Challenges of virus host prediction with Deep Learning

Florian Mock

FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

The basic idea

- Predict host of a virus based on its sequence
- Better suitable measures to contain epidemics
- We miss information how viruses adapt to there host
- Use Deep Learning, no features need to be defined
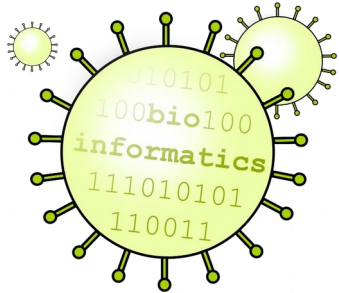
The basic idea

- Predict host of a virus based on its sequence
- Better suitable measures to contain epidemics
- We miss information how viruses adapt to there host
- Use Deep Learning, no features need to be defined
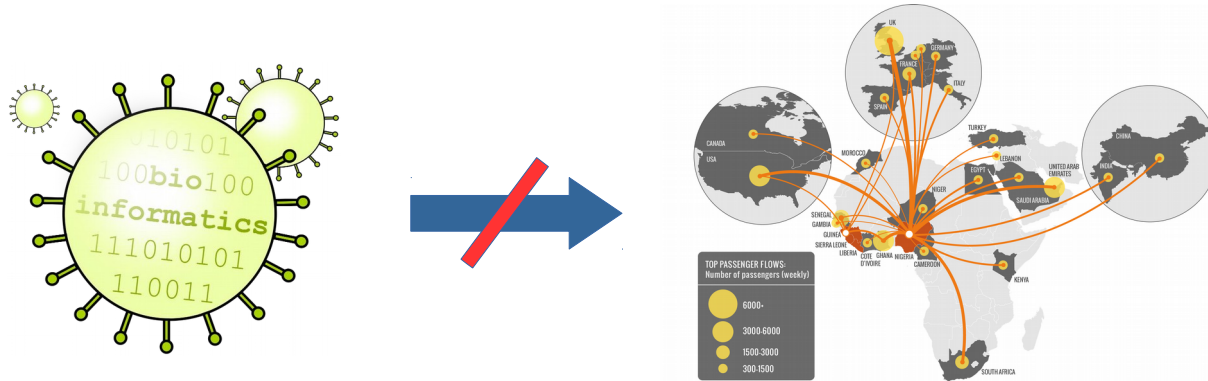
Work-flow

- Get data from database
- Curate and stratify data
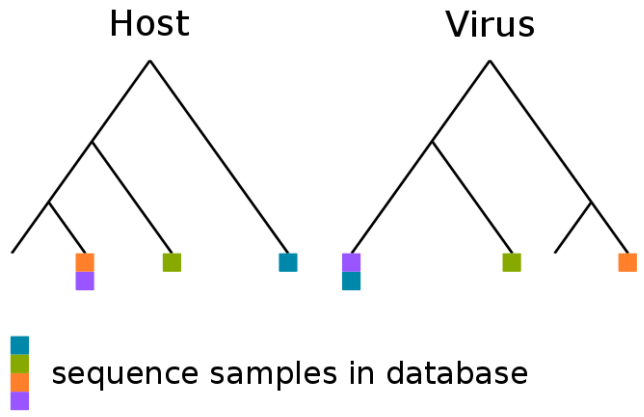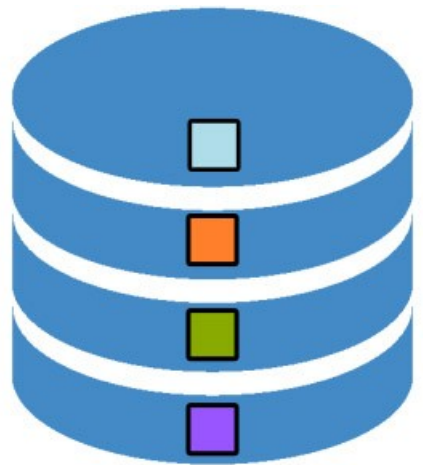- Split in training and test data set
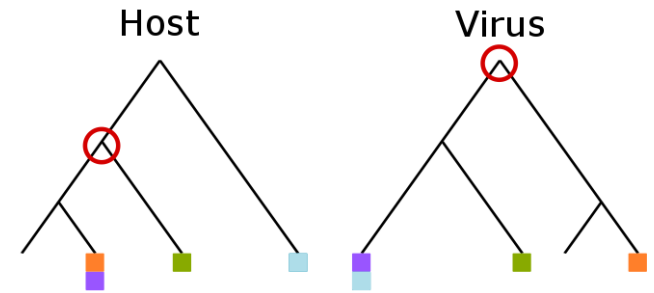- Train neural network
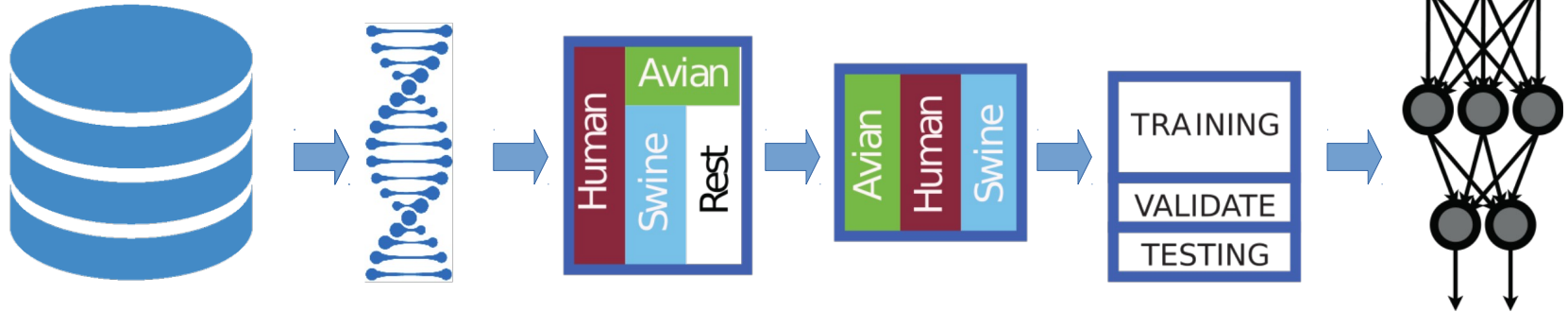
Database

