LANGUAGE DISCRIMINATION AND CLUSTERING

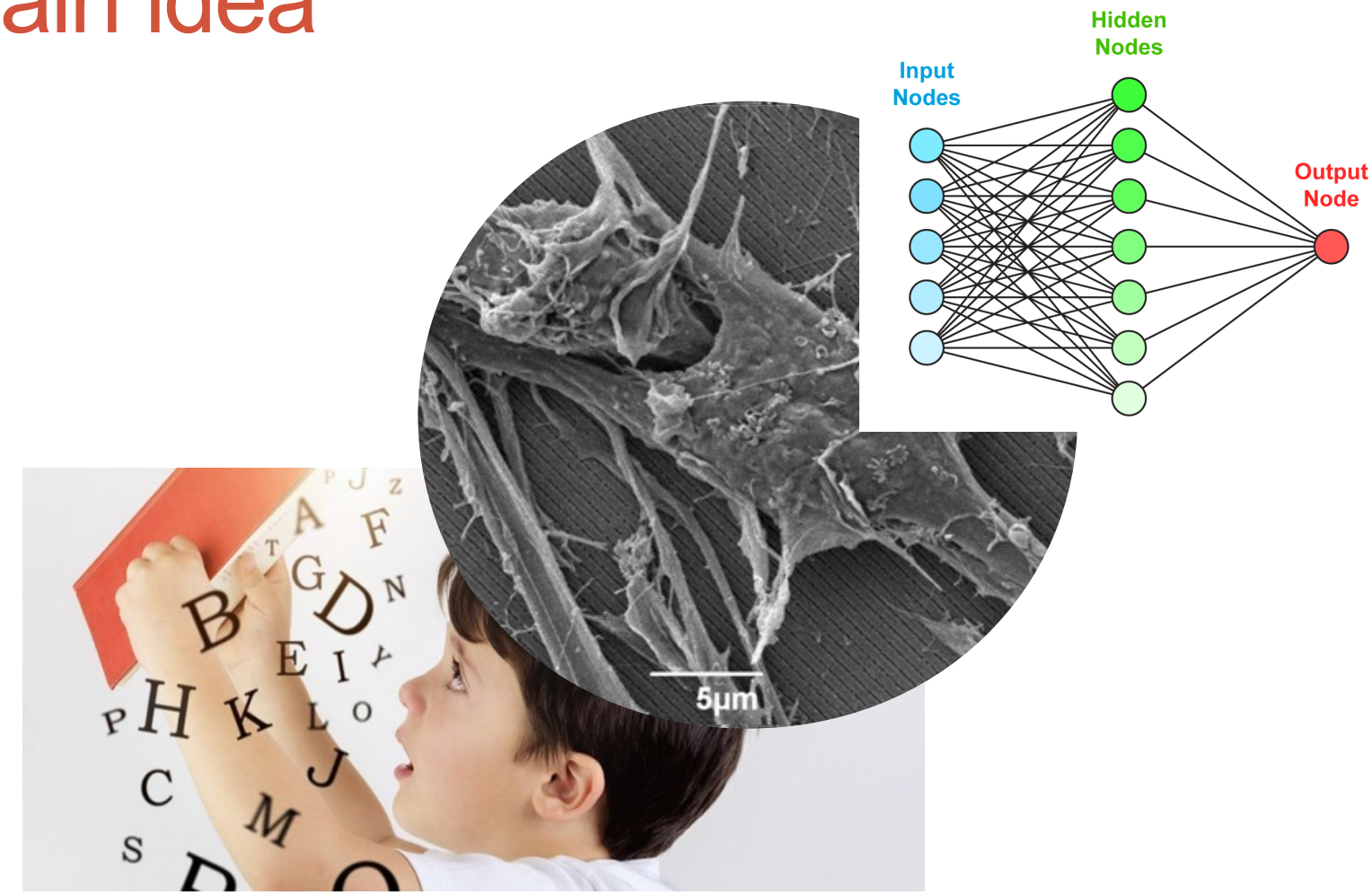
Angelo Mariano

in collaboration with
Giorgio Parisi (UNIRoma1) and Saverio Pascazio (UNIBa)

