Lecture Notes for

# Machine Learning in Python

Professor Eric Larson

## Lecture: RNN Demo

# Lecture Agenda

- Logistics
  - RNNs due **During Finals Time**
- Recurrent Networks
  - *Overview*
  - *Problem Types*
  - *Embeddings*
  - *Types of RNNs*
  - **Demo A**
  - **CNNs and RNNs**
  - **Demo B**
  - **Ethics Case Study**
  - **Course Retrospective**

# Class Overview, by topic

**Table Data Visualization**

Numpy, Pandas, Seaborn
Overviews with some in-depth discussion

**Dimension Reduction and Image Processing**

Scikit-learn, Scikit Image,
Intuition only, Some mathematics

**Linear and Logistic Regression**

Numpy, Recreate API for Scikit-learn
Detailed mathematics for simple optimization
intuition for advanced optimization

**Neural Networks and Back Prop.**

Numpy
Detailed mathematics for NN operations

**Wide and Deep Networks**

**Convolutional Networks**

**Recurrent Networks**

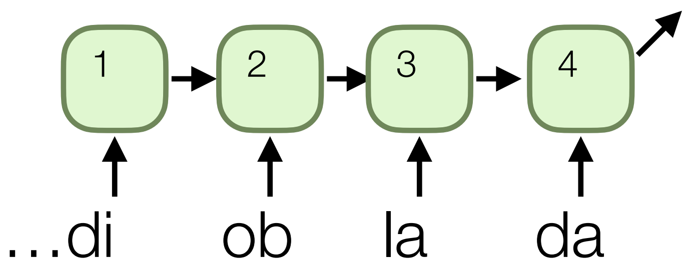
Keras, Tensorflow
Intuition, Detailed implement.

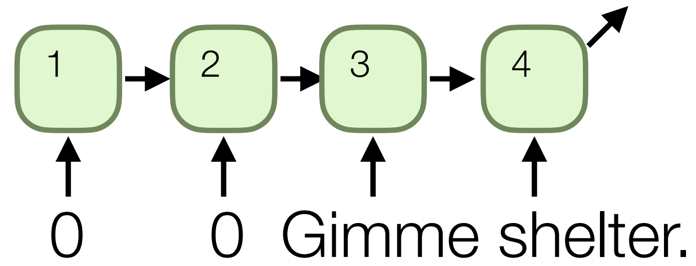
**Ethics in Language Models**

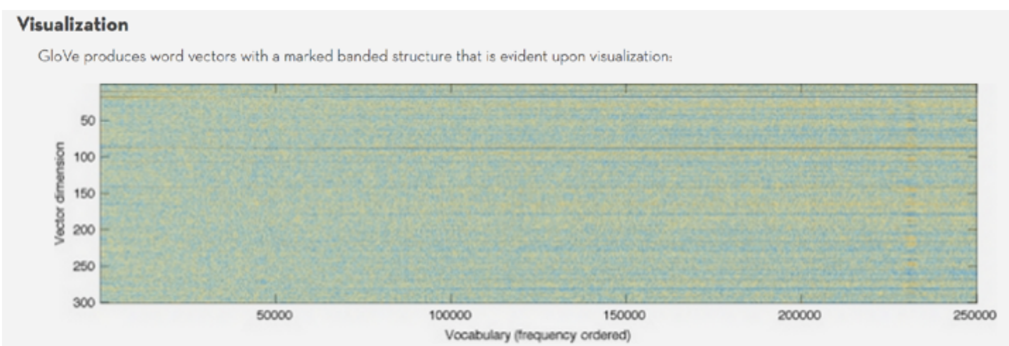
ConceptNet
Case studies

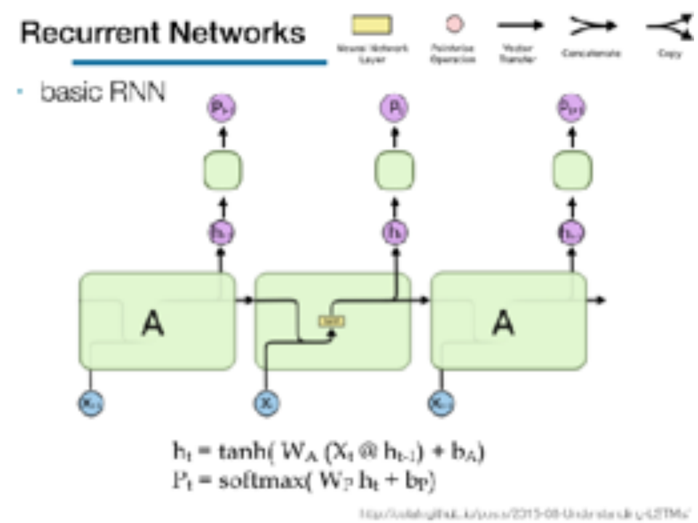
# Last Time

- padding/clipping



# Word Embeddings

- Many are pre-trained for you!!

## GloVe

Global Vectors for Word Representation

### Highlights

**1. Nearest neighbors**

The Euclidean distance (or cosine similarity) between two word vectors provides an effective method for measuring the linguistic or semantic similarity of the corresponding words. Sometimes, the nearest neighbors according to this metric reveal rare but relevant words that lie outside an average human's vocabulary. For example, here are the closest words to the target word *frog*:

0. *frog*
1. frogs
2. toad
3. litoria
4. leptodactylidae
5. rana
6. lizard
7. eleutherodactylus

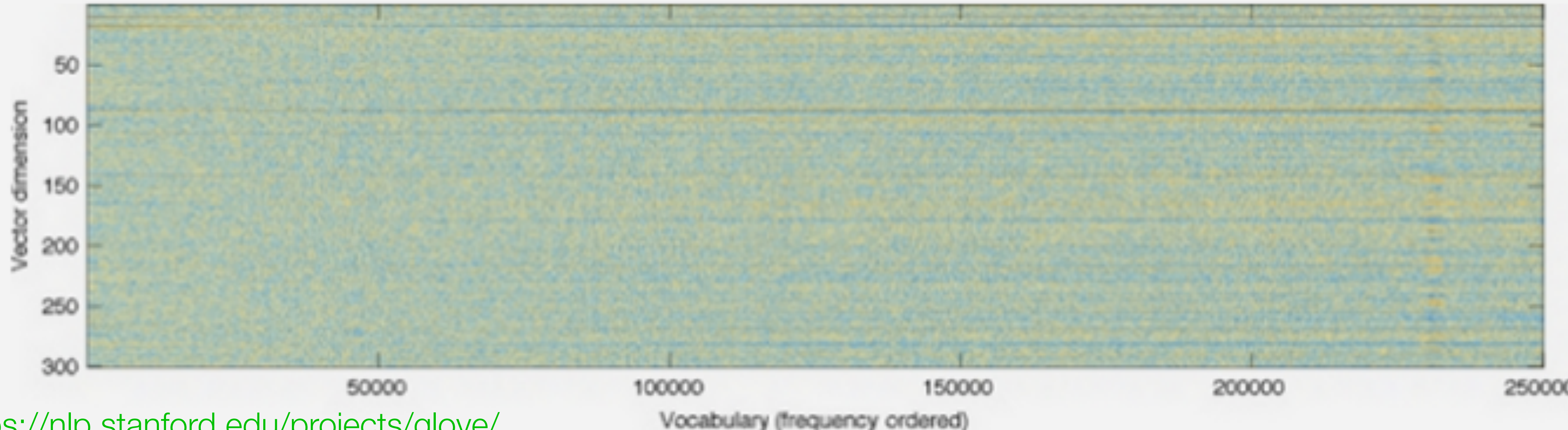
3. litoria

4. leptodactylidae

5. rana

7. eleutherodactylus

GloVe produces word vectors with a marked banded structure that is evident upon visualization:

