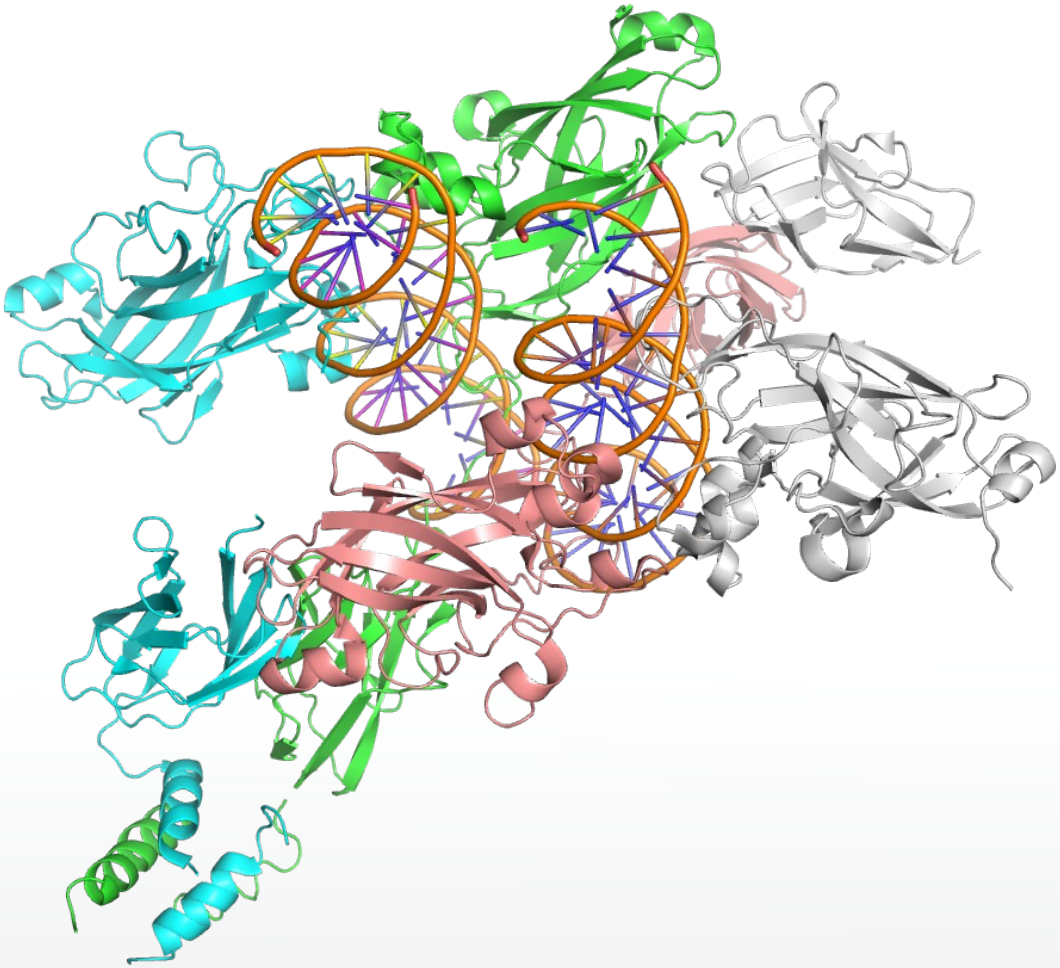
~~5~~ 6 things I have learned using neural networks (deep or not)

Claudio Mirabello

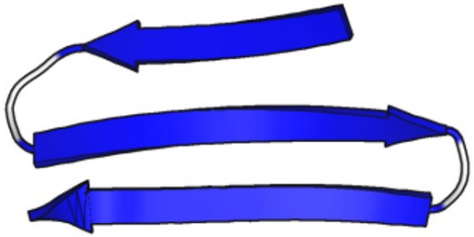
LINKÖPINGS UNIVERSITET

NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

What people are trying to predict about proteins



Prediction classic #1: Secondary Structure



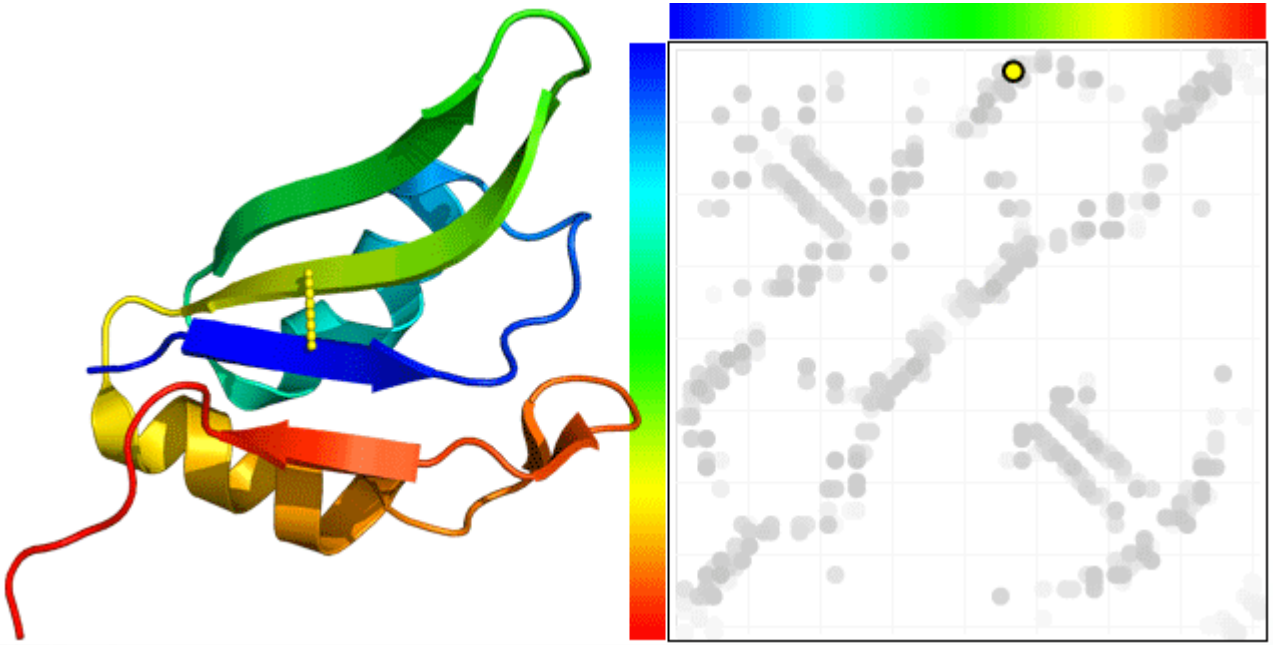
β-Sheet (3 strands)

α-helix

(or none of the above)

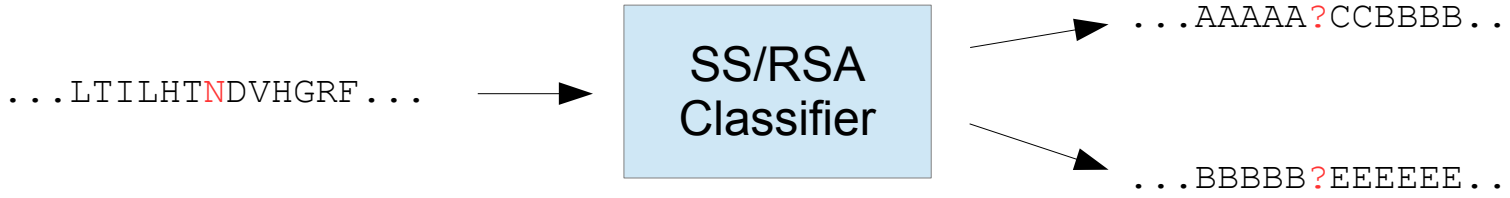
Prediction classic #2: Solvent Accessibility

