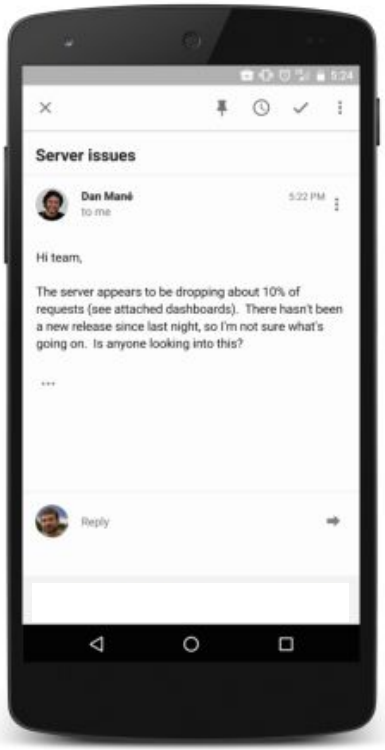
Deep Learning for Language Understanding (at Google Scale)

Anjuli Kannan
Software Engineer, Google Brain

Confidential + Proprietary

Text is just a sequence of words



["hi", "team", "the", "server", "appears", "to", "be", "dropping", "about", "10%", …]

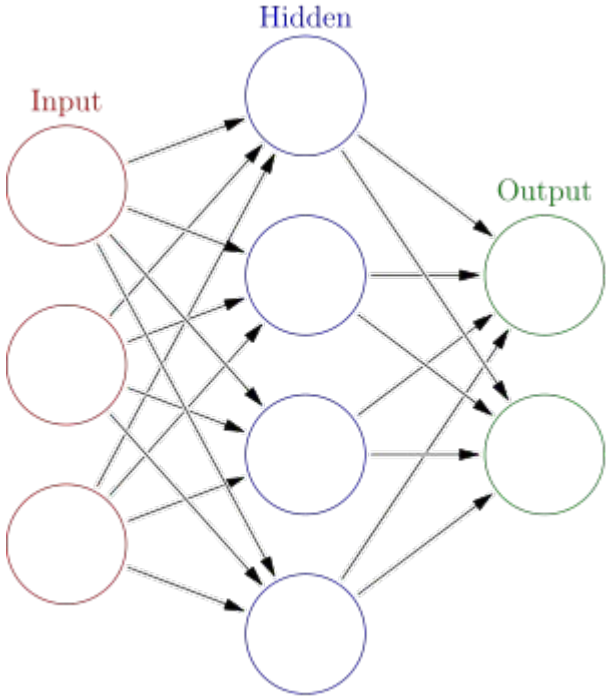
Confidential + Proprietary

About me

- My team: Google Brain
  - "Make machines intelligent, improve people's lives."
  - Research + software + applications
  - g.co/brain
- My work is at boundary of research and applications
- Focus on natural language understanding

Neural network basics

Neural network



...

Is a 4

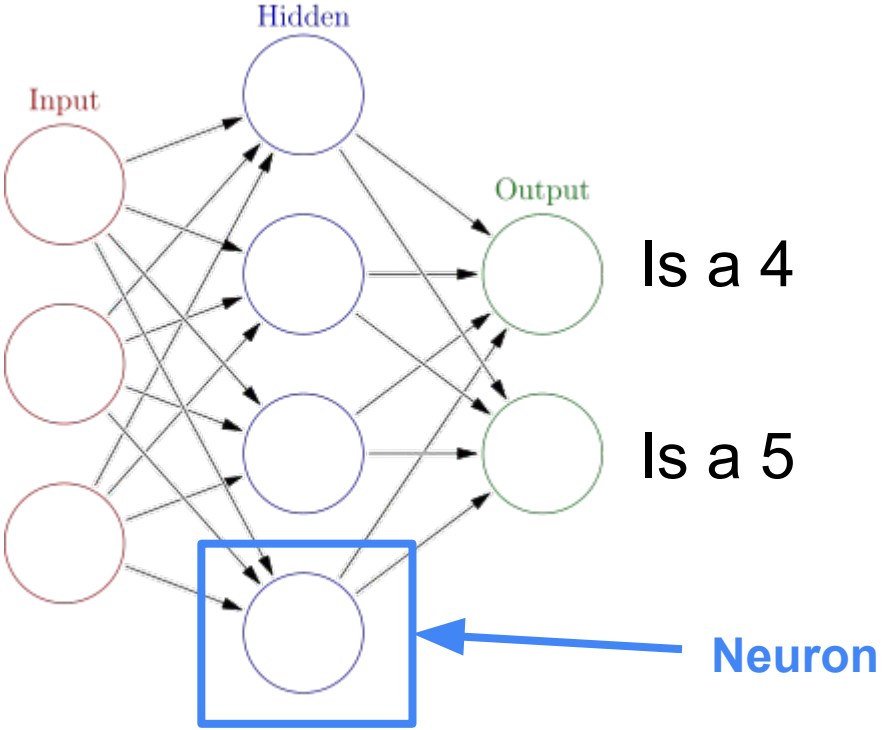
Is a 5

...

