models and algorithms #2

MGMT|from text analysis to actionable knowledge

kristoffer l nielbo
kln@cas.au.dk

DTL|IMC|AARHUS UNIVERSITY-CAS

IMC
INTERACTING MINDS CENTRE

the goal of **statistical learning** is to build a machine that can learn from data and automatically make the right decisions

---

unsupervised
identify latent classes in the data $\rightarrow$ lack theoretical 'ground truth'

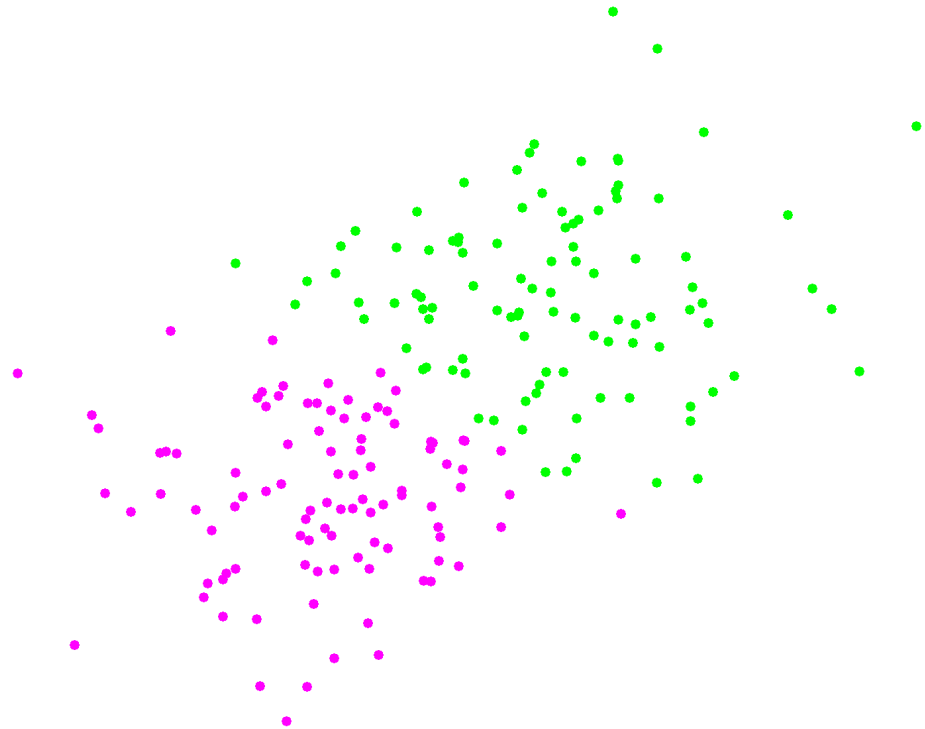
---

supervised
infer mapping between data & class-information $\rightarrow$ theoretical 'ground truth'

