contentSkip to site indexPoliticsSubscribeLog InSubscribeLog InToday's PaperAdvertisementSupported ORG byF.B.I. Agent Peter Strzok PERSON , Who Criticized Trump PERSON in Texts, Is FiredImagePeter Strzok, a top F.B.I. GPE counterintelligence agent who was taken off the special counsel investigation after his disparaging texts about President Trump PERSON were uncovered, was fired. CreditT.J. Kirkpatrick PERSON for The New York TimesBy Adam Goldman ORG and Michael S. SchmidtAug PERSON . 13 CARDINAL , 2018WASHINGTON CARDINAL — Peter Strzok PERSON , the F.B.I. GPE senior counterintelligence agent who disparaged President Trump PERSON in inflammatory text messages and helped oversee the Hillary Clinton PERSON email and Russia GPE investigations, has been fired for violating bureau policies, Mr. Strzok PERSON 's lawyer said Monday DATE .Mr. Trump and his allies seized on the texts — exchanged during the 2016 DATE campaign with a former F.B.I. GPE lawyer, Lisa Page — in PERSON assailing the Russia GPE investigation as an illegitimate "witch hunt." Mr. Strzok PERSON , who rose over 20 years DATE at the F.B.I. GPE to become one of its most experienced counterintelligence agents, was a key figure in the early months DATE of the inquiry.Along with writing the texts, Mr. Strzok PERSON was accused of sending a highly sensitive search warrant to his personal email account.The F.B.I. GPE had been under immense political pressure by Mr. Trump PERSON to dismiss Mr. Strzok PERSON , who was removed last summer DATE from the staff of the special counsel, Robert S. Mueller III PERSON . The president has repeatedly denounced Mr. Strzok PERSON in posts on

WEB CONTENT MINING
NAMED ENTITY RECOGNITION

Elina Tugaeva, Eliana Ruslanova, Maida Nazifi,
Nikolaj Kolbasko, Samail Guliev

Web Mining IE684 Course
Faculty of Business Informatics and Mathematics
University of Mannheim, 2020

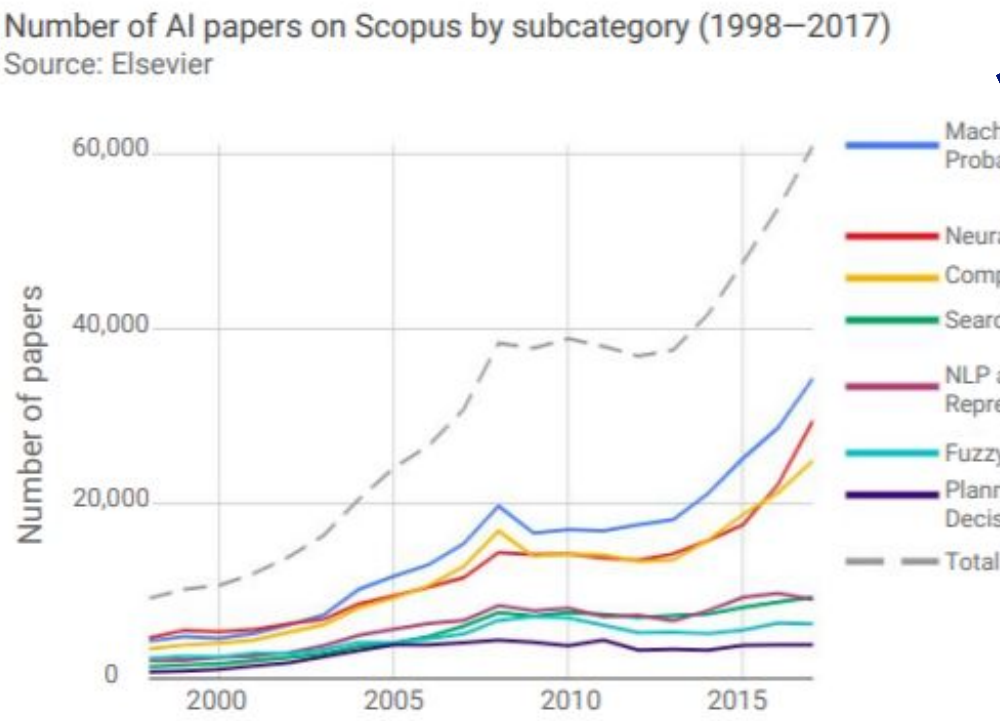
AGENDA

1. Motivation
2. Structure and size of the data set
3. CRF
4. LSTM-CRF
5. RNN - LSTM
6. Character embeddings
7. LSTM-CNN-CRF
8. BERT
9. SpaCy

AGENDA

1. Motivation
2. Structure and size of the data set
3. CRF
4. LSTM-CRF
5. RNN - LSTM
6. Character embeddings
7. LSTM-CNN-CRF
8. BERT
9. SpaCy

Motivation



Number of AI papers on Scopus by subcategory (1998—2017)

