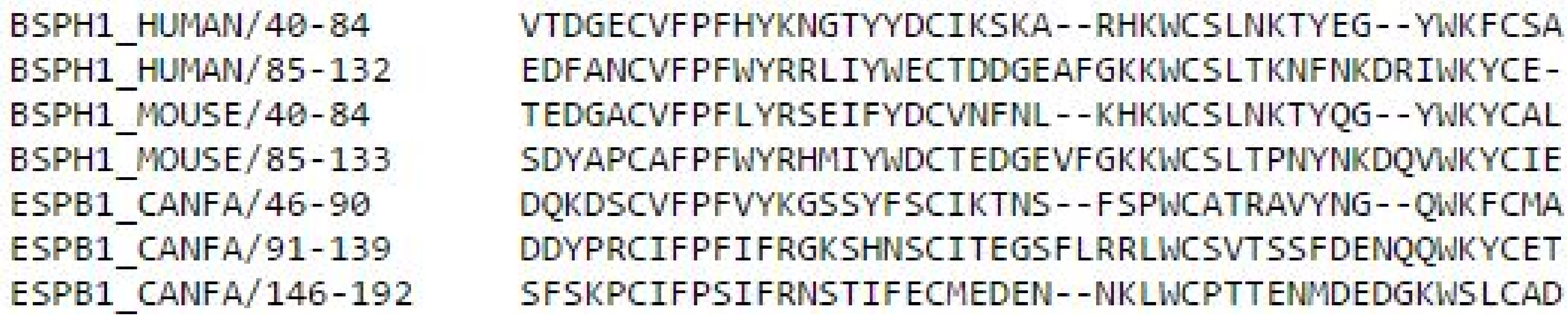
Protein Family Classification

MS. Umberto Di Fabrizio

COMPUTER SCIENCE
COLLEGE OF ENGINEERING
UIC

Purpose

- **Classify** proteins accordingly to their family.



PROTEIN NAME	*PROTEIN SEQUENCE*
BSPH1_HUMAN/40-84	VTDGECVFPFHYKNGTYYDCIKSKA--RHKWCSLNKTYEG--YWKFCSA
BSPH1_HUMAN/85-132	EDFANCVFPFWYRRLIYWECTDDGEAFGKKWCSLTKNFNKDRIWKYCE-
BSPH1_MOUSE/40-84	TEDGACVFPFLYRSEIFYDCVNFNL--KHKWCSLNKTYQG--YWKYCAL
BSPH1_MOUSE/85-133	SDYAPCAFPFWYRHMIYWDCTEDGEVFGKKWCSLTPNYNKDQVWKYCIE
ESPB1_CANFA/46-90	DQKDSCVFPFVYKGSSYFSCIKTNS--FSPWCATRAVYNG--QWKFCMA
ESPB1_CANFA/91-139	DDYPRCIFPFIFRGKSHNSCITEGSFLRRLWCSVTSSFDENQQWKYCET
ESPB1_CANFA/146-192	SFSKPCIFPSIFRNSTIFECMEDEN--NKLWCPTTENMDEDGKWSLCAD

↓

PROTEIN FAMILY

Fibronectin type-II 1

http://prosite.expasy.org/PS51092#TP

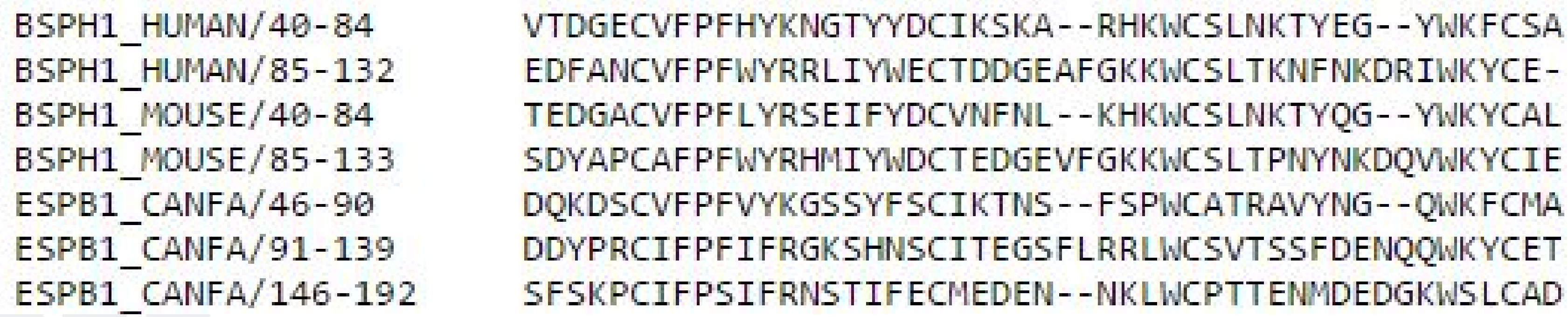
Outline

- Why protein classification?
- How to compare sequences?
- ANN - Pros & Cons
- Pipeline
  - Data preprocessing
  - Data collection
- Implementation
  - Backpropagation
  - Convolutional Neural Network
- Results