Prediction classic #3: Contact Maps



http://gremlin.bakerlab.org/gremlin_faq.php

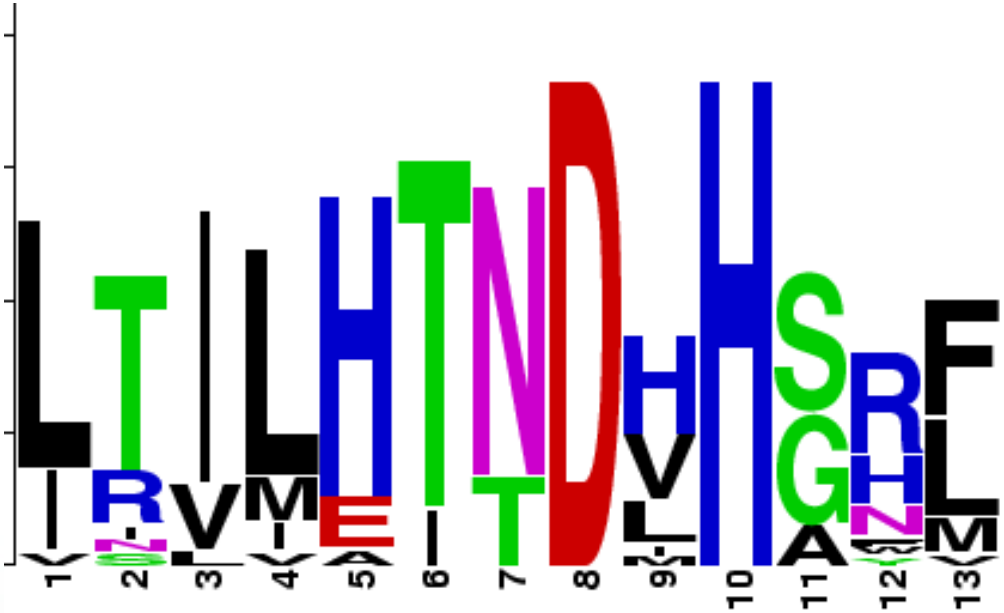
Old approach



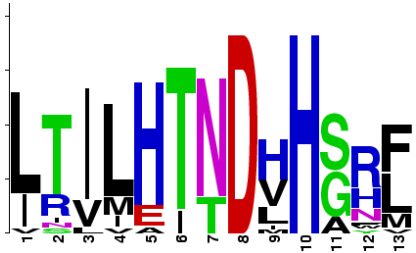
Current approach

```
CAV3 Hs NP 203123        --M-MAEEHTD-LEAQIVKDIHCKEIDLVNRDPKNINEDIVKVDFEDVIAEPVGTY-SFDGVWKVSYTTFTVSKYWCYRLI
CAV3 Mm NP 031643        --M-MTEEHTD-LEARIIKDIHCKEIDLVNRDPKNINEDIVKVDFEDVIAEPEGTY-SFDGVWKVSFTTFTVSKYWCYRLI
CAV3 Rn NP 062028        --M-MTEEHTD-LEARIIKDIHCKEIDLVNRDPKNINEDIVKVDFEDVIAEPEGTY-SFDGVWRVSYTTFTVSKYWCYRLI
CAV3 Gg NP 989701        --M-AEEQREL-EERIIIKDQHTKEIDLVNRDPKRINEDVVKVDFEDVIAEPVGTY-SFDGVWKGSYTTFTVSKYWCYRLI
CAV3 Dr NP 991301        --M-ADQYNT--NEEKILRDSHTKEIDLINRDPKQINEDVVKVDFEDVIAEPDGT-HSMDGVWKASYTTFTVSKYWCYRVI
CAV3 Tn CAG06808         --M-ADQYQYNANEEKIVKDSHTKEIDLINRDPKQINEDVVKVEFEDVIAEPDGT-HSLDGVWKLSYTTFTVSKYWCYRVI
CAV3 Xl AAH41289         --MAQIQQPEPAKQDKSNTEALTKEIDLVQRDPKKINQEVVQVDFEDVIAEPDGT-HSFDGVWKASSSTFTVTKYWCYRVI
CAV1b Dr AAN06979        --MDNDSIN-----EKTLQDVHTKEIDLVNRDPKHLNDDVVKVDFEDVIAEPAGTY-SFDGVWKASFTTFTVTKYWCYRLI
CAV1 Hs NP 001744        NNKAMADE----LSEKQVYDAHTKEIDLVNRDPKHLNDDVVKIDFEDVIAEPEGT-HSFHGIWKASFTTFTVTKYWFYRLI
CAV1 Bt NP 776429        NNKAMAEE----MNEKQVYDAHTKEIDLVNRDPKHLNDDVVKIDFEDVIAEPEGT-HSFDGIWKASFTTFTVTKYWFYRLI
CAV1 Ss NP 999603        NNKAMAEE----MNEKQVYDAHTKEIDLVNRDPKHLNDDVVKIDFEDVIAEPEGT-HSFDGIWKASFTTFTVTKYWFYRLI
CAV1 Oa AAT81146         NNKAMAEE----MNEKQVYDAHTKEIDLVNRDPKHLNDDVVKIDFEDVIAEPEGT-HSFDGIWKASFTTFTVTKYWFYRLI
CAV1 Cf NP 001003296     NNKAMAEE----MSEKQVYDAHTKEIDLVNRDPKHLNDDVVKIDFEDVIAEPEGT-HSFDGIWKASFTTFTVTKYWFYRLI
CAV1 Fc AAR16230         NNKAMAEE----INEKQVYDAHTKEIDLVNRDPKHLNDDVVKIDFEDVIAEPEGT-HSFDGIWKASFTTFTVTKYWFYRLI
CAV1 Mm NP 031642        NNKAMADE----VTEKQVYDAHTKEIDLVNRDPKHLNDDVVKIDFEDVIAEPEGT-HSFDGIWKASFTTFTVTKYWFYRLI
CAV1a Rn NP 113744       NNKAMADE----VNEKQVYDAHTKEIDLVNRDPKHLNDDVVKIDFEDVIAEPEGT-HSFDGIWKASFTTFTVTKYWFYRLI
CAV1 Dr NP 997816        NNKEMDNDS---INEKTLQDVHTKEIDLVNRDPKHLNDDVVKVDFEDVIAEPAGTY-SFDGVWKASFTTFTVTKYWCYRLI
CAV1 Gg AAR16241         NNKMMADE----LSEKAVHDVHTKEIDLVNRDPKHLNDDVVKIDFEDVIAEPEGT-HSFDGIWKASFTTFTVTKYWFYRLI
CAV1 Dr AAR16334         NNKEMDND-S--INEKTLQDVHTKEIDLVNRDPKHLNDDVVKVDFEDVIAEPAGTY-SFDGVWKASFTTFTVTKYWCYRLI
CAV1 Oc AAM19213         ---AMADE----VNEKQVYDAHTKEIDLVNRDPKHLNDDVVKIDFEDVIAEPEGT-HSFDGIWKASFTTFTVTKYWFYRLI
CAV1 Tr AAR16280         NNKDMDND-S--LNEKSMEDVHTKEIDLVNRDPKHLNDDVVKVDFEDVIAEPAGTY-SFDGVWKASFTTFTVTKYWCYRLI
CAV1 Tn AAR16324         NNKDMDND-S--LNEKSLEDVHTKEIDLVNRDPKHLNDDVVKVDFEDVIAEPAGTY-SFDGVWKASFTTFTVTKYWCYRLI
CAV Xl AAH70672          NNKTMADD-F--LTETEVRDSHTKEIDLVNRDPKHLNDDVVKIDFEDVIAEPDGT-HSFDGIWKTSFTTFTVTKYWFYRLI
CAV2 Hs NP 001224        VQLFMDDDSYSHHSGLEYADPEKFADSDQDRDPHRLNSHL-KLGFEDVIAEPVTT-HSFDKVWICSHALFEISKYVMYKFI
CAV2 Bt NP 001007809     VQLFMDDDSYSRHSSVDYADPDKFVDPGSDRDPHRLNSHL-KVGFEDVIAEPVST-HSFDKVWICSHALFEMSKYVIYKFI
CAV2 Mm NP 058596        VQLFMADDAYSHHSGVDYADPEKYVDSSHDRDPHQLNSHL-KLGFEDLIAEPETT-HSFDKVWICSHALFEISKYVMYKFI
CAV2 Ss AAR16299         VQLFMDDDSYSRHSGVDYADPKKFVDPGTDRDPHRLNSNL-KVGFEDVIAEPVST-HSFDKVWICSHALFEISKYVIYKFI
CAV2 Fc AAR16229         VQLFMDDDSYSRHSGVDYADPDKFADSGSDRDPHRLNSNL-KVGFEDVIAEPVST-HSFDKVWICSHALFEISKYVIYKFI
CAV2 Gg NP 001007087     TRIFMDDDNFPP-GGPALSEGEKCAEDGLERDPRGLNAHL-QLGFEDVIAEPELT-HSFDKVWICSHALFELSKYVLYKLI
CAV2 Dr NP 001002150     TRVIMDEDEFNRSIEPILGKKPNVYSEVQDRDPKDINKHL-KVGFEDIIAEPNST-HSFDRVWIGSHAVFELVKYVFYRII
CAV2 Rn NP 571989        VQLFMADDAYSHHSVVDYTDPEKYVDSSQDRDPHQLNSHL-KLGFEDLIAEPPTT-HSFDKVWICSHALFEISKYVIYKFI
CAV2 Cf NP 001003066     VQLCMDDDAYSHHSAVDFGDLEQLADSGSDRDPRRLNSHL-QVGFEDVIADAVST-HSFDKVWICSHALFEVSKYVIYKFI
CAV2 Tn AAR16323         TSIIMDEDEFNRSIEPILSQKAKASSSAPDRDPHDLNAQL-KVGFEDVIAEPASA-HSFDRVWIGSHAVFELVKFIFYRLI
CAV2 Tr CAA09081         CSIIMDEDEFNRSIEPILSKKARLYSSAPDRDPHDINAQL-KVGFEDVIAEPASA-HSFDRVWIGSSATFELVKFIFYRLI
CAV Ce NP 501743         WSRCQKGEGEQKEENIAIG------VDLVNRDANSMNNHV-QLNFEDIFGEADSQ-HSWDCVWRLNHTVFTAVRLFIYRLV
CAV Cb CAE69013          KCKTKKCENEQKEDHIAIG------MDFVNRDHNGLNNHV-QHNFDDIFGEADSQ-HTWEFMWRLNASVFNWVRLFVYRFI
CAV Ce NP 506357         HKPPNPEMEFD-IGVKNIAPVLIHKMNMDDRDPKD-SAQYLNTSFFEVFNEPSEQYHSIACVWTLSFKIFEIVRIYSYKII
```

MSAs to sequence profiles



Current approach

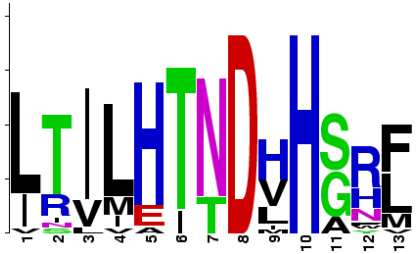


SS/RSA Classifier

...AAAAA?CCBBBB..

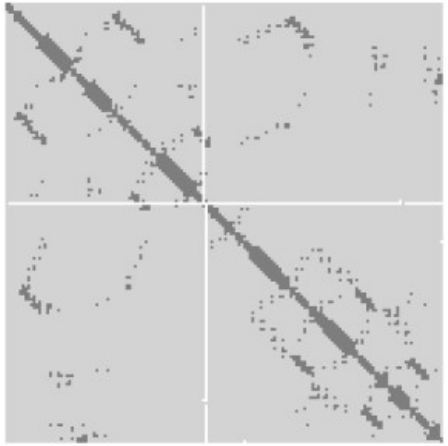
...BBBBB?EEEEEE..

Current approach

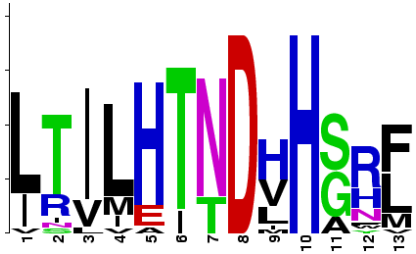


...HHHHHCCCEEEE...

Contact Map Predictor



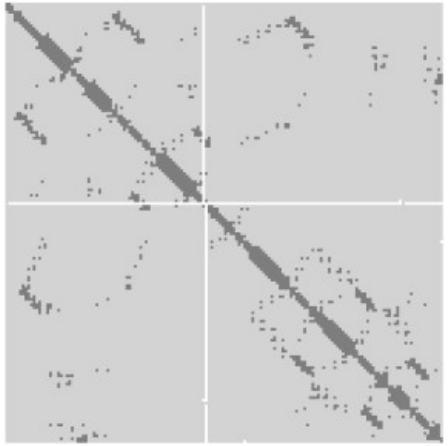
Current approach



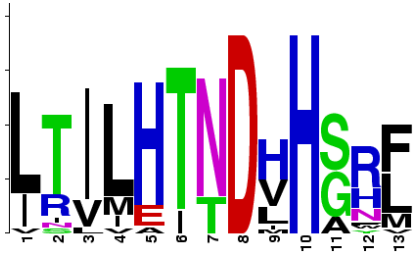
...HHHHHCCCEEEE...

...BBBBBBEEEEBB...

Contact Map Predictor



Current approach

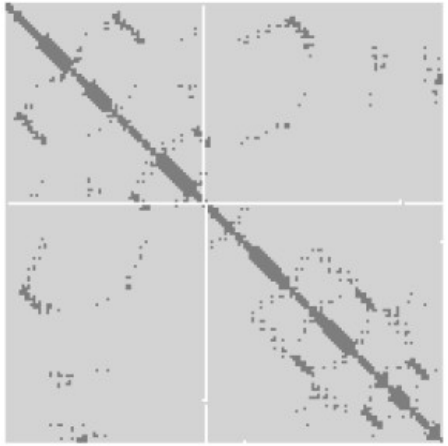


...HHHHHCCCEEEE...

