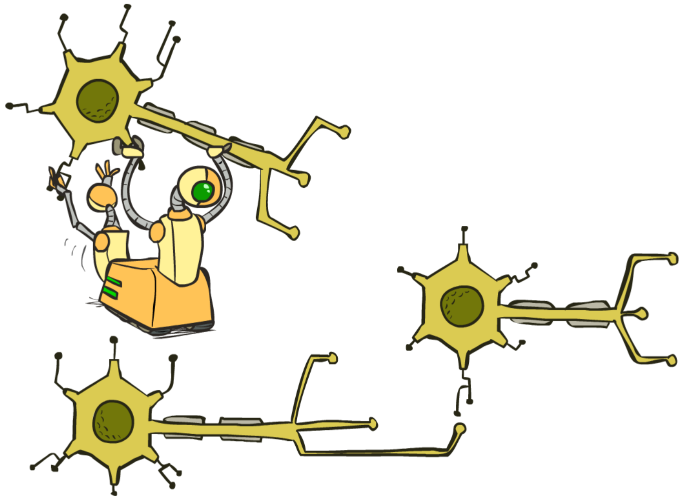
CSE 473: Artificial Intelligence

Hanna Hajishirzi

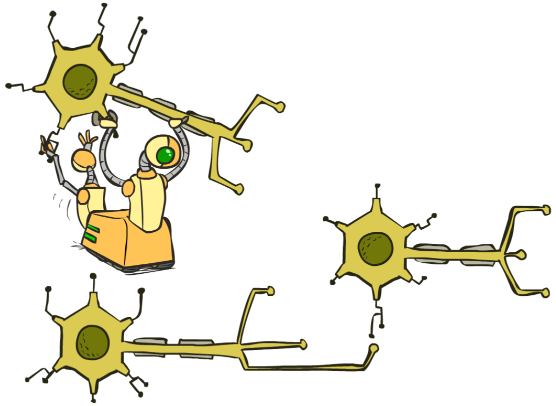
Neural Nets



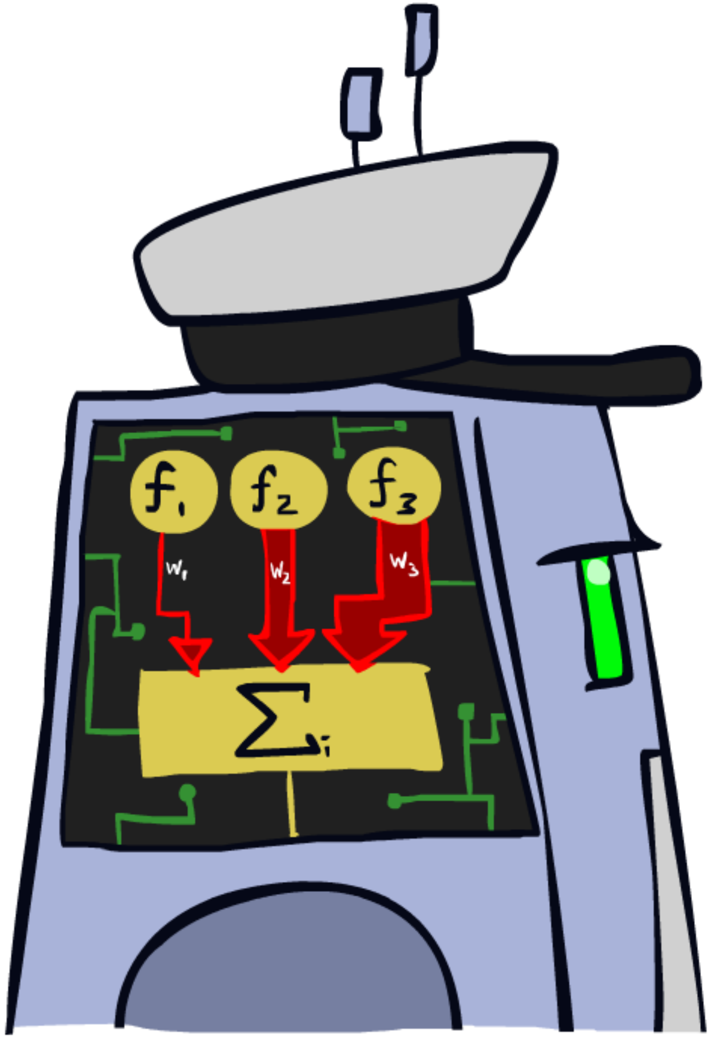
slides adapted from
Dan Klein, Pieter Abbeel ai.berkeley.edu
And Dan Weld, Luke Zettlemoyer

Trend in NLP

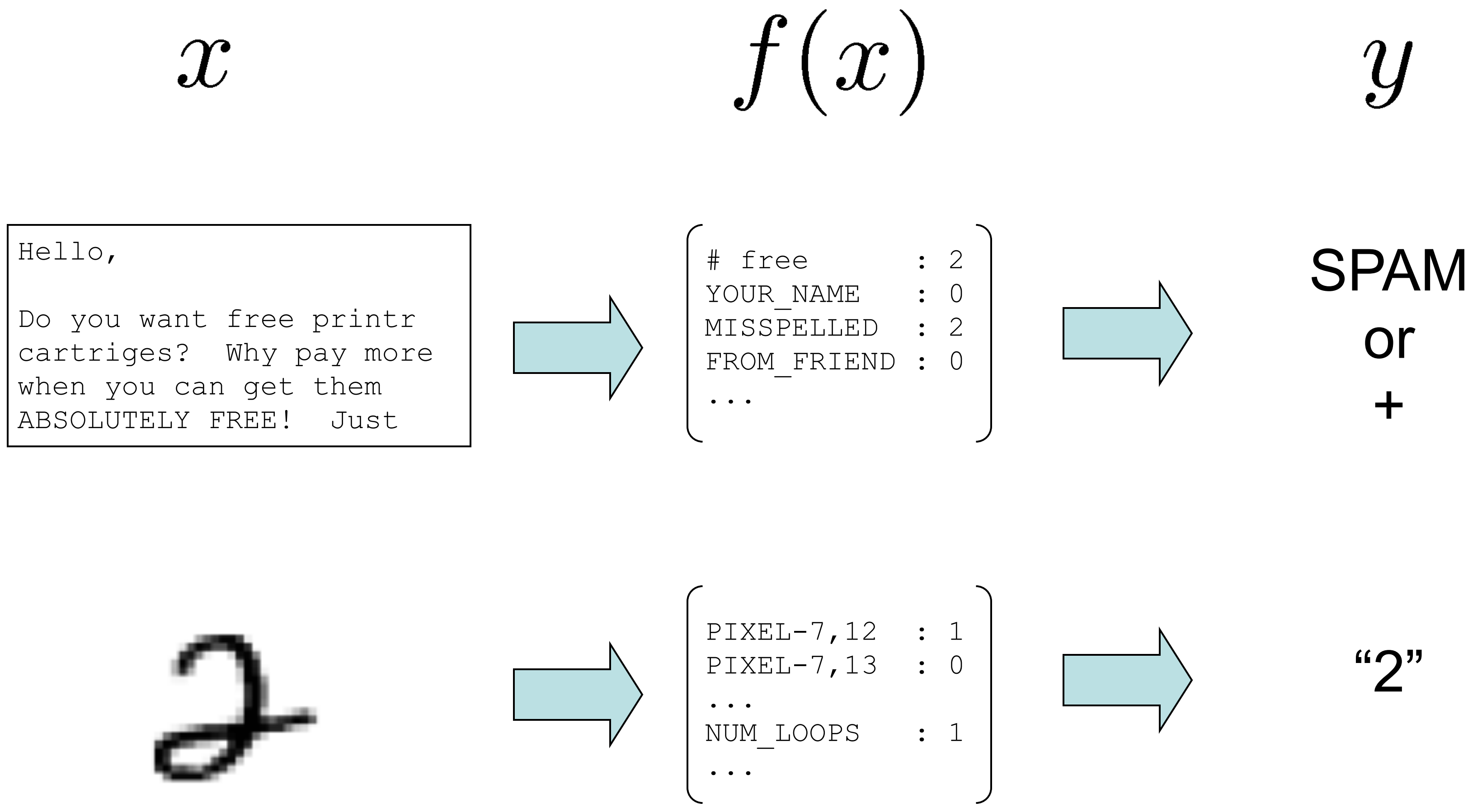
- Over time:
  - Learning HMMs (or related Probabilistic-based methods) with hand-designed features (tokens, syntactic features)
  - Recurrent Neural Networks:
    - replaces probabilistic dynamic model with neural functions (mostly non-linear functions)
  - Attention-based methods:
    - Adds the capability to go beyond Markov Models



Linear Classifiers



Feature Vectors



x $f(x)$ y

```
Hello,

Do you want free printr
cartriges?  Why pay more
when you can get them
ABSOLUTELY FREE!  Just
```

```
# free        : 2
YOUR_NAME     : 0
MISSPELLED    : 2
FROM_FRIEND   : 0
...
```

