Class-based Identification of 'Deviant' Semantic Features in Historical Corpora

Kristoffer L. Nielbo
kln@cas.au.dk

DTL|IMC|AU-CAS
Aarhus University, Denmark

statistical learning

goal
build a machine that can learn from data and automatically make the right decisions

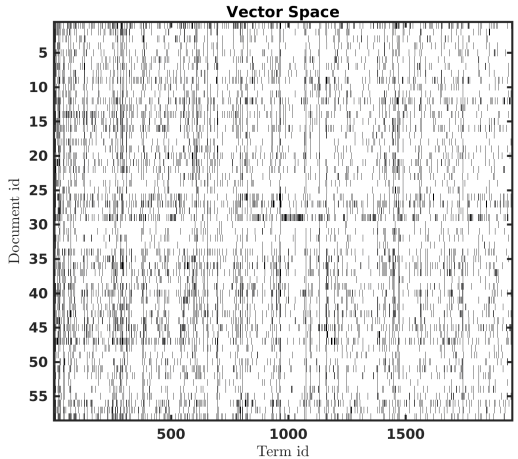
supervised
infer mapping between data & class-information $\rightarrow$ theoretical 'ground truth'
unsupervised
identify latent classes in the data $\rightarrow$ lack theoretical 'ground truth'

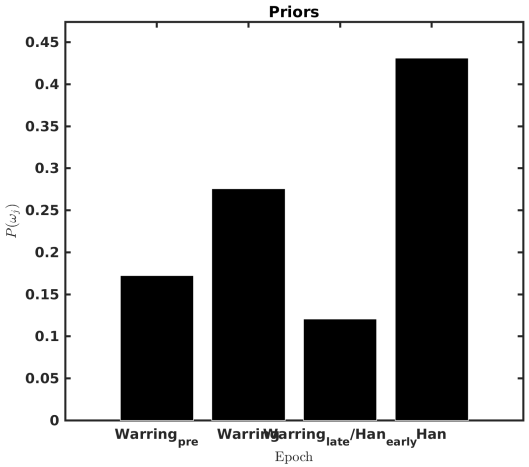
application to ctext corpus

goal
combine statistical learning and information theory in order to explore semantic relations between data (*Shangshu*) and theoretical classes (document dating)

naïve bayes
simple and well-performing Bayesian model for supervised learning, but *too constrained*
latent dirichlet allocation
simple and popular Bayesian model for unsupervised learning, but *too unconstrained*