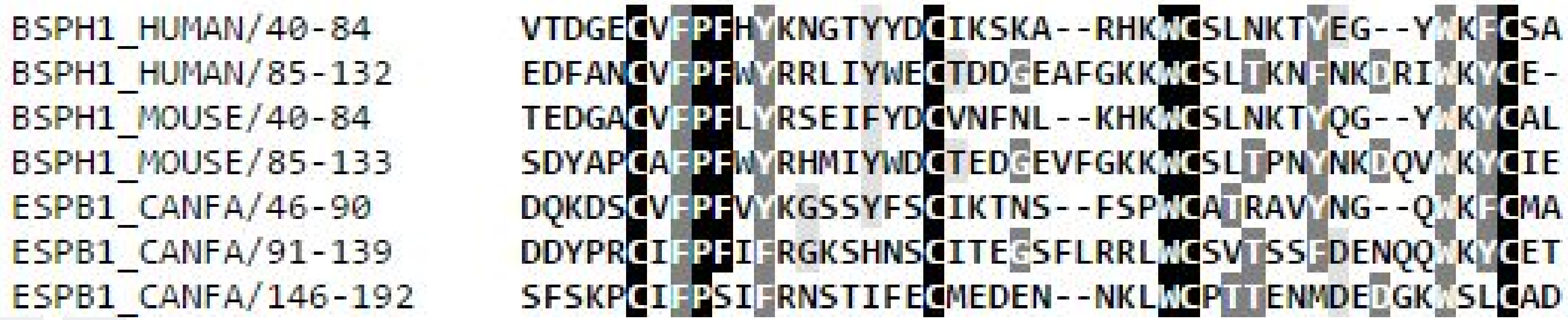
Why protein classification?

- **PURPOSE**: Preventing genetic disease, drug discovery, medical diagnosis.
- **PROBLEM**: Experiments are **costly** and **slow** thus they cannot keep pace with the amount of information available and which needs to be annotated.
- **CLUE:** Similar protein sequences exhibit almost the same biological function.

How to compare sequences?



```
BSPH1_HUMAN/40-84      VTDGECVFPFHYKNGTYYDCIKSKA--RHKWCSLNKTYEG--YWKFCSA
BSPH1_HUMAN/85-132     EDFANCVFPFWYRRLIYWECTDDGEAFGKKWCSLTKNFNKDRIWKYCE-
BSPH1_MOUSE/40-84      TEDGACVFPFLYRSEIFYDCVNFNL--KHKWCSLNKTYQG--YWKYCAL
BSPH1_MOUSE/85-133     SDYAPCAFPFWYRHMIYWDCTEDGEVFGKKWCSLTPNYNKDQVWKYCIE
ESPB1_CANFA/46-90      DQKDSCVFPFVYKGSSYFSCIKTNS--FSPWCATRAVYNG--QWKFCMA
ESPB1_CANFA/91-139     DDYPRCIFPFIFRGKSHNSCITEGSFLRRLWCSVTSSFDENQQWKYCET
ESPB1_CANFA/146-192    SFSKPCIFPSIFRNSTIFECMEDEN--NKLWCPTTENMDEDGKWSLCAD
```



```
BSPH1_HUMAN/40-84      VTDGECVFPFHYKNGTYYDCIKSKA--RHKWCSLNKTYEG--YWKFCSA
BSPH1_HUMAN/85-132     EDFANCVFPFWYRRLIYWECTDDGEAFGKKWCSLTKNFNKDRIWKYCE-
BSPH1_MOUSE/40-84      TEDGACVFPFLYRSEIFYDCVNFNL--KHKWCSLNKTYQG--YWKYCAL
BSPH1_MOUSE/85-133     SDYAPCAFPFWYRHMIYWDCTEDGEVFGKKWCSLTPNYNKDQVWKYCIE
ESPB1_CANFA/46-90      DQKDSCVFPFVYKGSSYFSCIKTNS--FSPWCATRAVYNG--QWKFCMA
ESPB1_CANFA/91-139     DDYPRCIFPFIFRGKSHNSCITEGSFLRRLWCSVTSSFDENQQWKYCET
ESPB1_CANFA/146-192    SFSKPCIFPSIFRNSTIFECMEDEN--NKLWCPTTENMDEDGKWSLCAD
```

ANN Pros&Cons

- Can extract hidden patterns (motif)

BUT

- How to encode Letters?
- Different length of sequences?
- Missing data?

