# Classifying cyphertexts

Rok Ivanšek

Univerza v Ljubljani

30. januar 2017

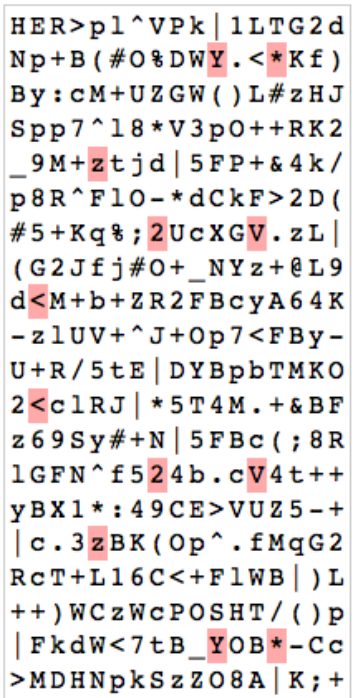
# Pregled

- Motivation
- Plan
  - Choosing texts
  - Procesing of texts
  - Encrypting texts
  - Feature engineering
  - ML techniques
  - Extra

# Motivation

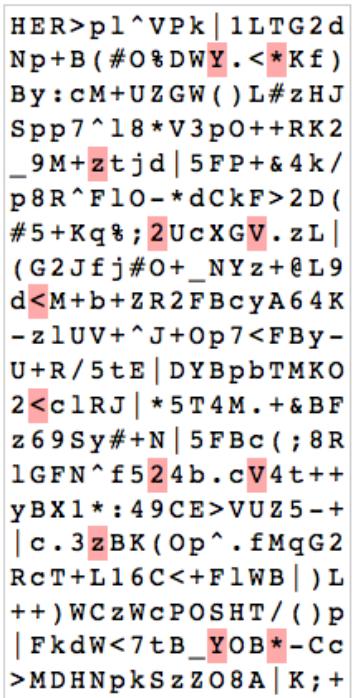
```
HER>pl^VPk|1LTG2d
Np+B(#O%DWY.<*Kf)
By:cM+UZGW()L#zHJ
Spp7^18*V3pO++RK2
_9M+ztjd|5FP+&4k/
p8R^Flo-*dCkF>2D(
#5+Kq%;2UcXGV.zL|
(G2Jfj#O+_NYz+@L9
d<M+b+ZR2FBcyA64K
-zlUV+^J+Op7<FBy-
U+R/5tE|DYBpbTMKO
2<clRJ|*5T4M.+&BF
z69Sy#+N|5FBc(;8R
lGFN^f524b.cV4t++
yBX1*:49CE>VUZ5-+
|c.3zBK(Op^.fMqG2
RcT+L16C<+FlWB|)L
++)WCzWcPOSHT/()p
|FkdW<7tB_YOB*-Cc
>MDHNpkSzZO8A|K;+
```

# Motivation




Cyphertexts are unreadable.

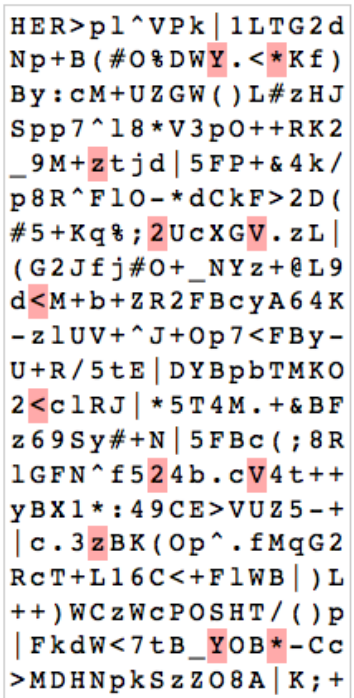
# Motivation




Cyphertexts are unreadable.

Patterns can be found in cyphertexts.

# Motivation




Cyphertexts are unreadable.

Patterns can be found in cyphertexts.

Are the patterns clear enought to allow for a good classification model?