supervised/nb

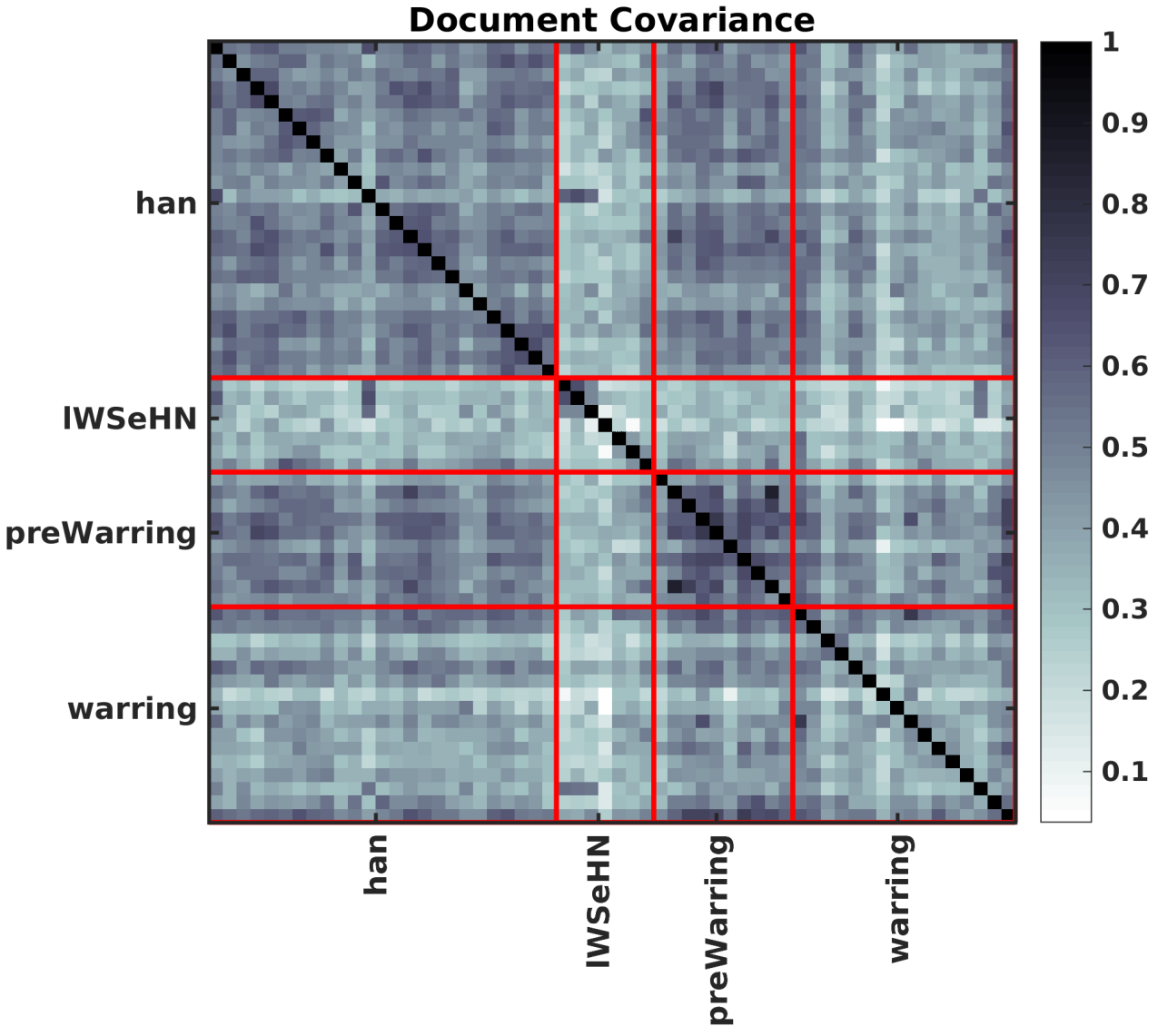


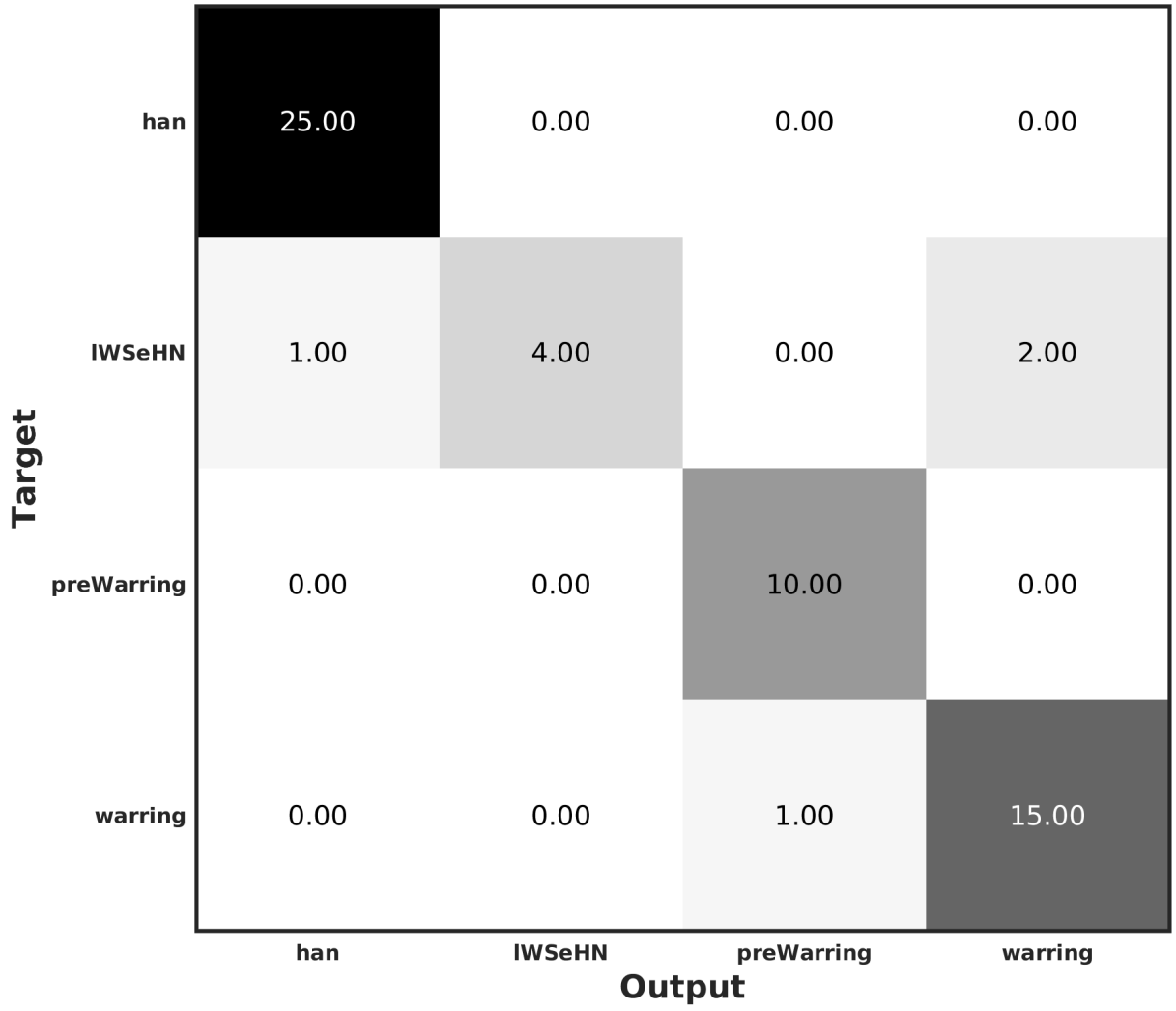


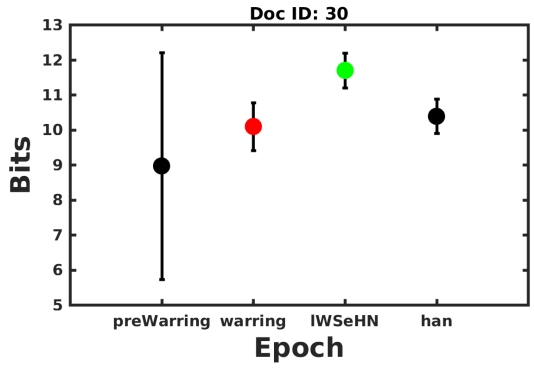
$\mathbf{d}_i$ is the $2K$-dimensional feature vectors, $d_{i1} = P(c_1)$, $d_{i2} = P(c_2)$, ...
$d_{i1958} = P(c_{1958})$

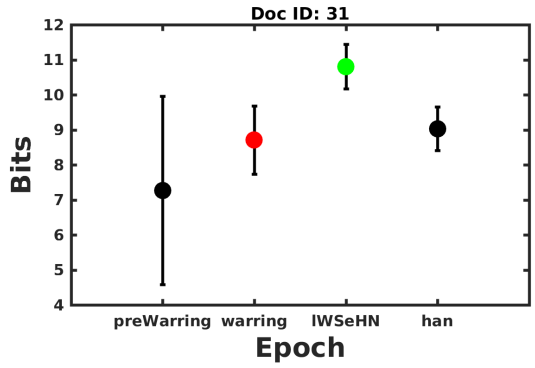
$w_j \in \{Warring_{pre}, Warring, Warring_{late}/Han_{early}, Han\}$

