Department of Computer Science
UNIVERSITY OF COLORADO **BOULDER**

Deep Learning for Question Answering

Jordan Boyd-Graber
University of Colorado Boulder
19. NOVEMBER 2014

Plan

Why Deep Learning

Review of Logistic Regression

Can't Somebody else do it? (Feature Engineering)

Deep Learning from Data

Tricks and Toolkits

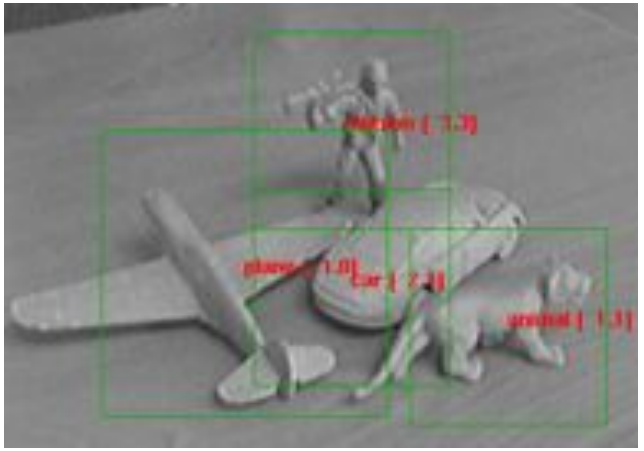
Toolkits for Deep Learning

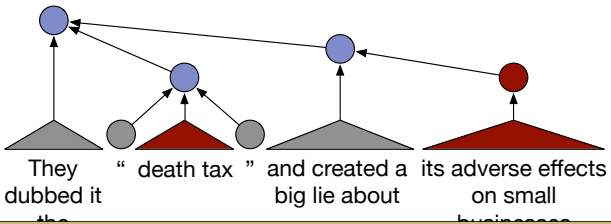
Quiz Bowl

Deep Quiz Bowl

Deep Learning was once known as "Neural Networks"

But it came back . . .





- More data
- Better tricks (regularization)
- Faster computers

And companies are investing . . .

Google Hires Brains that Helped Supercharge Machine Learning

BY ROBERT MCMILLAN 03.13.13 | 6:30 AM | PERMALINK

And companies are investing . . .

'Chinese Google' Opens Artificial-Intelligence Lab in Silicon Valley

BY DANIELA HERNANDEZ 04.12.13 | 6:30 AM | PERMALINK

And companies are investing . . .

Facebook's 'Deep Learning' Guru Reveals the Future of AI

BY CADE METZ 12.12.13 | 6:30 AM | PERMALINK

Plan

Why Deep Learning

Review of Logistic Regression

Can't Somebody else do it? (Feature Engineering)

Deep Learning from Data

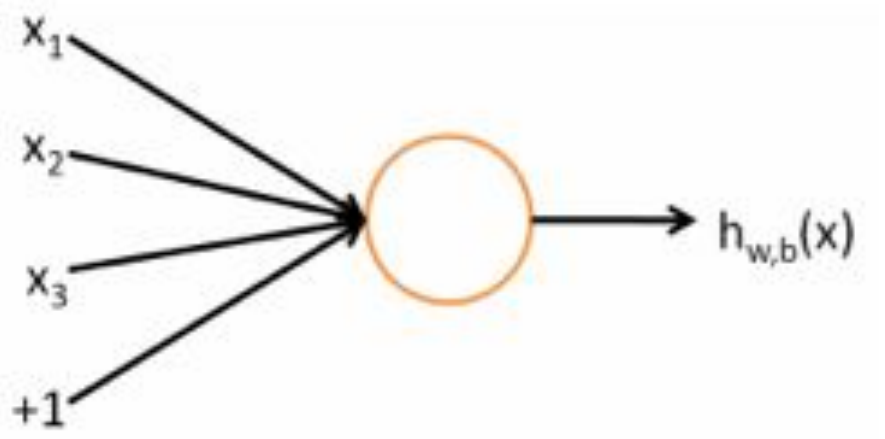
Tricks and Toolkits

Toolkits for Deep Learning

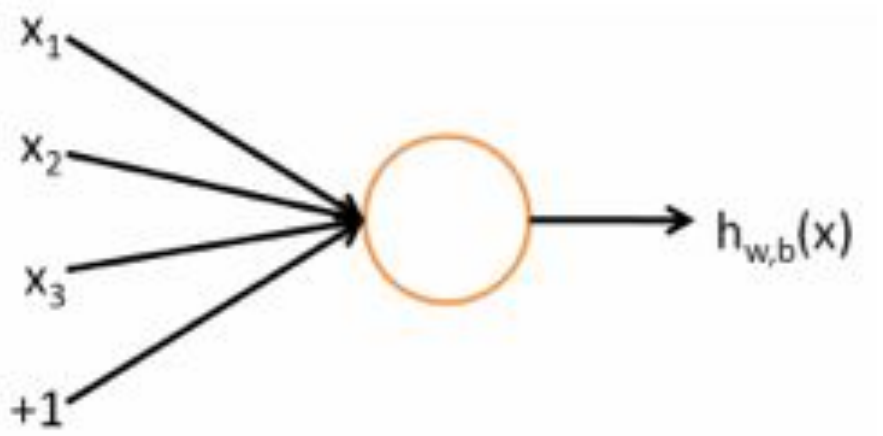
Quiz Bowl

Deep Quiz Bowl

Map inputs to output



Map inputs to output

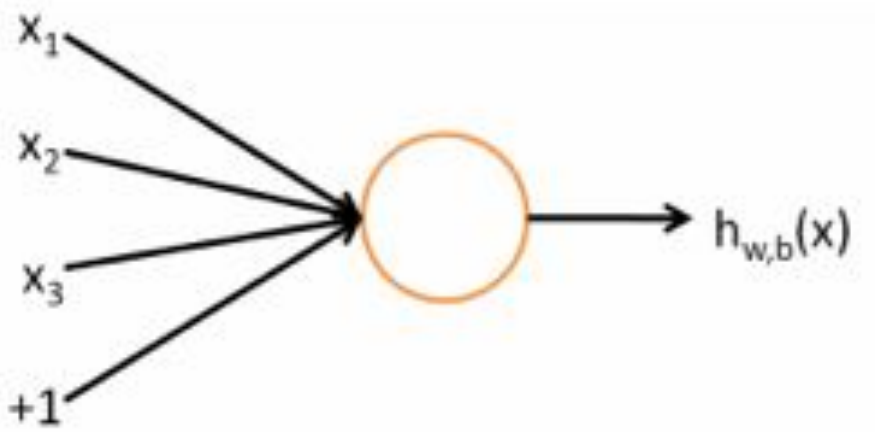


Input

Vector $x_1 \ldots x_d$

inputs encoded as real numbers

Map inputs to output



Input

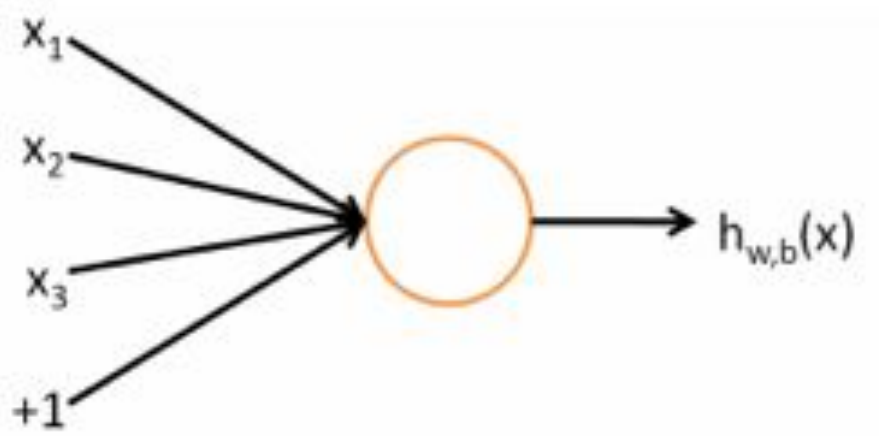
Vector $x_1 \ldots x_d$

Output

$$f\left(\sum_i W_i x_i + b\right)$$

multiply inputs by weights

Map inputs to output



Input

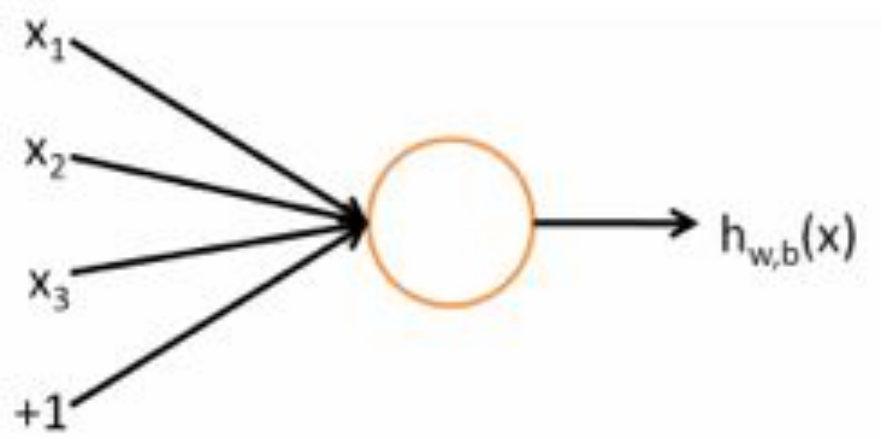
Vector $x_1 \ldots x_d$

Output

$$f\left(\sum_i W_i x_i + b\right)$$

add bias

Map inputs to output



Input

Vector $x_1 \ldots x_d$

Output

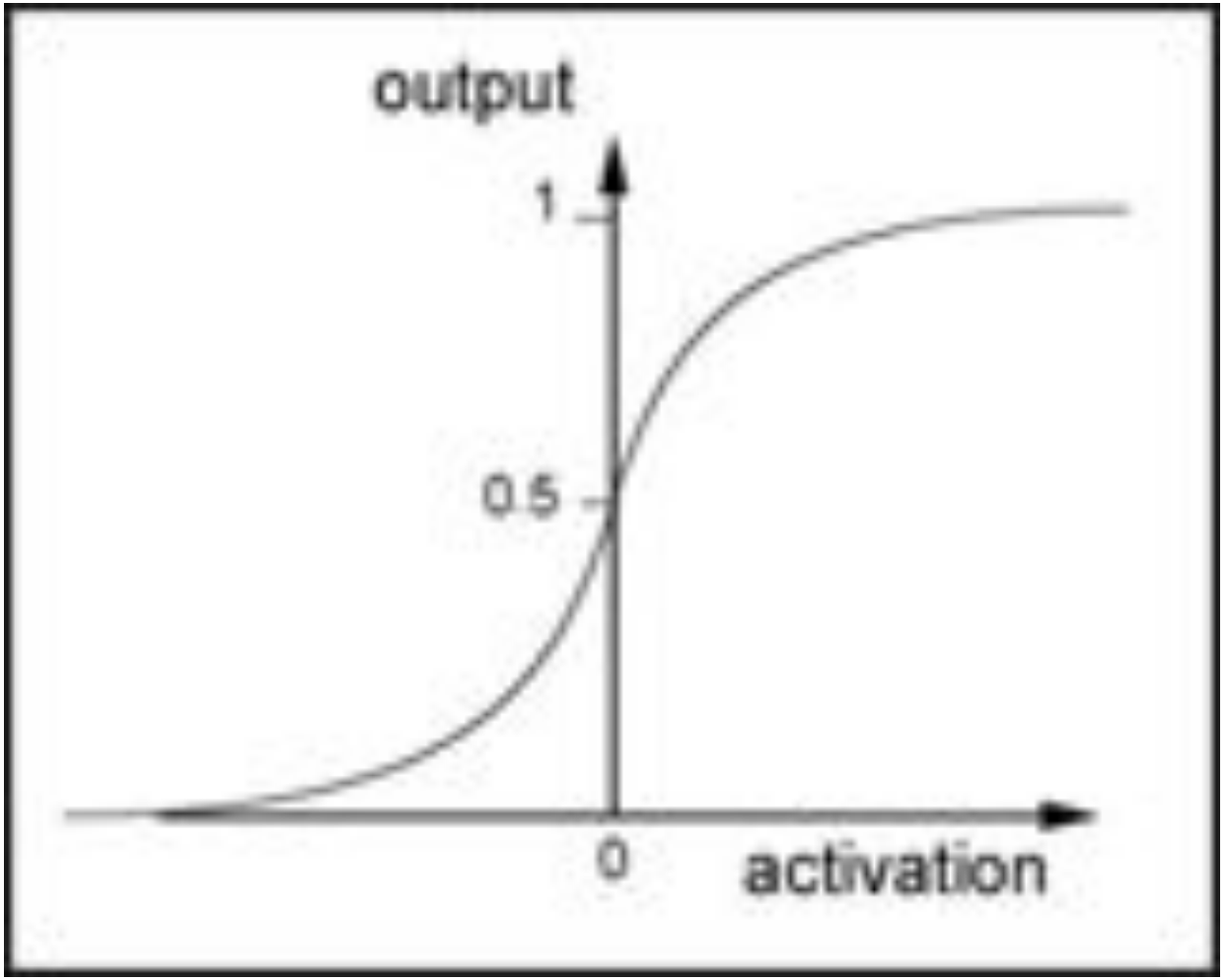
$$f\left(\sum_i W_i x_i + b\right)$$

Activation

$$f(z) \equiv \frac{1}{1 + \exp(-z)}$$

pass through nonlinear sigmoid

What's a sigmoid?



In the shallow end

- This is still logistic regression
- Engineering features x is difficult (and requires expertise)
- Can we learn how to represent inputs into final decision?

Plan

Why Deep Learning

Review of Logistic Regression

Can't Somebody else do it? (Feature Engineering)