Source: Elsevier

AGENDA

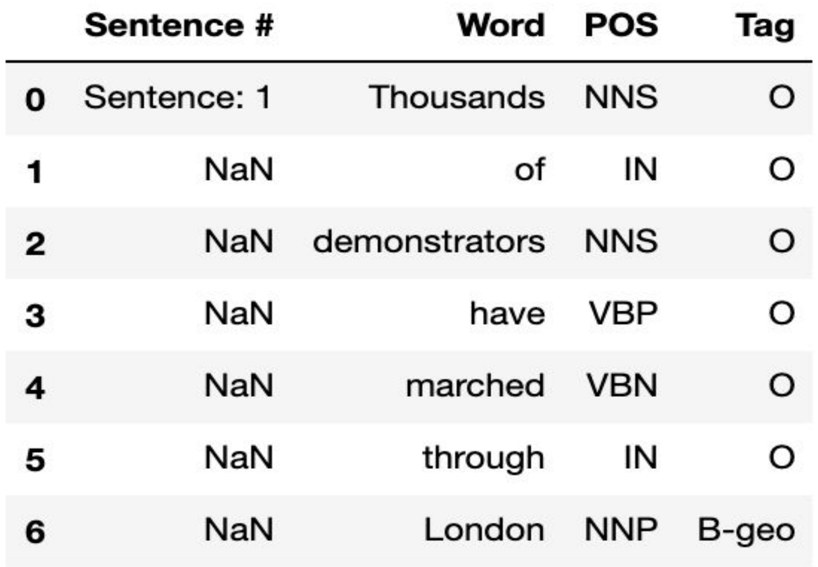
1. Motivation
2. Structure and size of the data set
3. CRF
4. LSTM-CRF
5. RNN - LSTM
6. Character embeddings
7. LSTM-CNN-CRF
8. BERT
9. SpaCy

Data Set

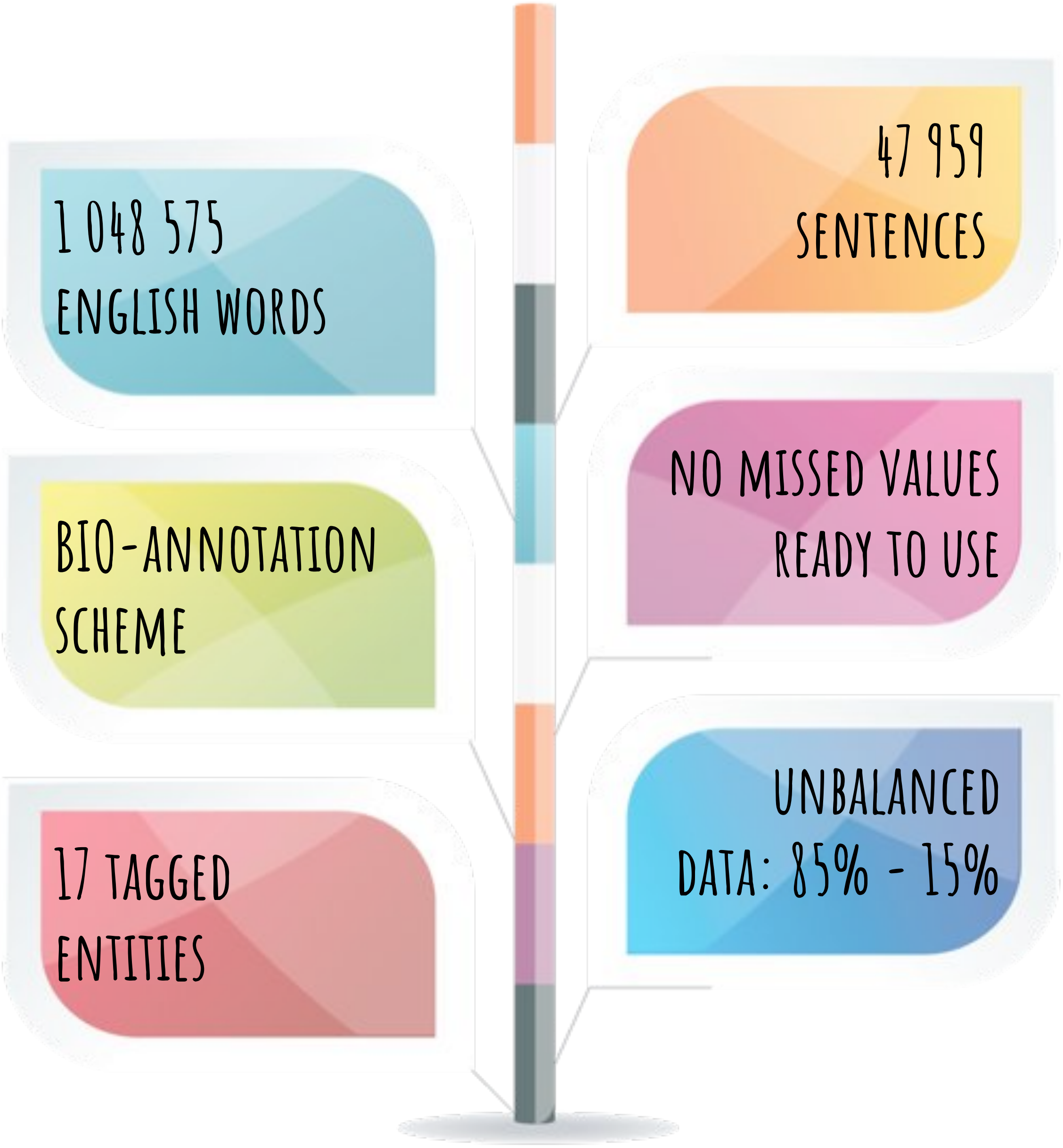
Annotated Corpus for

Named Entity Recognition

from Kaggle



	Sentence #	Word	POS	Tag
0	Sentence: 1	Thousands	NNS	O
1	NaN	of	IN	O
2	NaN	demonstrators	NNS	O
3	NaN	have	VBP	O
4	NaN	marched	VBN	O
5	NaN	through	IN	O
6	NaN	London	NNP	B-geo