# Choosing texts, preprocessing, encrypting

# Choosing texts, preprocessing, encrypting

- 20newsgroups corpus

# Choosing texts, preprocessing, encrypting

- 20newsgroups corpus
- Length of texts

# Choosing texts, preprocessing, encrypting

- 20newsgroups corpus
- Length of texts
- Preprocessing steps

# Choosing texts, preprocessing, encrypting

- 20newsgroups corpus
- Length of texts
- Preprocessing steps
- Choosing encryption algorithms

# Choosing texts, preprocessing, encrypting

- 20newsgroups corpus
- Length of texts
- Preprocessing steps
- Choosing encryption algorithms
- Raw data for modeling

# Feature engineering

# Feature engineering

- This is the most difficult task of this project

# Feature engineering

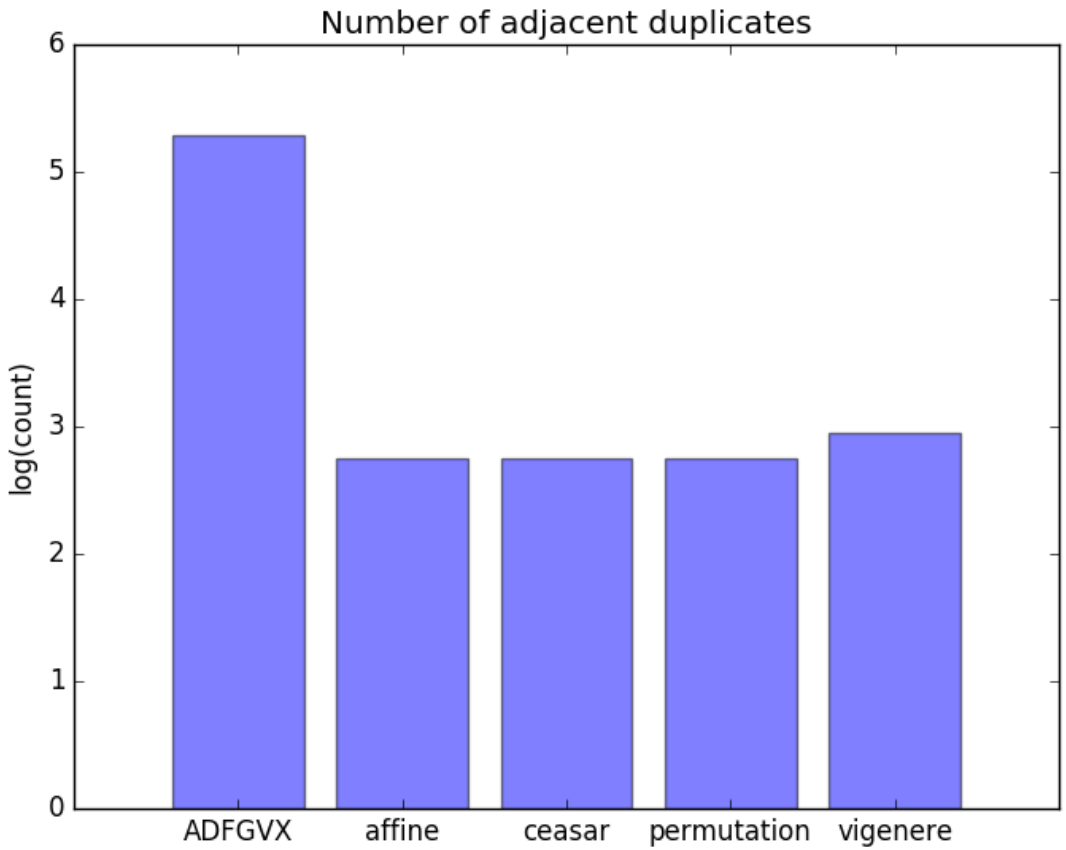
- This is the most difficult task of this project
- Features can make or break a model

# Feature engineering

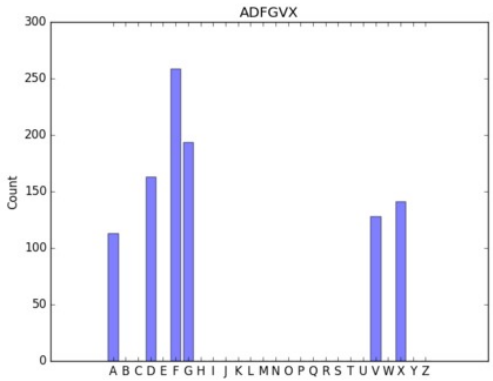
- This is the most difficult task of this project
- Features can make or break a model
- Examples of features