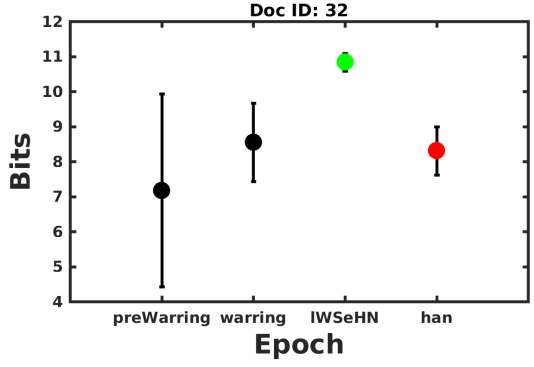
explore category boundaries $\sim$ **misclassification semantics**

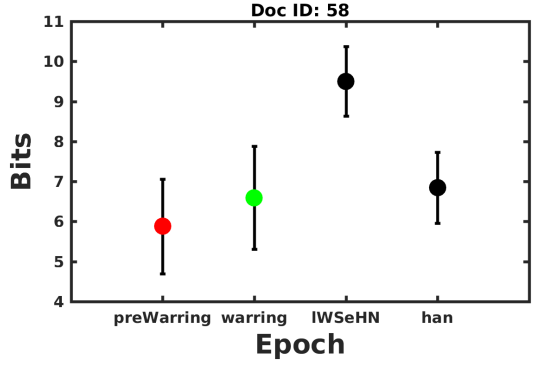

Document Covariance
han
lWSeHN
preWarring
warring
han
lWSeHN
preWarring
warring
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1