Deep Learning from Data

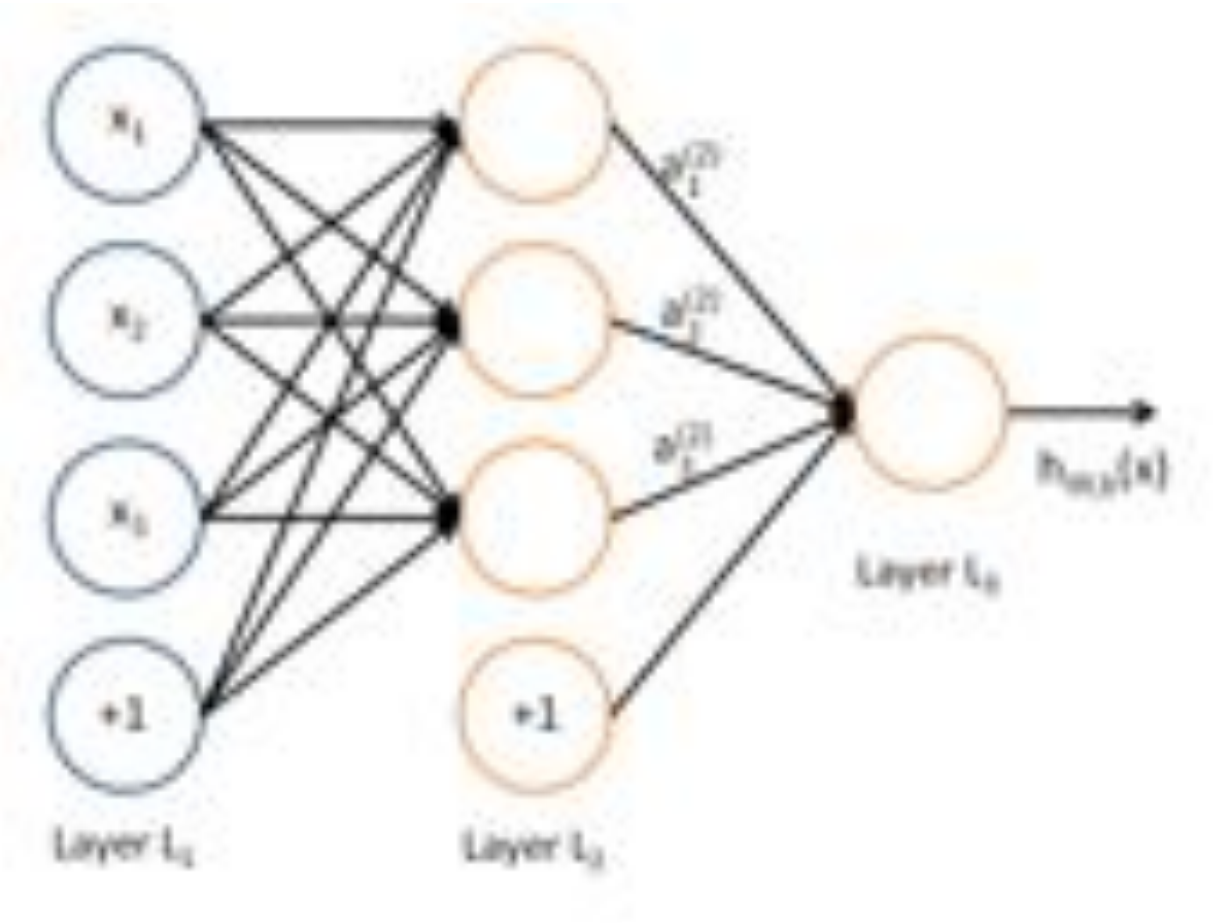
Tricks and Toolkits

Toolkits for Deep Learning

Quiz Bowl

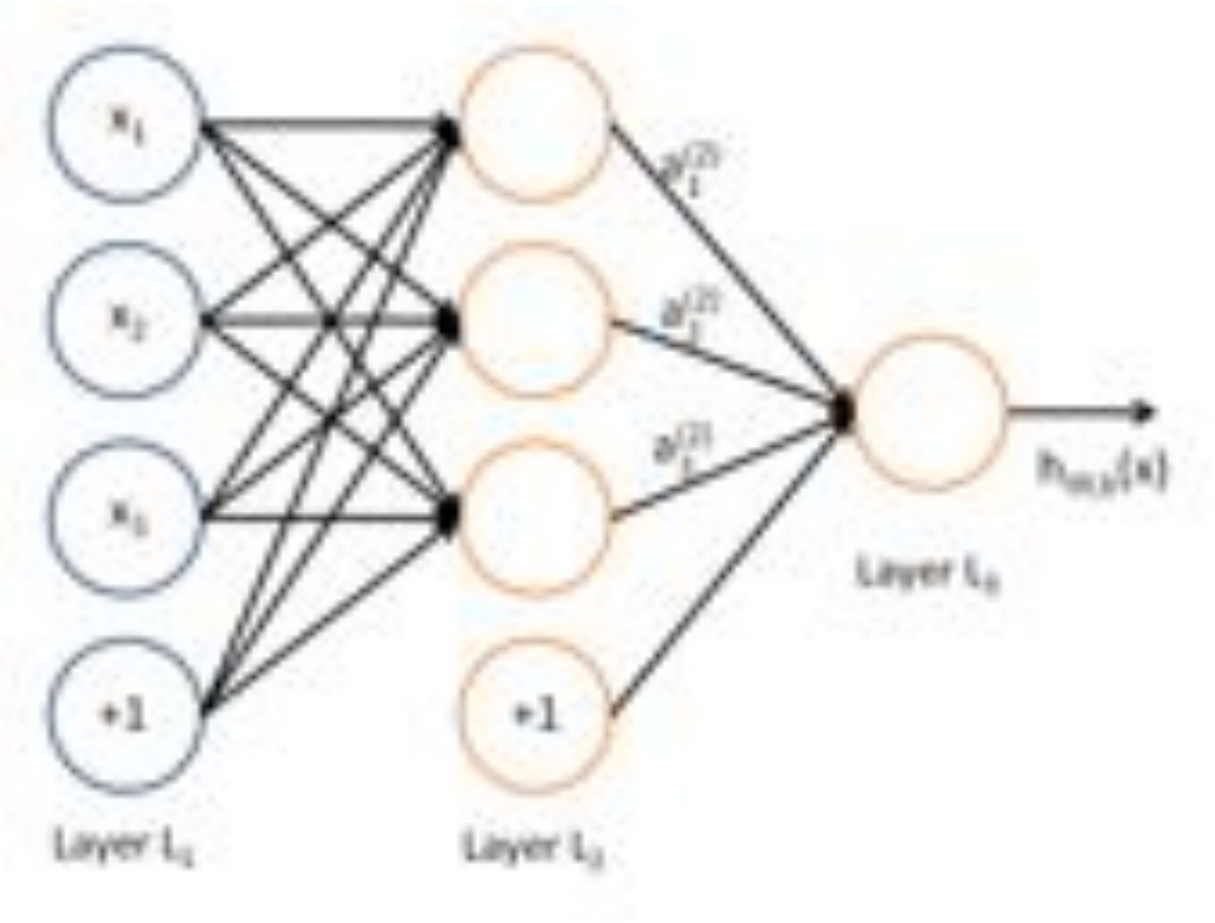
Deep Quiz Bowl

Learn the features and the function



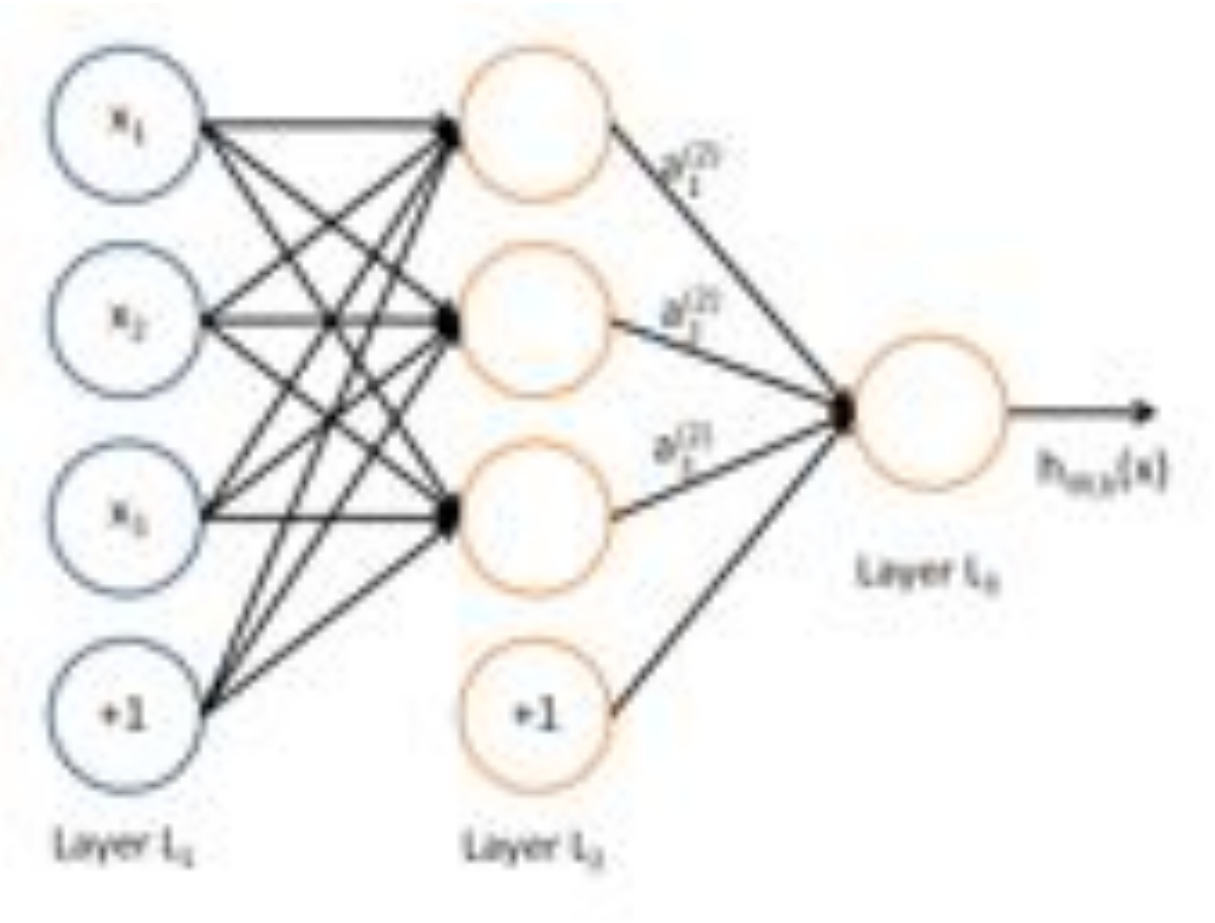
$$a_1^{(2)} = f\left(W_{11}^{(1)}x_1 + W_{12}^{(1)}x_2 + W_{13}^{(1)}x_3 + b_1^{(1)}\right)$$

Learn the features and the function



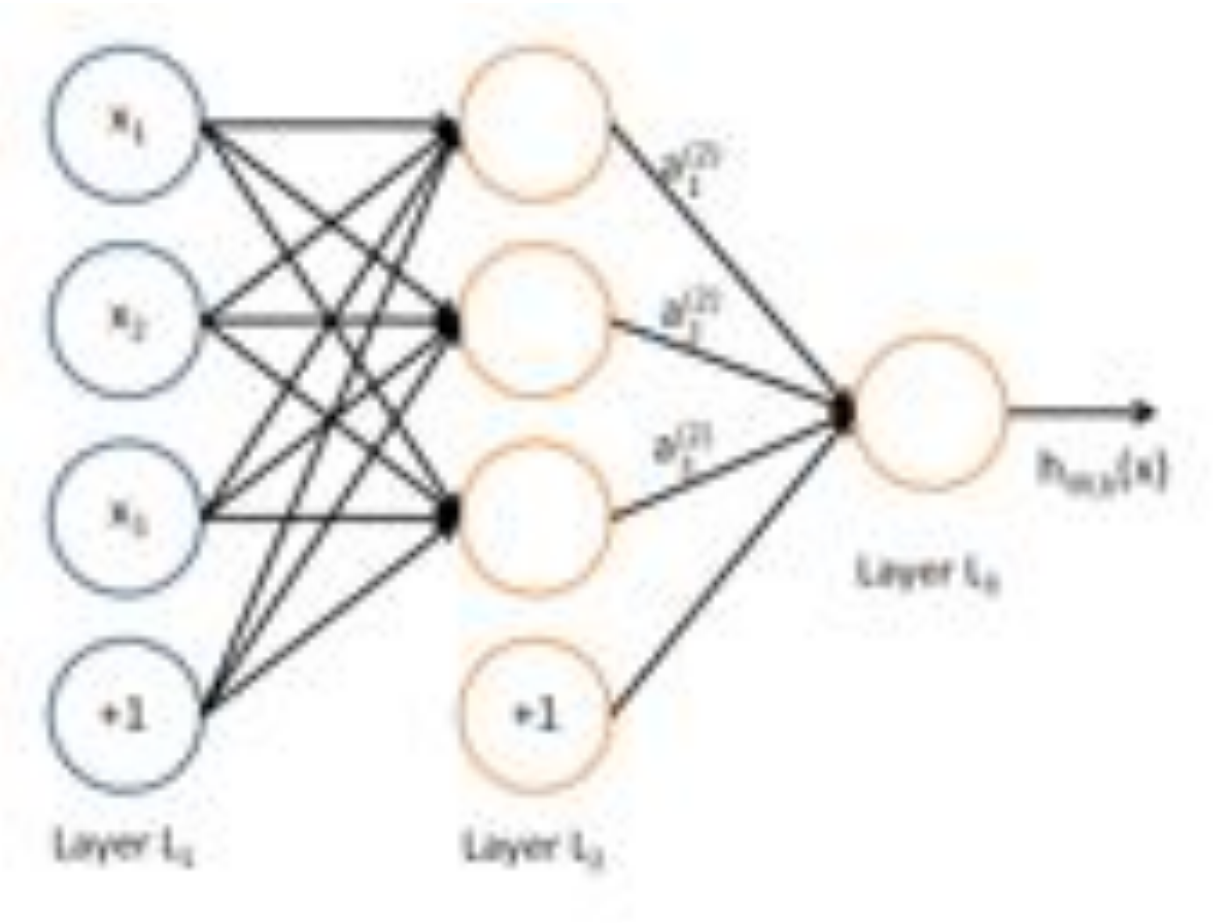
$$a_2^{(2)} = f\left(W_{21}^{(1)}x_1 + W_{22}^{(1)}x_2 + W_{23}^{(1)}x_3 + b_2^{(1)}\right)$$

Learn the features and the function



$$a_3^{(2)} = f\left(W_{31}^{(1)}x_1 + W_{32}^{(1)}x_2 + W_{33}^{(1)}x_3 + b_3^{(1)}\right)$$

Learn the features and the function



$$h_{W,b}(x) = a_1^{(3)} = f\left(W_{11}^{(2)} a_1^{(2)} + W_{12}^{(2)} a_2^{(2)} + W_{13}^{(2)} a_3^{(2)} + b_1^{(2)}\right)$$

Objective Function

- For every example x, y of our supervised training set, we want the label y to match the prediction $h_{W,b}(x)$.

$$J(W, b; x, y) \equiv \frac{1}{2}||h_{W,b}(x) - y||^2 \tag{1}$$

Objective Function

- For every example x, y of our supervised training set, we want the label y to match the prediction $h_{W,b}(x)$.

$$J(W, b; x, y) \equiv \frac{1}{2}||h_{W,b}(x) - y||^2 \tag{1}$$

- We want this value, summed over all of the examples to be as small as possible

Objective Function

- For every example x, y of our supervised training set, we want the label y to match the prediction $h_{W,b}(x)$.

$$J(W, b; x, y) \equiv \frac{1}{2}||h_{W,b}(x) - y||^2 \tag{1}$$

- We want this value, summed over all of the examples to be as small as possible
- We also want the weights not to be too large

$$\frac{\lambda}{2} \sum_{l}^{n_l - 1} \sum_{i=1}^{s_l} \sum_{j=1}^{s_{l+1}} \left(W_{ji}^{l} \right)^2 \tag{2}$$

Objective Function

- For every example x, y of our supervised training set, we want the label y to match the prediction $h_{W,b}(x)$.

$$J(W, b; x, y) \equiv \frac{1}{2}||h_{W,b}(x) - y||^2 \tag{1}$$

- We want this value, summed over all of the examples to be as small as possible
- We also want the weights not to be too large

$$\frac{\lambda}{2} \sum_{l}^{n_l - 1} \sum_{i=1}^{s_l} \sum_{j=1}^{s_{l+1}} \left(W_{ji}^{l} \right)^2 \tag{2}$$

Objective Function

- For every example x, y of our supervised training set, we want the label y to match the prediction $h_{W,b}(x)$.

$$J(W, b; x, y) \equiv \frac{1}{2}||h_{W,b}(x) - y||^2 \tag{1}$$

- We want this value, summed over all of the examples to be as small as possible
- We also want the weights not to be too large

$$\frac{\lambda}{2} \sum_{l}^{n_l - 1} \sum_{i=1}^{s_l} \sum_{j=1}^{s_{l+1}} \left(W_{ji}^{l} \right)^2 \tag{2}$$

Sum over all layers

Objective Function

- For every example x, y of our supervised training set, we want the label y to match the prediction $h_{W,b}(x)$.

$$J(W, b; x, y) \equiv \frac{1}{2}||h_{W,b}(x) - y||^2 \tag{1}$$

- We want this value, summed over all of the examples to be as small as possible
- We also want the weights not to be too large

$$\frac{\lambda}{2} \sum_{l}^{n_l - 1} \sum_{i=1}^{s_l} \sum_{j=1}^{s_{l+1}} \left(W_{ji}^{l} \right)^2 \tag{2}$$

Sum over all sources

Objective Function

- For every example x, y of our supervised training set, we want the label y to match the prediction $h_{W,b}(x)$.

$$J(W, b; x, y) \equiv \frac{1}{2}||h_{W,b}(x) - y||^2 \tag{1}$$

- We want this value, summed over all of the examples to be as small as possible
- We also want the weights not to be too large

$$\frac{\lambda}{2} \sum_{l}^{n_l - 1} \sum_{i=1}^{s_l} \sum_{j=1}^{s_{l+1}} \left(W_{ji}^{l} \right)^2 \tag{2}$$

Sum over all destinations

Objective Function

Putting it all together:

$$J(W,b) = \left[\frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}||h_{W,b}(x^{(i)}) - y^{(i)}||^2\right] + \frac{\lambda}{2}\sum_{l}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}\left(W_{ji}^{l}\right)^2 \tag{3}$$

Objective Function

Putting it all together:

$$J(W,b) = \left[\frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}||h_{W,b}(x^{(i)}) - y^{(i)}||^2\right] + \frac{\lambda}{2}\sum_{l}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}\left(W_{ji}^{l}\right)^2 \tag{3}$$

- Our goal is to minimize $J(W,b)$ as a function of W and b

Objective Function

Putting it all together:

$$J(W,b) = \left[\frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}||h_{W,b}(x^{(i)}) - y^{(i)}||^2\right] + \frac{\lambda}{2}\sum_{l}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}\left(W_{ji}^{l}\right)^2 \tag{3}$$

- Our goal is to minimize $J(W,b)$ as a function of W and b
- Initialize W and b to small random value near zero

Objective Function

Putting it all together:

$$J(W,b) = \left[\frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}||h_{W,b}(x^{(i)}) - y^{(i)}||^2\right] + \frac{\lambda}{2}\sum_{l}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}\left(W_{ji}^{l}\right)^2 \tag{3}$$

