Information mining

Taxonomies, classification, detection and linking

Web Search

Summary

- Introduction
- From data to information: Taxonomies and classes
- Classification, detection and linking

Importance of information mining

P. Jackson and I. Moulinier. 2002. *Natural Language Processing for Online Applications*

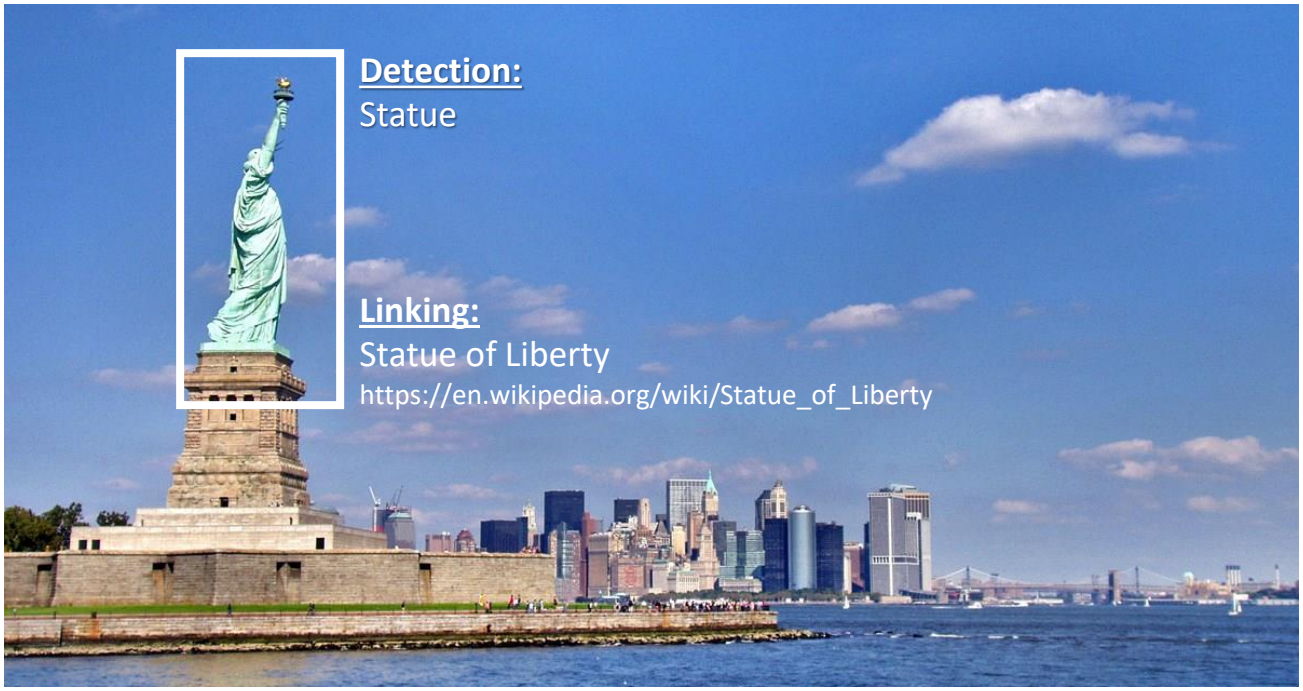
- "There is no question concerning the commercial value of being able to classify documents automatically by content. There are myriad potential applications of such a capability for corporate intranets, government departments, and Internet publishers"

- "Understanding the data is one of the keys to successful categorization, yet this is an area in which most categorization tool vendors are extremely weak. Many of the 'one size fits all' tools on the market have not been tested on a wide range of content types."

Real world tasks

- SPAM detection (fake opinions)
- Memes detection (not informational)
- Tampered images
- Sentiment detection (opinions)
- Emergency detection

Classification, detection, linking



Classification:
Sea side
Statue
City
Sky

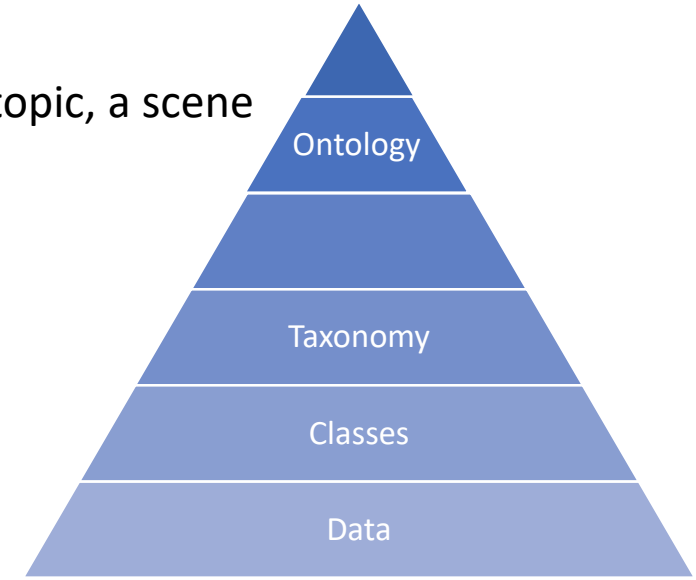
From data to information

Web Search

From data to information

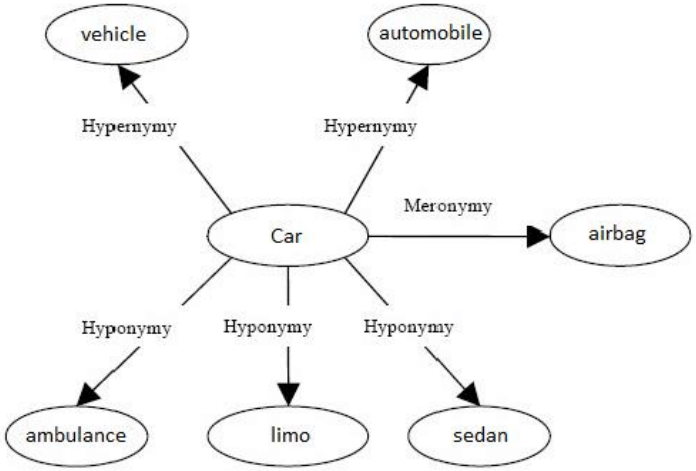
- A taxonomy is concerned with classifying and organizing hierarchically concepts of a specific domain.

- It is important to identify the list of items that need to be detected.
  - These items are domain specific, and can be a topic, a scene type, a visual object or a named entity.
  - They are normally associated to a class in a supervised learning task.



