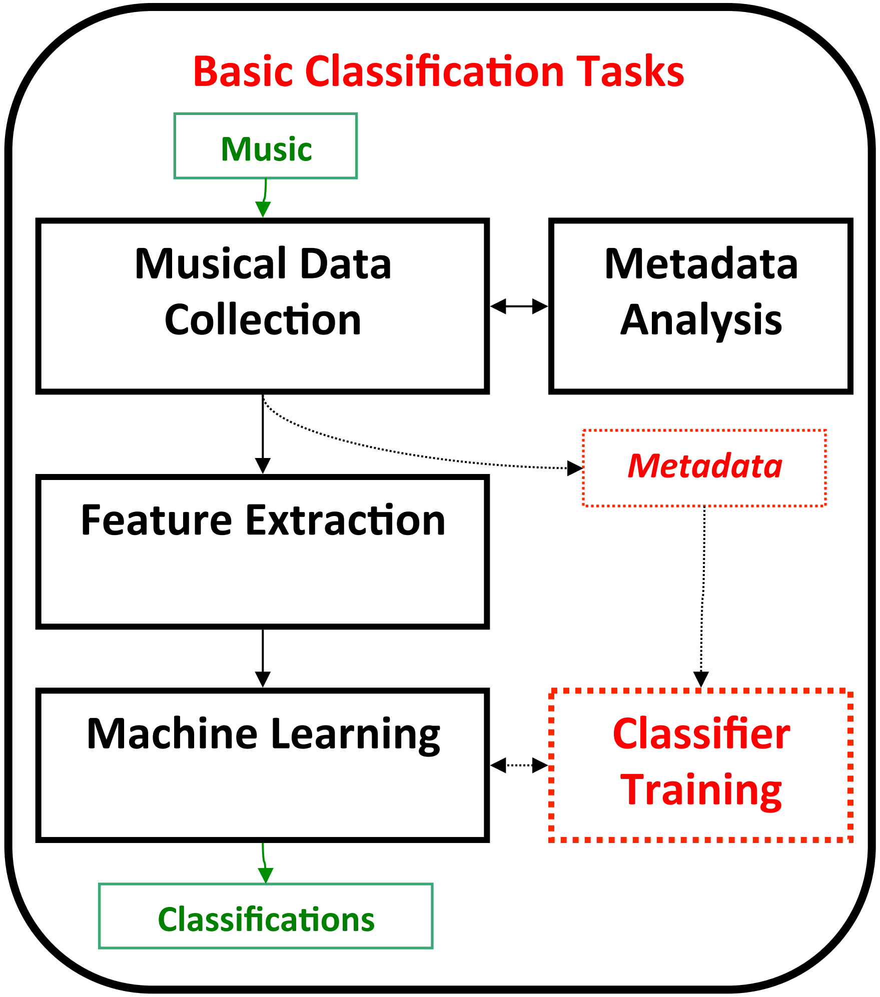
- Real-life applications:
  - Customer service phone routing
  - Voice recognition software

Music Information Retrieval Systems e.g., Automatic Music Classification

- Many areas of research in music information retrieval (MIR) involve using computers to classify music in various ways
  - Genre or style classification
  - Mood classification
  - Performer or composer identification
  - Music recommendation
  - Playlist generation
  - Hit prediction
  - Audio to symbolic transcription
  - etc.
- Such areas often share similar central procedures

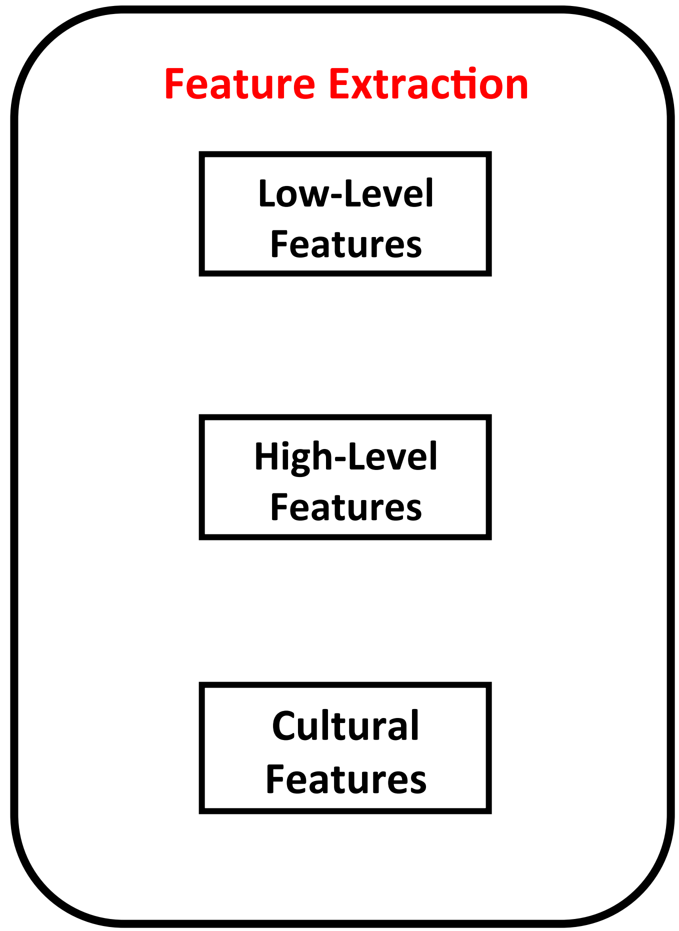
Music Information Retrieval Systems e.g., Automatic Music Classification