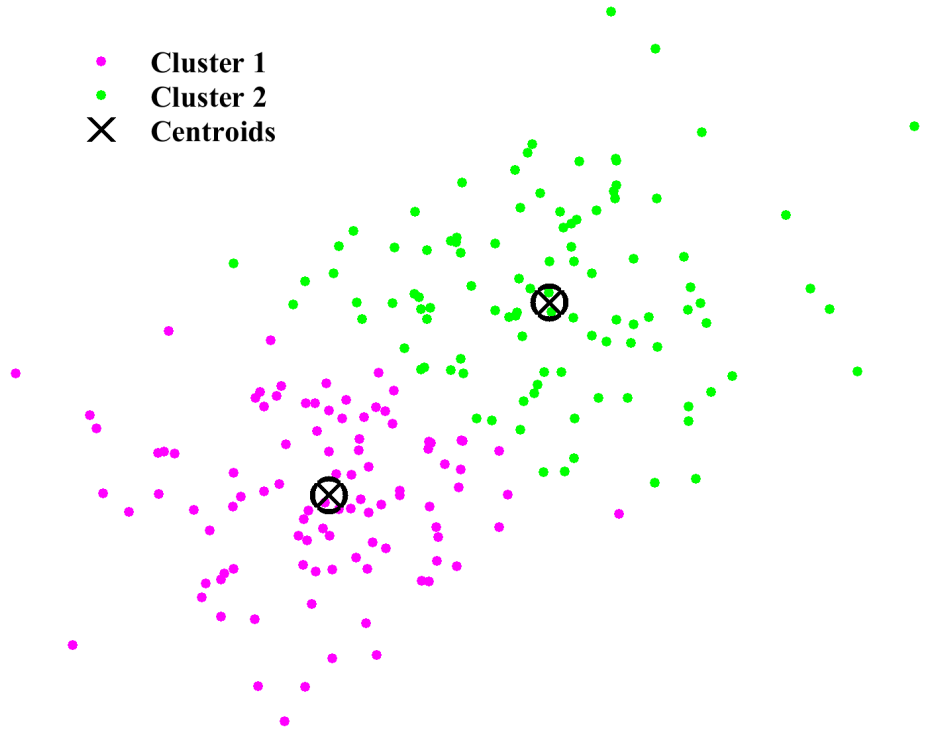
unsupervied learning|clustering
there are often groups of features or latent variables that identity subsets in a collection of documents

the 'clusters' can be used for understanding (e.g., thematic analysis) and utility (e.g., prototype documents)

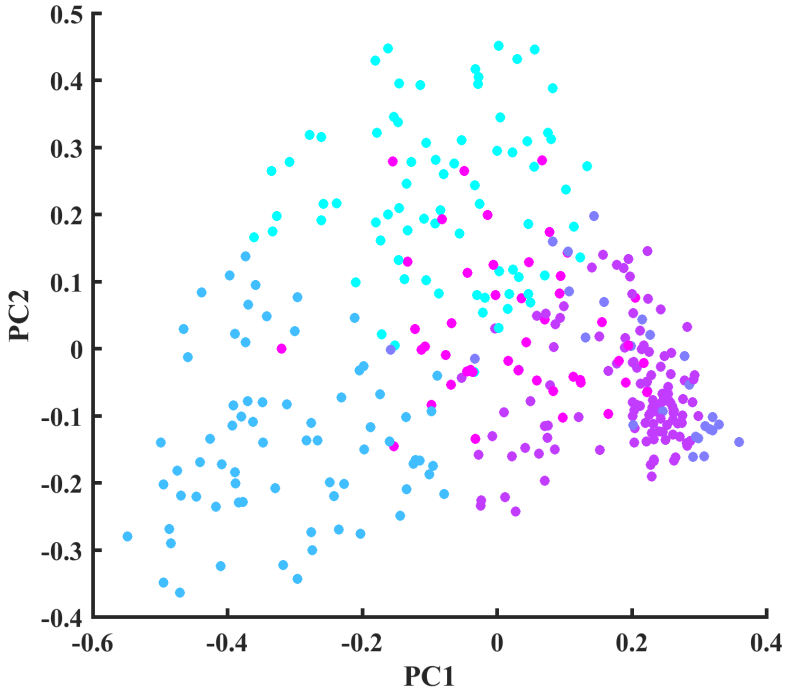
implicit assumption that we study differences in variables (e.g., terms) between homogeneous objects (e.g., documents)