Data Preprocessing

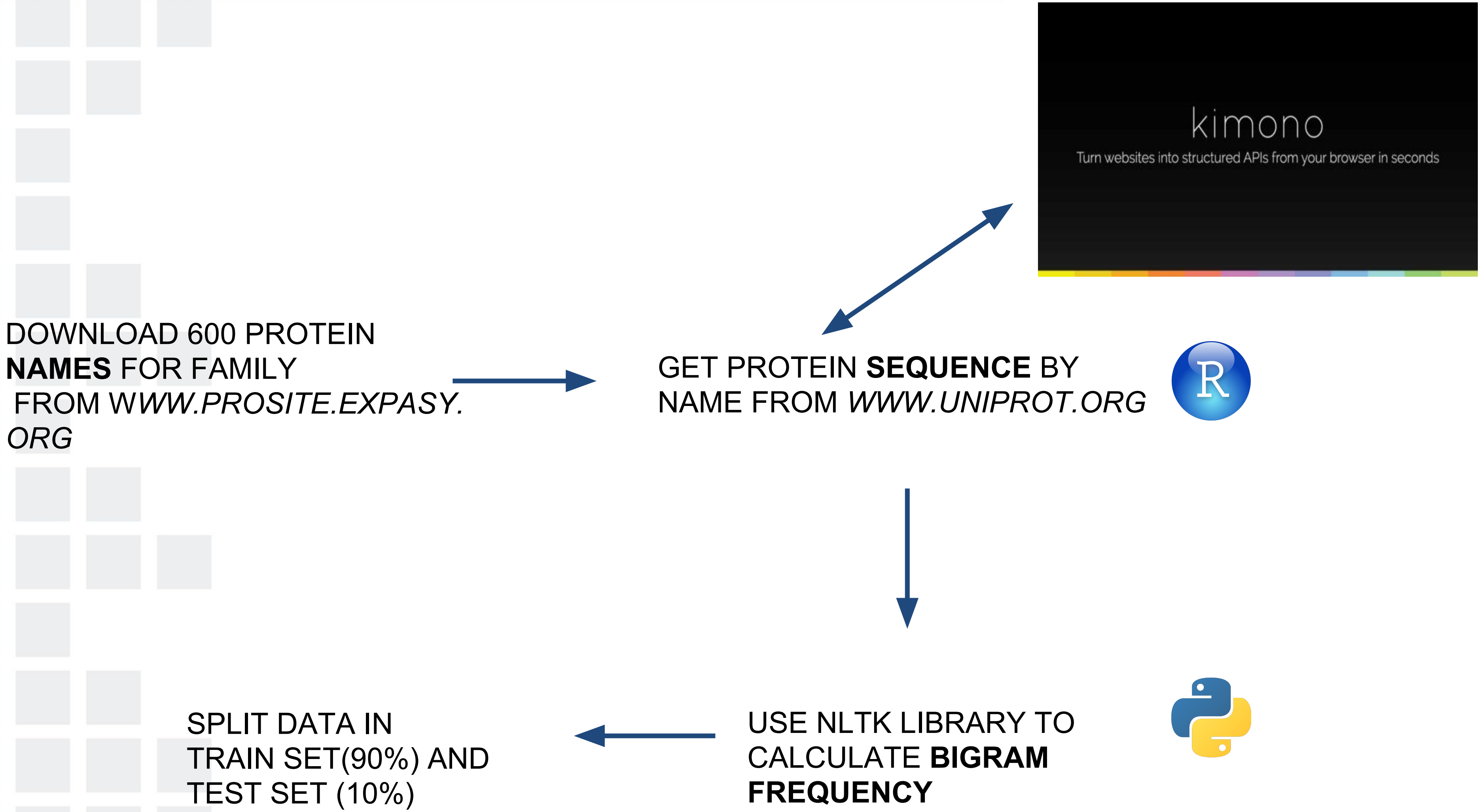
- Exploit NLP techniques:
  - For each protein sequence, analyze the frequency of bigrams.
  - Normalize the count of bigrams on the length of the sequence