WordNet: A lexical database

"WordNet® is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. "



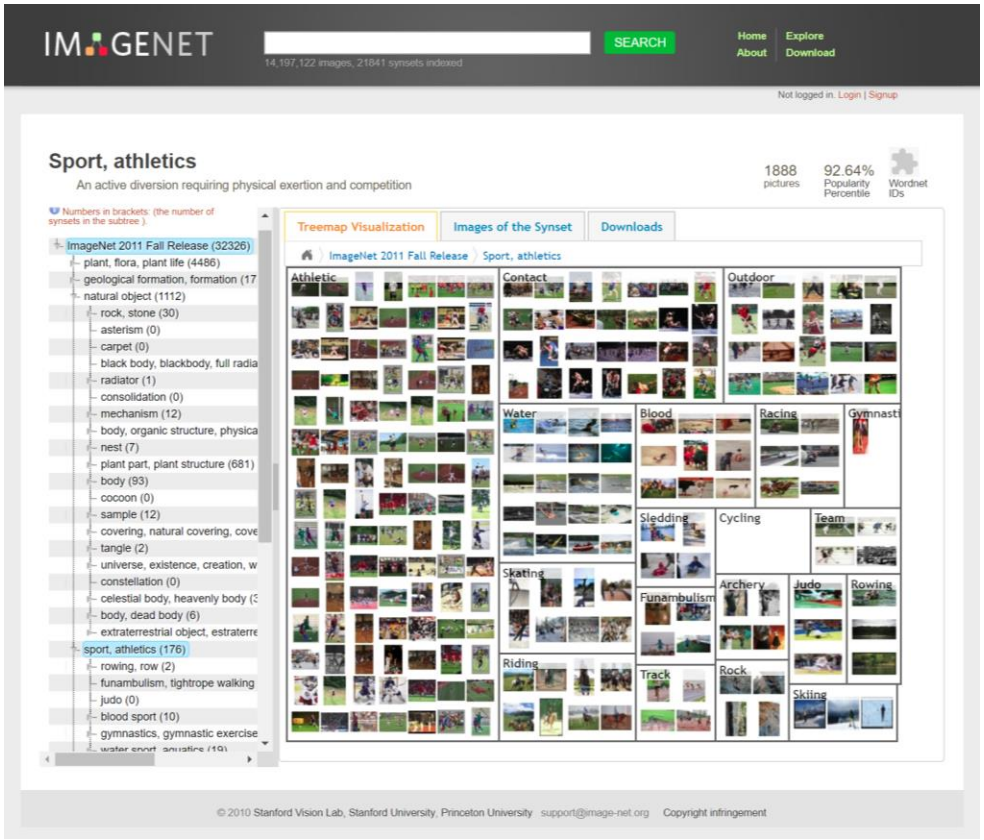
"WordNet interlinks specific senses of words. As a result, words that are found in close proximity to one another in the network are semantically disambiguated. Second, WordNet labels the semantic relations among words, whereas the groupings of words in a thesaurus does not follow any explicit pattern other than meaning similarity."

https://wordnet.princeton.edu/

ImageNet: A visual taxonomy

- Selected words of WordNet are illustrated in ImageNet.
- Currently, there are over 14.000 concepts illustrated.
- Roughly 1000 concepts are used by VOC.
- Great impact in advancing the state of the art.

http://image-net.org/explore.php

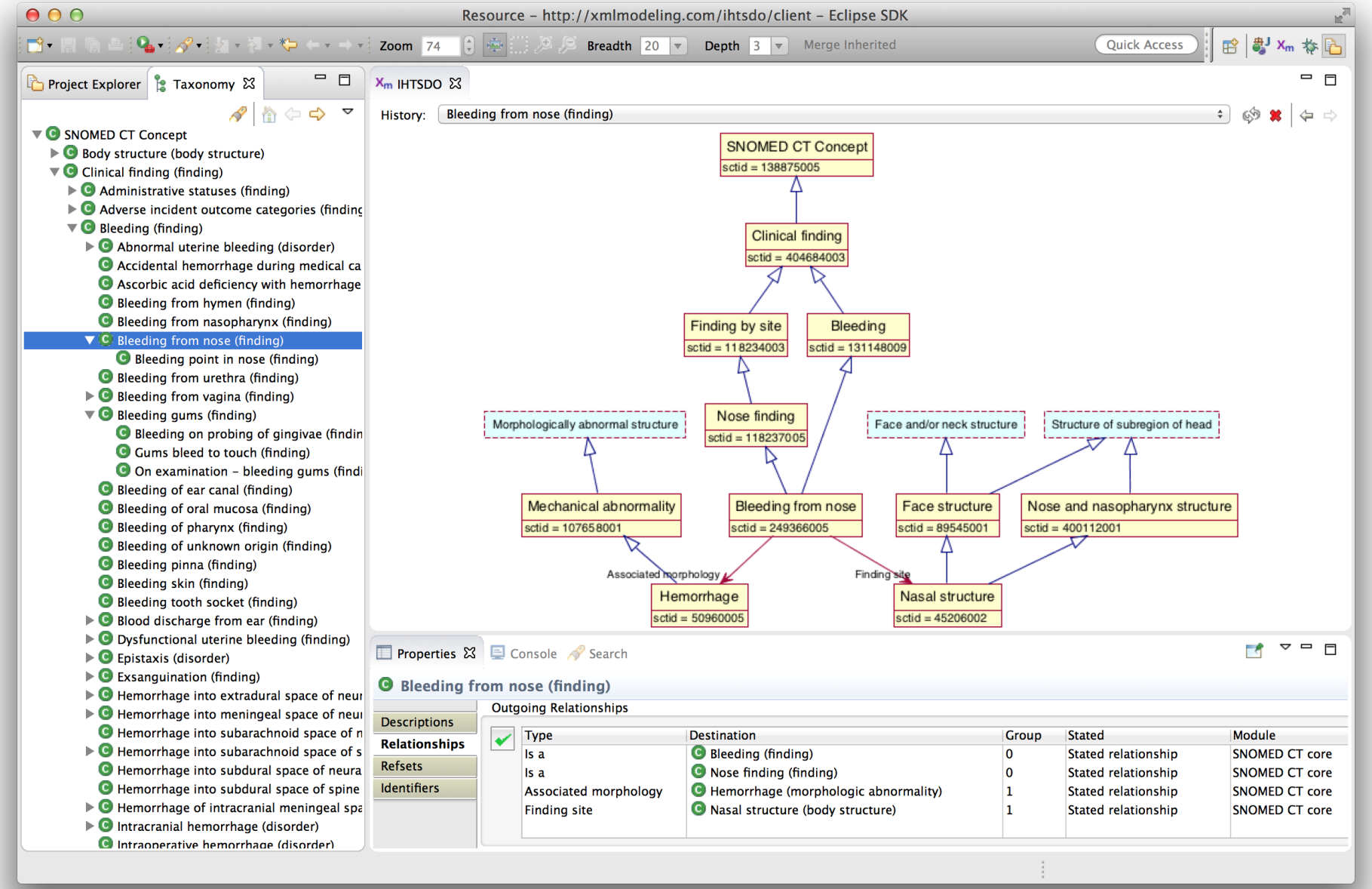


Domain specific taxonomies

- Domain specific terminologies are curated by domain experts and are designed with specific tasks and workflows in mind.