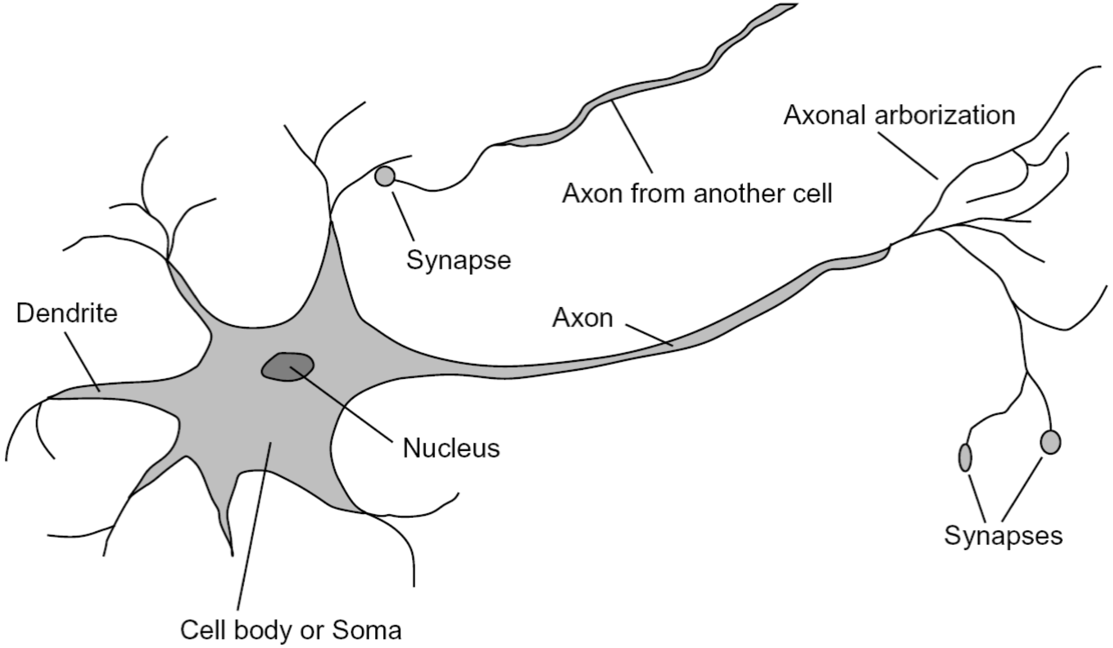
SPAM
or
+

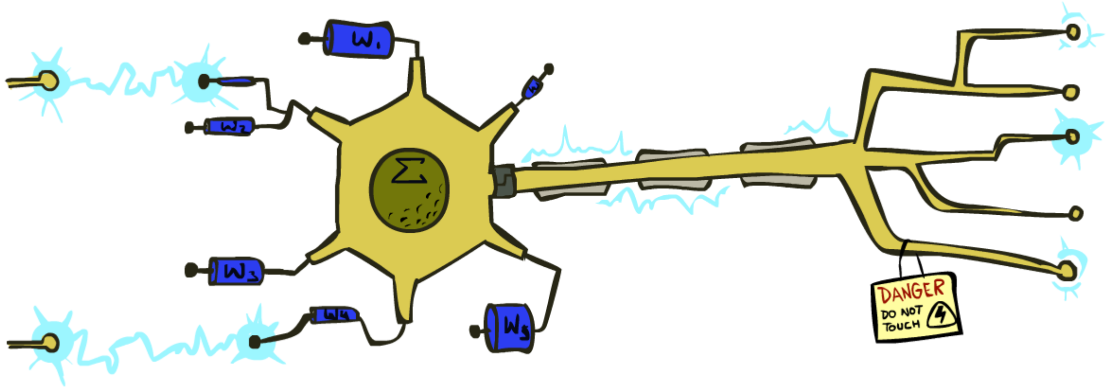
```
PIXEL-7,12   : 1
PIXEL-7,13   : 0
...
NUM_LOOPS    : 1
...
```

“2”

Some (Simplified) Biology

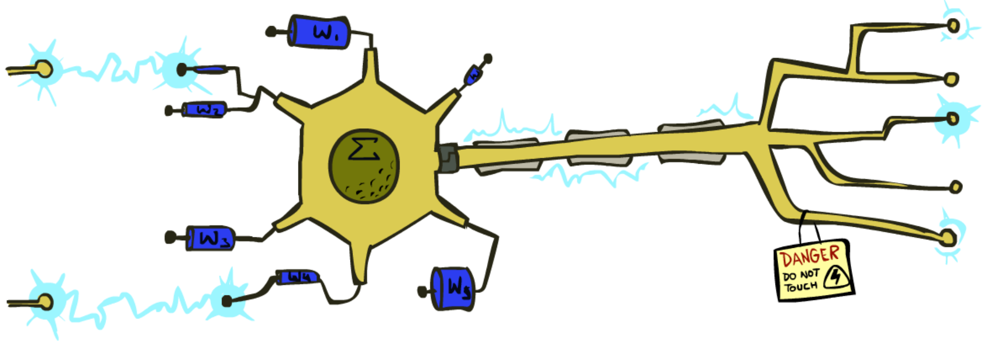
- Very loose inspiration: human neurons





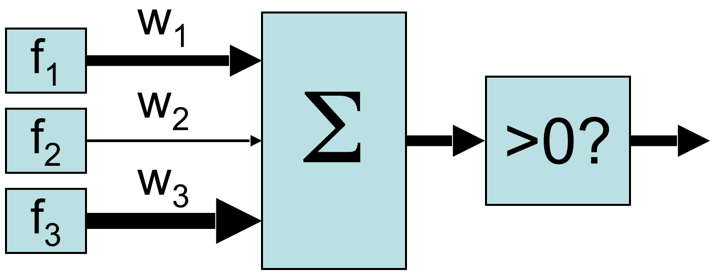
Linear Classifiers

- Inputs are feature values
- Each feature has a weight
- Sum is the activation



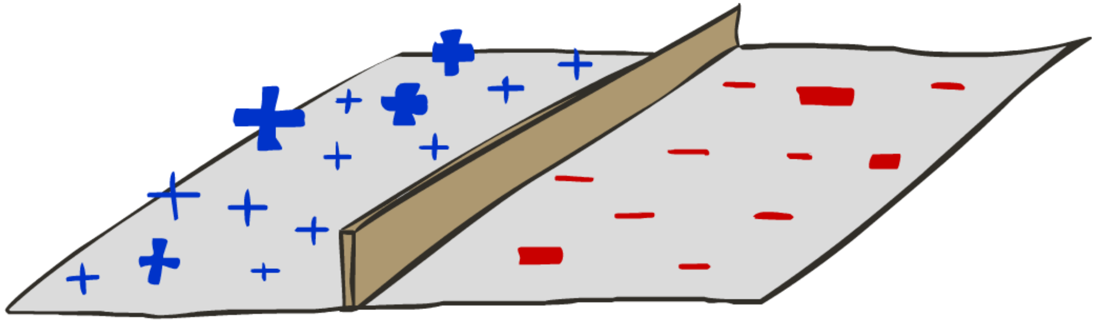
$$\text{activation}_w(x) = \sum_i w_i \cdot f_i(x) = w \cdot f(x)$$

- If the activation is:
  - Positive, output +1
  - Negative, output -1



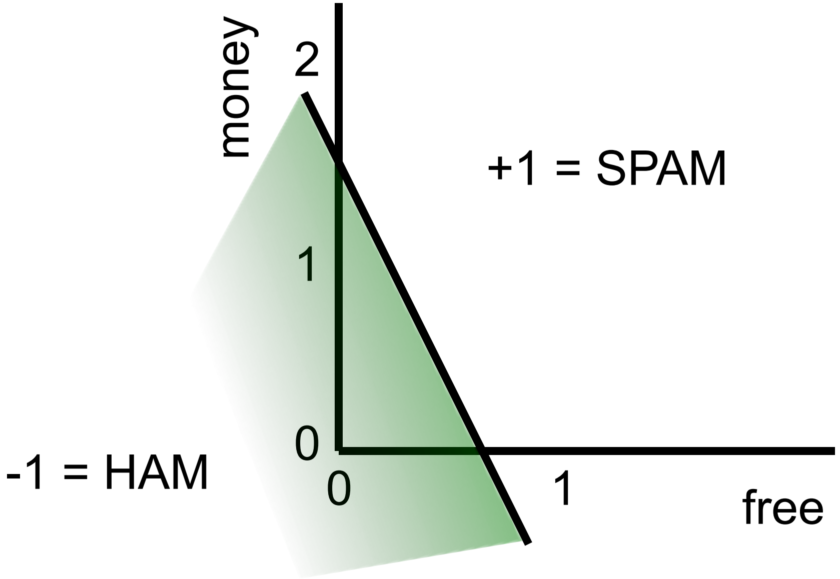
Binary Decision Rule

- In the space of feature vectors
  - Examples are points
  - Any weight vector is a hyperplane
  - One side corresponds to Y=+1
  - Other corresponds to Y=-1



$$w$$

```
BIAS  : -3
free  :  4
money :  2
...
```



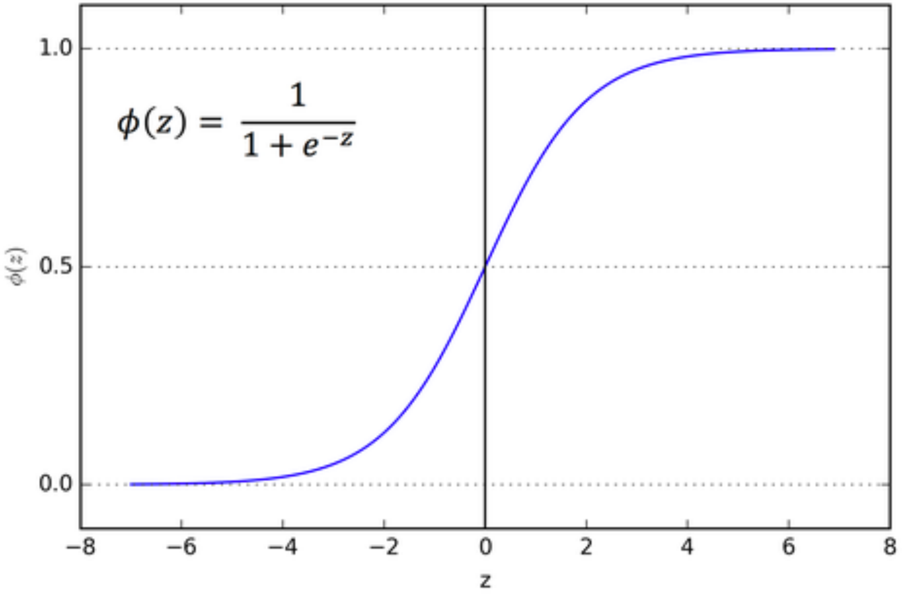
$$f \cdot w = 0$$

How to get probabilistic decisions?

- Activation: $z = w \cdot f(x)$
- If $z = w \cdot f(x)$ very positive → want probability going to 1
- If $z = w \cdot f(x)$ very negative → want probability going to 0

- Sigmoid function

$$\phi(z) = \frac{1}{1 + e^{-z}}$$



Best w?

