Information Extraction

Taxonomies, classification, detection and linking

Web Data Mining and Search

Outline

- Introduction
- From data to information: Taxonomies and classes
- Tasks:
  - Classification
  - Detection
  - Recognition
  - Linking
  - Relation extraction

Importance of information extraction

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

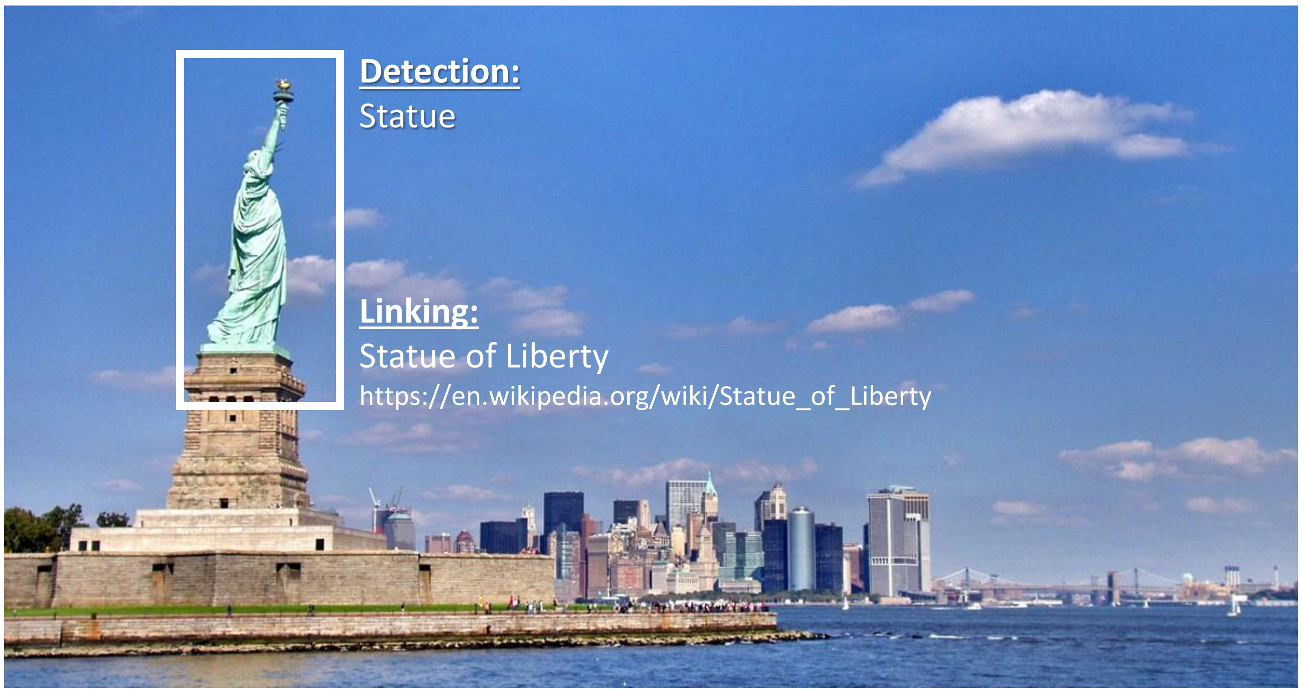
- "There is no question concerning the commercial value of being able to classify documents automatically by content. There are myriad potential applications of such a capability for corporate intranets, government departments, and Internet publishers"

- "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

Linking:

New York City
https://en.wikipedia.org/wiki/New_York_City

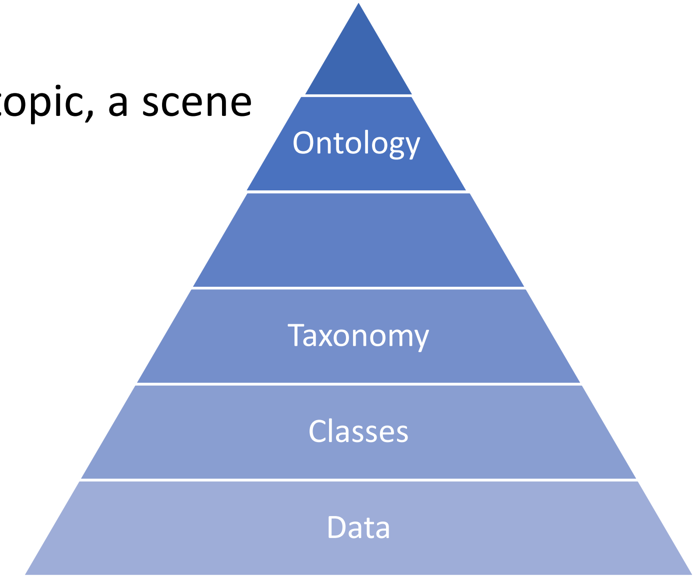
From data to information

Web Data Mining and Search

From data to information

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

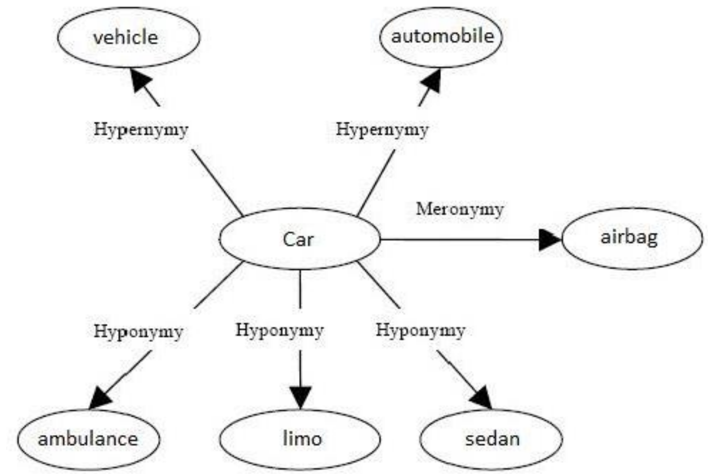
- It is important to identify the list of items that need to be detected.
  - These items are domain specific, and can be a topic, a scene type, a visual object or a named entity.
  - They are normally associated to a class in a supervised learning task.