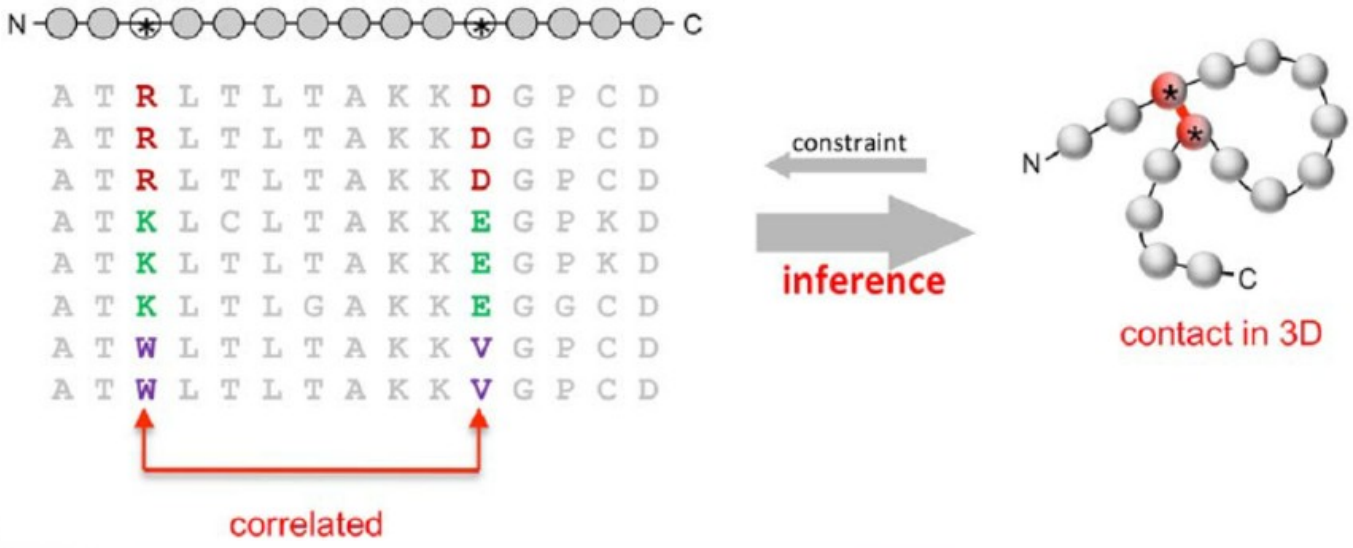
...BBBBBBEEEEBB...

Correlated mutations

Contact Map Predictor

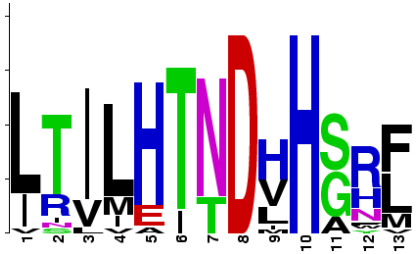


Correlated mutations



Marks, D., et al: Protein 3D Structure Computed from Evolutionary Sequence Variation (2011)

Current approach

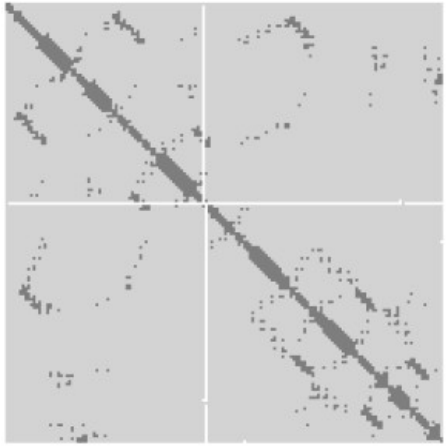


...HHHHHCCCEEEE...

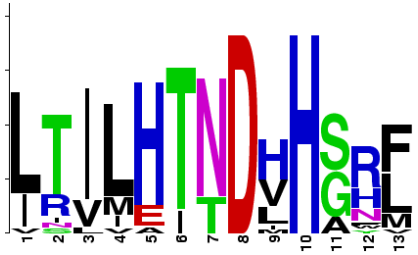
...BBBBBBEEEEBB...

Correlated mutations

Contact Map Predictor



“Deep” approach

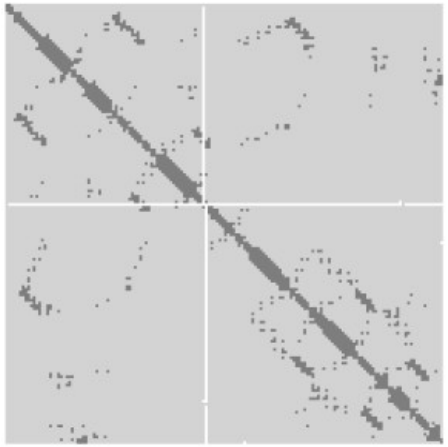


...HHHHHCCCEEEE...

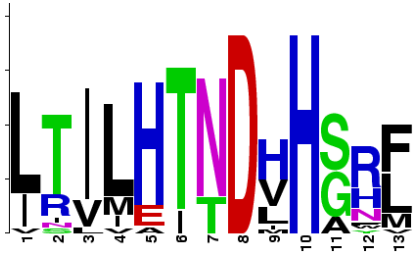
...BBBBBBEEEEBB...

Correlated mutations

Contact Map Predictor



“Deep” approach

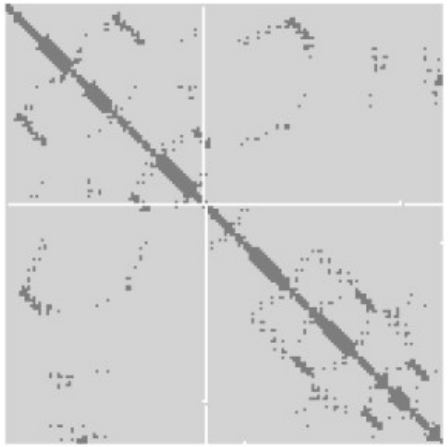


...HHHHHCCCEEEE...

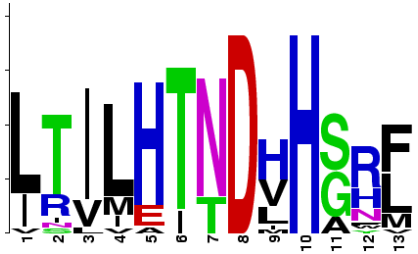
...BBBBBBEEEEBB...

Correlated mutations

Contact Map Predictor



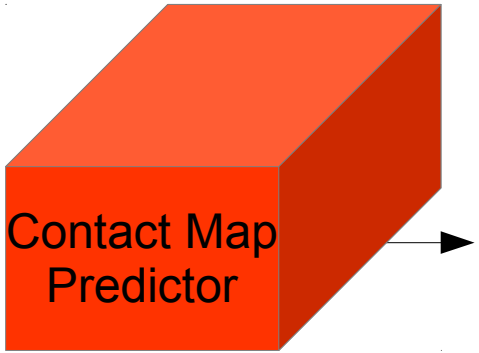
“Deeper” approach

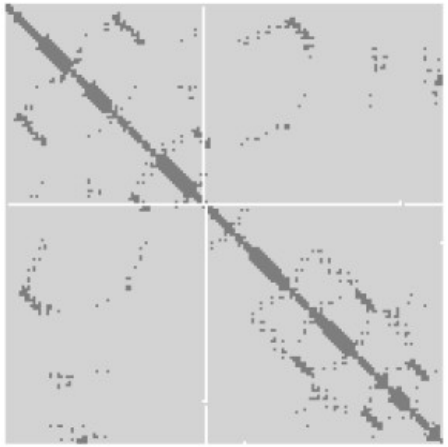


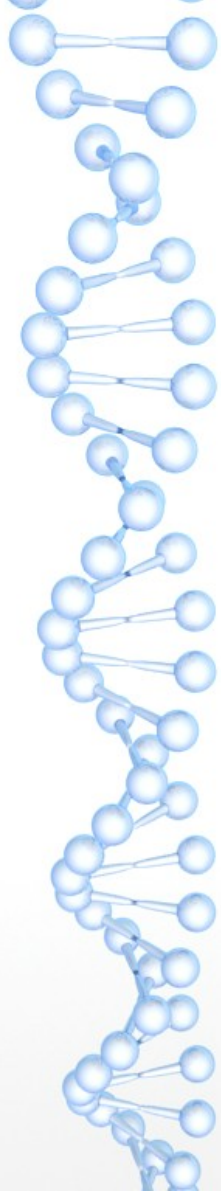
...HHHHHCCCEEEE...

...BBBBBBEEEEBB...

Correlated mutations

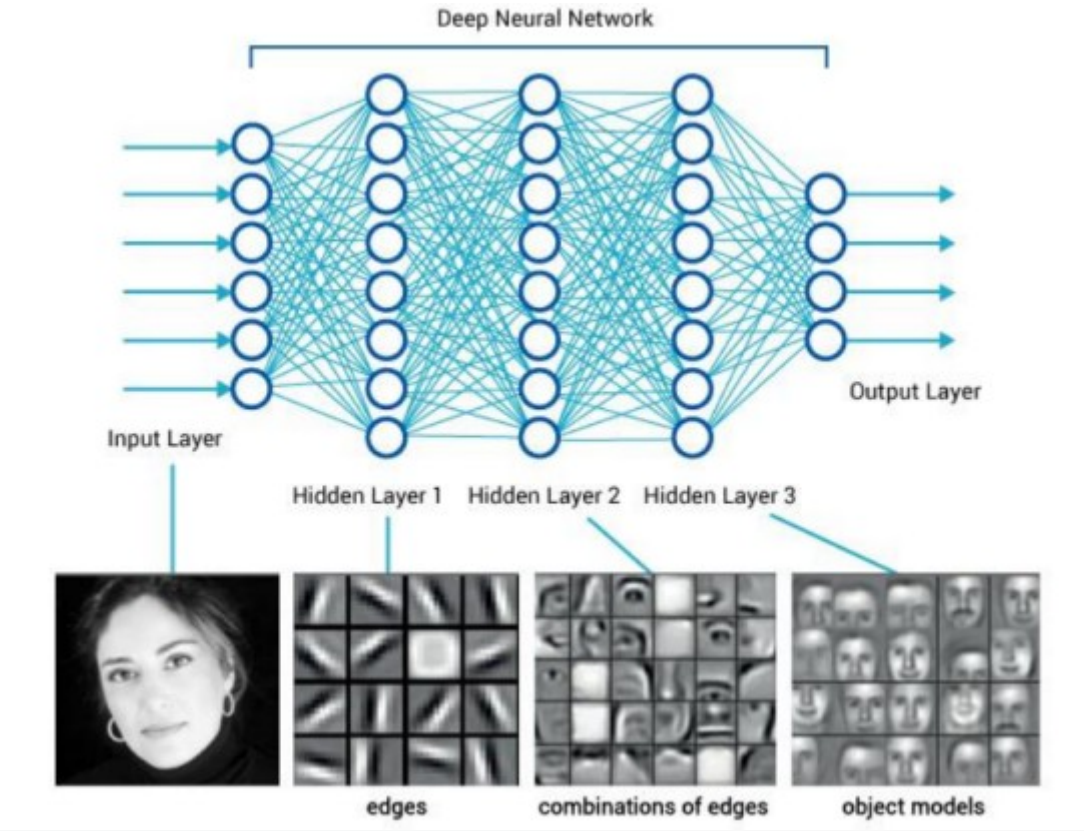




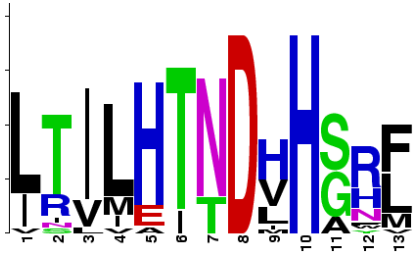


Things I've learned #1:
Deep Learning is not only about deeper architectures, but also about automatic feature extraction

Feature maps and automatic feature extraction



“Deep” approach

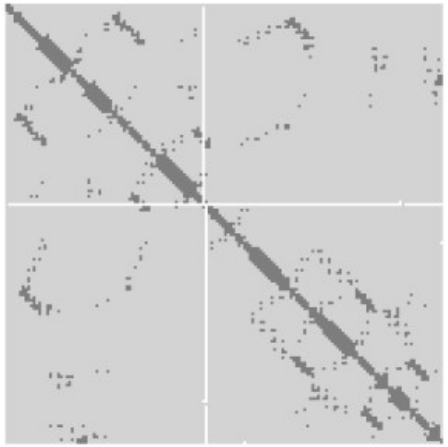


...HHHHHCCCEEEE...