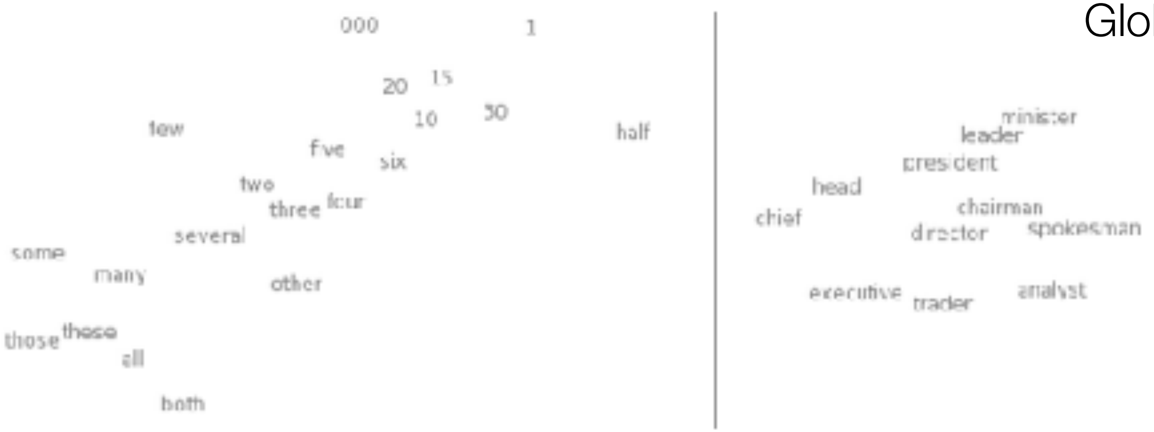



https://nlp.stanford.edu/projects/glove/

# Word Embeddings: proximity

**GloVe**

Global Vectors for Word Representation




t-SNE visualizations of word embeddings. Left: Number Region; Right: Jobs Region. From Turian *et al.* (2010), see complete image.

| FRANCE | JESUS | XBOX | REDDISH | SCRATCHED | MEGABITS |
|---|---|---|---|---|---|
| AUSTRIA | GOD | AMIGA | GREENISH | NAILED | OCTETS |
| BELGIUM | SATI | PLAYSTATION | BLUISH | SMASHED | MB/S |
| GERMANY | CHRIST | MSX | PINKISH | PUNCHED | BIT/S |
| ITALY | SATAN | IPOD | PURPLISH | POPPED | BAUD |
| GREECE | KALI | SEGA | BROWNISH | CRIMPED | CARATS |
| SWEDEN | INDRA | PSNUMBER | GREYISH | SCRAPED | KBIT/S |
| NORWAY | VISHNU | HD | GRAYISH | SCREWED | MEGAHERTZ |
| EUROPE | ANANDA | DREAMCAST | WHITISH | SECTIONED | MEGAPIXELS |
| HUNGARY | PARVATI | GEFORCE | SILVERY | SLASHED | GBIT/S |
| SWITZERLAND | GRACE | CAPCOM | YELLOWISH | RIPPED | AMPERES |

What words have embeddings closest to a given word? From Collobert *et al.* (2011)

The **chairman** called the **meeting** to order.

The **director** called the **conference** to order.

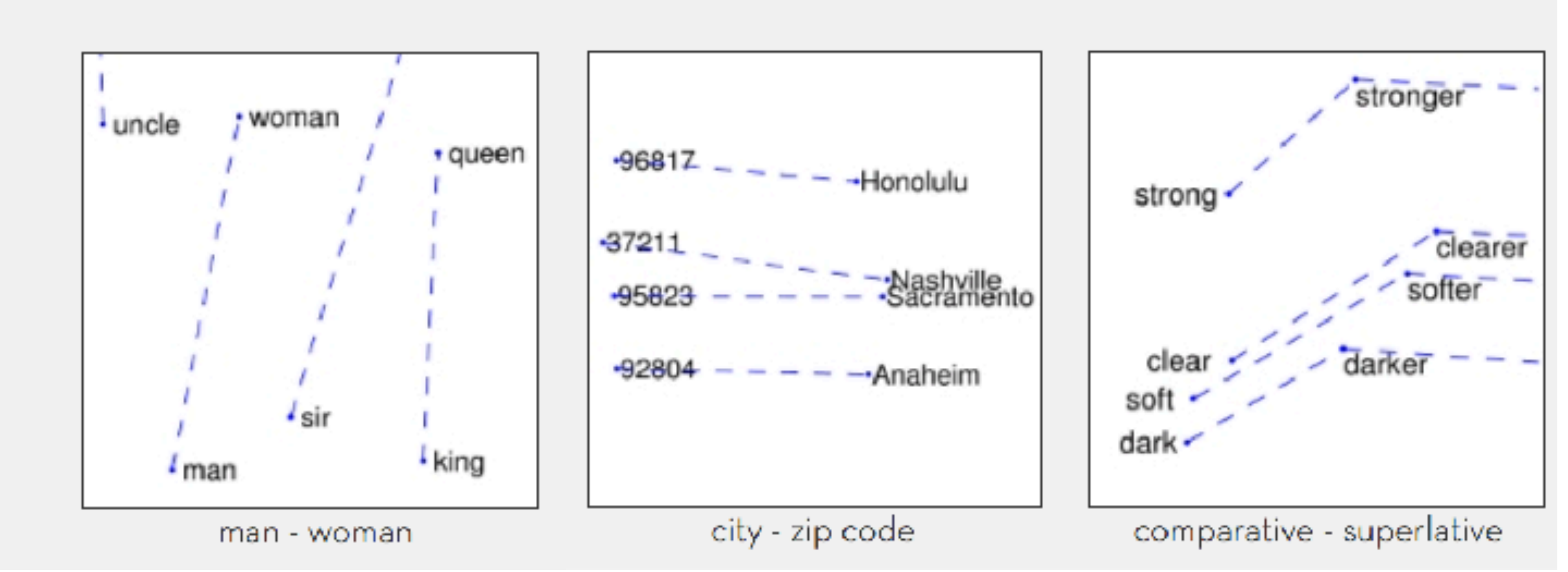
The **chief** called the **council** to order.

http://colah.github.io/posts/2014-07-NLP-RNNs-Representations/

# Word Embeddings: Analogy

**GloVe**

Global Vectors for Word Representation



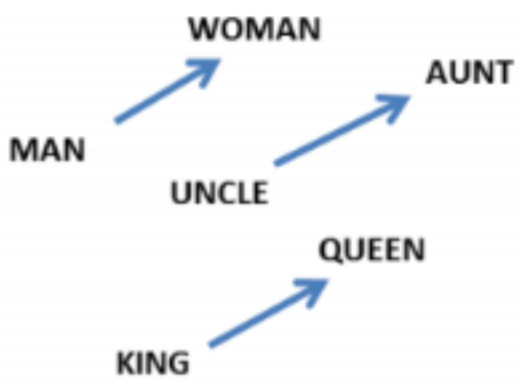

each axis **might** encode a different type of relationship

https://nlp.stanford.edu/projects/glove/ http://colah.github.io/posts/2014-07-NLP-RNNs-Representations/