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."

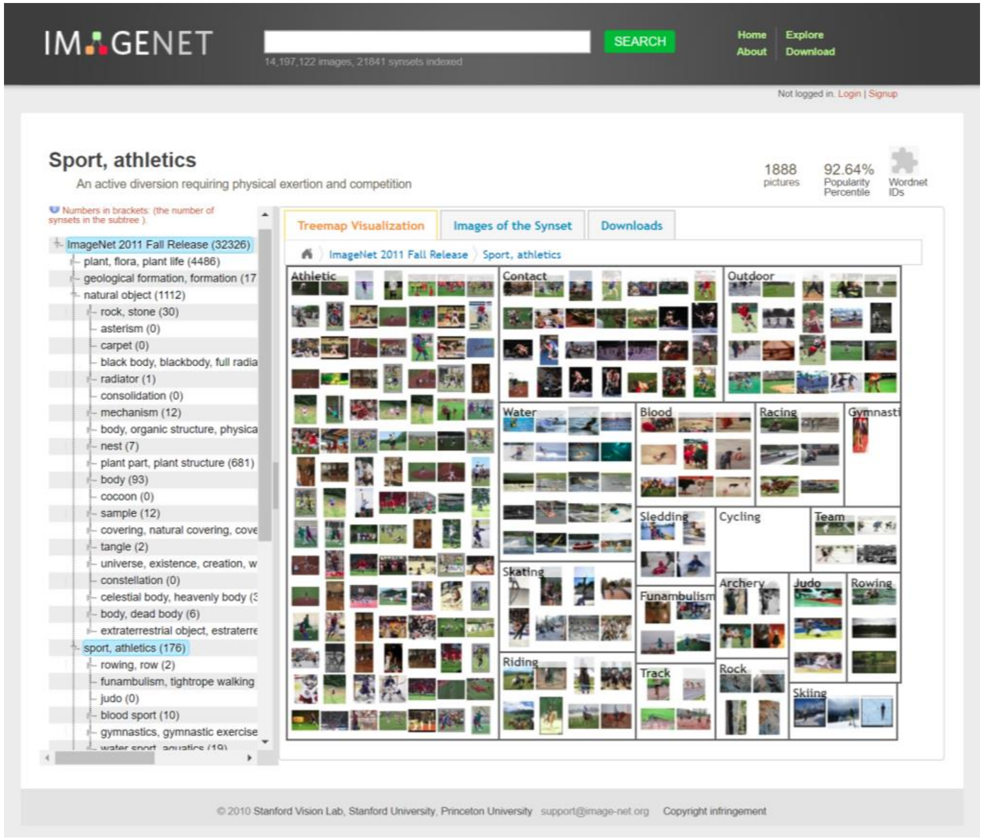


https://wordnet.princeton.edu/

ImageNet: A visual taxonomy

- Selected words of WordNet are illustrated in ImageNet.
- Currently, there are over 14.000 concepts illustrated.
- Roughly 1.000 concepts are used by VOC.
- Great impact in advancing the state of the art.