- Maximum likelihood estimation:

$$\max_{w} \quad ll(w) = \max_{w} \quad \sum_{i} \log P(y^{(i)}|x^{(i)};w)$$

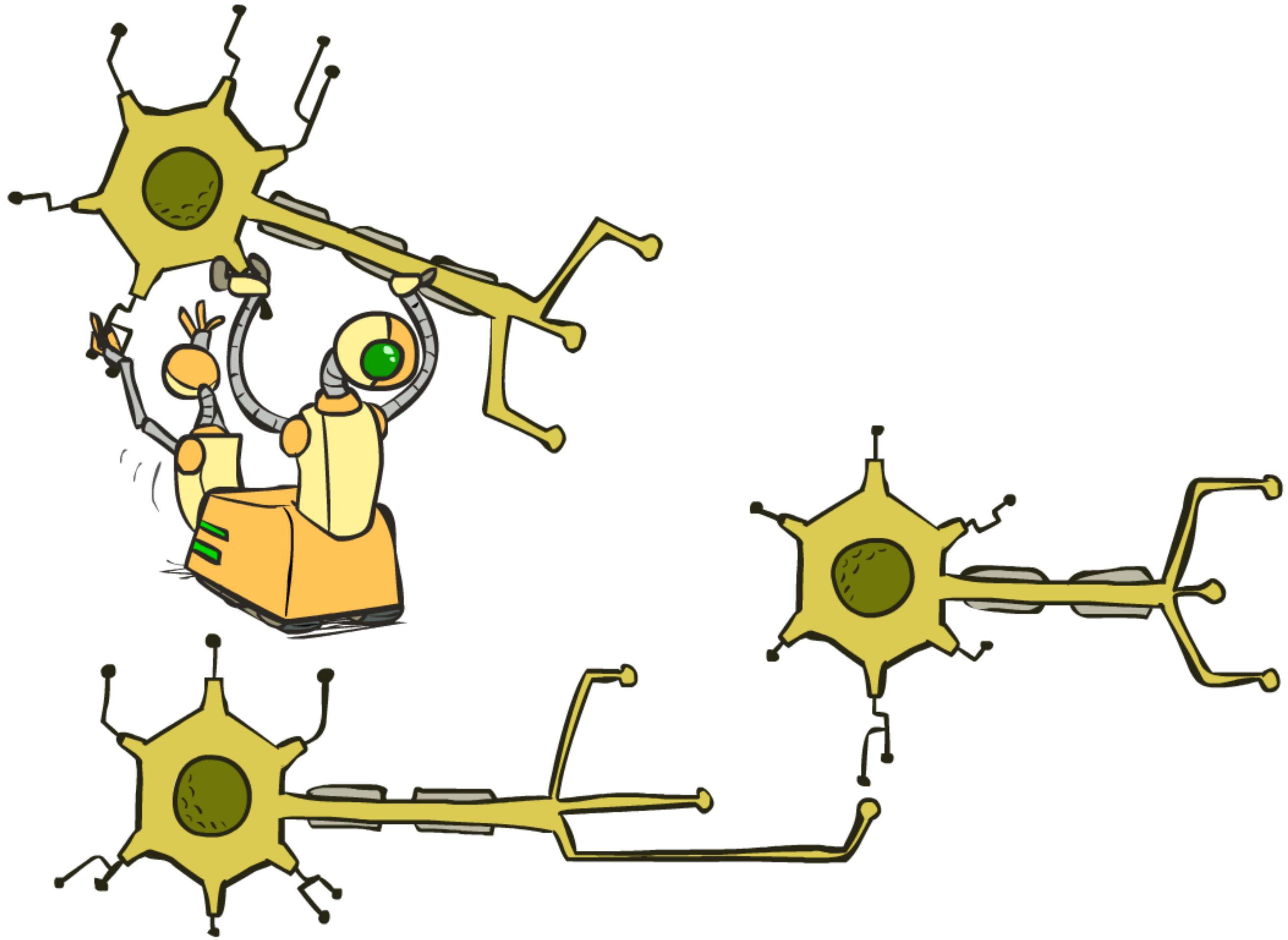
with:

$$P(y^{(i)} = +1|x^{(i)};w) = \frac{1}{1+e^{-w\cdot f(x^{(i)})}}$$

$$P(y^{(i)} = -1|x^{(i)};w) = 1 - \frac{1}{1+e^{-w\cdot f(x^{(i)})}}$$

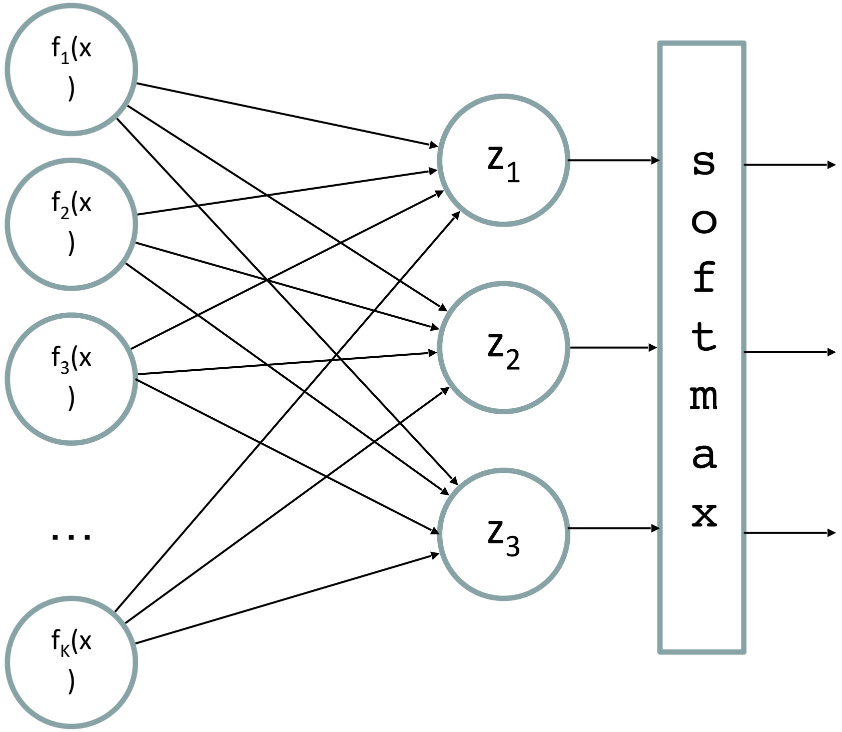
= Logistic Regression

Neural Networks



Multi-class Logistic Regression

- = special case of neural network

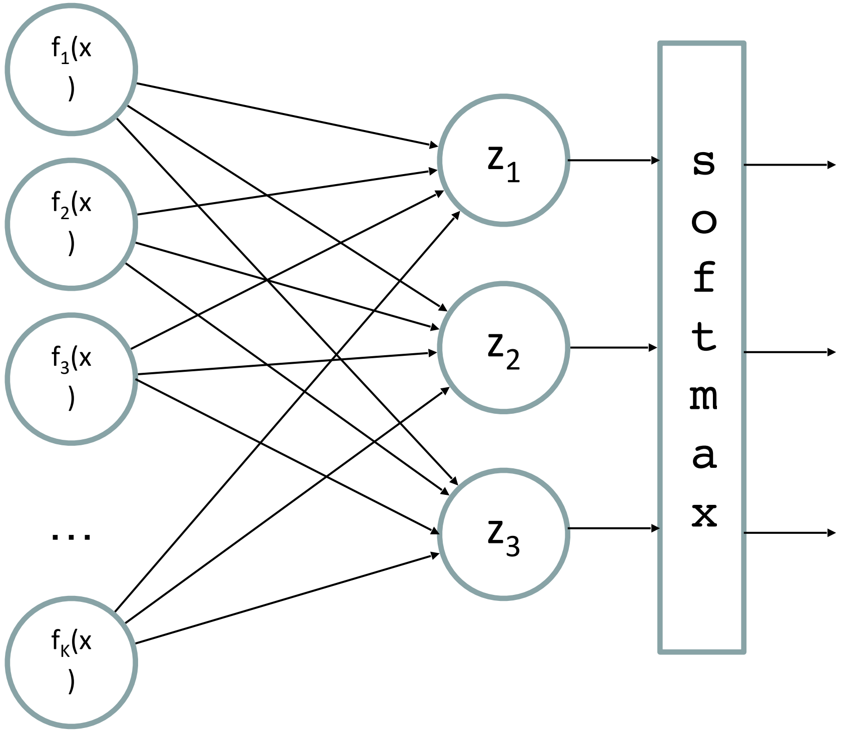


$$P(y_1|x;w) = \frac{e^{z_1}}{e^{z_1} + e^{z_2} + e^{z_3}}$$

$$P(y_2|x;w) = \frac{e^{z_2}}{e^{z_1} + e^{z_2} + e^{z_3}}$$

$$P(y_3|x;w) = \frac{e^{z_3}}{e^{z_1} + e^{z_2} + e^{z_3}}$$

Deep Neural Network = Also learn the features!

