Natural Language Processing

Lecture 21

Dirk Hovy

dirk.hovy@unibocconi.it

@dirk_hovy

Bocconi

Goals for Today

- Learn about **recurrent neural network architectures**
- Understand the difference to convolutional networks
- Learn about **different architecture**
- Understand the **attention mechanism**

Recurring Matters

Machine Translation

Text-to-Speech

Image Captions

Bocconi

Long-Term Trouble

SUBJECT

"Wenn er aber auf der Strasse der in Sammt und Seide gehüllten jetzt sehr ungenirt nach der neusten Mode gekleideten Regierungsräthin begegnet."

VERB

Mark Twain, *The Awful German Language*

Long-Term Trouble

NEGATION

This is not in any sense of the word a funny movie.

Sequence Tagging

PRON	VERB	ADP	DET	???	PUNCT
I	went	to	the	show	.

show {VERB, NOUN}

PART **show**

show

PRON **show**

show

show

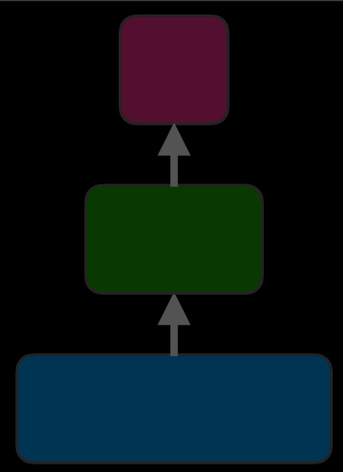
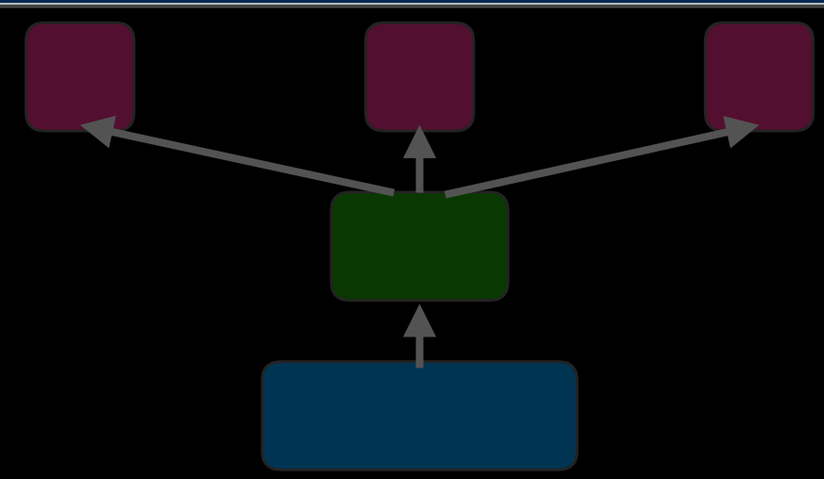
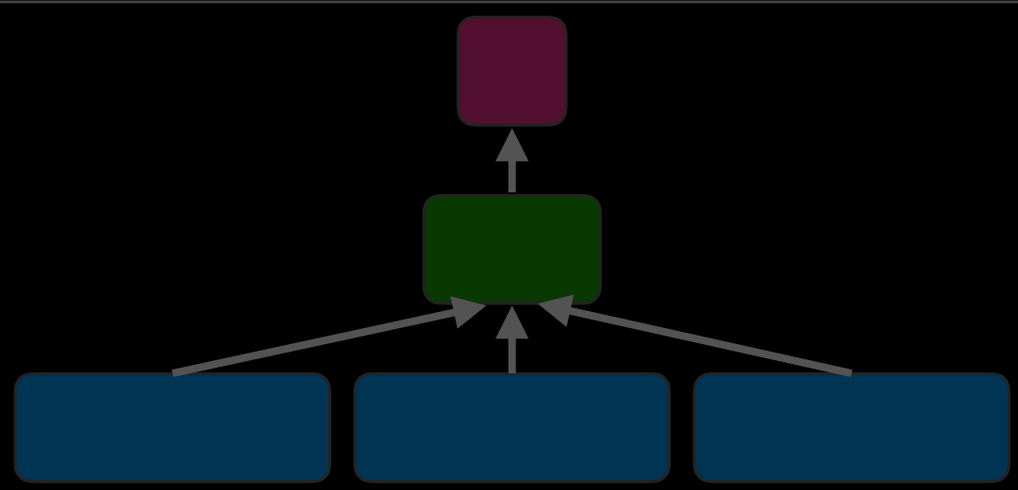
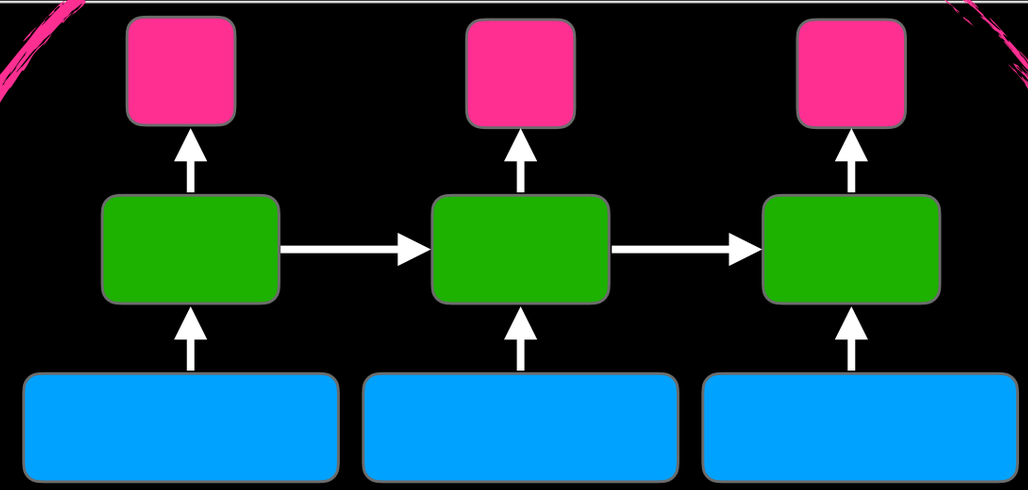
DET **show**

show

ADJ **show**

Structured prediction: depends on the POS of a previous word

Types of Text Classification

	Fixed length	Variable length
Fixed length	 Logistic Regression, Perceptron, Feed-Forward Network, Random Forest, Naive Bayes, SVM, …	 Multitask Learning, Decoder
Variable length	 Convolutional Neural Networks (CNN)	 Recurrent Neural Networks (RNN), Hidden Markov Models (HMM), Conditional Random Fields