- Our goal is to minimize $J(W,b)$ as a function of W and b
- Initialize W and b to small random value near zero
- Adjust parameters to optimize J

Plan

Why Deep Learning

Review of Logistic Regression

Can't Somebody else do it? (Feature Engineering)

Deep Learning from Data

Tricks and Toolkits

Toolkits for Deep Learning

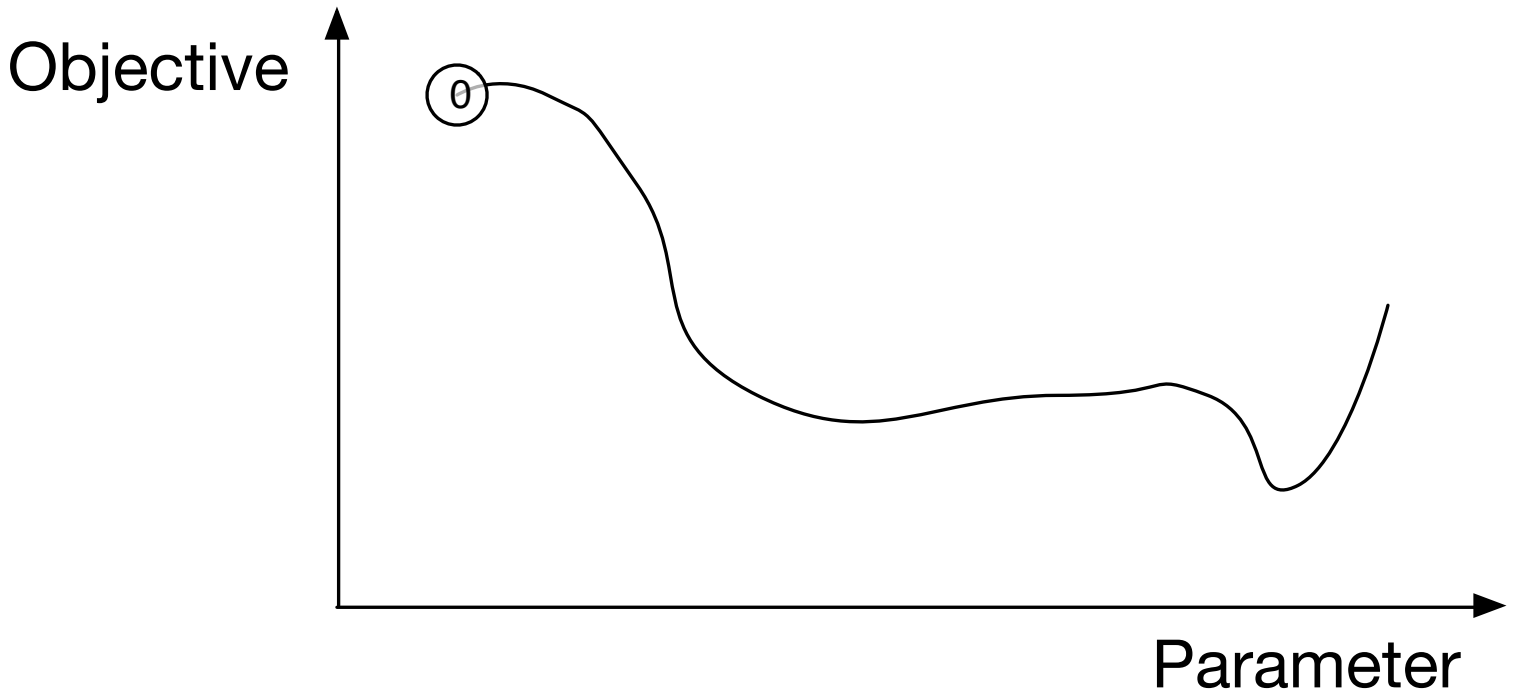
Quiz Bowl

Deep Quiz Bowl

Gradient Descent

Goal

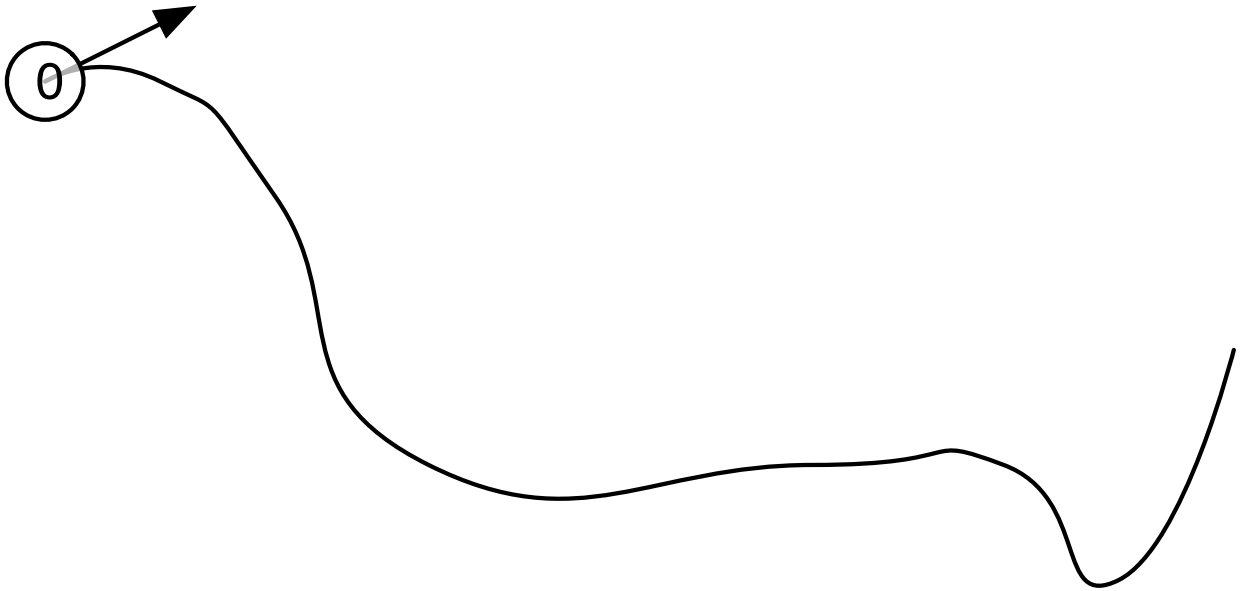
Optimize J with respect to variables W and b



Gradient Descent

Goal

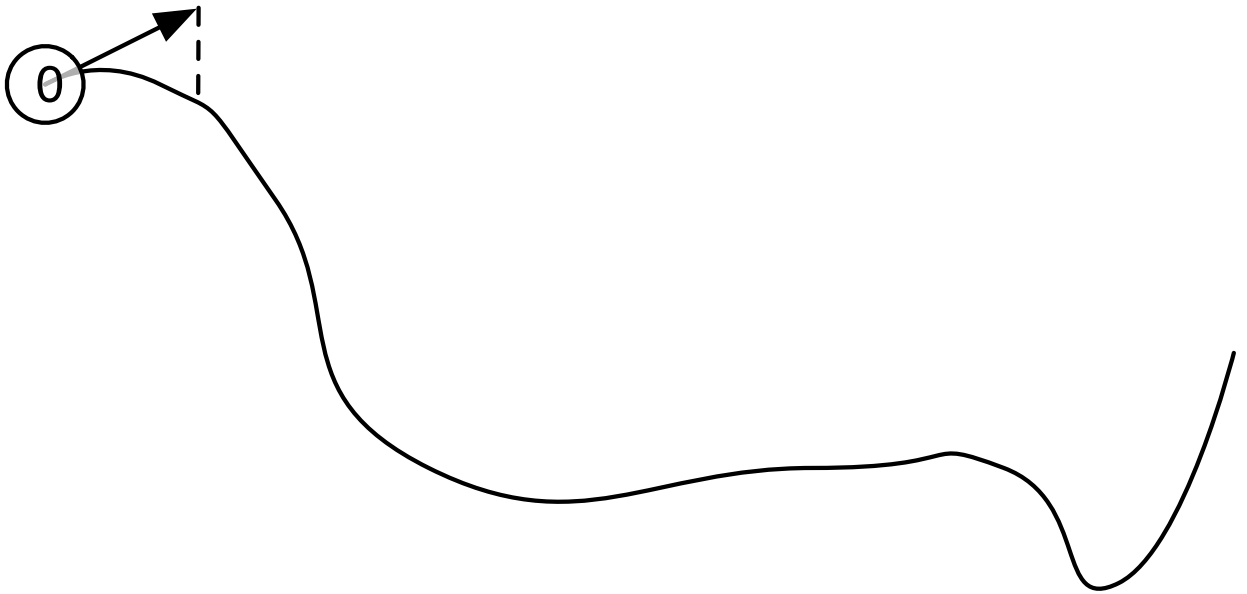
Optimize J with respect to variables W and b



Gradient Descent

Goal

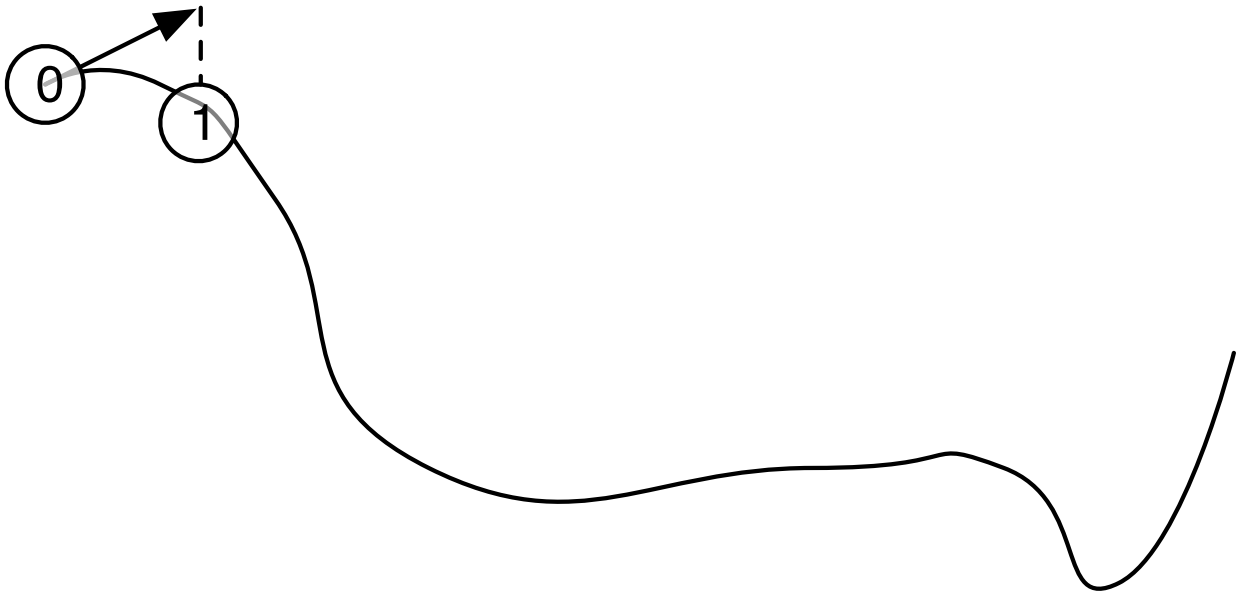
Optimize J with respect to variables W and b



Gradient Descent

Goal

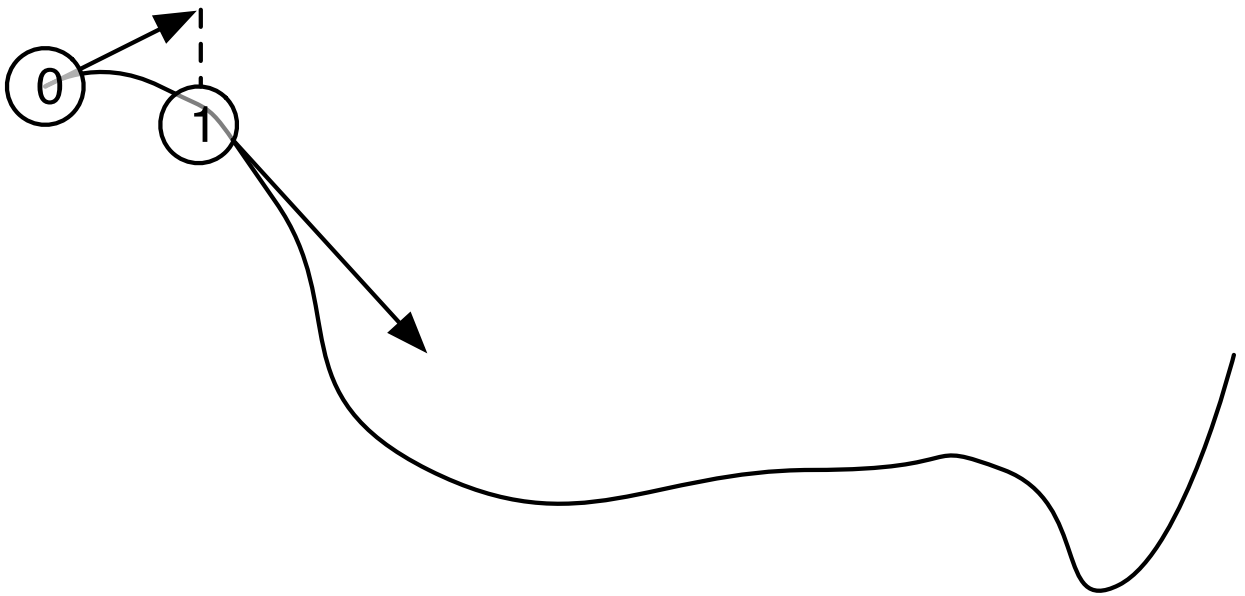
Optimize J with respect to variables W and b



Gradient Descent

Goal

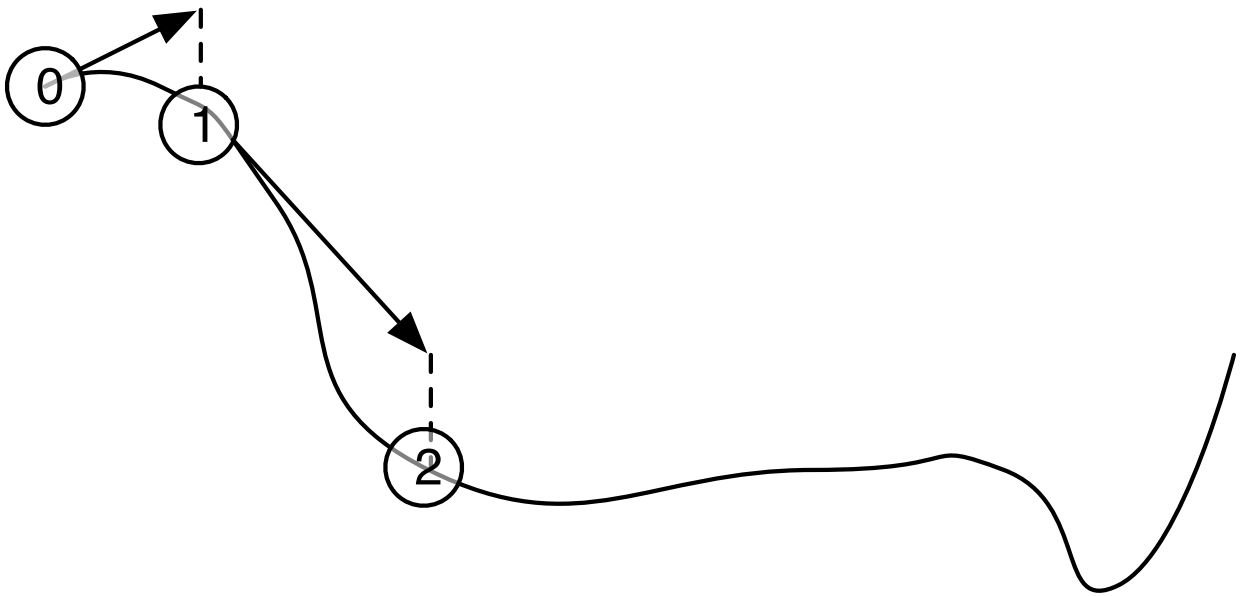
Optimize J with respect to variables W and b