Overview

- Context description
- Model
- Results
- Ultrametricity
- Conclusions

Main idea



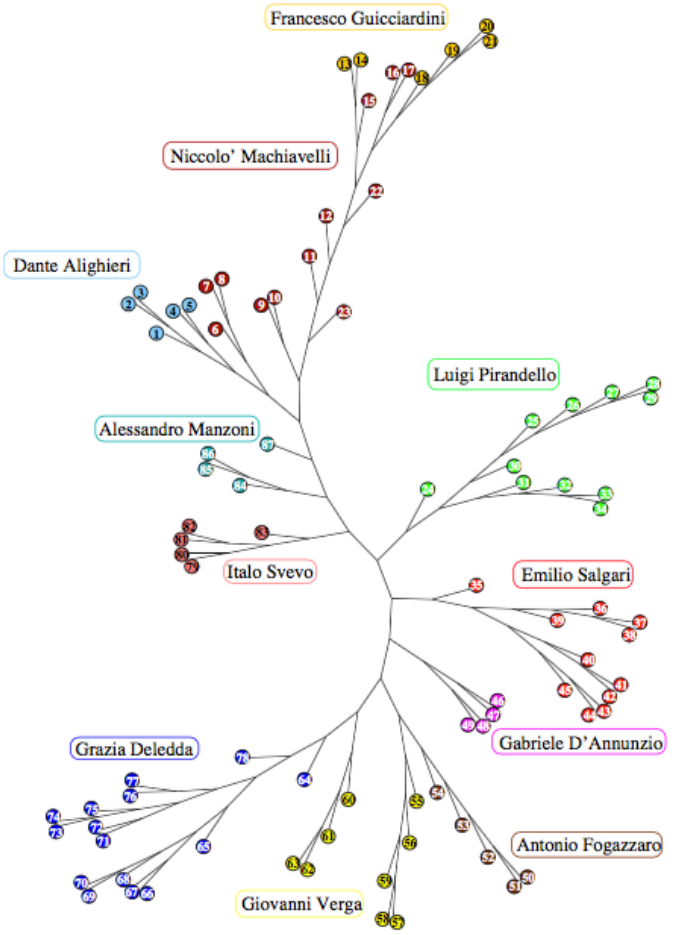
How does language training and discrimination work?

Context

Language trees and zipping, D.Benedetto, E.Caglioti, and V.Loreto, Phys. Rev. Lett. **88**, 048702 (2002)

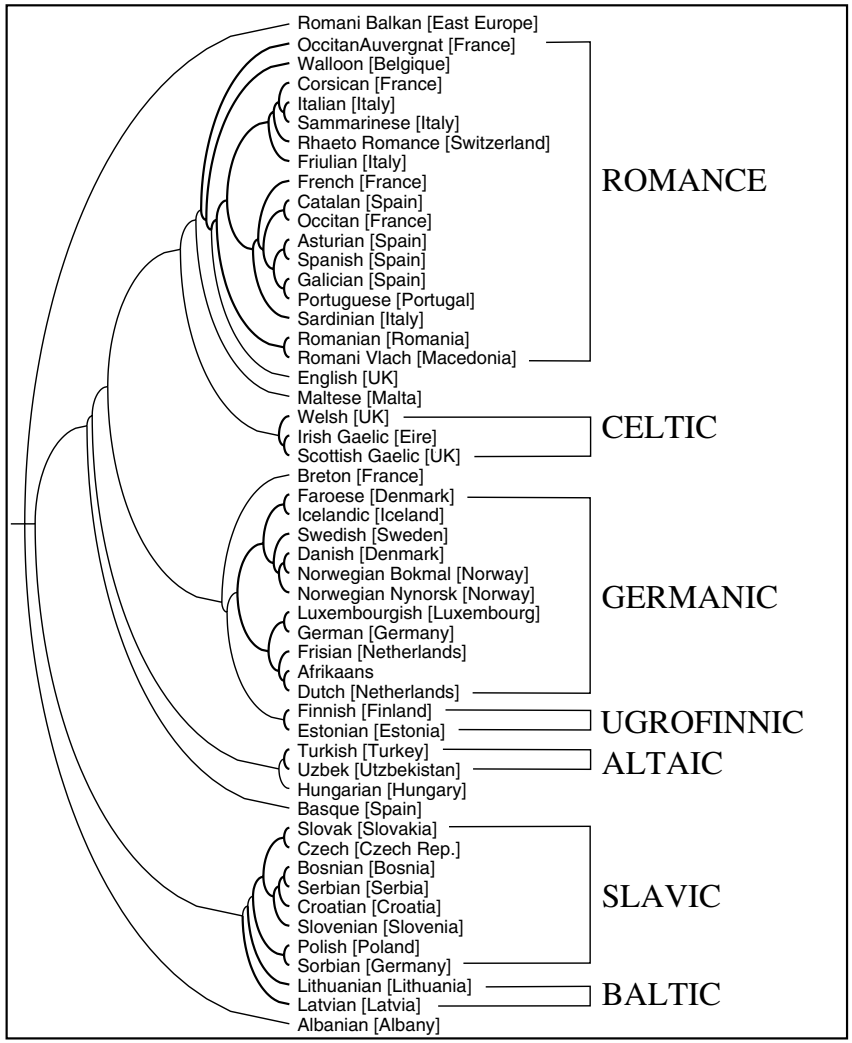
AUTHOR	N. of texts	N. of successes 1	N. of successes 2
Alighieri	8	8	8
D'Annunzio	4	4	4
Deledda	15	15	15
Fogazzaro	5	4	5
Guicciardini	6	5	6
Macchiavelli	12	12	12
Manzoni	4	3	4
Pirandello	11	11	11
Salgari	11	10	10
Svevo	5	5	5
Verga	9	7	9
TOTALS	90	84	89