Image: Wikipedia

Confidential + Proprietary

Neural network

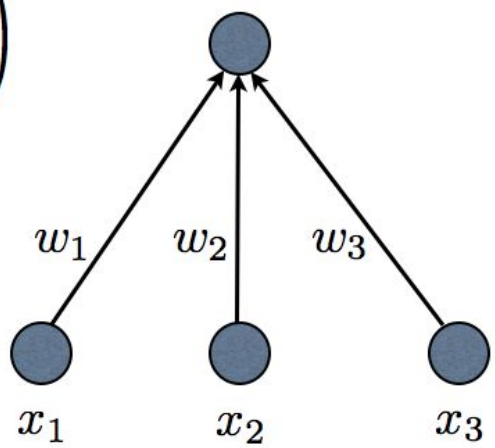


Confidential + Proprietary

Basic building block is the neuron

- **Different weights compute different functions**

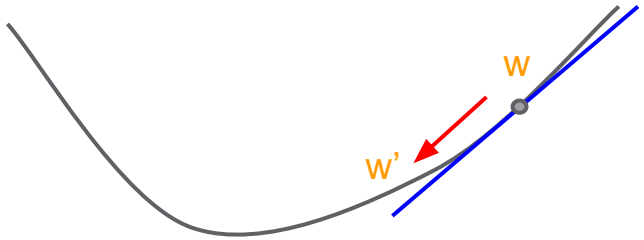
$$y_i = F\left(\sum_i w_i x_i\right)$$



Greg Corrado

Confidential + Proprietary

Gradient descent



Learning Rate

$$w' = w - \alpha \partial_w L(w)$$

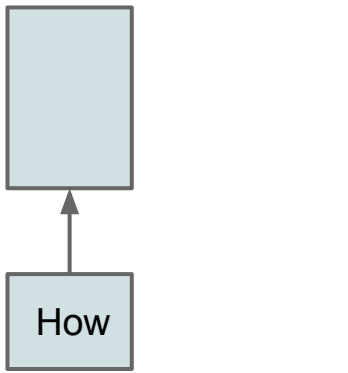
Slide: Vincent Vanhoucke

Recurrent neural networks

Recurrent neural networks can model sequences

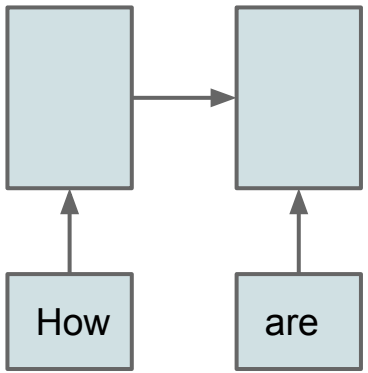
Confidential + Proprietary

Recurrent neural networks can model sequences



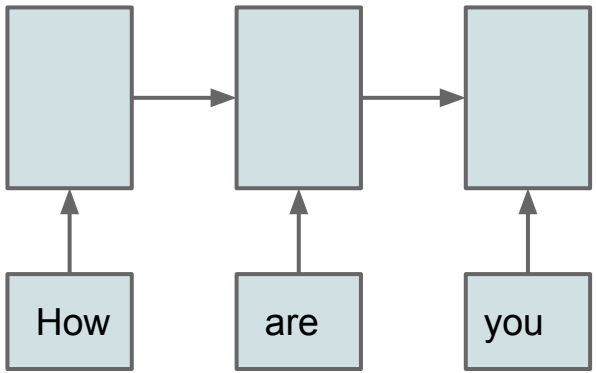
Message

Recurrent neural networks can model sequences



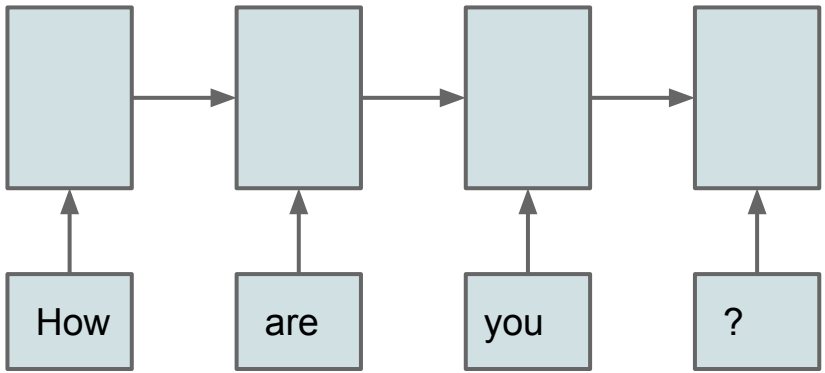
Message

Recurrent neural networks can model sequences



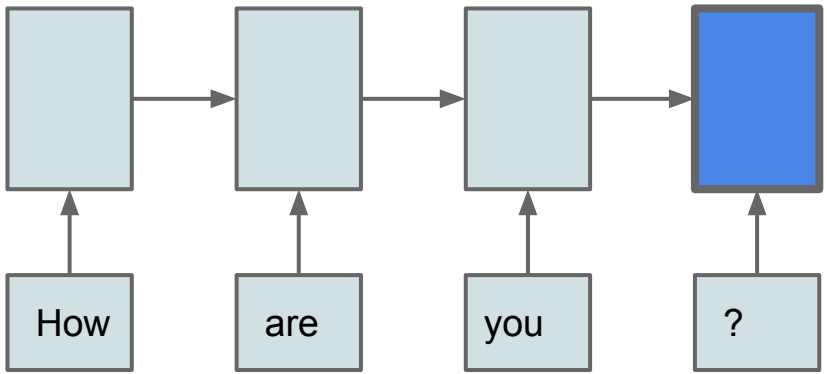
Message

Recurrent neural networks can model sequences



Message

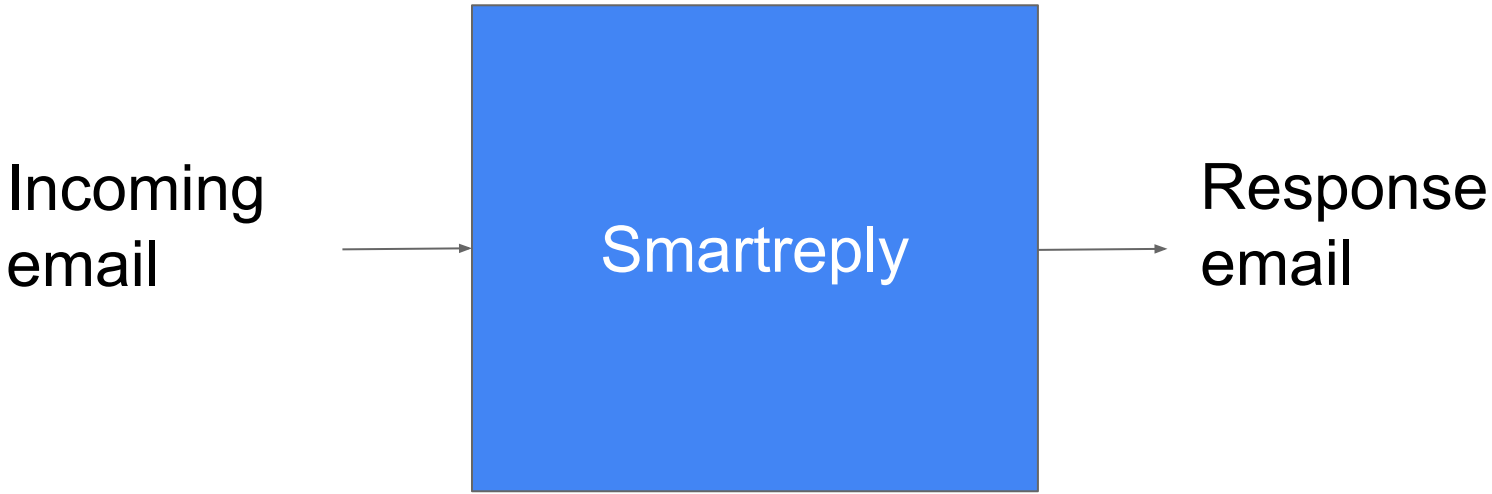
Recurrent neural networks can model sequences