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

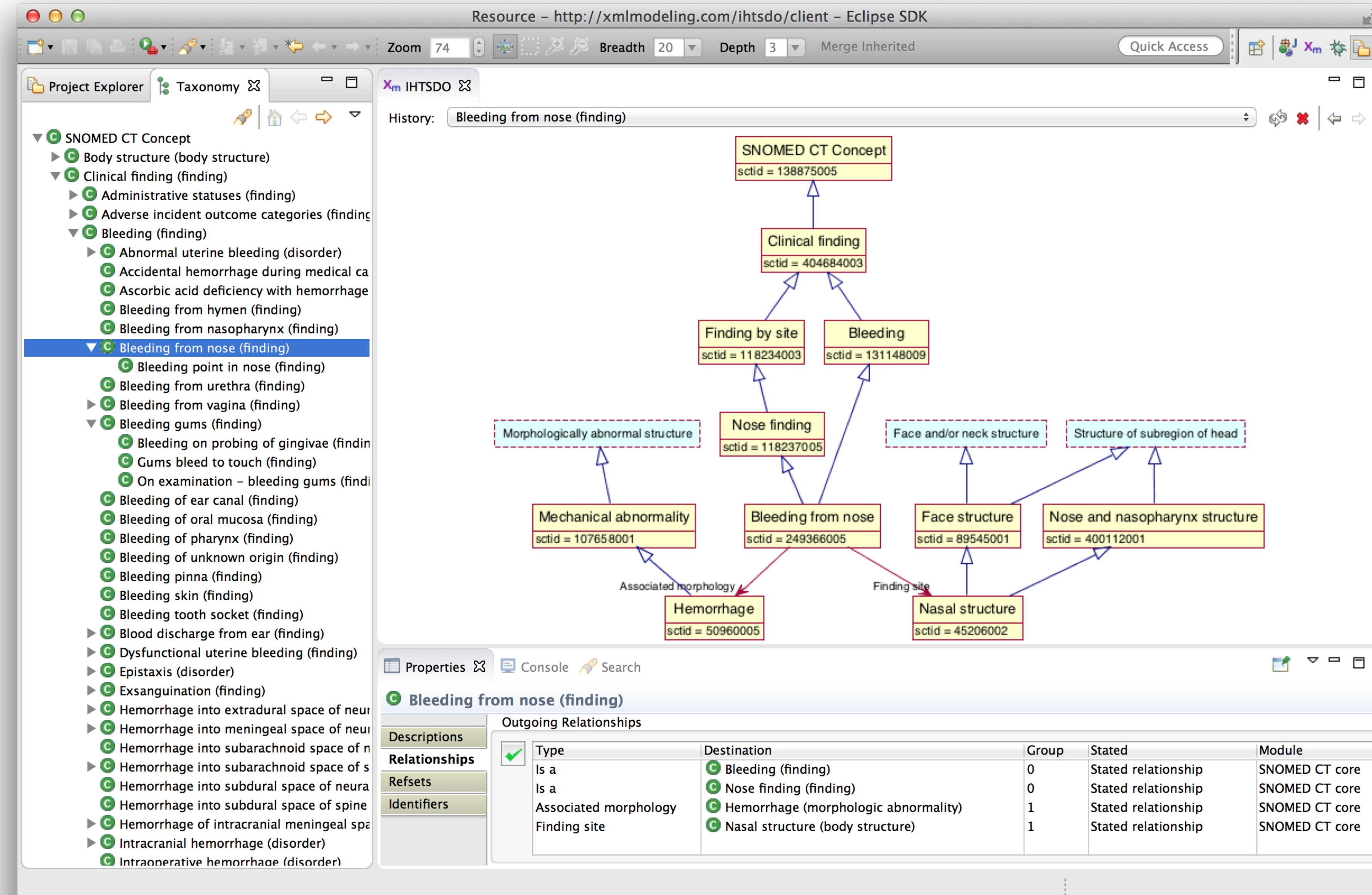


Domain specific taxonomies

- 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
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)
IHTSDO
History: Bleeding from nose (finding)
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