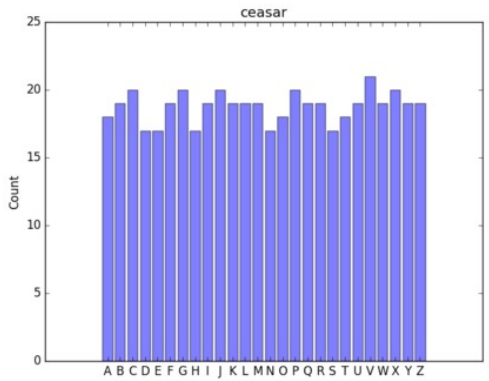
# Feature engineering

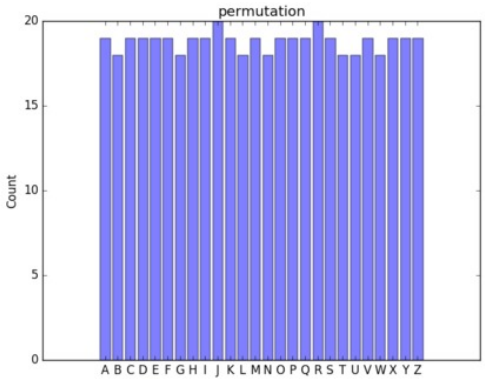
- This is the most difficult task of this project
- Features can make or break a model
- Examples of features

