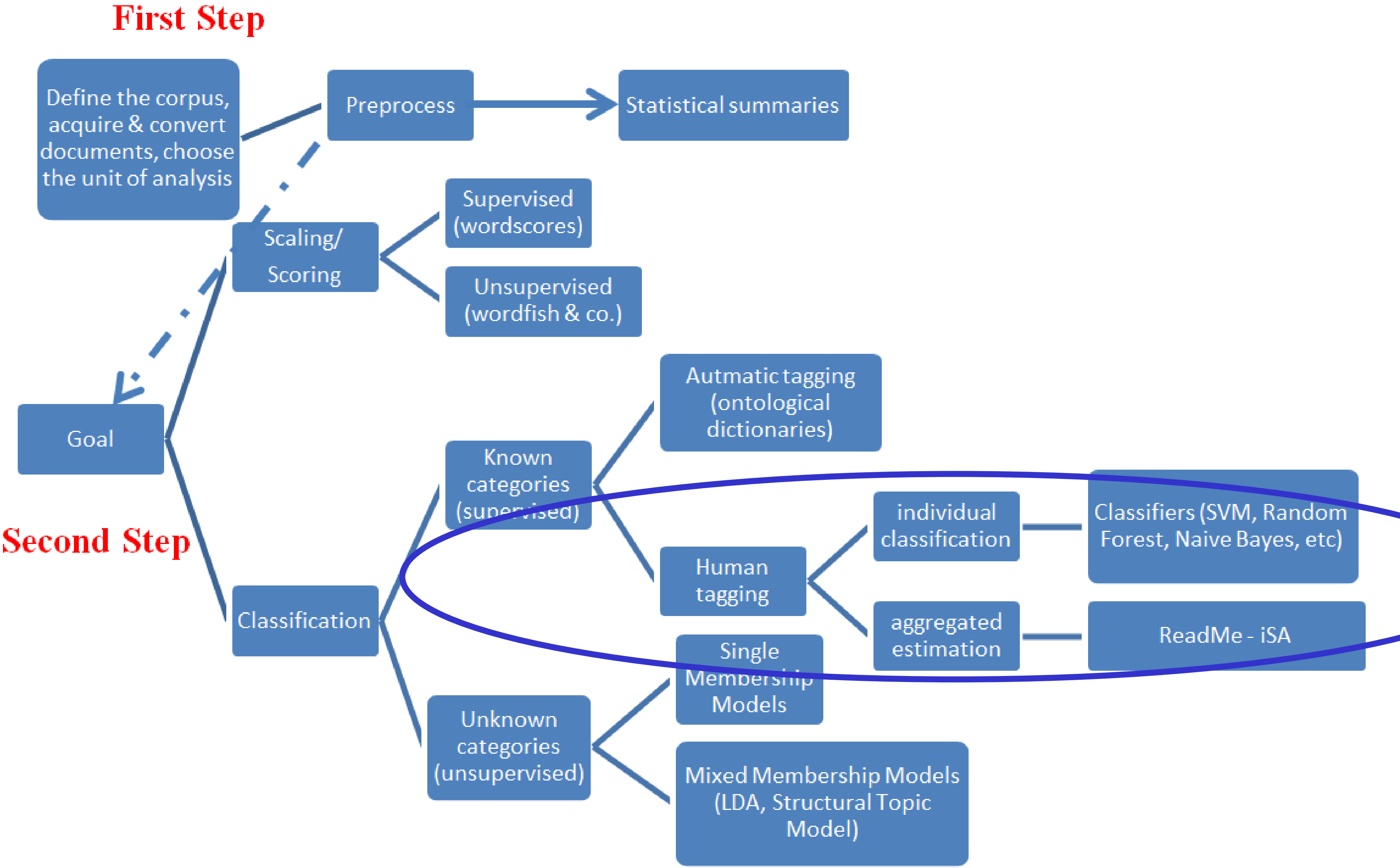
Big Data Analytics

Lecture 6 – part 2

Supervised classification methods with human tagging: an introduction

Our Course Map



First Step

Define the corpus, acquire & convert documents, choose the unit of analysis

Preprocess

Statistical summaries

Scaling/ Scoring

Supervised (wordscores)

Unsupervised (wordfish & co.)