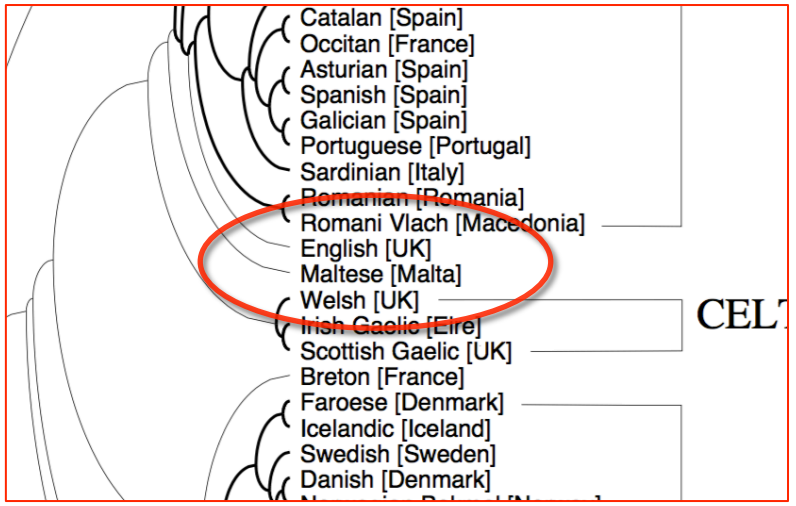
TABLE I: **Authorship attribution**: For each author depicted we report the number of different texts considered and two measures of success. Number of success 1 and 2 are the numbers of times another text of the same author was ranked in the first position or in one of the first two positions respectively.



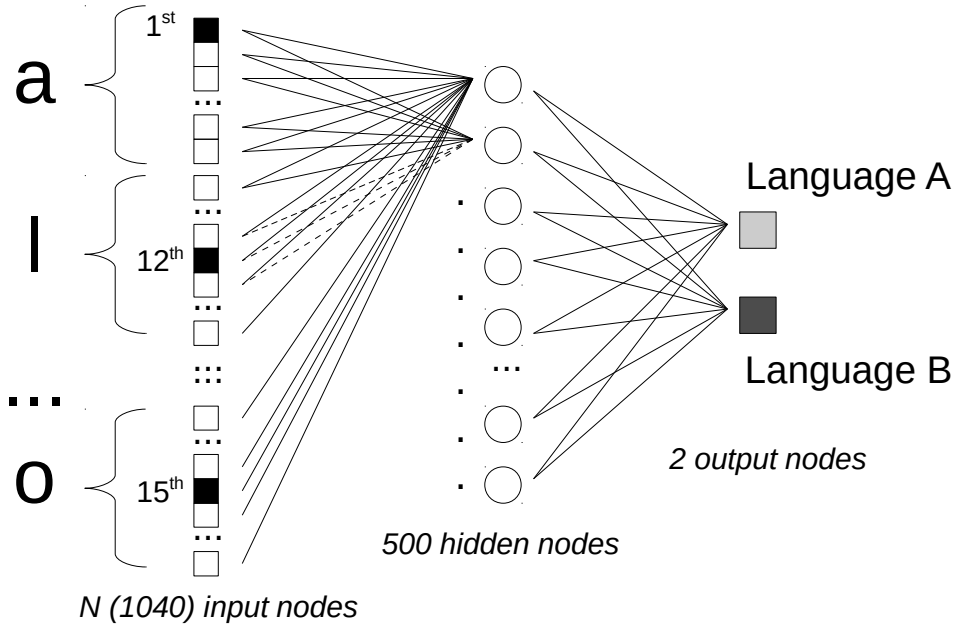
Context

Language trees and zipping, D.Benedetto, E.Caglioti, and V.Loreto, Phys. Rev. Lett. **88**, 048702 (2002)





Our model



- Feed-forward neural network with backpropagation
- Source: 20k sentences (10k for language) from 21 indo-european languages taken from Leipizig Linguistic Corpora http://corpora.uni-leipzig.de/
- Short sentences of 40 chars:

  “Alice thought she might as well wait, as she had nothing else to do, and perhaps after all it might tell her something worth hearing.”