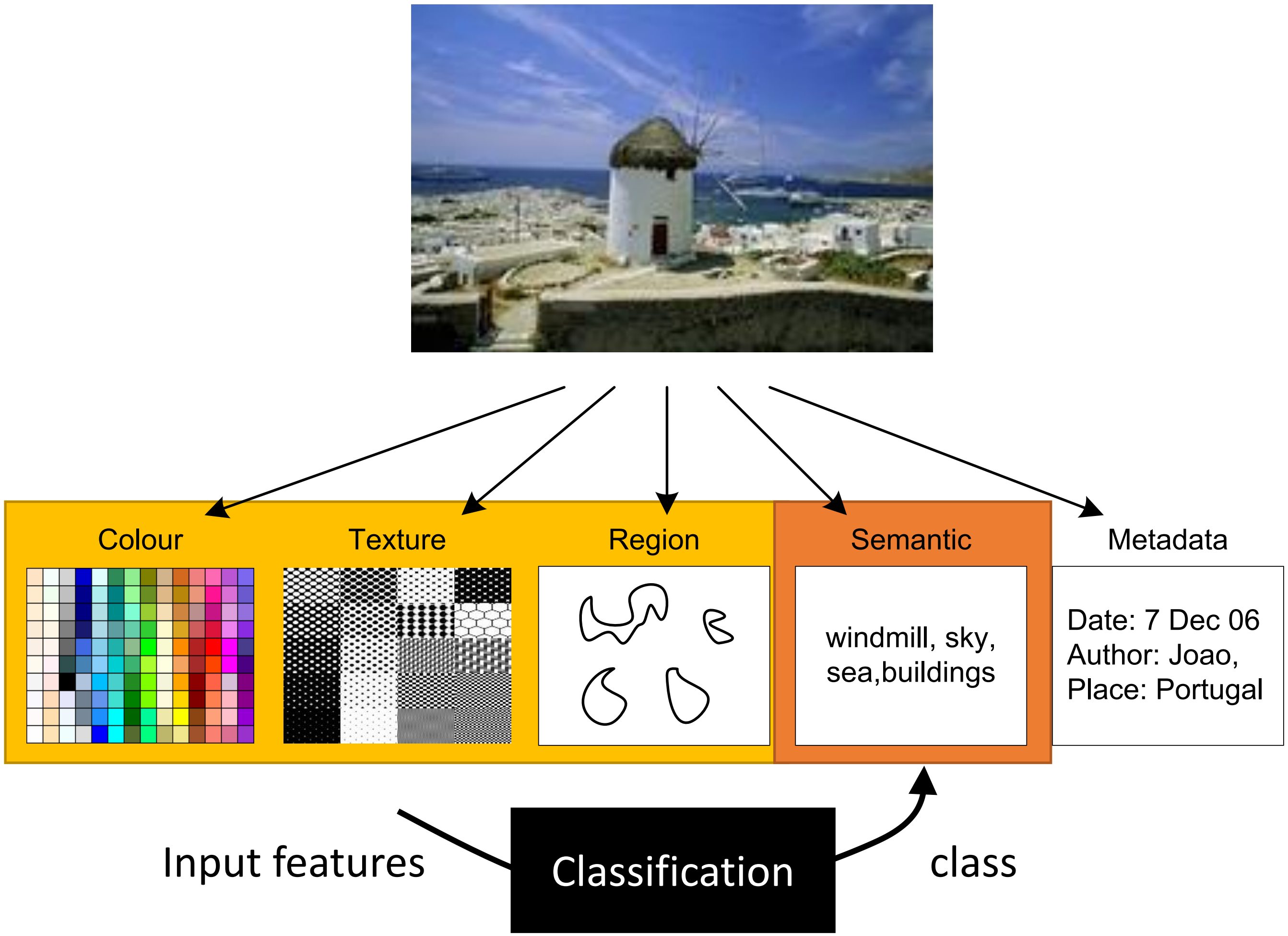
Taxonomies and classification

- 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

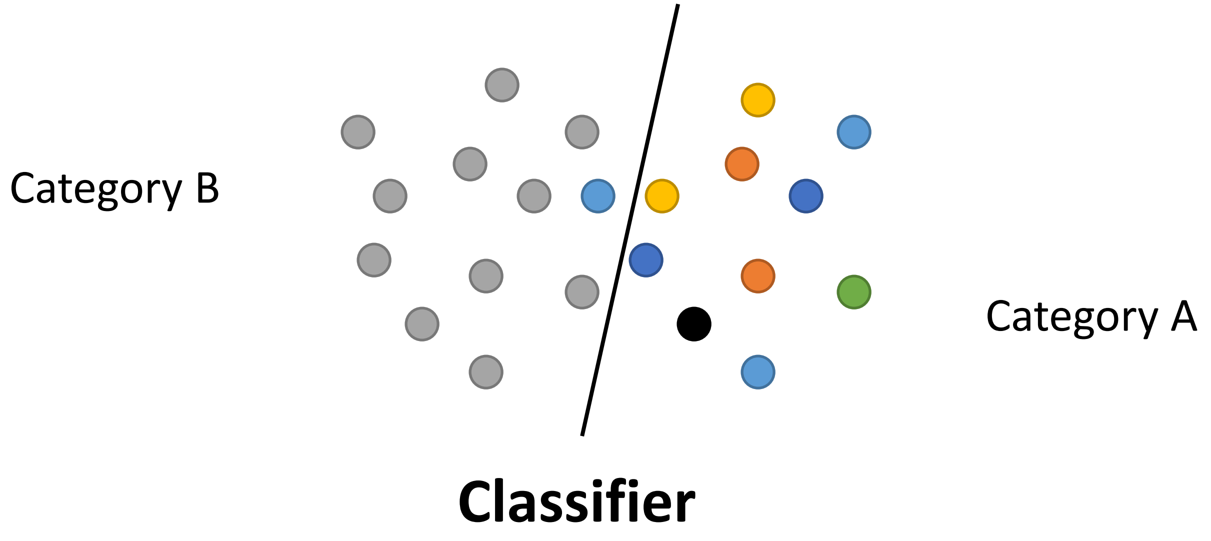
Web Data Mining and Search

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.

