Hierarchical Text Categorization for MALACH

J. Scott Olsson
olsson@math.umd.edu

w/ Doug Oard of UMIACS, CLIS

MALACH

Multilingual Access to Large spoken ArCHives

- UMD, JHU, IBM, Charles University, the University of West Bohemia, the Visual History Foundation
- Many areas of study (automatic speech transcription, machine translation, cross-language information retrieval, text categorization, etc.)

Goal: "to improve access to large multilingual collections of recorded speech in oral history archives"

The Shoah VHF Dataset

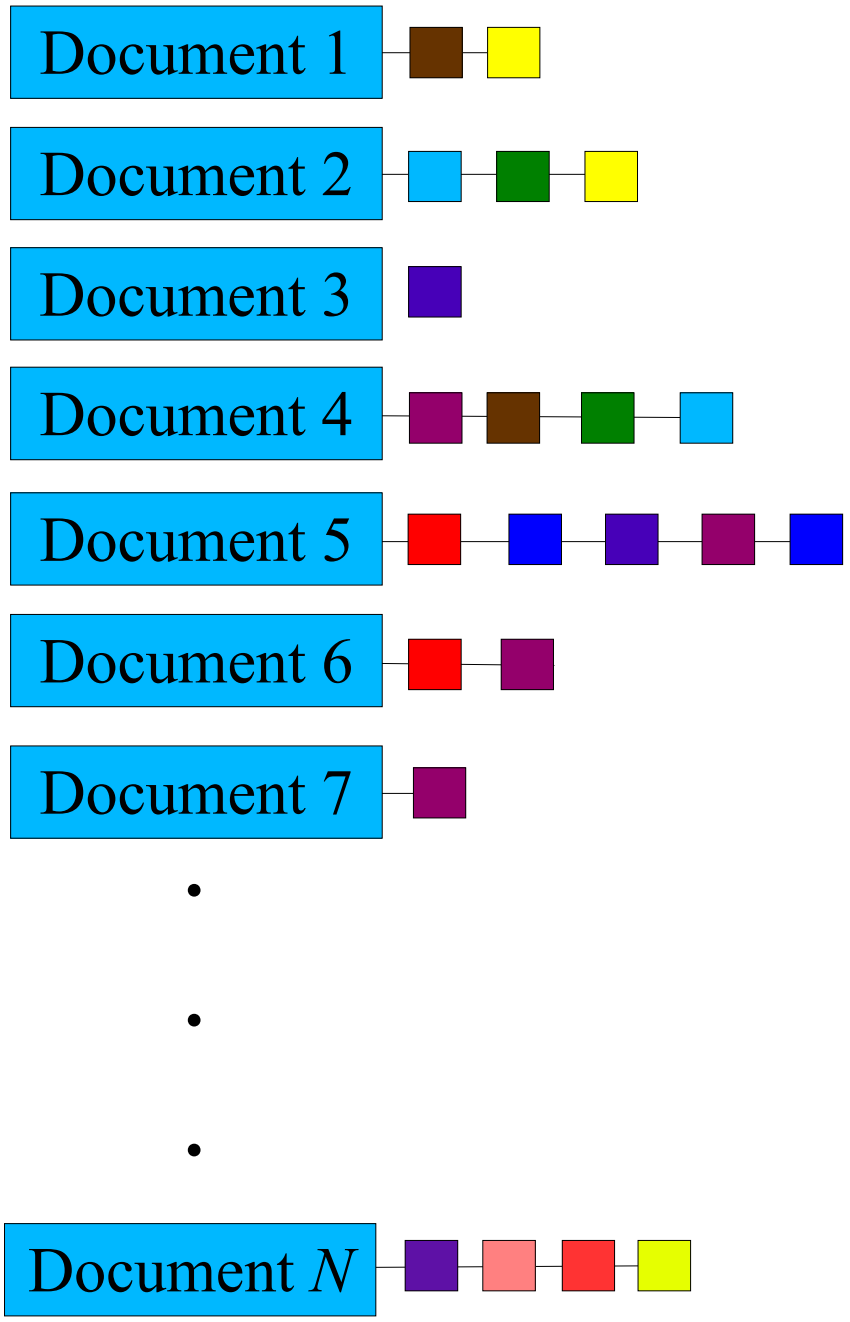
Audiovideo testimonies of Holocaust survivors

Unique features: size, coherency

- 51,649 testimonies, roughly 2.5 hours each, in 32 languages.
- The largest, most complex, coherent digital video library on Earth (180+ Tb).