Gradient Descent

Goal

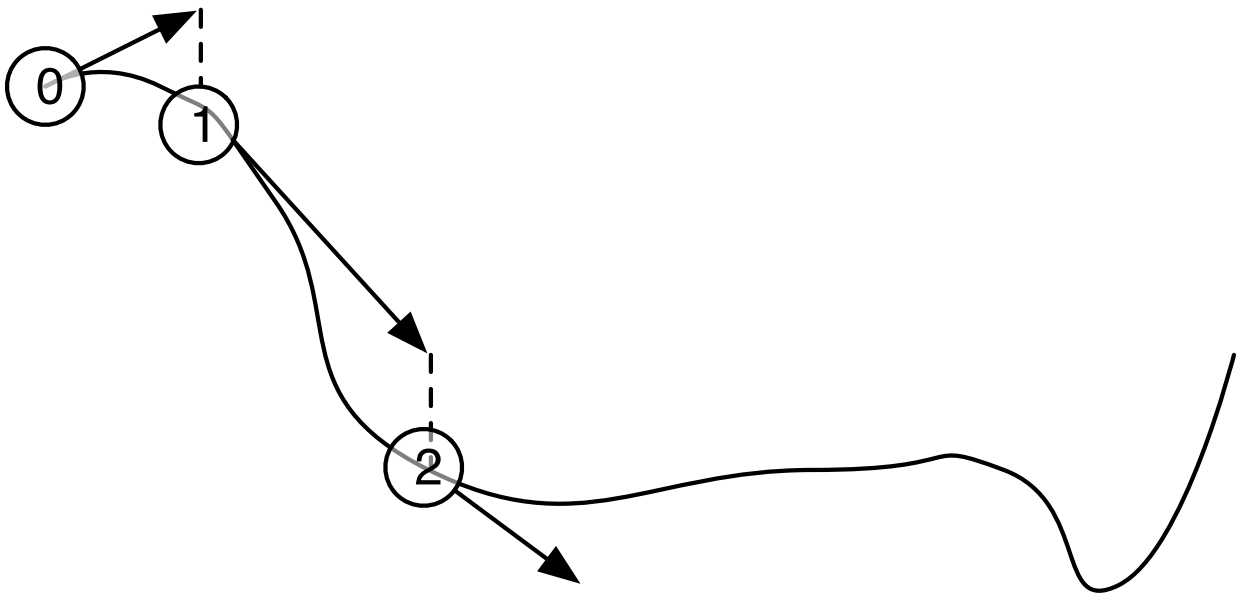
Optimize J with respect to variables W and b



Gradient Descent

Goal

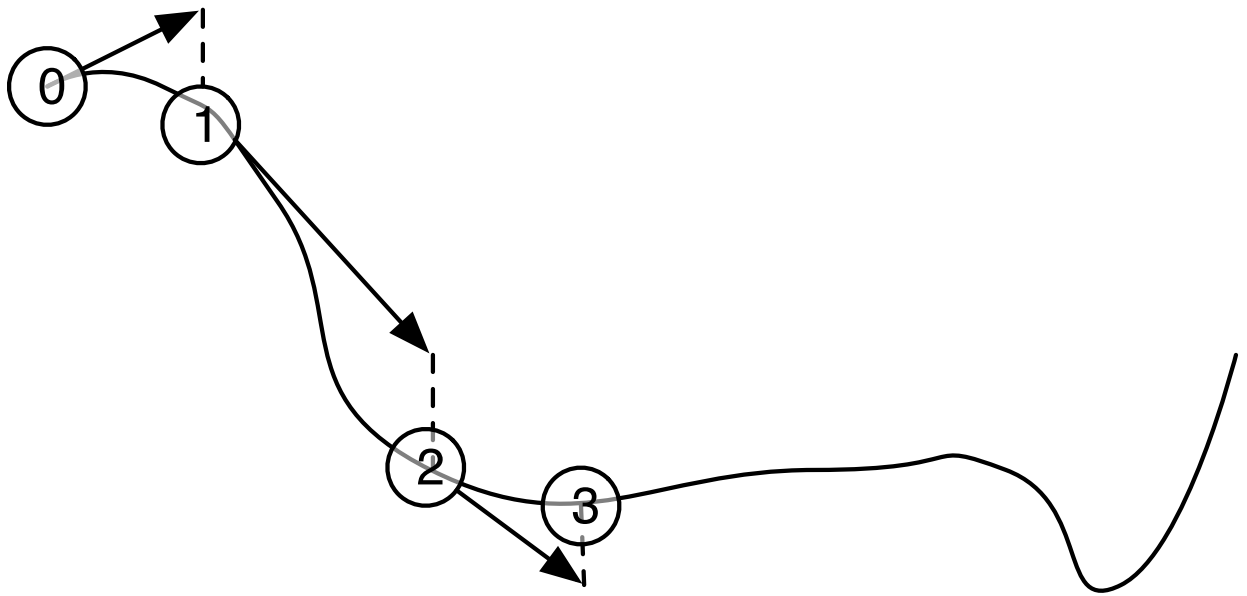
Optimize J with respect to variables W and b



Gradient Descent

Goal

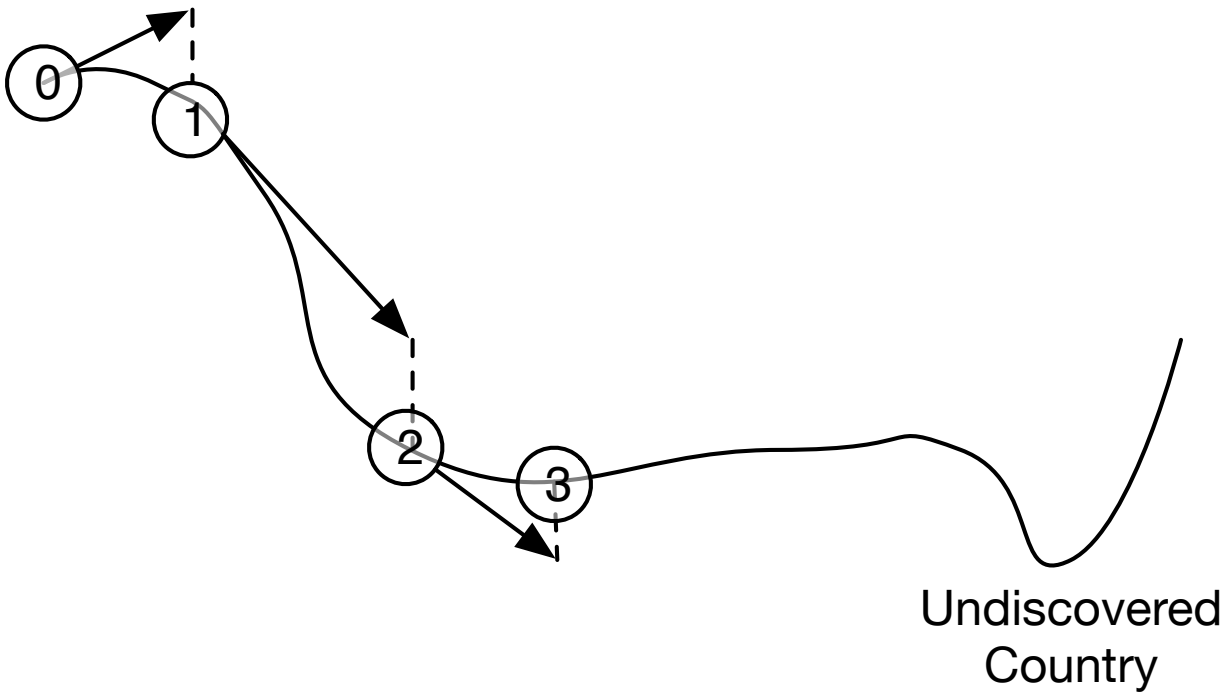
Optimize J with respect to variables W and b



Gradient Descent

Goal

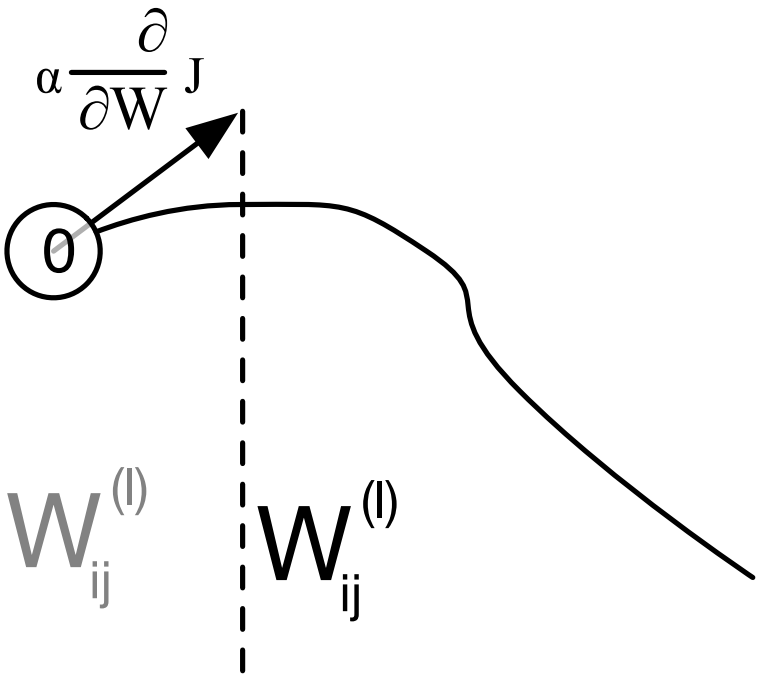
Optimize J with respect to variables W and b



Gradient Descent

Goal

Optimize J with respect to variables W and b



Backpropigation

- For convenience, write the input to sigmoid

$$z_i^{(l)} = \sum_{j=1}^{n} W_{ij}^{(l-1)} x_j + b_i^{(l-1)} \tag{4}$$

Backpropigation

- For convenience, write the input to sigmoid

$$z_i^{(l)} = \sum_{j=1}^{n} W_{ij}^{(l-1)} x_j + b_i^{(l-1)} \tag{4}$$

- The gradient is a function of a node's error $\delta_i^{(l)}$

Backpropigation

- For convenience, write the input to sigmoid

$$z_i^{(l)} = \sum_{j=1}^{n} W_{ij}^{(l-1)} x_j + b_i^{(l-1)} \tag{4}$$

- The gradient is a function of a node's error $\delta_i^{(l)}$
- For output nodes, the error is obvious:

$$\delta_i^{(n_l)} = \frac{\partial}{\partial z_i^{(n_l)}} ||y - h_{w,b}(x)||^2 = -\left(y_i - a_i^{(n_l)}\right) \cdot f'\left(z_i^{(n_l)}\right) \frac{1}{2} \tag{5}$$

Backpropigation

- For convenience, write the input to sigmoid

$$z_i^{(l)} = \sum_{j=1}^{n} W_{ij}^{(l-1)} x_j + b_i^{(l-1)} \tag{4}$$

- The gradient is a function of a node's error $\delta_i^{(l)}$
- For output nodes, the error is obvious:

$$\delta_i^{(n_l)} = \frac{\partial}{\partial z_i^{(n_l)}} ||y - h_{w,b}(x)||^2 = -\left(y_i - a_i^{(n_l)}\right) \cdot f'\left(z_i^{(n_l)}\right) \frac{1}{2} \tag{5}$$

- Other nodes must "backpropagate" downstream error based on connection strength

$$\delta_i^{(l)} = \left(\sum_{j=1}^{s_{t+1}} W_{ji}^{(l+1)} \delta_j^{(l+1)} \right) f'(z_i^{(l)}) \tag{6}$$

Backpropigation

- For convenience, write the input to sigmoid

$$z_i^{(l)} = \sum_{j=1}^{n} W_{ij}^{(l-1)} x_j + b_i^{(l-1)} \tag{4}$$

- The gradient is a function of a node's error $\delta_i^{(l)}$
- For output nodes, the error is obvious:

$$\delta_i^{(n_l)} = \frac{\partial}{\partial z_i^{(n_l)}} ||y - h_{w,b}(x)||^2 = -\left(y_i - a_i^{(n_l)}\right) \cdot f'\left(z_i^{(n_l)}\right) \frac{1}{2} \tag{5}$$

- Other nodes must "backpropagate" downstream error based on connection strength

$$\delta_i^{(l)} = \left(\sum_{j=1}^{s_{t+1}} W_{ji}^{(l+1)} \delta_j^{(l+1)} \right) f'(z_i^{(l)}) \tag{6}$$

Backpropigation

- For convenience, write the input to sigmoid

$$z_i^{(l)} = \sum_{j=1}^{n} W_{ij}^{(l-1)} x_j + b_i^{(l-1)} \tag{4}$$

- The gradient is a function of a node's error $\delta_i^{(l)}$
- For output nodes, the error is obvious:

$$\delta_i^{(n_l)} = \frac{\partial}{\partial z_i^{(n_l)}} ||y - h_{w,b}(x)||^2 = -\left(y_i - a_i^{(n_l)}\right) \cdot f'\left(z_i^{(n_l)}\right) \frac{1}{2} \tag{5}$$

- Other nodes must "backpropagate" downstream error based on connection strength

$$\delta_i^{(l)} = \left(\sum_{j=1}^{s_{t+1}} W_{ji}^{(l+1)} \delta_j^{(l+1)} \right) f'(z_i^{(l)}) \tag{6}$$

Partial Derivatives

- For weights, the partial derivatives are

$$\frac{\partial}{\partial W_{ij}^{(l)}} J(W, b; x, y) = a_j^{(l)} \delta_i^{(l+1)} \tag{7}$$

- For the bias terms, the partial derivatives are

$$\frac{\partial}{\partial b_i^{(l)}} J(W, b; x, y) = \delta_i^{(l+1)} \tag{8}$$

- But this is just for a single example . . .

Full Gradient Descent Algorithm

1. Initialize $U^{(l)}$ and $V^{(l)}$ as zero
2. For each example $i1 \dots m$
   - 2.1 Use backpropagation to compute $\nabla_W J$ and $\nabla_b J$
   - 2.2 Update weight shifts $U^{(l)} = U^{(l)} + \nabla_{W^{(l)}} J(W, b; x, y)$
   - 2.3 Update bias shifts $V^{(l)} = V^{(l)} + \nabla_{b^{(l)}} J(W, b; x, y)$
3. Update the parameters

$$W^{(l)} = W^{(l)} - \alpha \left[\left(\frac{1}{m} U^{(l)} \right) \right] \tag{9}$$

$$b^{(l)} = b^{(l)} - \alpha \left[\frac{1}{m} V^{(l)} \right] \tag{10}$$

4. Repeat until weights stop changing

Plan

Why Deep Learning

Review of Logistic Regression

Can't Somebody else do it? (Feature Engineering)

Deep Learning from Data

Tricks and Toolkits

Toolkits for Deep Learning

Quiz Bowl

Deep Quiz Bowl

Tricks

- **Stochastic gradient**: compute gradient from a few examples
- **Hardware**: Do matrix computations on GPUs
- **Dropout**: Randomly set some inputs to zero
- **Initialization**: Using an autoencoder can help representation

Plan

Why Deep Learning

Review of Logistic Regression

Can't Somebody else do it? (Feature Engineering)

Deep Learning from Data

Tricks and Toolkits

Toolkits for Deep Learning

Quiz Bowl

Deep Quiz Bowl

- **Theano**: Python package (Yoshua Bengio)
- **Torch7**: Lua package (Yann LeCunn)
- **ConvNetJS**: Javascript package (Andrej Karpathy)
- Both automatically compute gradients and have numerical optimization
- Working group this summer at UMD

Plan

Why Deep Learning

Review of Logistic Regression

Can't Somebody else do it? (Feature Engineering)

Deep Learning from Data

Tricks and Toolkits

Toolkits for Deep Learning

Quiz Bowl

Deep Quiz Bowl

Humans doing Incremental Classification

- Game called "quiz bowl"
- Two teams play each other
  - Moderator reads a question
  - When a team knows the answer, they signal ("buzz" in)
  - If right, they get points; otherwise, rest of the question is read to the other team
- Hundreds of teams in the US alone

Humans doing Incremental Classification

- Game called "quiz bowl"
- Two teams play each other
  - Moderator reads a question
  - When a team knows the answer, they signal ("buzz" in)
  - If right, they get points; otherwise, rest of the question is read to the other team
- Hundreds of teams in the US alone
- Example . . .

Sample Question 1

With Leo Szilard, he invented a doubly-eponymous

Sample Question 1

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of

Sample Question 1

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity, so

Sample Question 1

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity, so Debye adjusted the theory for low temperatures. His summation convention automatically sums repeated indices in tensor products. His name is attached to the A and B coefficients

Sample Question 1

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity, so Debye adjusted the theory for low temperatures. His summation convention automatically sums repeated indices in tensor products. His name is attached to the A and B coefficients for spontaneous and stimulated emission, the subject of one of his multiple groundbreaking 1905 papers. He further developed the model of statistics sent to him by

Sample Question 1

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity, so Debye adjusted the theory for low temperatures. His summation convention automatically sums repeated indices in tensor products. His name is attached to the A and B coefficients for spontaneous and stimulated emission, the subject of one of his multiple groundbreaking 1905 papers. He further developed the model of statistics sent to him by Bose to describe particles with integer spin. For 10 points, who is this German physicist best known for formulating the

Sample Question 1

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity, so Debye adjusted the theory for low temperatures. His summation convention automatically sums repeated indices in tensor products. His name is attached to the A and B coefficients for spontaneous and stimulated emission, the subject of one of his multiple groundbreaking 1905 papers. He further developed the model of statistics sent to him by Bose to describe particles with integer spin. For 10 points, who is this German physicist best known for formulating the special and general theories of relativity?