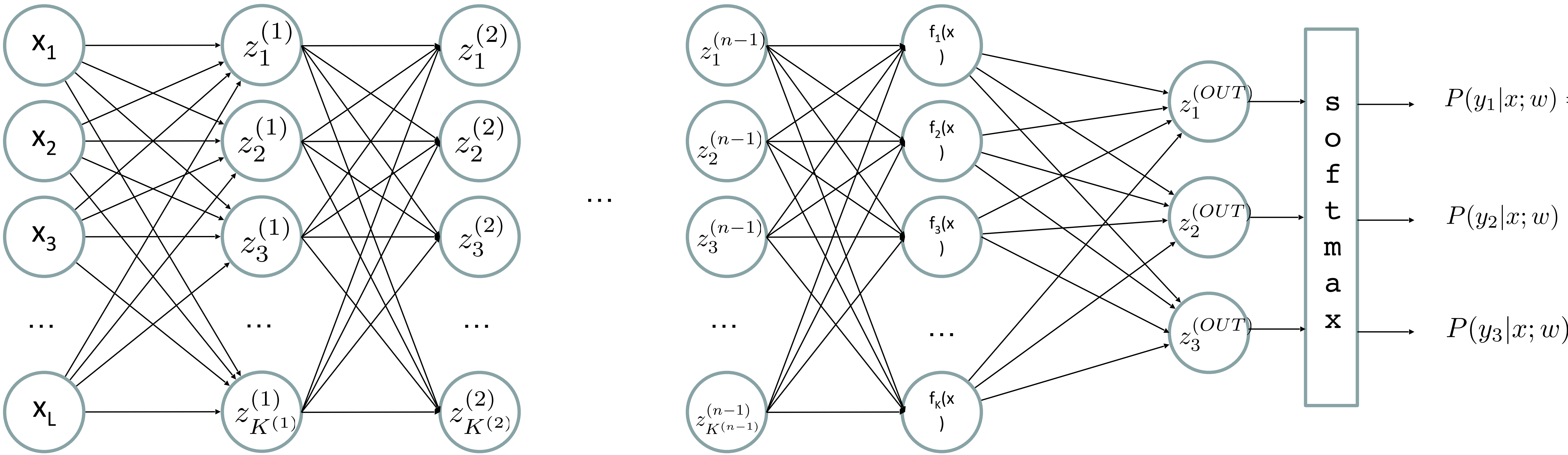


$$P(y_1|x;w) = \frac{e^{z_1}}{e^{z_1} + e^{z_2} + e^{z_3}}$$

$$P(y_2|x;w) = \frac{e^{z_2}}{e^{z_1} + e^{z_2} + e^{z_3}}$$

$$P(y_3|x;w) = \frac{e^{z_3}}{e^{z_1} + e^{z_2} + e^{z_3}}$$

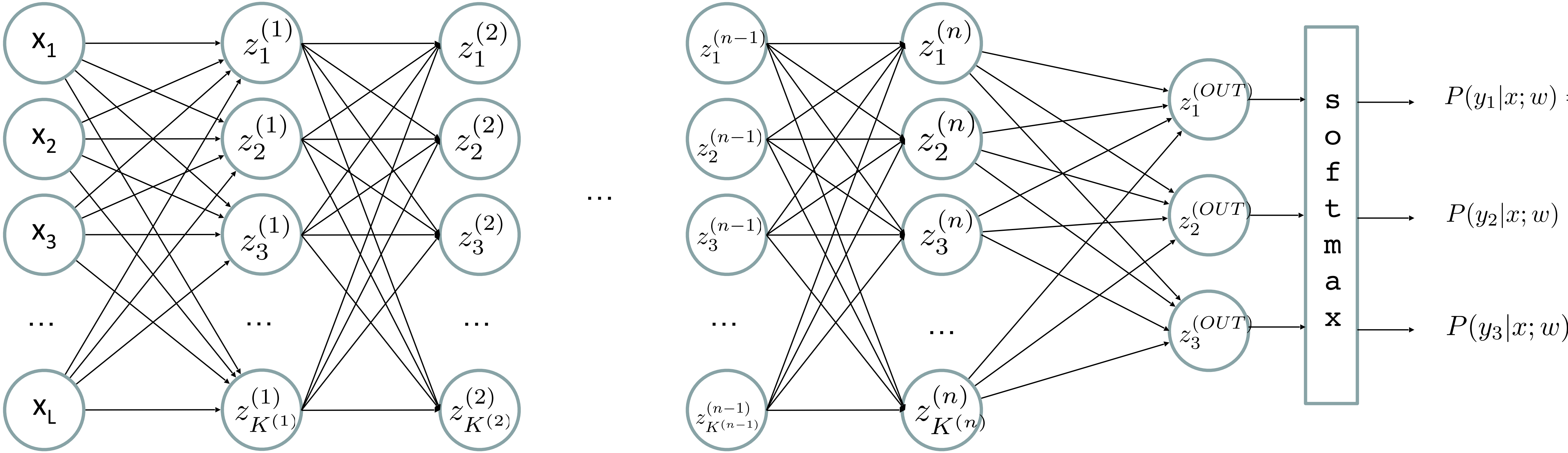
Deep Neural Network = Also learn the features!



$$z_i^{(k)} = g(\sum_j W_{i,j}^{(k-1,k)} z_j^{(k-1)})$$

g = nonlinear activation function

Deep Neural Network = Also learn the features!



$$z_i^{(k)} = g(\sum_j W_{i,j}^{(k-1,k)} z_j^{(k-1)})$$

g = nonlinear activation function

Deep Neural Network: Also Learn the Features!

- Training the deep neural network is just like logistic regression:

$$\max_w \quad ll(w) = \max_w \quad \sum_i \log P(y^{(i)} | x^{(i)}; w)$$

Neural Networks Properties

- Theorem (Universal Function Approximators). A two-layer neural network with a sufficient number of neurons can approximate any continuous function to any desired accuracy.

- Practical considerations
  - Can be seen as learning the features
  - Large number of neurons
    - Danger for overfitting
    - (hence early stopping!)

Fun Neural Net Demo Site

- Demo-site:
  - http://playground.tensorflow.org/

Automatic Differentiation

- Automatic differentiation software
  - e.g. Theano, TensorFlow, PyTorch, Chainer
  - Only need to program the function g(x,y,w)
  - Can automatically compute all derivatives w.r.t. all entries in w
  - This is typically done by caching info during forward computation pass of f, and then doing a backward pass = “backpropagation”
  - Autodiff / Backpropagation can often be done at computational cost comparable to the forward pass
- Need to know this exists
- How this is done? -- outside of scope of CSE573

Summary of Key Ideas

- Optimize probability of label given input $\max_w \quad ll(w) = \max_w \quad \sum_i \log P(y^{(i)}|x^{(i)}; w)$

- Continuous optimization
  - Gradient ascent:
    - Compute steepest uphill direction = gradient (= just vector of partial derivatives)
    - Take step in the gradient direction
    - Repeat (until held-out data accuracy starts to drop = "early stopping")

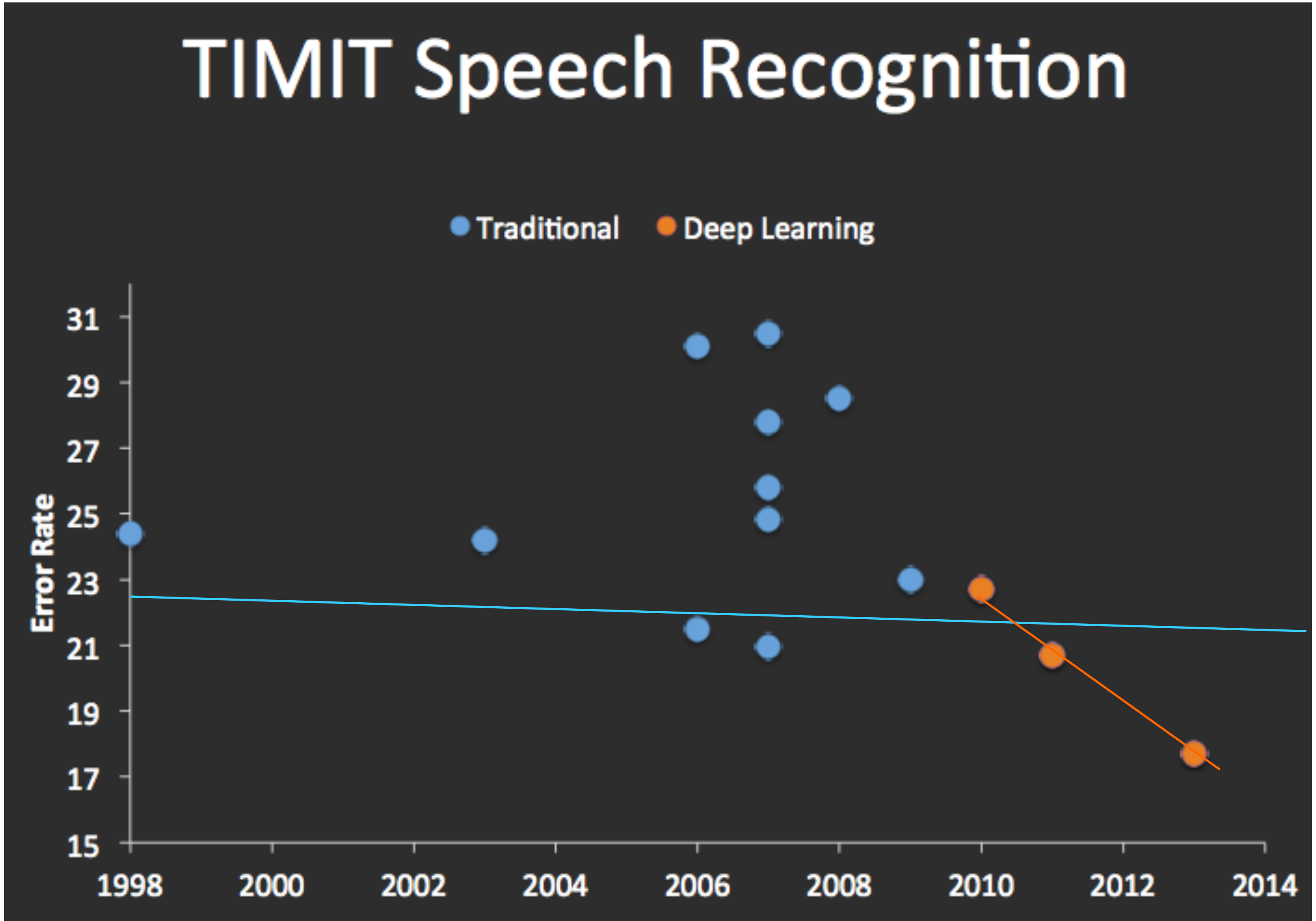
- Deep neural nets
  - Last layer = still logistic regression
  - Now also many more layers before this last layer
    - = computing the features
    - → the features are learned rather than hand-designed
  - Universal function approximation theorem
    - `If` neural net is large enough
    - `Then` neural net can represent any continuous mapping from input to output with arbitrary accuracy
    - But remember: need to avoid overfitting / memorizing the training data → early stopping!
  - Automatic differentiation gives the derivatives efficiently (how? = outside of scope of 473)