Database

Phylogenetic Tree of Life

Host   Virus

sequence samples in database

FRIEDRICH-SCHILLER-
UNIVERSITÄT
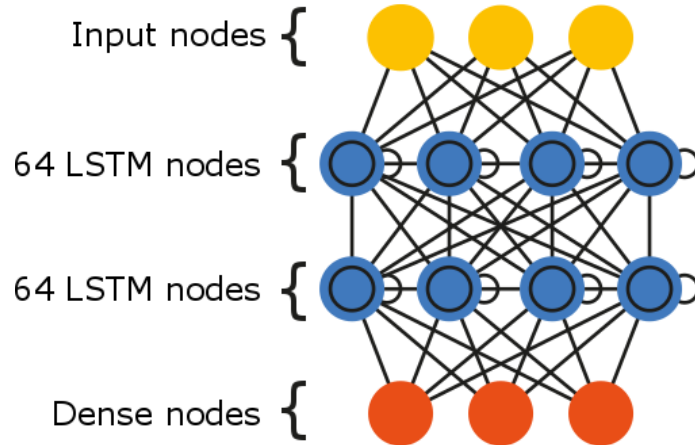JENA

Database

Phylogenetic Tree of Life

Host   Virus

- sequence samples in database
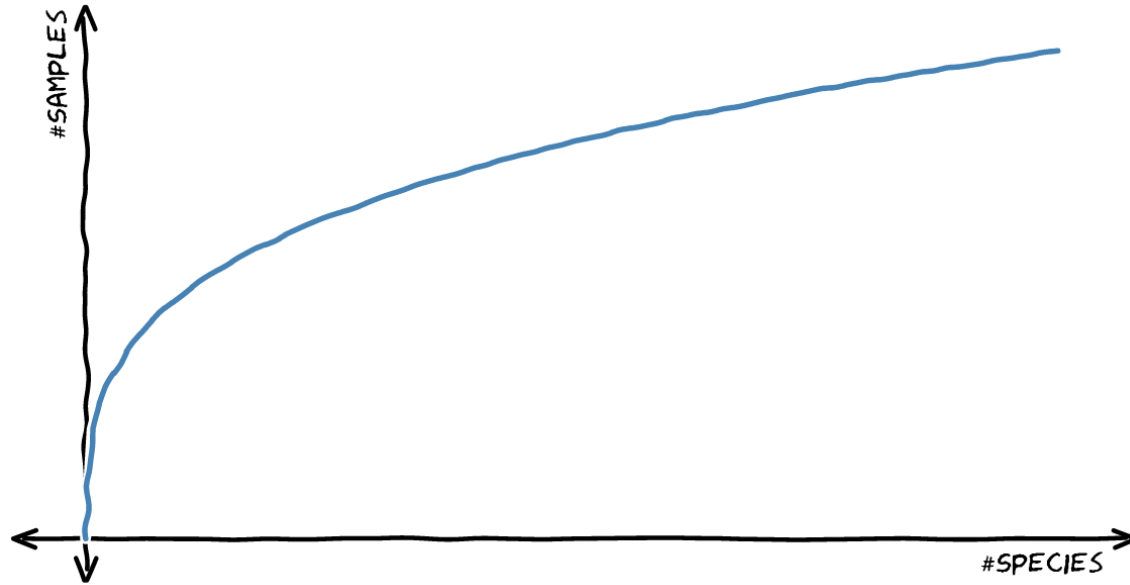- ⭕ select clade of interest
- filtered samples