Seq= ACTGTGCAT

Number of inputs? length(Seq)^2
In our case (20+3)^2=529 inputs

	A	**C**	**T**	**G**
A	0	1	1	0
C	1	0	1	0
T	0	0	0	2
G	0	1	1	0

Data Collection



kimono

Turn websites into structured APIs from your browser in seconds

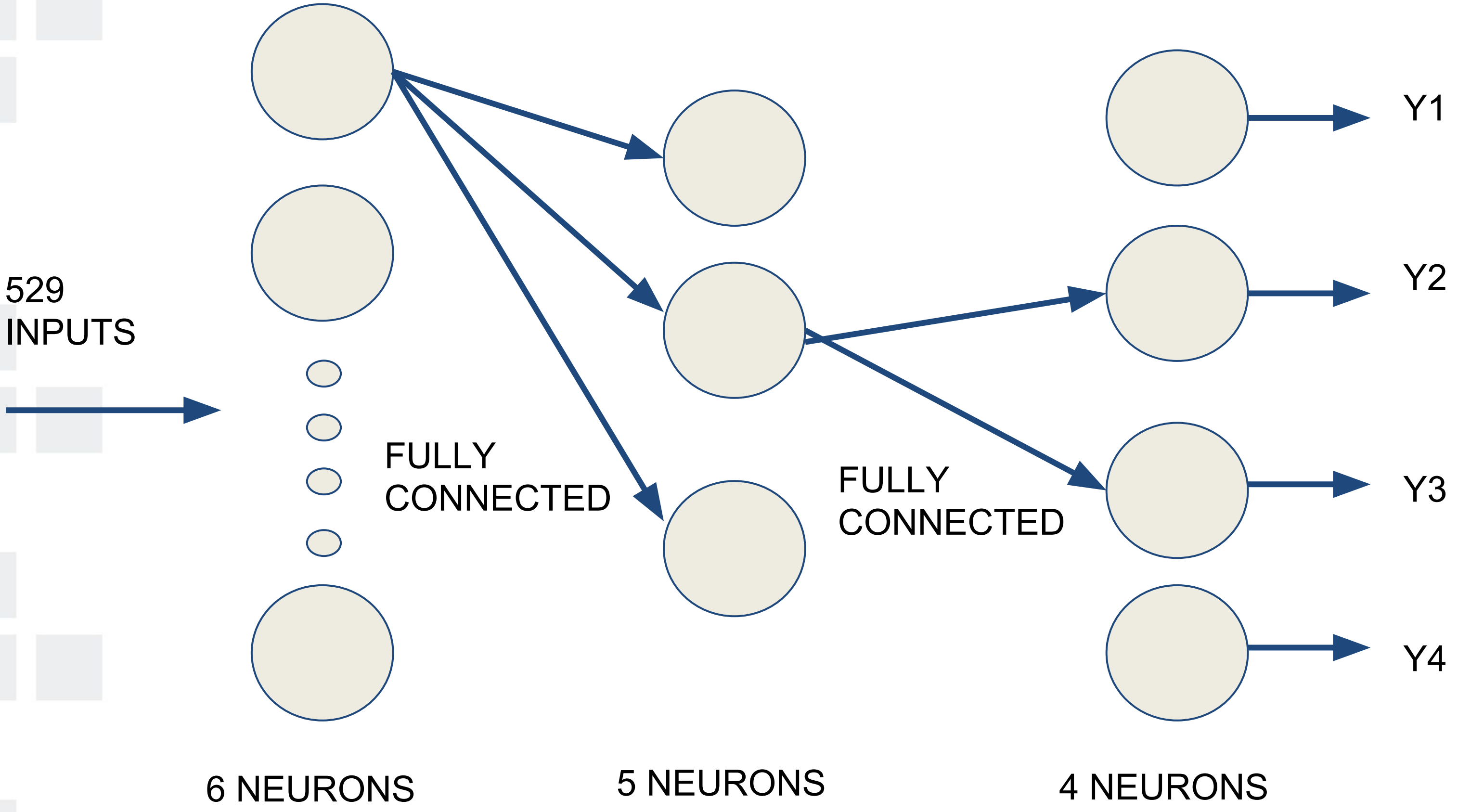
DOWNLOAD 600 PROTEIN **NAMES** FOR FAMILY FROM *WWW.PROSITE.EXPASY.ORG*

GET PROTEIN **SEQUENCE** BY NAME FROM *WWW.UNIPROT.ORG*

USE NLTK LIBRARY TO CALCULATE **BIGRAM FREQUENCY**

SPLIT DATA IN TRAIN SET(90%) AND TEST SET (10%)