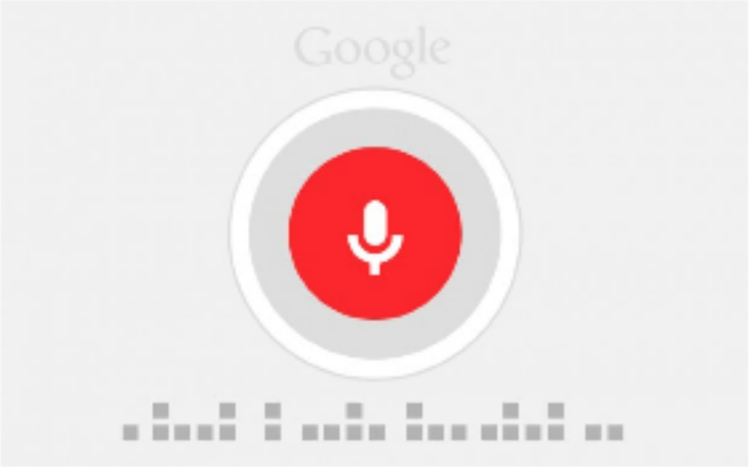
How well does it work?

Speech and Natural Language Processing

- Different approaches to:
  - Modeling sequences of tokens
- Language Modeling: $P(x_t|x_{t-1})$
- Applications:
  - Machine Translation
  - Document Classification
    - Sentiment
    - Document types
  - Question Answering
  - etc

Speech Recognition

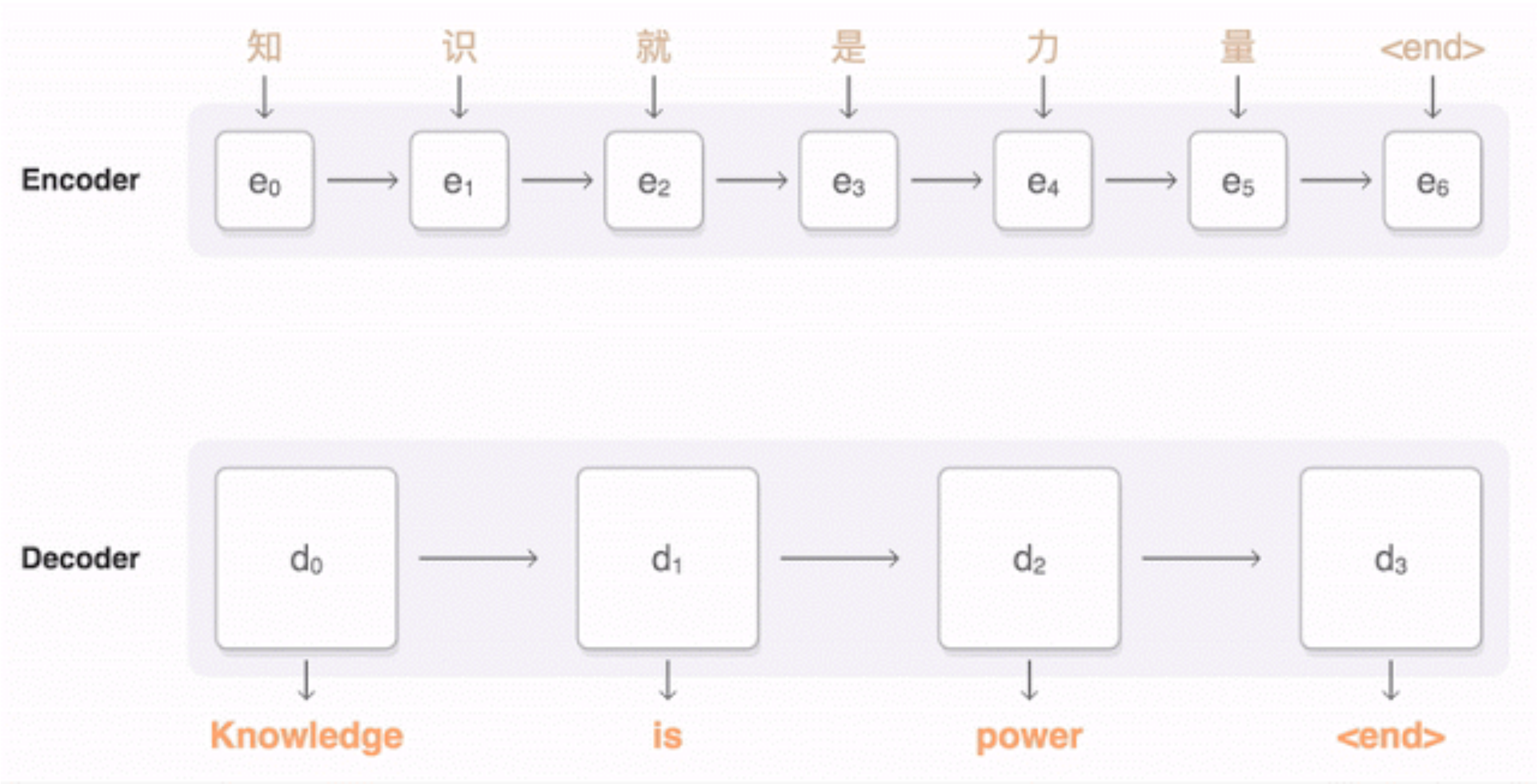




graph credit Matt Zeiler, Clarifai

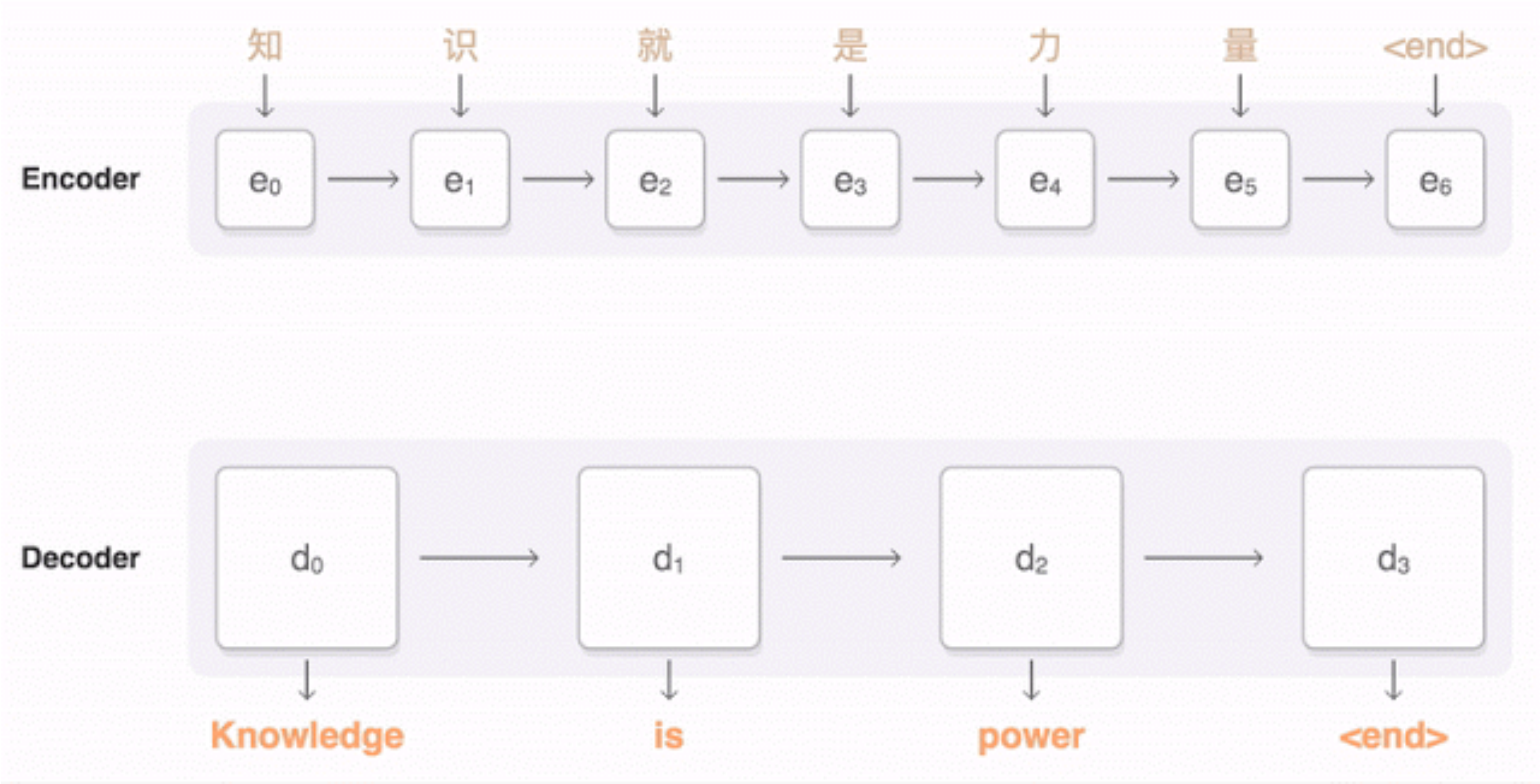
Machine Translation

Google Neural Machine Translation (in production)