- In the medical domain, the SNOMED-CT is intended to describe medical conditions, procedures, admin, etc.
  - http://browser.ihtsdotools.org/

- In the computer science domain the ACM Computing Classification Scheme is widely used to classify published articles.
  - https://dl.acm.org/ccs/ccs.cfm


Resource – http://xmlmodeling.com/ihtsdo/client – Eclipse SDK
Zoom 74
Breadth 20
Depth 3
Merge Inherited
Quick Access
Project Explorer
Taxonomy
IHTSDO
History: Bleeding from nose (finding)
SNOMED CT Concept
Body structure (body structure)
Clinical finding (finding)
Administrative statuses (finding)
Adverse incident outcome categories (finding
Bleeding (finding)
Abnormal uterine bleeding (disorder)
Accidental hemorrhage during medical ca
Ascorbic acid deficiency with hemorrhage
Bleeding from hymen (finding)
Bleeding from nasopharynx (finding)
Bleeding from nose (finding)
Bleeding point in nose (finding)
Bleeding from urethra (finding)
Bleeding from vagina (finding)
Bleeding gums (finding)
Bleeding on probing of gingivae (findin
Gums bleed to touch (finding)
On examination – bleeding gums (findi
Bleeding of ear canal (finding)
Bleeding of oral mucosa (finding)
Bleeding of pharynx (finding)
Bleeding of unknown origin (finding)
Bleeding pinna (finding)
Bleeding skin (finding)
Bleeding tooth socket (finding)
Blood discharge from ear (finding)
Dysfunctional uterine bleeding (finding)
Epistaxis (disorder)
Exsanguination (finding)
Hemorrhage into extradural space of neur
Hemorrhage into meningeal space of neur
Hemorrhage into subarachnoid space of n
Hemorrhage into subarachnoid space of s
Hemorrhage into subdural space of neura
Hemorrhage into subdural space of spine
Hemorrhage of intracranial meningeal spa
Intracranial hemorrhage (disorder)
Intraoperative hemorrhage (disorder)
SNOMED CT Concept
sctid = 138875005
Clinical finding
sctid = 404684003
Finding by site
sctid = 118234003
Bleeding
sctid = 131148009
Morphologically abnormal structure
Nose finding
sctid = 118237005
Face and/or neck structure
Structure of subregion of head
Mechanical abnormality
sctid = 107658001
Bleeding from nose
sctid = 249366005
Face structure
sctid = 89545001
Nose and nasopharynx structure
sctid = 400112001
Associated morphology
Finding site
Hemorrhage
sctid = 50960005
Nasal structure
sctid = 45206002
Properties
Console
Search
Bleeding from nose (finding)
Descriptions
Relationships
Refsets
Identifiers
Outgoing Relationships
Type
Destination
Group
Stated
Module
Is a
Bleeding (finding)
0
Stated relationship
SNOMED CT core
Is a
Nose finding (finding)
0
Stated relationship
SNOMED CT core
Associated morphology
Hemorrhage (morphologic abnormality)
1
Stated relationship
SNOMED CT core
Finding site
Nasal structure (body structure)
1
Stated relationship
SNOMED CT core

Wikipedia as a database

- Wikipedia contains large amounts of information largely unstructured but structured as a taxonomy.
- **DBPedia** aims to create a rigorous database out of Wikipedia.
- A key application is to link data to Wikipedia entries.

https://en.wikipedia.org/wiki/Portal:Contents

Which and how many are detectable?

- An important question to ask is which and how many items of the taxonomy are detectable in data?
- A few (well separated ones)? -> Easy!
- A zillion closely related ones? -> Not so easy…
  - Think: Yahoo! Directory, Library of Congress classification, legal applications
  - Quickly gets difficult!
    - Classifier combination is always a useful technique
      - Voting, bagging, or boosting multiple classifiers
    - Much literature on hierarchical classification
      - Definitely helps for scalability, even if not in accuracy
    - May need a hybrid automatic/manual solution

Taxonomies and classification

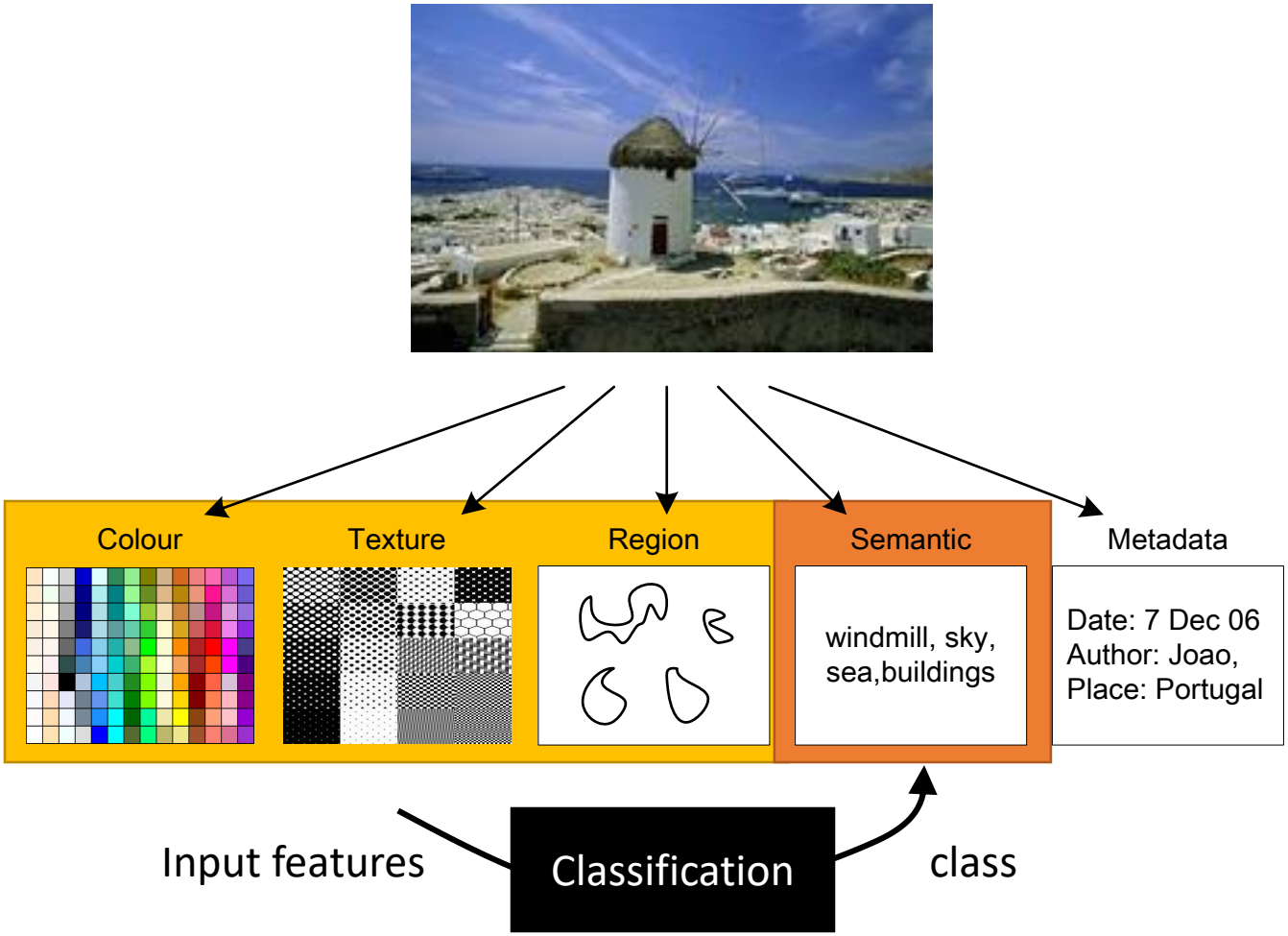
- In practice, only a few elements of the taxonomy should be used as classes for classification
  - Only the ones offering a stable document class representation.
- The ultimate goal is to link information to an entry on a taxonomy capturing the target domain.
- Ultimately more complete domain representation should be used, e.g. an ontology.