# Word Embeddings: Analogy

## GloVe

Global Vectors for Word Representation




From Mikolov *et al.* (2013a)

$$W(\text{``woman''}) - W(\text{``man''}) \simeq W(\text{``aunt''}) - W(\text{``uncle''})$$

$$W(\text{``woman''}) - W(\text{``man''}) \simeq W(\text{``queen''}) - W(\text{``king''})$$

| Relationship | Example 1 | Example 2 | Example 3 |
|---|---|---|---|
| France - Paris | Italy: Rome | Japan: Tokyo | Florida: Tallahassee |
| big - bigger | small: larger | cold: colder | quick: quicker |
| Miami - Florida | Baltimore: Maryland | Dallas: Texas | Kona: Hawaii |
| Einstein - scientist | Messi: midfielder | Mozart: violinist | Picasso: painter |
| Sarkozy - France | Berlusconi: Italy | Merkel: Germany | Koizumi: Japan |
| copper - Cu | zinc: Zn | gold: Au | uranium: plutonium |
| Berlusconi - Silvio | Sarkozy: Nicolas | Putin: Medvedev | Obama: Barack |
| Microsoft - Windows | Google: Android | IBM: Linux | Apple: iPhone |
| Microsoft - Ballmer | Google: Yahoo | IBM: McNealy | Apple: Jobs |
| Japan - sushi | Germany: bratwurst | France: tapas | USA: pizza |

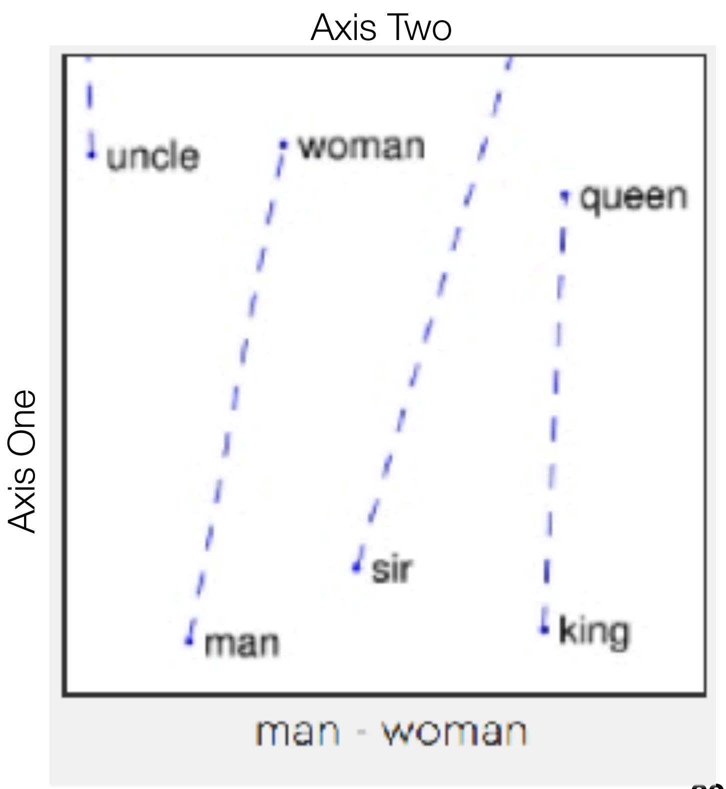
Relationship pairs in a word embedding. From Mikolov *et al.* (2013b).

https://nlp.stanford.edu/projects/glove/

http://colah.github.io/posts/2014-07-NLP-RNNs-Representations/

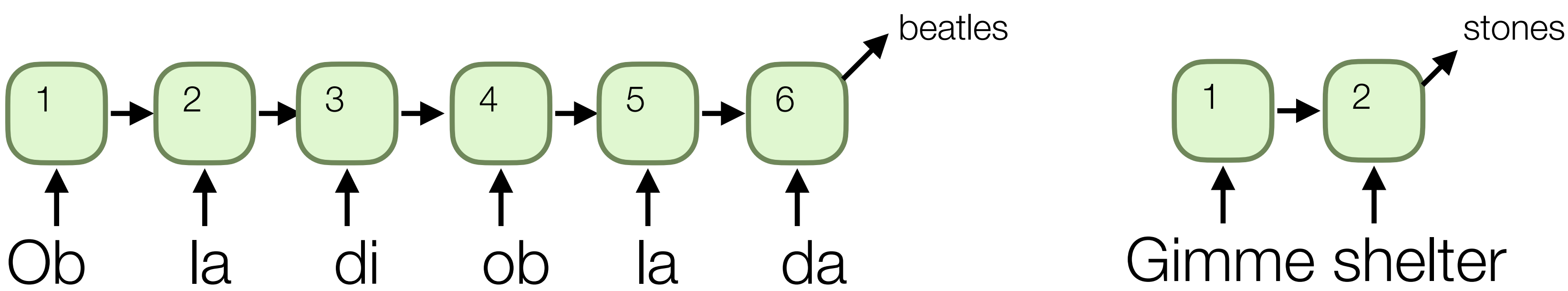
# Self Test: Analogy

- Each axis on the embedding plot below corresponds to:
- A. a weight inside the embedding layer
- B. an average of weights inside the embedding layer
- C. the average of the one hot encoding for a word
- D. an output of the embedding layer