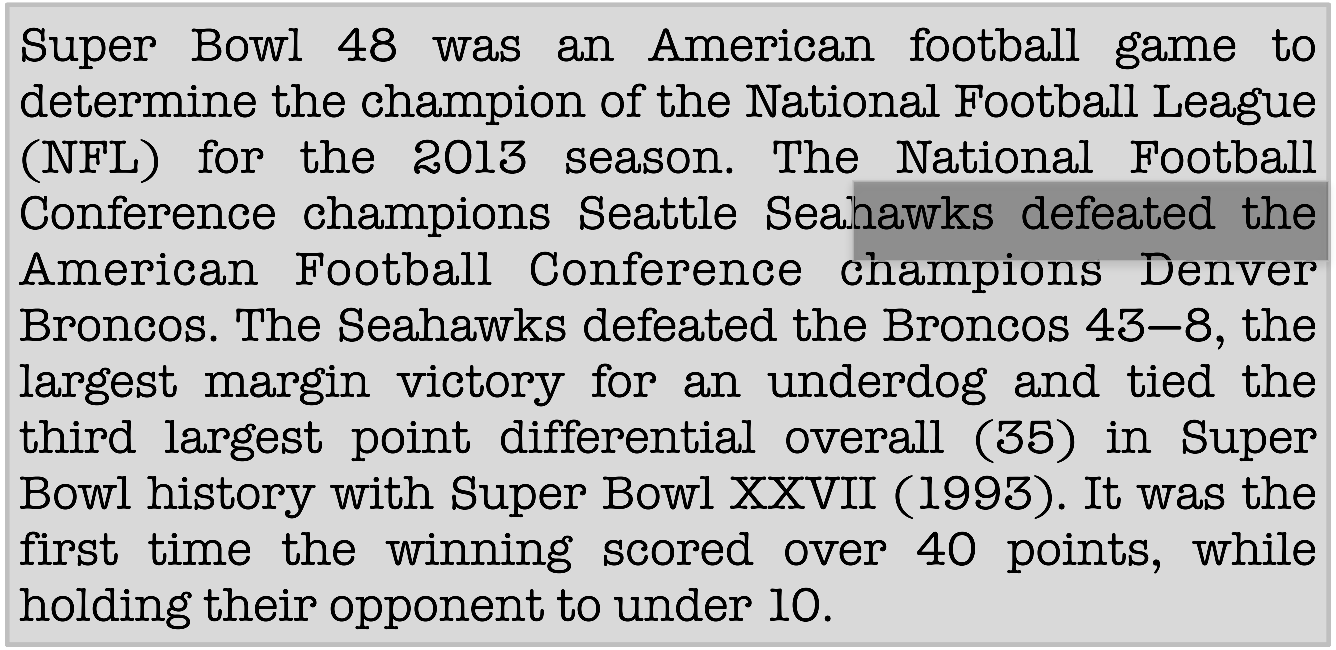
Machine Translation

Google Neural Machine Translation (in production)



Question Answering

Context



Super Bowl 48 was an American football game to determine the champion of the National Football League (NFL) for the 2013 season. The National Football Conference champions Seattle Seahawks defeated the American Football Conference champions Denver Broncos. The Seahawks defeated the Broncos 43–8, the largest margin victory for an underdog and tied the third largest point differential overall (35) in Super Bowl history with Super Bowl XXVII (1993). It was the first time the winning scored over 40 points, while holding their opponent to under 10.

Question

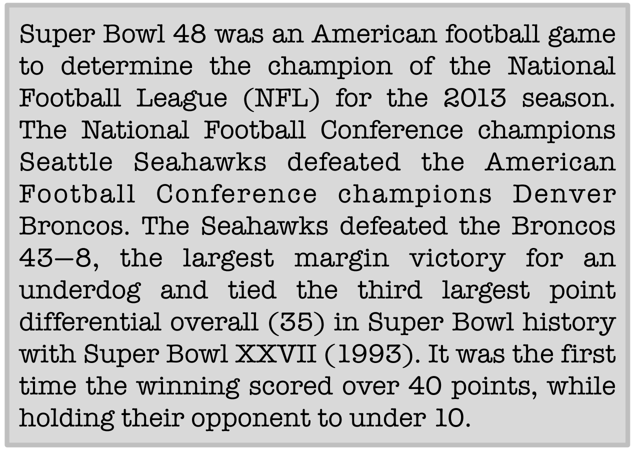
Which NFL team represented the NFC at Super Bowl 48?

Answer

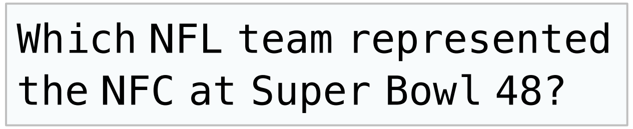
Seattle Seahawks

Pipeline Approach for Question Answering

- Feature engineering
- Classifying phrases



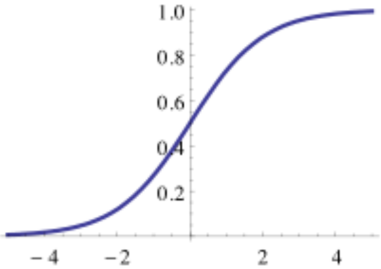
Super Bowl 48 was an American football game to determine the champion of the National Football League (NFL) for the 2013 season. The National Football Conference champions Seattle Seahawks defeated the American Football Conference champions Denver Broncos. The Seahawks defeated the Broncos 43–8, the largest margin victory for an underdog and tied the third largest point differential overall (35) in Super Bowl history with Super Bowl XXVII (1993). It was the first time the winning scored over 40 points, while holding their opponent to under 10.



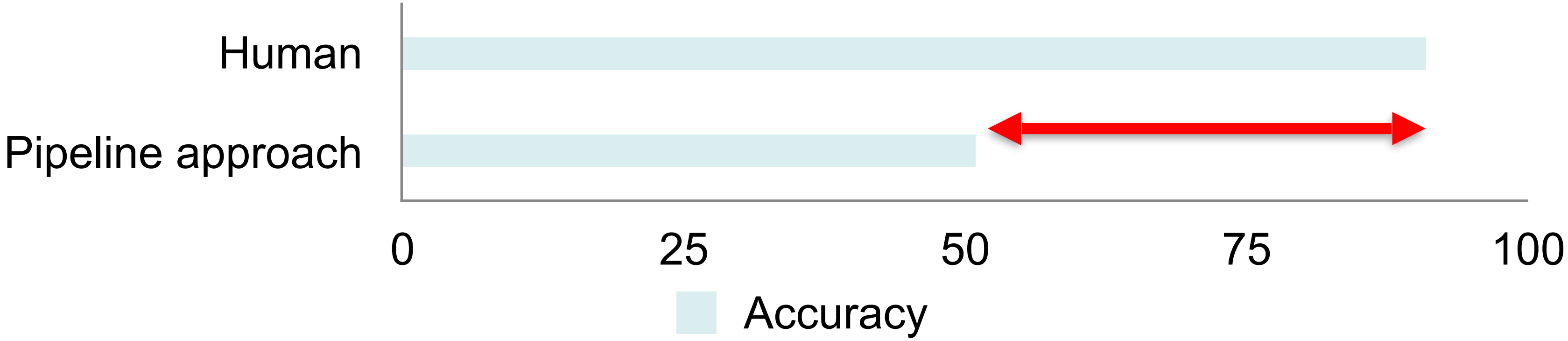
Which NFL team represented the NFC at Super Bowl 48?

words, types, frequencies
dependency relations

$$f_1, f_2, \ldots, f_n$$



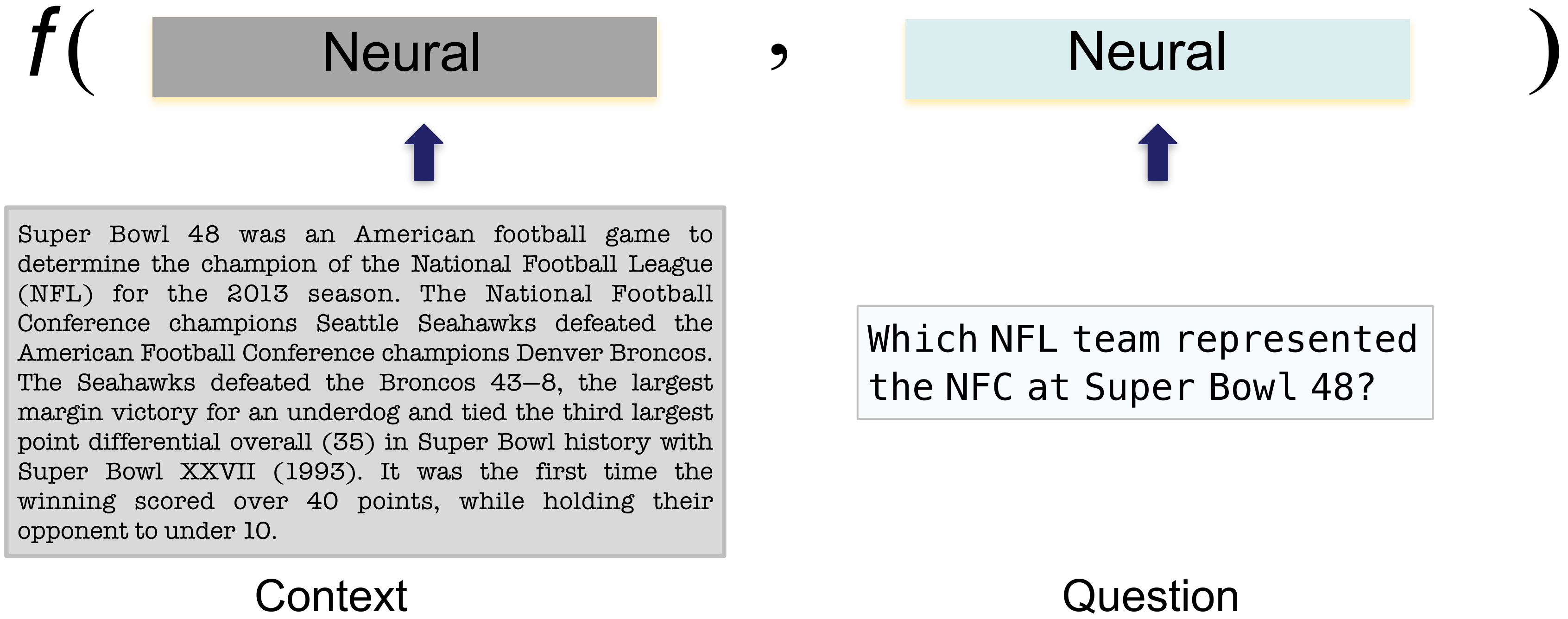
Pipeline Approach Results



- Dataset: Stanford Question Answering Dataset (SQuAD) [Rajpurkar et al 2016]:
  - 100k Wikipedia documents with question
- Accuracy: percentage of correctly predicted phrases