1 048 575 English words

47 959 sentences

BIO-annotation scheme

No missed values ready to use

17 tagged entities

Unbalanced data: 85% - 15%

AGENDA

1. Motivation
2. Structure and size of the data set
3. CRF
4. LSTM-CRF
5. RNN - LSTM
6. Character embeddings
7. LSTM-CNN-CRF
8. BERT
9. SpaCy

Conditional random field

- A probabilistic, discriminative classification model for sequences
- Directly model the association between the observed features and labels for those features: defines a posterior probability of a label sequence given an input observation sequence

$$p(s_1, .., s_m | x_1, .., x_m, w) = \frac{exp(w\Phi(x, s))}{\sum exp(w\Phi(x, s))}$$

$$L = \sum_{i=1}^{n} p(s^i | x^i, w) - \frac{\lambda_2}{2} \times |w|^2 - \lambda_1 \times |w|$$

REGULARIZED

F-SCORE

65%

F-SCORE

56%

AGENDA

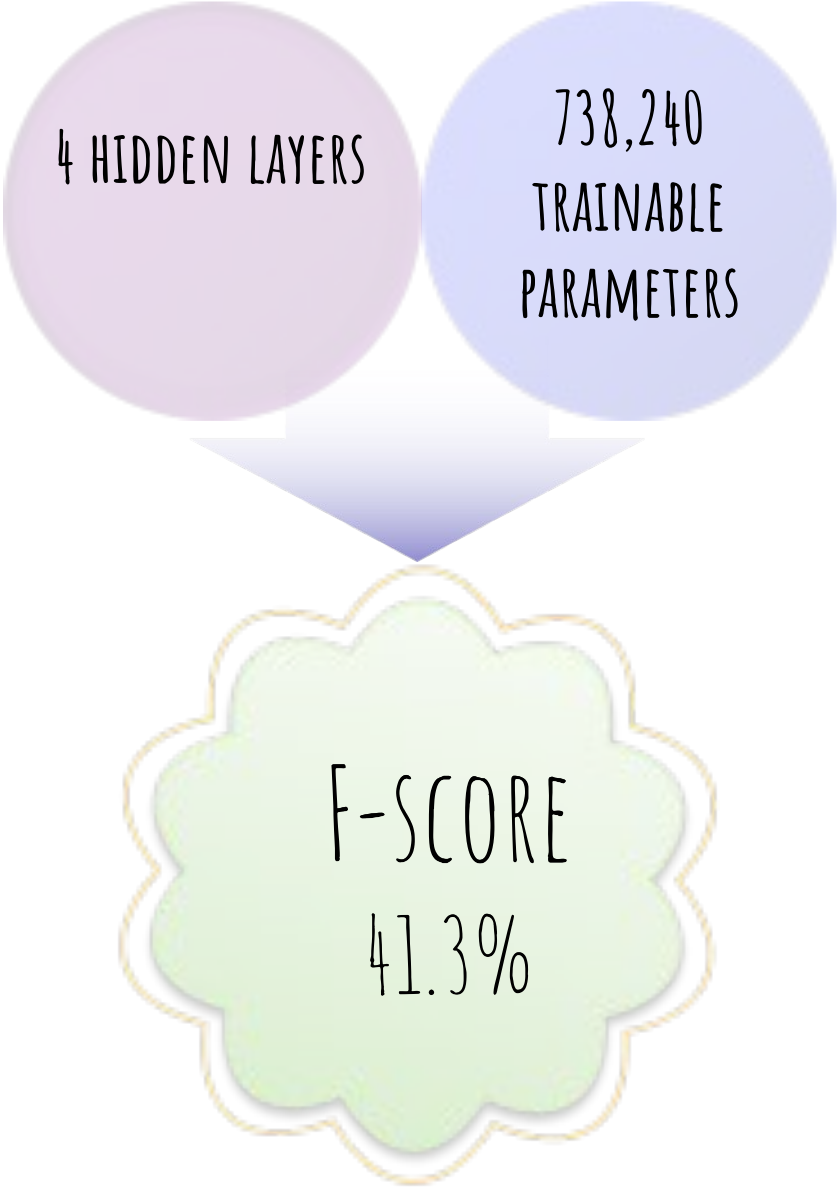
1. Motivation
2. Structure and size of the data set
3. CRF
4. RNN - LSTM
5. LSTM-CRF
6. Character embeddings
7. LSTM-CNN-CRF
8. BERT
9. SpaCy

RNN

- Deep learning Method known for good results in NER Tasks
    - BiDirectional RNN