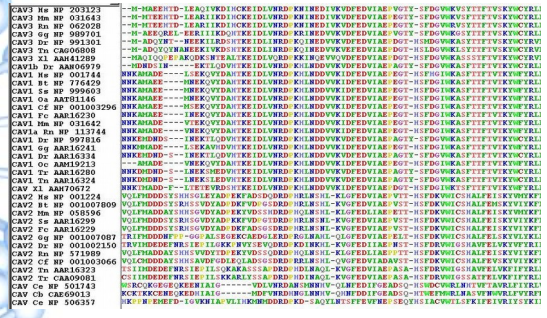
...BBBBBBEEEEBB...

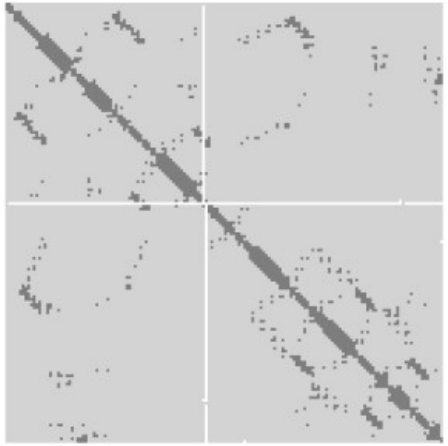
Correlated mutations

Contact Map Predictor



Can we input the MSA as it is?



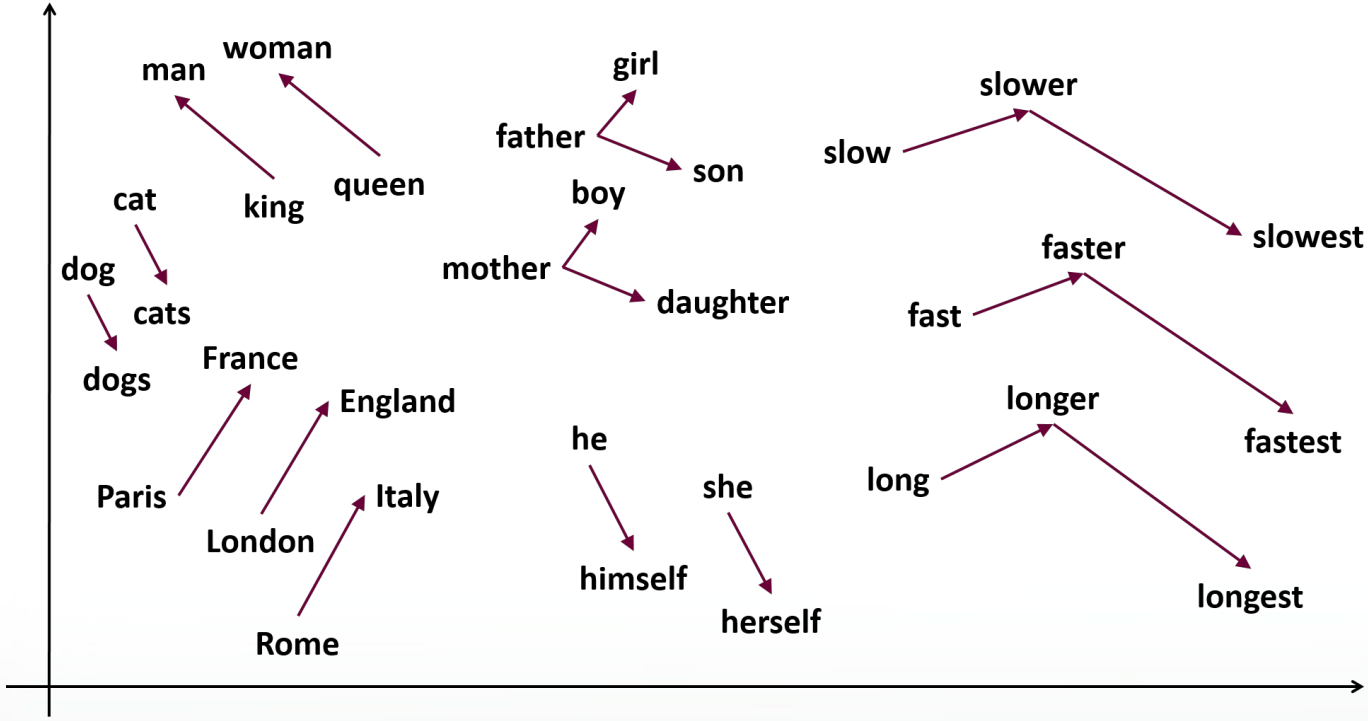


Things I've learned #2:
Don't be afraid of getting creative when designing your model