Backpropagation

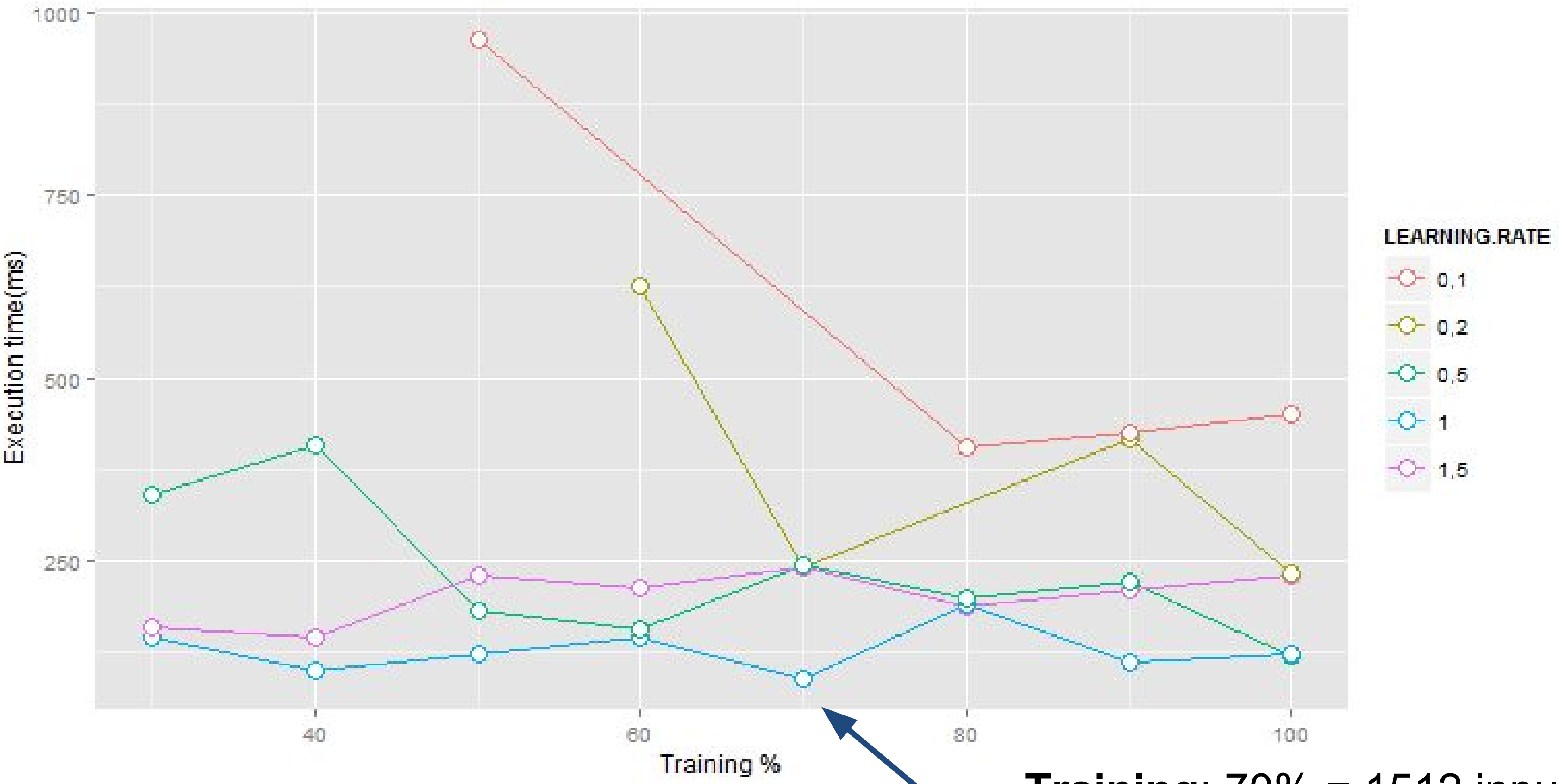


4 FAMILIES:
0001
0010
0100
1000

LEARNING RATE:
[0.1-1.5]