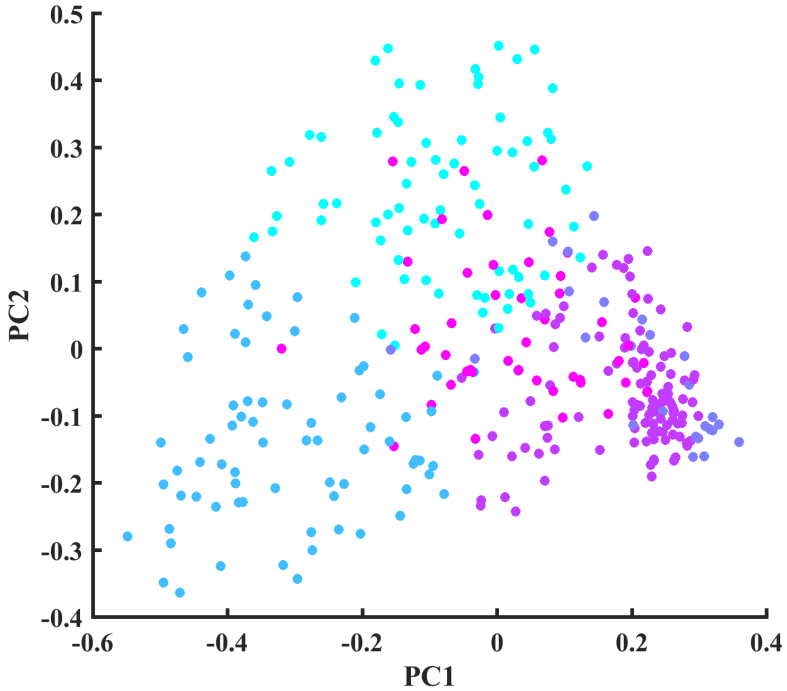


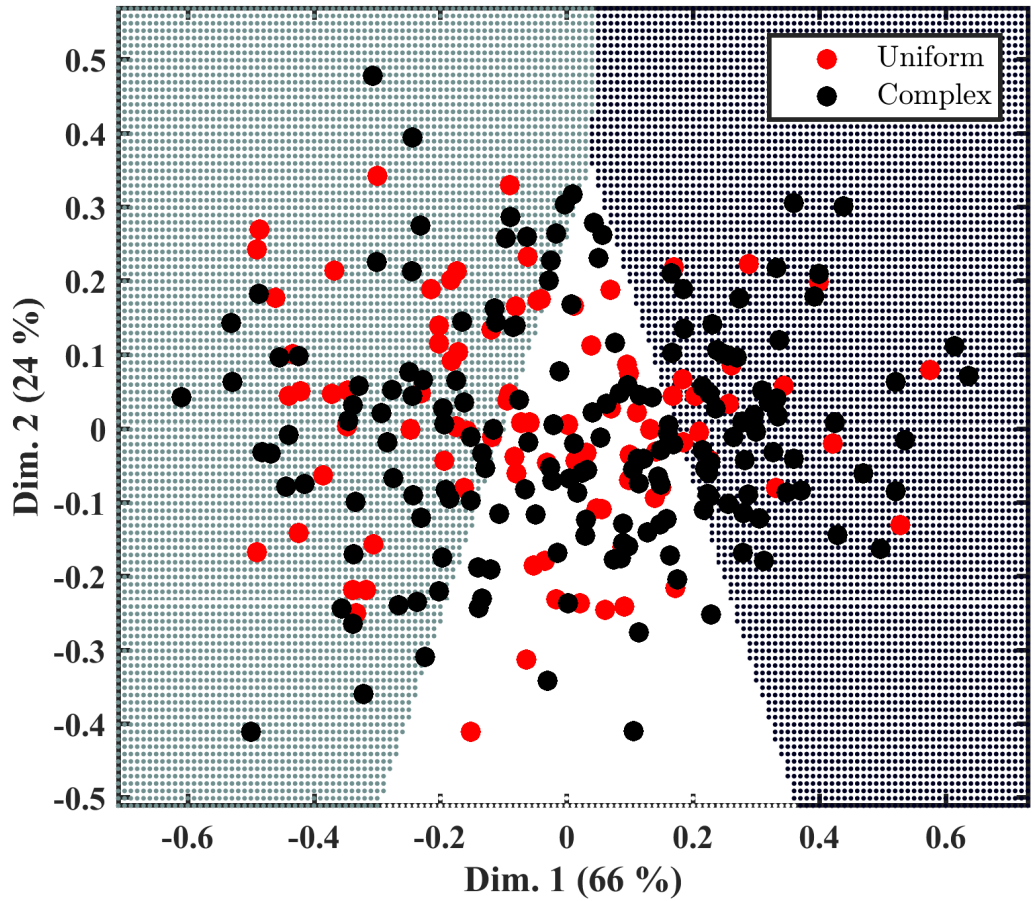
systematic differences between documents result in non-random sub-corpora (e.g., genre, author characteristics, community effects)