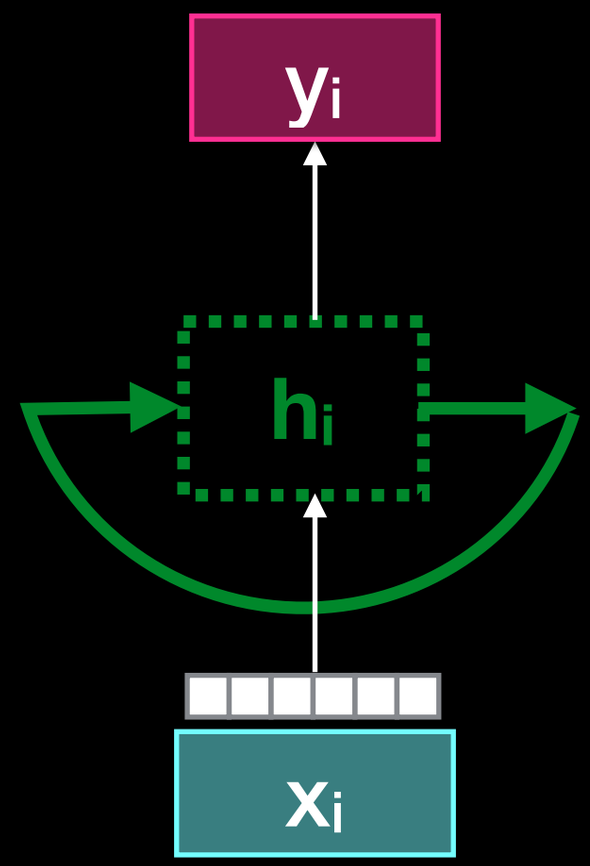
Recurrent Networks

Bocconi

Recurrence

$$y_i = f(h_i)$$

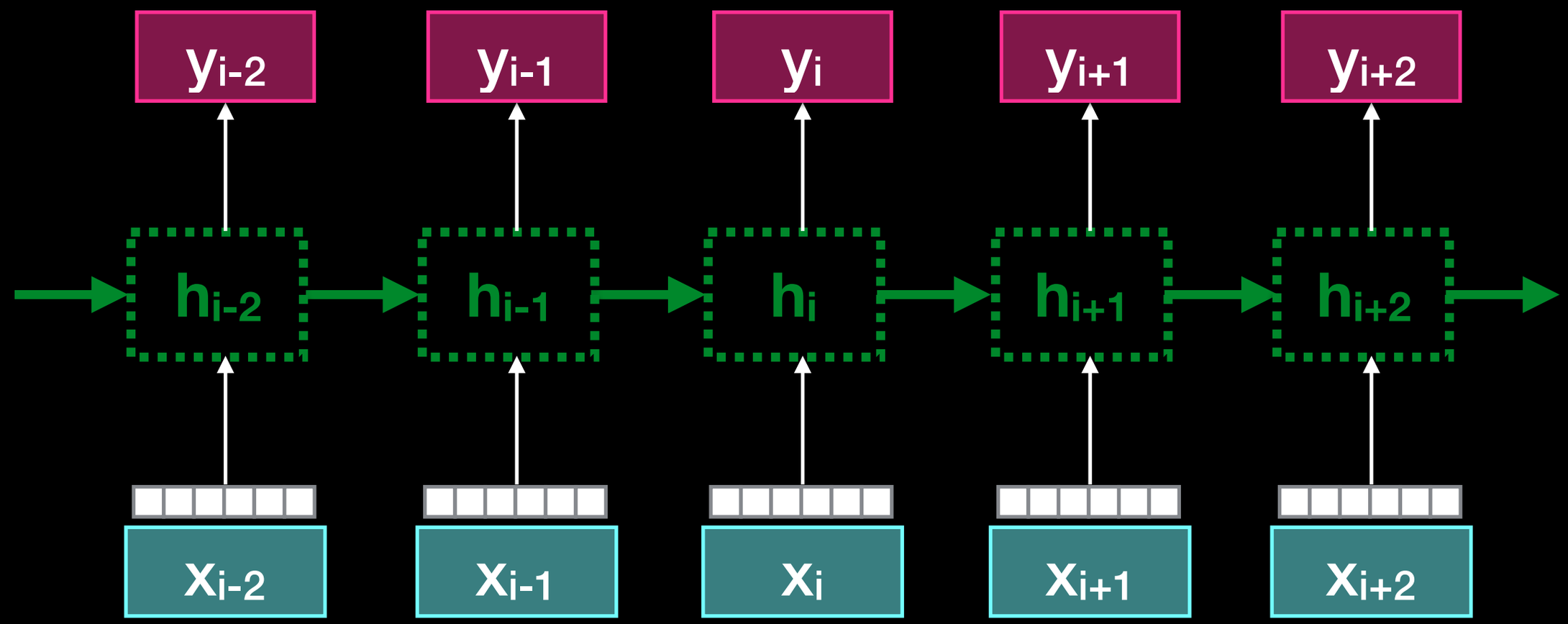
$$h_i = s(h_{i-1}, x_i)$$



...Unrolled

$$y_i = f(h_i)$$

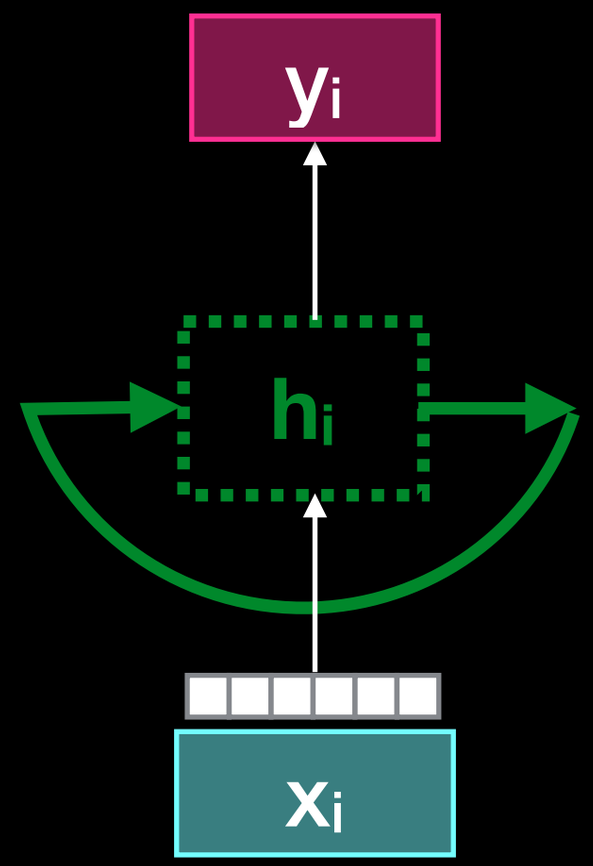
$$h_i = s(h_{i-1}, x_i)$$



Concretely

$$y_i = f(h_i) \;\; = h_i$$

$$h_i = s(h_{i-1}, x_i) = tanh(W_1 h_{i-1} + W_2 x_i + b)$$



SIMPLE RNN

Recap: LMs

TRIGRAM MODEL

$$P(w_1, w_2, \ldots, w_n) \approx \prod_{i=1}^{N} P(w_i | w_{i-2}, w_{i-1})$$