  becomes
  “alicethoughtshemightaswellwaitasshehadno”
- N = 26 * 40 input nodes represented by ASCII chars

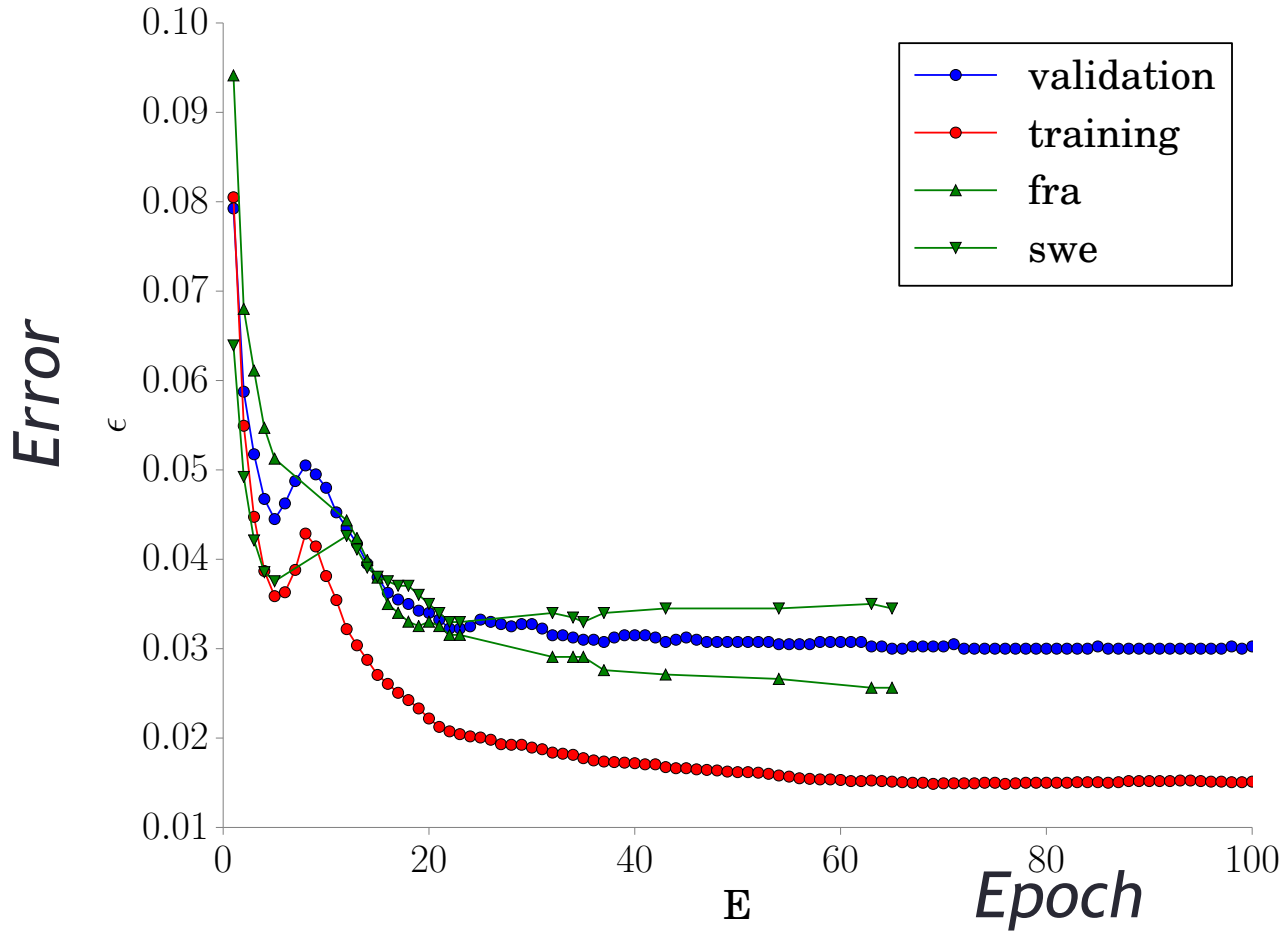
Our model

Artificial Neural Network:

$$f(x) = G(b^{(2)} + W^{(2)}(s(b^{(1)} + W^{(1)}x)))$$

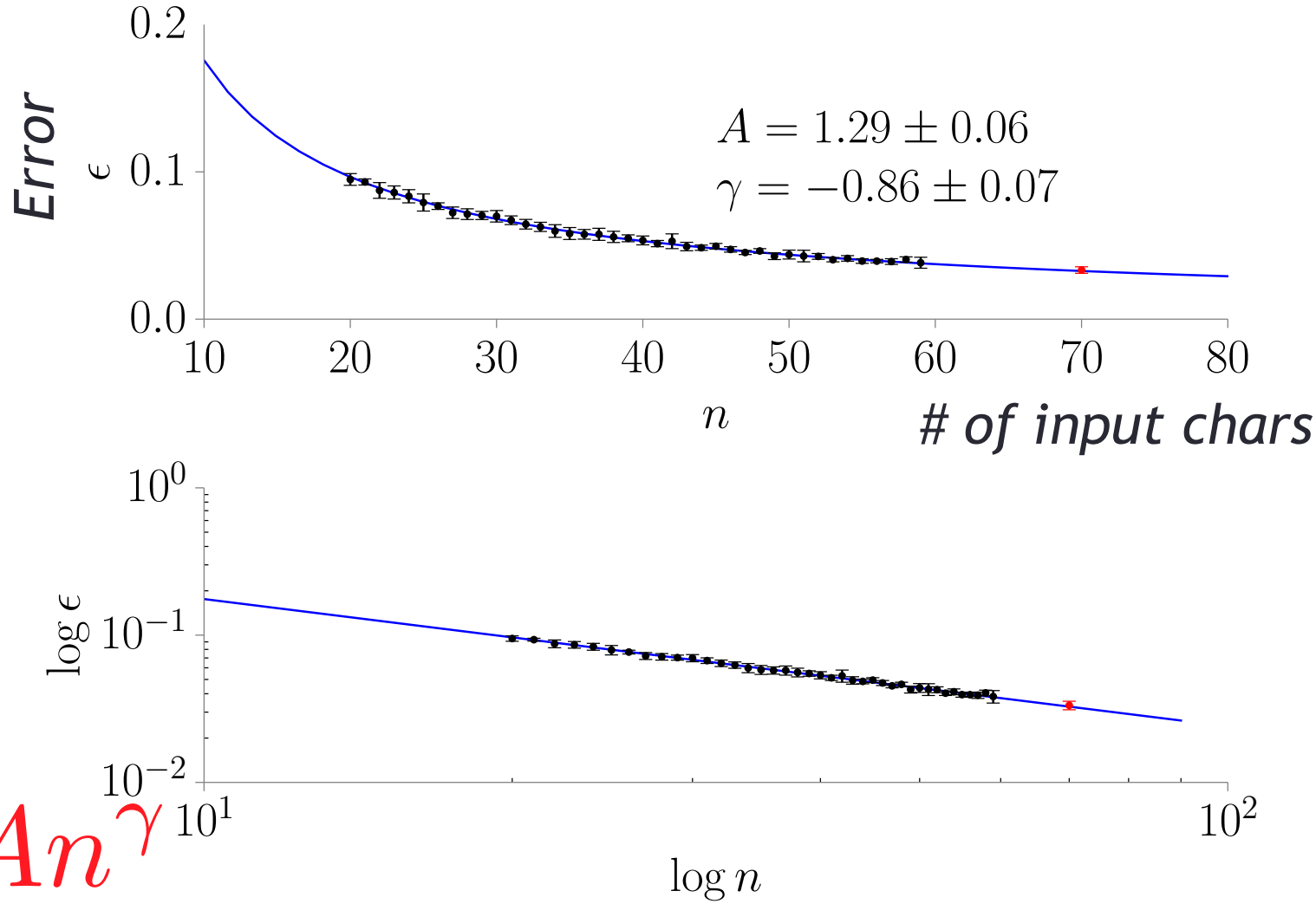
- G and s activation functions:
  - s = tanh
  - G = softmax
- b bias vectors
- W transfer matrices
- At every epoch params b and W are updated according to minimizing a cost function