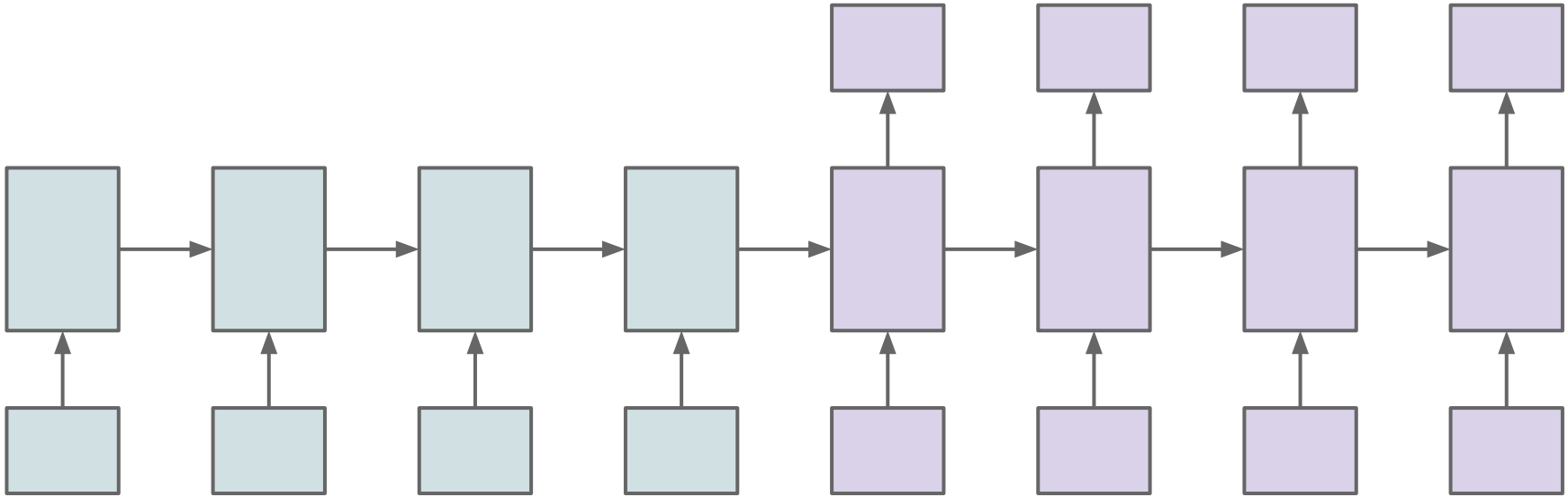
Internal state is a fixed length encoding of the message