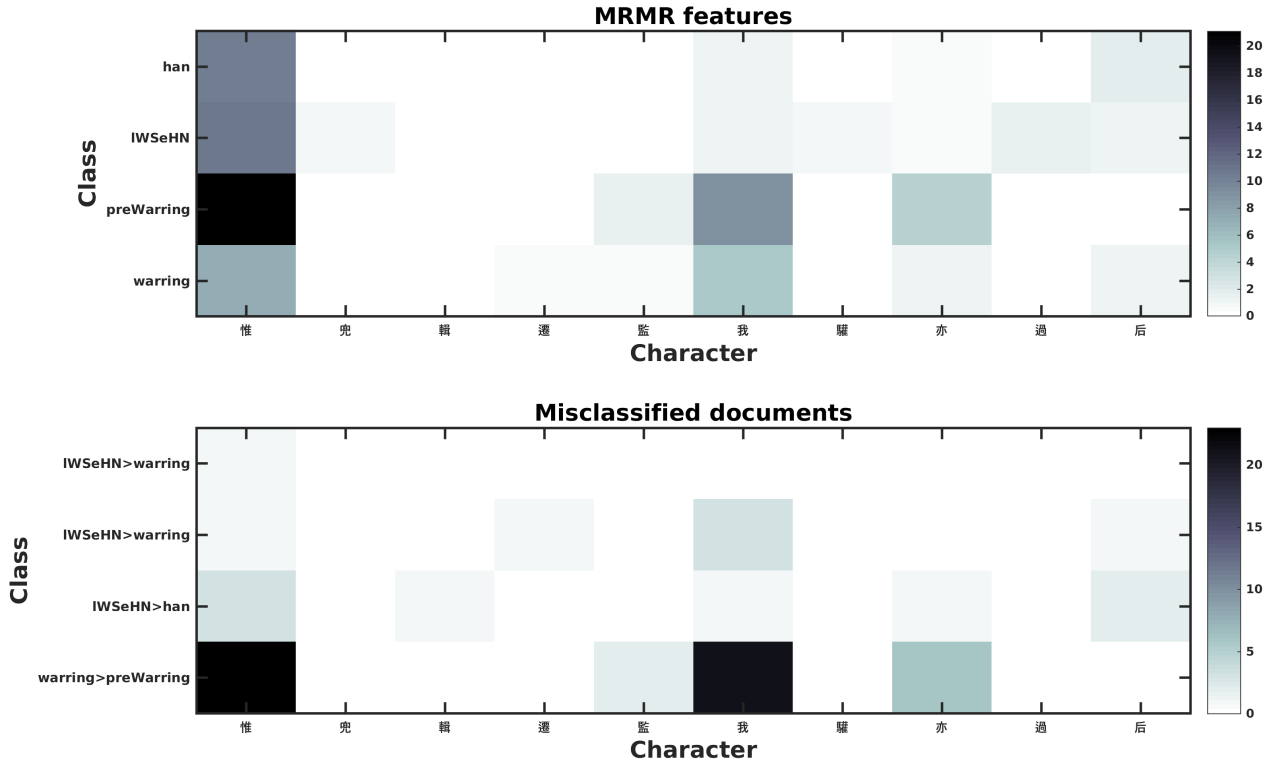

Target
han
lWSeHN
preWarring
warring
25.00
0.00
0.00
0.00
1.00
4.00
0.00
2.00
0.00
0.00
10.00
0.00
0.00
0.00
1.00
15.00
han
lWSeHN
preWarring
warring
Output