Classification

Web Search

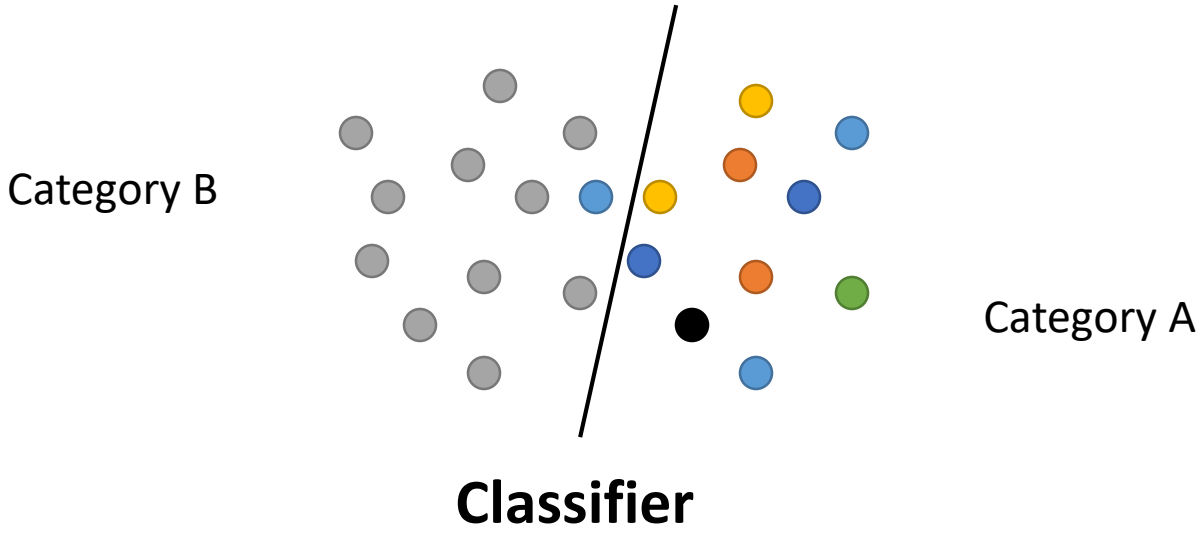
Section 12.2

Document classification



Classification task

- For new unseen documents, we wish to classify documents with one of the known classes.
- New documents are represented in some feature space and then a machine learning algorithm classifies the new documents.

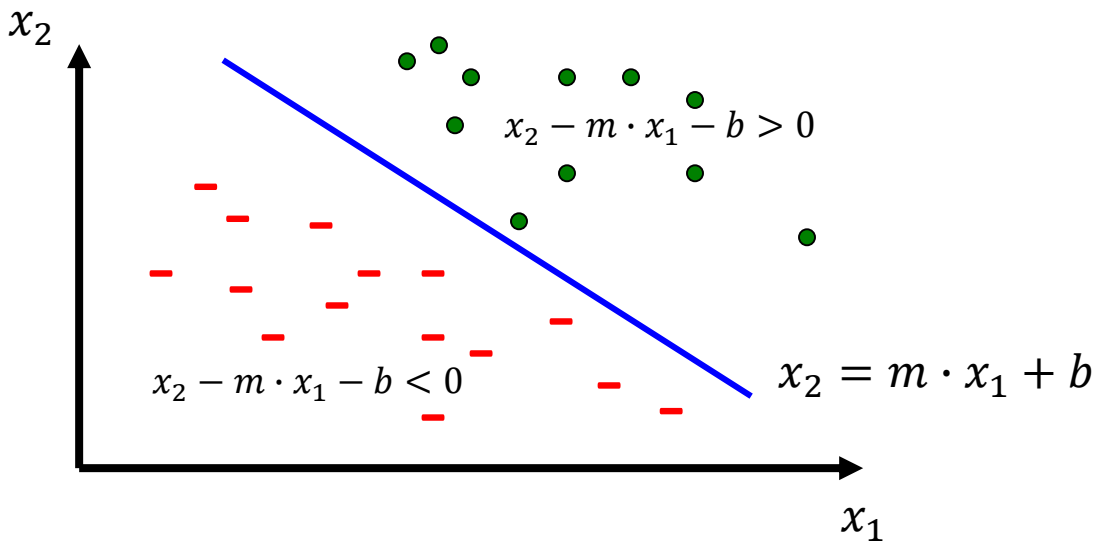


Perceptron

- All sample vectors $\boldsymbol{x^{(j)}}$ have their corresponding label $\boldsymbol{y^{(j)} = \{+1, -1\}}$

- **The perceptron performs a binary prediction $\hat{y}$ based on the observed data x :**

$$\hat{y} = f(x) = \begin{cases} +1 & , if\ x_2 - m \cdot x_1 - b \geq 0 \\ -1 & , if\ x_2 - m \cdot x_1 - b < 0 \end{cases}$$

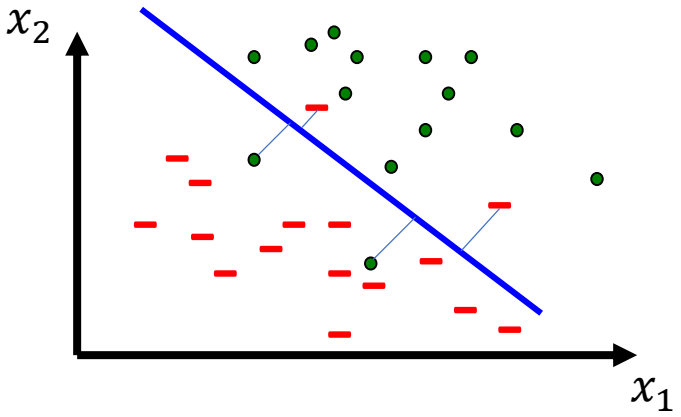


Model error

- The Mean Square Error (MSE) measures the error between the true labels and the predicted labels