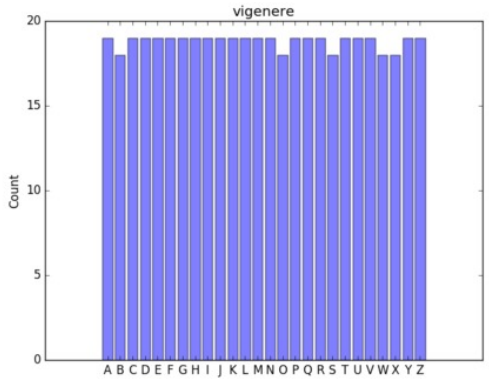


# Feature engineering



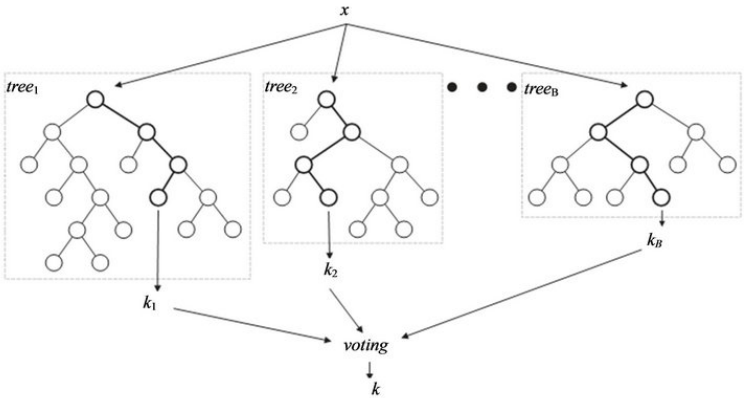



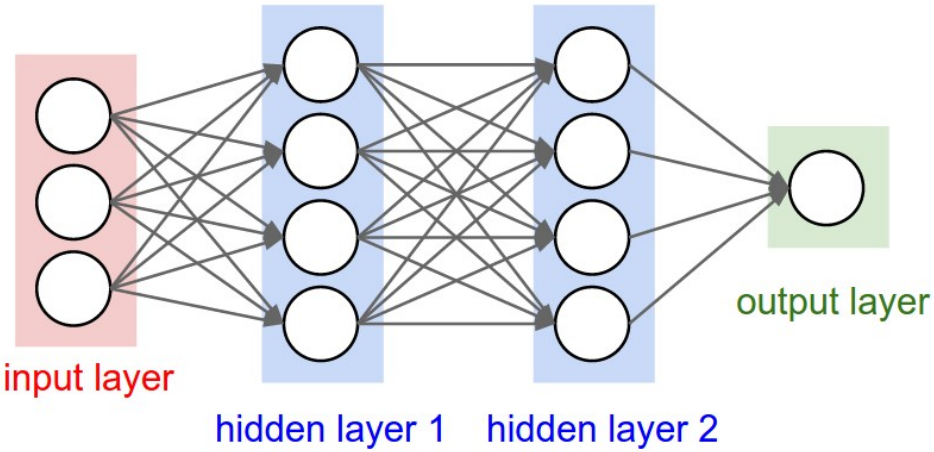



# ML techniques, conclusion

# ML techniques, conclusion

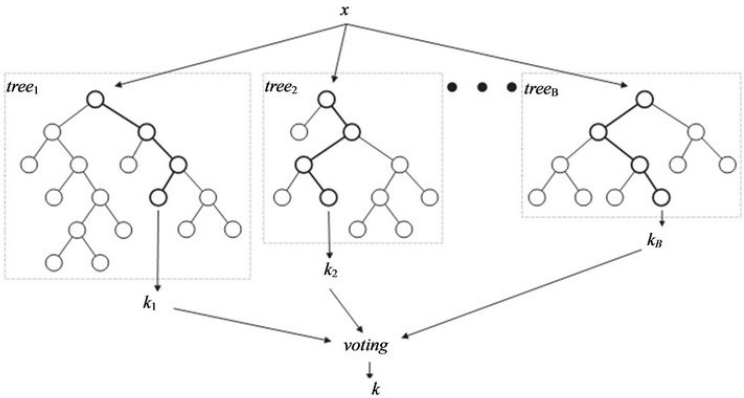
- Random forest, Neural network

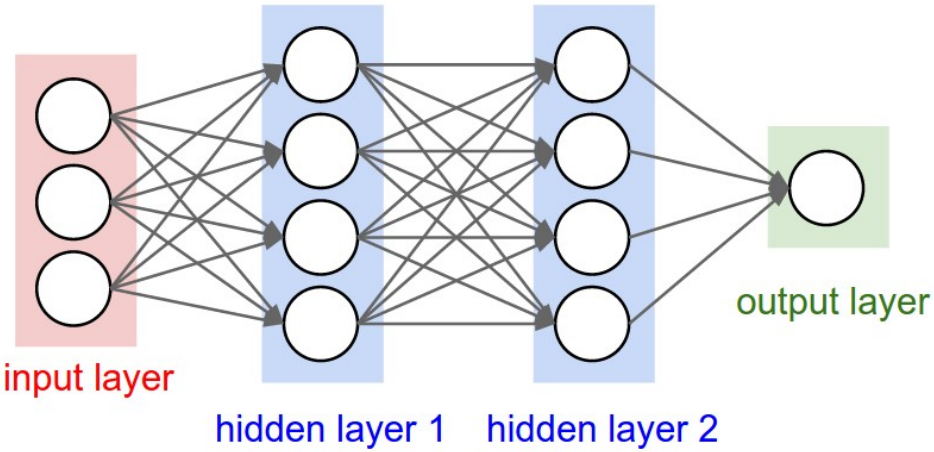


# ML techniques, conclusion

- Random forest, Neural network

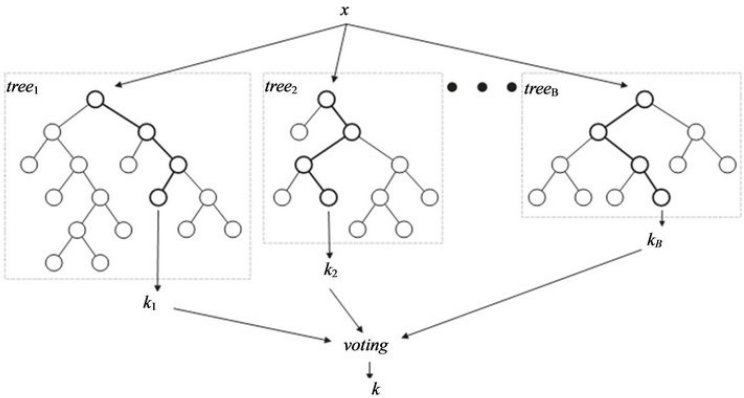