Results: training



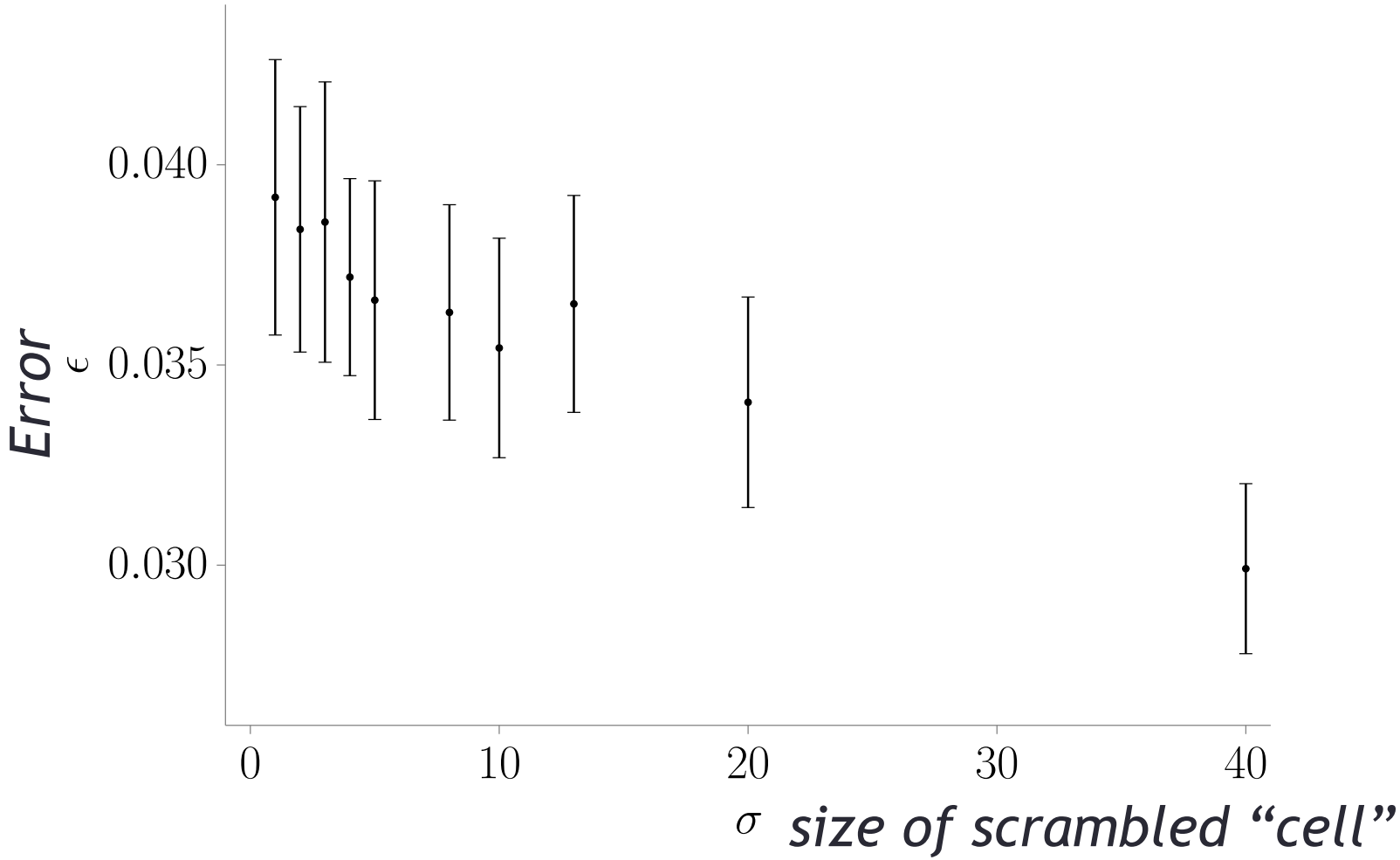
They are averaged over a 5-fold cross-validation technique

Results: dependency on input nodes



$$\epsilon = An^{\gamma}$$

Results: dependency on sequence of chars



Towards a distance

Let us define the following distance:

$$d = \frac{1}{\epsilon} - 1 \in \mathbb{R}_+$$

A distance satisfies:

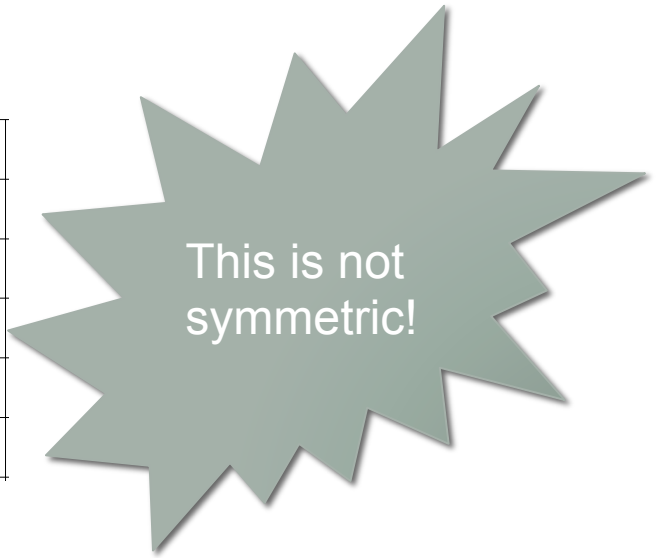
$$\begin{aligned} &d(x,y) = 0 \quad \iff \quad x = y \\ &d(x,y) = d(y,x) \\ &d(x,y) \leq d(x,z) + d(z,y). \quad \forall z \in \mathcal{S} \end{aligned}$$