Sequence-to-sequence models

Suppose we want to generate email replies

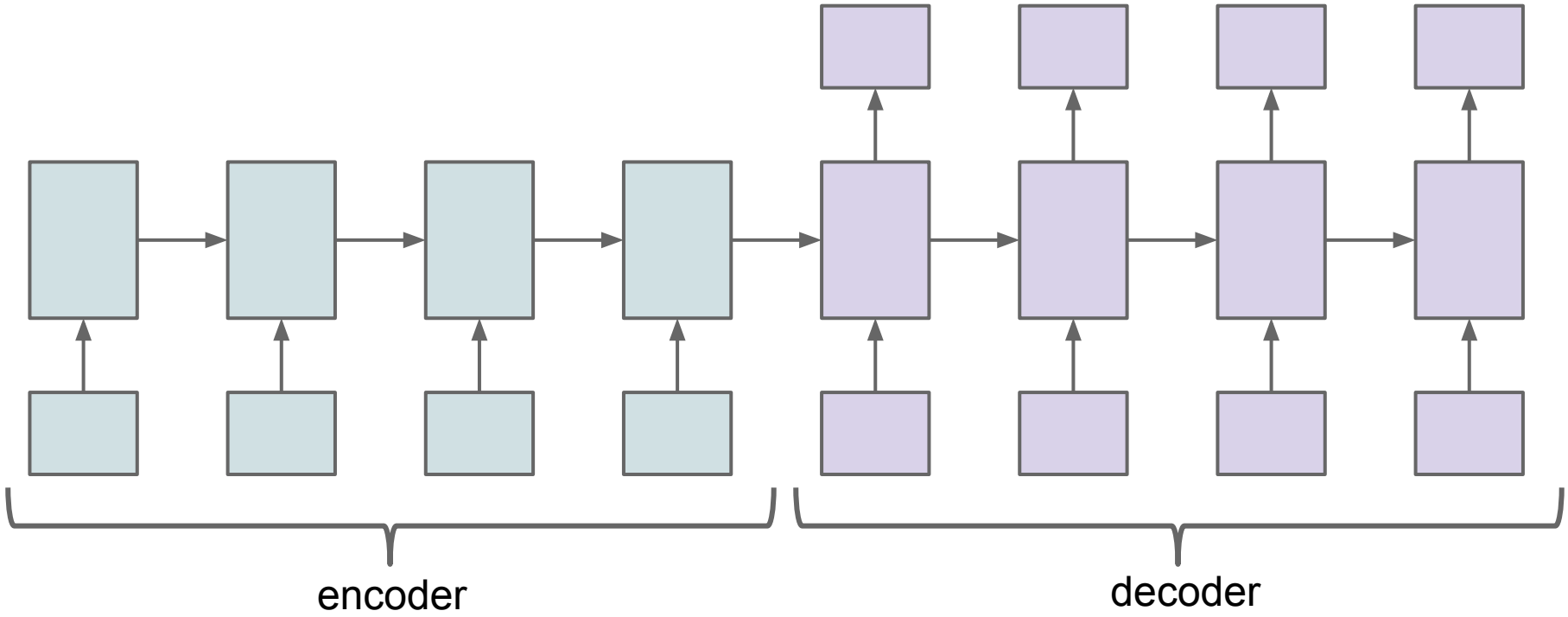


Sequence-to-sequence model

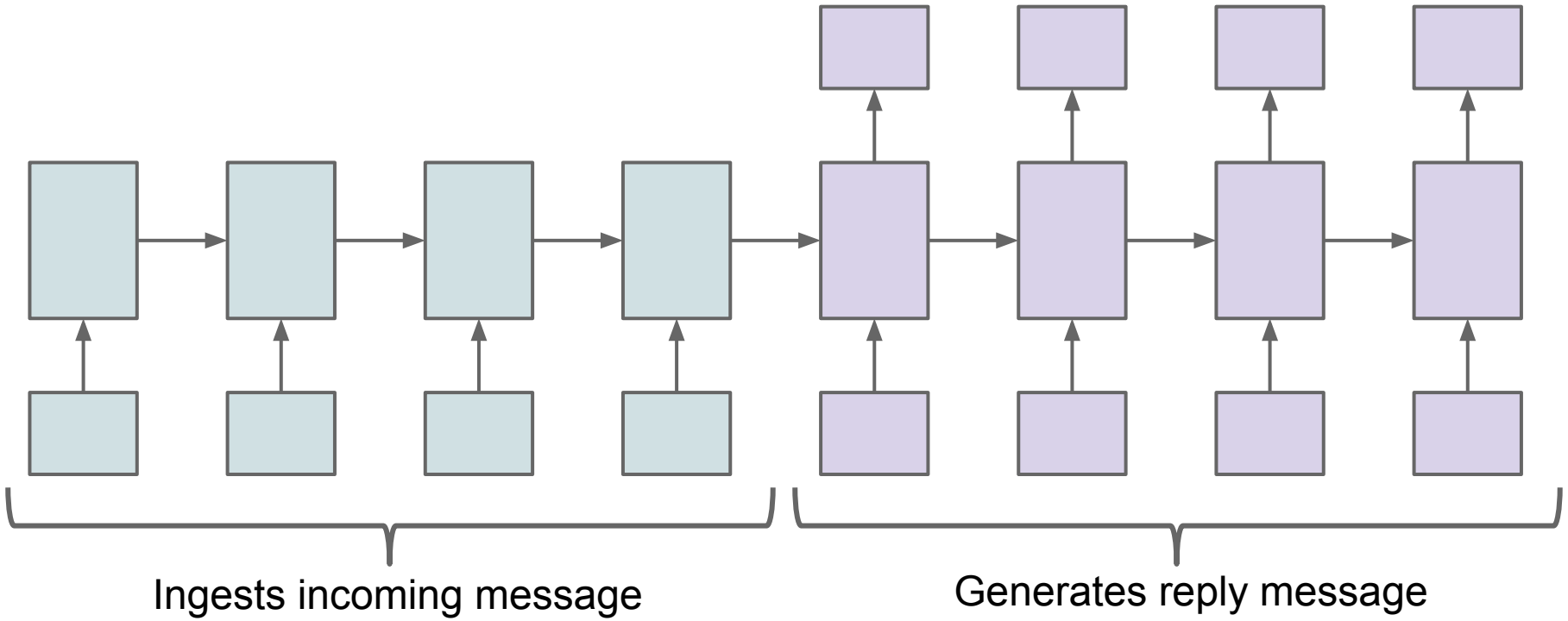


Sutskever et al, NIPS 2014

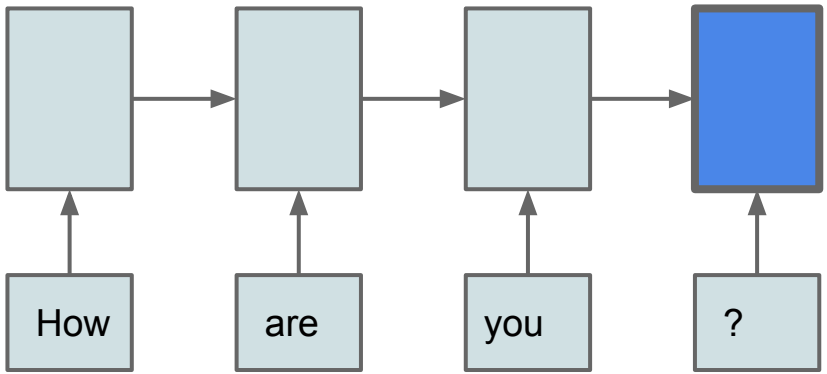
Sequence-to-sequence model