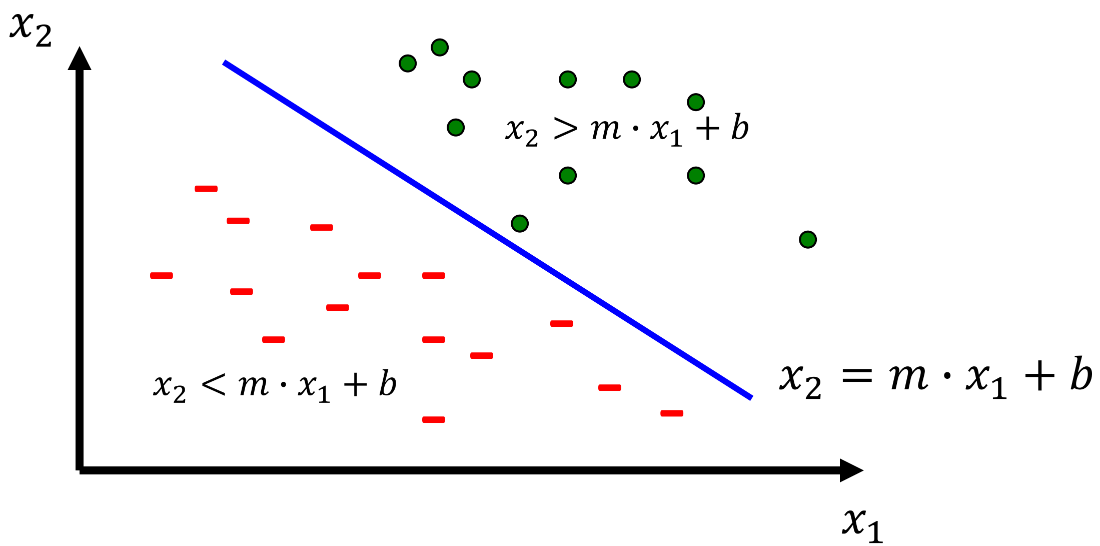


Perceptron

- 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 :**

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

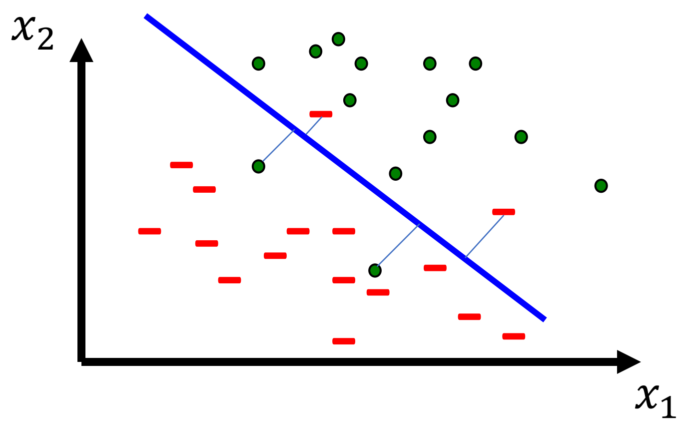


Model error

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

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

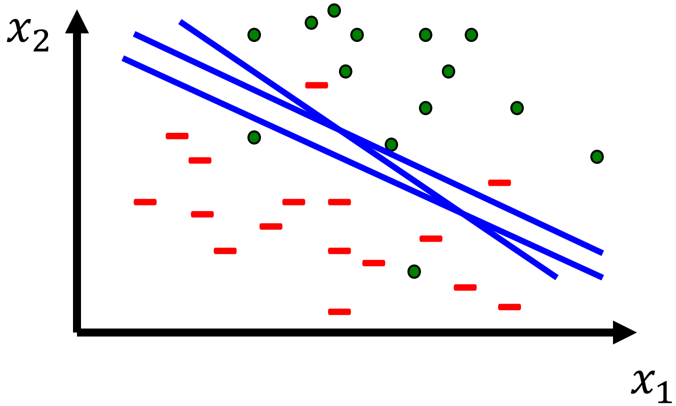
$$x_2 - m \cdot x_1 - b = 0$$

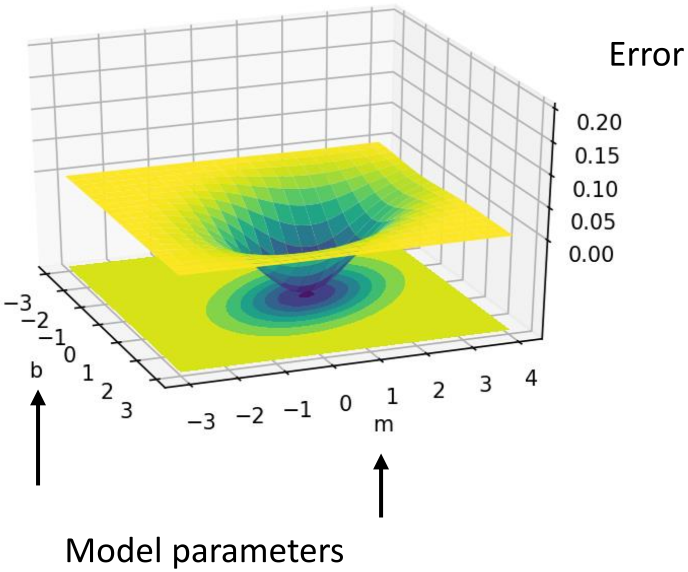


Minimizing the error

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

$$x_2 = x_1 \cdot \boldsymbol{m} + \boldsymbol{b}$$





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
```

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

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

$$update_m = error \cdot x_1$$

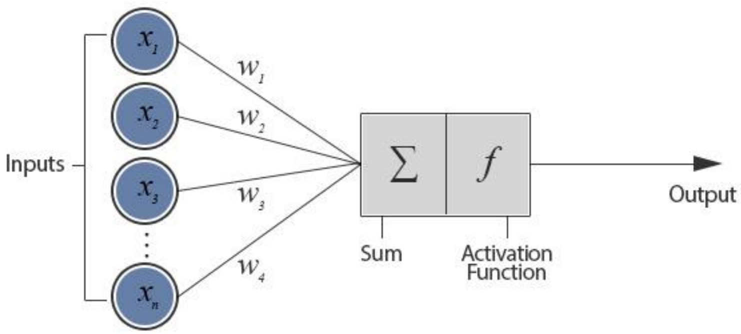
$$m = m - update_m \cdot learning_{rate}$$

Perceptron: general formulation

- **Binary classification:**

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

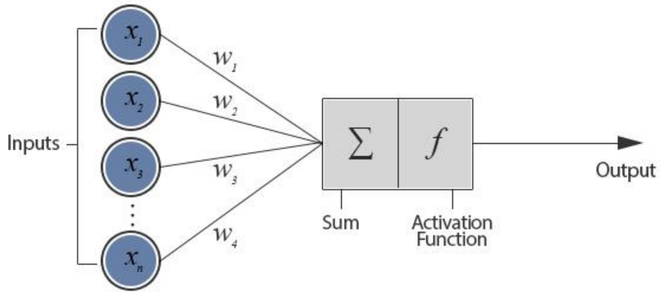
$$\hat{y} = f(z) = \begin{cases} +1 & , if\ z \geq 0 \\ -1 & , if\ z < 0 \end{cases}$$



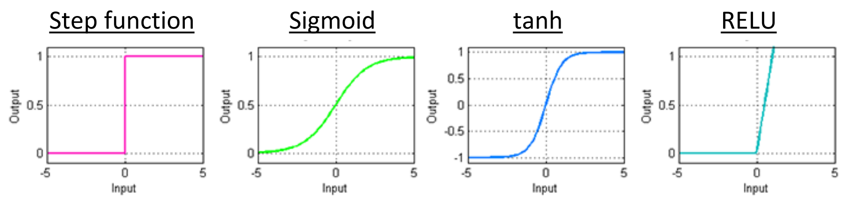
- **Input:** Vectors $\boldsymbol{x}^{(j)}$ and labels $\boldsymbol{y}^{(j)}$
  - Vectors $\boldsymbol{x}^{(j)}$ are real valued where $\|\boldsymbol{x}\|_2 = \mathbf{1}$
- **Goal:** Find vector $\boldsymbol{w} = (\boldsymbol{w_1}, \boldsymbol{w_2}, \ldots, \boldsymbol{w_d})$
  - Each $\boldsymbol{w_i}$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

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

* abragência, precisão 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

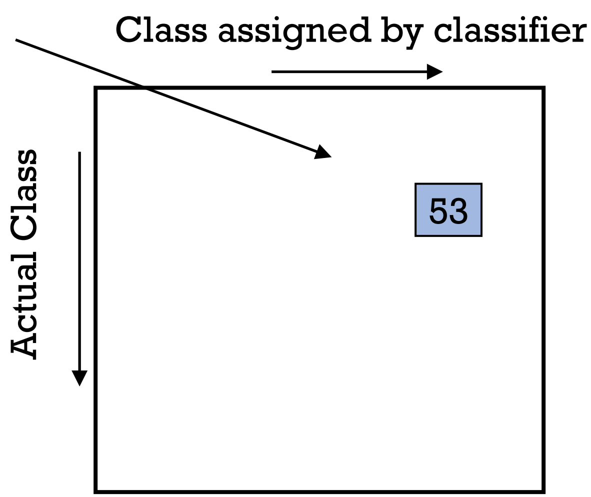
	Truth: yes	Truth: no
Classifier: yes	100	20
Classifier: no	20	1860

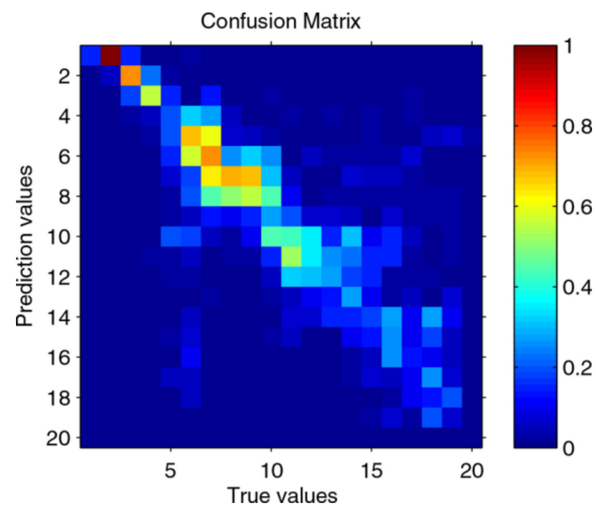
- Macroaveraged precision: (0.5 + 0.9)/2 = 0.7
- Microaveraged precision: 100/120 = .83

- 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 and recognition

Web Data Mining and Search

Detection and recognition

- Detecting and recognizing "things" in natural language and images is a necessary first step in many more complex tasks.

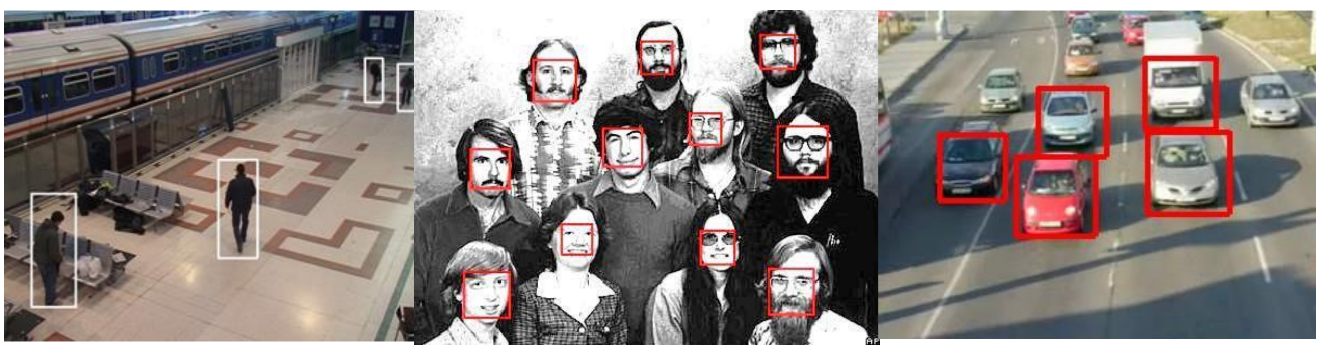