Goal

Second Step

Classification

Known categories (supervised)

Autmatic tagging (ontological dictionaries)

Human tagging

individual classification

Classifiers (SVM, Random Forest, Naive Bayes, etc)

aggregated estimation

ReadMe - iSA

Unknown categories (unsupervised)

Single Membership Models

Mixed Membership Models (LDA, Structural Topic Model)

References

- Grimmer, Justin, and Stewart, Brandon M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3): 267-297
- Curini, Luigi, and Robert Fahey (2020). Sentiment Analysis and Social Media. In Luigi Curini and Robert Franzese (eds.), *SAGE Handbook of Research Methods is Political Science & International Relations*, London, Sage, chapter 29
- Barberá, Pablo et al. (2020) Automated Text Classification of News Articles: A Practical Guide, *Political Analysis*, DOI: 10.1017/pan.2020.8

Supervised Learning (classification) Methods

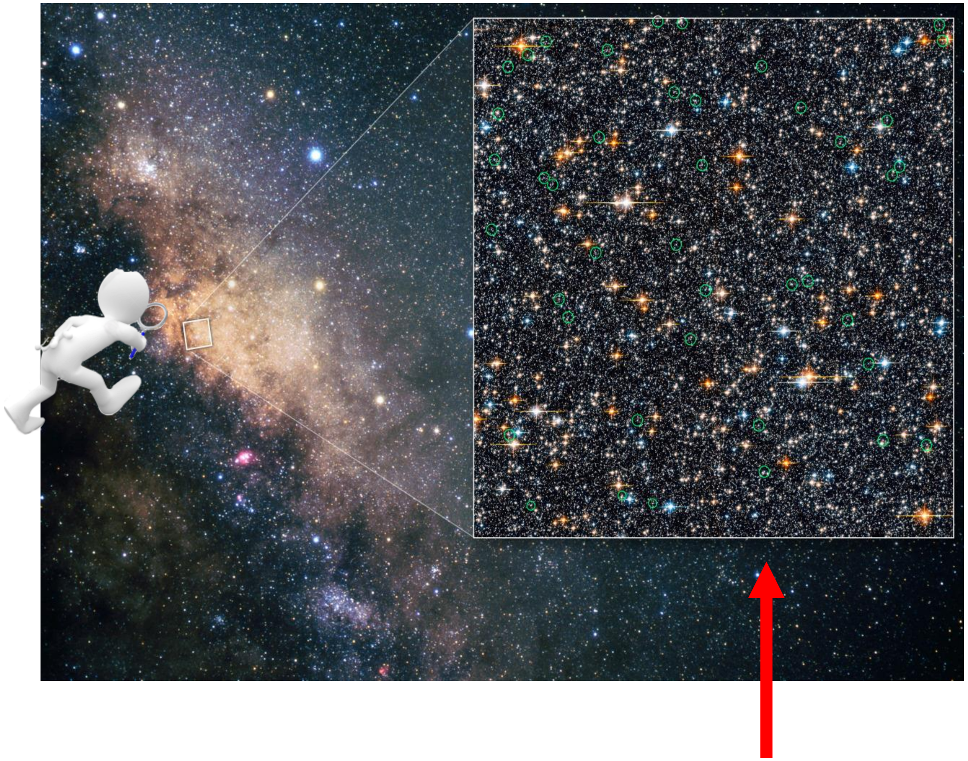
The idea of supervised learning is simple: human coders categorize a set of documents (the “**training-set**” or **“labelled-set”**) by hand into a set of pre-defined categories (such as positive, negative, neutral for example)

The algorithm “learns” how to sort the documents into categories using the **training set and words**

Then, it classifies the remaining set of document not classified by hand (the **“test-set”** or **“unlabelled set”**) using the characteristics (i.e., words) of the unread documents to place them into the categories