Sample Question 1

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity, so Debye adjusted the theory for low temperatures. His summation convention automatically sums repeated indices in tensor products. His name is attached to the A and B coefficients for spontaneous and stimulated emission, the subject of one of his multiple groundbreaking 1905 papers. He further developed the model of statistics sent to him by Bose to describe particles with integer spin. For 10 points, who is this German physicist best known for formulating the special and general theories of relativity?

Albert Einstein

Sample Question 1

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity, so Debye adjusted the theory for low temperatures. His summation convention

Faster = Smarter

1. Colorado School of Mines
2. Brigham Young University
3. California Institute of Technology
4. Harvey Mudd College
5. University of Colorado

Albert Einstein

Humans doing Incremental Classification

- This is **not** Jeopardy (Watson)
- There are buzzers, but players can only buzz at the end of a question
- Doesn't discriminate knowledge
- Quiz bowl questions are pyramidal

Humans doing Incremental Classification

- Thousands of questions are written every year
- Large question databases
- Teams practice on these questions (some online, e.g. IRC)
- How can we learn from this?

System for Incremental Classifiers

- Treat this as a MDP
- Action: **buzz** now or **wait**

System for Incremental Classifiers

- Treat this as a MDP
- Action: **buzz** now or **wait**
  1. **Content Model** is constantly generating guesses
  2. **Oracle** provides examples where it is correct
  3. The **Policy** generalizes to test data
  4. **Features** represent our state

content model oracle policy features

System for Incremental Classifiers

- Treat this as a MDP
- Action: **buzz** now or **wait**
  1. **Content Model** is constantly generating guesses
  2. **Oracle** provides examples where it is correct
  3. The **Policy** generalizes to test data
  4. **Features** represent our state

content model oracle policy features

Content Model

content model oracle policy features

- Bayesian generative model with answers as latent state
- Unambiguous Wikipedia pages
- Unigram term weightings (naïve Bayes, BM25)
- Maintains posterior distribution over guesses
- Always has a guess of what it should answer
  - policy will tell us when to trust it

Vector Space Model

Qatar

From Wikipedia, the free encyclopedia

For other places with the same name, see Qatar (disambiguation).

Qatar (/ˈkɑːtɑːr/, /ˈkʌtər/ or /kəˈtɑːr/;[6] Arabic: قطر *Qatar* [qaṭ'ar]; local
the **State of Qatar** (Arabic: دولة قطر *Dawlat Qatar*), is a sovereign Arab
the small Qatar Peninsula on the northeastern coast of the Arabian Penir
to the south, with the rest of its territory surrounded by the Persian Gulf.
from the nearby island kingdom of Bahrain. In 2013, Qatar's total populat
and 1.5 million expatriates.[8]

Vector Space Model

Qatar

From Wikipedia, the free encyclopedia

For other places with the same name, see Qatar (disambiguation).

Qatar (/ˈkɑːtɑːr/, /ˈkʌtər/ or /kɑːˈtɑːr/;[6] Arabic: قطر *Qatar* [qaṭ'ar]; local
the **State of Qatar** (Arabic: دولة قطر *Dawlat Qatar*), is a sovereign Arab
the small Qatar Peninsula on the northeastern coast of the Arabian Penir
to the south, with the rest of its territory surrounded by the Persian Gulf.
from the nearby island kingdom of Bahrain. In 2013, Qatar's total populat
and 1.5 million expatriates.[8]

arabian

persian

gulf

kingdom

expatriates

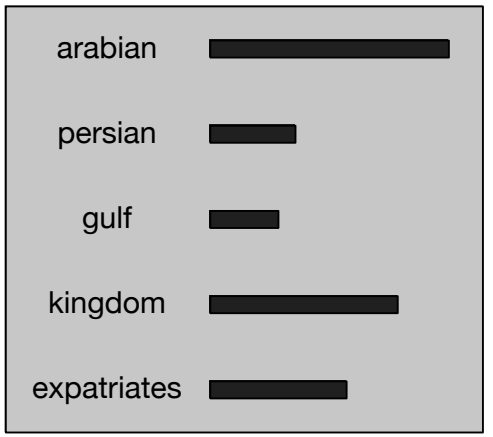
Vector Space Model

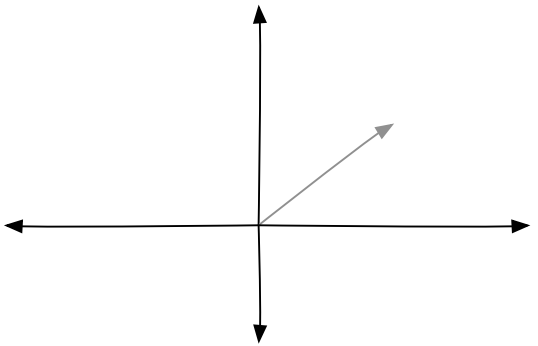
Qatar

From Wikipedia, the free encyclopedia

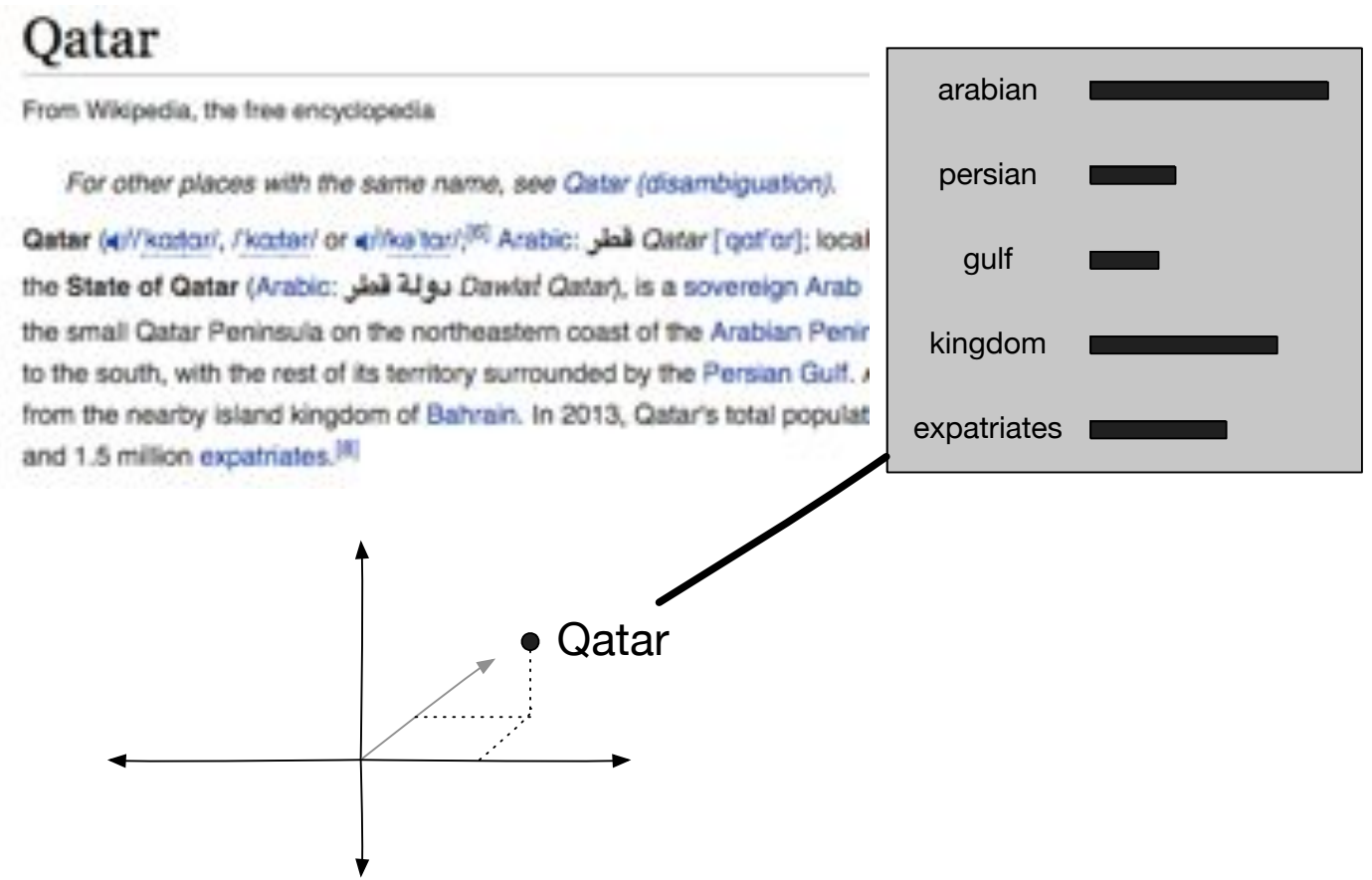
For other places with the same name, see Qatar (disambiguation).

Qatar (/ˈkɑːtɑːr/, /ˈkɑtər/ or /kɑˈtɑːr/;[5] Arabic: قطر *Qatar* [qɑtˤɑr]; local
the **State of Qatar** (Arabic: دولة قطر *Dawlat Qatar*), is a sovereign Arab
the small Qatar Peninsula on the northeastern coast of the Arabian Penir
to the south, with the rest of its territory surrounded by the Persian Gulf.
from the nearby island kingdom of Bahrain. In 2013, Qatar's total populat
and 1.5 million expatriates.[8]





Vector Space Model



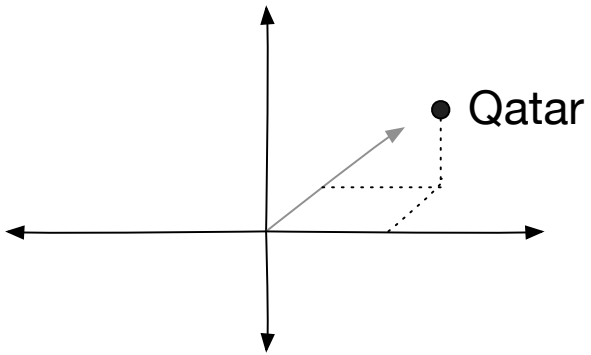
Vector Space Model

Qatar

From Wikipedia, the free encyclopedia

For other places with the same name, see Qatar (disambiguation).

Qatar (/ˈkɑːtɑːr/, /kʌtɑːr/ or /kəˈtɑːr/;[6] Arabic: قطر *Qatar* [qɑtˤɑr]; local
the **State of Qatar** (Arabic: دولة قطر *Dawlat Qatar*), is a sovereign Arab
the small Qatar Peninsula on the northeastern coast of the Arabian Penir
to the south, with the rest of its territory surrounded by the Persian Gulf.
from the nearby island kingdom of Bahrain. In 2013, Qatar's total populat
and 1.5 million expatriates.[8]



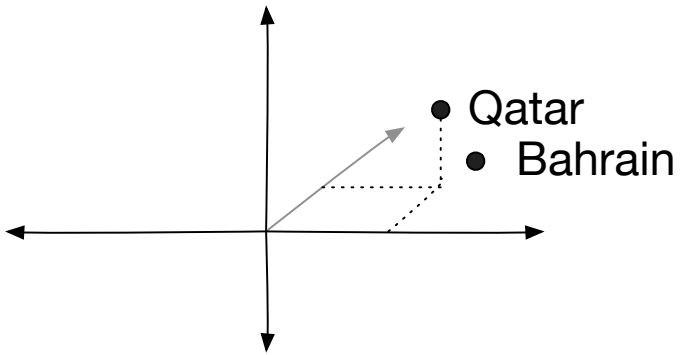
Vector Space Model

Qatar

From Wikipedia, the free encyclopedia

For other places with the same name, see Qatar (disambiguation).

Qatar (/ˈkɑːtɑːr/, /ˈkʌtər/ or /kɑːˈtɑːr/;[6] Arabic: قطر *Qatar* [qatˤar]; local
the **State of Qatar** (Arabic: دولة قطر *Dawlat Qatar*), is a sovereign Arab
the small Qatar Peninsula on the northeastern coast of the Arabian Penir
to the south, with the rest of its territory surrounded by the Persian Gulf.
from the nearby island kingdom of Bahrain. In 2013, Qatar's total populat
and 1.5 million expatriates.[8]



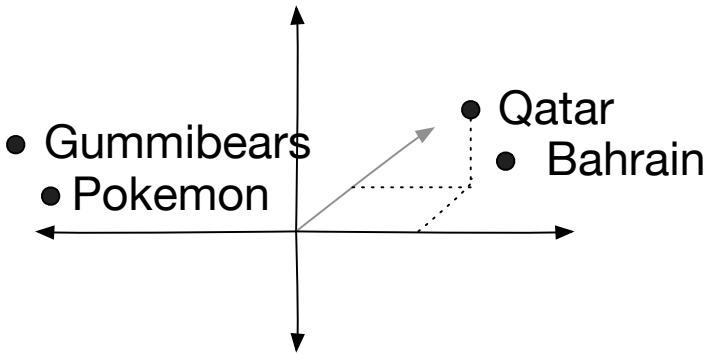
Vector Space Model

Qatar

From Wikipedia, the free encyclopedia

For other places with the same name, see Qatar (disambiguation).

Qatar (/ˈkɑːtɑːr/, /ˈkʌtər/ or /kɑːˈtɑːr/;[6] Arabic: قطر *Qatar* [qɑtˤɑr]; local
the **State of Qatar** (Arabic: دولة قطر *Dawlat Qatar*), is a sovereign Arab
the small Qatar Peninsula on the northeastern coast of the Arabian Penir
to the south, with the rest of its territory surrounded by the Persian Gulf.
from the nearby island kingdom of Bahrain. In 2013, Qatar's total populat
and 1.5 million expatriates.[8]



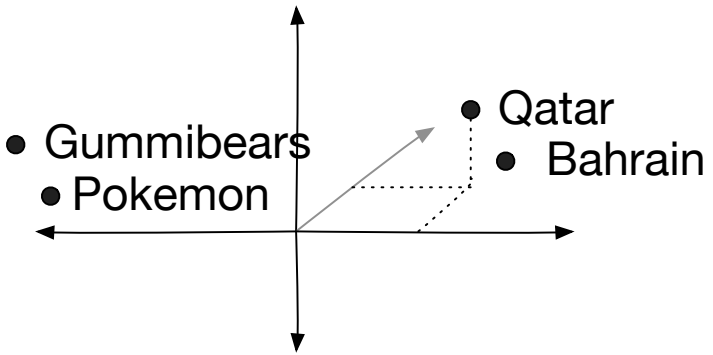
Vector Space Model

Qatar

From Wikipedia, the free encyclopedia

For other places with the same name, see Qatar (disambiguation).

Qatar (/ˈkɑːtɑːr/, /ˈkʌtər/ or /kɑːˈtɑːr/;[6] Arabic: قطر *Qatar* [qaṭ'ar]; local
the **State of Qatar** (Arabic: دولة قطر *Dawlat Qatar*), is a sovereign Arab
the small Qatar Peninsula on the northeastern coast of the Arabian Penir
to the south, with the rest of its territory surrounded by the Persian Gulf.
from the nearby island kingdom of Bahrain. In 2013, Qatar's total populat
and 1.5 million expatriates.[8]



?

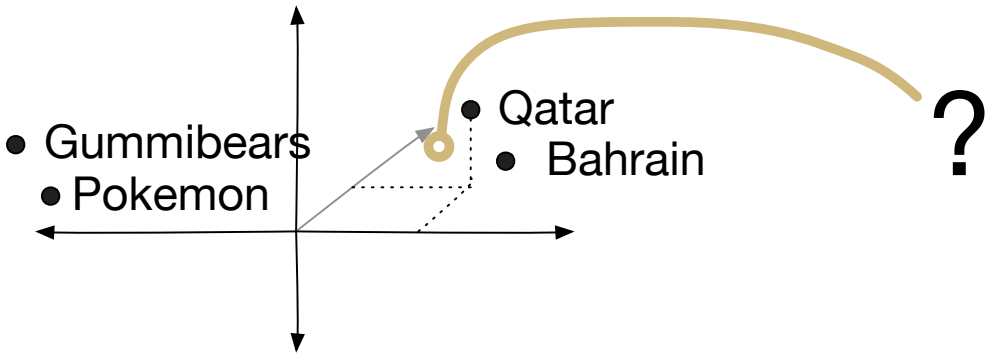
Vector Space Model

Qatar

From Wikipedia, the free encyclopedia

For other places with the same name, see Qatar (disambiguation).