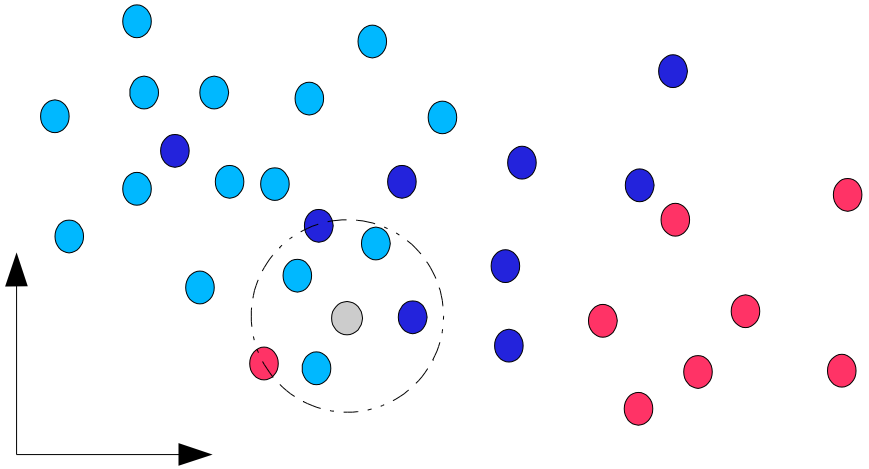
This differs *significantly* from other available datasets.