*** * The weather today is fine STOP**

$$P(S) = P(w_1, \ldots, w_n) = P(\texttt{The} | \texttt{* *})$$

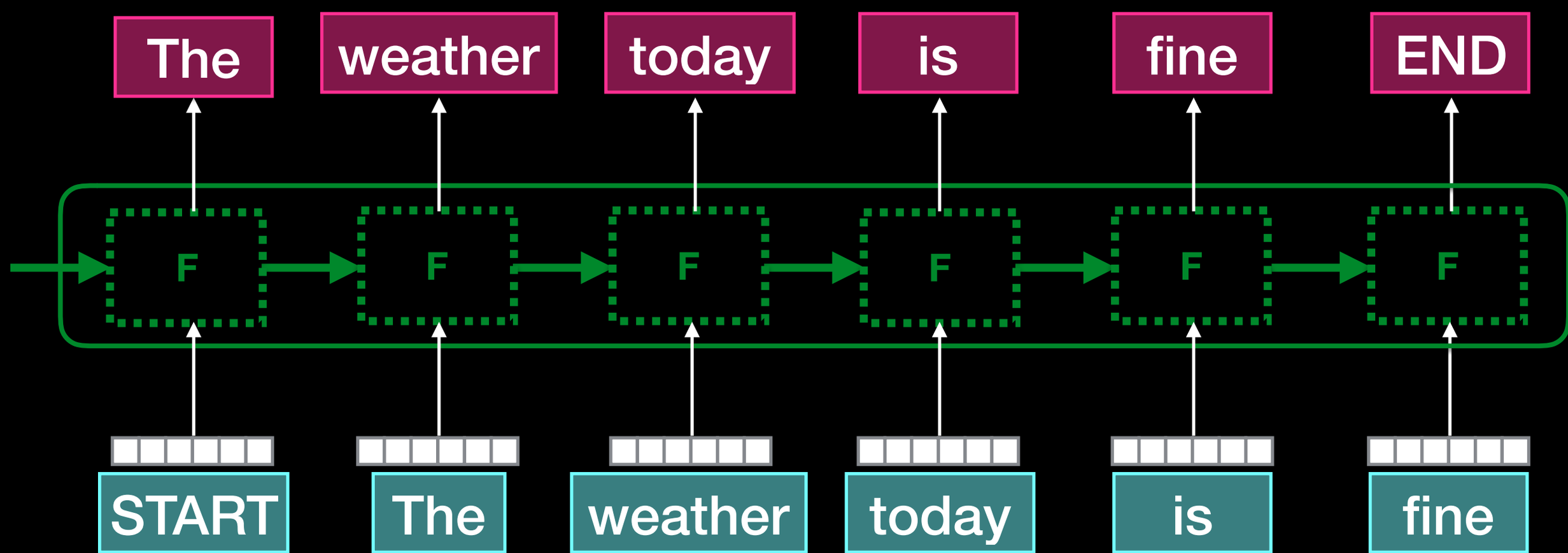
CHAIN RULE

$$\times P(\texttt{weather} | \texttt{* The})$$
$$\times P(\texttt{today} | \texttt{The weather})$$
$$\times P(\texttt{is} | \texttt{weather today})$$
$$\times P(\texttt{fine} | \texttt{today is})$$
$$\times P(\texttt{STOP} | \texttt{is fine})$$

Neural LMs

INIFINTE MODEL

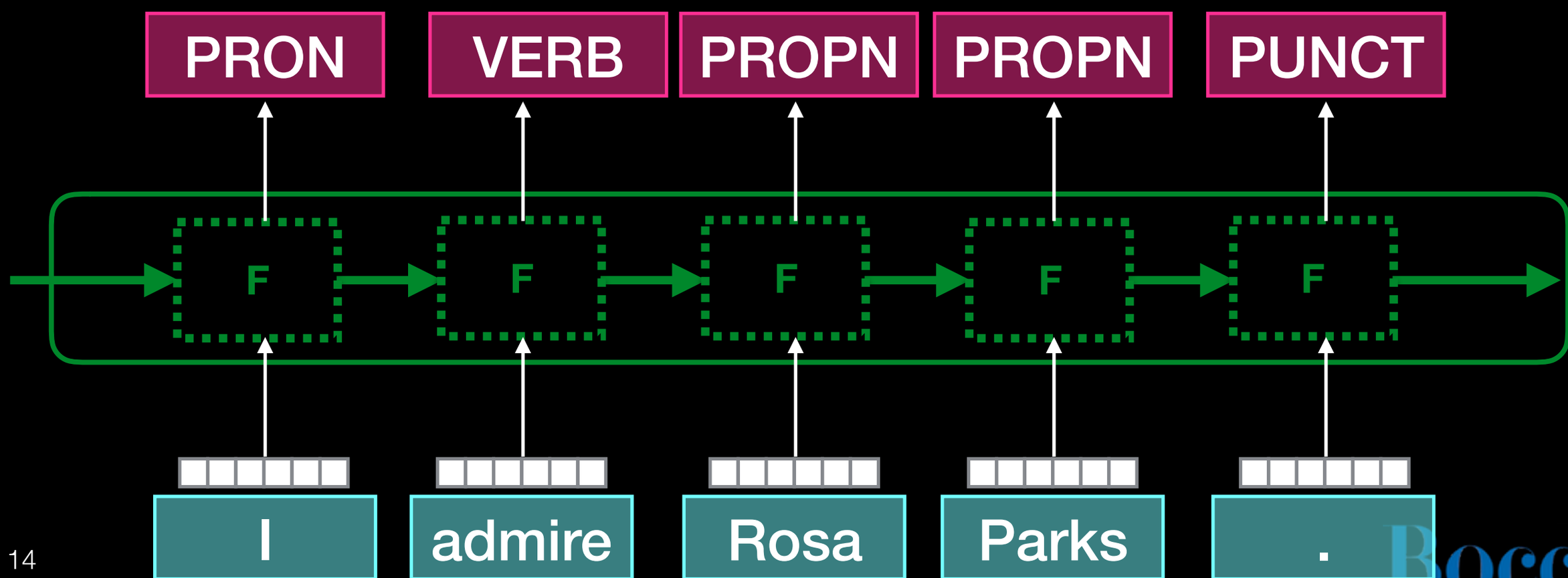
$$P(w_1, w_2, \ldots, w_n) \approx \prod_{i=1}^{N} P(w_i | w_1, w_{i-1})$$