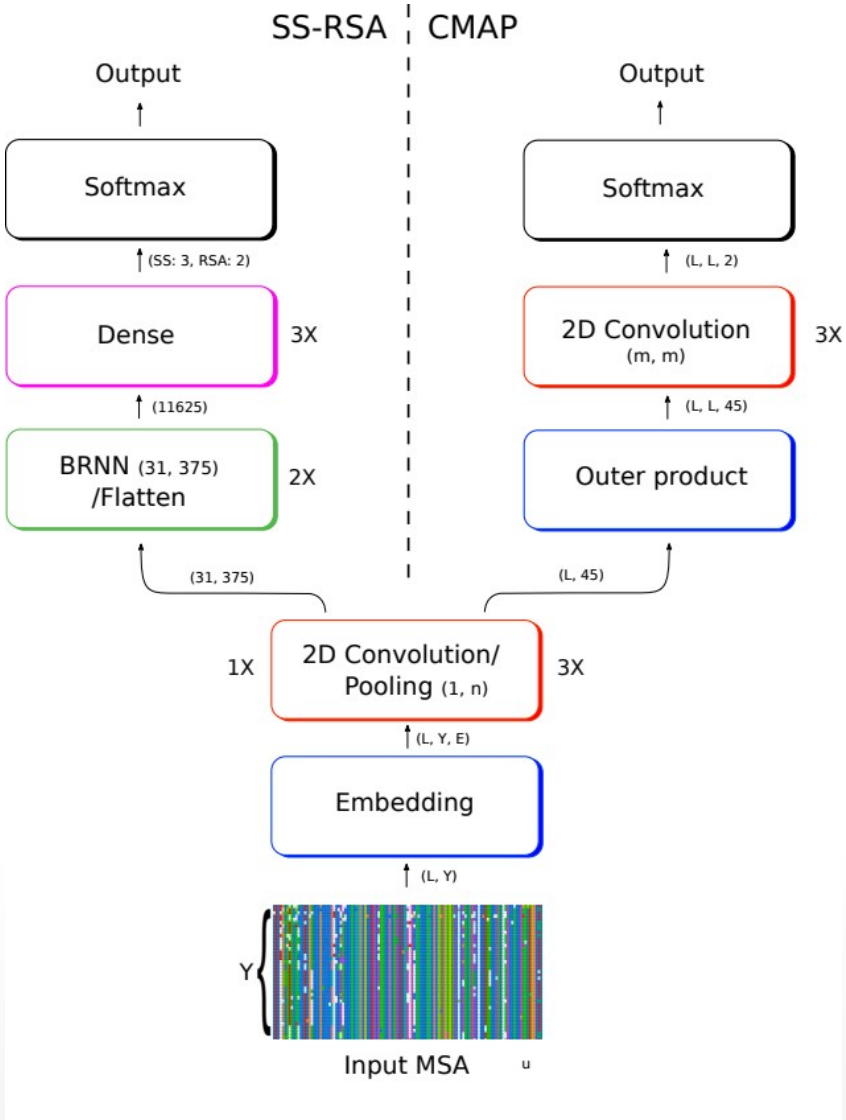
MSAs aren’t an easy input

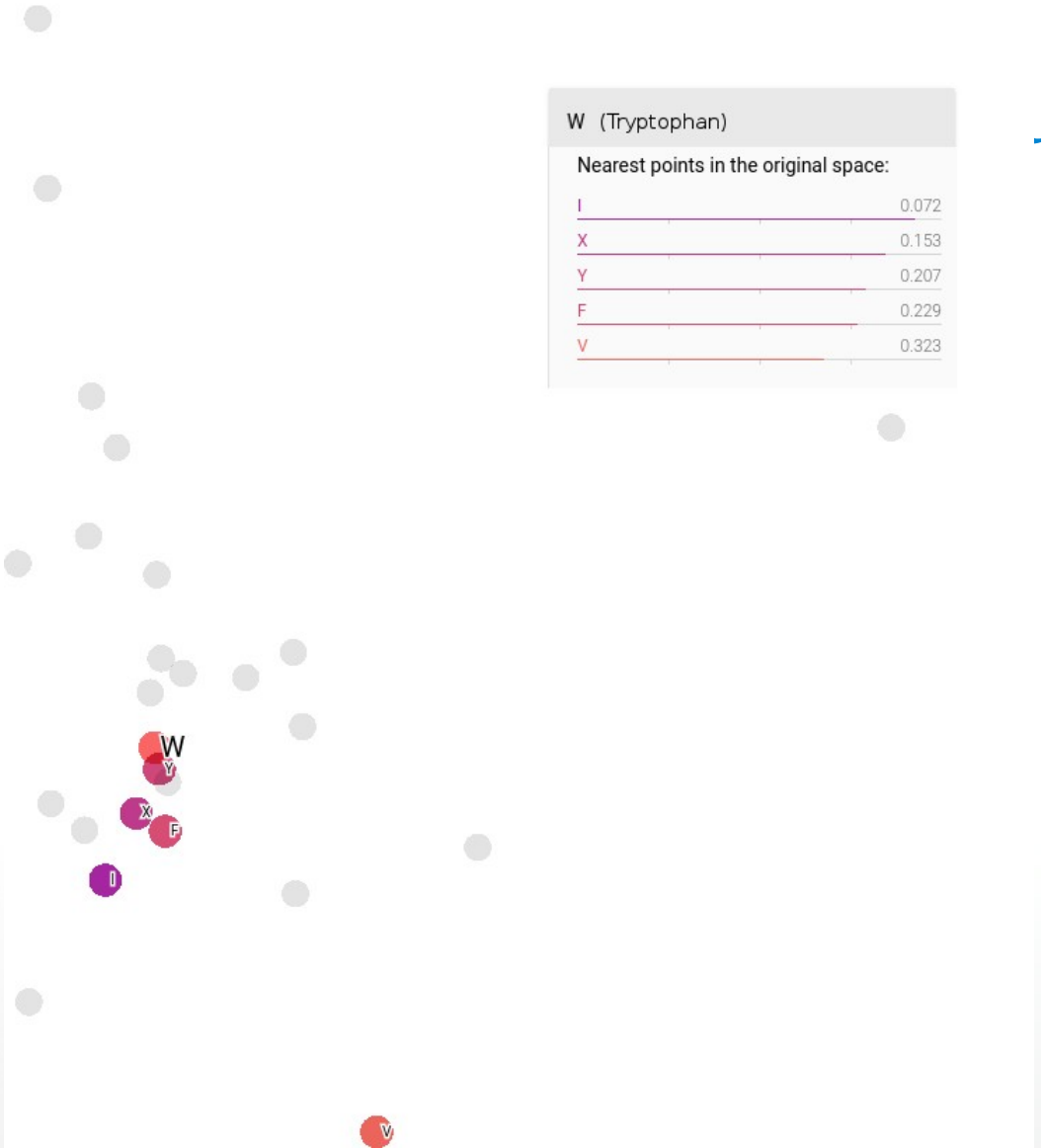
```
CAV3 Hs NP 203123      --M-MAEEHTD-LEAQIVKDIHCKEIDLVNRDPKNINEDIVKVDFEDVIAEPVGTY-SFDGVWKVSYTTFTVSKYWCYRLI
CAV3 Mm NP 031643      --M-MTEEHTD-LEARIIKDIHCKEIDLVNRDPKNINEDIVKVDFEDVIAEPEGTY-SFDGVWKVSFTTFTVSKYWCYRLI
CAV3 Rn NP 062028      --M-MTEEHTD-LEARIIKDIHCKEIDLVNRDPKNINEDIVKVDFEDVIAEPEGTY-SFDGVWRVSYTTFTVSKYWCYRLI
CAV3 Gg NP 989701      --M-AEEQREL-EERIIIKDQHTKEIDLVNRDPKRINEDVVKVDFEDVIAEPVGTY-SFDGVWKGSYTTFTVSKYWCYRLI
CAV3 Dr NP 991301      --M-ADQYNT--NEEKILRDSHTKEIDLINRDPKQINEDVVKVDFEDVIAEPDGT-HSMDGVWKASYTTFTVSKYWCYRVI
CAV3 Tn CAG06808       --M-ADQYQYNANEEKIVKDSHTKEIDLINRDPKQINEDVVKVEFEDVIAEPDGT-HSLDGVWKLSYTTFTVSKYWCYRVI
CAV3 Xl AAH41289       --MAQIQQPEPAKQDKSNTEALTKEIDLVQRDPKKINQEVVQVDFEDVIAEPDGT-HSFDGVWKASSSTFTVTKYWCYRVI
CAV1b Dr AAN06979      --MDNDSIN-----EKTLQDVHTKEIDLVNRDPKHLNDDVVKVDFEDVIAEPAGTY-SFDGVWKASFTTFTVTKYWCYRLI
CAV1 Hs NP 001744      NNKAMADE----LSEKQVYDAHTKEIDLVNRDPKHLNDDVVKIDFEDVIAEPEGT-HSFHGIWKASFTTFTVTKYWFYRLI
CAV1 Bt NP 776429      NNKAMAEE----MNEKQVYDAHTKEIDLVNRDPKHLNDDVVKIDFEDVIAEPEGT-HSFDGIWKASFTTFTVTKYWFYRLI
CAV1 Ss NP 999603      NNKAMAEE----MNEKQVYDAHTKEIDLVNRDPKHLNDDVVKIDFEDVIAEPEGT-HSFDGIWKASFTTFTVTKYWFYRLI
CAV1 Oa AAT81146       NNKAMAEE----MNEKQVYDAHTKEIDLVNRDPKHLNDDVVKIDFEDVIAEPEGT-HSFDGIWKASFTTFTVTKYWFYRLI
CAV1 Cf NP 001003296   NNKAMAEE----MSEKQVYDAHTKEIDLVNRDPKHLNDDVVKIDFEDVIAEPEGT-HSFDGIWKASFTTFTVTKYWFYRLI
CAV1 Fc AAR16230       NNKAMAEE----INEKQVYDAHTKEIDLVNRDPKHLNDDVVKIDFEDVIAEPEGT-HSFDGIWKASFTTFTVTKYWFYRLI
CAV1 Mm NP 031642      NNKAMADE----VTEKQVYDAHTKEIDLVNRDPKHLNDDVVKIDFEDVIAEPEGT-HSFDGIWKASFTTFTVTKYWFYRLI
CAV1a Rn NP 113744     NNKAMADE----VNEKQVYDAHTKEIDLVNRDPKHLNDDVVKIDFEDVIAEPEGT-HSFDGIWKASFTTFTVTKYWFYRLI
CAV1 Dr NP 997816      NNKEMDNDS---INEKTLQDVHTKEIDLVNRDPKHLNDDVVKVDFEDVIAEPAGTY-SFDGVWKASFTTFTVTKYWCYRLI
CAV1 Gg AAR16241       NNKMMADE----LSEKAVHDVHTKEIDLVNRDPKHLNDDVVKIDFEDVIAEPEGT-HSFDGIWKASFTTFTVTKYWFYRLI
CAV1 Dr AAR16334       NNKEMDND-S--INEKTLQDVHTKEIDLVNRDPKHLNDDVVKVDFEDVIAEPAGTY-SFDGVWKASFTTFTVTKYWCYRLI
CAV1 Oc AAM19213       ---AMADE----VNEKQVYDAHTKEIDLVNRDPKHLNDDVVKIDFEDVIAEPEGT-HSFDGIWKASFTTFTVTKYWFYRLI
CAV1 Tr AAR16280       NNKDMDND-S--LNEKSMEDVHTKEIDLVNRDPKHLNDDVVKVDFEDVIAEPAGTY-SFDGVWKASFTTFTVTKYWCYRLI
CAV1 Tn AAR16324       NNKDMDND-S--LNEKSLEDVHTKEIDLVNRDPKHLNDDVVKVDFEDVIAEPAGTY-SFDGVWKASFTTFTVTKYWCYRLI
CAV Xl AAH70672        NNKTMADD-F--LTETEVRDSHTKEIDLVNRDPKHLNDDVVKIDFEDVIAEPDGT-HSFDGIWKTSFTTFTVTKYWFYRLI
CAV2 Hs NP 001224      VQLFMDDDSYSHHSGLEYADPEKFADSDQDRDPHRLNSHL-KLGFEDVIAEPVTT-HSFDKVWICSHALFEISKYVMYKFI
CAV2 Bt NP 001007809   VQLFMDDDSYSRHSSVDYADPDKFVDPGSDRDPHRLNSHL-KVGFEDVIAEPVST-HSFDKVWICSHALFEMSKYVIYKFI
CAV2 Mm NP 058596      VQLFMADDAYSHHSGVDYADPEKYVDSSHDRDPHQLNSHL-KLGFEDLIAEPETT-HSFDKVWICSHALFEISKYVMYKFI
CAV2 Ss AAR16299       VQLFMDDDSYSRHSGVDYADPKKFVDPGTDRDPHRLNSNL-KVGFEDVIAEPVST-HSFDKVWICSHALFEISKYVIYKFI
CAV2 Fc AAR16229       VQLFMDDDSYSRHSGVDYADPDKFADSGSDRDPHRLNSNL-KVGFEDVIAEPVST-HSFDKVWICSHALFEISKYVIYKFI
CAV2 Gg NP 001007087   TRIFMDDDNFPP-GGPALSEGEKCAEDGLERDPRGLNAHL-QLGFEDVIAEPELT-HSFDKVWICSHALFELSKYVLYKLI
CAV2 Dr NP 001002150   TRVIMDEDEFNRSIEPILGKKPNVYSEVQDRDPKDINKHL-KVGFEDIIAEPNST-HSFDRVWIGSHAVFELVKYVFYRII
CAV2 Rn NP 571989      VQLFMADDAYSHHSVVDYTDPEKYVDSSQDRDPHQLNSHL-KLGFEDLIAEPPTT-HSFDKVWICSHALFEISKYVIYKFI
CAV2 Cf NP 001003066   VQLCMDDDAYSHHSAVDFGDLEQLADSGSDRDPRRLNSHL-QVGFEDVIADAVST-HSFDKVWICSHALFEVSKYVIYKFI
CAV2 Tn AAR16323       TSIIMDEDEFNRSIEPILSQKAKASSSAPDRDPHDLNAQL-KVGFEDVIAEPASA-HSFDRVWIGSHAVFELVKFIFYRLI
CAV2 Tr CAA09081       CSIIMDEDEFNRSIEPILSKKARLYSSAPDRDPHDINAQL-KVGFEDVIAEPASA-HSFDRVWIGSSATFELVKFIFYRLI
CAV Ce NP 501743       WSRCQKGEGEQKEENIAIG------VDLVNRDANSMNNHV-QLNFEDIFGEADSQ-HSWDCVWRLNHTVFTAVRLFIYRLV
CAV Cb CAE69013        KCKTKKCENEQKEDHIAIG------MDFVNRDHNGLNNHV-QHNFDDIFGEADSQ-HTWEFMWRLNASVFNWVRLFVYRFI
CAV Ce NP 506357       HKPPNPEMEFD-IGVKNIAPVLIHKMNMDDRDPKD-SAQYLNTSFFEVFNEPSEQYHSIACVWTLSFKIFEIVRIYSYKII
```

Using word embeddings to represent residues



http://www.samyzaf.com/ML/nlp/nlp.html


SS-RSA
CMAP
Output
Softmax
(SS: 3, RSA: 2)
Dense
3X
(11625)
BRNN (31, 375)
/Flatten
2X
(31, 375)
Output
Softmax
(L, L, 2)
2D Convolution
(m, m)
3X
(L, L, 45)
Outer product
(L, 45)
1X
2D Convolution/
Pooling (1, n)
3X
(L, Y, E)
Embedding
(L, Y)
Y
Input MSA


W (Tryptophan)
Nearest points in the original space:
I 0.072
X 0.153
Y 0.207
F 0.229
V 0.323
W
Y
X
F
I
V


K
Nearest points in the original space:
H 0.029
R 0.065
D 0.182
N 0.188
. 0.201
D
N
K
R
H

Things I've learned #3:
Be careful with how you build your training/testing sets

k-fold cross-validation is nice and all, but...