Sequence-to-sequence model

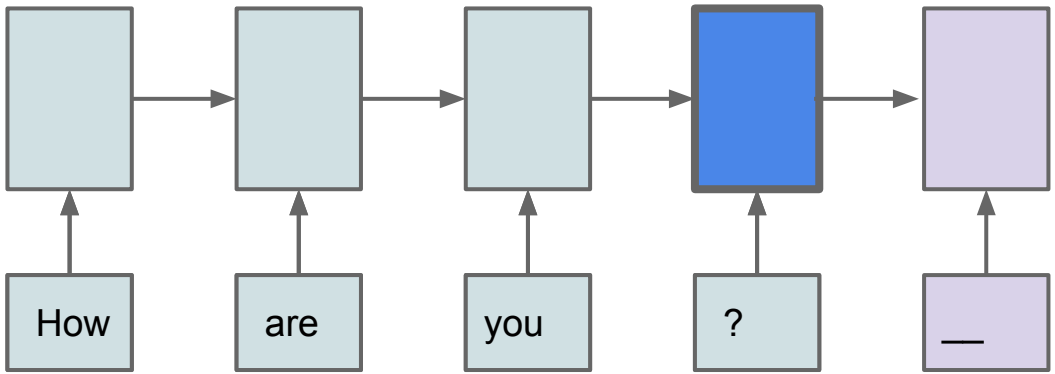


Encoder ingests the incoming message

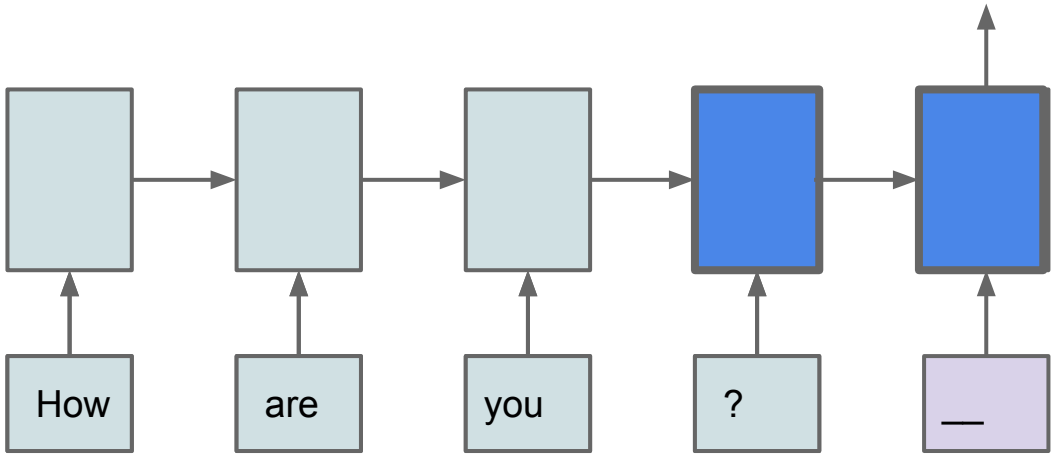


Internal state is a fixed length encoding of the message

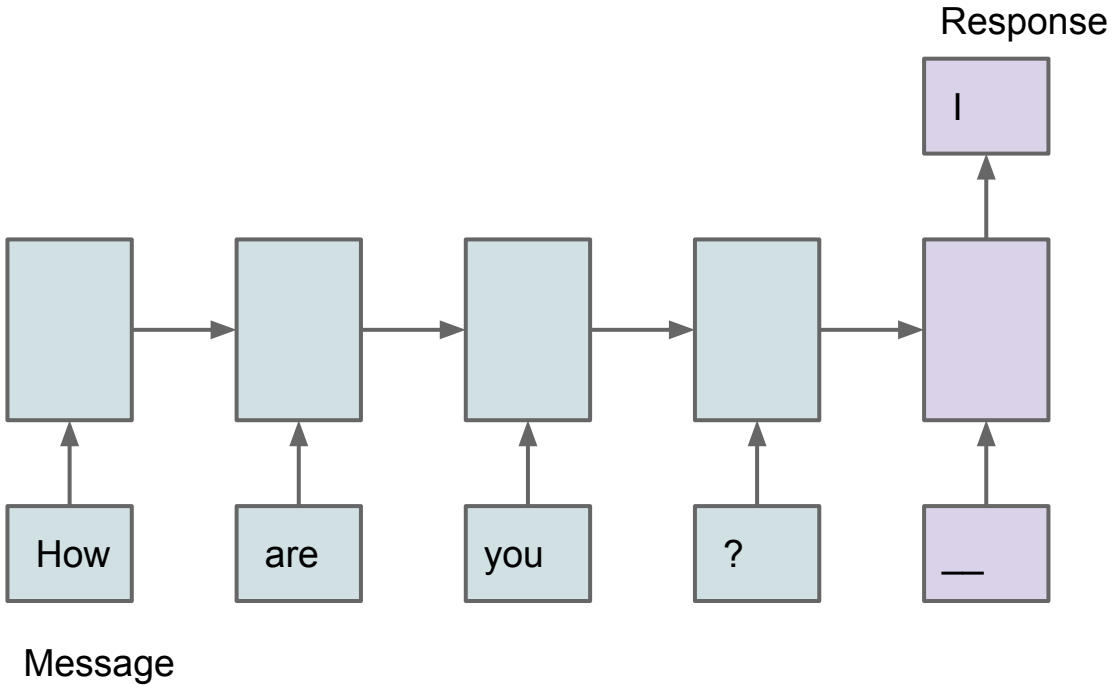
Decoder is initialized with final state of encoder