Automatic Text Categorization

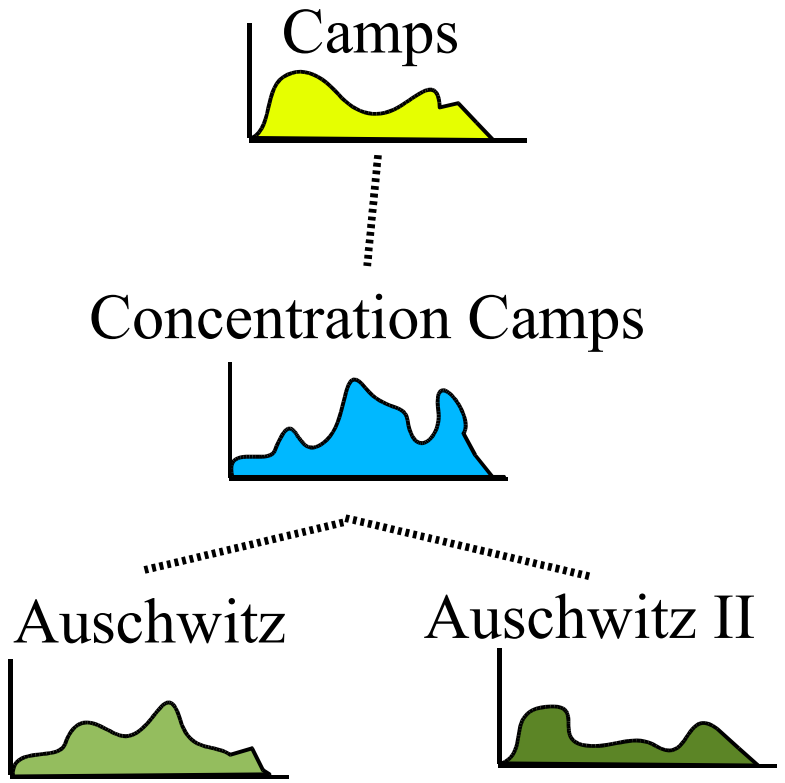


Two Models

A Flat Model: *k* Nearest Neighbors