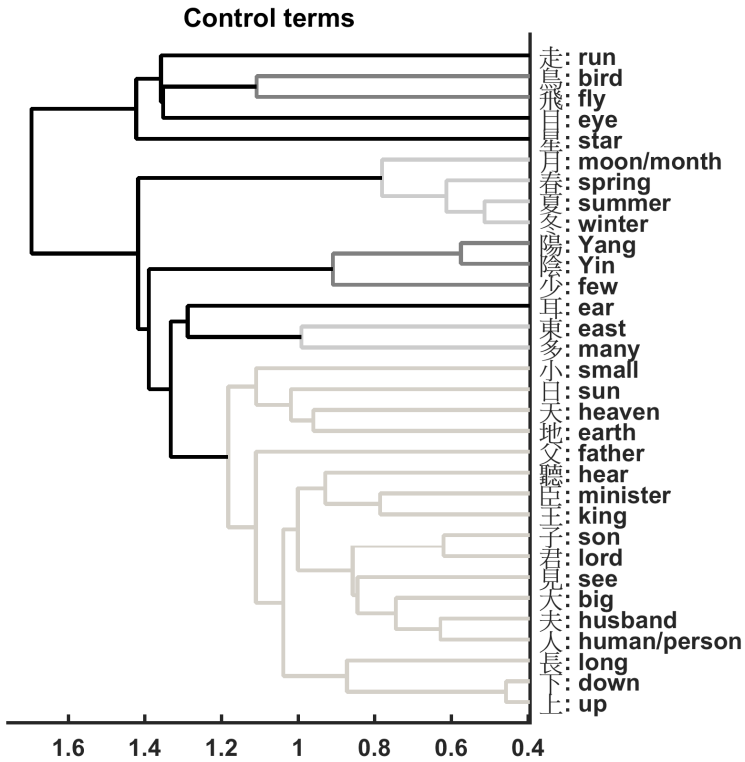


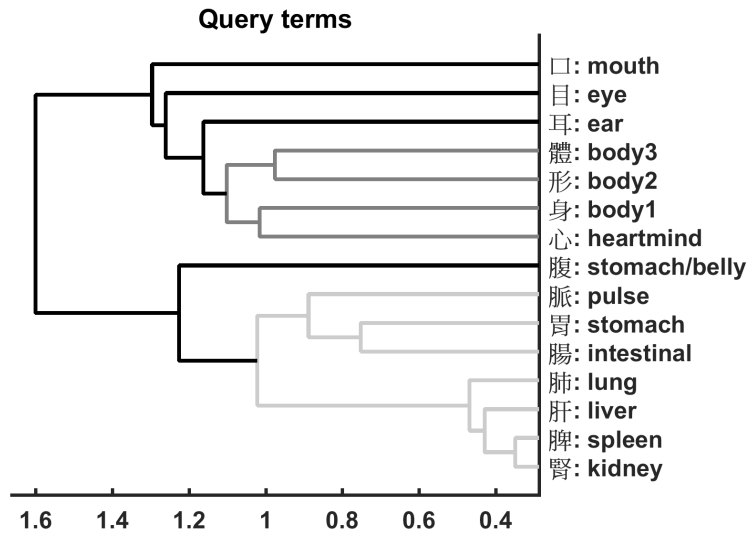
partition documents (or other objects) into homogeneous subsets using document similarity/distance from prototype