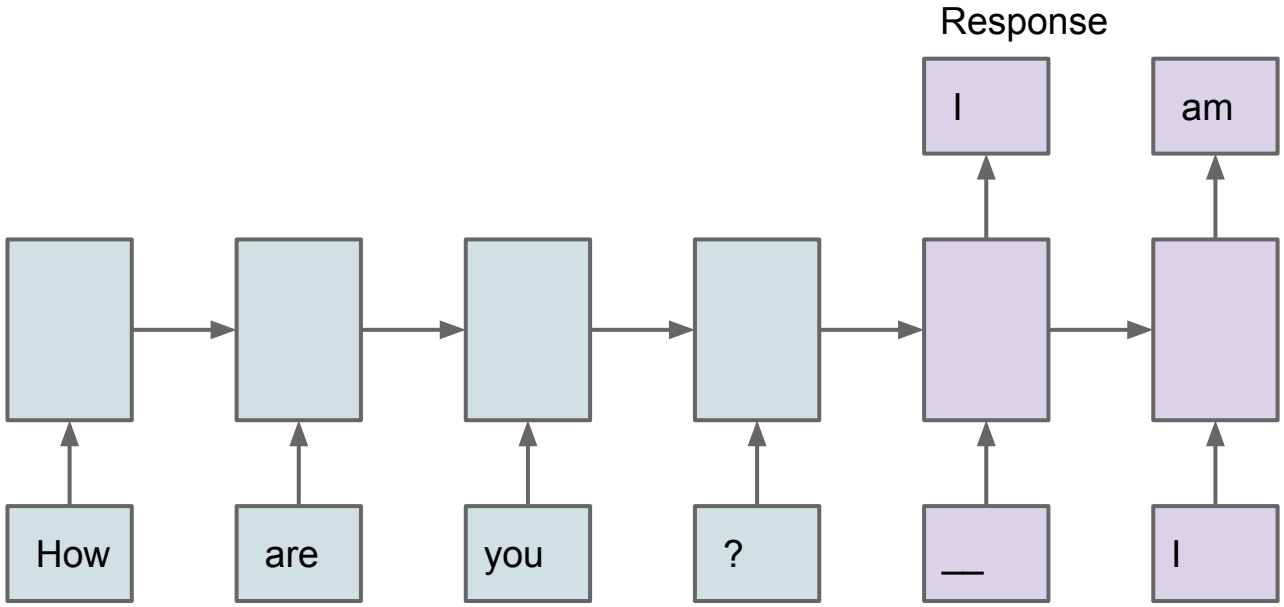
Decoder is initialized with final state of encoder