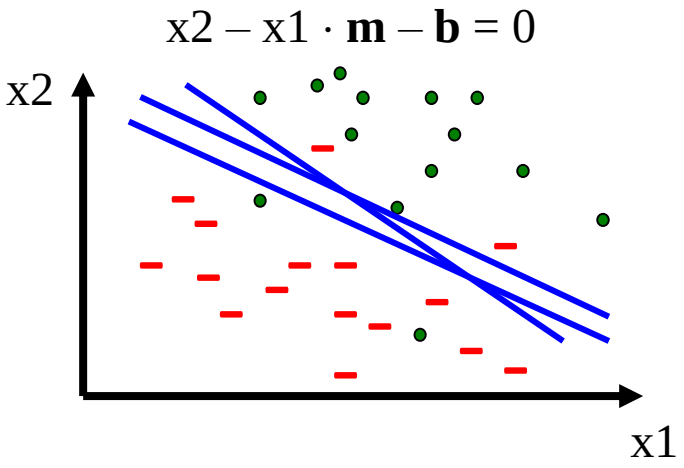
$$MSE = \frac{1}{TotalSamples} \sum_{i}^{TotalSamples} (label_i - predictedLabel_i)^2$$

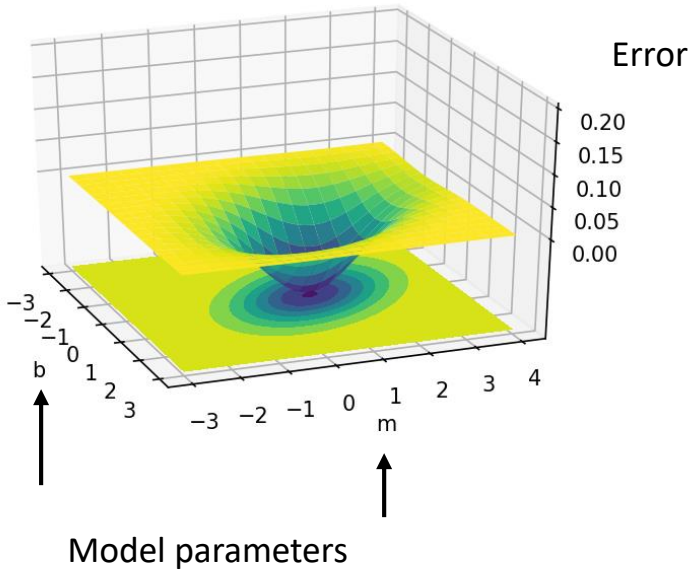
$$x_2 - m \cdot x_1 - b = 0$$



Minimizing the error

$$MeanSquareError = \frac{1}{TotalSamples} \sum_{i}^{TotalSamples} (label_i - predictedLabel_i)^2$$





Learning to minimize the model error

- Initialize the model with random weights
- Compute the model predictions
- Compute the error of each prediction
- Update the model with the samples incorrectly classified.

Observation	Prediction	Error	Update
-1	-1	0	0
-1	+1	-1	-1*x
+1	-1	+1	+1*x
+1	+1	0	0

Learning algorithm

```
[ ]: b=0
     m=0
     model = [m,b]

     max_iters = 30
     mean_square_error = []
     for iter in range(0,max_iters):

         # Compute the model predictions
         predicted_labels = ((observations_x2 - m*observations_x1 - b ) >= 0)*2-1

         # Compute the model error
         error_of_all_samples = (true_labels-predicted_labels)/2

         # Update the model parameters
         update_m = np.mean(error_of_all_samples*observations_x1)
         update_b = np.mean(error_of_all_samples)

         m = m - update_m*0.1
         b = b - update_b*0.1
```

$$\hat{y} = f(x) = \begin{cases} +1 & ,if\ x_2 - m \cdot x_1 - b \geq 0 \\ -1 & ,if\ x_2 - m \cdot x_1 - b < 0 \end{cases}$$

$$error = (y - \hat{y})/2 = \begin{cases} +1 \\ 0 \\ -1 \end{cases}$$

$$update_m = error \cdot x_1$$

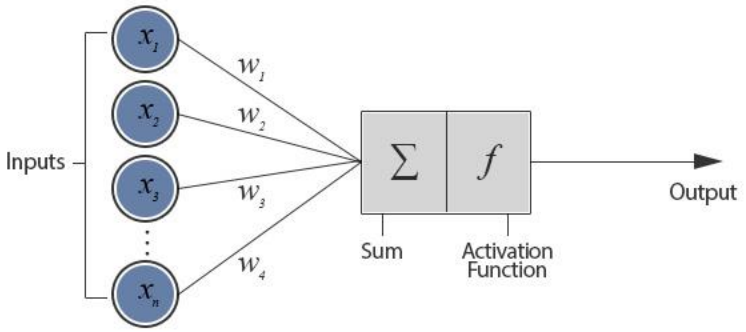
$$m = m - update_m \cdot learning_{rate}$$

Perceptron: general formulation

- **Binary classification:**

$$z = w_0 + w_1 x_1 + \ \ldots + w_n x_n$$

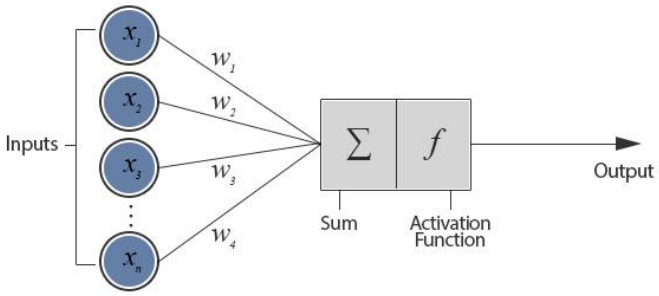
$$\hat{y} = f(z) = \begin{cases} +1 & , if\ z \geq 0 \\ -1 & , if\ z < 0 \end{cases}$$



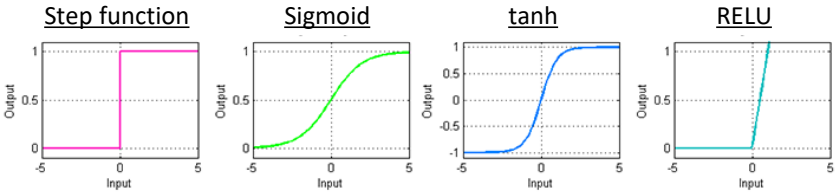
- **Input:** Vectors $\boldsymbol{x^{(j)}}$ and labels $\boldsymbol{y^{(j)}}$
  - Vectors $\boldsymbol{x^{(j)}}$ are real valued where $\|\boldsymbol{x}\|_2 = \mathbf{1}$
- **Goal:** Find vector $\boldsymbol{w = (w_1, w_2, \ldots, w_d)}$
  - Each $\boldsymbol{w_i}$ is a real number

Activation functions

- The perceptron was initially proposed with the step function.
- Historically, other activation functions have been studied.
- It can be shown that the perceptron with the sigmoid activation function corresponds to the logistic regression model.



