# Lecture reviews — Week 07

J.-C. Chappelier & M. Rajman

Laboratoire d'Intelligence Artificielle
Faculté I&C

©EPFL
J.-C. Chappelier & M. Rajman

# Week 7 keypoints

- supervised/unsupervised
- preprocessing is key
- baseline methods:
  - classification: Naive Bayes, (Logistic regression,) KNN
  - clustering: K-means, dendrograms
  - dim. reduction: PCA, UMAP
- don't forget evaluation keypoints (see lesson 2)

©EPFL
J.-C. Chappelier & M. Rajman

# Week 7 – study case

1 2 ... 6324

a abacus abbey ... cat ... mouse ... Zulu Zygomatic

$D_1$ : 37 0 0 ... 10 ... 12 ... 0 0

Some financial company offers you to work on
"*fraud detection using Natural Language Technology applied to client documents*".

① Some preliminary work has already been performed by a former intern who created document vectors based on an indexing set of 6'324 terms and reduced them to vectors of size 100 using PCA.

Reviewing his/her work and report, you found a graph related to the corresponding singular values.
Next slide shows a (rescaled) zoom on the first 550 left-most points in that graph.

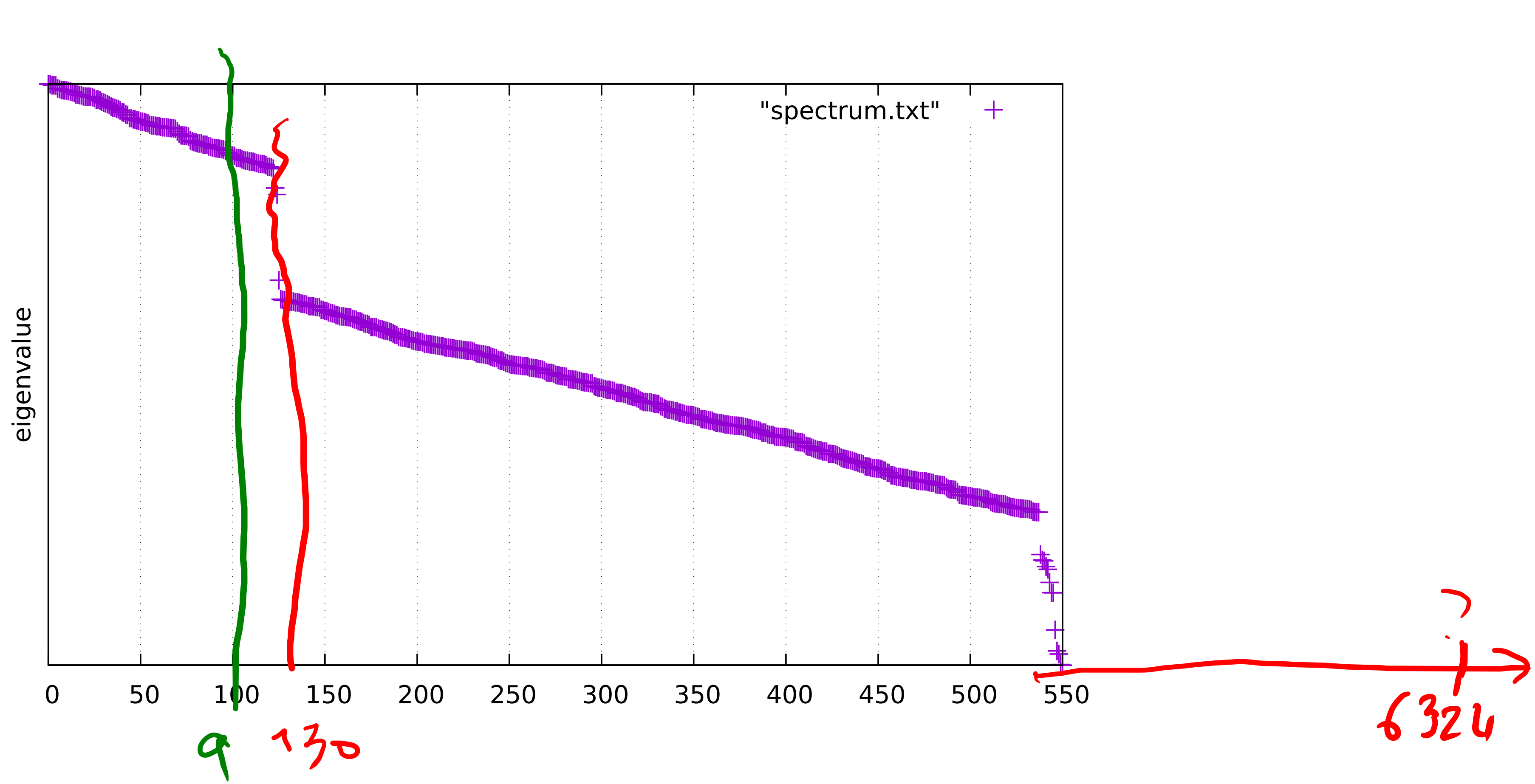
$D_N$

©EPFL
J.-C. Chappelier & M. Rajman

# Week 7 – study case

$\lambda_i$

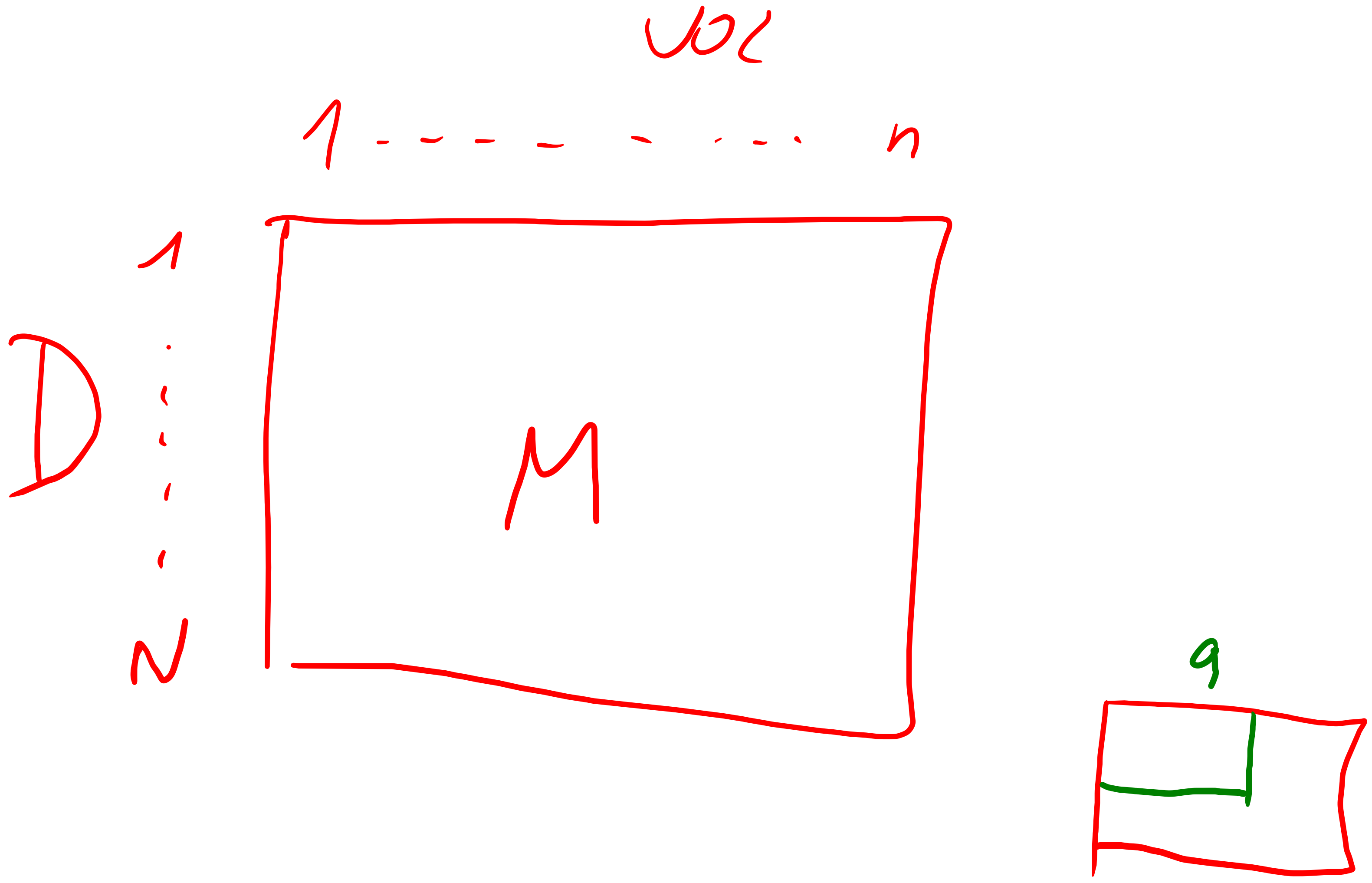
$$\frac{\lambda_i}{\sum \lambda_i}$$




**a)** What is the abscissa (x-value, horizontal axis) of the right-most point in the original complete graph (not reported here)?

**b)** What do you think about the intern's methodology for selecting the dimension of the vector space? Would you have performed differently? If yes, how?

©EPFL
J.-C. Chappelier & M. Rajman
EPFL




$$MM^t \longrightarrow U_q \Lambda_q V_q^t$$

# Week 7 – study case

features $f_i$