0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
PC2
-0.6
-0.4
-0.2
0
0.2
0.4
PC1
KEYWORDS
brother
fight
mujahidin
ummah
women
abu
brother
children
love
people
Al Qaeda
AQAP
Inspire
issue
magazine
America
attack
bomb
people
world
Allah*
Islam
message
Muslim
time


0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
PC2
-0.6
-0.4
-0.2
0
0.2
0.4
PC1
KEYWORDS
brother
fight
mujahidin
ummah
women
abu
brother
children
love
people
Al Qaeda
AQAP
Inspire
issue
magazine
America
attack
bomb
people
world
Allah*
Islam
message
Muslim
time

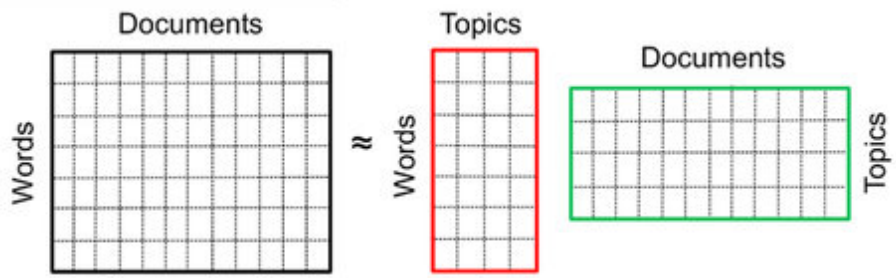

Uniform
Complex
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.5
Dim. 2 (24 %)
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
Dim. 1 (66 %)





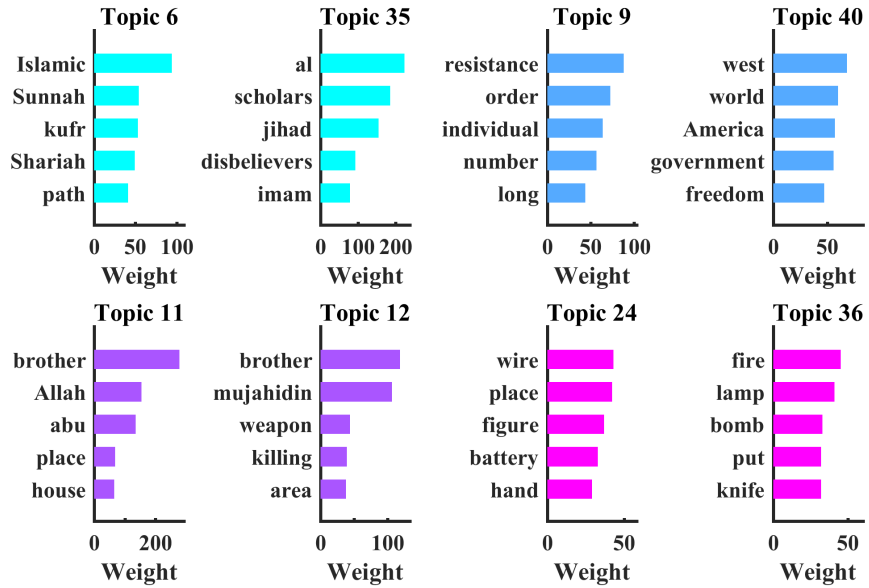
– with hierarchical clustering you cut or prune the tree at some level to define clusters