Qatar (/ˈkɑːtɑːr/, /ˈkʌtər/ or /kɑːˈtɑːr/;[6] Arabic: قطر *Qatar* [qɑtˤɑr]; local
the **State of Qatar** (Arabic: دولة قطر *Dawlat Qatar*), is a sovereign Arab
the small Qatar Peninsula on the northeastern coast of the Arabian Penir
to the south, with the rest of its territory surrounded by the Persian Gulf.
from the nearby island kingdom of Bahrain. In 2013, Qatar's total populat
and 1.5 million expatriates.[8]



Oracle

content model oracle policy features

Revealed Text	Content Model	Answer	Oracle
The National Endowment for the Arts, War o	Martha Graham	Lyndon Johnson	✗
The National Endowment for the Arts, War on Poverty, and Medicare were establish by, for 10 points, what Texan	George Bush	Lyndon Johnson	✗
The National Endowment for the Arts, War on Poverty, and Medicare were establish by, for 10 points, what Texan who defeated Barry Goldwater, promoted the Great Society, and succeeded John F. Kennedy?	Lyndon Johnson	Lyndon Johnson	✓

- As each token is revealed, look at content model's guess
- If it's right, positive instance; otherwise negative
- Nearly optimal policy to buzz whenever correct (upper bound)

Policy

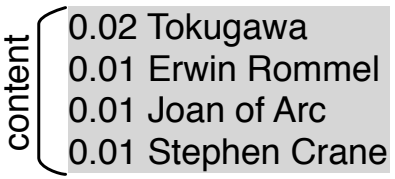
content model oracle **policy** features

- Mapping: state $\mapsto$ action
- Use oracle as example actions
- Learned as classifier (Langford et al., 2005)
- At test time, use the same features as for training
  - Question text (so far)
  - Guess
  - Posterior distribution
  - Change in posterior

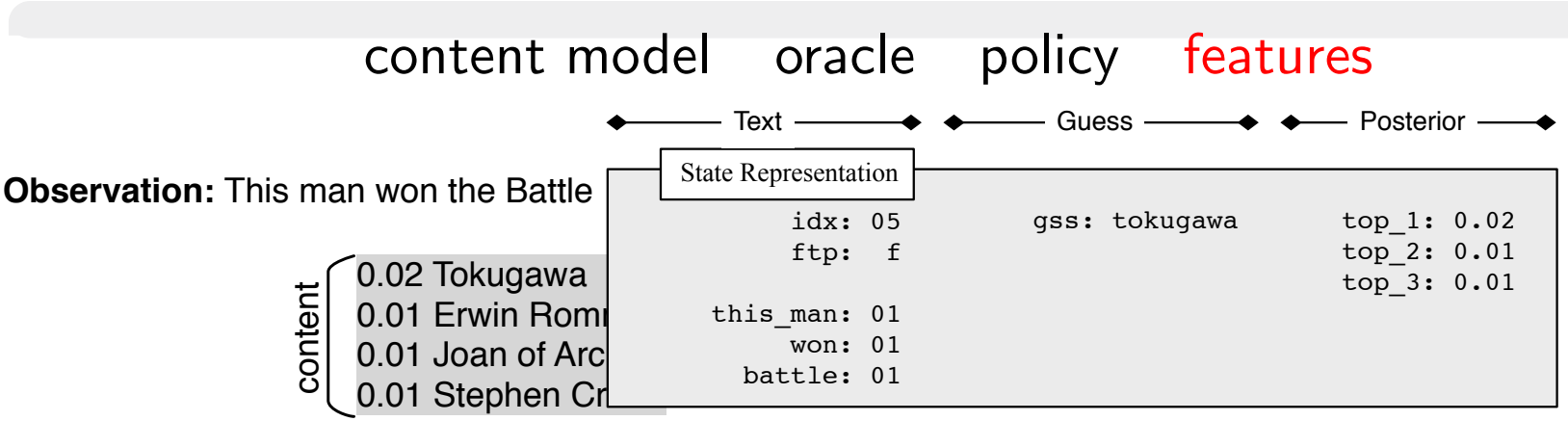
Features (by example)

content model oracle policy features

Observation: This man won the Battle

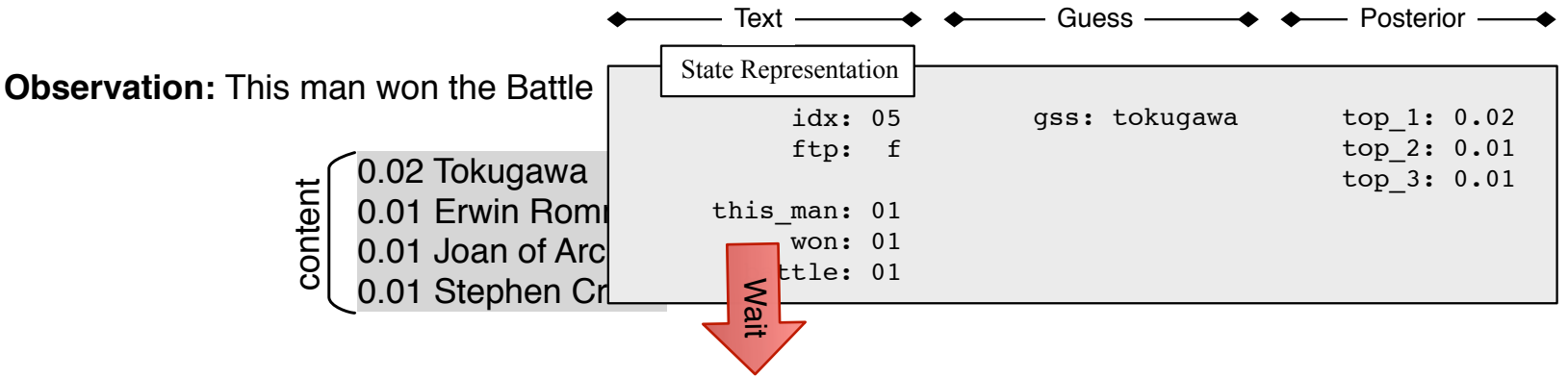


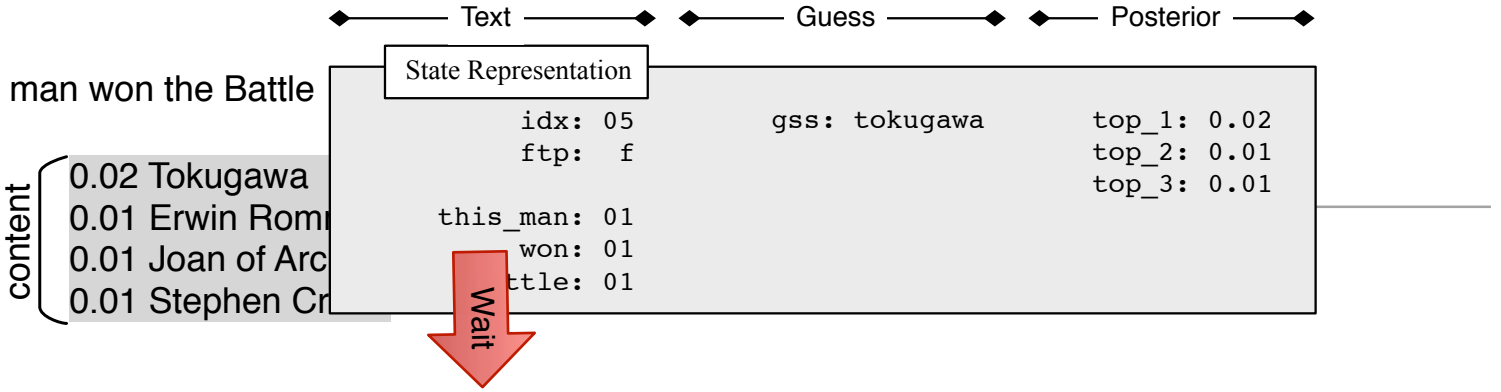
Features (by example)



Features (by example)

content model oracle policy features

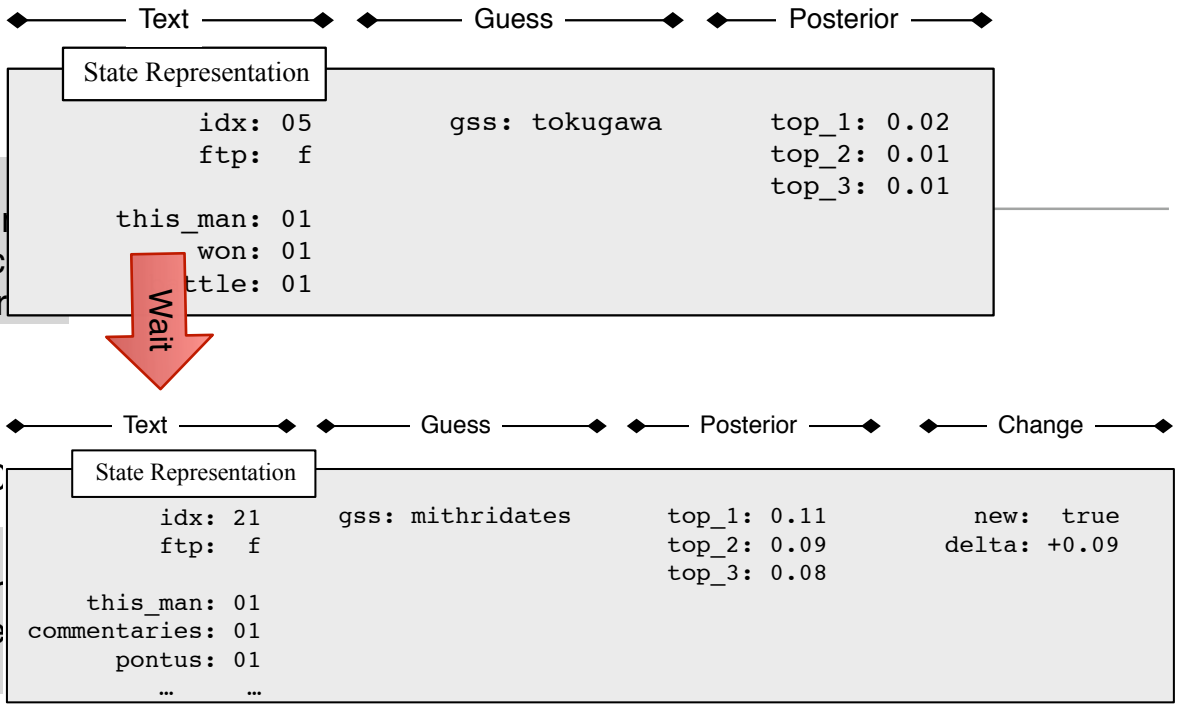


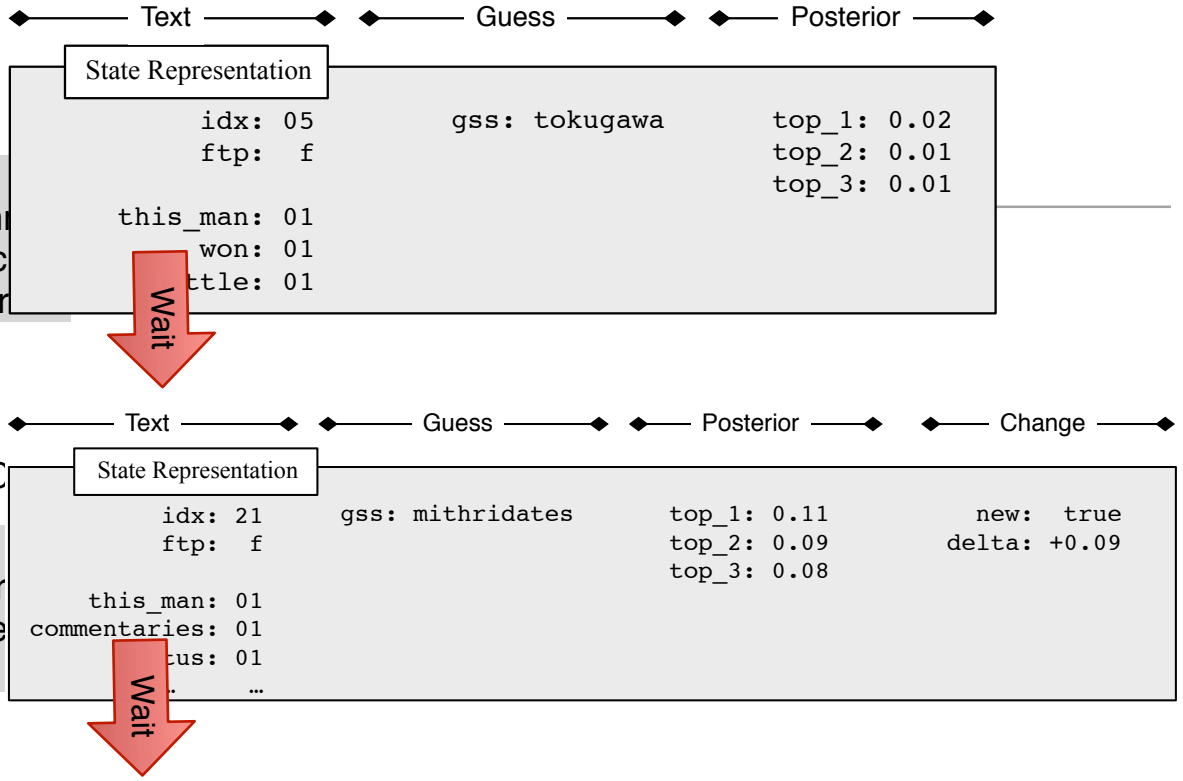


Observation: This man won the Battle

Observation: This man won the Battle of Zela over Pontus. He wrote about his victory at Alesia in his Commentaries on the

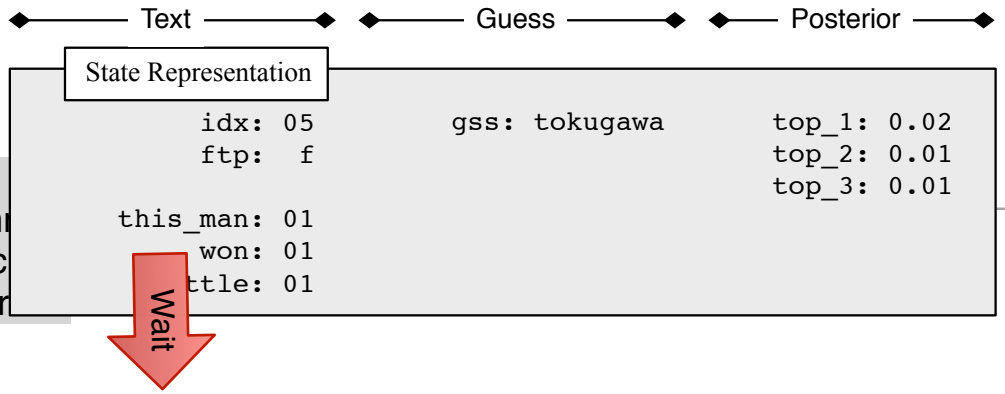
0.11 Mithridates
0.09 Julius Caesar
0.08 Alexander the Great
0.07 Sulla


Text
Guess
Posterior
State Representation
idx: 05
gss: tokugawa
top_1: 0.02
ftp: f
top_2: 0.01
top_3: 0.01
this_man: 01
won: 01
Wait
Observation: This man won the Battle
content
0.02 Tokugawa
0.01 Erwin Rom
0.01 Joan of Arc
0.01 Stephen Cr
Text
Guess
Posterior
Change
State Representation
idx: 21
gss: mithridates
top_1: 0.11
new: true
ftp: f
top_2: 0.09
delta: +0.09
top_3: 0.08
this_man: 01
commentaries: 01
pontus: 01
Observation: This man won the Battle
wrote about his victory at Alesia in his C
0.11 Mithridates
0.09 Julius Caesar
0.08 Alexander the
0.07 Sulla



Observation: This man won the Battle

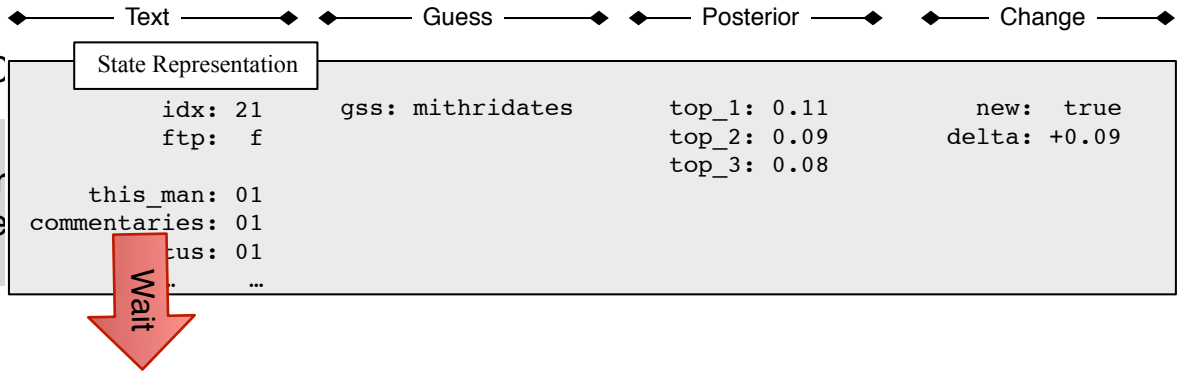
Observation: This man won the Battle
wrote about his victory at Alesia in his C