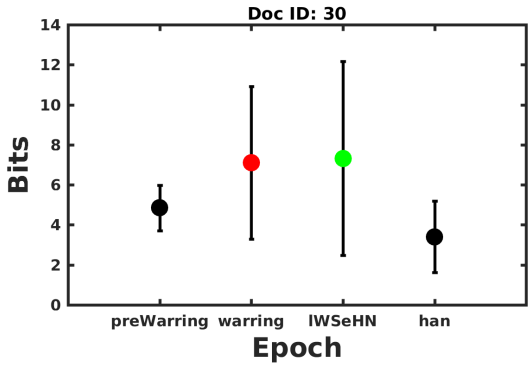

Doc ID: 30
Bits
13
12
11
10
9
8
7
6
5
preWarring
warring
lWSeHN
han
Epoch

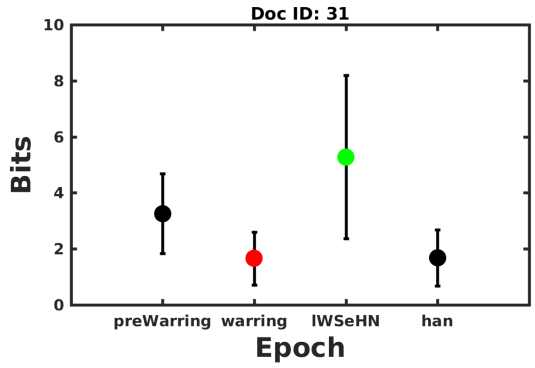

Doc ID: 31
Bits
12
11
10
9
8
7
6
5
4
preWarring
warring
lWSeHN
han
Epoch

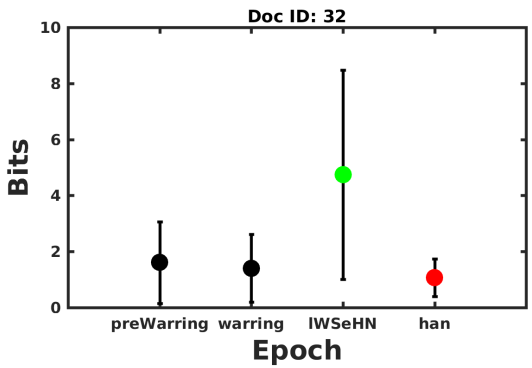

Doc ID: 32
Bits
12
11
10
9
8
7
6
5
4
preWarring
warring
lWSeHN
han
Epoch

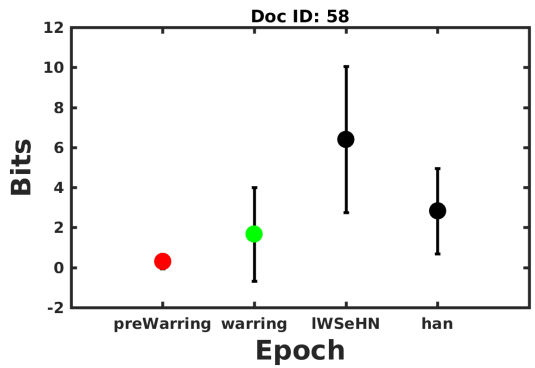

Doc ID: 58
Bits
11
10
9
8
7
6
5
4
preWarring
warring
lWSeHN
han
Epoch


MRMR features
Class
han
lWSeHN
preWarring
warring
惟
兜
輯
遷
監
我
驩
亦
過
后
Character
20
18
16
14
12
10
8
6
4
2
0
Misclassified documents
Class
lWSeHN>warring
lWSeHN>warring
lWSeHN>han
warring>preWarring
惟
兜
輯
遷
監
我
驩
亦
過
后
Character
20
15
10
5
0