# Practical Logistics: Sequence Length

- option A: dynamic length sequences



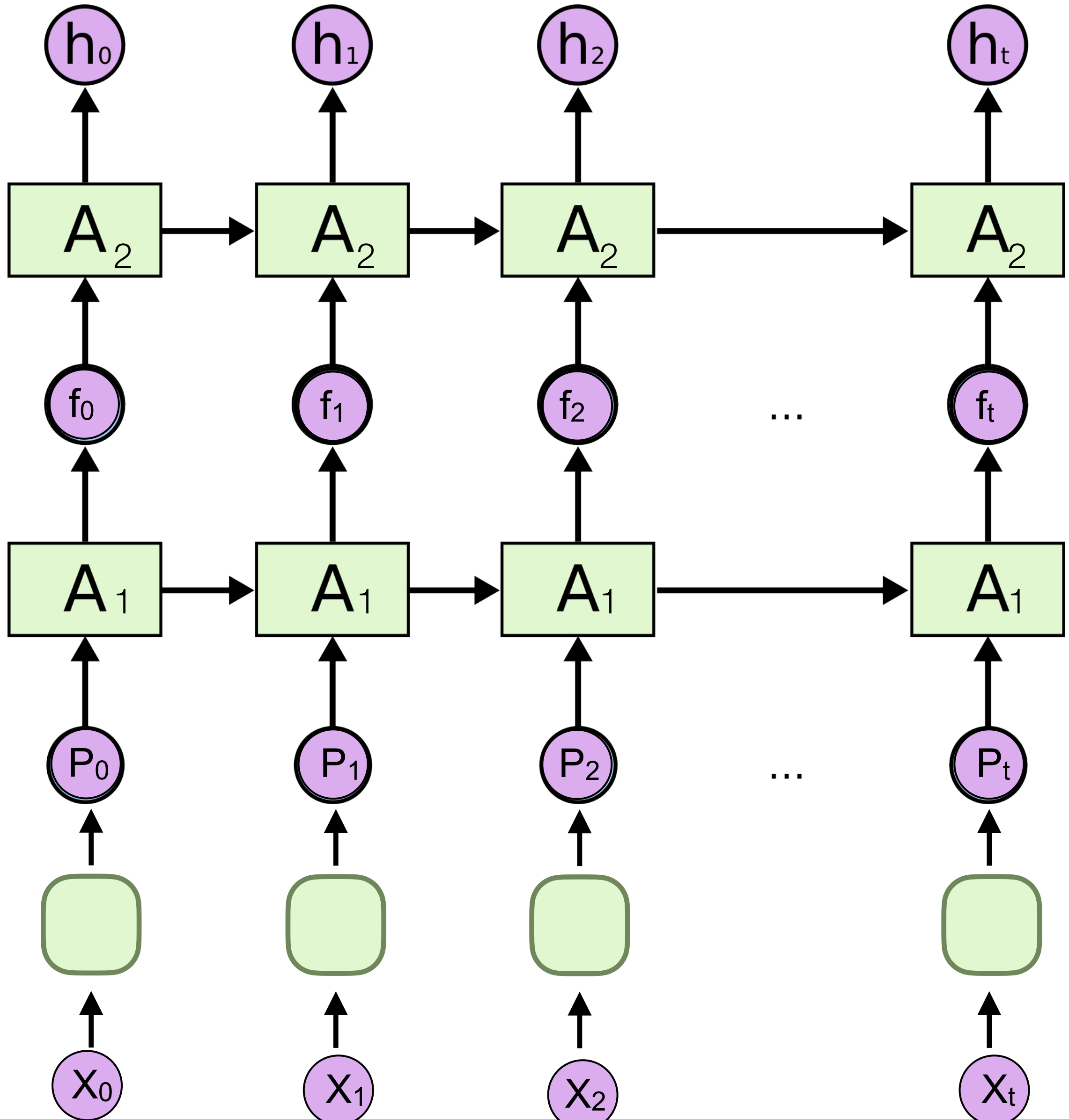

- option B: padding/clipping


- main difference:

  **speed based on computation graph design**

# Sequence Stacking

**keras**:
return sequences

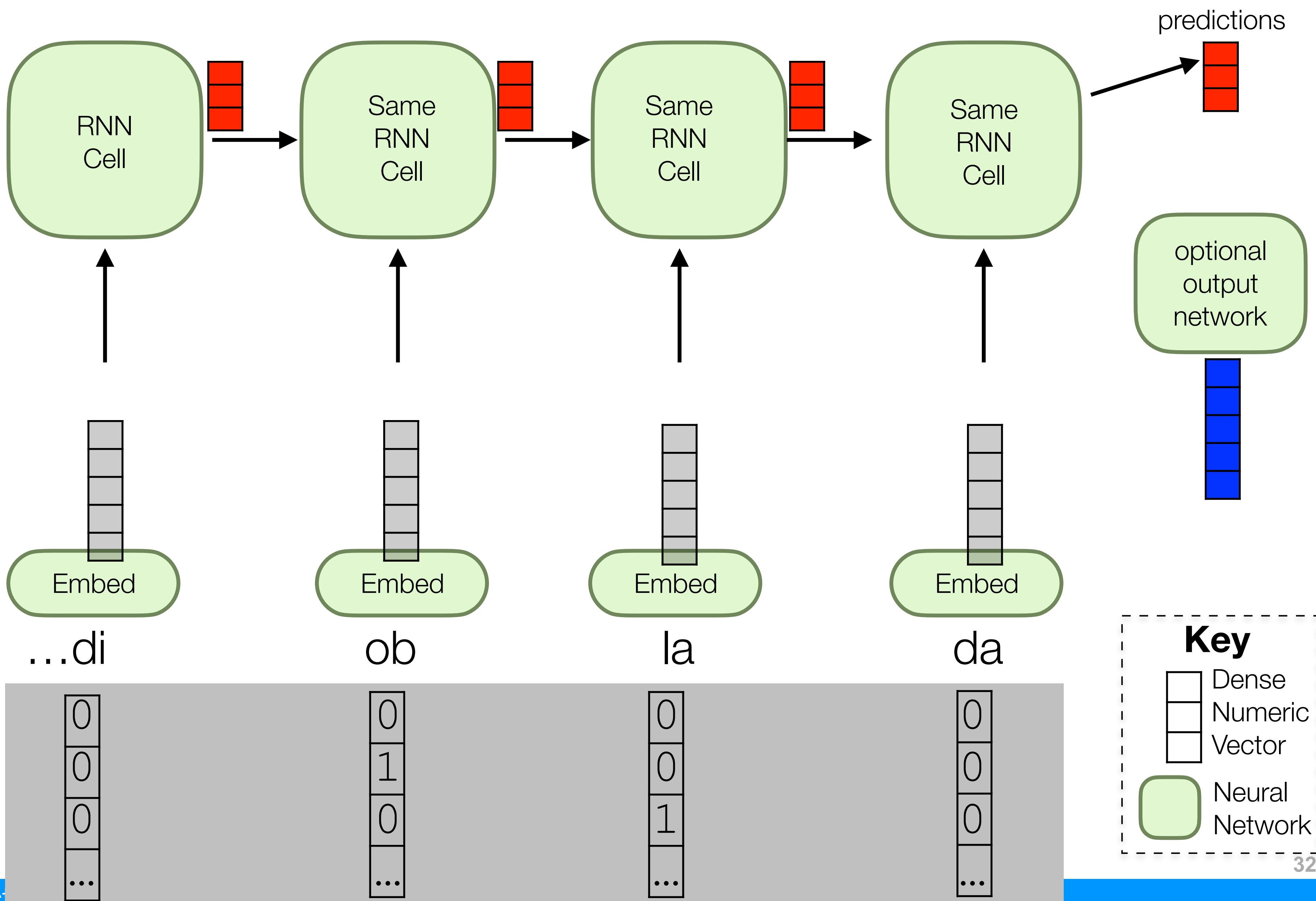



**keras**:
add final dense layer for prediction from last element only

**keras RNN default**:
return last element only
*go to dense layer for prediction*

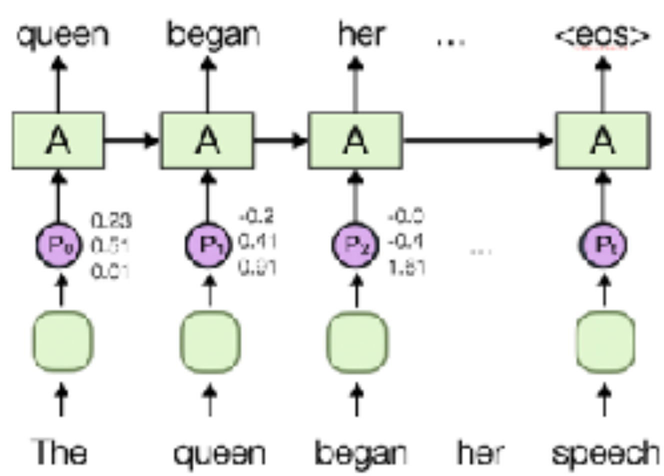
**keras embed**:
length of embedding is meant for sequences