Neural Approach [ICLR’17]

Find a function that assigns a high score to the the correct answer given the context and question

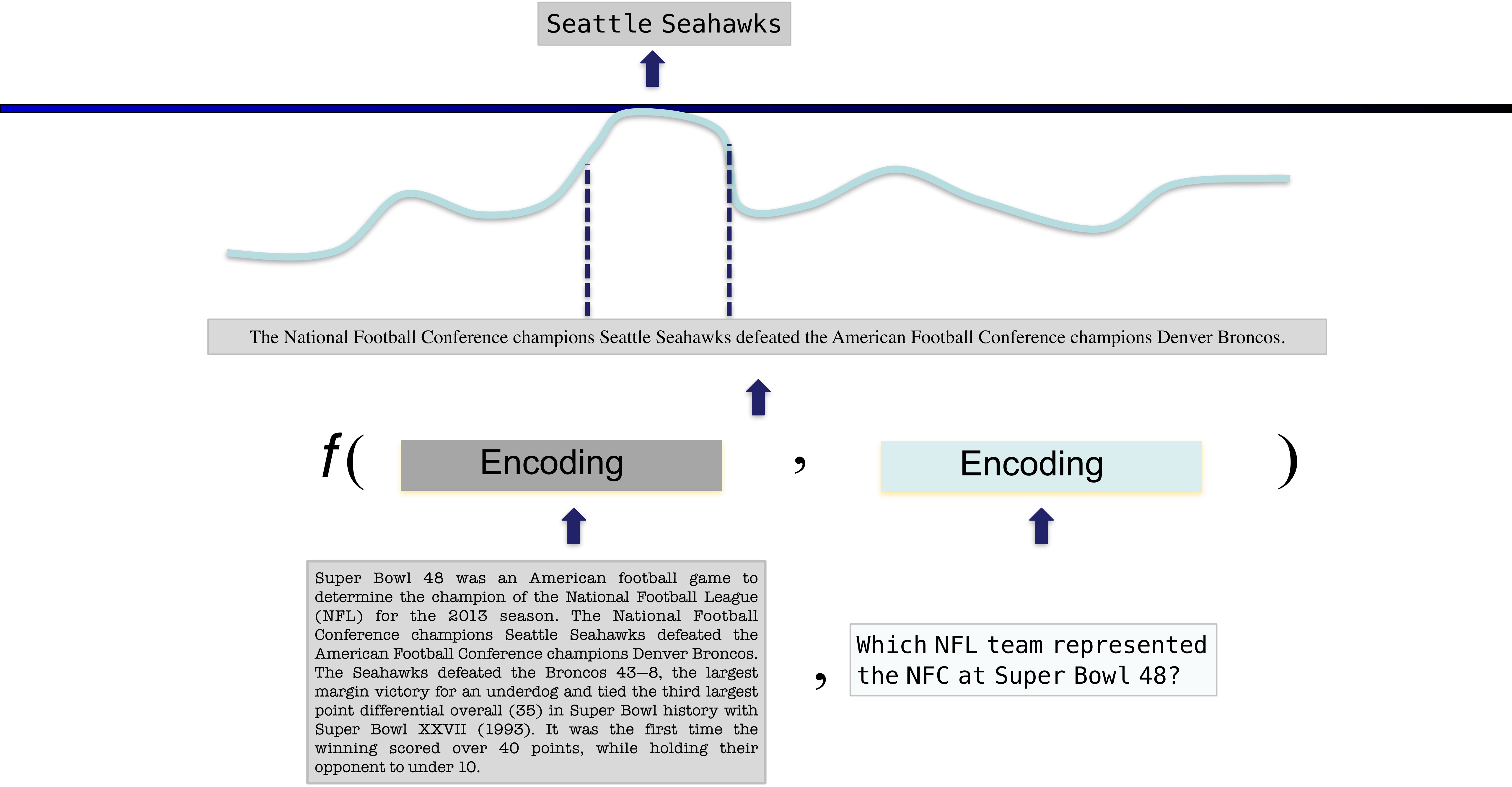


Super Bowl 48 was an American football game to determine the champion of the National Football League (NFL) for the 2013 season. The National Football Conference champions Seattle Seahawks defeated the American Football Conference champions Denver Broncos. The Seahawks defeated the Broncos 43–8, the largest margin victory for an underdog and tied the third largest point differential overall (35) in Super Bowl history with Super Bowl XXVII (1993). It was the first time the winning scored over 40 points, while holding their opponent to under 10.

Which NFL team represented the NFC at Super Bowl 48?

Context

Question



Seattle Seahawks

The National Football Conference champions Seattle Seahawks defeated the American Football Conference champions Denver Broncos.

$f($ Encoding , Encoding $)$

Super Bowl 48 was an American football game to determine the champion of the National Football League (NFL) for the 2013 season. The National Football Conference champions Seattle Seahawks defeated the American Football Conference champions Denver Broncos. The Seahawks defeated the Broncos 43–8, the largest margin victory for an underdog and tied the third largest point differential overall (35) in Super Bowl history with Super Bowl XXVII (1993). It was the first time the winning scored over 40 points, while holding their opponent to under 10.

,

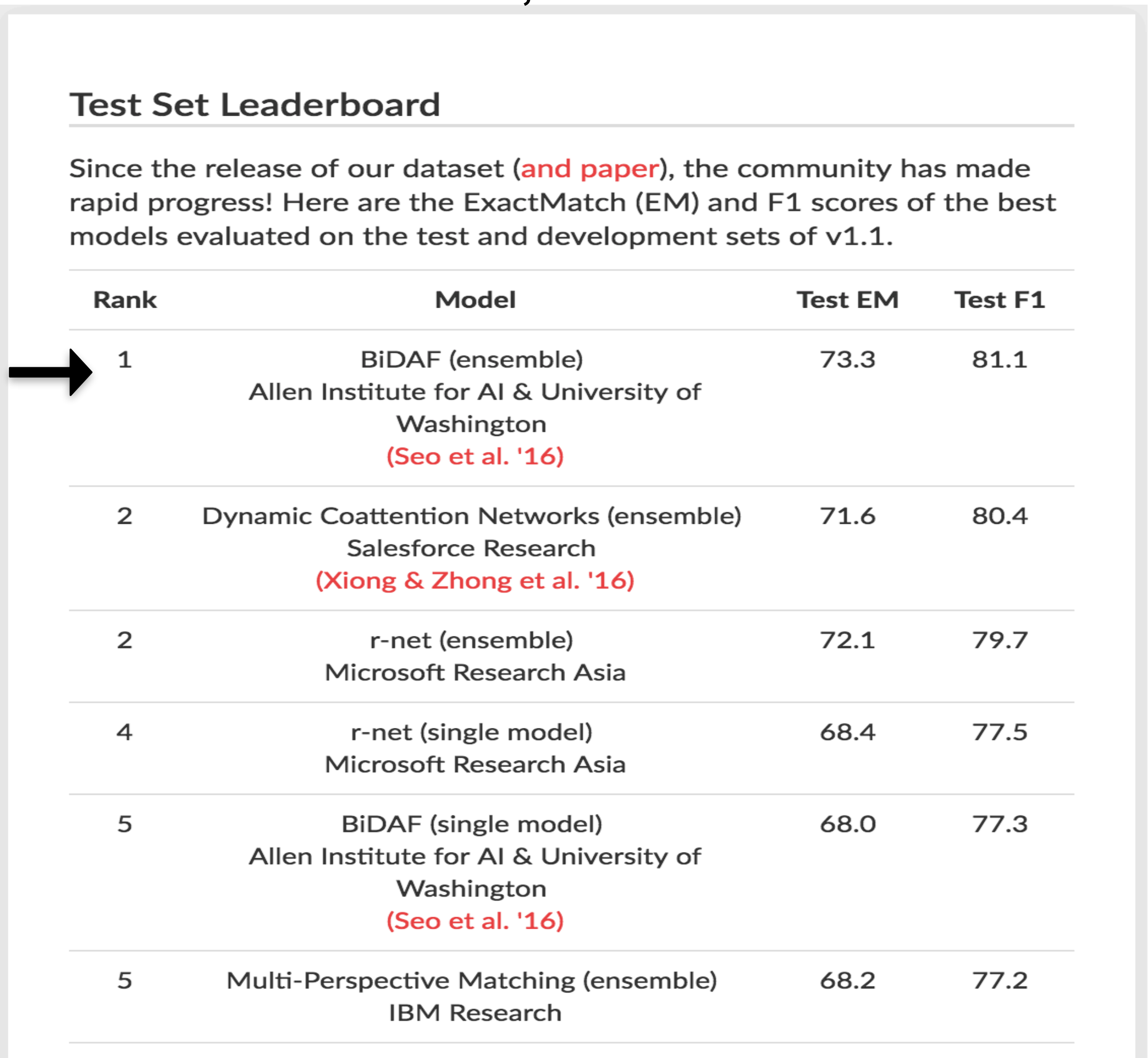
Which NFL team represented the NFC at Super Bowl 48?