PREDICT NEXT WORD GIVEN HISTORY

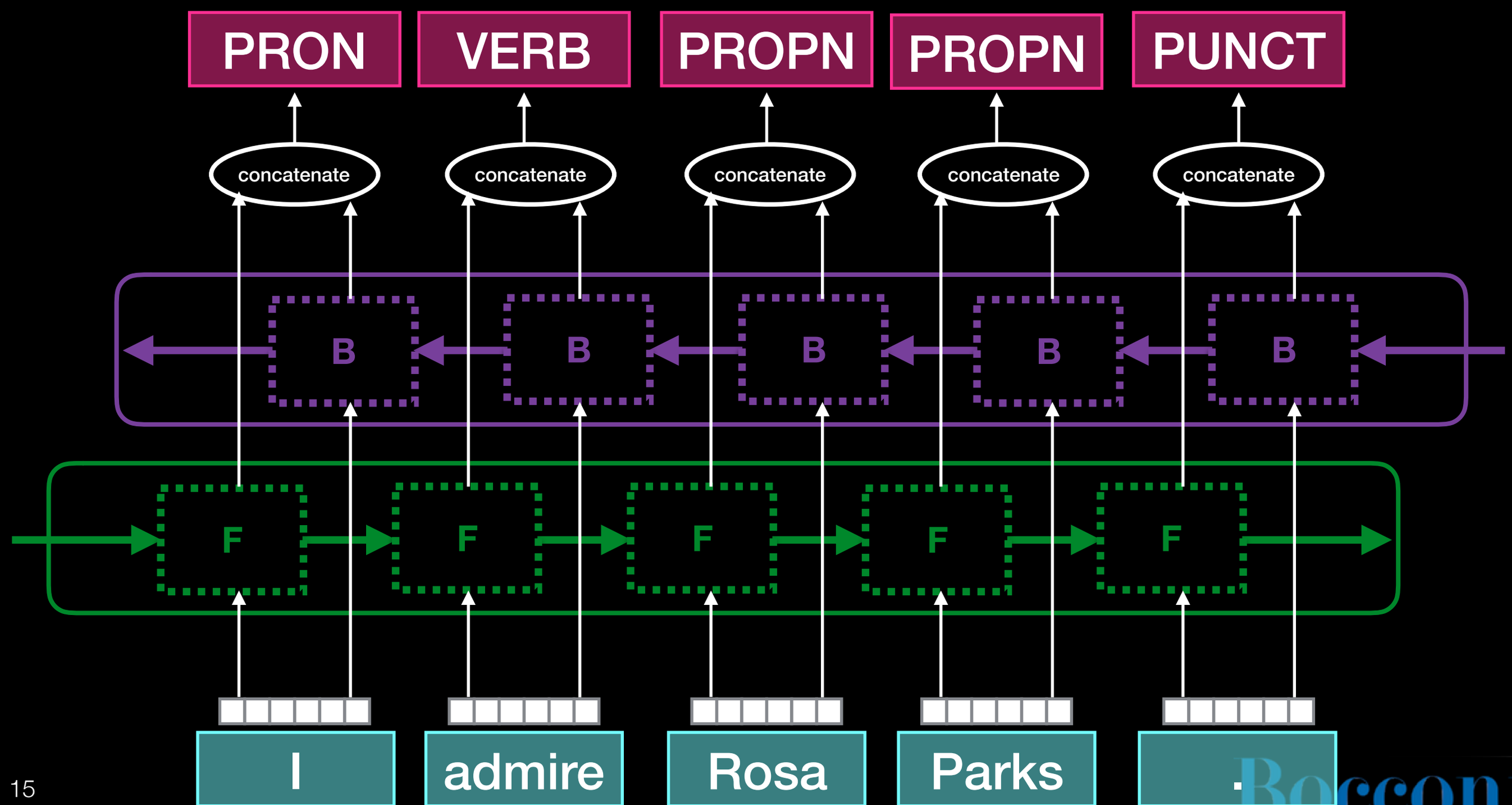
RNN Tagging

Structured Prediction



Bidirectional-RNN

Structured Prediction

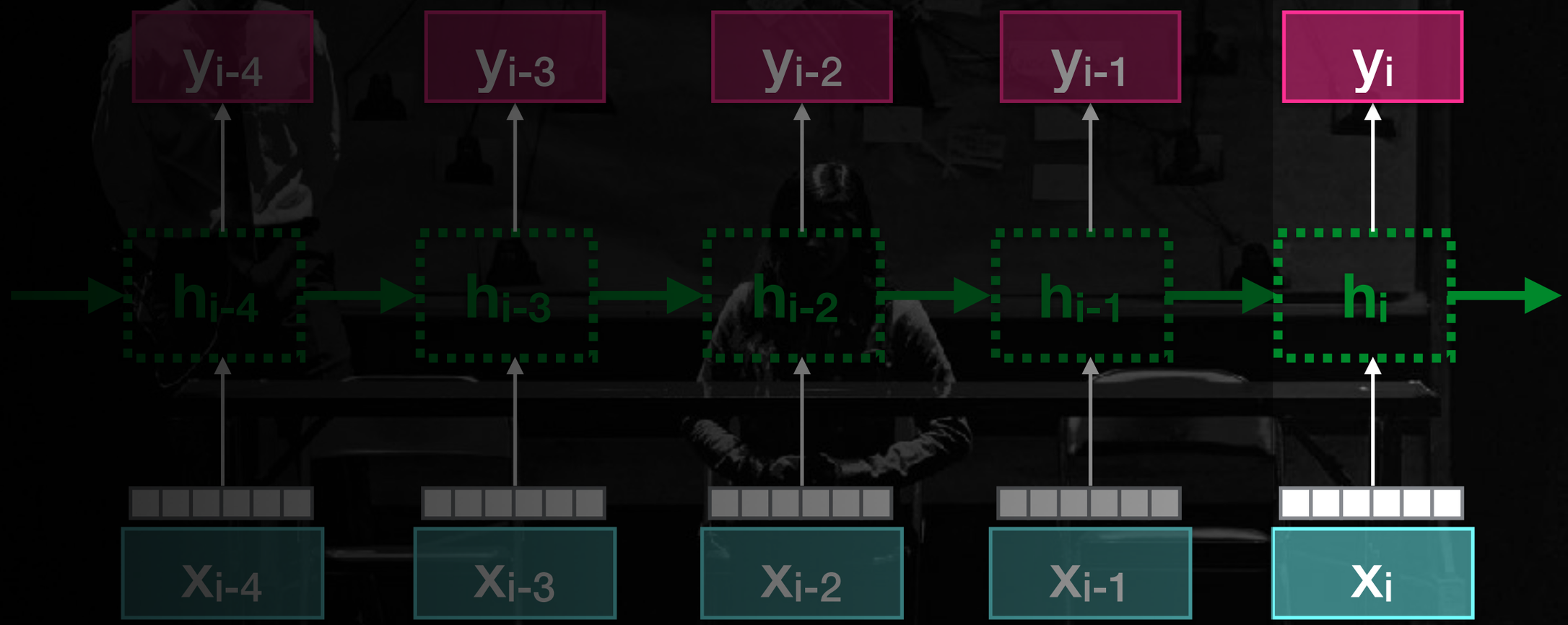


Special Recurrent Networks

Bocconi

Vanishing Memory

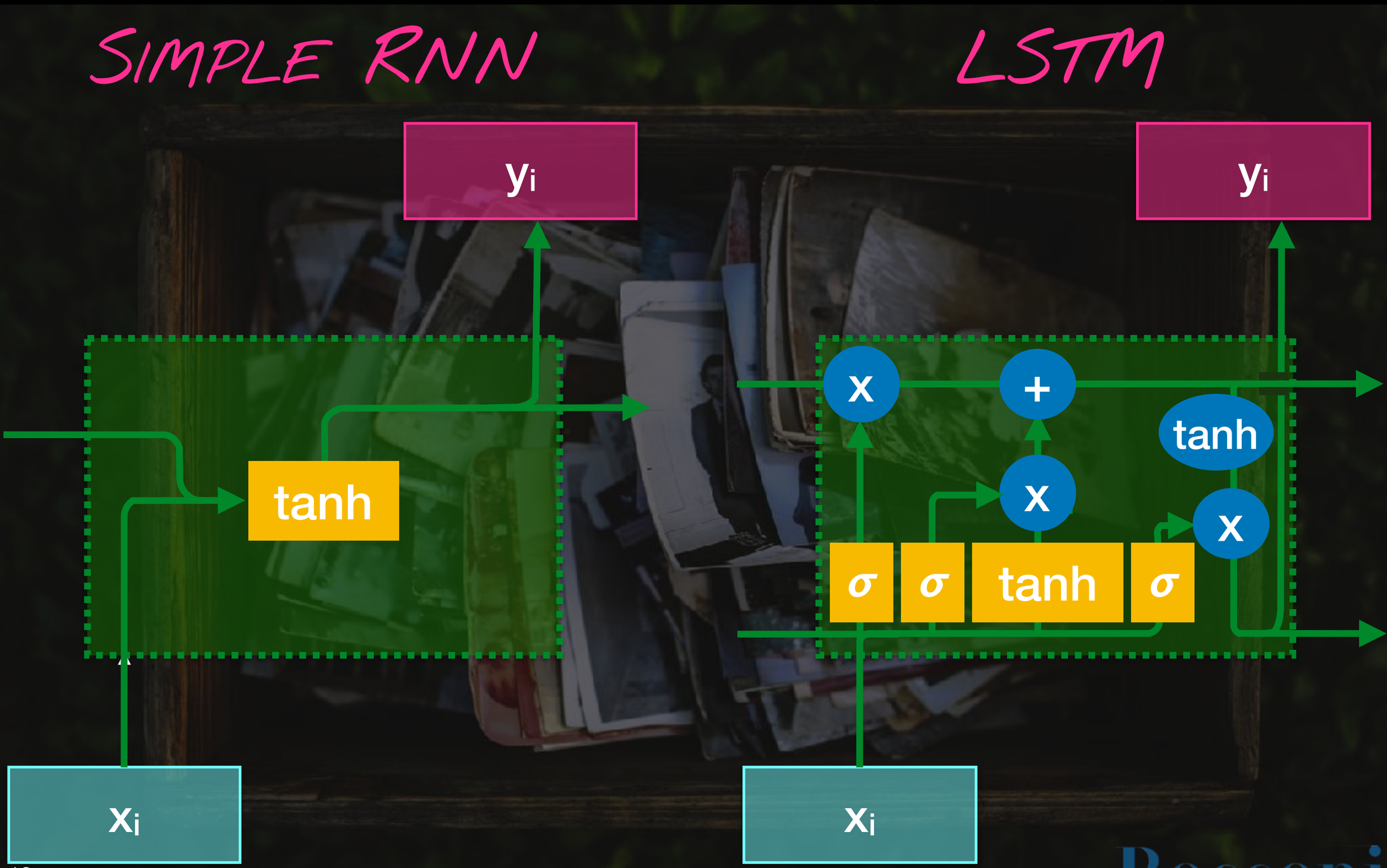
WHERE WERE YOU MARCH 3, 2016?