- Embedding layer fed into RNN with a Dropout Layer = 0.5

- long-term dependency problem



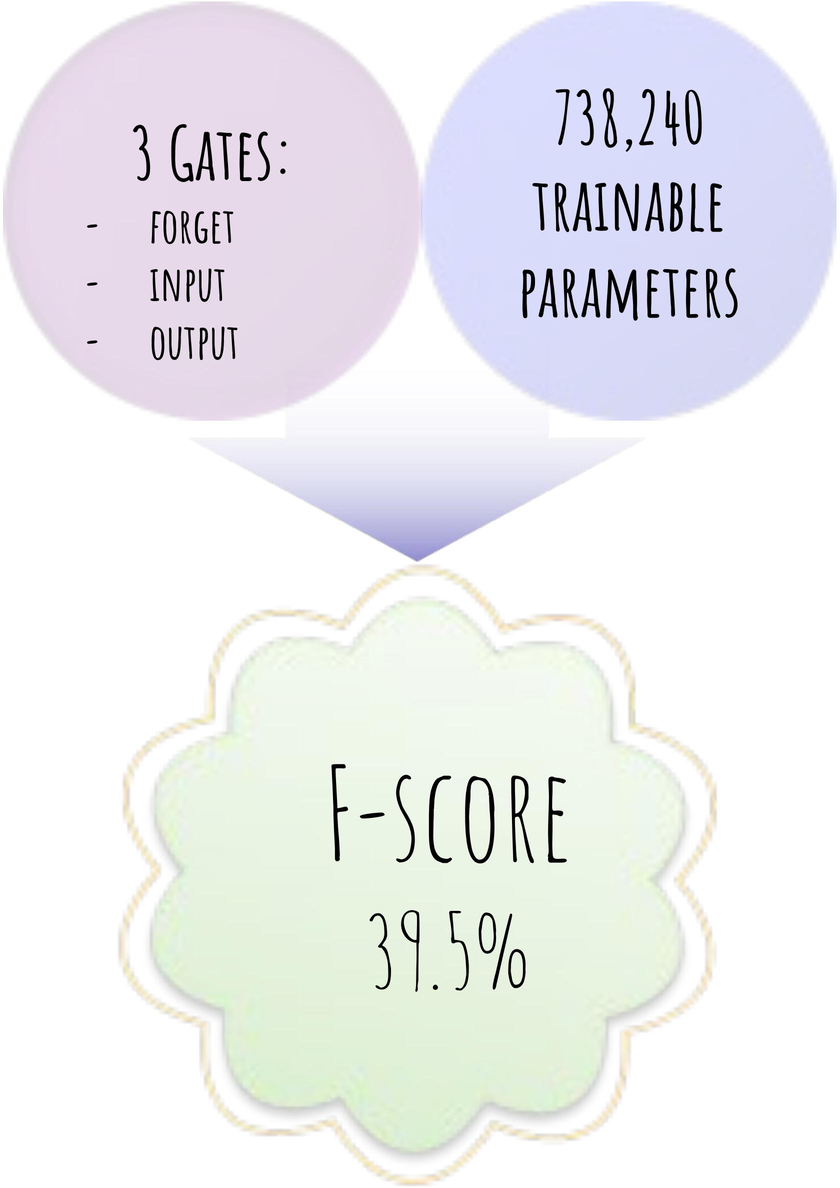
4 hidden layers

738,240 trainable parameters

F-score 41.3%

LSTM

- Trains model using back-propagation
- tanh activation function
- Even though not good general performance, results in geo and per tags
- combination with CRF or CNN



3 Gates:

- forget
- input
- output

738,240 trainable parameters

F-score 39.5%

AGENDA

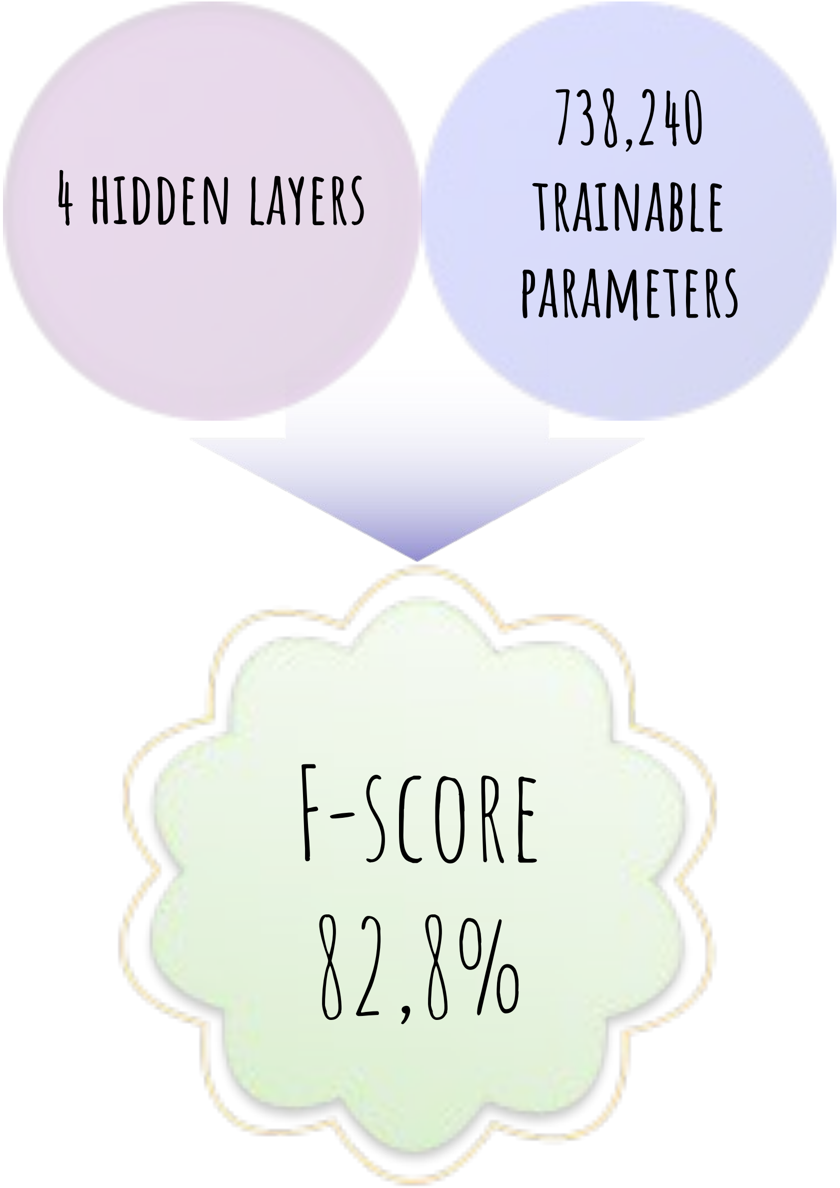
1. Motivation
2. Structure and size of the data set
3. CRF
4. RNN - LSTM
5. LSTM-CRF
6. Character embeddings
7. LSTM-CNN-CRF
8. BERT
9. SpaCy