vector dim 130

② Before considering more sophisticated Deep-Learning methods, you wisely decide to start with a simple baseline, namely a Naive Bayes model (on the former representation).

document after PCA :

**a)** Based on your former answer, what is the input of the Naive Bayes module?
What is the output? ⟶ fraud / not fraud
What are the parameters?
What is needed for training such a model?

$P(\text{fraud})$

$P(f_i \mid \text{fraud})$

+ same with ¬fraud

**b)** Concretely, what probability should be computed as an output from the (very simple excerpt of) client document:

*My salary is about 10'000 CHF and I don't pay any tax.*

1) preprocessing ?

2) project ⟶ vector $f_i$

3) classif : $P(\text{fraud}) \cdot \prod_{i=1}^{130} P(f_i \mid \text{fraud})$ same for ¬fraud.

©EPFL
J.-C. Chappelier & M. Rajman

# Week 7 – study case

③ From your first analysis of the baseline results, you realize that single tokens do not adequately capture dependencies that clearly appear at the syntactic level (for instance the one between "*don't*" and "*pay*" in the former example).
Using some syntactic parser, you are able to transform the former example sentence

$P(\neg fraud)$

*My salary is about 10'000 CHF and I don't pay any tax.*

into:

$P(f'_i \mid fraud)$

*SALARY-10K-RANGE not_pay tax* $\quad f'$

**a)** What probability would then be computed as the resulting output by the Naive Bayes model in such a case?

**b)** Compared to former Naive Bayes model, what is <u>the</u> main fundamental reason why you can reasonably expect the results to be better?

$$P(C) \cdot \prod_{i=1}^{3} P(f'_i \mid C) \qquad C \in \{fraud, \neg fraud\}$$

©EPFL