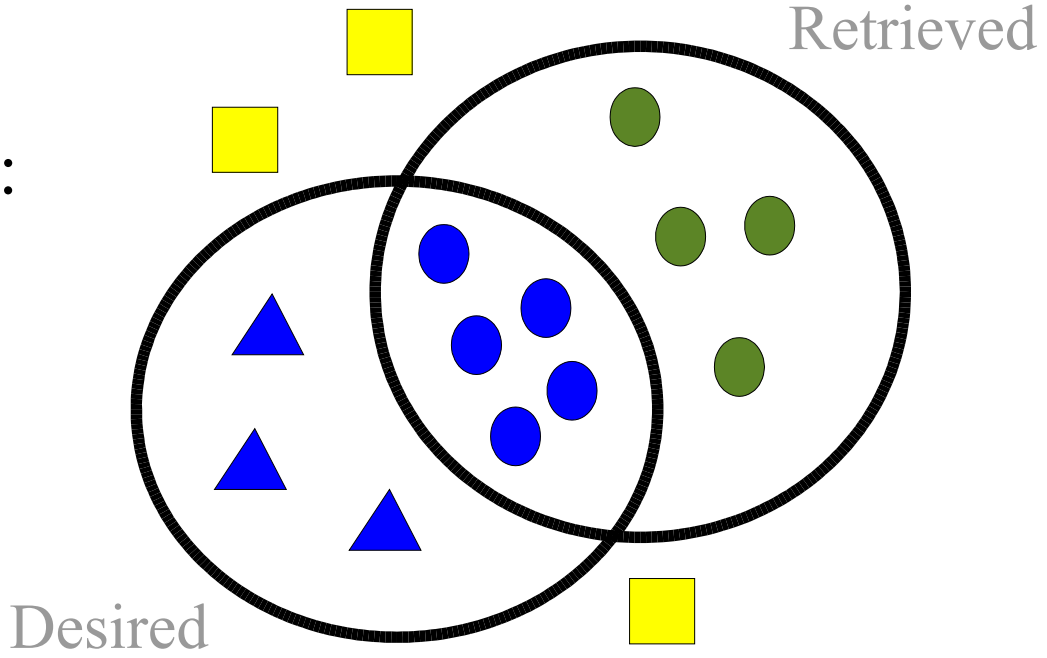
A Hierarchical Model: HPLC



Measuring Success

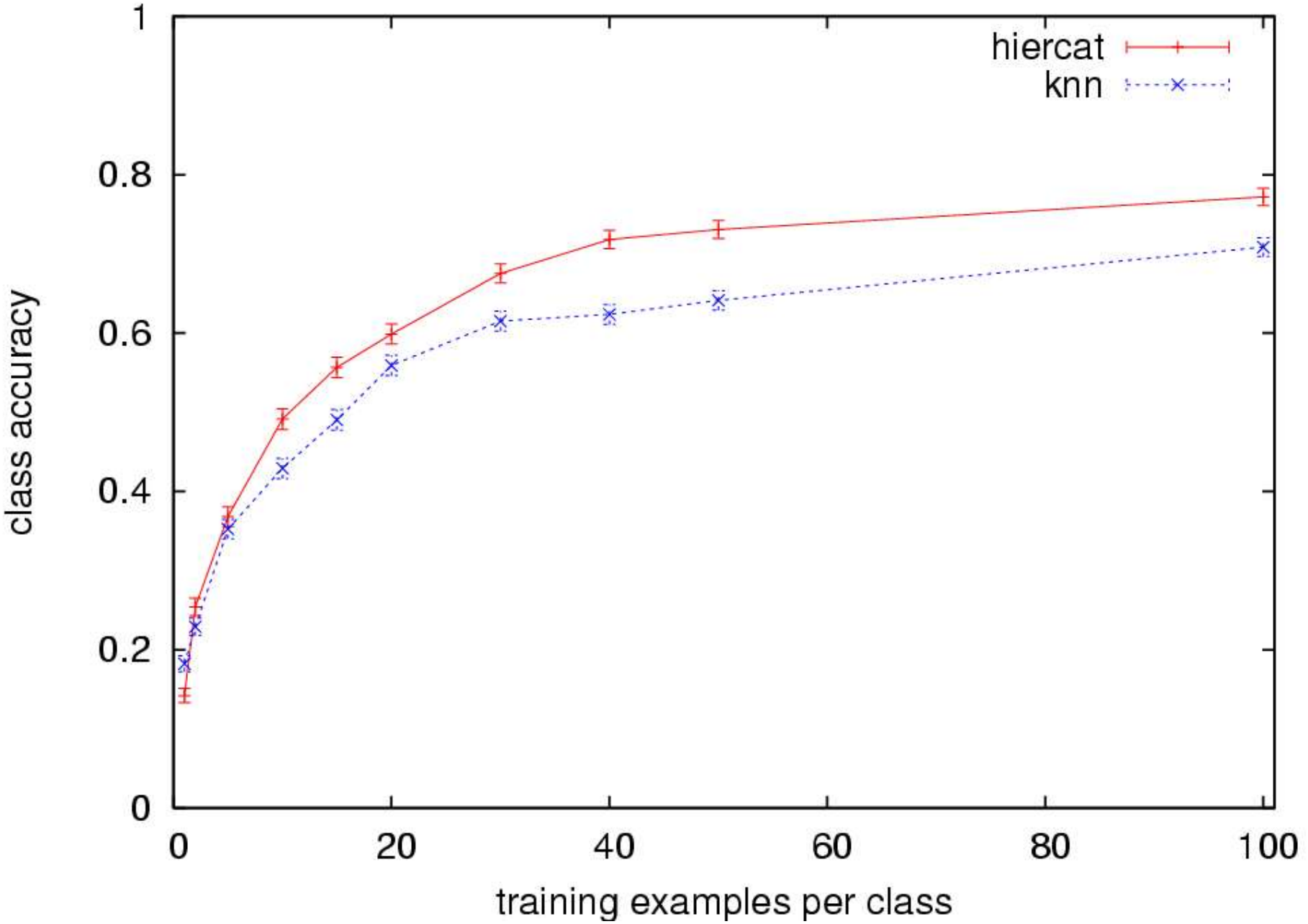
Binary measures of success:

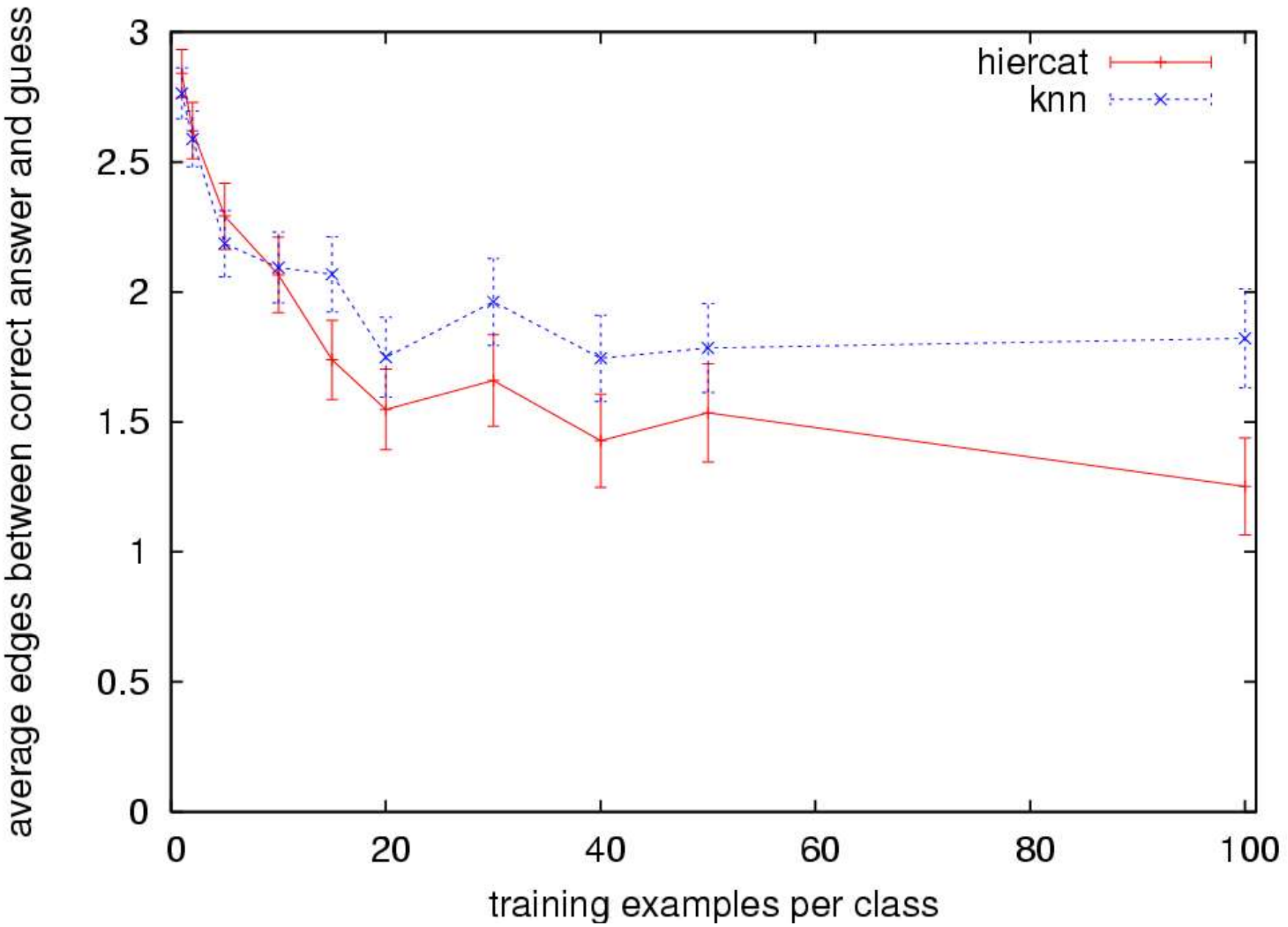
- accuracy
- precision, recall
- F-measure