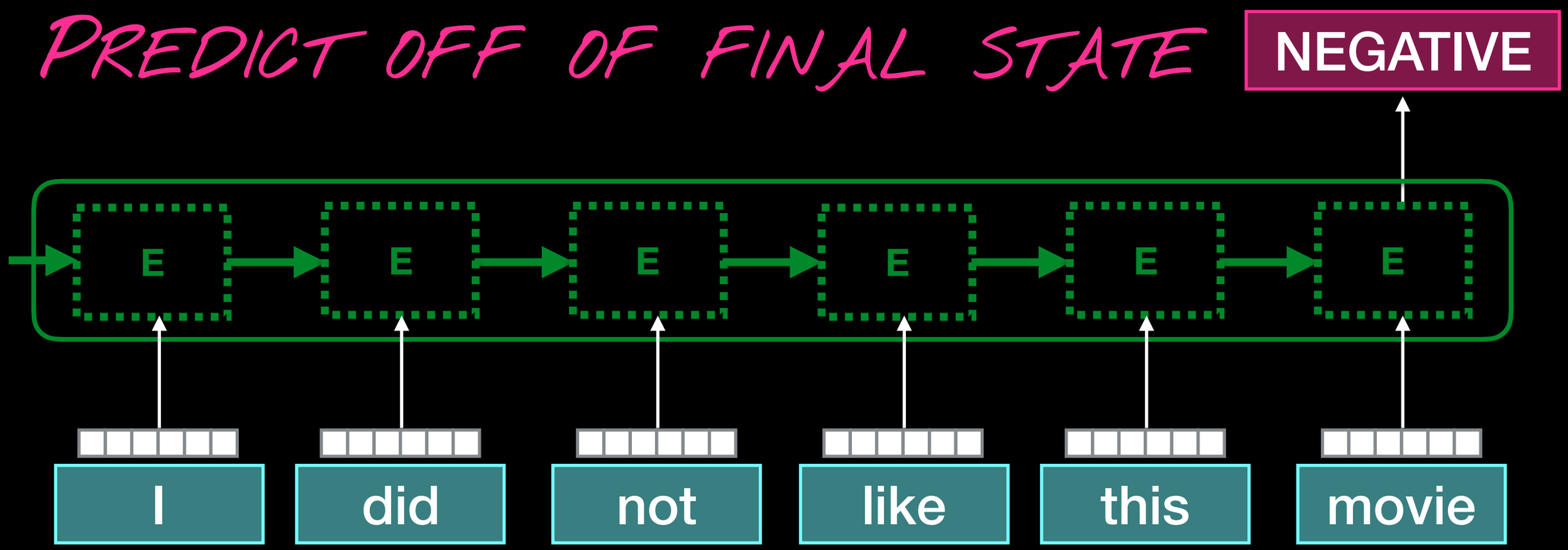
PROBLEM WITH LONG SEQUENCES

Bocconi

Selective Forgetting

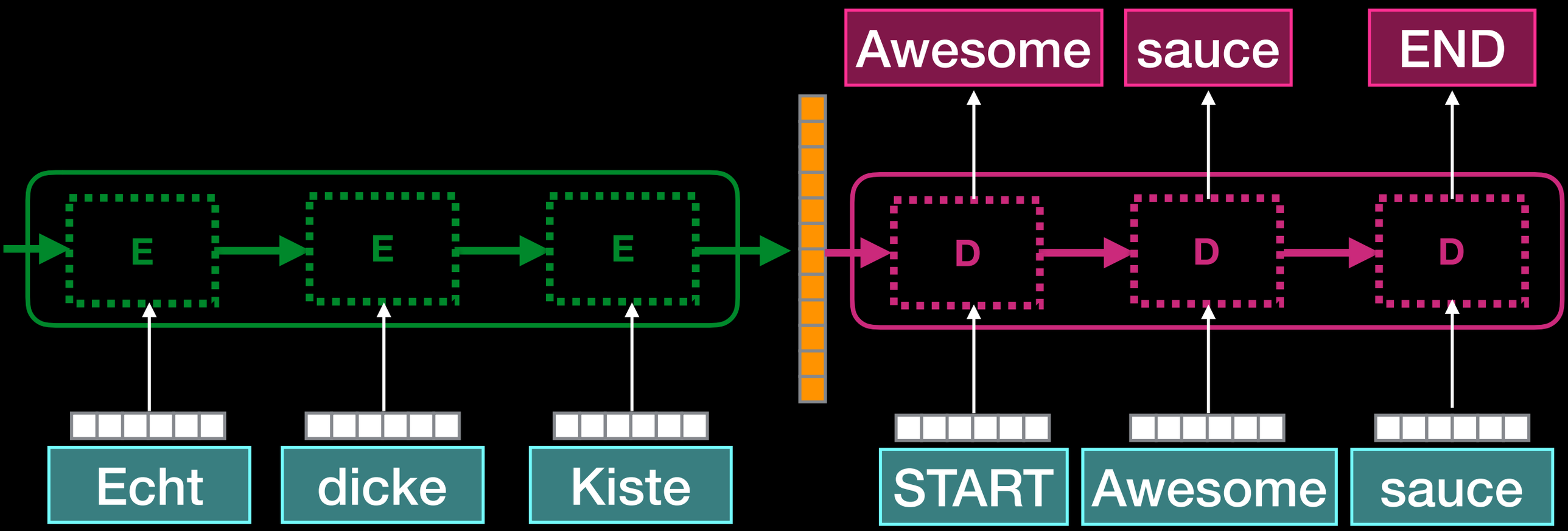


Acceptor


PREDICT OFF OF FINAL STATE
NEGATIVE
E
E
E
E
E
E
I
did
not
like
this
movie

Encoder-Decoder

...and generate output from it



Gobble up sequence into a vector...

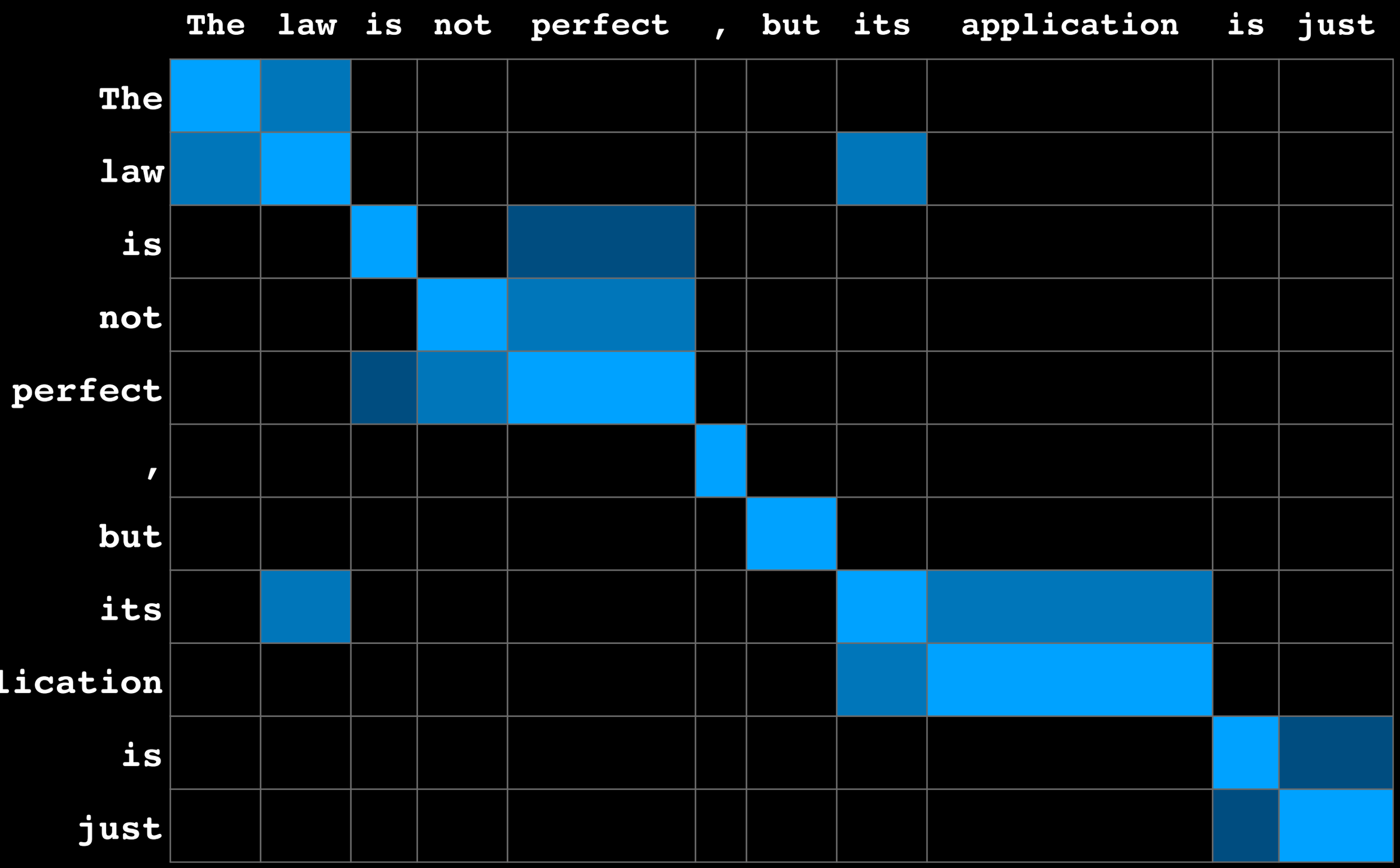
The Attention Mechanism

Bocconi

Attention!