- Musical data collection
  - The instances (basic entities) to classify
  - Audio recordings, scores, cultural data, etc.
- Feature extraction
  - Features represent characteristic information about instances
  - Must provide sufficient information to segment instances among classes (categories)
- Machine learning
  - Algorithms ("classifiers" or "learners") learn to associate feature patterns of instances with their classes



Audio, Types of features

- Low-level
  - Associated with signal processing and basic auditory perception
  - e.g. spectral flux or RMS
  - Usually not intuitively musical
- High-level
  - Musical abstractions
  - e.g. meter or pitch class distributions
- Cultural
  - Sociocultural information outside the scope of auditory or musical content
  - e.g. playlist co-occurrence or purchase correlations



Where are we ? ➔ Three major sections for classification

- We can divide the large variety of classification approaches into roughly three major types

1. Discriminative
   - directly estimate a decision rule/boundary
   - e.g., support vector machine, decision tree, logistic regression

2. Generative:
   - build a generative statistical model
   - e.g., Bayesian networks, Naïve Bayes classifier

3. Instance based classifiers
   - Use observation directly (no models)
   - e.g. K nearest neighbors

ICML '06 Proceedings of the 23rd international conference on Machine learning

A study comparing Classifiers

An Empirical Comparison of Supervised Learning Algorithms

Rich Caruana CARUANA@CS.CORNELL.EDU
Alexandru Niculescu-Mizil ALEXN@CS.CORNELL.EDU
Department of Computer Science, Cornell University, Ithaca, NY 14853 USA

Abstract

A number of supervised learning methods have been introduced in the last decade. Unfortunately, the last comprehensive empirical evaluation of supervised learning was the Statlog Project in the early 90's. We present a large-scale empirical comparison between ten supervised learning methods: SVMs, neural nets, logistic regression, naive bayes, memory-based learning, random forests, decision trees, bagged trees, boosted trees, and boosted stumps. We also examine the effect that calibrating the models via Platt Scaling and Isotonic Regression has on their performance. An important aspect of our study is

This paper presents results of a large-scale empirical comparison of ten supervised learning algorithms using eight performance criteria. We evaluate the performance of SVMs, neural nets, logistic regression, naive bayes, memory-based learning, random forests, decision trees, bagged trees, boosted trees, and boosted stumps on eleven binary classification problems using a variety of performance metrics: accuracy, F-score, Lift, ROC Area, average precision, precision/recall break-even point, squared error, and cross-entropy. For each algorithm we examine common variations, and thoroughly explore the space of parameters. For example, we compare ten decision tree styles, neural nets of many sizes, SVMs with many kernels, etc.

Because some of the performance metrics we examine

A study comparing Classifiers
→ 11 binary classification problems

PROBLEM	#ATTR	TRAIN SIZE	TEST SIZE	%POZ
ADULT	14/104	5000	35222	25%
BACT	11/170	5000	34262	69%
COD	15/60	5000	14000	50%
CALHOUS	9	5000	14640	52%
COV_TYPE	54	5000	25000	36%
HS	200	5000	4366	24%
LETTER.P1	16	5000	14000	3%
LETTER.P2	16	5000	14000	53%
MEDIS	63	5000	8199	11%
MG	124	5000	12807	17%
SLAC	59	5000	25000	50%

Ratio among label

A study comparing Classifiers

➔ 11 binary classification problems / 8 metrics

ACC ROC

Table 2. Normalized scores for each learning algorithm by metric (average over eleven problems)