Architecture

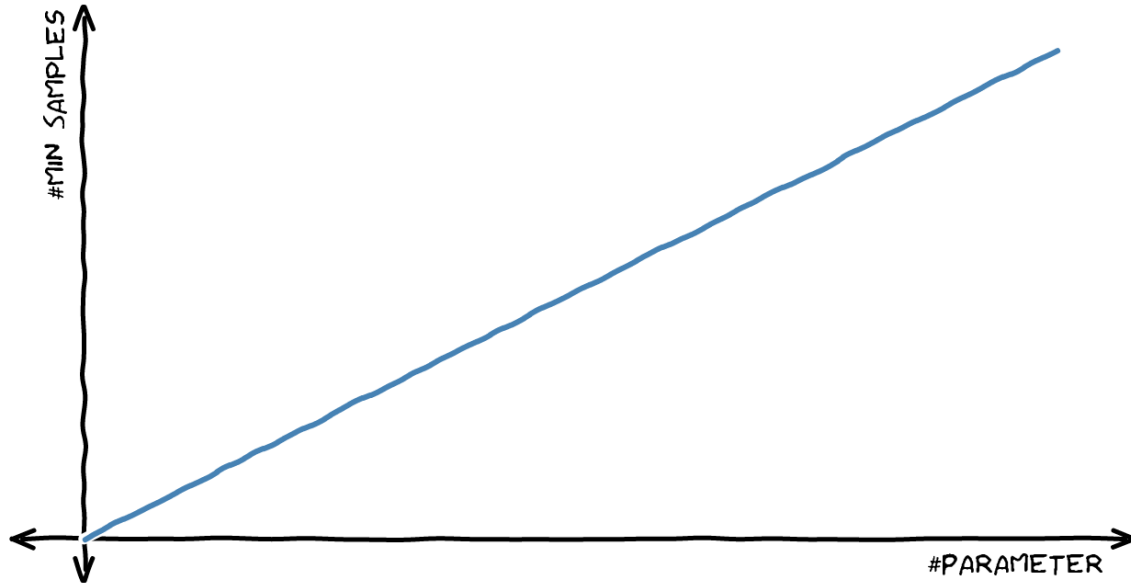# Main Problem

Main Problem

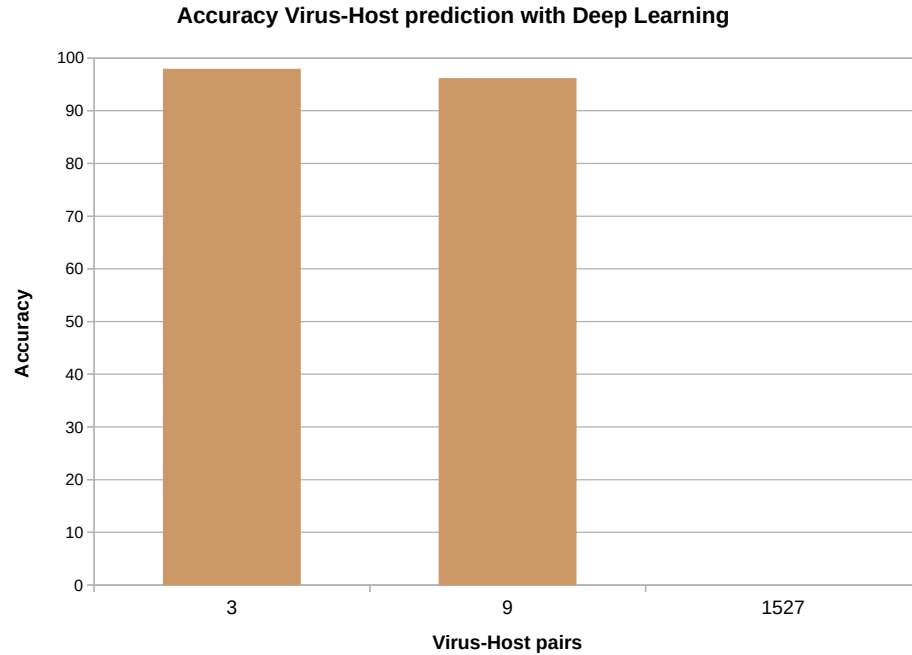WHAT WE GOT

(Graph: #SAMPLES on the y-axis, #SPECIES on the x-axis, with a curve rising steeply then flattening.)

Main Problem

Results

Accuracy Virus-Host prediction with Deep Learning

Results

Accuracy Virus-Host prediction with Deep Learning

FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

Main Problem

Main Problem

WHAT WE GOT

(Graph: y-axis labeled #SAMPLES, x-axis labeled #SPECIES, showing a curve that rises steeply then levels off)

FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA
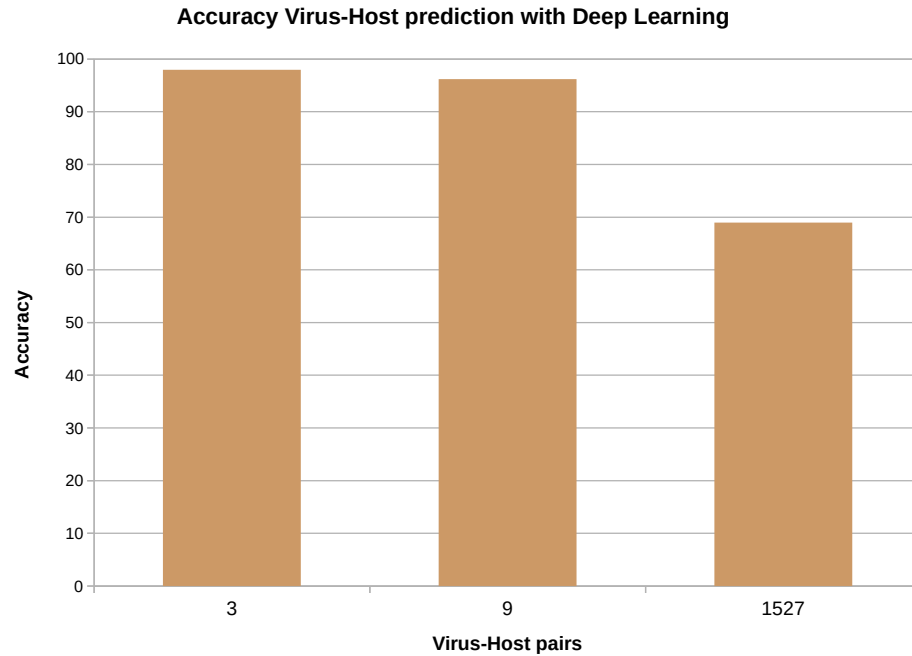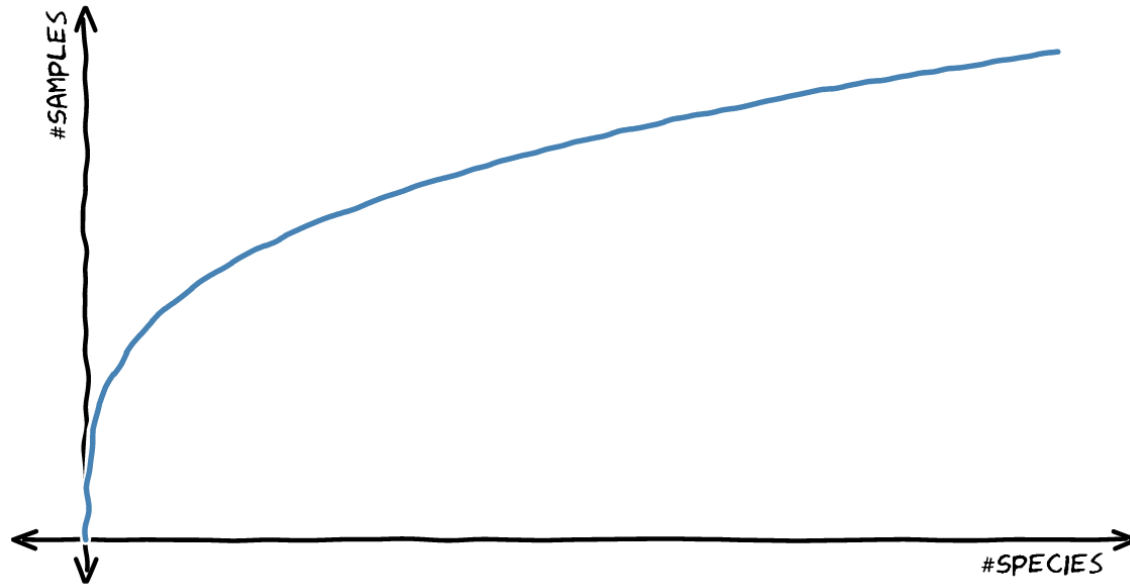