Observation: This man won the Battle

content

0.02 Tokugawa
0.01 Erwin Romm
0.01 Joan of Arc
0.01 Stephen Cr

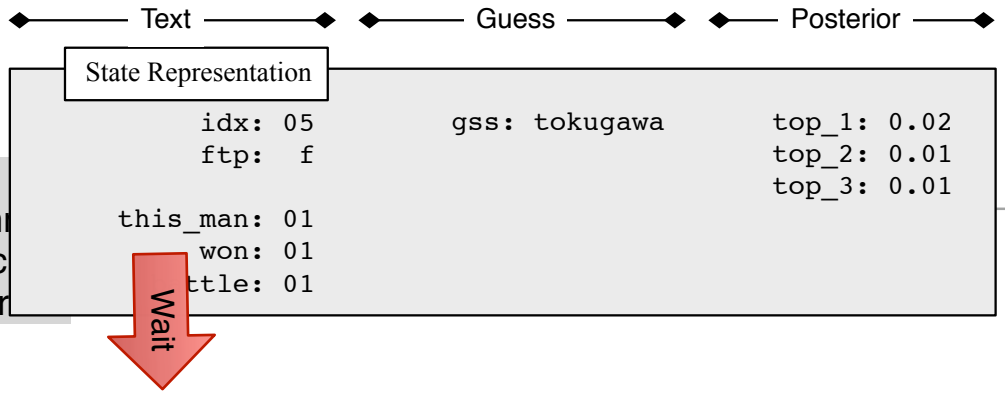


Observation: This man won the Battle
wrote about his victory at Alesia in his C

0.11 Mithridates
0.09 Julius Caesar
0.08 Alexander the
0.07 Sulla

Observation: This man won the Battle of Zela over Pontus. He wrote about his victory at Alesia in his Commentaries on the Gallic Wars. FTP, name this Roman

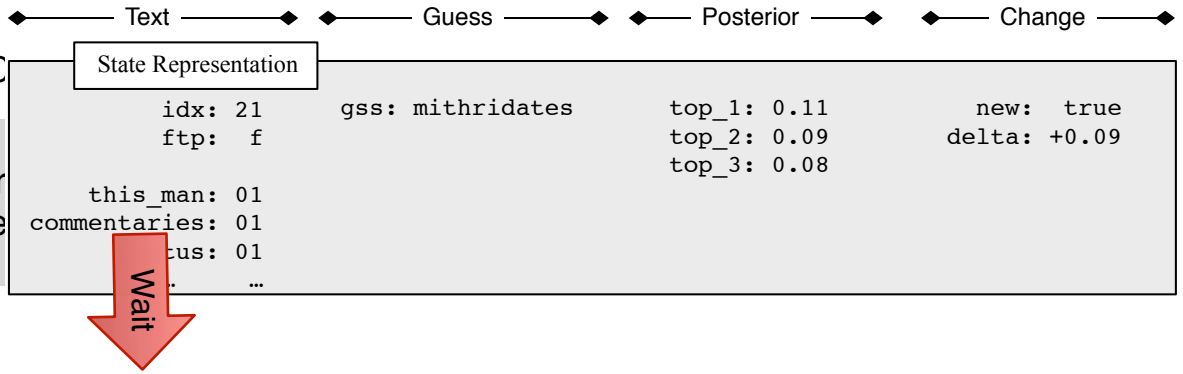
0.89 Julius Caesar
0.02 Augustus
0.01 Sulla
0.01 Pompey



Observation: This man won the Battle

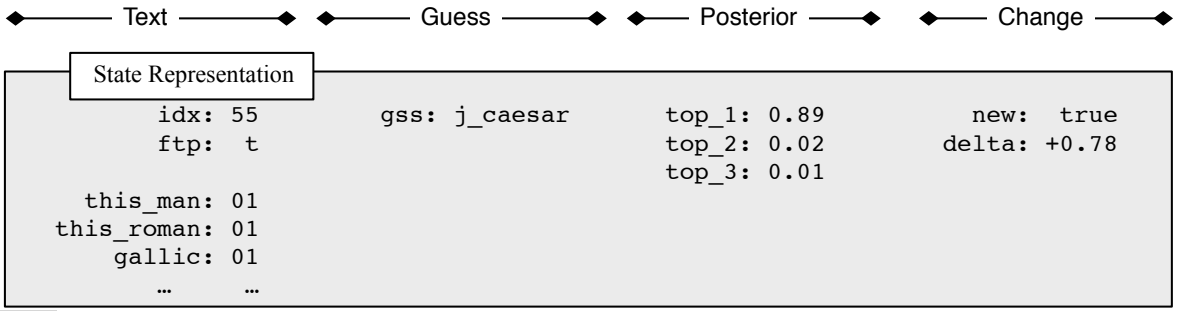
content

0.02 Tokugawa
0.01 Erwin Romm
0.01 Joan of Arc
0.01 Stephen Cr



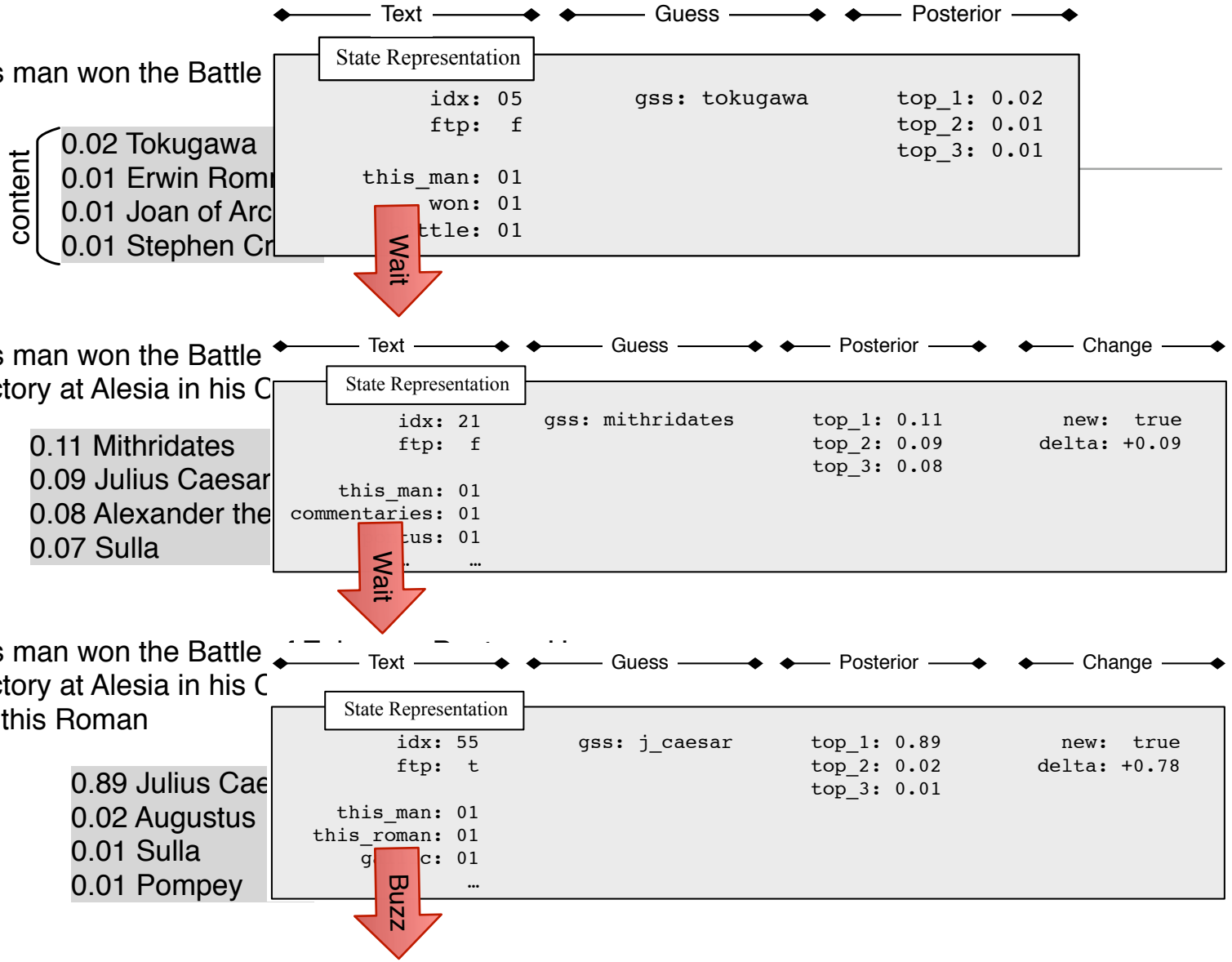
Observation: This man won the Battle
wrote about his victory at Alesia in his C

0.11 Mithridates
0.09 Julius Caesar
0.08 Alexander the
0.07 Sulla



Observation: This man won the Battle
wrote about his victory at Alesia in his C
Wars. FTP, name this Roman

0.89 Julius Cae
0.02 Augustus
0.01 Sulla
0.01 Pompey



Answer: **Julius Caesar**

Simulating a Game

- Present tokens incrementally to algorithm, see where it buzzes
- Compare to where humans buzzed in
- Payoff matrix (wrt Computer)

	Computer	Human	Payoff
1	first and wrong	right	−15
2	—	first and correct	−10
3	first and wrong	wrong	−5
4	first and correct	—	+10
5	wrong	first and wrong	+5
6	right	first and wrong	+15

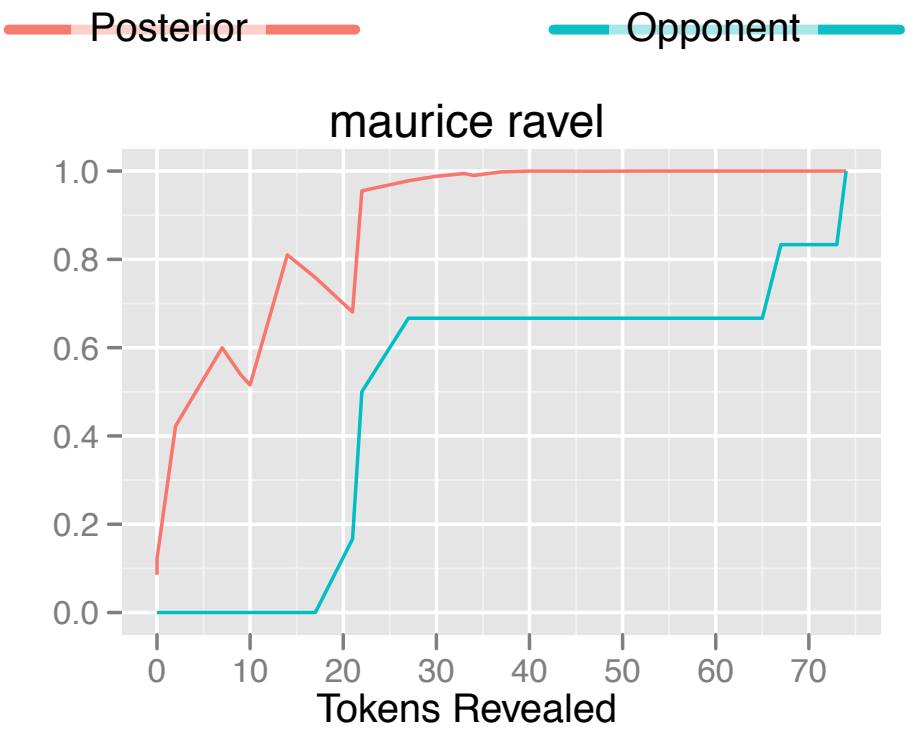
Interface

- Users could "veto" categories or tournaments
- Questions presented in canonical order
- Approximate string matching (w/ override)

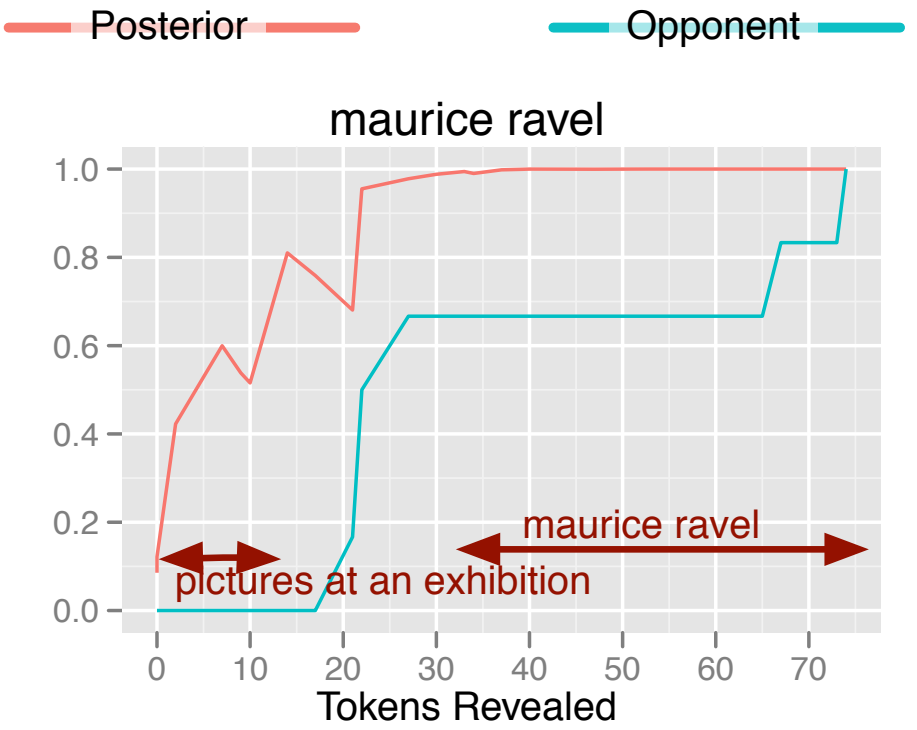
Interface

- Started on Amazon Mechanical Turk
- 7000 questions were answered in the first day
- Over 43000 questions were answered in the space of two weeks
- Total of 461 unique users
- Leaderboard to encourage users