Long term short memory CRF

- CRF built on top of a Bi-LSTM:

  CRF layer is basically an optimisation on top of Bi-LSTM layer
- Embedding layer in a vector form is passed to the CRF algorithm:

  => little effort on feature engineering



4 hidden layers

738,240 trainable parameters

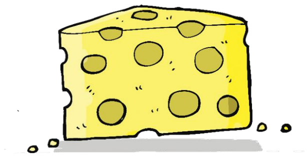
F-score 82,8%

AGENDA

1. Motivation
2. Structure and size of the data set
3. CRF
4. RNN - LSTM
5. LSTM-CRF
6. Character embeddings
7. LSTM-CNN-CRF
8. BERT
9. SpaCy

CHARACTER EMBEDDINGS

CHEESE



C	H	E	E	S	E
1	2	3	3	4	3

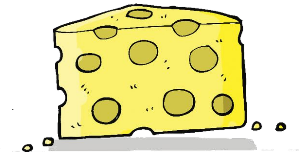
1	2	3	3	4	3

1	2	3	3	4	3	0	0	0	0

MAX SIZE OF WORD VECTOR = 10

CHARACTER EMBEDDINGS

CHEESE



C	H	E	E	S	E
1	2	3	3	4	3

1	2	3	3	4	3

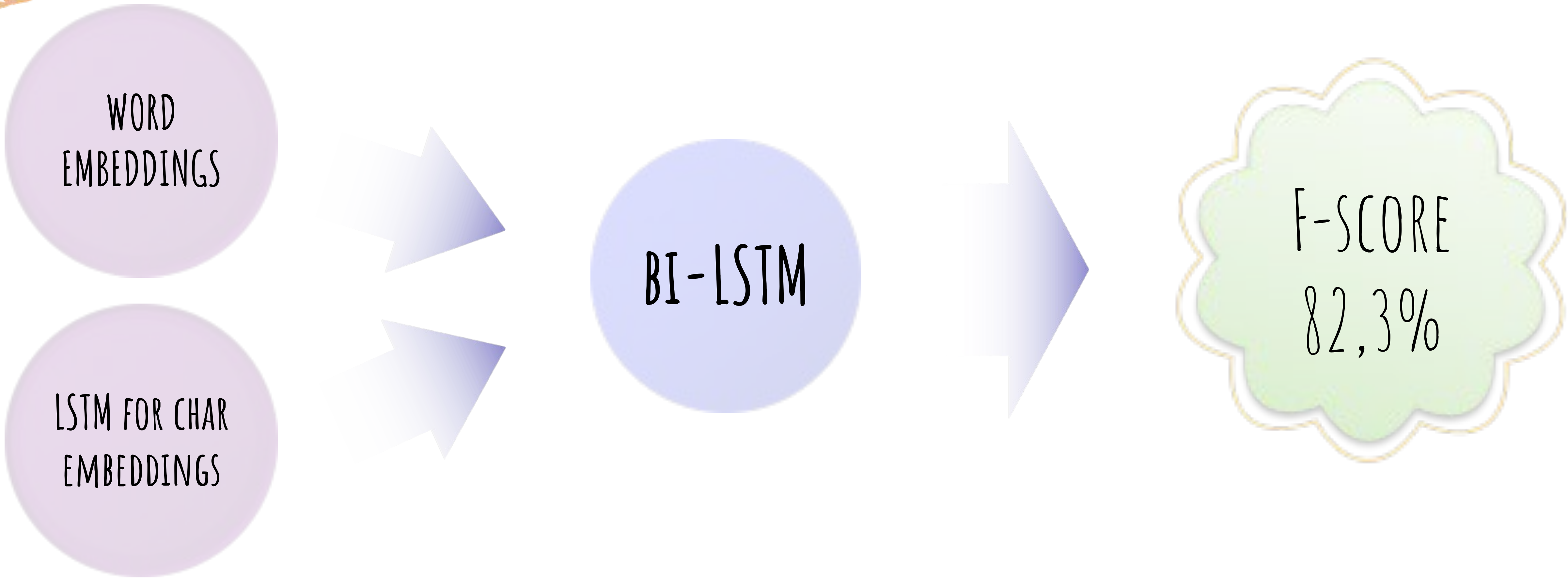
1	2	3	3	4

,

3	0	0	0	0

Max size of word vector = 5

LSTM for char embeddings



Word embeddings

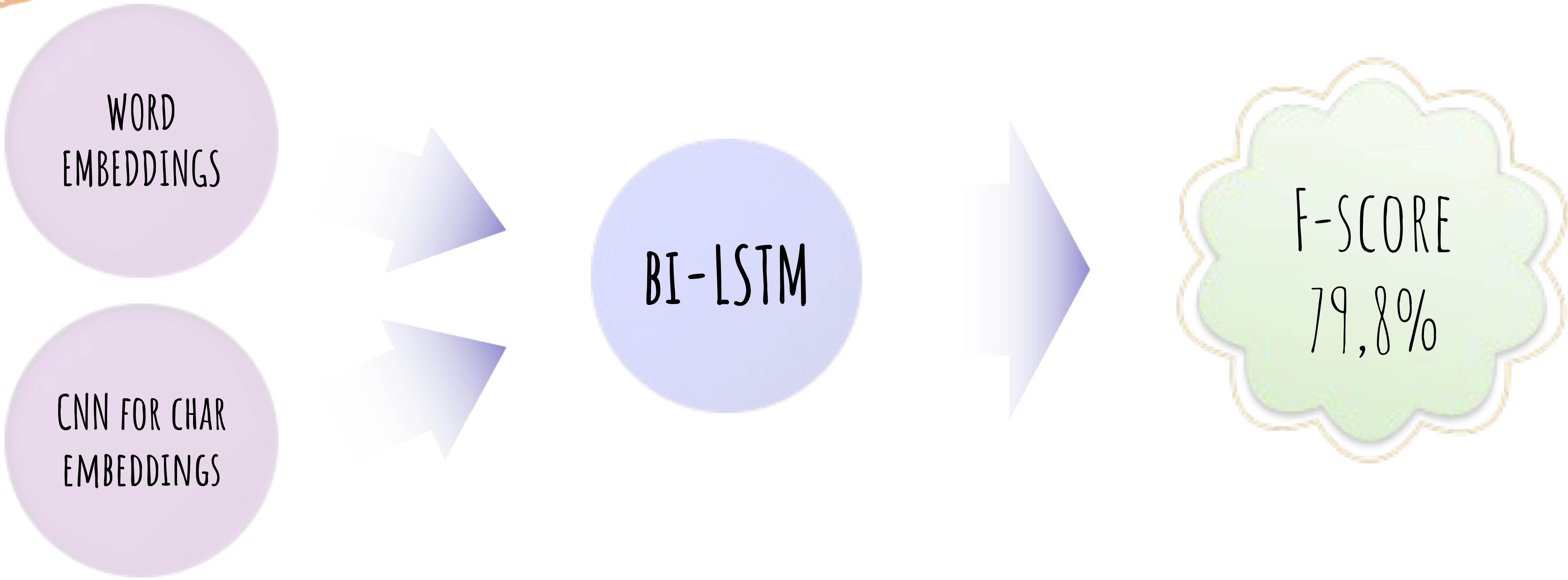
LSTM for char embeddings

BI-LSTM

F-score 82,3%

Time for one epoch: 120 seconds

CNN for char embeddings

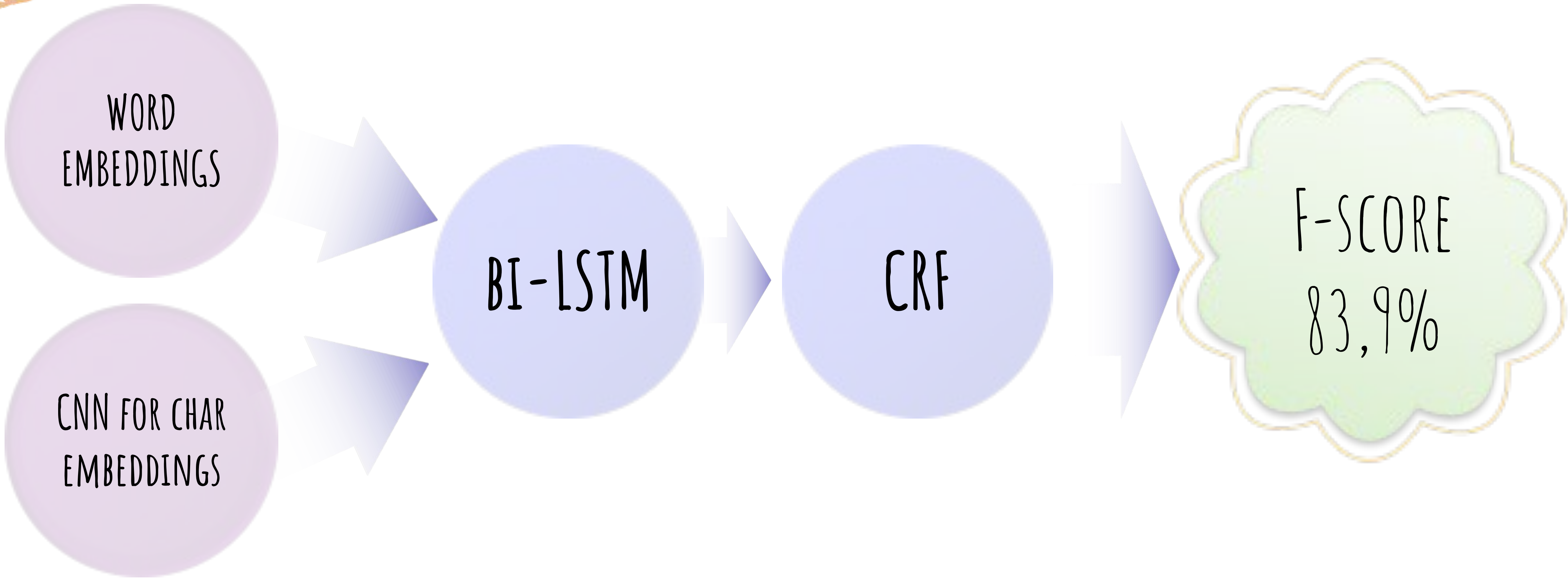


Time for one epoch: 85 seconds

AGENDA

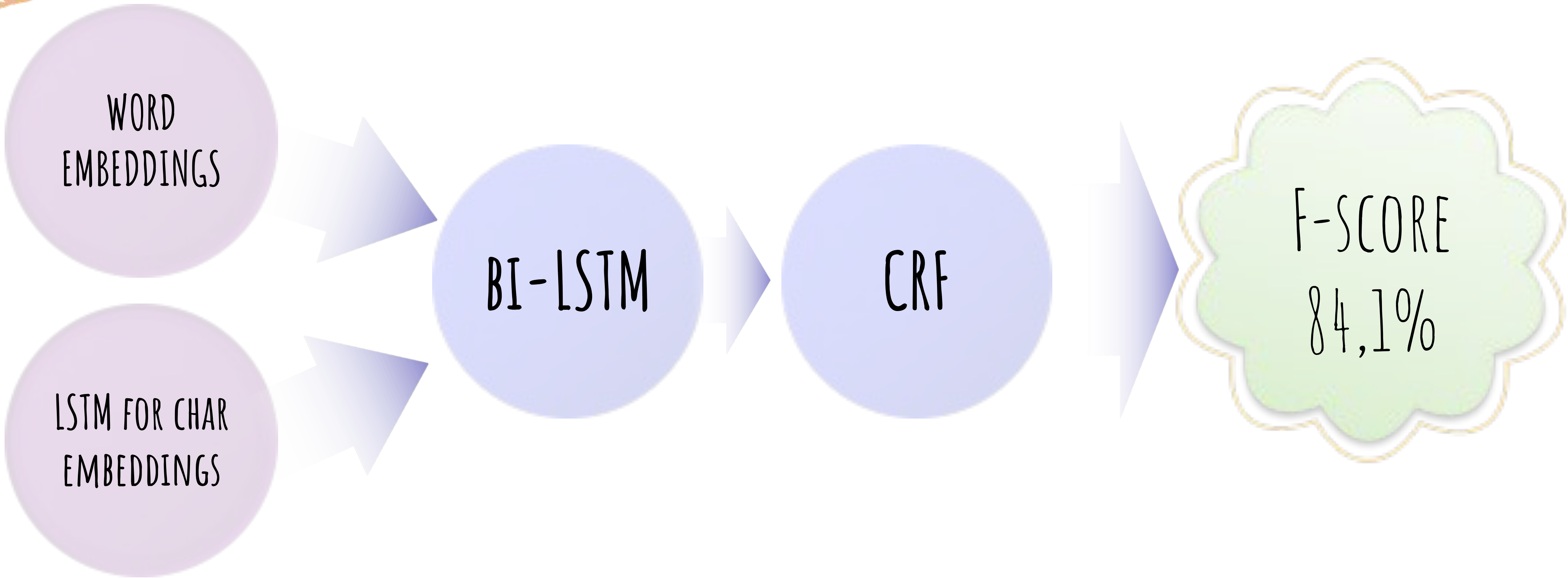
1. Motivation
2. Structure and size of the data set
3. CRF
4. RNN - LSTM
5. LSTM-CRF
6. Character embeddings
7. LSTM-CNN-CRF
8. BERT
9. SpaCy

LSTM-CNN-CRF



TIME FOR ONE EPOCH: 130 SECONDS

LSTM-LSTM-CRF



Word embeddings

LSTM for char embeddings

Bi-LSTM

CRF

F-score 84,1%

Time for one epoch: 305 seconds

AGENDA

1. Motivation
2. Structure and size of the data set
3. CRF
4. RNN - LSTM
5. LSTM-CRF
6. Character embeddings
7. LSTM-CNN-CRF
8. BERT
9. SpaCy

BERT

- Huggingface Library
- Basic Cased Model
- 4 Epochs
- AdamW Optimizer
    - ○ Learning Rate: 3e-5
    - ○ Epsilon 1e-8
- In the original Paper: F1 Score ~ 92%

Fine-Tuning AdamW

#Words	Accuracy	F1
10000	0.87167	0.35335
100000	0.94249	0.7303
All	0.96268	0.83785

No Fine-Tuning AdamW

#Words	Accuracy	F1
10000	0.08524	0.0254
100000	0.80744	0.00514
All	0.84254	0.27026