A four-step procedure

1. Data preparation: separating the training set from the test set in the corpus



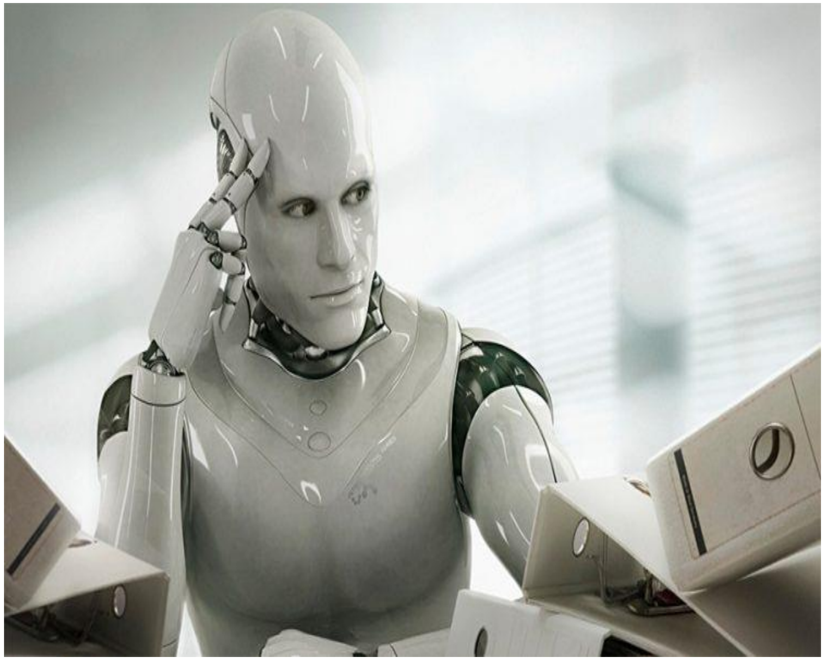
the training set

2. Human classification
of the training set on a base of a list of pre-defined categories

A four-steps procedure

3. Cogito ergo sum! The algorithm learns from the human classification done in the training set

4. Let's classify! The well-educated algorithm is now ready to classify all the texts in the test-set



The algorithms that we will employ belong to the Machine Learning class

Machine learning

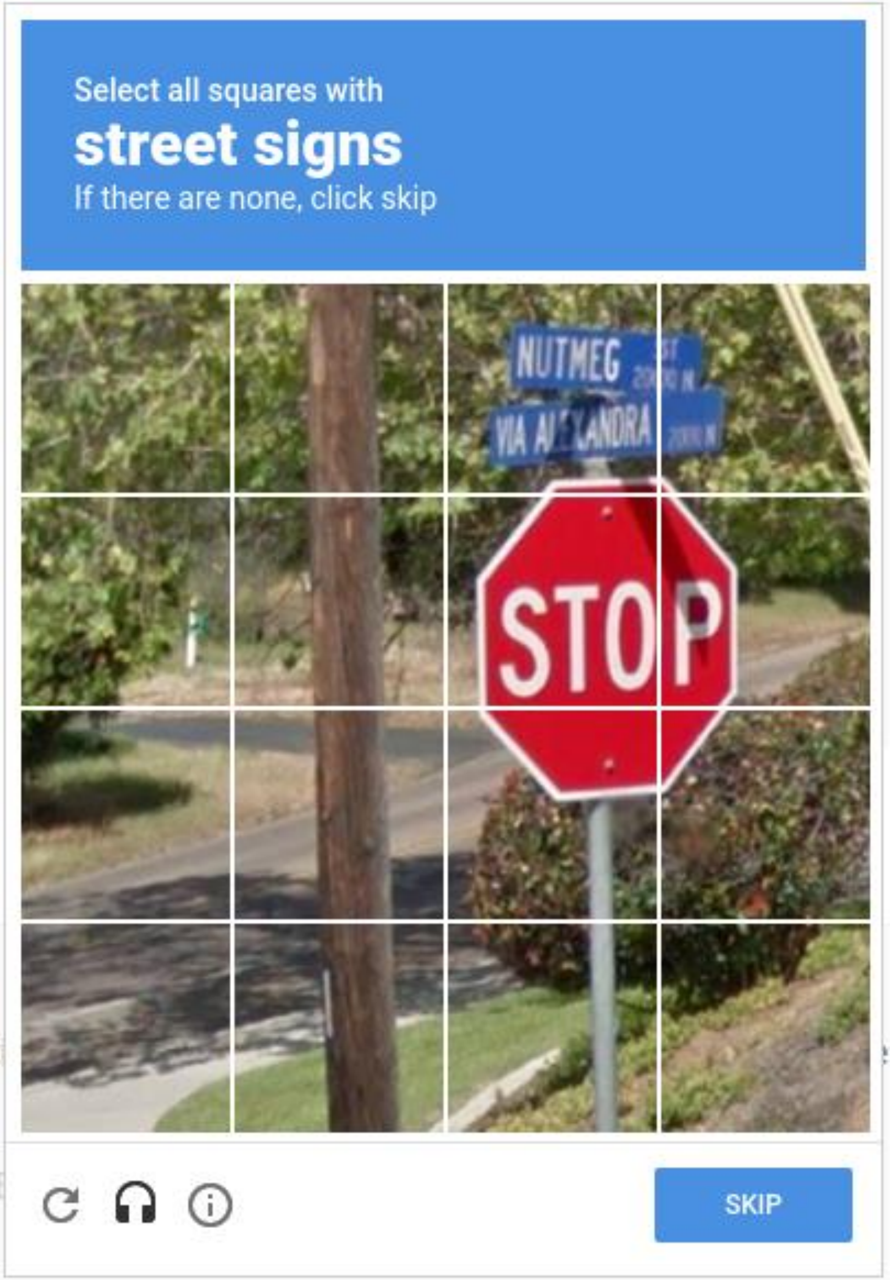
Machine learning is defined as the “field of study that gives computers the ability to learn without being explicitly programmed” (Samuel 1959)