- The "things" that can be detected/recognized include:

Type	Tag	Sample Categories	Example sentences
People	PER	people, characters	**Turing** is a giant of computer science.
Organization	ORG	companies, sports teams	The **IPCC** warned about the cyclone.
Location	LOC	regions, mountains, seas	The **Mt. Sanitas** loop is in **Sunshine Canyon**.
Geo-Political Entity	GPE	countries, states, provinces	**Palo Alto** is raising the fees for parking.
Facility	FAC	bridges, buildings, airports	Consider the **Golden Gate Bridge**.
Vehicles	VEH	planes, trains, automobiles	It was a classic **Ford Falcon**.

Figure 18.1 A list of generic named entity types with the kinds of entities they refer to.

Object detection

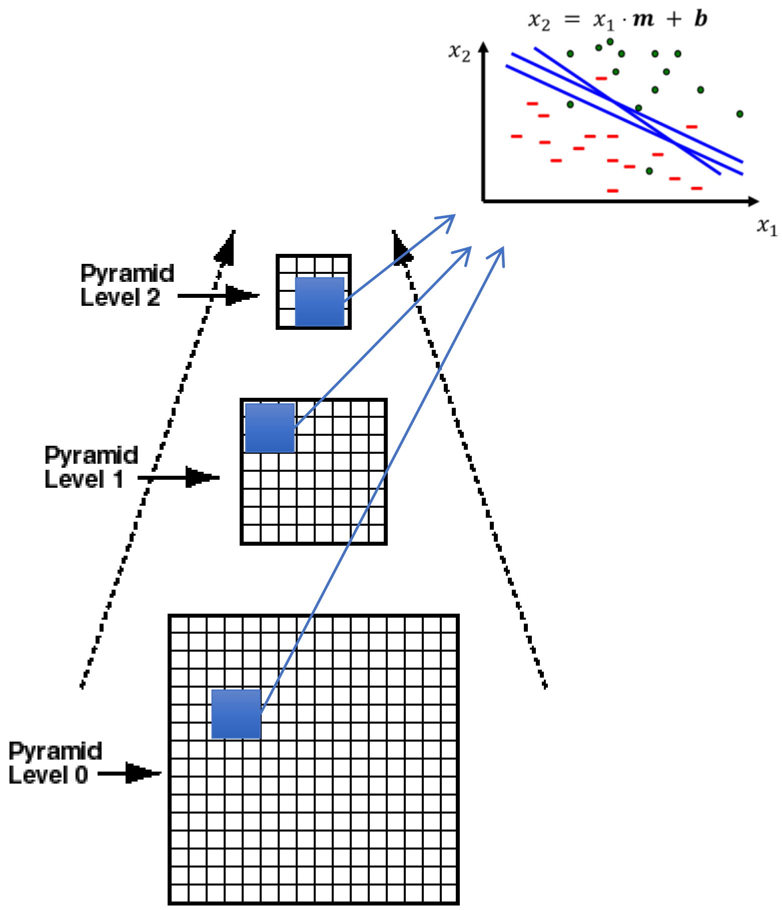
- How to detect a face, a person or a car in a picture?
- How to find pictures of BigBen or the Eiffel Tower?



Object detection

- To solve the problem of finding objects at multiple scales:
  - The image is scalled multiple times
  - At each scale the entire image is scanned for faces on each possible position.

- This is the convolution operation that we will study later in the course.



Named entity recognition

- Recognizing a named entity is an important task to extract the meaning of a sentence or a natural language document.

- Ambiguity can exist in the form of polysemy and synonyms.

Name	Possible Categories
Washington	Person, Location, Political Entity, Organization, Vehicle
Downing St.	Location, Organization
IRA	Person, Organization, Monetary Instrument
Louis Vuitton	Person, Organization, Commercial Product

Figure 18.2 Common categorical ambiguities associated with various proper names.

[$_{\text{PER}}$ Washington] was born into slavery on the farm of James Burroughs.
[$_{\text{ORG}}$ Washington] went up 2 games to 1 in the four-game series.
Blair arrived in [$_{\text{LOC}}$ Washington] for what may well be his last state visit.
In June, [$_{\text{GPE}}$ Washington] passed a primary seatbelt law.
The [$_{\text{VEH}}$ Washington] had proved to be a leaky ship, every passage I made...

Figure 18.3 Examples of type ambiguities in the use of the name *Washington*.

Named entity recognition