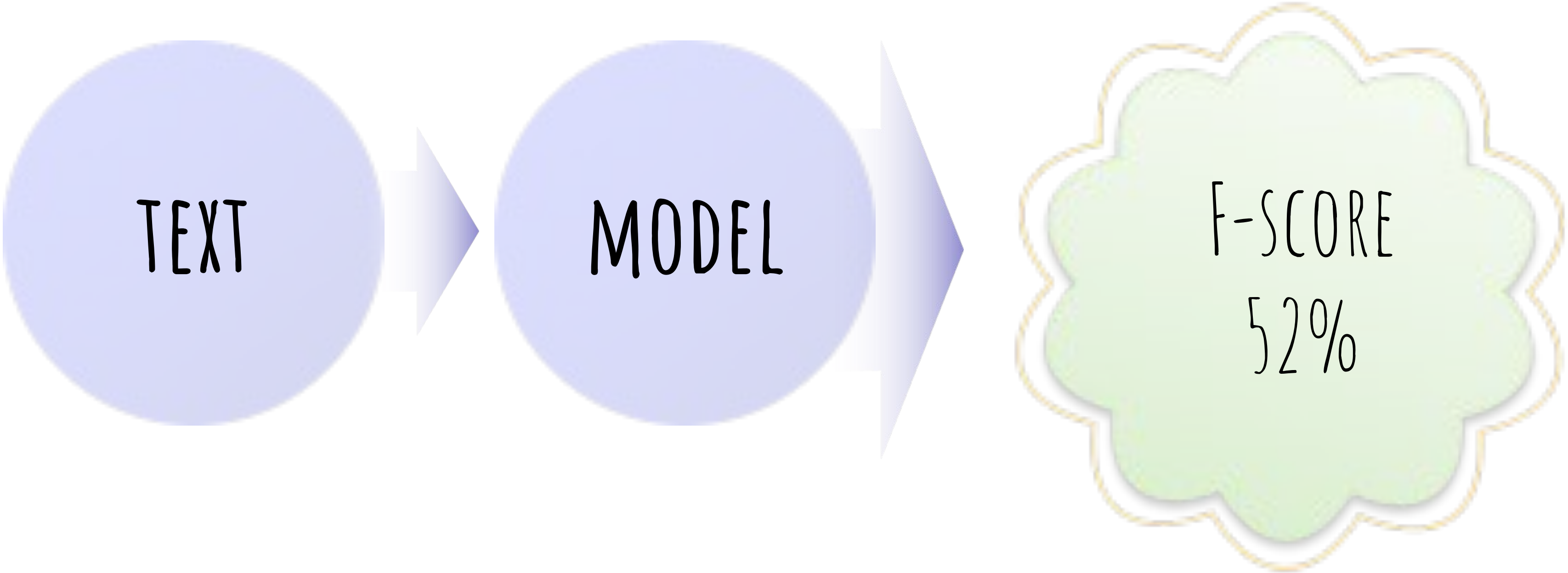
Max

F-score

85.44%

PRETRAINED BERT-NER

- Input -> Text
- Benchmark
- HUGGINGFACE LIBRARY
- BASELINE



TEXT

MODEL

F-SCORE

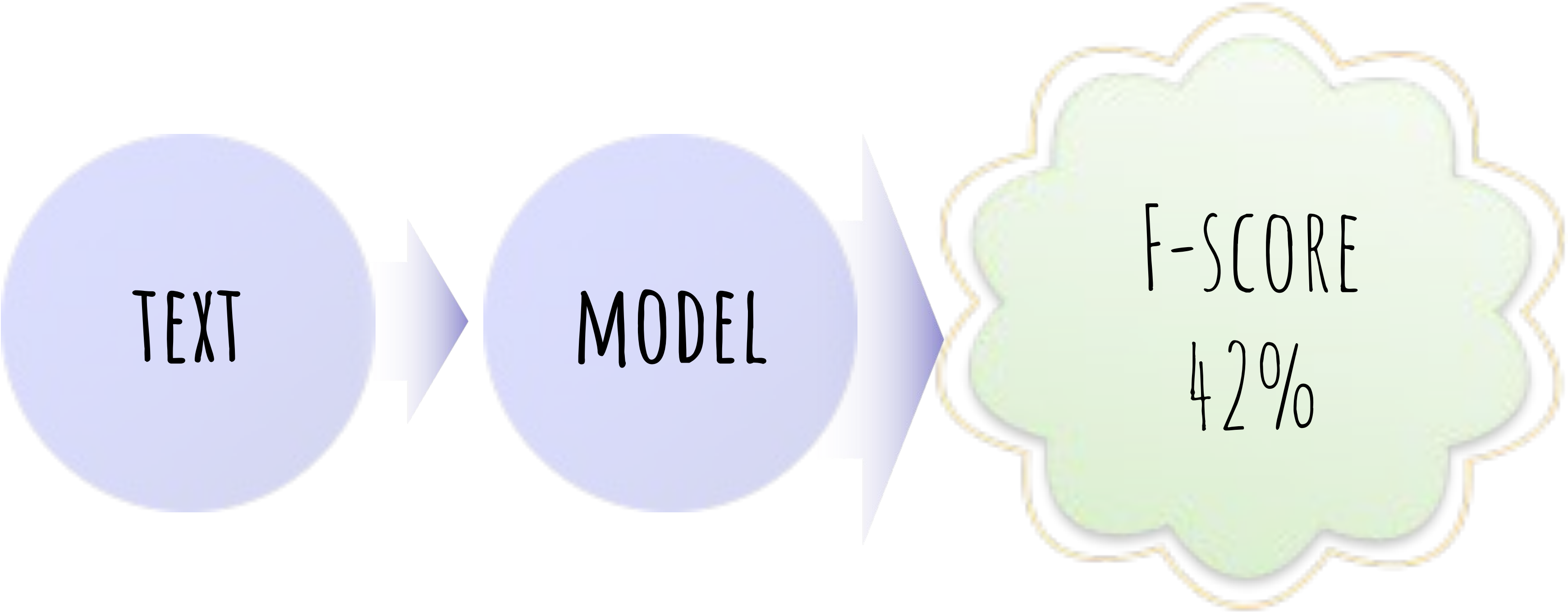
52%

AGENDA

1. Motivation
2. Structure and size of the data set
3. CRF
4. RNN - LSTM
5. LSTM-CRF
6. Character embeddings
7. LSTM-CNN-CRF
8. BERT
9. SpaCy

SPACY

- INPUT -> TEXT
- BENCHMARK
- NLP LIBRARY
- BASELINE



TEXT

MODEL

F-SCORE
42%

Thank you for attention!

Do you have any questions?