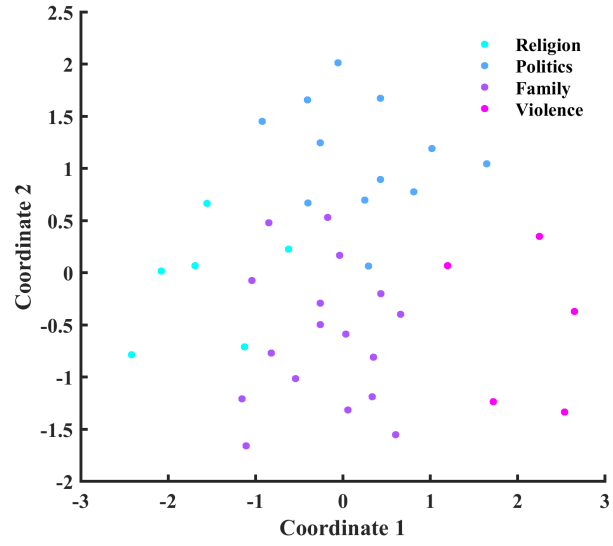
we can extract latent semantic variables that generated the documents by reverse engineering the process *words* & *doc* $\Rightarrow$ *topics*

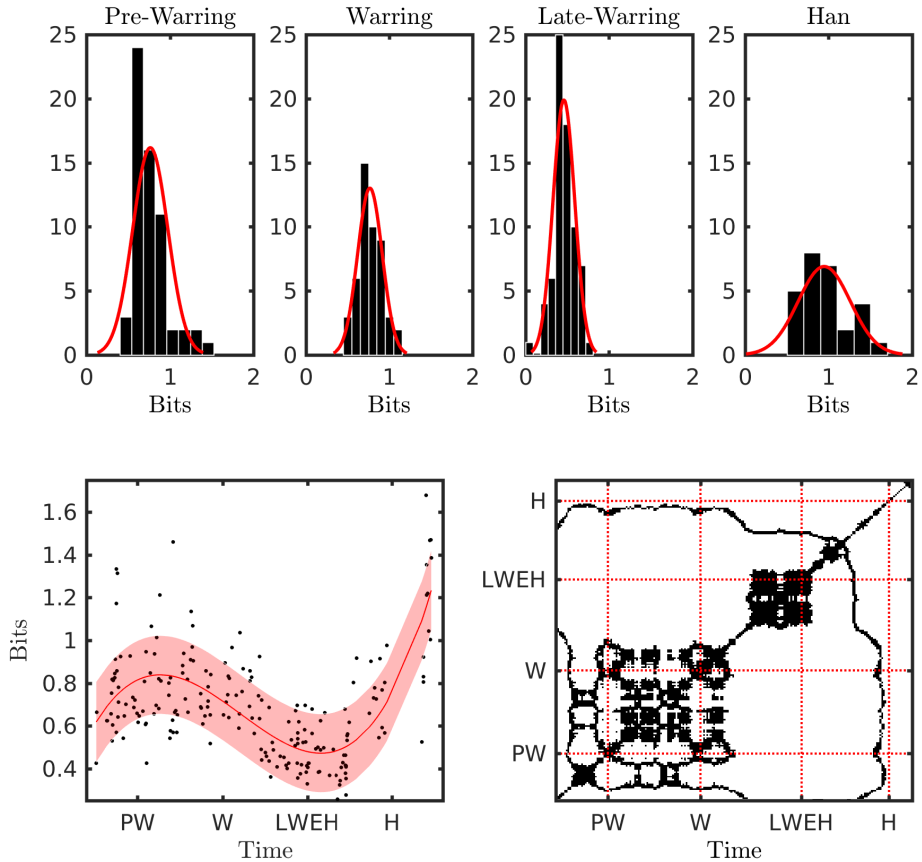


topic modeling a set of unsupervised mixed models where each document is more or less likely within each clusters

1. **discover** thematic structure
2. **annotate** documents
3. **use** the annotations to visualize, organize, summarize, ...