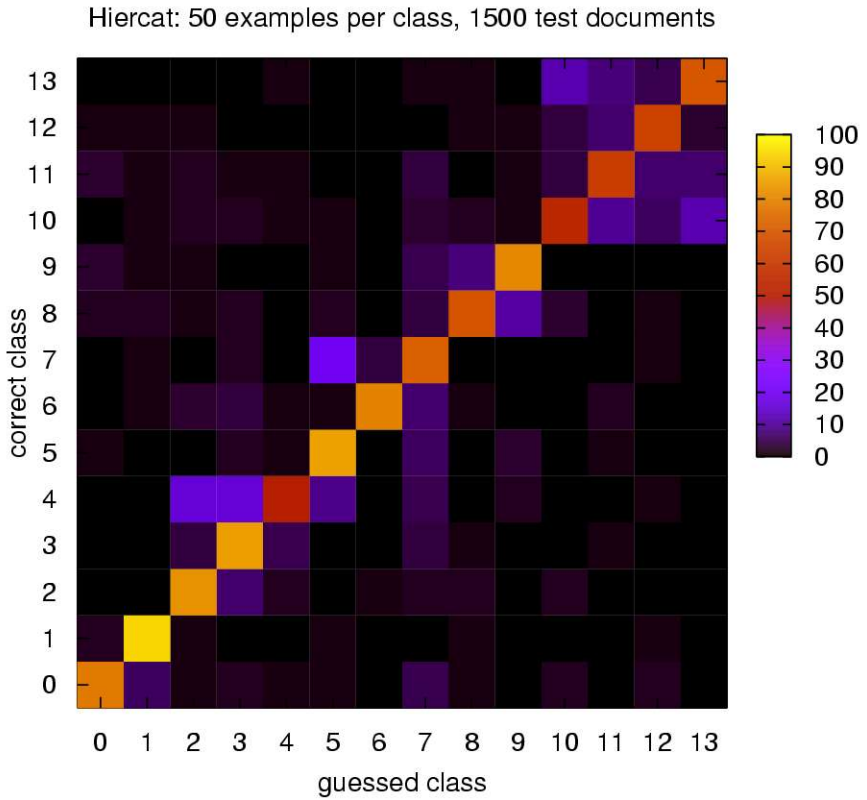


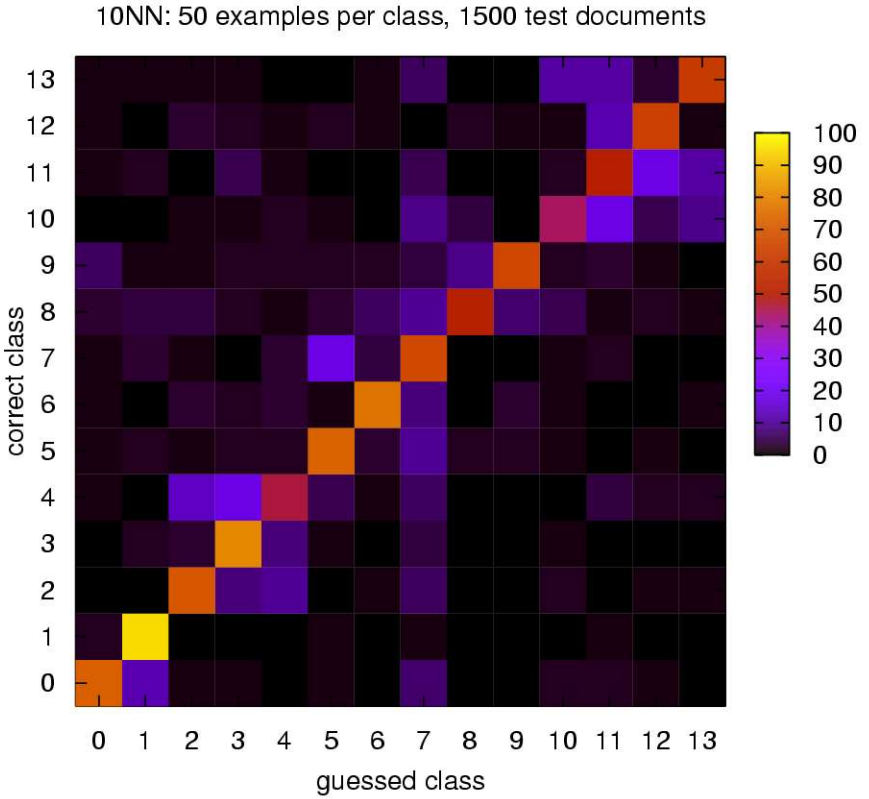
We can do better: hierarchical measures

How *close* is our guess to the correct answer?


hiercat
knn
class accuracy
training examples per class
0
0.2
0.4
0.6
0.8
1
0
20
40
60
80
100


average edges between correct answer and guess
training examples per class
hiercat
knn
0
0.5
1
1.5
2
2.5
3
0
20
40
60
80
100





0 rec.sport.baseball
1 rec.sport.hockey
2 alt.atheism
3 soc.religion.christian
4 talk.religion.misc
5 talk.politics.guns
6 talk.politics.mideast
7 talk.politics.misc
8 rec.autos
9 rec.motorcycles
10 comp.graphics
11 comp.os.ms-windows.misc
12 comp.sys.ibm.pc.hardware
13 comp.windows.x

Questions?

olsson@math.umd.edu

http://www.math.umd.edu/~olsson/hiercat/