Predicting the Gender of Indonesian Names

Ali Akbar Septiandri

PYCON
INDONESIA 2018

About me

Part-time lecturer at **Universitas Al Azhar Indonesia**
Data scientist at **Airy**

@aliakbars on Twitter, Instagram, & Github

QUIZ!

*Can you guess what is the gender of a person named **Putri**?*

What about ***Joko****?*

Disclaimer

Not a political campaign

Bowo?

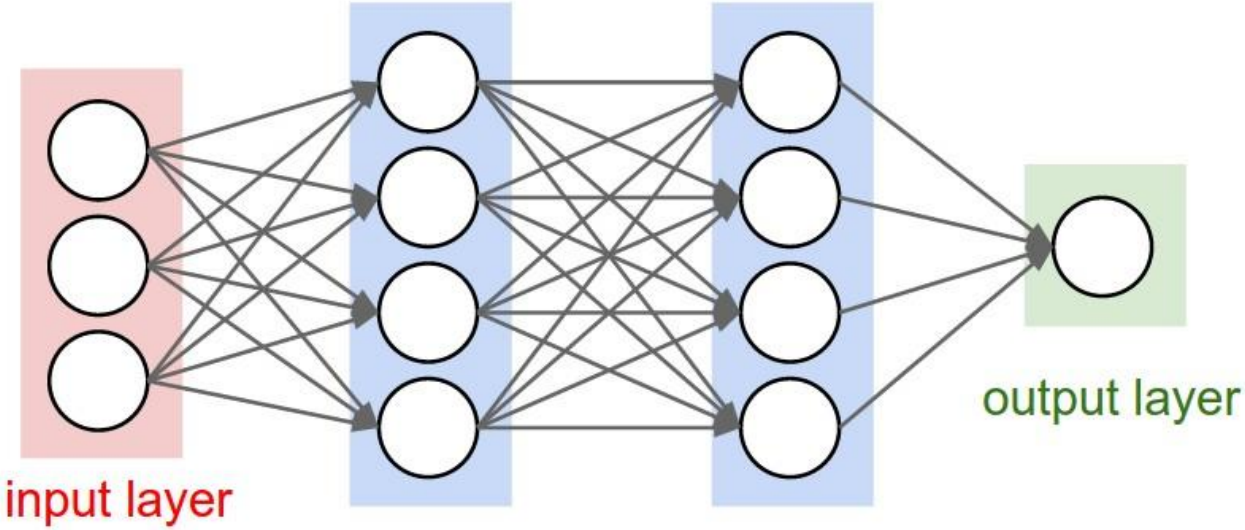
“ Annisa?

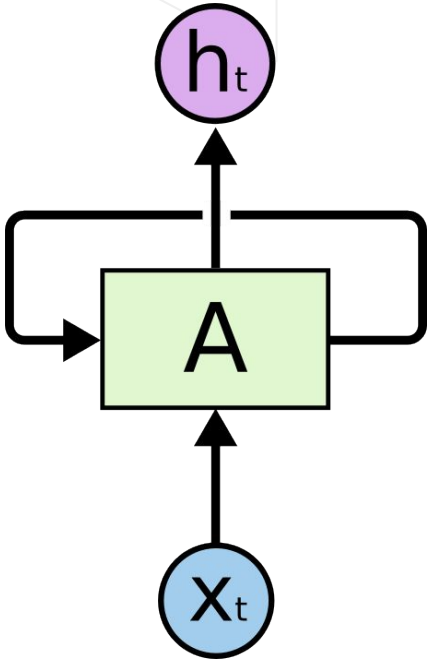
Dwi? Tri?

Can machines classify them?