- To detect or recognize a named entity, one needs to run a classifier over the sequence of tokens of the natural language sentence.

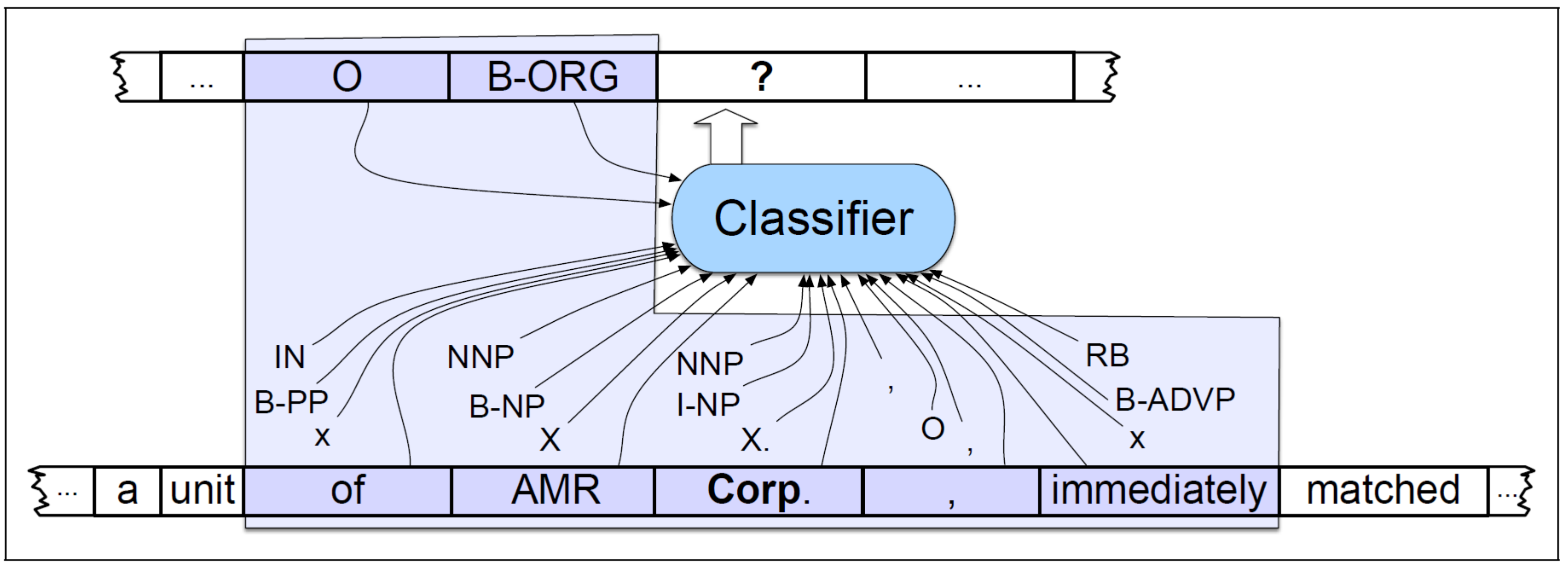
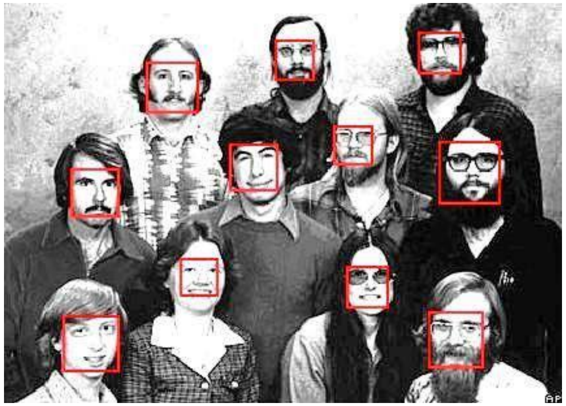


Figure 18.7 Named entity recognition as sequence labeling. The features available to the classifier during training and classification are those in the boxed area.

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?

Linking

Web Data Mining and Search

Entity linking

- Entity linking concerns the task of linking a mention or an image region to the unique identifier of the entity represented in that piece of data.

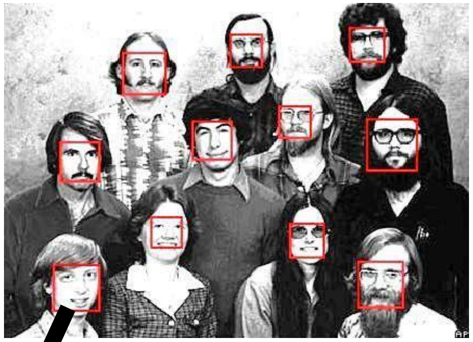
- “The task of entity linking is to associate an occurrence in text with the representation of some real-world entity in an ontology, a list of entities in the world, like a gazeteer.”

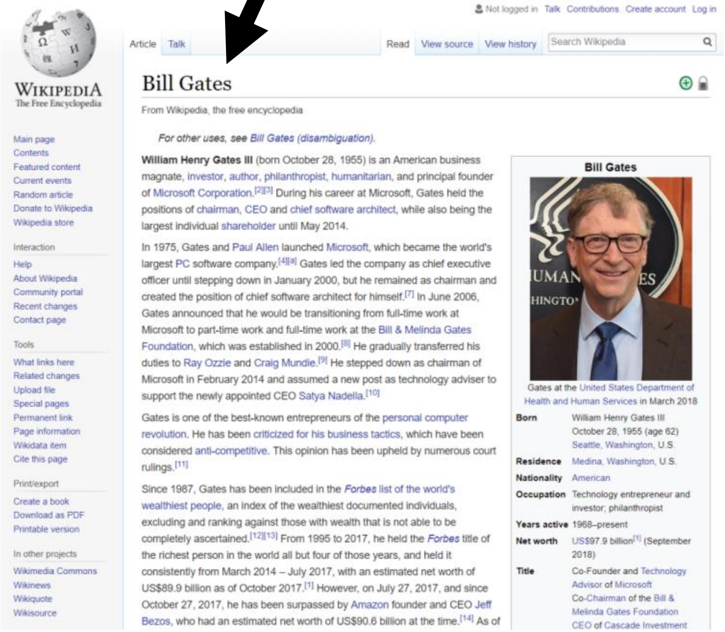
- “Perhaps the most common ontology used for this task is Wikipedia, in which each Wikipedia page acts as the unique id for a particular entity.”

Person recognition

- Once a face is detected, the goal is to link the face image to 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.





Not logged in Talk Contributions Create account Log in

Article Talk Read View source View history Search Wikipedia

WIKIPEDIA The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export
Create a book
Download as PDF
Printable version

In other projects
Wikimedia Commons
Wikinews
Wikiquote
Wikisource

Bill Gates

From Wikipedia, the free encyclopedia

For other uses, see Bill Gates (disambiguation).

William Henry Gates III (born October 28, 1955) is an American business magnate, investor, author, philanthropist, humanitarian, and principal founder of Microsoft Corporation.[2][3] During his career at Microsoft, Gates held the positions of chairman, CEO and chief software architect, while also being the largest individual shareholder until May 2014.