Context

Question

Question Answering Leaderboard

Jan 1, 2017

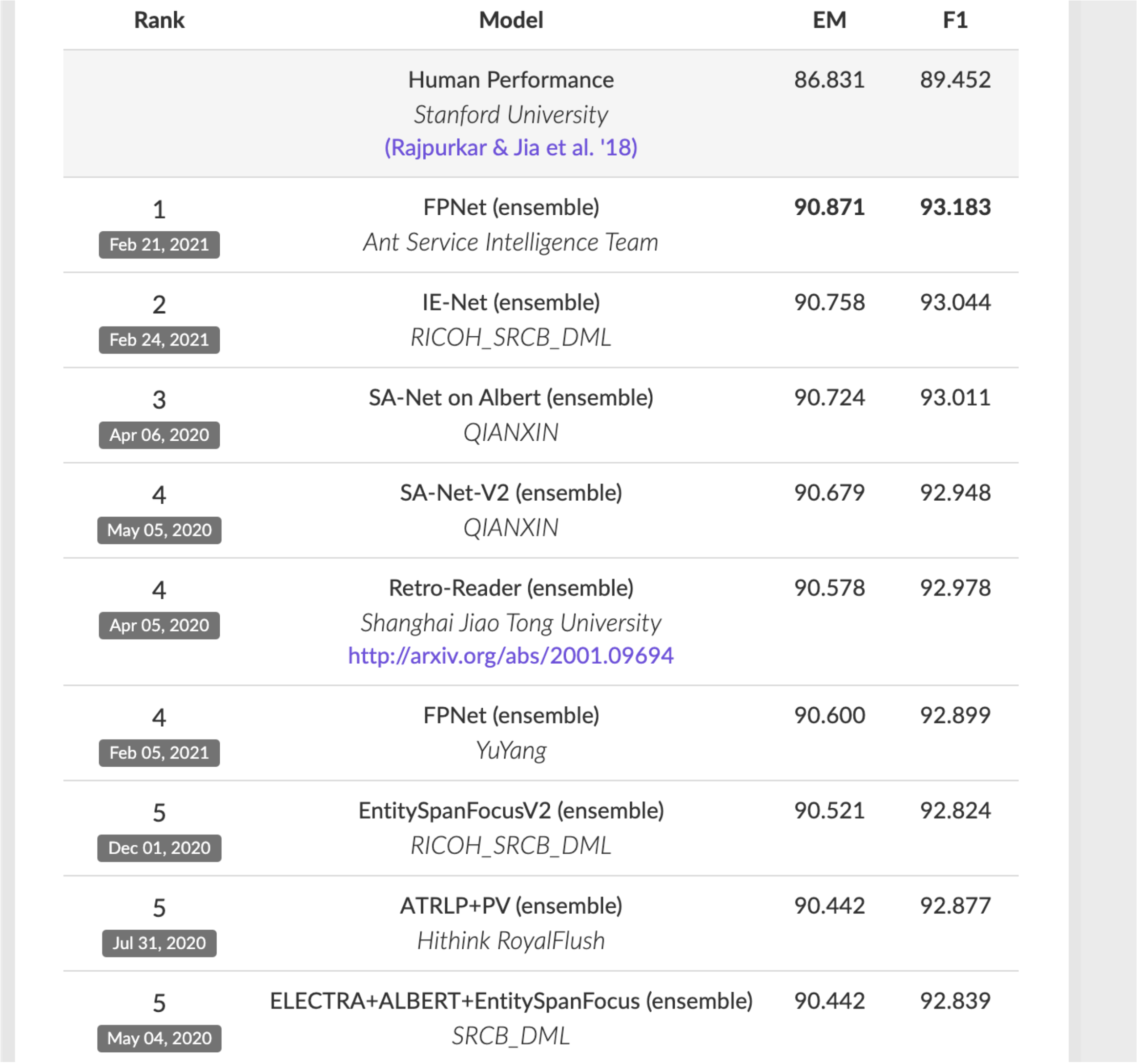


Test Set Leaderboard

Since the release of our dataset (and paper), the community has made rapid progress! Here are the ExactMatch (EM) and F1 scores of the best models evaluated on the test and development sets of v1.1.

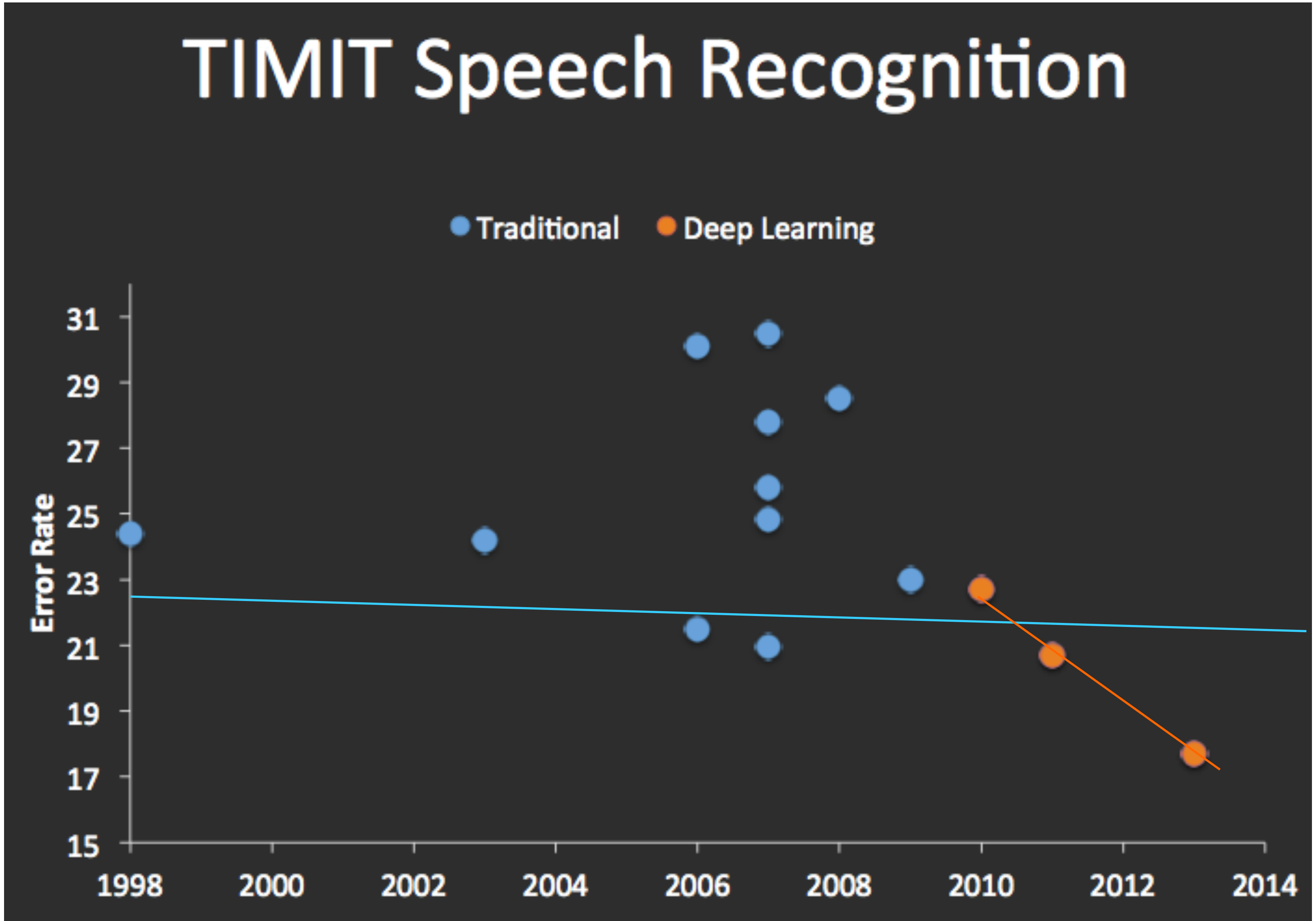
Rank	Model	Test EM	Test F1
1	BiDAF (ensemble) Allen Institute for AI & University of Washington (Seo et al. '16)	73.3	81.1
2	Dynamic Coattention Networks (ensemble) Salesforce Research (Xiong & Zhong et al. '16)	71.6	80.4
2	r-net (ensemble) Microsoft Research Asia	72.1	79.7
4	r-net (single model) Microsoft Research Asia	68.4	77.5
5	BiDAF (single model) Allen Institute for AI & University of Washington (Seo et al. '16)	68.0	77.3
5	Multi-Perspective Matching (ensemble) IBM Research	68.2	77.2

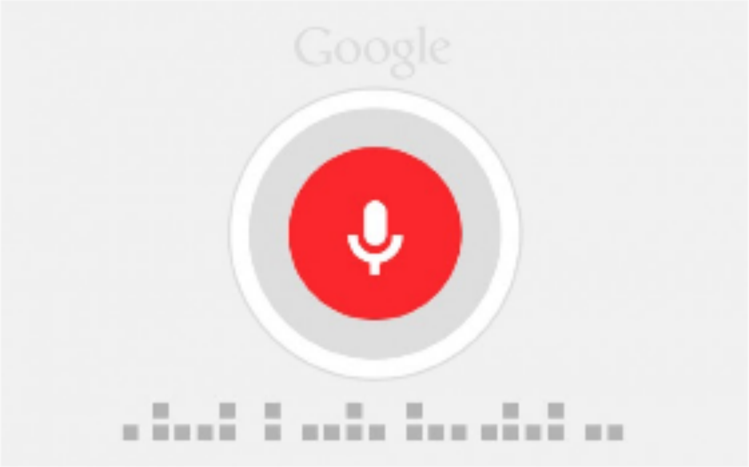
March 8, 2021



Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Feb 21, 2021	FPNet (ensemble) *Ant Service Intelligence Team*	**90.871**	**93.183**
2 Feb 24, 2021	IE-Net (ensemble) *RICOH_SRCB_DML*	90.758	93.044
3 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	90.724	93.011
4 May 05, 2020	SA-Net-V2 (ensemble) *QIANXIN*	90.679	92.948
4 Apr 05, 2020	Retro-Reader (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694	90.578	92.978
4 Feb 05, 2021	FPNet (ensemble) *YuYang*	90.600	92.899
5 Dec 01, 2020	EntitySpanFocusV2 (ensemble) *RICOH_SRCB_DML*	90.521	92.824
5 Jul 31, 2020	ATRLP+PV (ensemble) *Hithink RoyalFlush*	90.442	92.877
5 May 04, 2020	ELECTRA+ALBERT+EntitySpanFocus (ensemble) *SRCB_DML*	90.442	92.839

Speech Recognition





graph credit Matt Zeiler, Clarifai

Thanks!