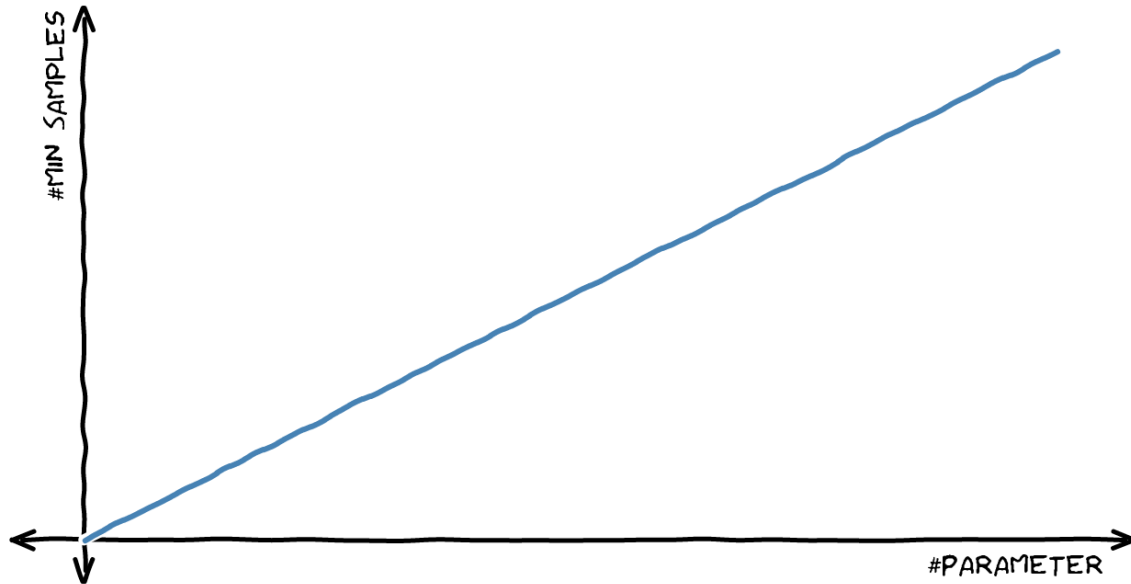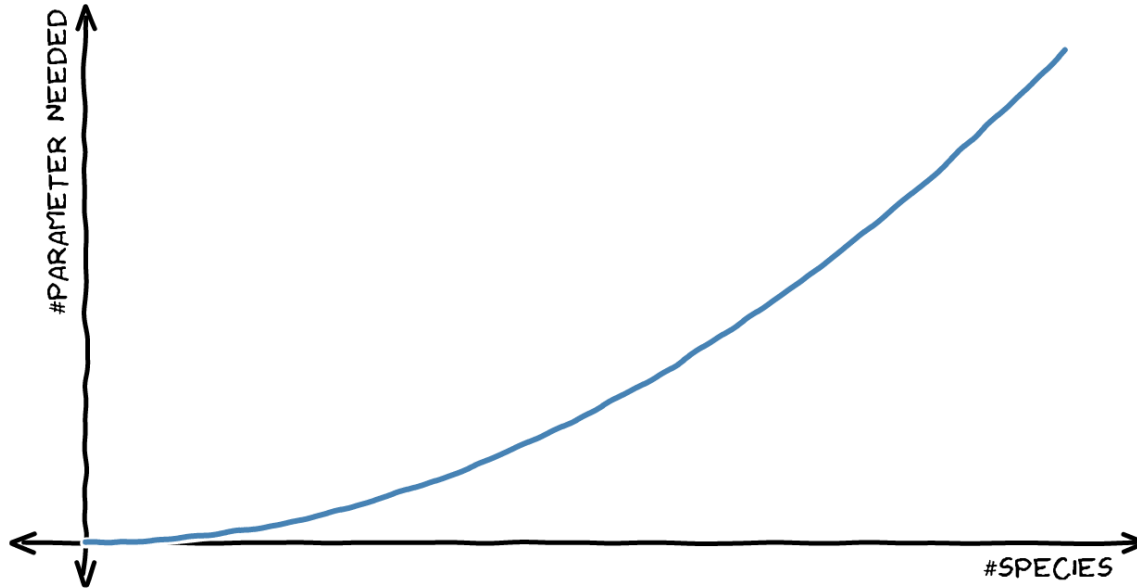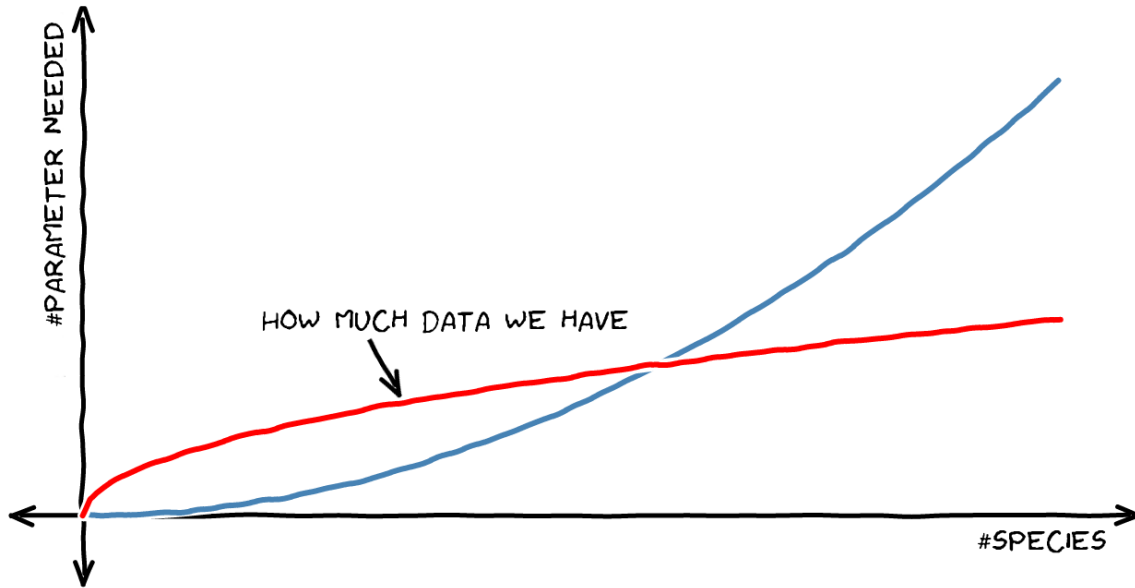
Main Problem