In 1975, Gates and Paul Allen launched Microsoft, which became the world's largest PC software company.[4][a] Gates led the company as chief executive officer until stepping down in January 2000, but he remained as chairman and created the position of chief software architect for himself.[7] In June 2006, Gates announced that he would be transitioning from full-time work at Microsoft to part-time work and full-time work at the Bill & Melinda Gates Foundation, which was established in 2000.[8] He gradually transferred his duties to Ray Ozzie and Craig Mundie.[9] He stepped down as chairman of Microsoft in February 2014 and assumed a new post as technology adviser to support the newly appointed CEO Satya Nadella.[10]

Gates is one of the best-known entrepreneurs of the personal computer revolution. He has been criticized for his business tactics, which have been considered anti-competitive. This opinion has been upheld by numerous court rulings.[11]

Since 1987, Gates has been included in the *Forbes* list of the world's wealthiest people, an index of the wealthiest documented individuals, excluding and ranking against those with wealth that is not able to be completely ascertained.[12][13] From 1995 to 2017, he held the *Forbes* title of the richest person in the world all but four of those years, and held it consistently from March 2014 – July 2017, with an estimated net worth of US$89.9 billion as of October 2017.[1] However, on July 27, 2017, and since October 27, 2017, he has been surpassed by Amazon founder and CEO Jeff Bezos, who had an estimated net worth of US$90.6 billion at the time.[14] As of

Bill Gates

Gates at the United States Department of Health and Human Services in March 2018

Born	William Henry Gates III October 28, 1955 (age 62) Seattle, Washington, U.S.
Residence	Medina, Washington, U.S.
Nationality	American
Occupation	Technology entrepreneur and investor; philanthropist
Years active	1968–present
Net worth	US$97.9 billion[1] (September 2018)
Title	Co-Founder and Technology Advisor of Microsoft Co-Chairman of the Bill & Melinda Gates Foundation CEO of Cascade Investment

Summary

- Information extracton tasks
  - https://web.stanford.edu/~jurafsky/slp3/18.pdf

- From data to information: Taxonomies and classes
  - WordNet, ImageNet, SNOMED-CT

- Linear classifier:
  - http://d2l.ai/chapter_linear-networks/index.html