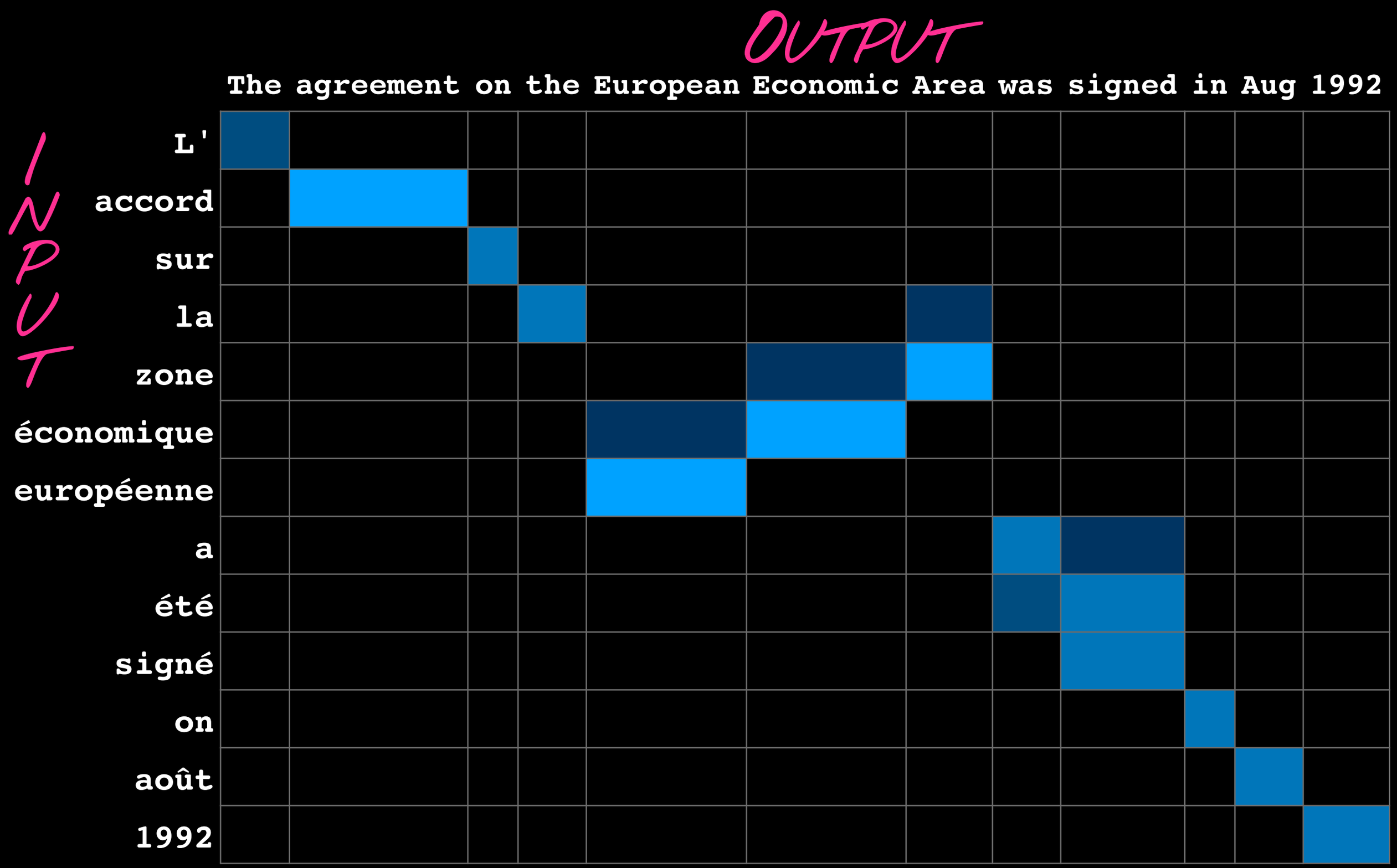
- Learn syntactic and semantic relations between words in
  - the input and output (RNNs)
  - only the input (CNNs)
- Good for machine translation (word alignment) and classification (complex expressions)