Topic 6
Islamic
Sunnah
kufr
Shariah
path
0 50 100
Weight
Topic 35
al
scholars
jihad
disbelievers
imam
0 100 200
Weight
Topic 9
resistance
order
individual
number
long
0 50 100
Weight
Topic 40
west
world
America
government
freedom
0 50
Weight
Topic 11
brother
Allah
abu
place
house
0 200
Weight
Topic 12
brother
mujahidin
weapon
killing
area
0 100
Weight
Topic 24
wire
place
figure
battery
hand
0 50
Weight
Topic 36
fire
lamp
bomb
put
knife
0 50
Weight


Religion
Politics
Family
Violence
Coordinate 2
Coordinate 1
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-3
-2
-1
0
1
2
3



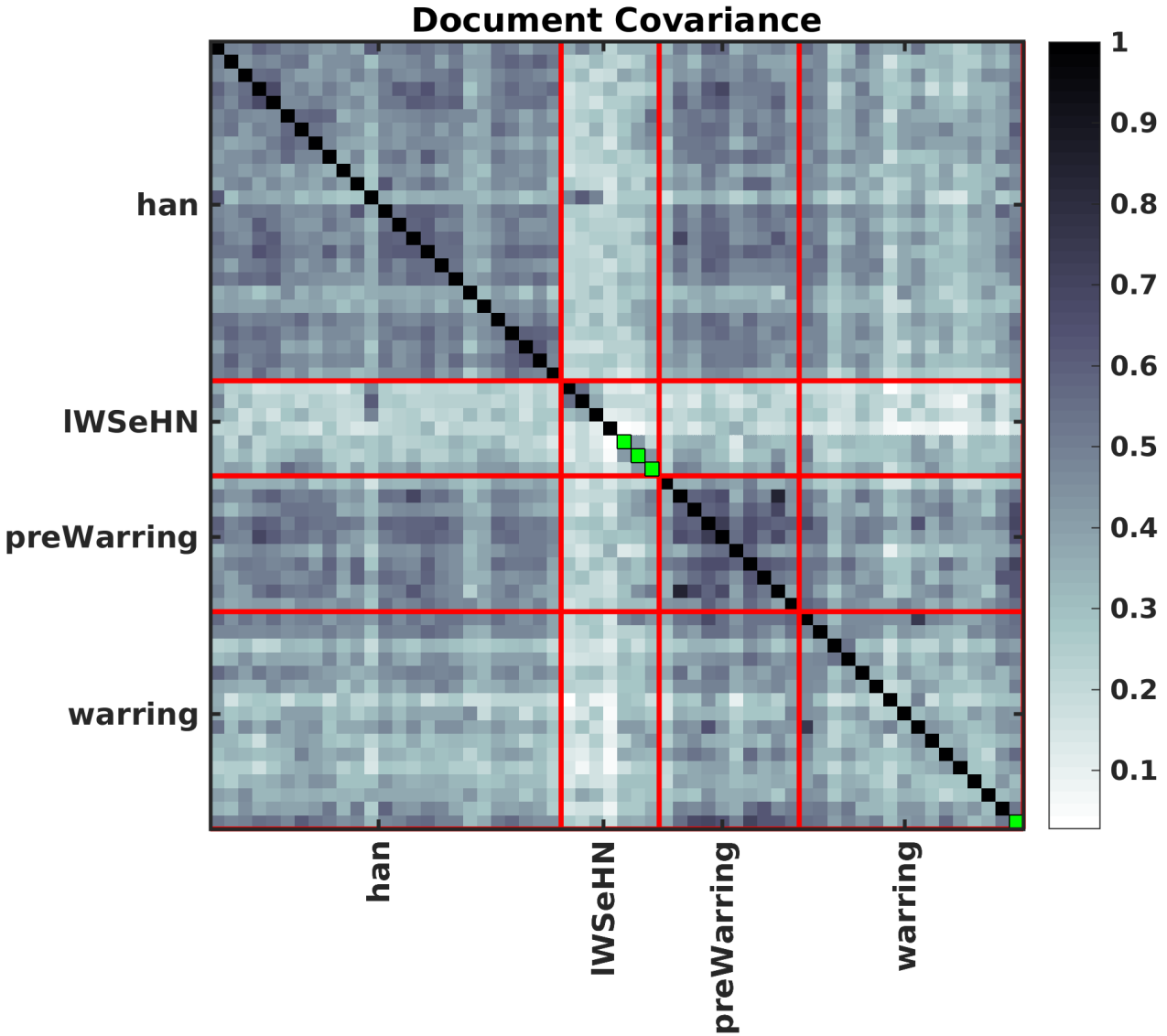
represent each document as a distribution on topics: $\theta_{d=1\ldots M}$

semantic innovation $\sim$ relative entropy between documents:
$D_{KL}(P \parallel Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$