WHAT WE NEED

Main Problem

WHAT WE REALLY NEED

Main Problem

WHAT WE REALLY NEED

(graph with y-axis labeled "#PARAMETER NEEDED" and x-axis labeled "#SPECIES"; red curve annotated "HOW MUCH DATA WE HAVE")

FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

Text mining with CNNs

Text mining with CNNs

Coding

-	A	C	G	T	N
0	0.2	0.4	0.6	0.8	1
⬜	🔲	🔲	🔲	⬛	⬛

FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA
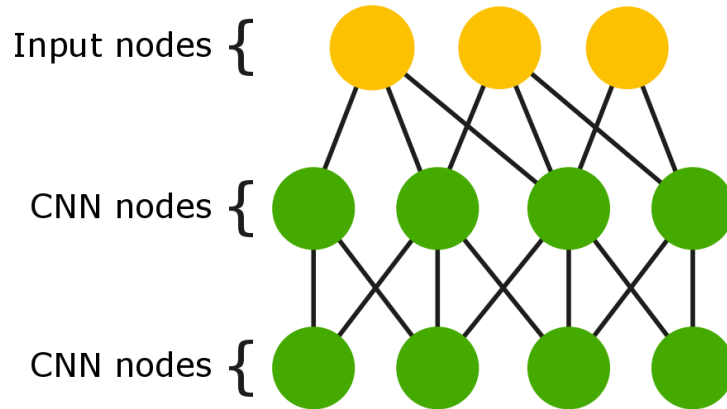
Text mining with CNNs