# General recurrent flow (many to one)



# Self Test

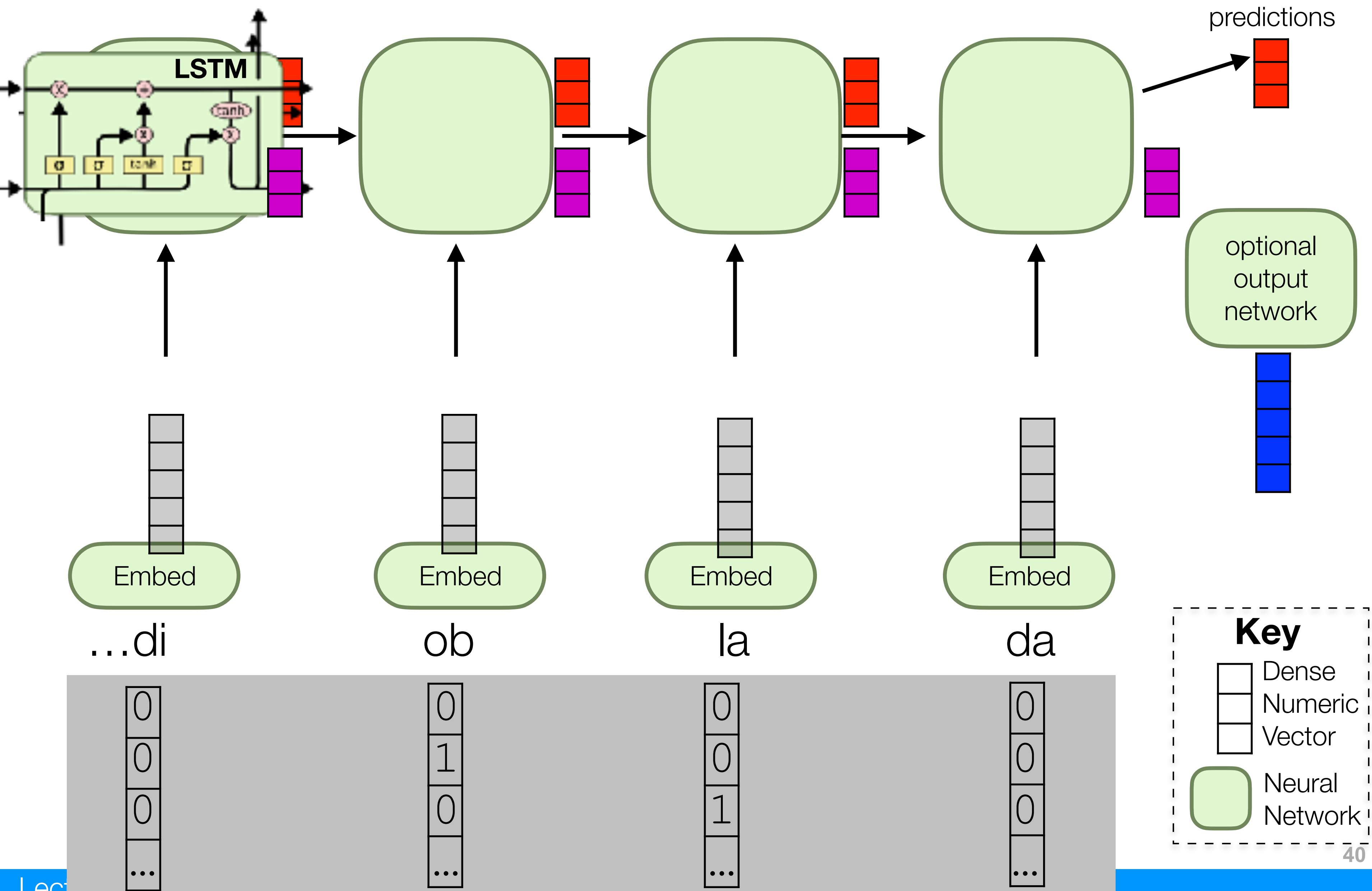
- T/F: In Recurrent Neural Networks that are “rolled out”, each RNN cell can be run in parallel.
    - A. **True**, state vectors can be added later
    - B. **True**, but parallelization must use forward backward (like Viterbi)
    - C. **False**, state vectors must be found sequentially
    - D. **False**, input changes due to sequential nature of $X_t$



# Commonly Used RNN Nodes


I like this version better.

# General recurrent flow (many to one)



- gated recurrent units

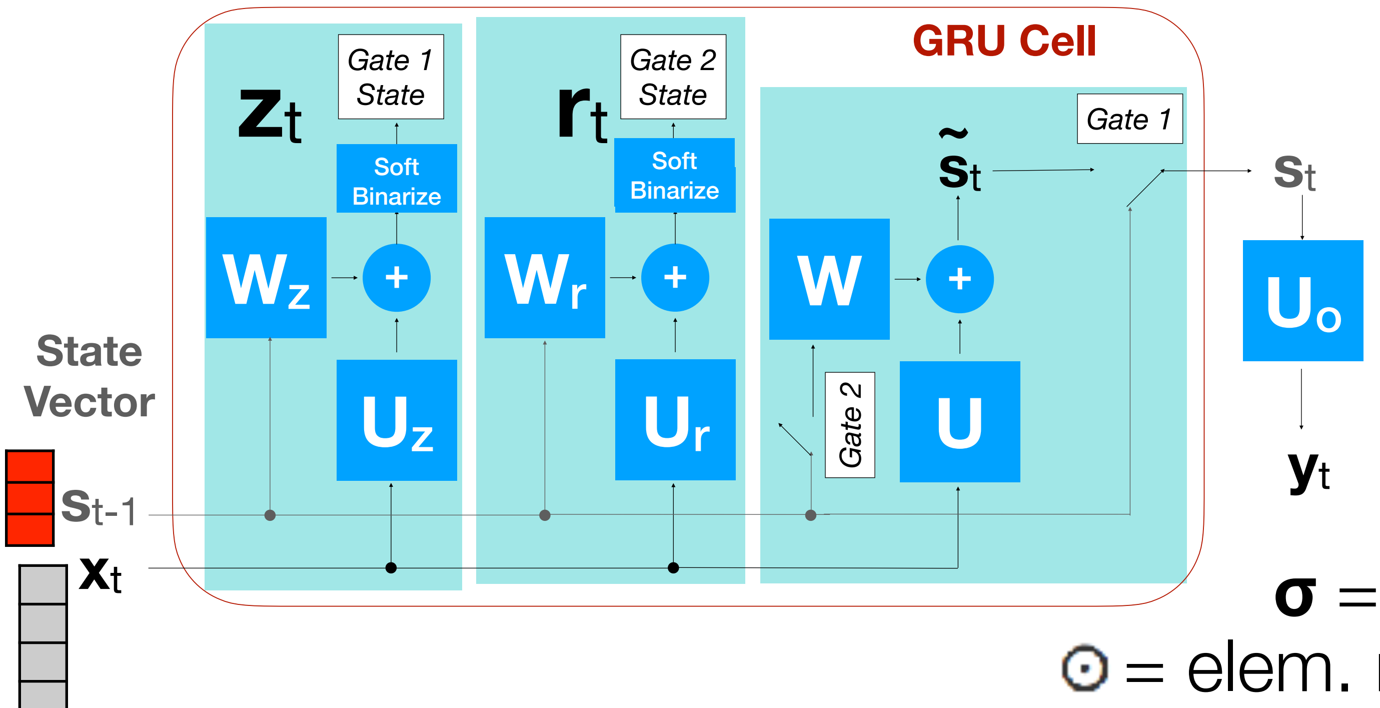
Selectivity controls, gates (**0 or 1**)

$$r_t = \sigma(W_r \boxed{s_{t-1}} + U_r \boxed{x_t} + b_r)$$
$$z_t = \sigma(W_z \boxed{s_{t-1}} + U_z \boxed{x_t} + b_z)$$

past state current input

$$\tilde{s}_t = \phi(\overbrace{W(r_t \odot s_{t-1})}^{\text{selectively remember}} + \overbrace{U x_t}^{\text{with influence}} + b)$$

$$s_t = \underbrace{z_t \odot s_{t-1}}_{\text{remember only past}} + \underbrace{(1 - z_t) \odot \tilde{s}_t}_{\text{OR remember with input}}$$