A metric is called an ultrametric if it satisfies:

$$d(x,z) \leq \max(d(x,y), d(y,z))$$

Towards a distance

ε	cat	ces	dan	deu	ell
cat	94.93	1.07	2.75	2.92	0.69
ces	2.56	94.39	2.72	3.16	1.11
dan	2.91	2.03	94.41	6.24	1.15
deu	2.32	2.16	8.44	96.13	0.83
ell	0.83	0.35	0.82	0.82	99.19



Linkage algorithms

- There are many ways of clustering in order to find philogenetic structure:
  - complete
  - single
  - centroid
  - average where distance between two custers U and V is

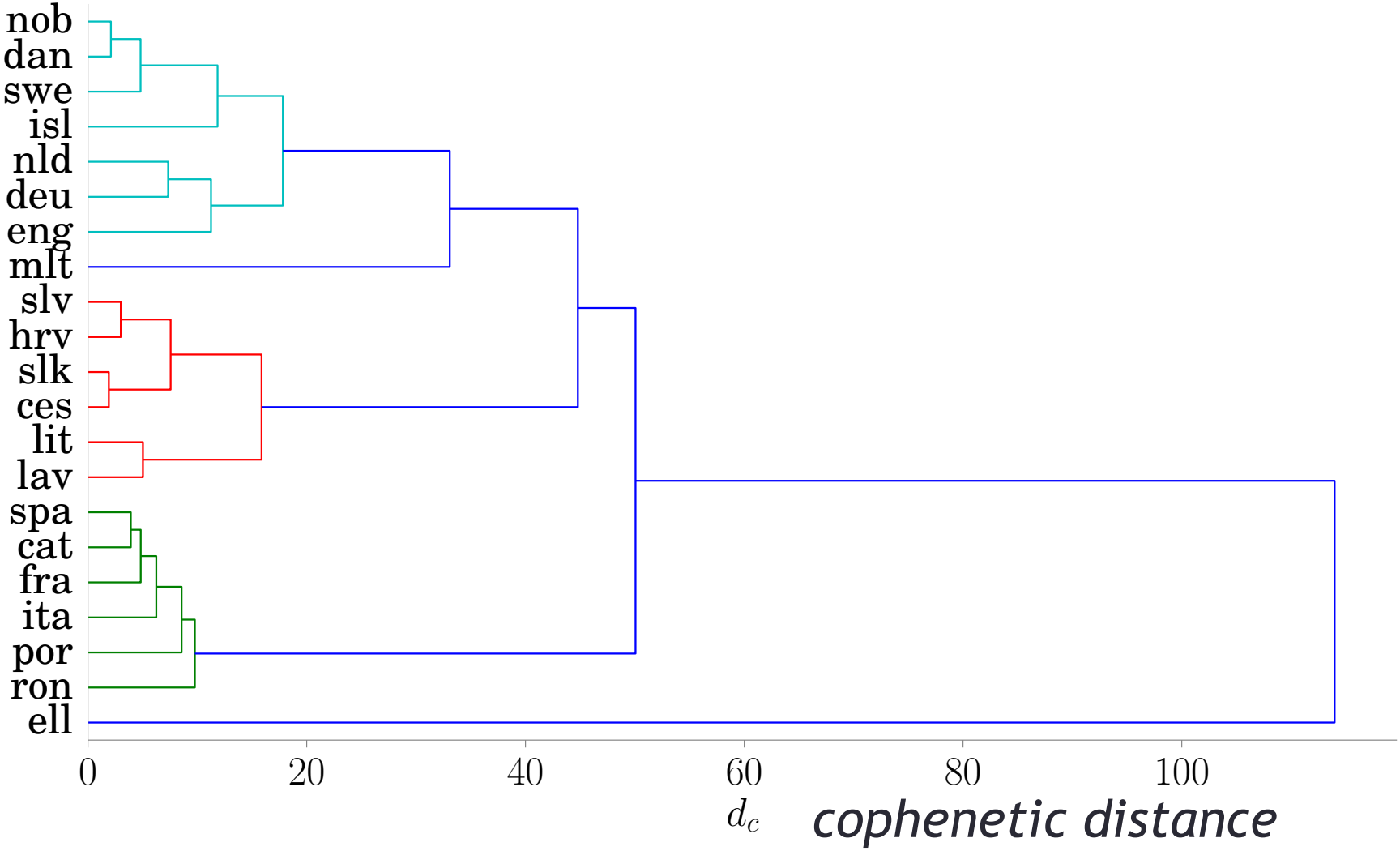
$$d(U,V) = \sum_{\substack{i \in U \\ j \in V}} \frac{d(i,j)}{|U| * |V|}$$

- A good measure is the cophenetic coefficient[1]:

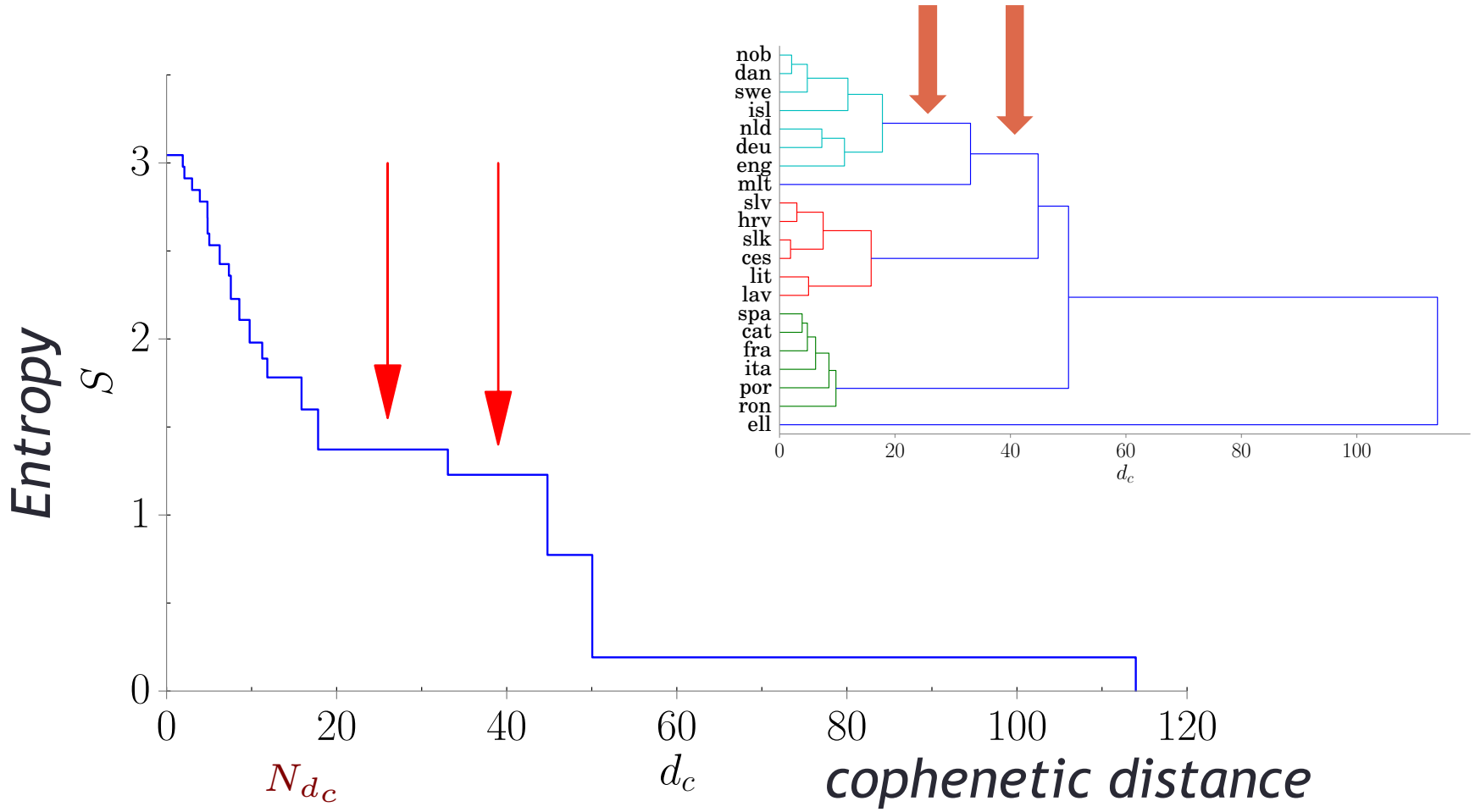
$$c = \frac{\sum_{i<j}(d(i,j) - \langle d \rangle)(d_c(i,j) - \langle d_c \rangle)}{\sqrt{\sum_{i<j}(d(i,j) - \langle d \rangle)^2 \sum_{i<j}(d_c(i,j) - \langle d_c \rangle)^2}}$$

[1] *A Quantitative Clustering Approach to Ultrametricity in Spin Glasses*, S.Ciliberti and E.Marinari, J. Stat. Phys. **115**, 2004

Our analysis: language dendrogram



Our analysis: dendrogram cut



$$S(d_c) = -\sum_{k=1}^{N_{d_c}} P_{d_c}(k) \ln P_{d_c}(k)$$

Conclusions and perspectives

- Simple model (implemented from little or zero background code)
- Hard task of classification
- Open to improvements
- Can be applied to any sequence of symbols

Thank you for your attention