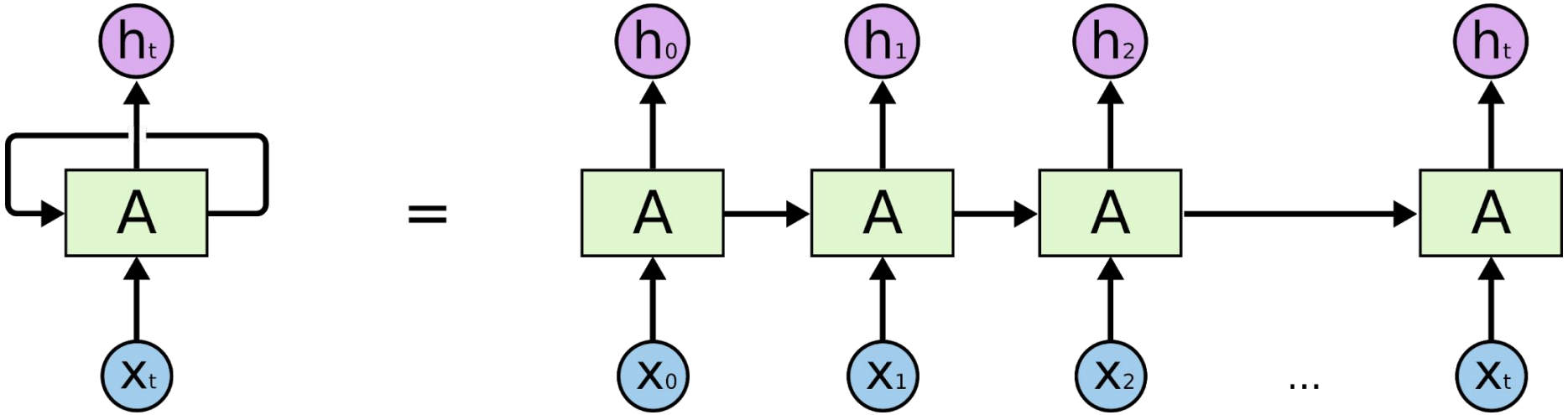
Of course...

You might see the following figures several times throughout this conference...