$\sigma$ = sigmoid

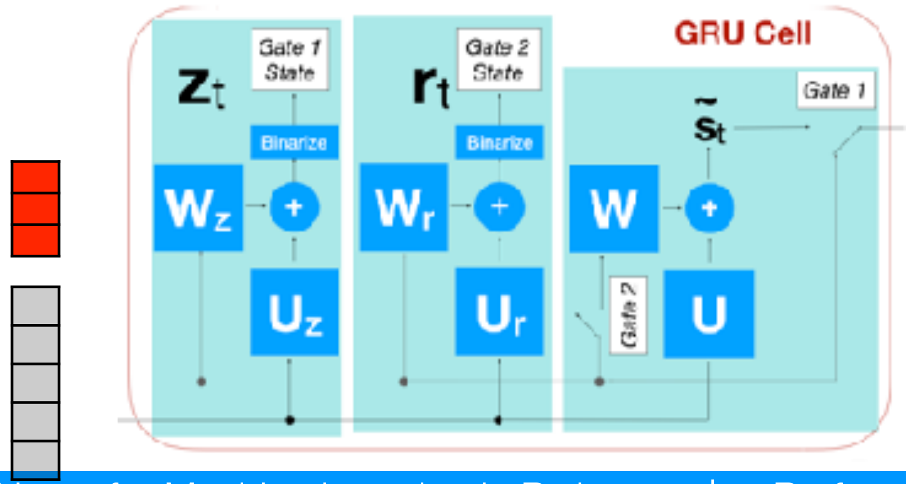
$\odot$ = elem. multiplication

# Self Test

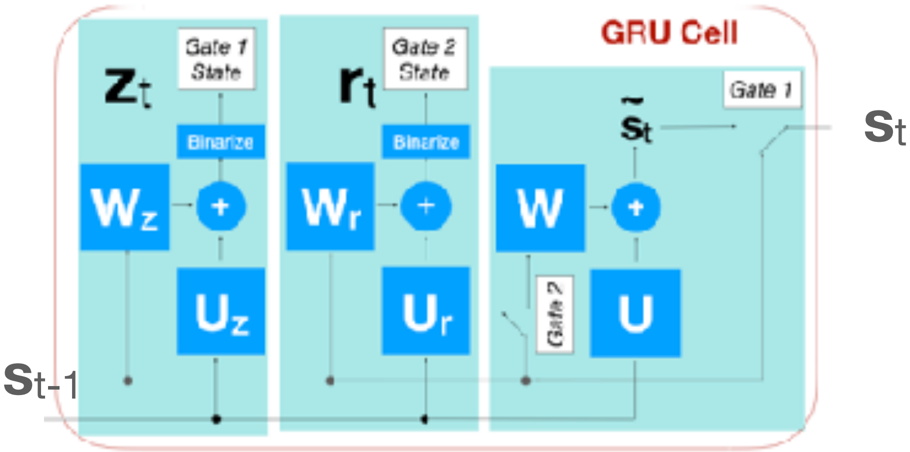
- What element of the GRU helps with vanishing and exploding gradients?
- A. derivative of $\sigma$
- B. no activation function
- C. derivative of $\phi$
- D. $\phi$

$$r_t = \sigma(W_r s_{t-1} + U_r x_t + b_r)$$
$$z_t = \sigma(W_z s_{t-1} + U_z x_t + b_z)$$

$$\tilde{s}_t = \phi(W(r_t \odot s_{t-1}) + U x_t + b)$$
$$s_t = z_t \odot s_{t-1} + (1 - z_t) \odot \tilde{s}_t$$



# Derivative of GRU




$$r_t = \sigma(W_r s_{t-1} + U_r x_t + b_r)$$
$$z_t = \sigma(W_z s_{t-1} + U_z x_t + b_z)$$

$$\tilde{s}_t = \phi(W(r_t \odot s_{t-1}) + U x_t + b)$$
$$s_t = \boxed{z_t \odot s_{t-1}} + \boxed{(1 - z_t) \odot \tilde{s}_t}$$

Product Rule

Product Rule

**To back propagate, we need sensitivity of state vector, w.r.t previous state**

$$\partial s_t \,/\, \partial s_{t-1} = (\partial z_t \text{ x } s_{t-1}) + (\partial s_{t-1} \text{ x } z_t) + \partial \tilde{s}_t - (\partial z_t \text{ x } \tilde{s}_t) - (\partial \tilde{s}_t \text{ x } z_t)$$

likely vanish

hard to vanish unless $z_t = 0$

could vanish, depending on ϕ

likely vanish

could vanish, depending on ϕ

- LSTM prototype

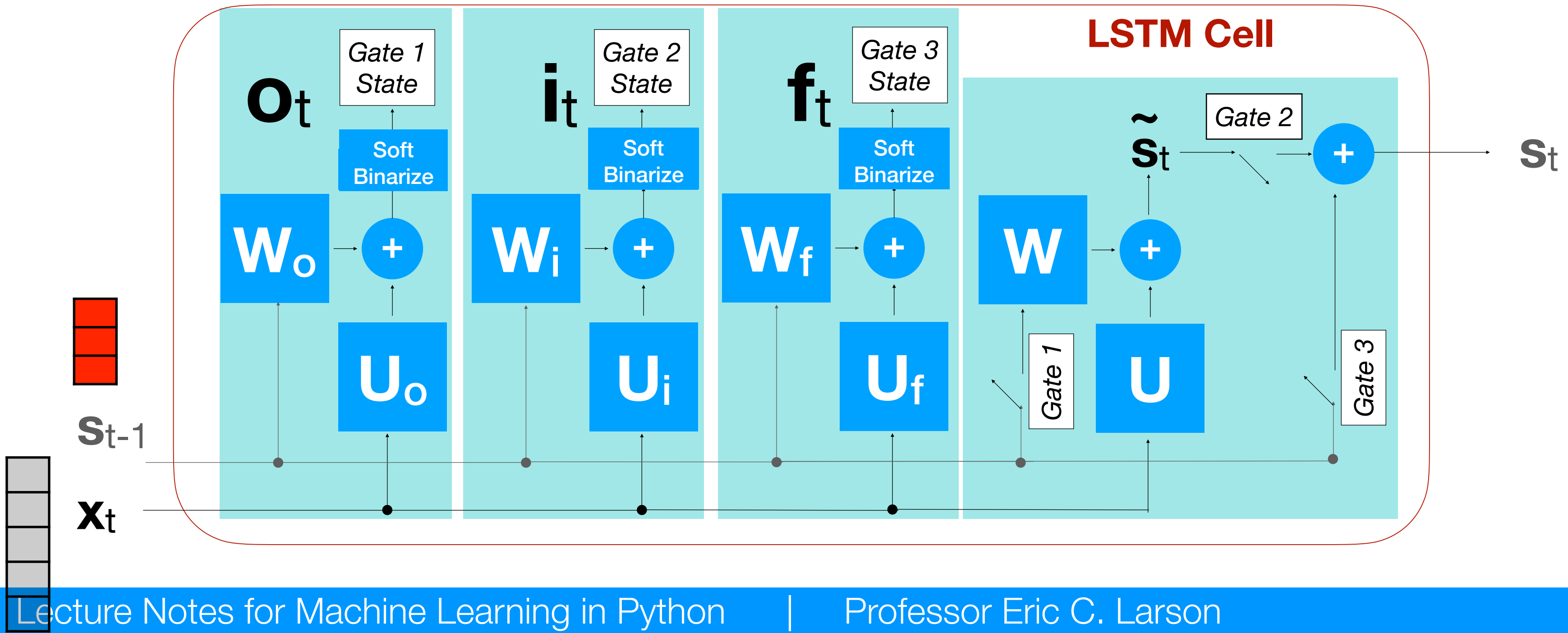
Selectivity controls (**gates, 0 or 1**)