Activation functions



Note regarding model training

- Robustly training a model for Web data is a complex task.
- In most of the cases, we will use pre-trained models.
- These models were trained on large-scale data.
- These pre-trained models are robust and reliable.

Per-class evaluation measures

		Ground-truth	
		True	False
Method	True	True positive	False positive
	False	False negative	True negative

- **Recall**: Fraction of docs in class i classified correctly:

$$Recall = \frac{truePos}{truePos + falseNeg}$$

- **Precision**: Fraction of docs assigned class i that are actually about class i:

$$Precision = \frac{truePos}{truePos + falsePos}$$

- **Accuracy**: Fraction of docs classified correctly:

$$Accuracy = \frac{truePos + trueNeg}{truePos + falsePos + trueNeg + falseNeg}$$

* precisão, abragência e exatidão.

Micro- vs. Macro-Averaging

- If we have more than one class, how do we combine multiple performance measures into one quantity?
- **Macroaveraging**: Compute performance for each class, then average.
- **Microaveraging**: Collect decisions for all classes, compute contingency table, evaluate.

Micro- vs. Macro-Averaging: Example

Class 1

	Truth: yes	Truth: no
Classifier: yes	10	10
Classifier: no	10	970

Class 2

	Truth: yes	Truth: no
Classifier: yes	90	10
Classifier: no	10	890

Micro Ave. Table

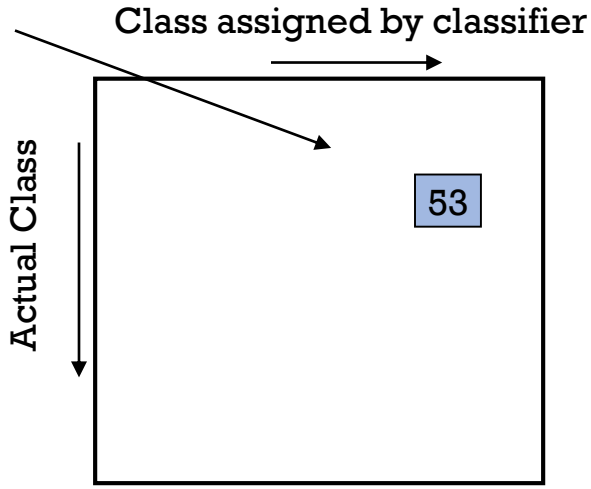
	Truth: yes	Truth: no
Classifier: yes	100	20
Classifier: no	20	1860

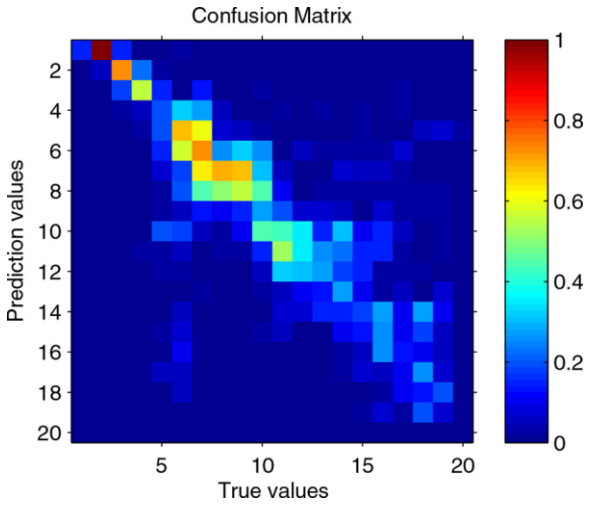
- Macroaveraged precision: (0.5 + 0.9)/2 = 0.7
- Microaveraged precision: 100/120 = .83

- Microaveraged score is dominated by score on common classes

Good practice: Make a confusion matrix

- This (i, j) entry means 53 of the docs actually in class i were put in class j by the classifier.

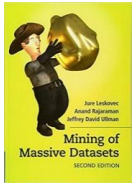




- In a perfect classification, only the diagonal has non-zero entries
- Look at common confusions and how they might be addressed

Detection

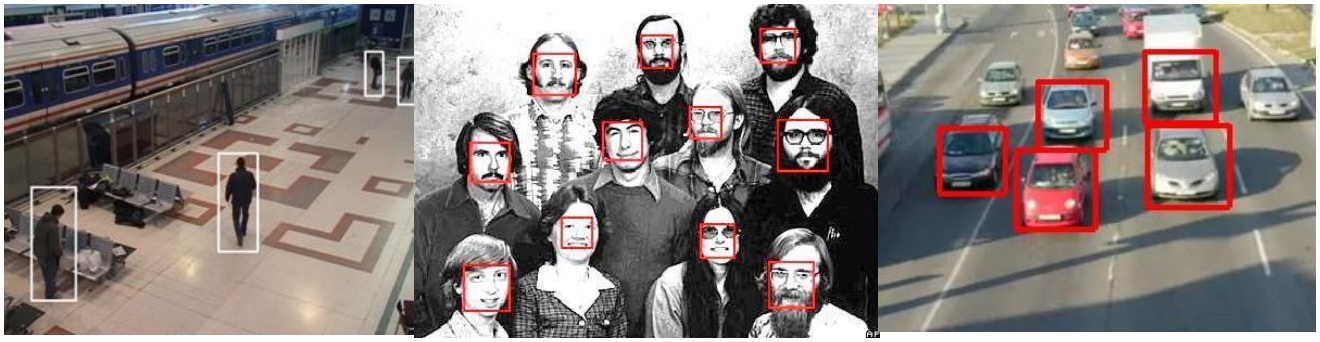
Web Search



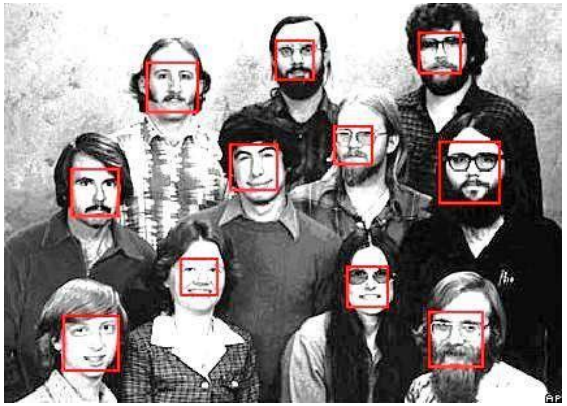
Section 12.2

Detection

- How to detect a face, a person or a car in a picture?
- How to find pictures of BigBen or the Eiffel Tower?