In this context “learning” can be viewed as the use of statistical techniques to enable computer systems to progressively improve their performance on a specific task from data without being explicitly programmed (Goldberg and Holland 1988)

Machine

To be able to le become better at it, a machine

- ✓ …be provide tion (inputs) and the desi learn a general rule to the outputs



Select all squares with

street signs

If there are none, click skip

Machine learning

In our case, our aim is to do *text classification*

Therefore, **machine learning algorithms** (when dealing with text classification methods) refer to those techniques that learn how to map a set of inputs (e.g., features within documents) to a predicted class as the output in a pre-coded *training set (via human intervention)* before classifying the data in the *test set*

Supervised Learning (classification) Methods

Despite the fact that the methods to do supervised classification are diverse, they share a **common structure** that usefully unifies the methods

The dfm *S* representing our corpus has *N* rows and *L* columns, with each document *j* being represented by a vector S_j of length L

The value of each element of the vector S_j may either be the frequency with which that feature appeared in document *j*, or a binary value – 1 if the feature appeared at all, 0 if it did not

Of course, more than one text in the corpus can be represented by the same vector S_j

Supervised Learning (classification) Methods

This dfm is then divided into two subsets. One of length *n*, which is called the *training set*, and the remaining one of size *N-n*, the *test set*

We denote by (D_j, S_j) the couple that contains the coded value D_j for text S_j

Clearly, D_j is a value in *D* for the texts in the training set (such as positive, negative, neutral, if you think about a Sentiment classification) and "NA" for the uncoded texts in the test set

The texts in the training set are assumed to be classified without error

Supervised Learning (classification) Methods

Any machine learning algorithm will then try to predict the category D_j to which a given document *j* belongs given that its features are represented by the vector S_j

This model can be represented as $P(D|S)$ – for a given document *j* and set of categories *M*, it will find the value of *m*, i.e. the classification, which maximises the model $P(D_{m=0,1,\dots,M}|S_j)$

Expressing this as a matrix model, the classification algorithm is:

$P(D) = P(D|S)P(S)$ – where *P(D)* is a vector of length M+1, *P(D|S)* is a matrix of conditional probabilities and *P(S)* is a vector representing the distribution of text vectors across the corpus of texts

Supervised Learning (classification) Methods

More in details, each supervised learning algorithm assumes that there is some (unobserved) function that describes the (**true**) relationship between the words S and the categories M in the training-set

It then attempts to learn this relationship by approaching the “true” function with a **classification function** $\hat{P}(D|S)$ extracted from the subset of observations in the training set