$$o_t = \sigma(W_o s_{t-1} + U_o x_t + b_o)$$
$$i_t = \sigma(W_i s_{t-1} + U_i x_t + b_i)$$
$$f_t = \sigma(W_f s_{t-1} + U_f x_t + b_f)$$

selectively remember past　　with influence

$$\tilde{s}_t = \phi(W(\boxed{o_t \odot s_{t-1}}) + \boxed{U x_t + b})$$
$$s_t = \boxed{f_t \odot s_{t-1}} + \boxed{i_t \odot \tilde{s}_t}$$

selectively remember past　　with past weighted influence



# Recurrent Networks: Gen 2 LSTM

- LSTM in TensorFlow

Selectivity controls (**gates, 0 or 1**)

$$i_t = \sigma(W_i h_{t-1} + U_i x_t + b_i)$$

$$o_t = \sigma(W_o h_{t-1} + U_o x_t + b_o)$$

$$f_t = \sigma(W_f h_{t-1} + U_f x_t + b_f)$$

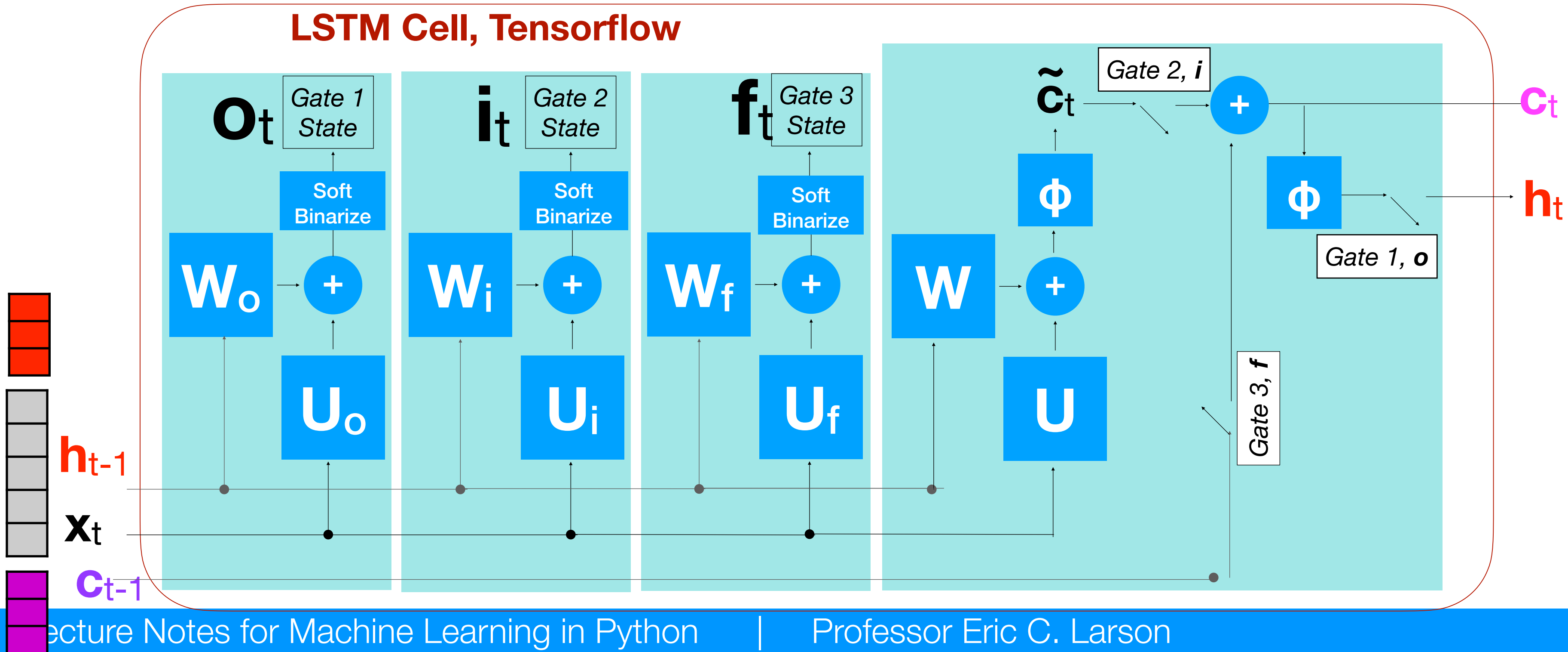
explicit remembering state

$$\tilde{c}_t = \phi(W h_{t-1} + U x_t + b)$$

$$c_t = \underbrace{\boxed{f_t \odot c_{t-1}}}_{\text{remember previous state}} + \underbrace{\boxed{i_t \odot \tilde{c}_t}}_{\text{update with output, } h_t}$$

$$h_t = \boxed{o_t \odot \phi(c_t)}$$

get next ht for selecting gates



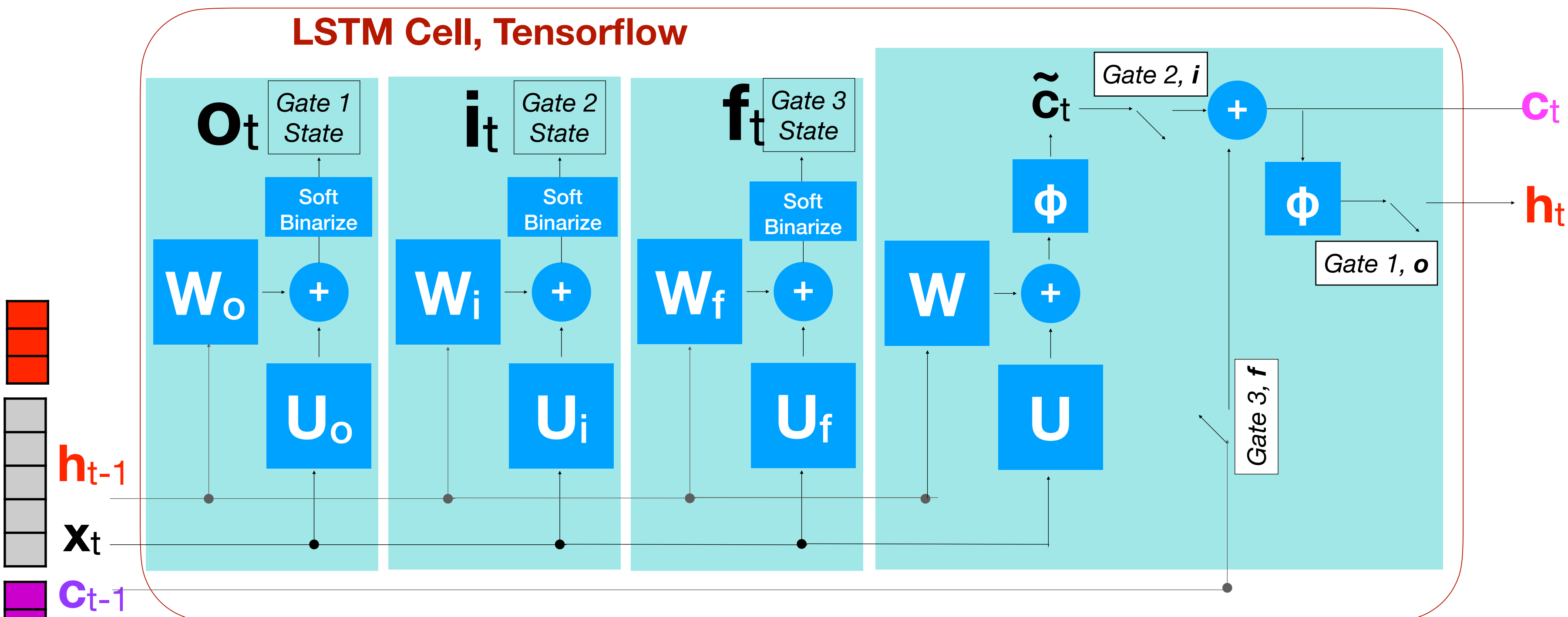
# LSTM Dropout

$$
\begin{aligned}
i_t &= \sigma(W_i h_{t-1} + U_i x_t + b_i) \\
o_t &= \sigma(W_o h_{t-1} + U_o x_t + b_o) \\
f_t &= \sigma(W_f h_{t-1} + U_f x_t + b_f)
\end{aligned}
$$

Recurrent Dropout

Input Dropout

The days of training **without** using **dropout** are **over**.