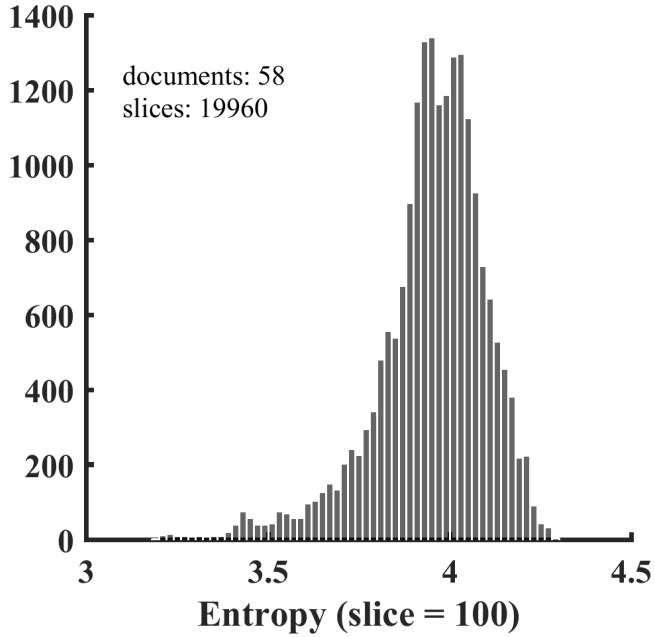

Doc ID: 30
Bits
14
12
10
8
6
4
2
0
preWarring
warring
lWSeHN
han
Epoch

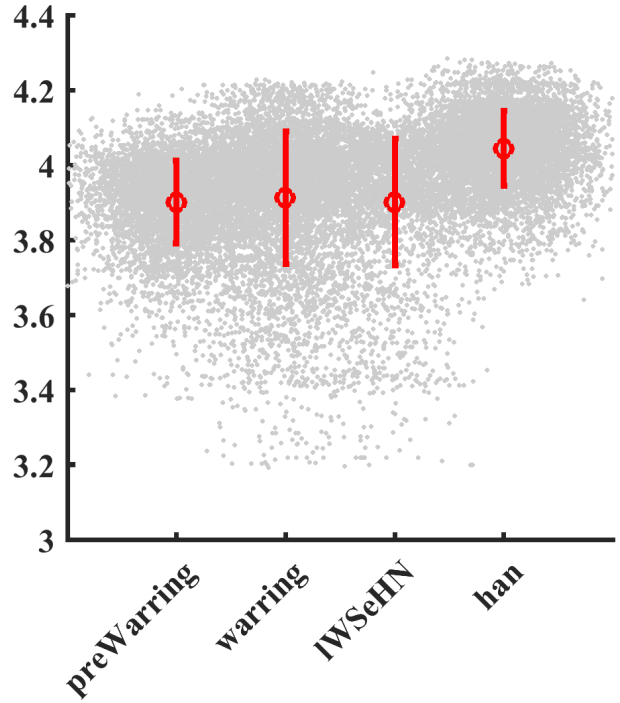

Doc ID: 31
Bits
10
8
6
4
2
0
preWarring
warring
lWSeHN
han
Epoch


Doc ID: 32
Bits
10
8
6
4
2
0
preWarring
warring
lWSeHN
han
Epoch


Doc ID: 58
Bits
12
10
8
6
4
2
0
-2
preWarring
warring
lWSeHN
han
Epoch

unsupervised/lda



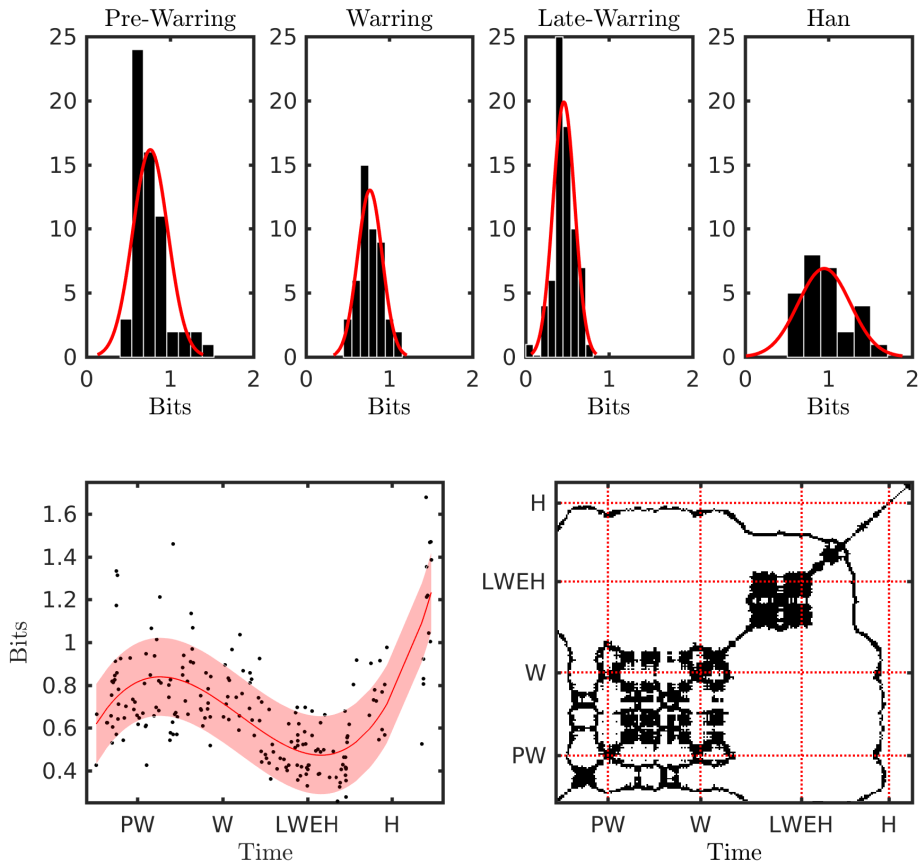


"a rose is a rose is a rose" is less lexically dense than "a rose is red and thorny"

lexical density $\sim$ text predictability: $H(X) = \sum_{i=1}^{n} p_i \log_2 p_i$

$$H(\text{a rose is a rose is a rose}) < H(\text{a rose is red and thorny})$$

$$H(\text{a rose is a rose is a rose}) = H(\text{erea oisasessar oiors})$$



lda to the rescue: θ_i probability distribution of k latent variables in document i

disruption between document is the relative entropy: $D_{KL}(P \parallel Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$

in summary

goal
creative use of statistical learning to support humanistic inquiry

supervised
study semantics on the boundaries of theoretical classes relying on less constrained human interpretation
unsupervised
use theoretical classes to study semantic evolution of cultural system without unconstrained human interpretation