TOTAL WEIGHTS: 529*6+6*5+5*4=3224

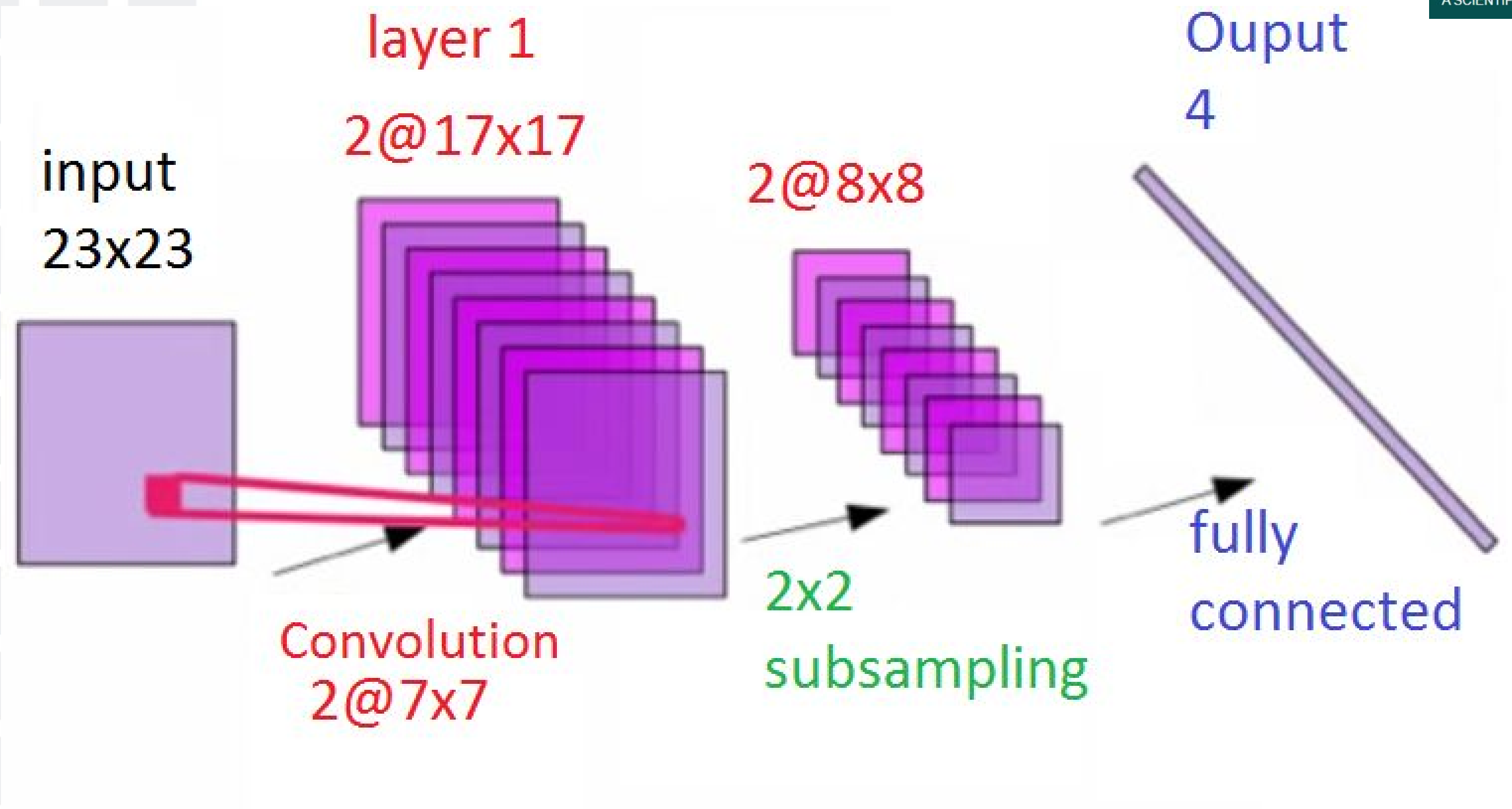
BackProp Results: 95%



Training: 70% = 1512 inputs
Epochs: 5
Execution time: 98(ms)

Convolutional NN: 95%

torch
A SCIENTIFIC COMPUTING FRAMEWORK FOR LUAJIT



EPOCHS: [1,5]
LEARNING RATE: [0.01,0.03,0.05]
FEATURE MAPS: [2,3]
TRAINING % = 70

TOTAL WEIGHTS: 2*7*7+2*2*2+64*2*4=618
EXECUTION TIME: 669(ms)

Future Work

- Evaluate the performances of the LAMSTAR

Questions?