Face detection



- As done previously for classification...
  - Positive and negative examples must be gathered
  - Features must be computed for each image
  - A classifier must be estimated

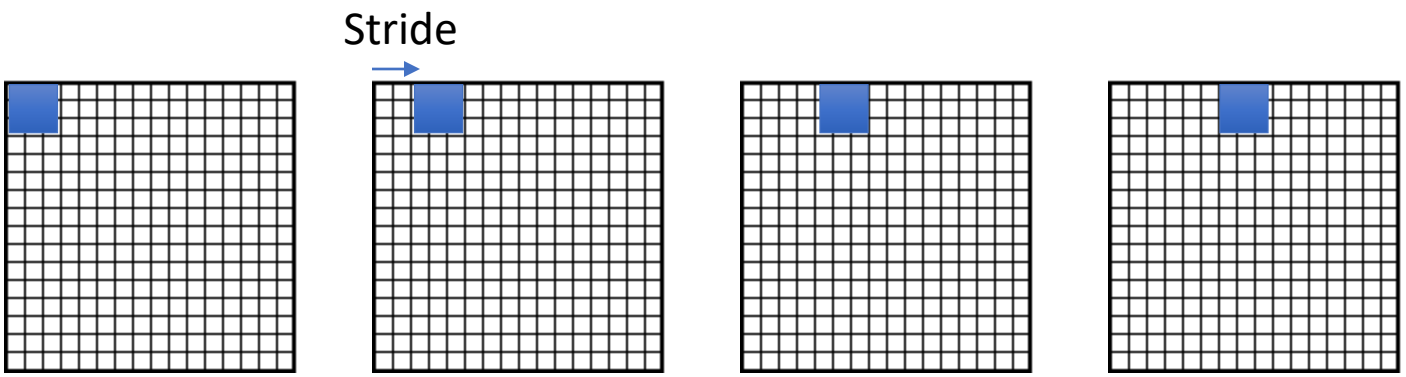
- Positive examples should cover a wide variation of poses, illuminations, instances, etc.

- Challenges:
  - Where is the face?
  - What's the face size?
  - Given an image patch, how to classify the patch as a face or not?

Image scanning

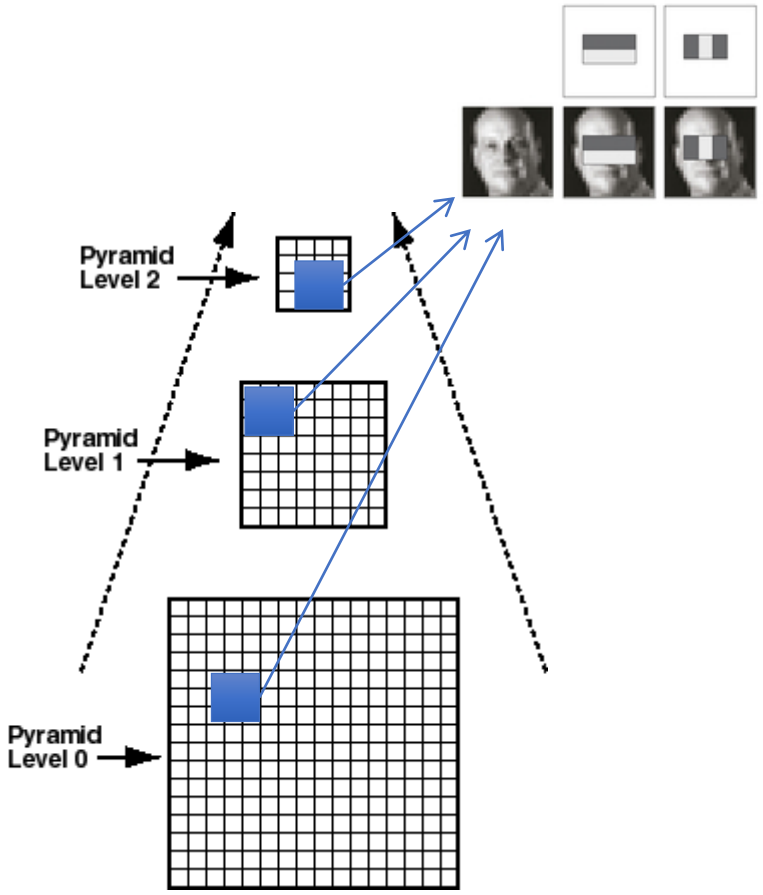
- Image is scanned for faces in all possible positions.
  - Step between different positions can be more than 1.

```
for (x=1; x < N; x += stride_x)
        for (y=1; y < M; y += stride_y)
                test_face(x, y, img);
```



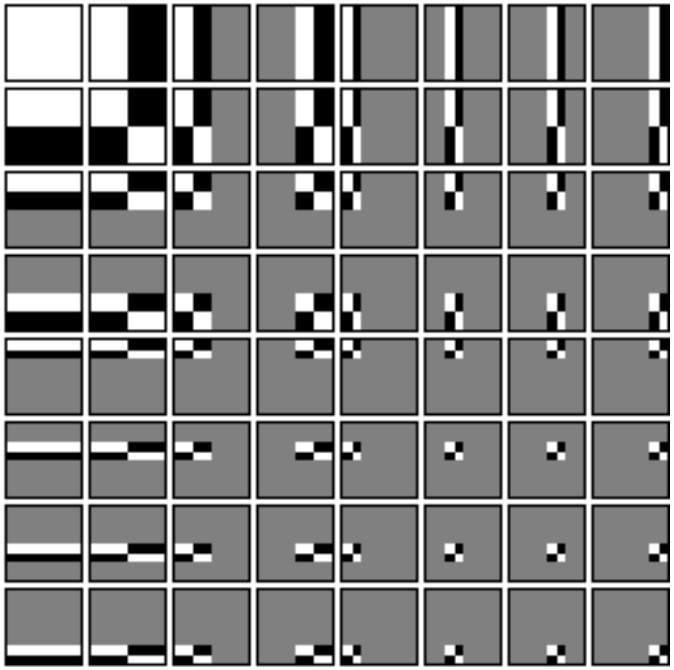
Pyramid multi-scaling

- To solve the problem of finding faces at multiple scales image is scalled multiple times
- At each scale the described process is done to find faces at different positions and at different scales



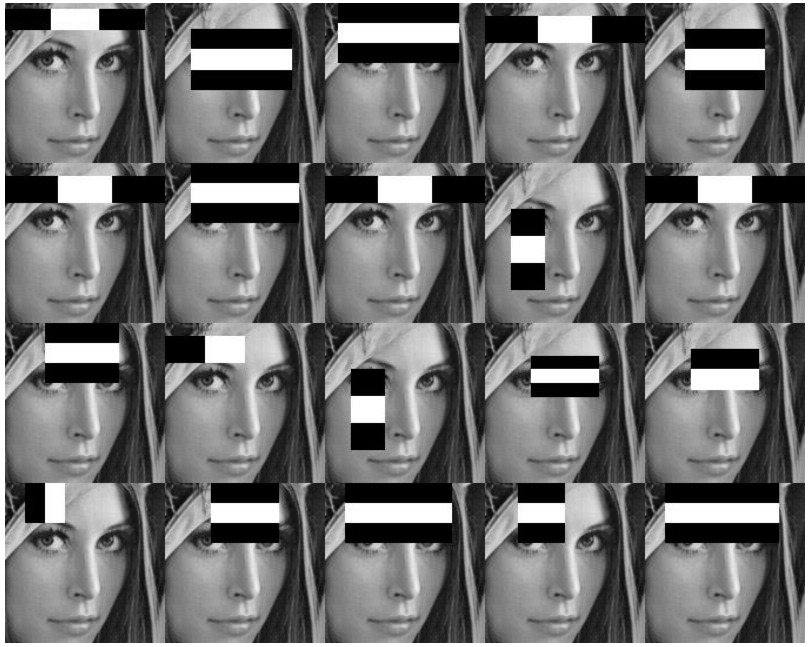
Haar Wavelet basis

- The most popular feature for this task is the Haar Wavelet
- Each basis function is convolved with the image region and its intensity is computed
- The output is a vector of basis functions intensities