MODEL	CAL	ACC	FSC	LFT	ROC	APR	BEP	RMS	MXE	MEAN	OPT-SEL
BST-DT	PLT	.843*	.779	**.939**	**.963**	**.938**	.929*	**.880**	**.896**	**.896**	**.917**
RF	PLT	.872*	.805	.934*	.957	.931	**.930**	.851	.858	.892	.898
BAG-DT	–	.846	.781	.938*	.962*	.937*	.918	.845	.872	.887*	.899
BST-DT	ISO	.826*	.860*	.929*	.952	.921	.925*	.854	.815	.885	.917*
RF	–	**.872**	.790	.934*	.957	.931	**.930**	.829	.830	.884	.890
BAG-DT	PLT	.841	.774	.938*	.962*	.937*	.918	.836	.852	.882	.895
RF	ISO	.861*	**.861**	.923	.946	.910	.925	.836	.776	.880	.895
BAG-DT	ISO	.826	.843*	.933*	.954	.921	.915	.832	.791	.877	.894
SVM	PLT	.824	.760	.895	.938	.898	.913	.831	.836	.862	.880
ANN	–	.803	.762	.910	.936	.892	.899	.811	.821	.854	.885
SVM	ISO	.813	.836*	.892	.925	.882	.911	.814	.744	.852	.882
ANN	PLT	.815	.748	.910	.936	.892	.899	.783	.785	.846	.875
ANN	ISO	.803	.836	.908	.924	.876	.891	.777	.718	.842	.884
BST-DT	–	.834*	.816	**.939**	**.963**	**.938**	.929*	.598	.605	.828	.851
KNN	PLT	.757	.707	.889	.918	.872	.872	.742	.764	.815	.837
KNN	–	.756	.728	.889	.918	.872	.872	.729	.718	.810	.830
KNN	ISO	.755	.758	.882	.907	.854	.869	.738	.706	.809	.844
BST-STMP	PLT	.724	.651	.876	.908	.853	.845	.716	.754	.791	.808
SVM	–	.817	.804	.895	.938	.899	.913	.514	.467	.781	.810
BST-STMP	ISO	.709	.744	.873	.899	.835	.840	.695	.646	.780	.810
BST-STMP	–	.741	.684	.876	.908	.853	.845	.394	.382	.710	.726
DT	ISO	.648	.654	.818	.838	.756	.778	.590	.589	.709	.774

Ratio of Positive Class (binary case)

- Class imbalance issue num AP << num AN

$$\underbrace{\text{AP}}_{\text{actual positive}} \qquad \underbrace{\text{AN}}_{\text{actual neg}}$$

- Balanced accuracy: $= \frac{1}{2}\left(\frac{TP}{P} + \frac{TN}{N}\right)$

TP+FP TN+FN

Labeled

Confusion Matrix

	actual + (AP)	actual − (AN)
predicted+	TP	FP
predicted−	FN	TN

precision

Recall

TP: true positive

FP: false positive

$$\text{Accuracy} = \frac{\#\,\text{Correct Predicted}}{\#\,\text{all test Examples}}$$

$$= \frac{TP + TN}{TP + FP + TN + FN}$$

$$\left\{ \begin{array}{l} \underline{\text{Precision}}\_\text{Pos} = \dfrac{TP}{P} \\ \underline{\text{Recall}}\text{-Pos} = \dfrac{TP}{TP+FN} \end{array} \right\} \Rightarrow F_1 = \frac{2\,\text{Rec} \times \text{Prec}}{\text{Rec} + \text{Prec}}$$

Ratio of Positive Class (binary case)

If $\frac{\text{Actual P}}{\text{AP} + \text{AN}}$ very small (e.g. < 1%)

(1, 99)
pos neg

⇒ a classifier can predict every example as Neg

⇒ ①

	AP ①	AN ㊾
predict P	0	0
predict N	1	99

⇒ Accuracy = $\frac{99}{100}$ = 0.99

⇒ Balanced Acc =

Bad-neg-classifier

① Balanced Acc $= \frac{1}{2}\left(\frac{TP}{P} + \frac{TN}{N}\right)$

$= \frac{1}{2}\left(\frac{0}{0+\varepsilon} + \frac{99}{100}\right) = 0.495$

$[0, 1]$

another classifier

②

	AP	AN
PP	1	0
PN	0	99

Balanced Acc $= \frac{1}{2}\left(\frac{1}{1} + \frac{99}{99}\right) = 1$

$\text{Acc} = \frac{1+99}{1+0+99+0} = 1$