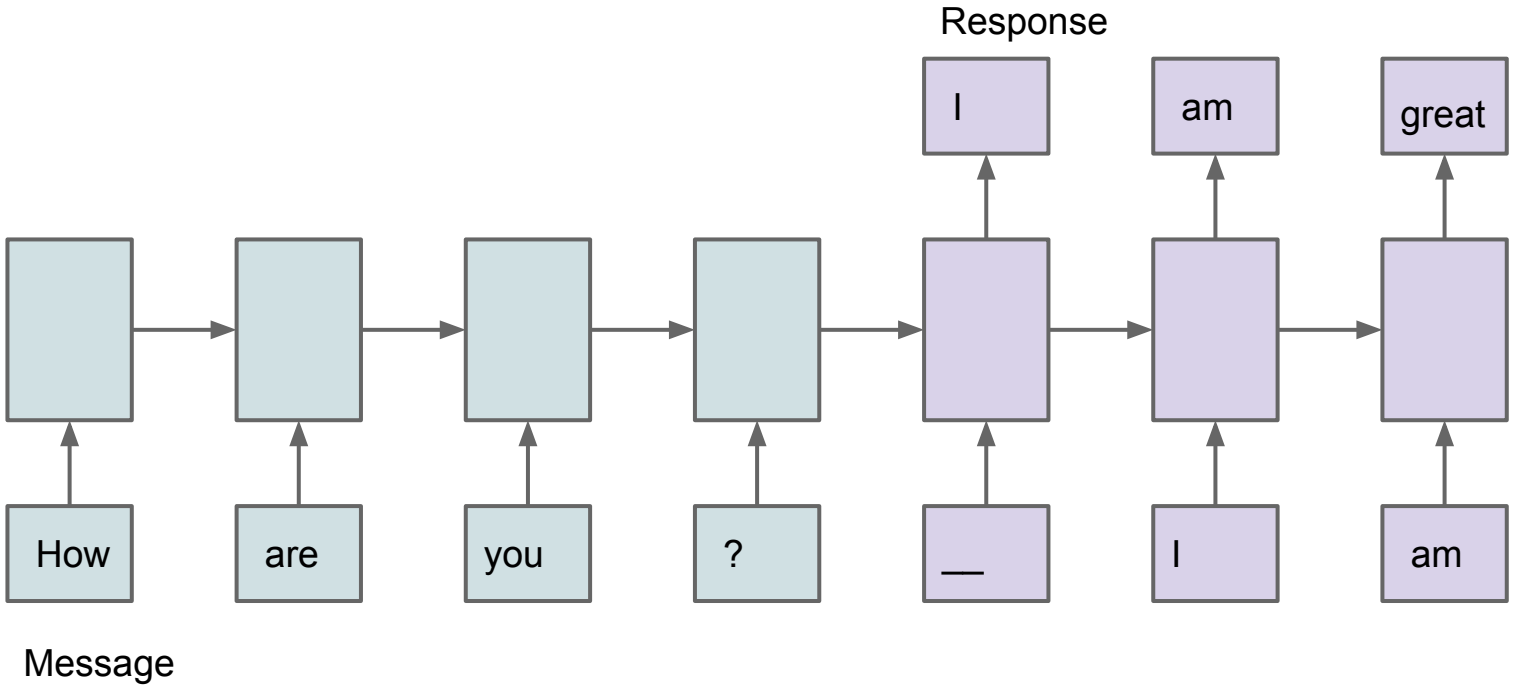
Decoder predicts next word



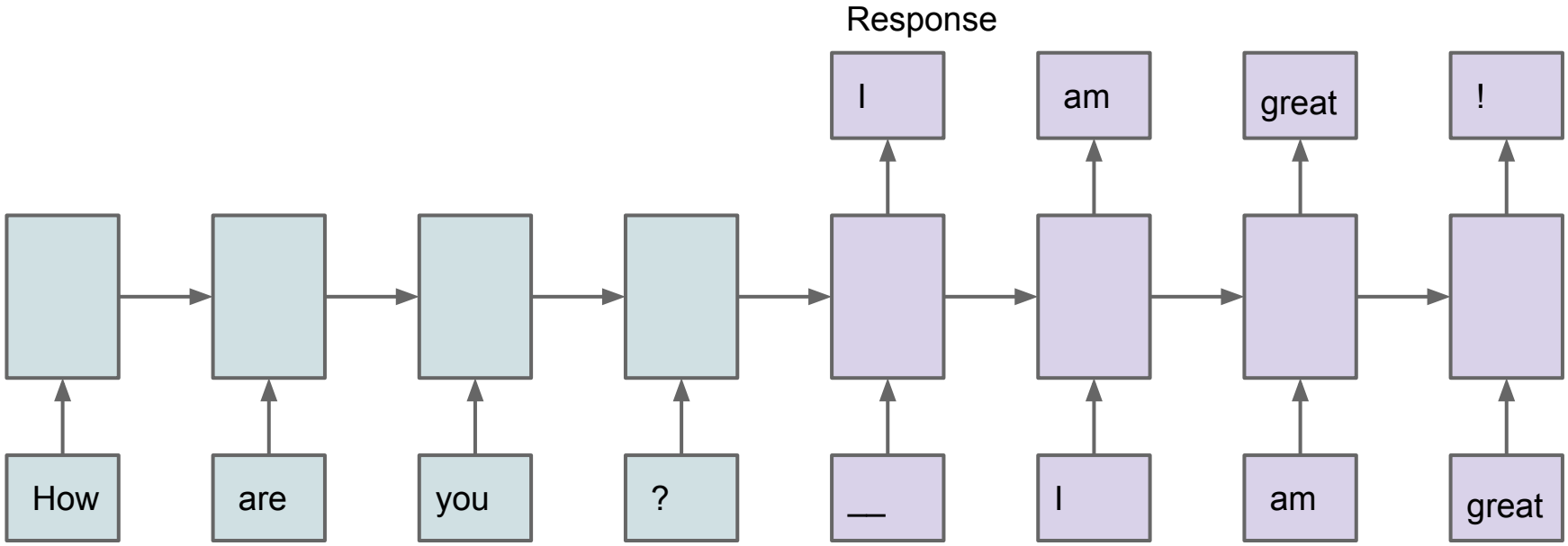
Decoder predicts next word



Decoder predicts next word



Decoder predicts next word



Vinyals & Le, ICML DL 2015
Kannan et al, KDD 2016

What the model can do

message = Can you do Tuesday or Wednesday?

responses:

```
I can do Tuesday
I can do Wednesday
I can do Wednesday .
I can do Tuesday .
Tuesday works for me .
Wednesday works for me .
Tuesday works for me
Tuesday works .
Tuesday works for me !
Yes I can
```

What the model can do

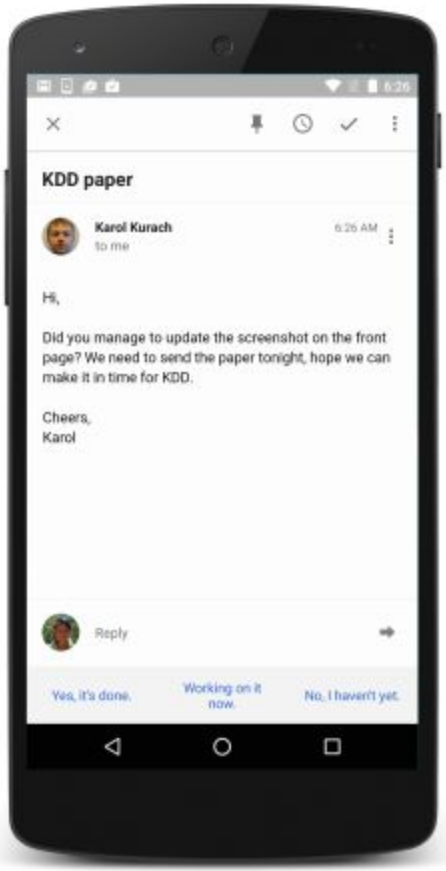
message = I feel so gross this morning. I think I ate something bad last night. Feel like I'm going to barf.

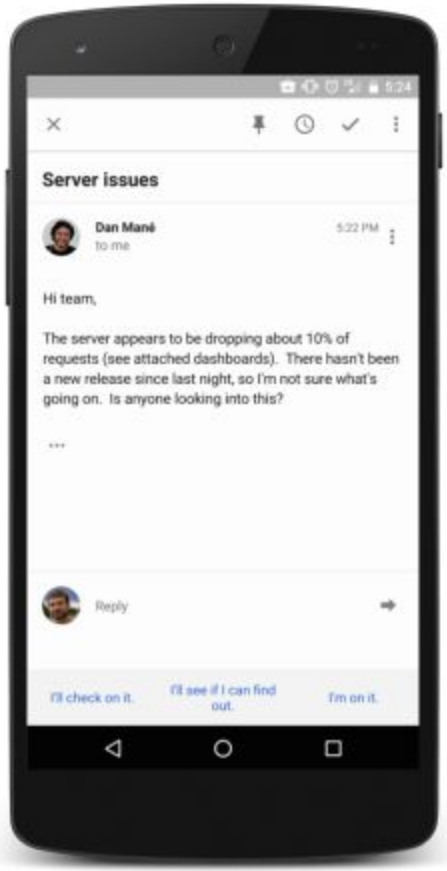
```
responses:

Feel better !
What 's wrong ?
I 'm sorry .
I 'm sorry :(
I 'm sorry. :(
Feel better .
Oh no !
What did you eat ?
I 'm sorry
That sucks .
```