CNN with Attention



FIND LONG-RANGE DEPENDENCIES

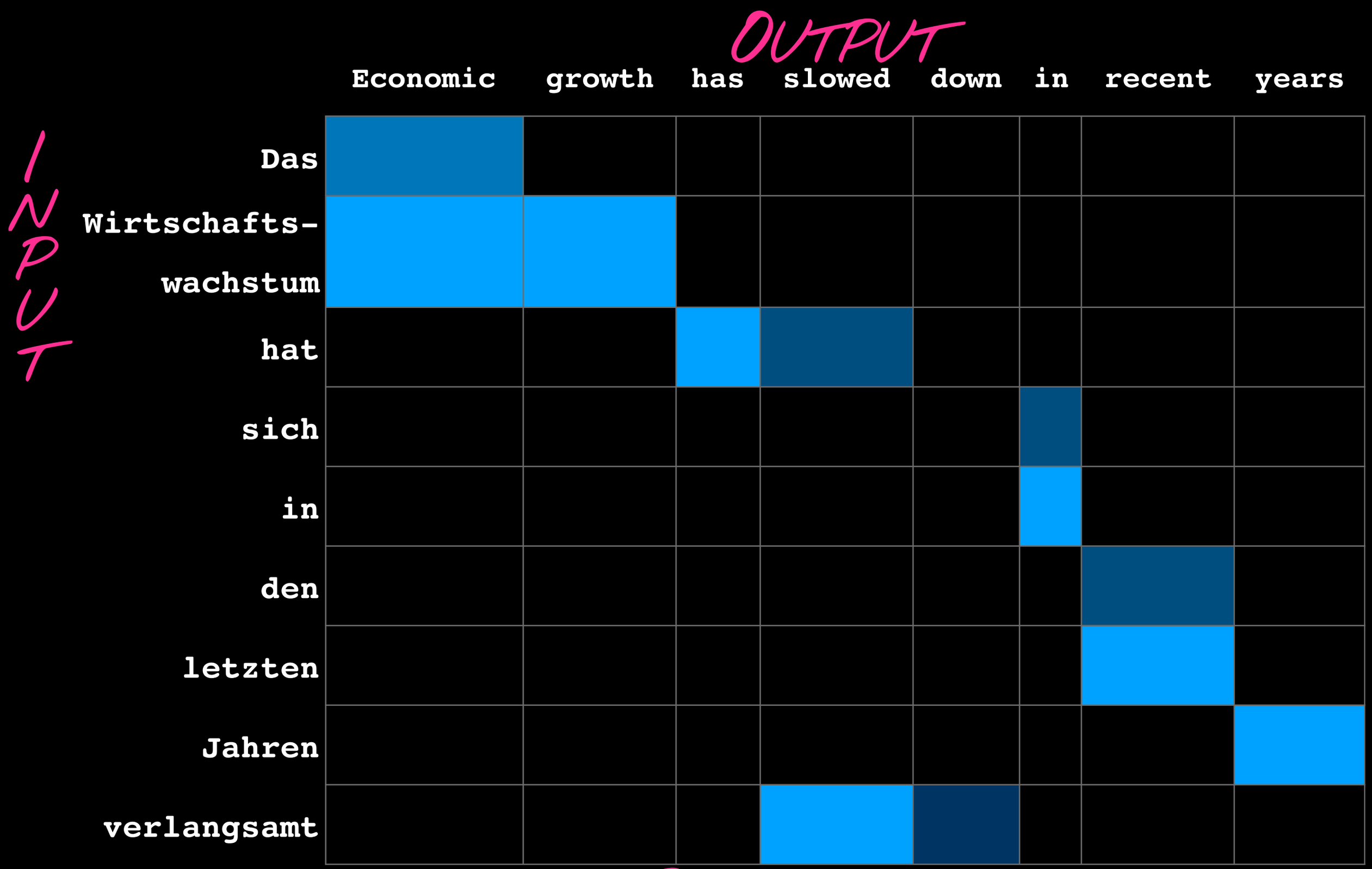
RNN with Attention



LEARN REORDERING

Bocconi

RNN with Attention

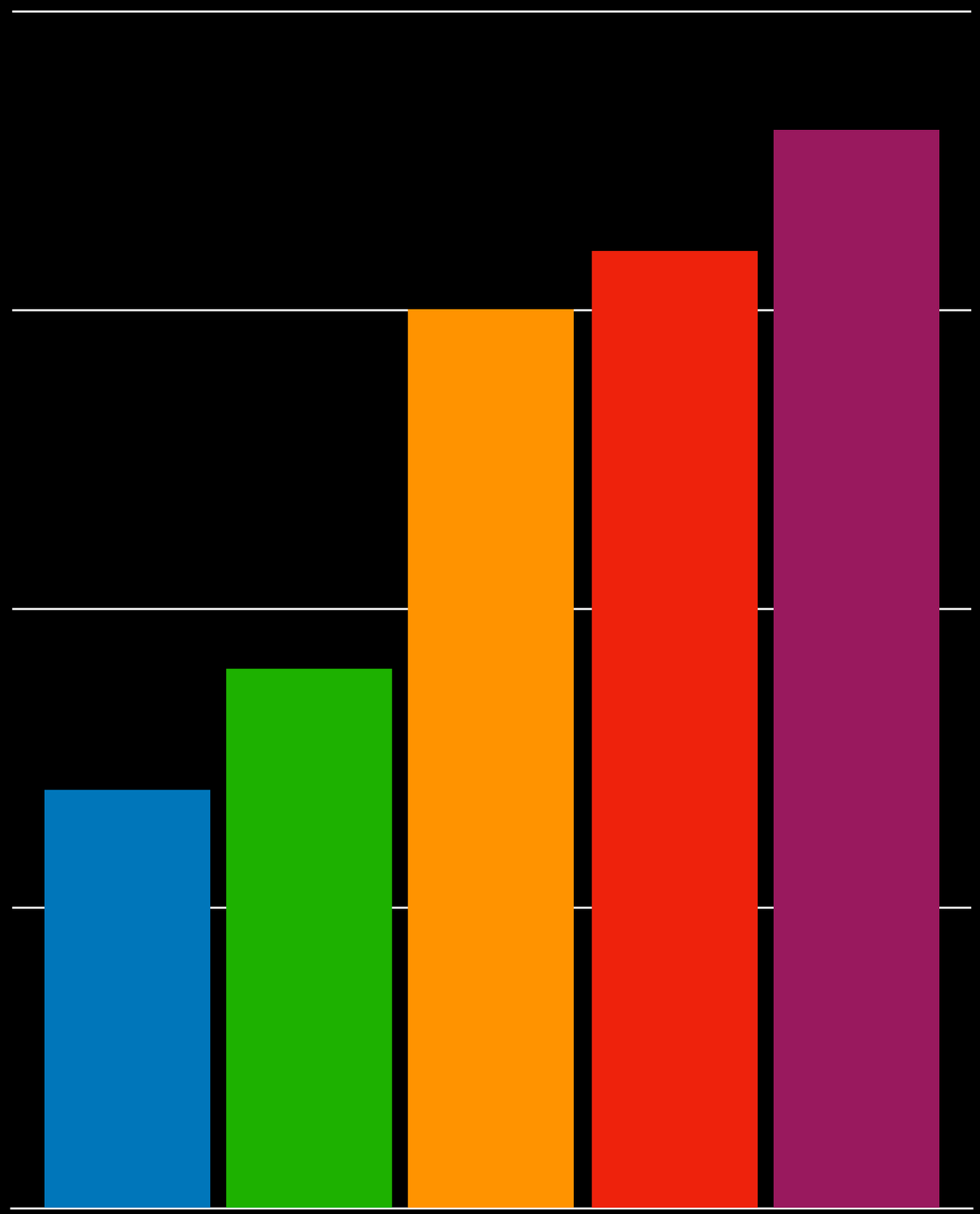


LEARN REORDERING

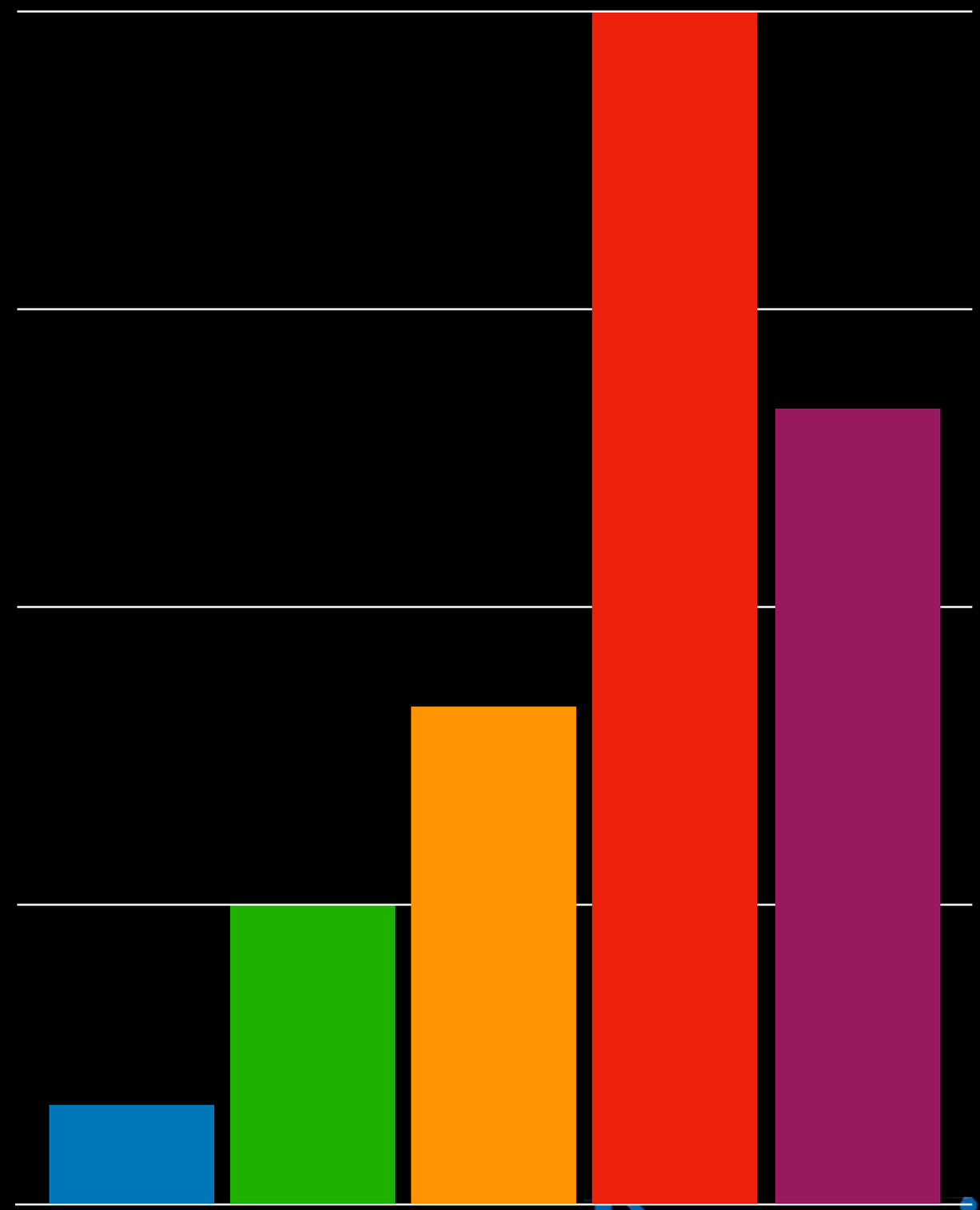
Bocconi

Performance

Logistic Regression | Feed-Forward | CNN | RNN | CNN+Attention



performance



parameters

Wrapping up

Bocconi

Take Home Points

- **Recurrent Neural Nets** address long-range dependencies
- Condition each word on all previous ones:
  - better **LMs**
  - better **sequence labels**
- **Bidirectional RNNs** condition on following words
- **LSTMs** learn to forget useless input
- **Attention** improves coherence

Bocconi