$\hat{P}(D|S)$ is then used to infer properties (categories) of the test set (the unlabeled set), using the actual S_j in the test set for each document in the test set

Supervised Learning (classification) Methods

That is, the value of P(D) for the data in the test set is obtained by replacing S with the actual S_j in the test set and assigning D_j = argmax $\hat{P}(D|S = S_j)$

In a naive model, the elements of the matrix P(D|S) can be estimated by taking the proportion of all texts in the training set that are hand-coded as D = D_m which have S = $S_j$as feature vector

Any other model (Support Vector Machines, etc.) will do essentially the same thing in more sophisticated ways (as we will learn)

Supervised Learning (classification) Methods

Summing up…

Supervised learning models **share the same goal**: learn the potentially complicated relationships that relate (combinations of) features x to the outcome of interest y in general, using information available in the set of observations for which the pair (x; y) is fully observed (i.e., in the training-set)

Supervised Learning vs. Dictionary methods

Supervised learning can be therefore conceptualized as a **generalization of dictionary methods**, where features associated with each categories (and their relative weight) are learned from the data **via human intervention**

The feature space is thus likely to be both larger and more comprehensive than that used in a dictionary

The end result is that much more information drives the subsequent classification of text

Supervised Learning vs. Dictionary methods

Moreover, compared to dictionary methods:

Supervised learning is **necessarily domain specific** and therefore avoids the problems of applying dictionaries outside of their intended area of use

Second, human involvement is crucial to understand the correct meaning of a text (double meaning sentences, specific jargons, neologisms, irony)

Finally, supervised learning methods are much easier to validate, with clear statistics that summarize model performance (as we will discuss)

Supervised Learning (classification) Methods

We have two broad classes of Supervised Learning Methods:

a) those who aim to classify the **individual documents** into categories

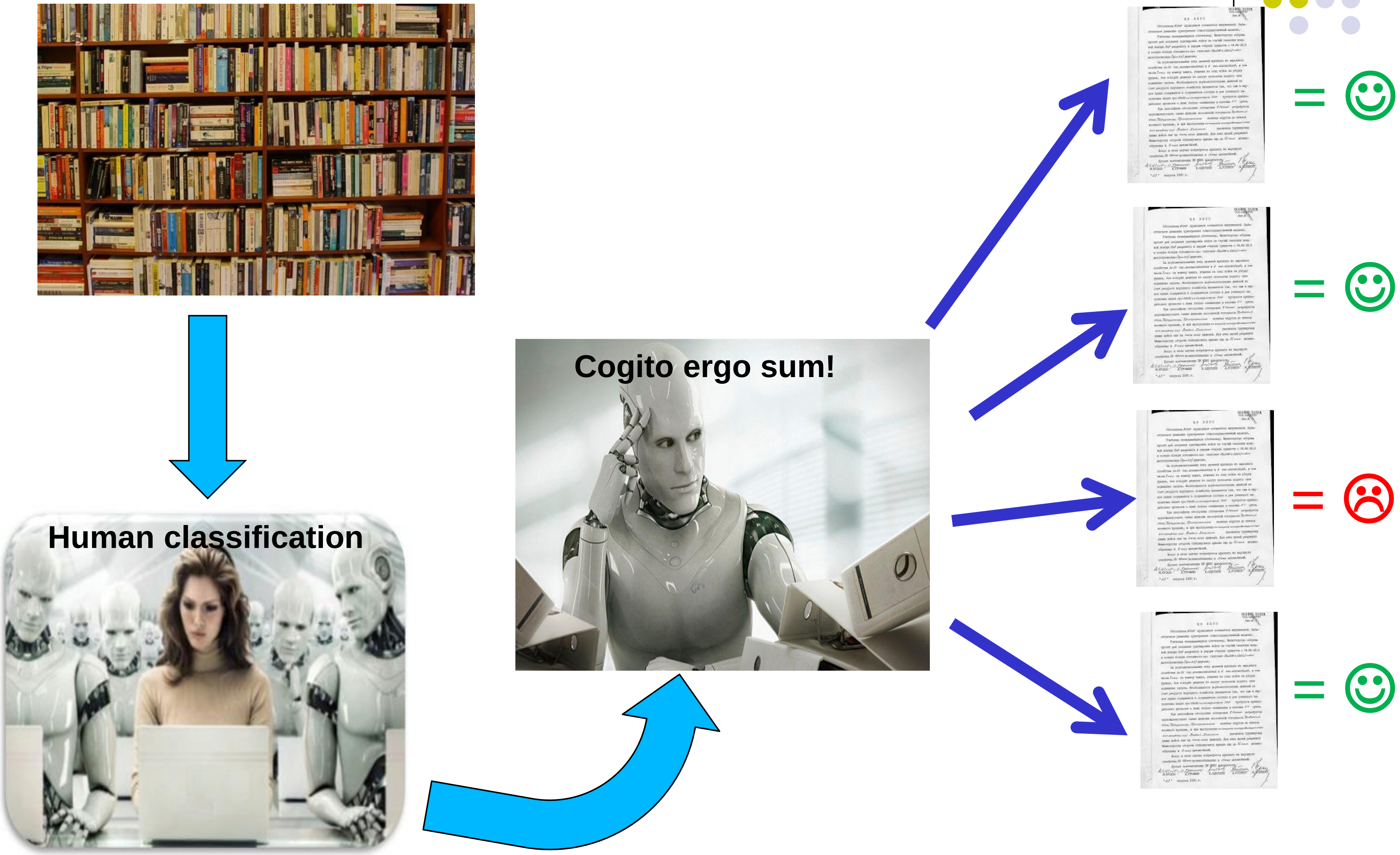
b) those who aim to measure the **proportion of documents** in each category