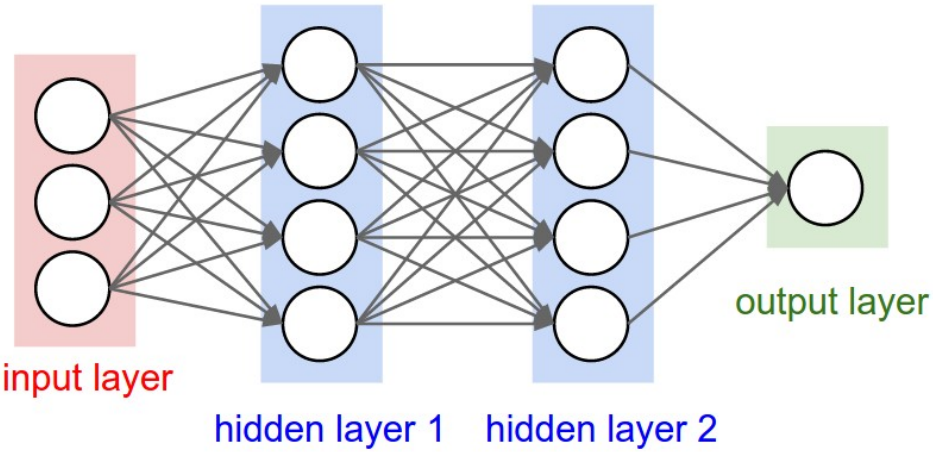




- Test the model

# ML techniques, conclusion

- Random forest, Neural network

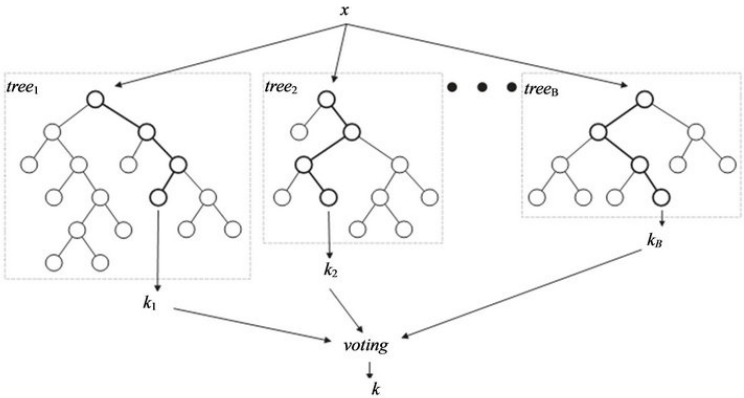


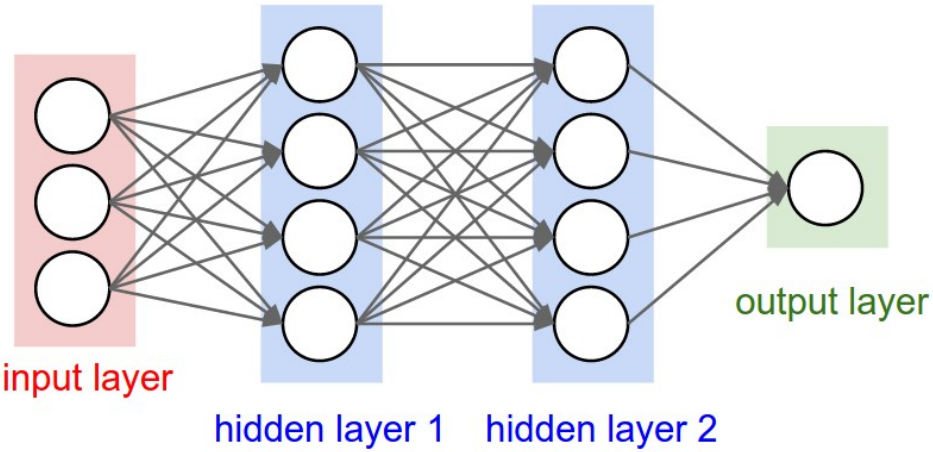




- Test the model
- Meassure accuracy

# ML techniques, conclusion

- Random forest, Neural network




- Test the model
- Meassure accuracy
- Try on more complicated, modern cyphers

# Thank you for your time