Training sample

Validation sample

(2F08 “Fear of Flying”)

ELSEVIER

Gene

Volume 642, 5 February 2018, Pages 74-83

Research paper

Protein secondary structure prediction based on the fuzzy support vector machine with the hyperplane optimization

Shangxin Xie [a], Zhong Li [a], Hailong Hu [a, b]

Show more

https://doi.org/10.1016/j.gene.2017.11.005

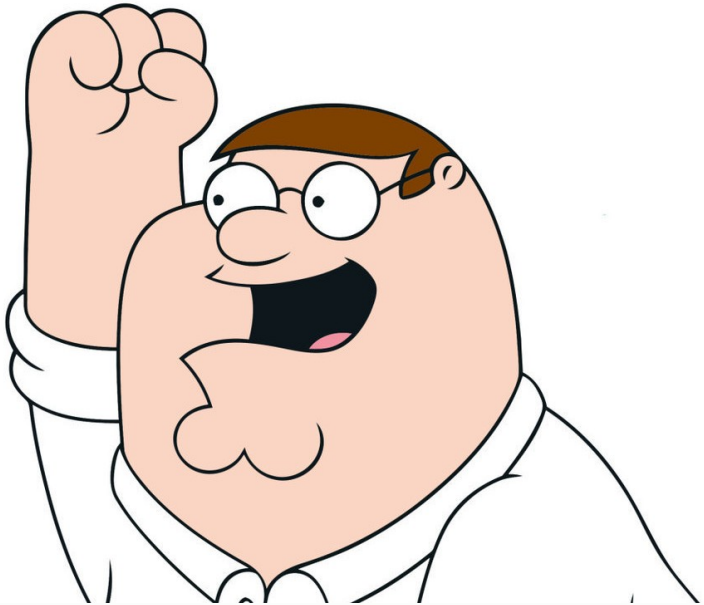
Get rights and content

To reduce the protein sequence similarity (< 25%) between the training set and the test set, we removed 1966 protein sequences from the training set based on the sequence-based structure similarity comparison, obtaining 5986 protein sequences as the training set for the FSVM.

< 25% sequence similarity

Training sample

Validation sample



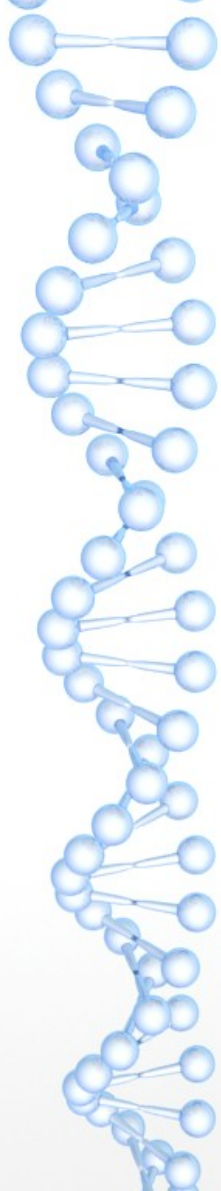
<25% sequence similarity

Porter, PaleAle 4.0: high-accuracy prediction of protein secondary structure and relative solvent accessibility FREE

Claudio Mirabello, Gianluca Pollastri Author Notes

Bioinformatics, Volume 29, Issue 16, 15 August 2013, Pages 2056–2058,
https://doi.org/10.1093/bioinformatics/btt344
Published: 14 June 2013 **Article history** ▾

resolution + r_value/20 or fixed at 10 for NMR structures) and redundancy, and reduced it at a 25% sequence identity threshold, resulting in 9152 proteins. We further selected the 7522 proteins with a quality better than 4 (Full_set), but



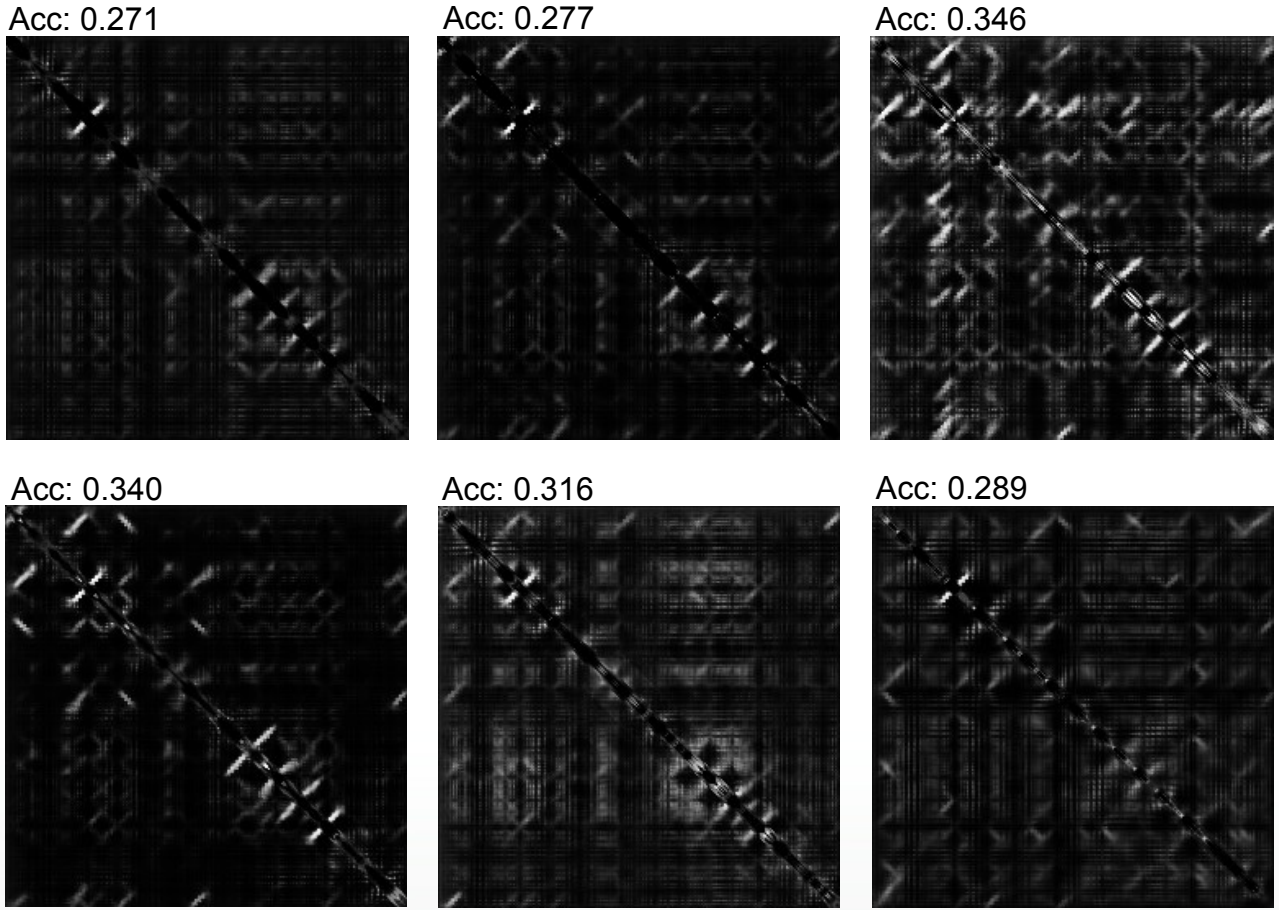
Things I've learned #4:

Not everything you try
will work for your problem
(and it will be frustrating)

Some of the things I've tried this far

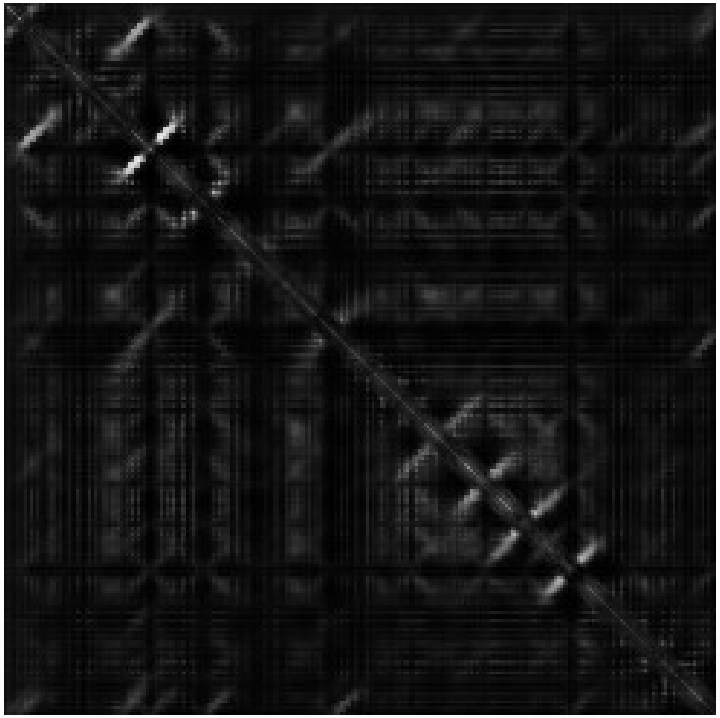
- ✓ Dropout (for SS/RSA)
- ✗ L1, L2 Regularization
- ✓ ✗ ResNets
- ✗ Batch Normalization
- ✗ Very deep networks
- ✗ Other stuff I have forgotten about

Things I've learned #5:
Put many weak models together and you will get a model that is better than any of them


Acc: 0.271
Acc: 0.277
Acc: 0.346
Acc: 0.340
Acc: 0.316
Acc: 0.289

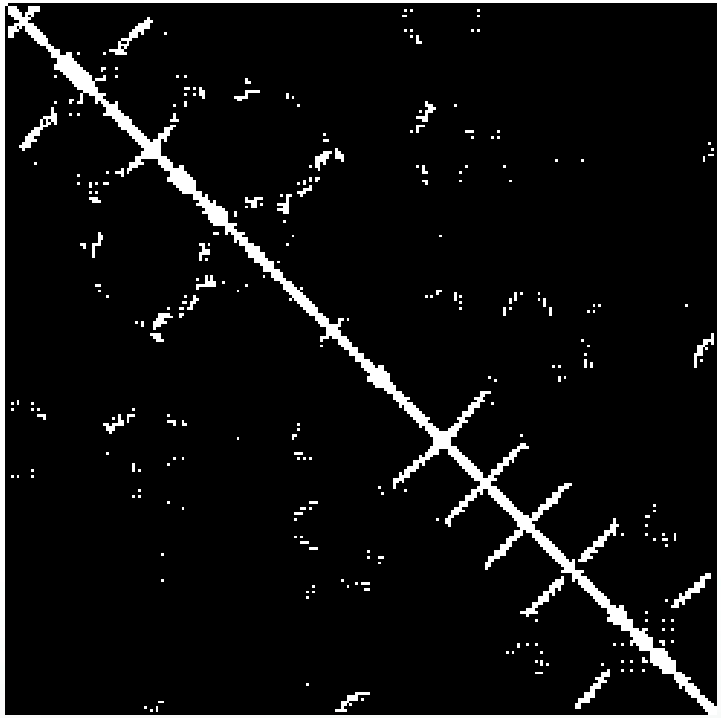
Ensembling the outputs improves the performance