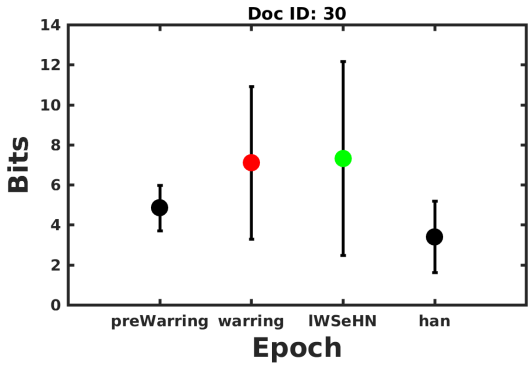
supervised learning|classfication
documents often come with associated metadate in the form of classes or labels (e.g., genres, departments, companies) $\rightarrow$ approximate a function that can map documents features onto class information

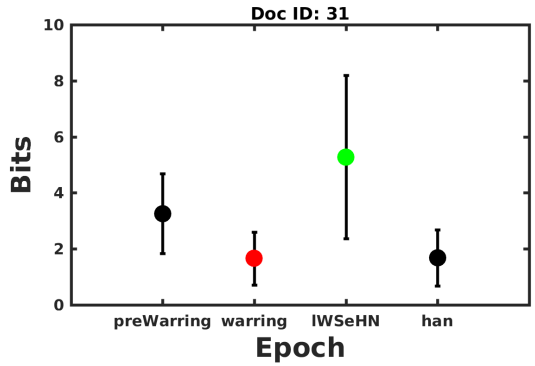
train a classifier on labeled documents, test on unseen documents, and generalize to new instances

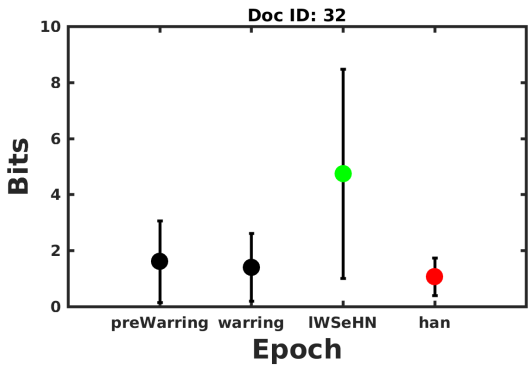
while clustering (unsupervised learning) searches for groups within the corpus, classification learns to map a collection of documents onto a categorical class values

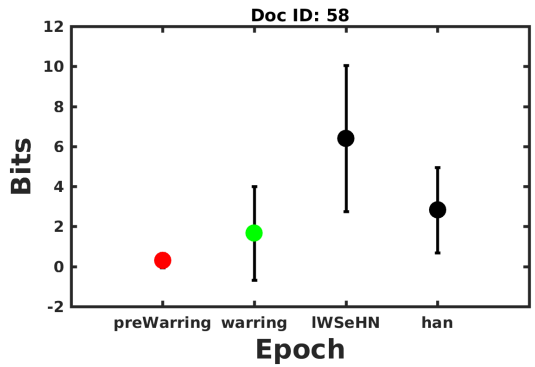


IMC
INTERACTING MINDS CENTRE


Doc ID: 30
Bits
Epoch
preWarring
warring
lWSeHN
han


Doc ID: 31
Bits
Epoch
preWarring
warring
lWSeHN
han


Doc ID: 32
Bits
Epoch
preWarring
warring
lWSeHN
han


Doc ID: 58
Bits
Epoch
preWarring
warring
lWSeHN
han