Text mining with CNNs

Coding

-	A	C	G	T	N
0	0.2	0.4	0.6	0.8	1

working principle

Input nodes

CNN nodes

CNN nodes

+ Complexity

+ Complexity
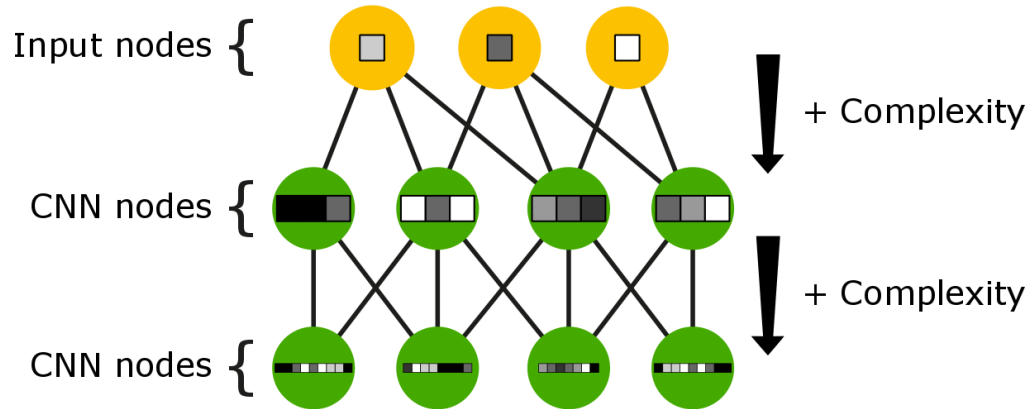
Architecture reloaded

Architecture reloaded

Input nodes {

64 LSTM nodes {

64 LSTM nodes {

Dense nodes {

Architecture reloaded

Input nodes {

64 CNN nodes {

64 CNN nodes {

64 LSTM nodes {

64 LSTM nodes {

Dense nodes {

FRIEDRICH-SCHILLER-
**UNIVERSITÄT
JENA**

Architecture reloaded

Input nodes — Input nodes
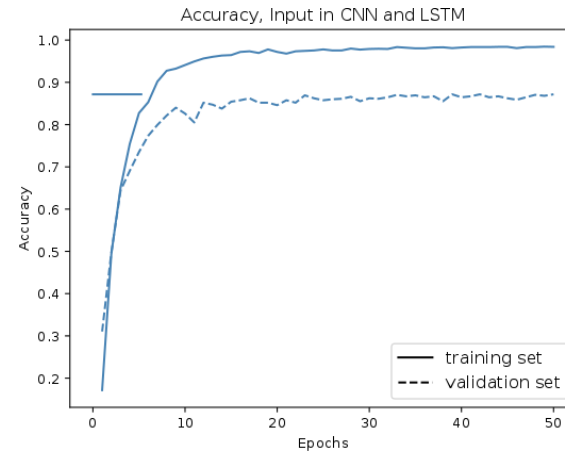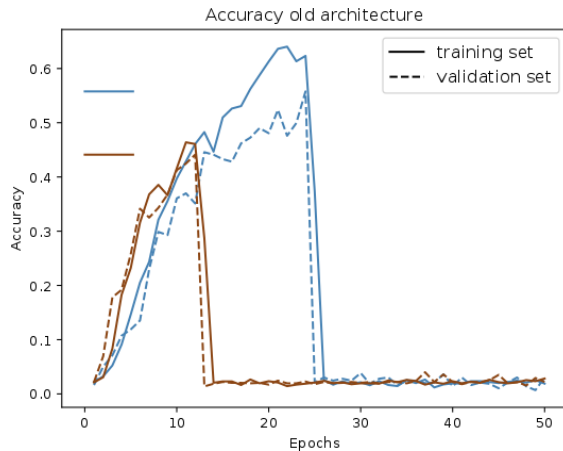
64 LSTM nodes — 64 CNN nodes