Summary

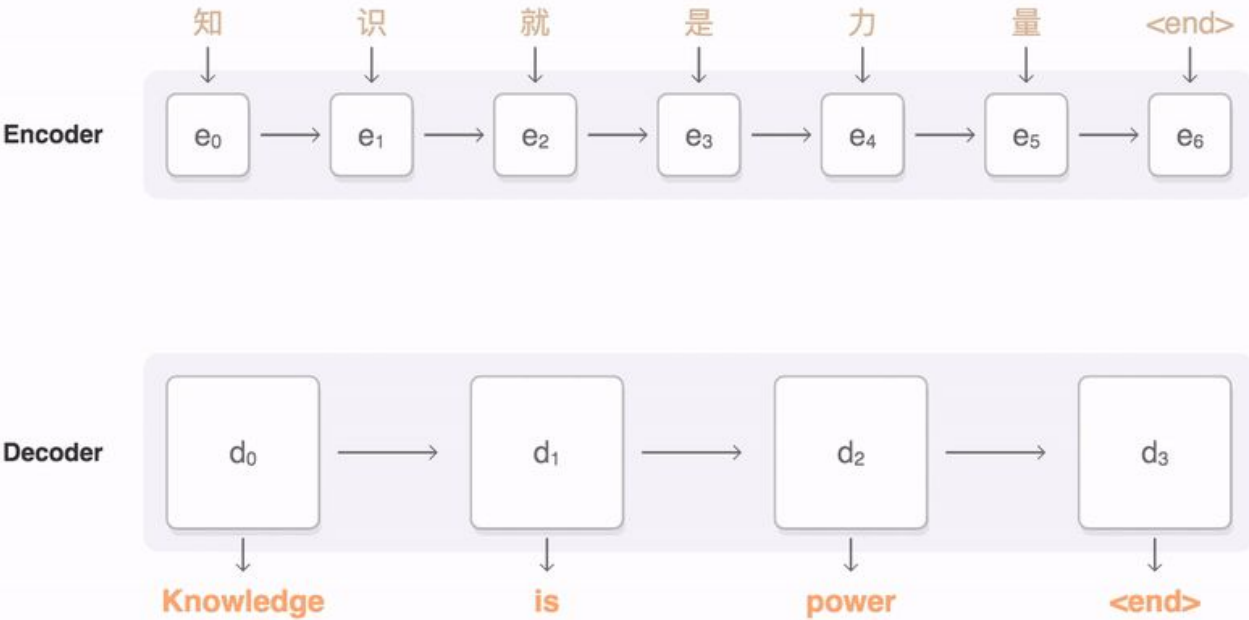
- Neural networks learn **feature representations** from raw data
- **Recurrent neural networks** have statefulness which allows them to model sequences of data such as text
- The **sequence-to-sequence** model contains two recurrent neural networks: one to encode an input sequence and one to generate an output sequence

Smartreply


KDD paper
Karol Kurach
to me
6:26 AM
Hi,
Did you manage to update the screenshot on the front page? We need to send the paper tonight, hope we can make it in time for KDD.
Cheers,
Karol
Reply
Yes, it's done.
Working on it now.
No, I haven't yet.


Server issues
Dan Mané
to me
5:22 PM
Hi team,
The server appears to be dropping about 10% of requests (see attached dashboards). There hasn't been a new release since last night, so I'm not sure what's going on. Is anyone looking into this?
Reply
I'll check on it.
I'll see if I can find out.
I'm on it.

Google Translate



Authors

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, Jeffrey Dean

Research: Speech recognition

STATE-OF-THE-ART SPEECH RECOGNITION WITH SEQUENCE-TO-SEQUENCE MODELS

Chung-Cheng Chiu, Tara N. Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J. Weiss, Kanishka Rao, Ekaterina Gonina, Navdeep Jaitly, Bo Li, Jan Chorowski, Michiel Bacchiani

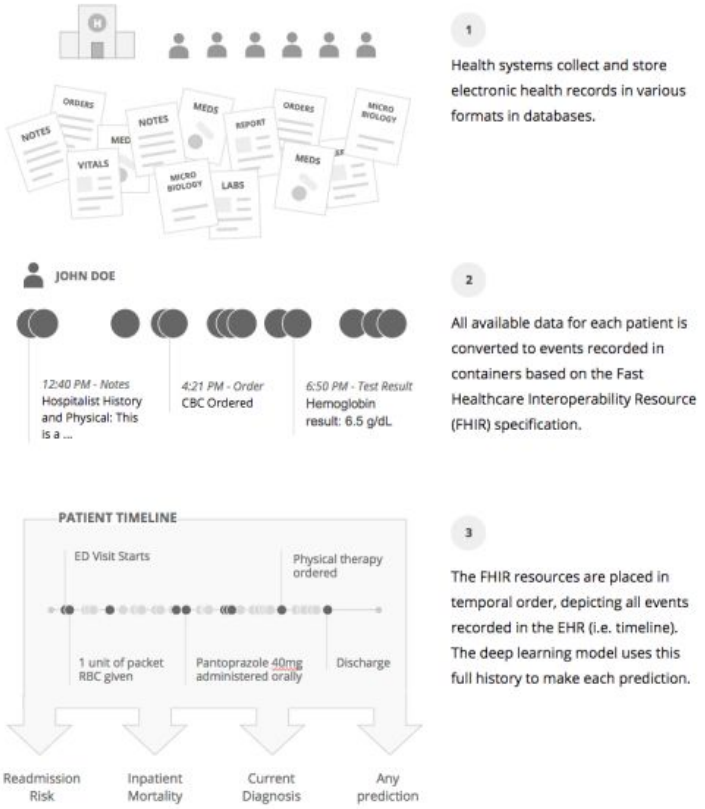
Research: Electronic health records

Scalable and accurate deep learning for electronic health records

Alvin Rajkomar*[1,2], Eyal Oren*[1], Kai Chen[1], Andrew M. Dai[1], Nissan Hajaj[1], Peter J. Liu[1], Xiaobing Liu[1], Mimi Sun[1], Patrik Sundberg[1], Hector Yee[1], Kun Zhang [1], Yi Zhang[1], Gavin E. Duggan[1], Gerardo Flores[1], Michaela Hardt[1], Jamie Irvine[1], Quoc Le[1], Kurt Litsch[1], Jake Marcus[1], Alexander Mossin[1], Justin Tansuwan[1], De Wang[1], James Wexler[1], Jimbo Wilson[1], Dana Ludwig[2], Samuel L. Volchenboum[4], Katherine Chou[1], Michael Pearson[1], Srinivasan Madabushi[1], Nigam H. Shah[3], Atul J. Butte[2], Michael Howell[1], Claire Cui[1], Greg Corrado[1], and Jeff Dean[1]

[1]Google Inc, Mountain View, California
[2]University of California, San Francisco, San Francisco, California
[3]Stanford University, Stanford, California
[4]University of Chicago Medicine, Chicago, Illinois

January 2018



Google

What's next?

Resources

- All tensorflow tutorials: https://www.tensorflow.org/versions/master/tutorials/index.html
- Sequence-to-sequence tutorial (machine translation): https://www.tensorflow.org/versions/master/tutorials/seq2seq
- Chris Olah's blog: http://colah.github.io/

Thank you!