Applying a supervised learning model

Individual approaches estimate the category of each text in a new corpus of texts (test-set) aiming at *minimizing the probability of error in individual class assignment*…

…while **aggregated approaches** estimate the class proportions into the new corpus of texts (test-set), rather than assembling the results of several individual classifications, aiming at *minimizing the error between the estimated proportions and true proportions*

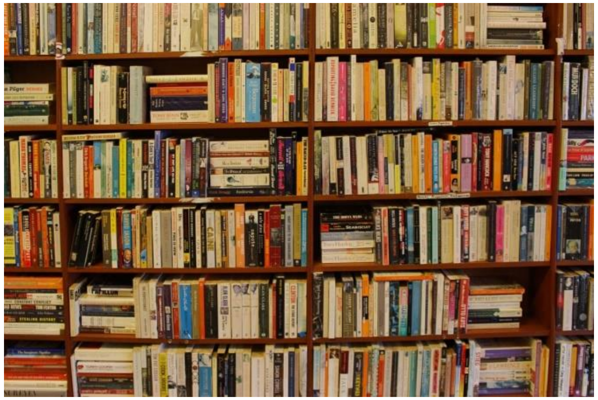
Supervised learning: individual classification



Human classification

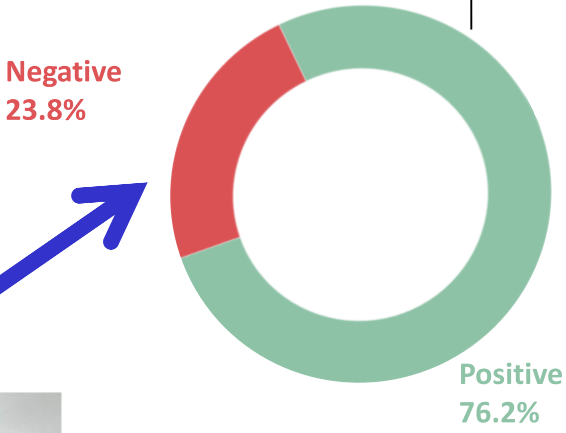
Cogito ergo sum!

Supervised learning: proportional classification



Human classification

Cogito ergo sum!



Applying a supervised learning model

There is an important **theoretical (and statistical) difference** between individual and proportional classification:

- ✓ for some social science application, only the proportion of documents in a category is needed, not the categories of each individual document
- ✓ The opposite happens in some other application

Once again, there is no “a best method” out there. It depends on your research topic (remember the 4 rules!)

In our course, we will discuss mainly about individual classification approaches. Hopefully we will have time to discuss about proportional classification algorithm as well later on