Acc: 0.46

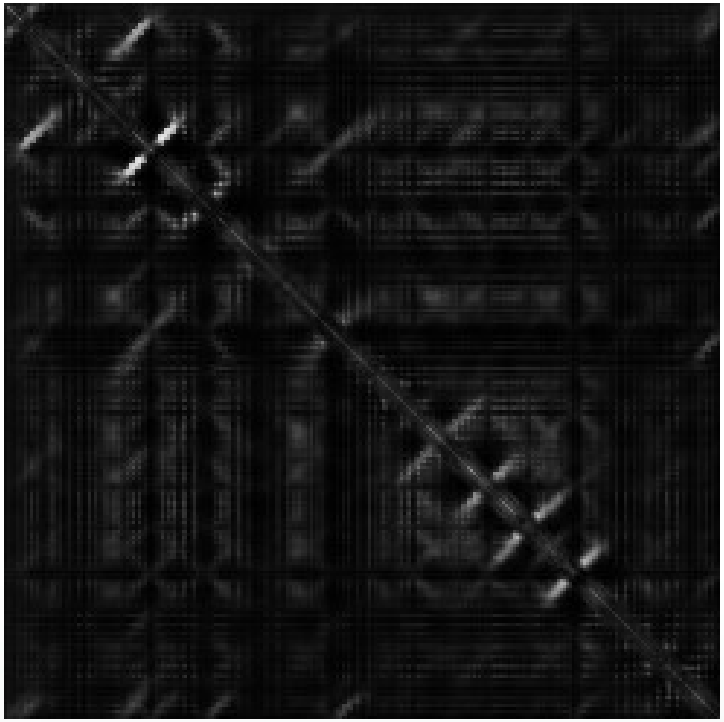


Ensembling the outputs improves the performance

Native

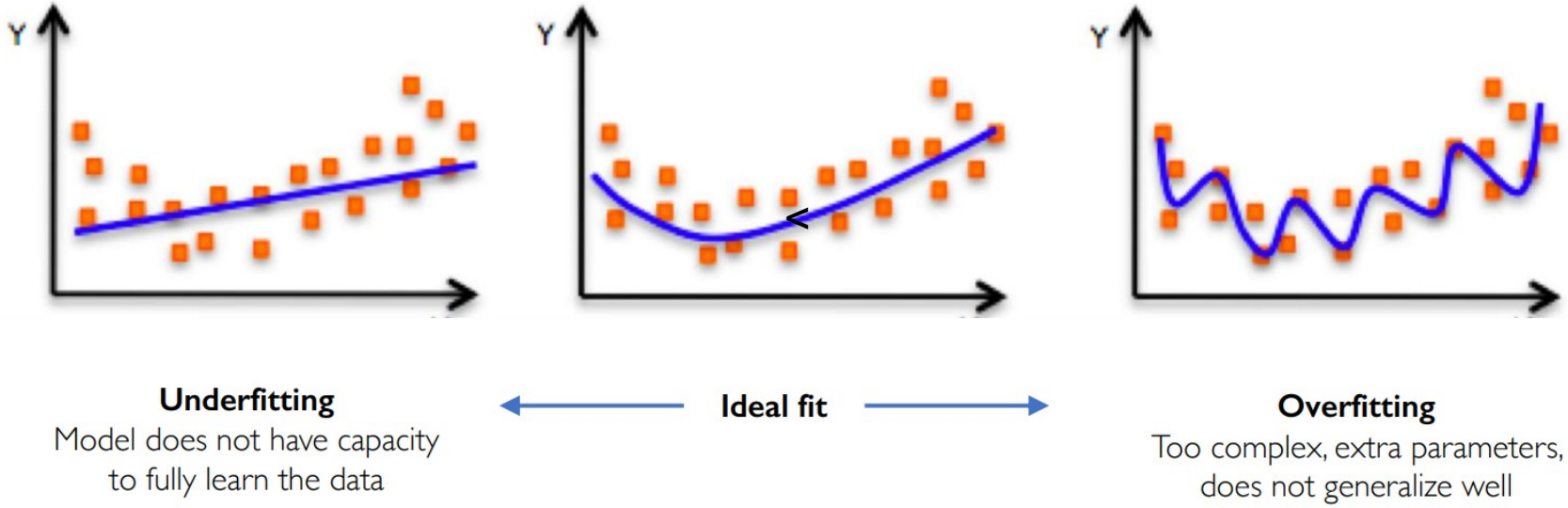


Ensembling the outputs improves the performance

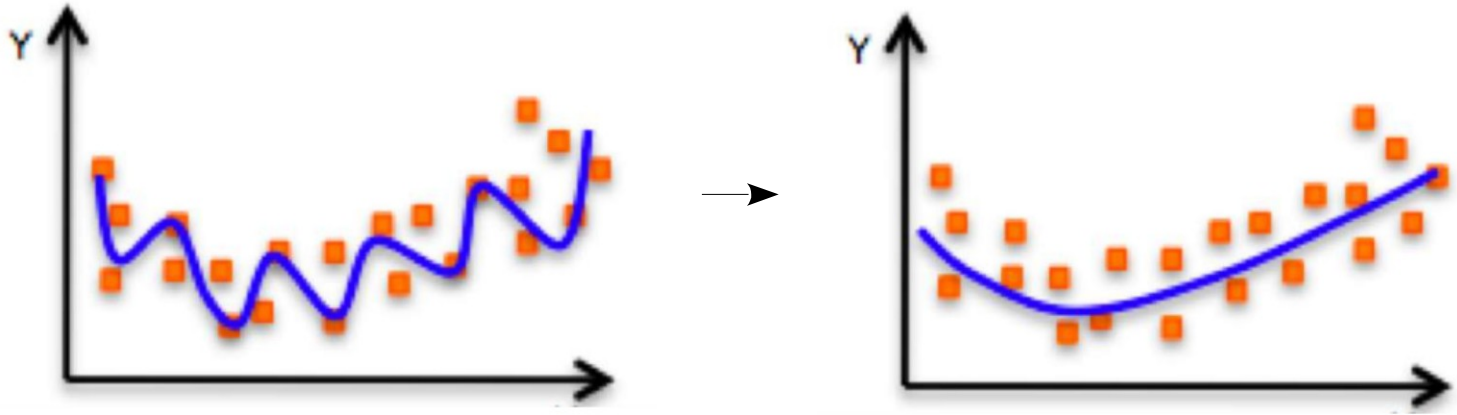


Things I've learned #6:
Overfit first, ask questions later

Overfitting is not necessarily a bad thing at first



If you can’t overfit a model to your data something is wrong

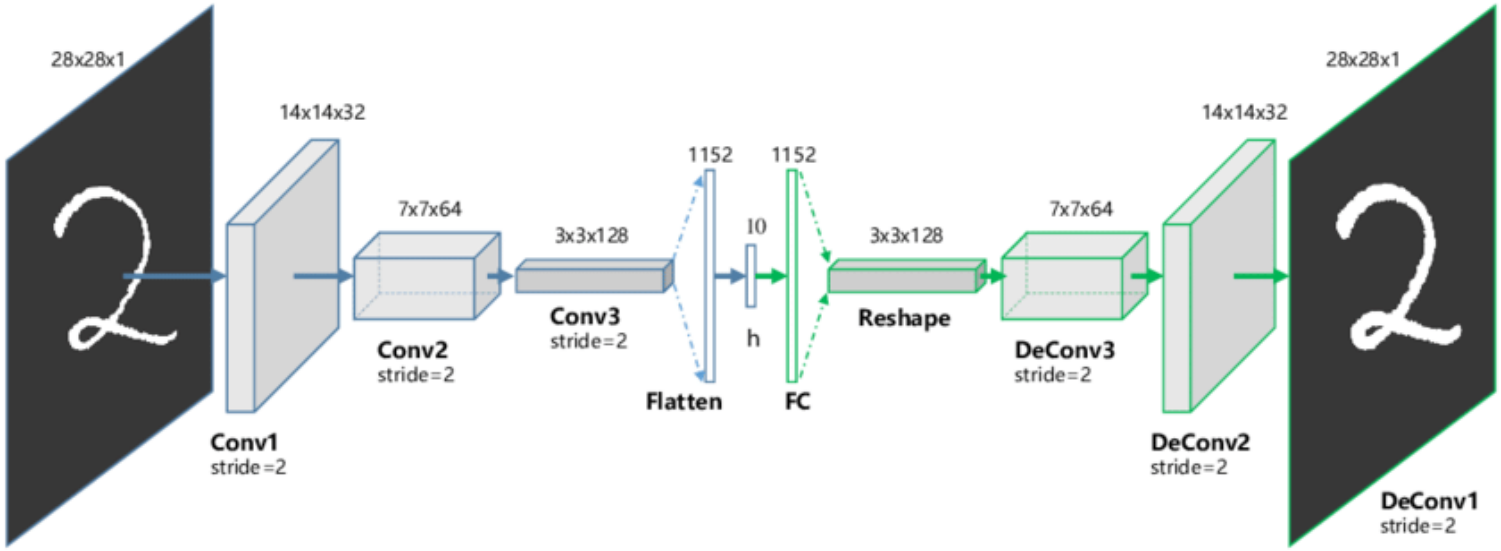


Is your model learning anything at all or making random predictions?

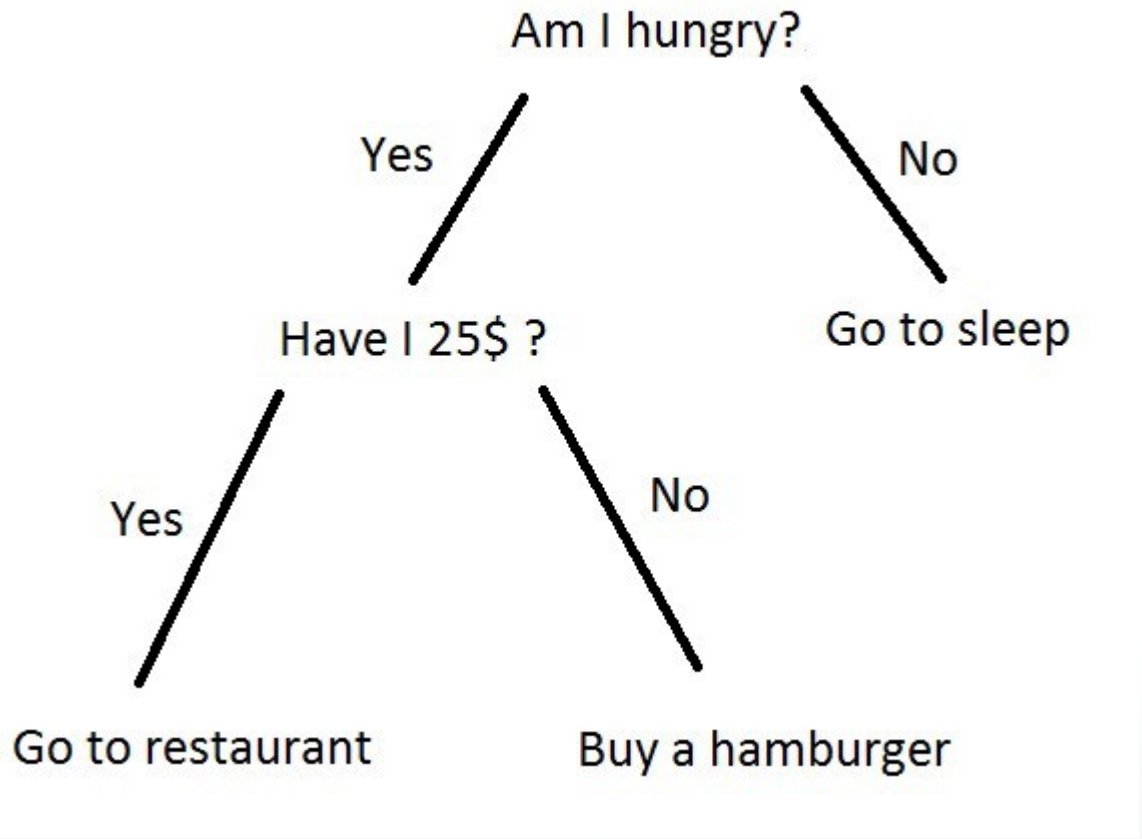
What if I have a lot of inputs and too few samples?