# What to choose?

- There is no hard and fast rule
  - try both
  - basic LSTM has had great success
  - GRU also sometimes is easier to train
  - you will see many variations
    - peephole LSTM
    - hierarchical LSTM
    - and many more…

# Final Demo - Part A

Many to one:

- Simple RNNs
- GRUs
- LSTMs

More examples:

https://github.com/tensorflow/tensorflow/tree/r0.11/tensorflow/examples/skflow
http://r2rt.com/recurrent-neural-networks-in-tensorflow-i.html
http://machinelearningmastery.com/sequence-classification-lstm-recurrent-neural-networks-python-keras/

Seq2Seq:

https://github.com/tensorflow/tensorflow/blob/r0.11/tensorflow/examples/skflow/neural_translation_word.py

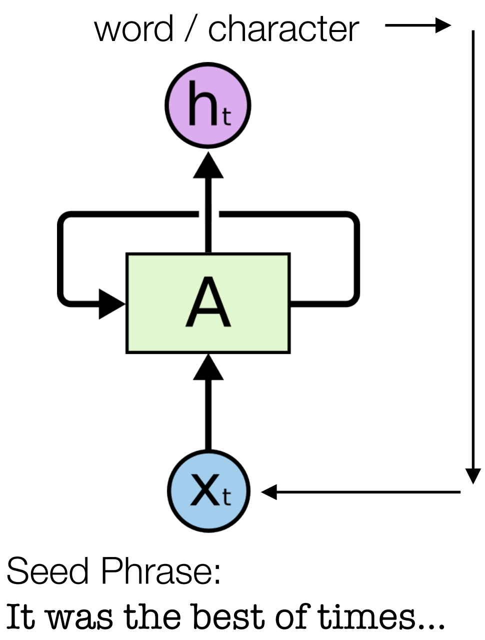
# Recurrent Generation


WHEN VISITING A NEW HOUSE, IT'S GOOD TO CHECK WHETHER THEY HAVE AN ALWAYS-ON DEVICE TRANSMITTING YOUR CONVERSATIONS SOMEWHERE.

**if time: just for fun**

# Generating Outputs

- Highly “sophisticated” steps:
  - train an RNN to generate the **next** word/character from the **current** word/character
  - train on a corpus of text
    - Shakespeare
    - Movie Scripts
    - Whatever!
  - seed with random word, feed output words as input to next node
  - rinse, repeat



# Training Data: Plays by Shakespeare

```
First Citizen:
Before we proceed any further, hear me speak.

All:
Speak, speak.

First Citizen:
You are all resolved rather to die than to famish?

All:
Resolved. resolved.

First Citizen:
First, you know Caius Marcius is chief enemy to the people.

All:
We know't, we know't.

First Citizen:
Let us kill him, and we'll have corn at our own price.
Is't a verdict?

All:
No more talking on't; let it be done: away, away!

Second Citizen:
One word, good citizens.

ervices he has done for his country?

First Citizen:
Very well; and could be content to give him good
report fort, but that he pays himself with being proud.

Second Citizen:
Nay, but speak not maliciously.
```

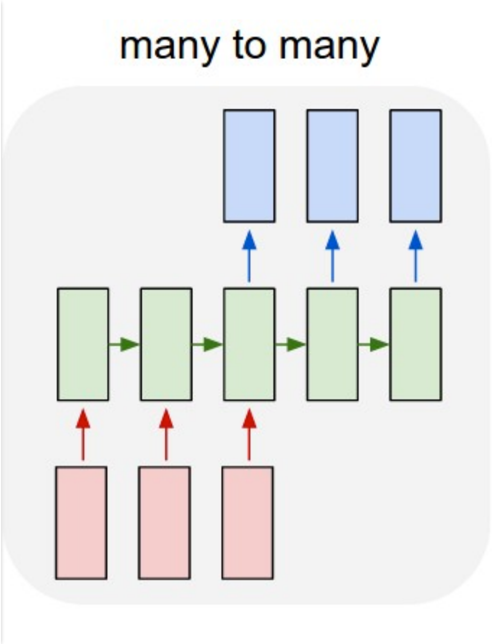
```
First Citizen:
We are accounted poor citizens, the patricians g
What authority surfeits on would relieve us: if
would yield us but the superfluity, while it wer
wholesome, we might guess they relieved us human
but they think we are too dear: the leanness tha
afflicts us, the object of our misery, is as an
inventory to particularise their abundance; our
sufferance is a gain to them Let us revenge this
our pikes, ere we become rakes: for the gods kno
speak this in hunger for bread, not in thirst fo

Second Citizen:
Would you proceed especially against Caius Marci

All:
Against him first: he's a very dog to the common

Second Citizen:
Consider you what s
```

# Generating Shakespeare




```
ATOOOS

UIEAOUYOUZZZZZZUZAAAYAYf n fsflflrurctuateot t ta's  a  wtutss

ESGNANO:
Whith then, a do makes and them and to sees,
I wark on this ance may string take thou honon
To sorriccorn of the bairer, whither, all
I'd see if yiust the would a peid.

LARYNGLe:
To would she troust they fould.

PENMES:
Thou she so the havin to my shald woust of
As tale we they all my forder have
As to say heant thy wansing thag and
Whis it thee shath his breact, I be and might, she
Tirs you desarvishensed and see thee: shall,
What he hath with that is all time,
And sen the have would be sectiens, way thee,
They are there to man shall with me to the mon,
And mere fear would be the balte, as time an at
And the say oun touth, thy way womers thee.
```

# Generating Movies

**more data**: star wars + star trek + tarantino + the matrix

```
DENT'SUEENCK

Bartholomew of the TIE FIGHTERS are stunned. There is a crowd and armored
switcheroos.

PICARD
(continuing)
Couns two dim is tired. In order to the sentence...

The sub    bottle appears on the screen into a small shuttle shift of the
ceiling. The DAMBA FETT splash fires and matches them into the top, transmit to stable high above
upon their statels,
falling from an alien shaft.

ANAKIN and OBI-WAN stand next to OBI-WAN down the control plate of smoke at the TIE fighter. They
stare at the centre of the station loose into a comlink cover -- comes up to the General, the
GENERAL HUNTAN AND FINNFURMBARD from the PICADOR to a beautiful Podracisly.

ENGINEER
Naboo from an army seventy medical
security team area re-weilergular.

EXT.
```

# Movie written by algorithm turns out to be hilarious and intense

For *Sunspring*'s exclusive debut on Ars, we talked to the filmmakers about collaborating with an AI.

ANNALEE NEWITZ - 6/9/2016, 5:30 AM