64 LSTM nodes — 64 CNN nodes

Dense nodes
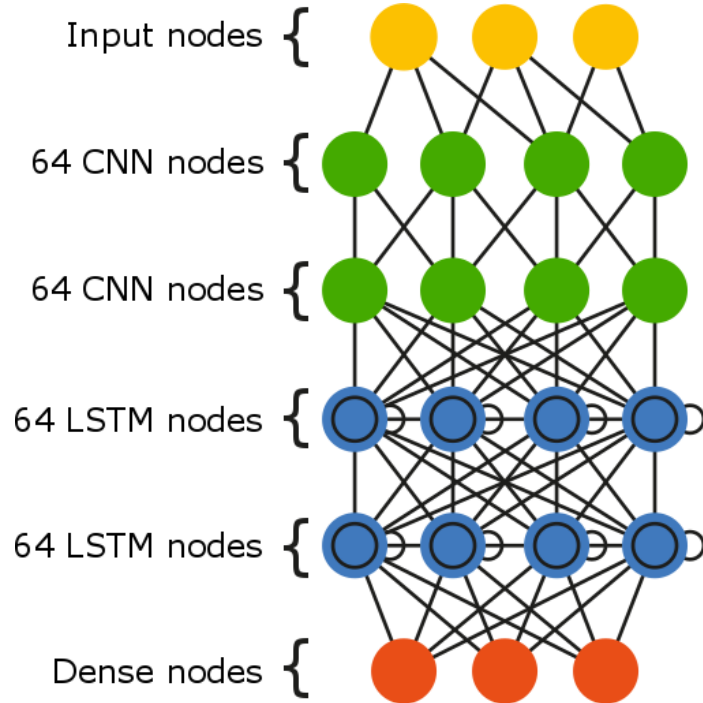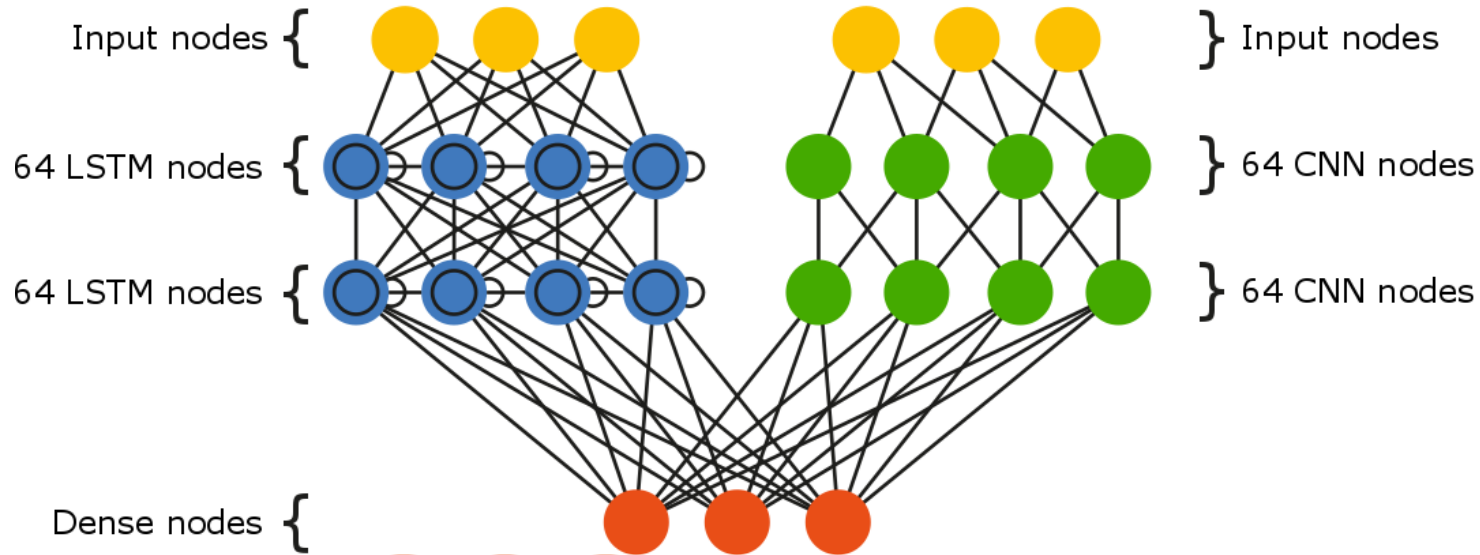
Results

- Both more stable training behavior
- One is training faster
- No accuracy gain

Thank you