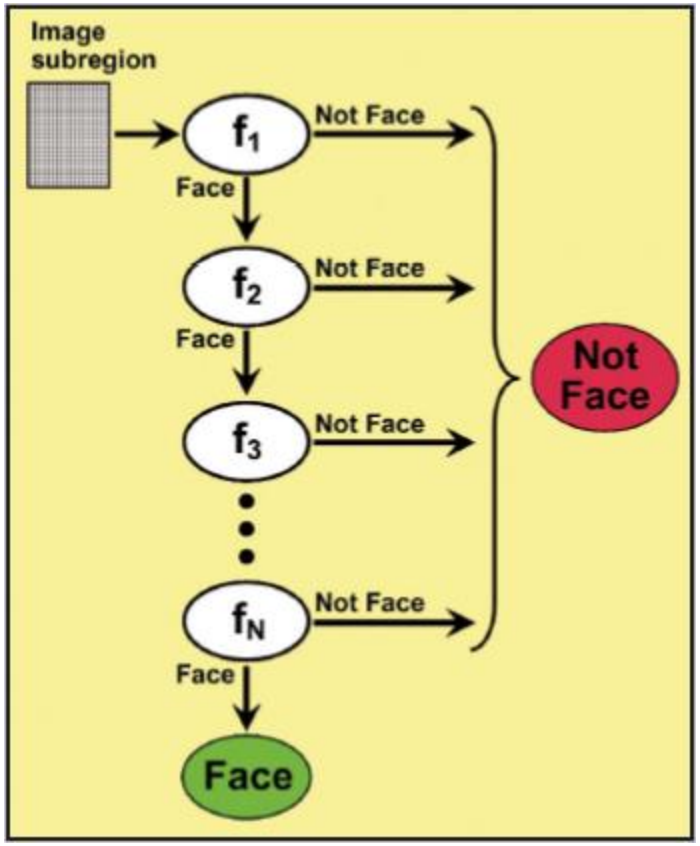
Haar features computation

- Compute the intensity of each Haar basis function for every position in the processed window



Fast object detection: AdaBoost

- To increase the search efficiency, a cascaded classifier is used
  - It combines multiple weak classifiers
    - Classifiers with low accuracy
  - In the first level only the two most important Haar functions are used
- The number of coefficients in each weak classifier increases with the cascade level
  - This allows rejecting most of the true negatives
  - The following levels handle less subregions but use more coefficients to increase precision

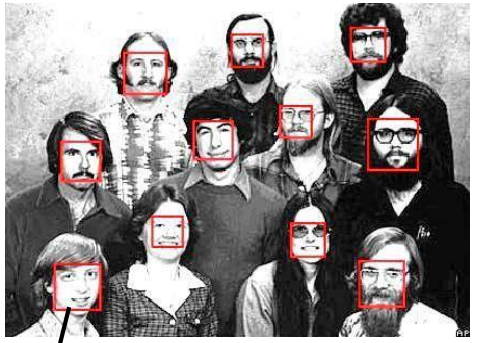


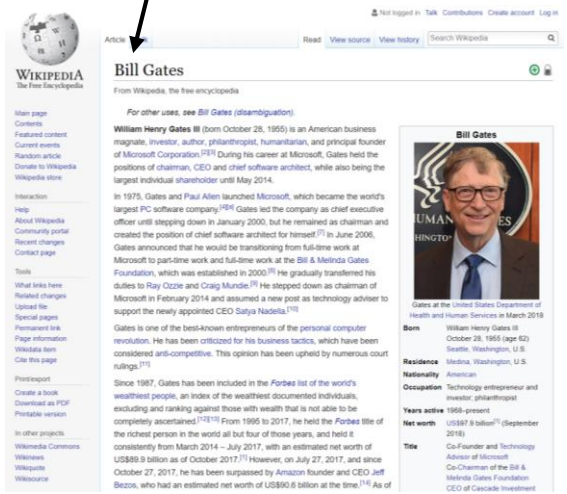
Linking (or recognition)

Web Search

Face recognition

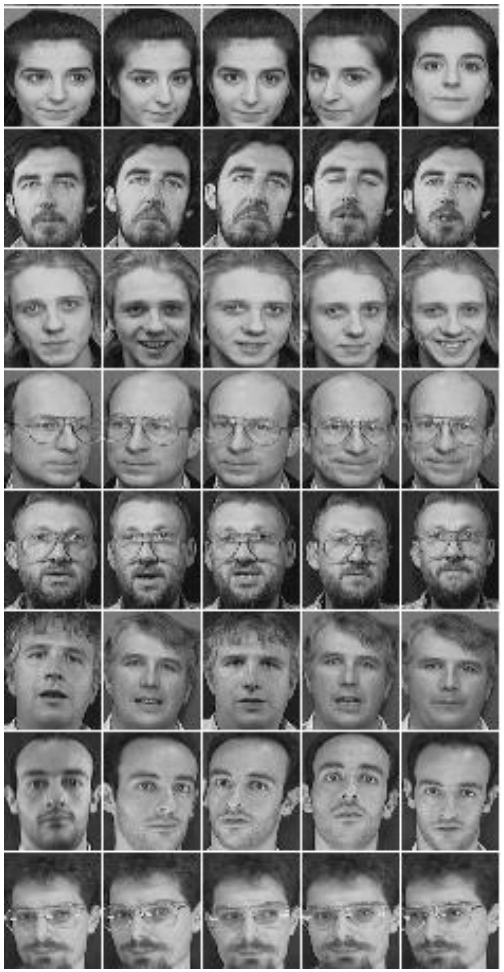
- Once a face is detected, the goal is to recognize the person.
  - Ideally linking the face to some named entity in a taxonomy.
- The image face needs to be classified into one of the existing classes, i.e. one of the known persons.





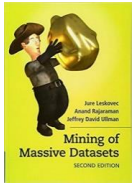
Eigenface space

- The eigenfaces method was the first to successfully address the problem.
  - Based on the SVD algorithm presented in the Recommendation
- It creates a space where face images are represented by a vector and arranged by their similarity to the person of interest.
- Classifiers trained in this space can recognize persons.



Summary

- Information mining tasks
- From data to information: Taxonomies and classes
  - WordNet, ImageNet, SNOMED-CT
- Classification



section 12.2

- Detection
- Linking (recognition)