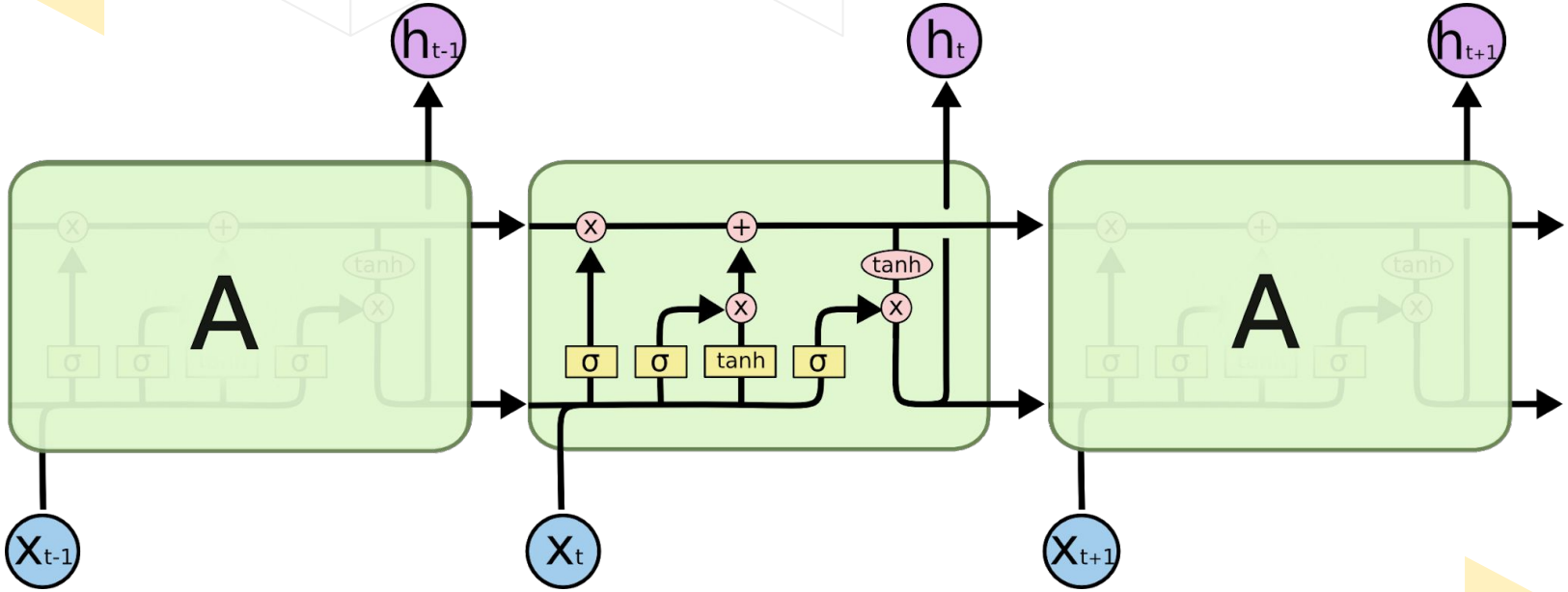

input layer
hidden layer 1
hidden layer 2
output layer



Source:
http://colah.github.io/posts/2015-08-Understanding-LSTMs/



Source:
http://colah.github.io/posts/2015-08-Understanding-LSTMs/



Source:
http://colah.github.io/posts/2015-08-Understanding-LSTMs/

We'll get back to that later...

Let's try a simpler approach!

Bag of Words

D1: “send us your password”
D2: “send us your review”
D3: “review us”

send	us	your	review	password
1	1	1	0	1
1	1	1	1	0
0	1	0	1	0

~~Bag of Words~~

Bag of Characters

We can do better!

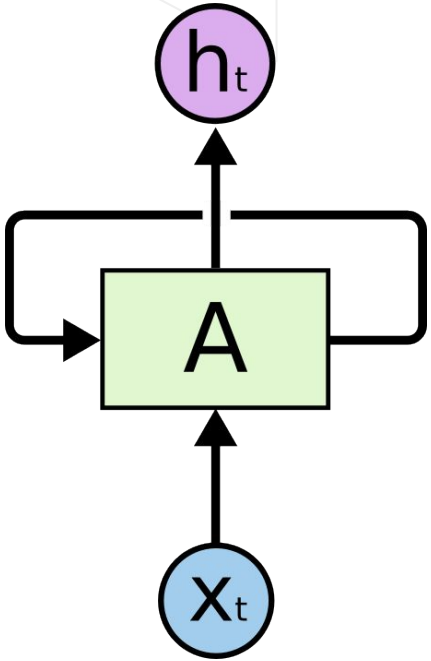
- Bag of characters will **only list A-Z**
- What if we want to get **subsequent characters**?
- **n-grams**

n-grams

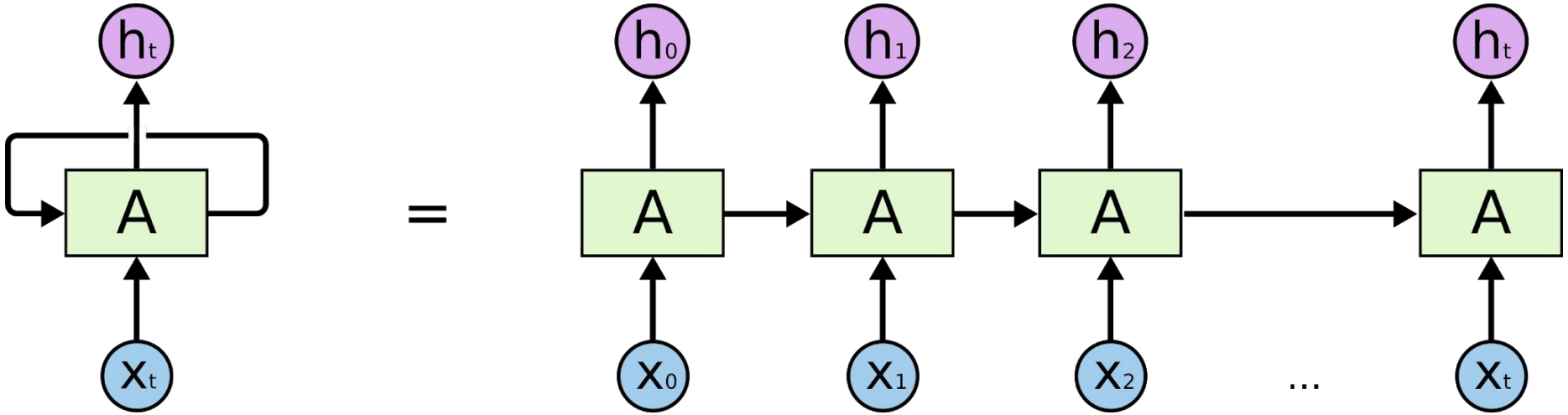
- Given “akbar”
- 1-gram: {a, b, k, r}
- 2-gram: {ak, kb, ba, ar}
- 3-gram: {akb, kba, bar}