- Do you really need a deep network?
- Really? Are you sure?
- Rule of thumb: you want at least ~10 samples per each weight (Yaser S. Abu-Mostafa, Caltech)
- Reduce the number of inputs to your network, use simpler networks
- Other tricks? (more on that in a second)
- Use your domain knowledge
- Check the literature!

Autoencoders to compress inputs



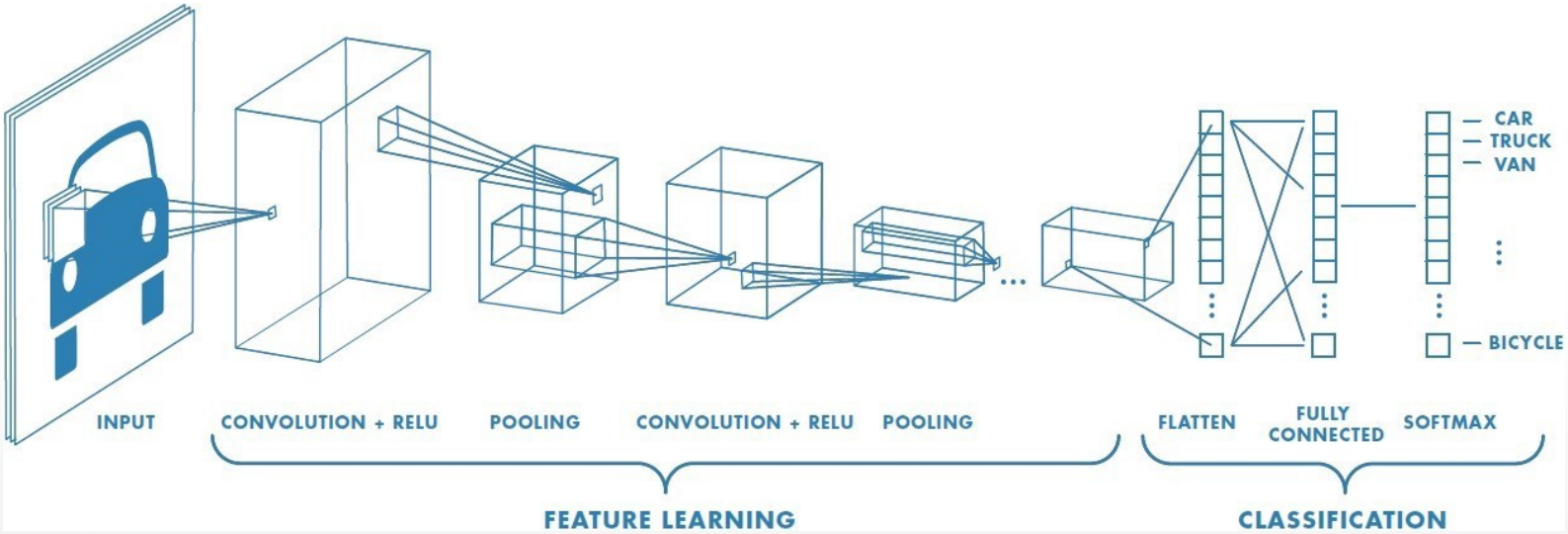
Decision trees for feature selection



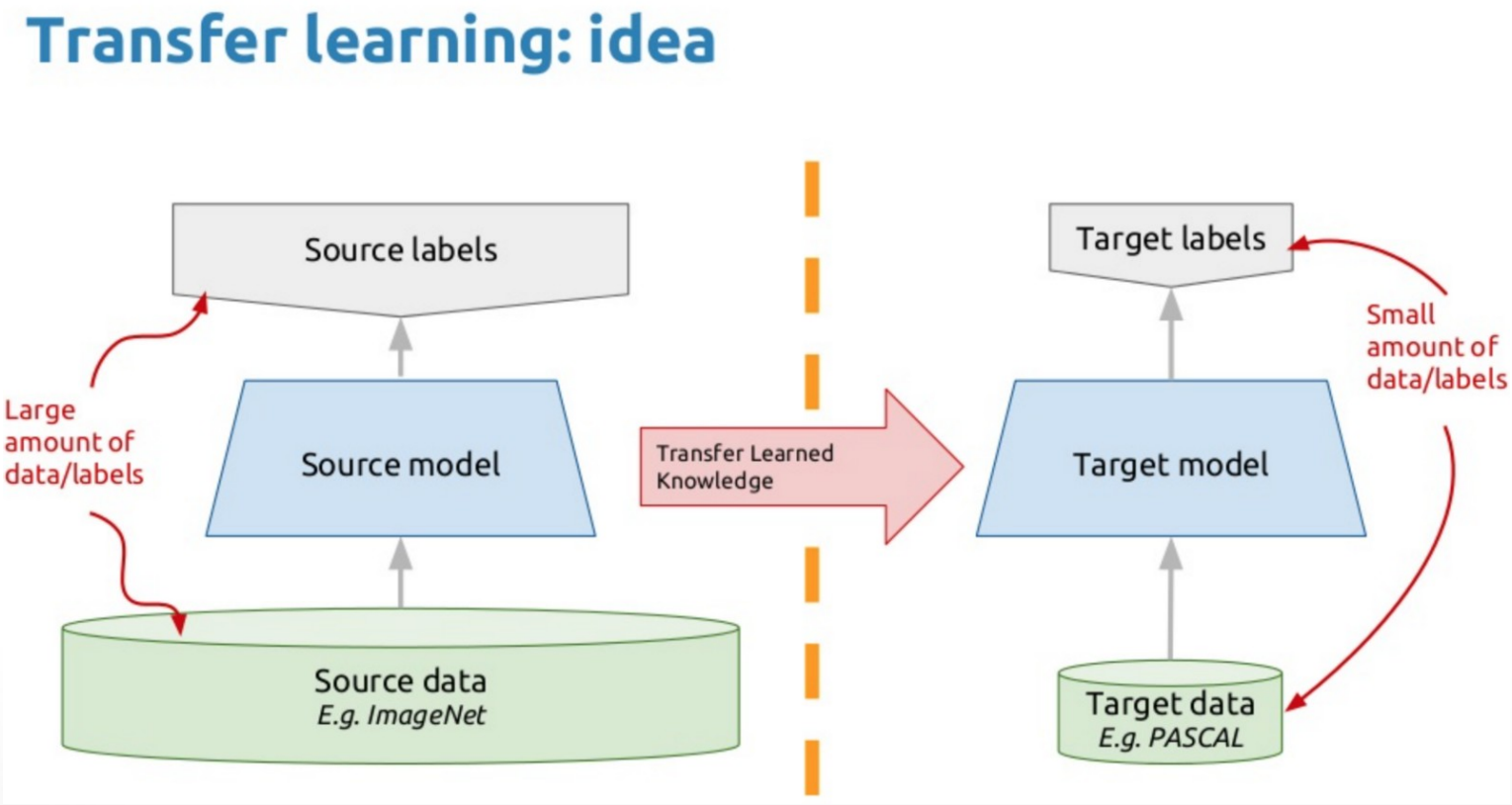
Egor Dezhic

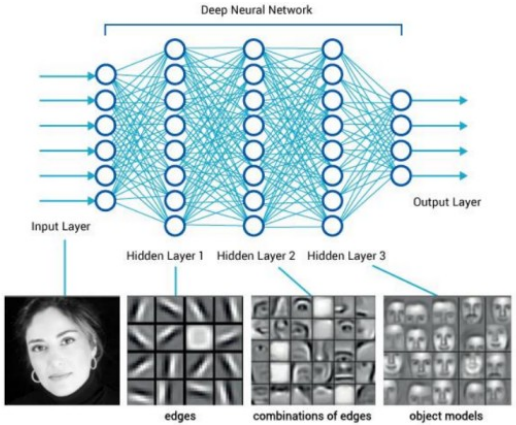
Transfer learning

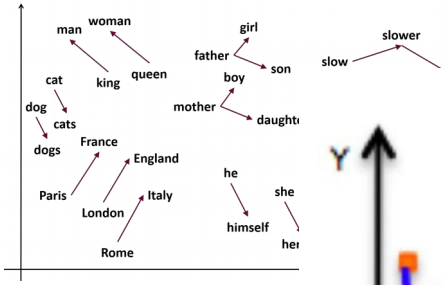
- Idea: use another, larger dataset containing similar data to do the initial training
- When the network has “learned” the features, you can move on to refine the training on your dataset

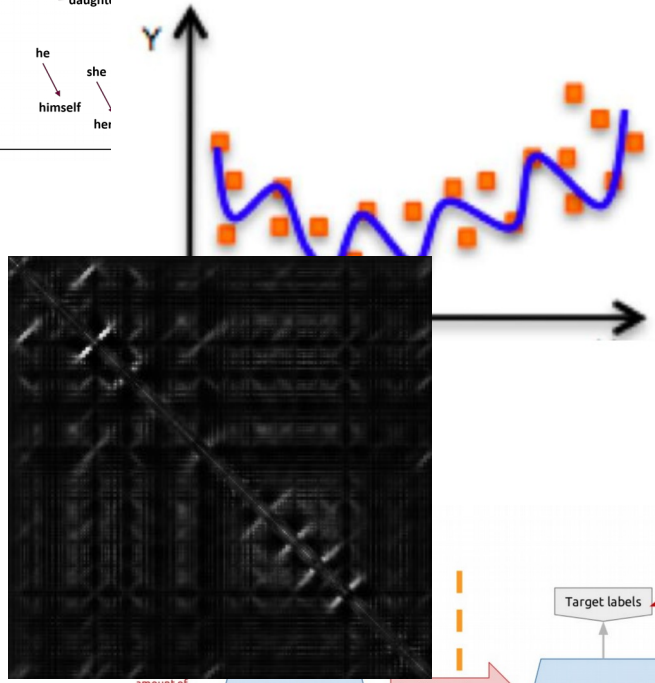


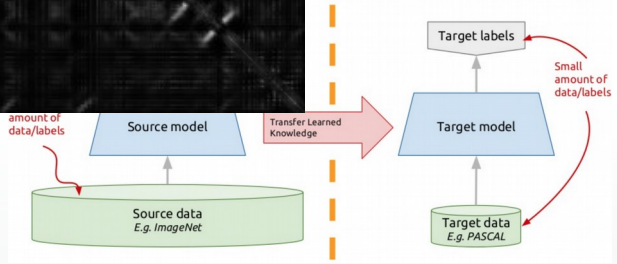
Transfer learning




Deep Neural Network
Input Layer
Hidden Layer 1
Hidden Layer 2
Hidden Layer 3
Output Layer
edges
combinations of edges
object models


man
woman
king
queen
cat
dog
cats
dogs
France
England
Paris
London
Italy
Rome
girl
father
son
boy
mother
he
she
himself
slow
slower


Y


Target labels
Small amount of data/labels
amount of data/labels
Source model
Transfer Learned Knowledge
Target model
Source data
E.g. ImageNet
Target data
E.g. PASCAL