LSTM

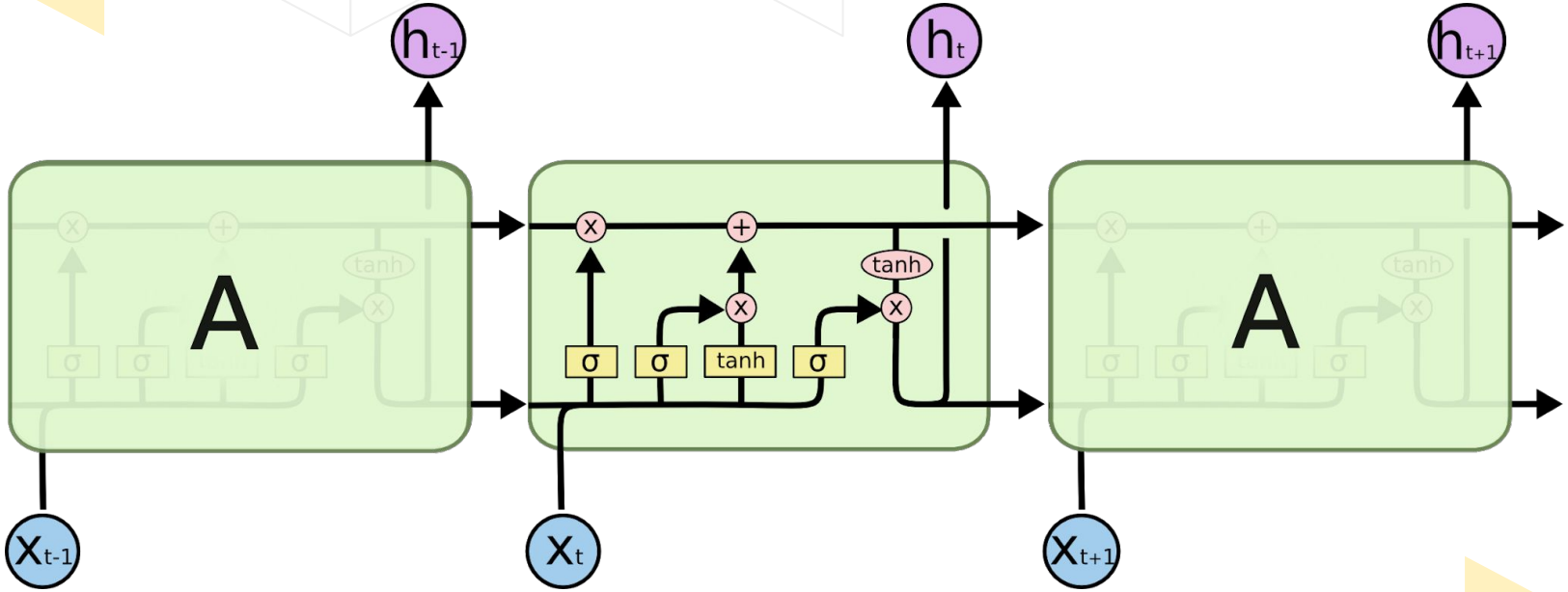
A deep learning approach



Source:
http://colah.github.io/posts/2015-08-Understanding-LSTMs/



Source:
http://colah.github.io/posts/2015-08-Understanding-LSTMs/



Source:
http://colah.github.io/posts/2015-08-Understanding-LSTMs/

What you need to know...

1. Embedding layer
2. Sequence padding
3. You can return the full output sequence (probably for machine translation)

Key takeaways

- Start from something simple
- LSTM is very powerful for sequential data
- Don’t forget to tune your hyperparameters!

Thank you!

@aliakbars

https://github.com/aliakbars/pycon-id-2018