Error Analysis



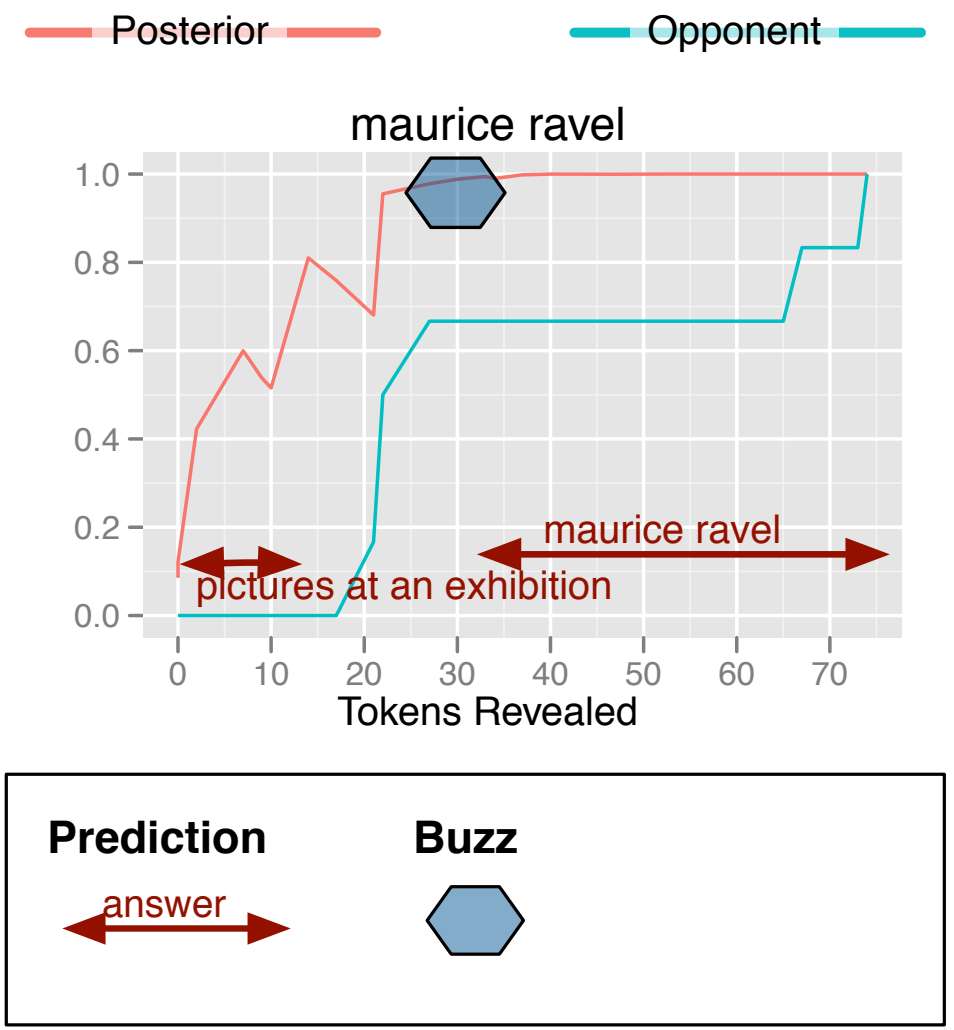
Error Analysis



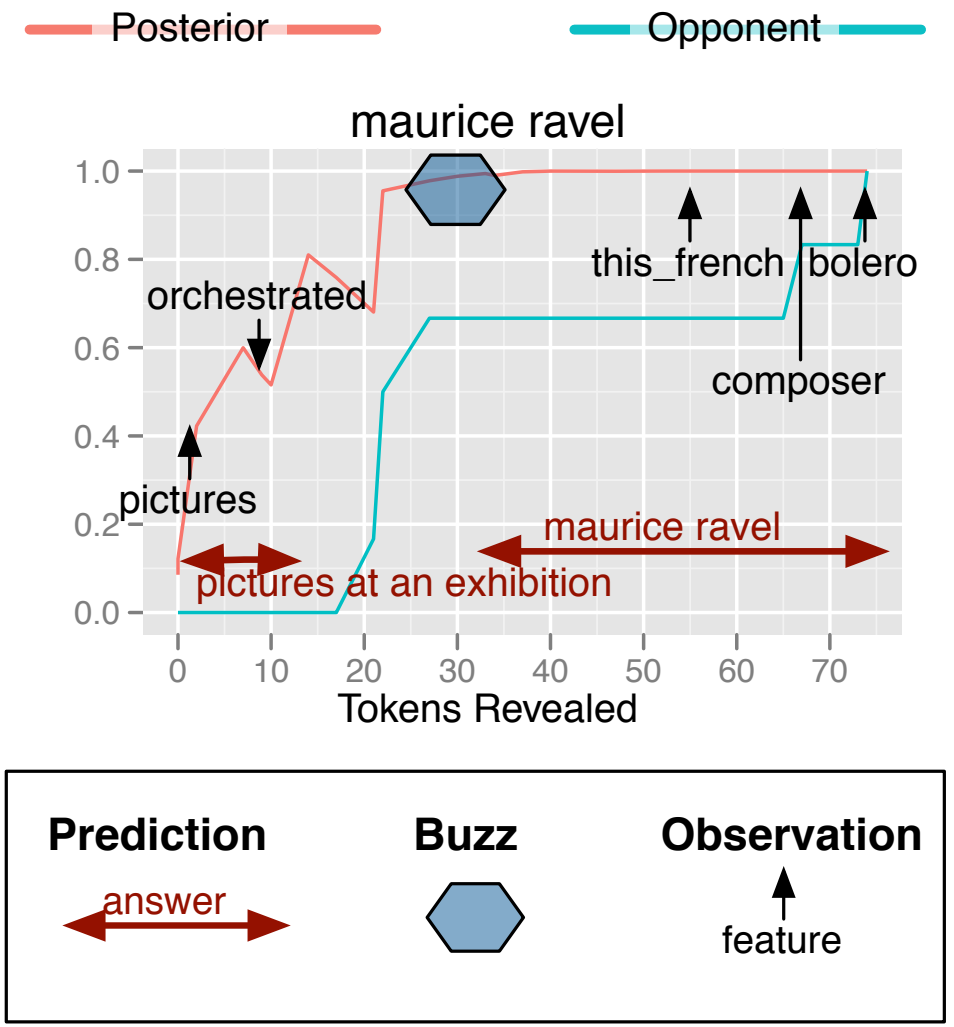
Prediction

answer

Error Analysis

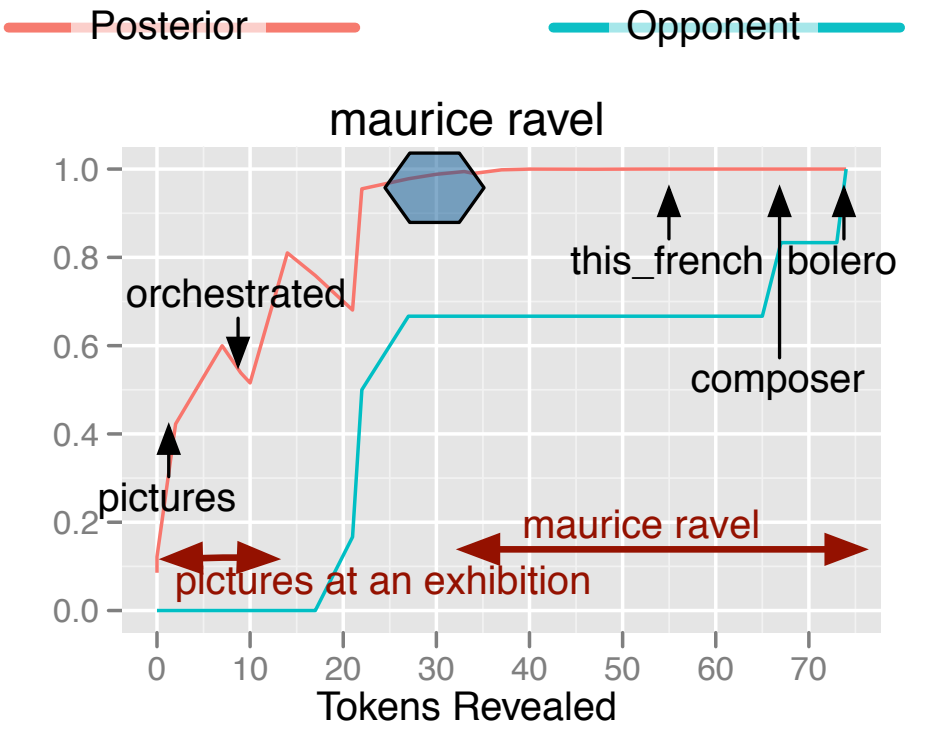


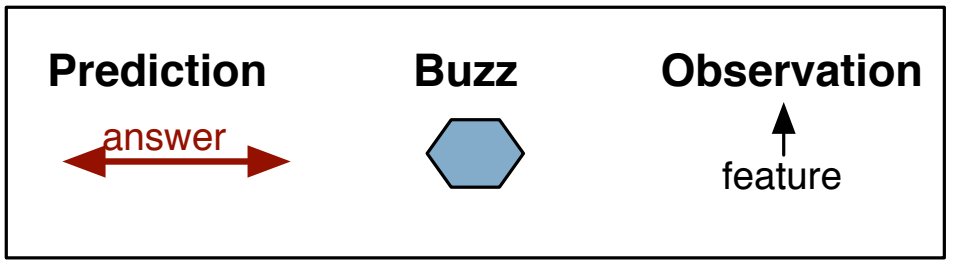
Error Analysis



Error Analysis

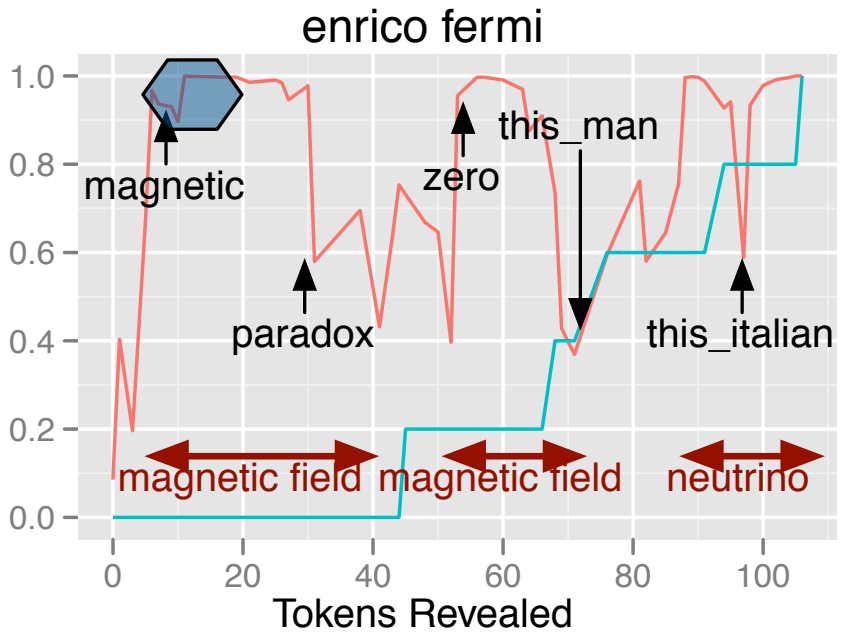
- Too slow
- Coreference and correct question type
- Not enough information / not weighting later clues higher

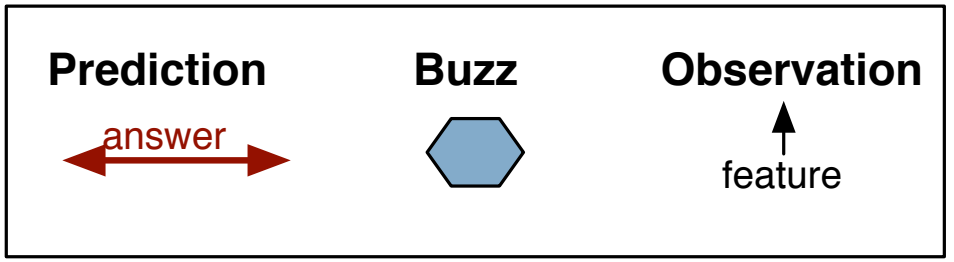




Error Analysis

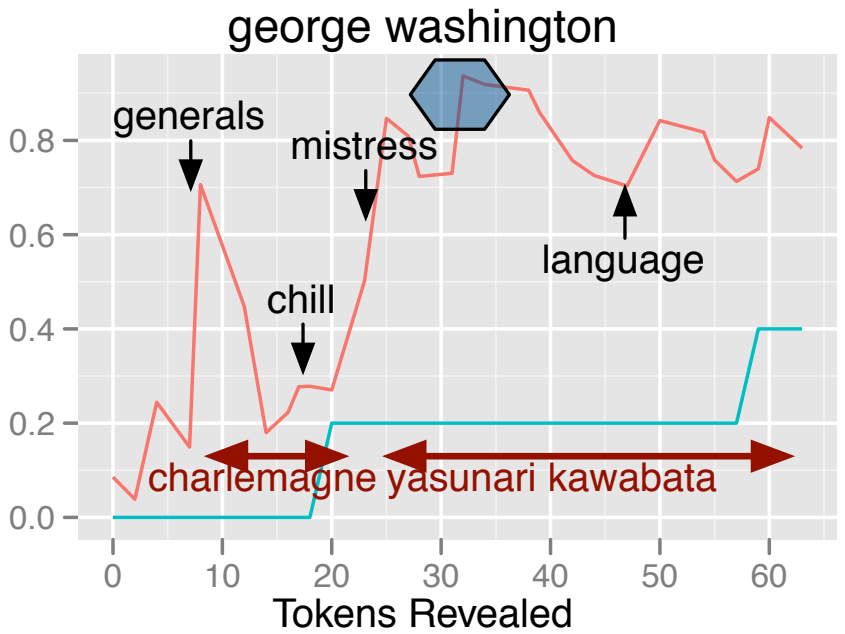
- Too slow
- Coreference and correct question type
- Not enough information / not weighting later clues higher

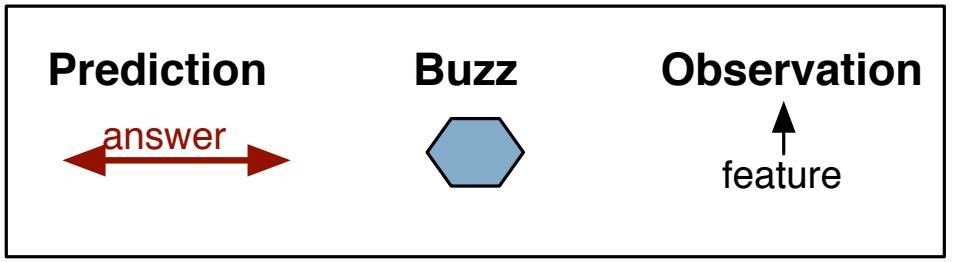




Error Analysis

- Too slow
- Coreference and correct question type
- Not enough information / not weighting later clues higher





How can we do better?

- Use order of words in a sentence "this man shot Lee Harvey Oswald" very different from "Lee Harvey Oswald shot this man"
- Use relationship between questions ("China" and "Taiwan")
- Use learned features and dimensions, not the words we start with

How can we do better?

- Use order of words in a sentence "this man shot Lee Harvey Oswald" very different from "Lee Harvey Oswald shot this man"
- Use relationship between questions ("China" and "Taiwan")
- Use learned features and dimensions, not the words we start with
- Recursive Neural Networks (Socher et al., 2012)
- First-time a *learned* representation has been applied to question answering

Plan

Why Deep Learning

Review of Logistic Regression

Can't Somebody else do it? (Feature Engineering)

Deep Learning from Data

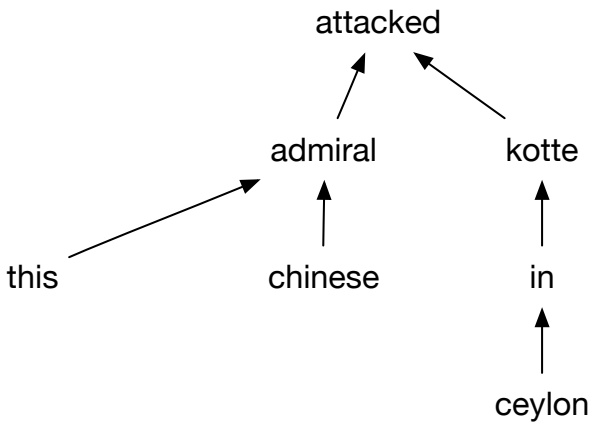
Tricks and Toolkits

Toolkits for Deep Learning

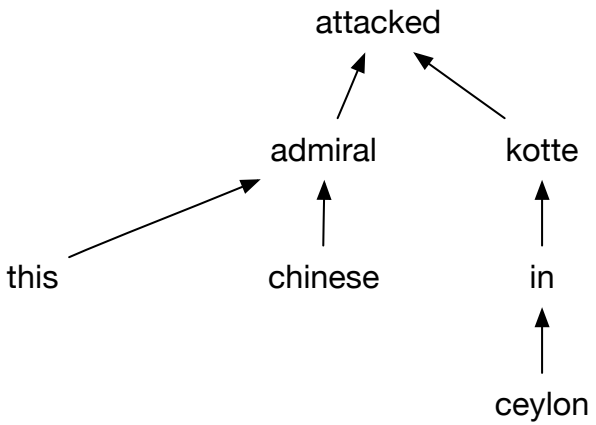
Quiz Bowl

Deep Quiz Bowl

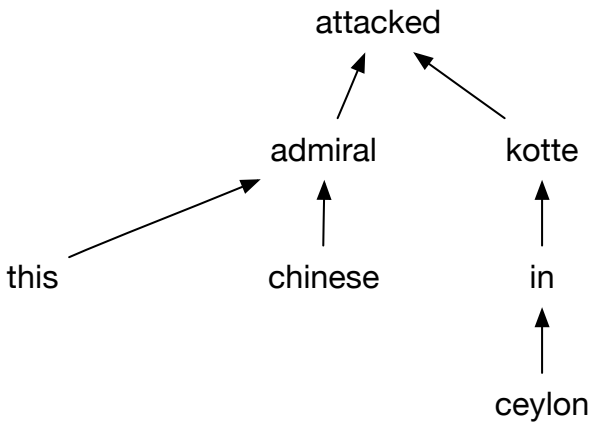
Using Compositionality

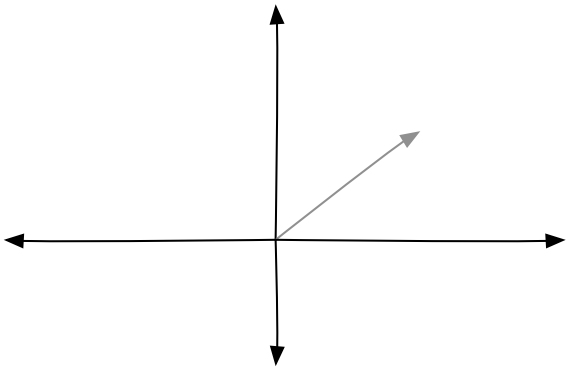


Using Compositionality

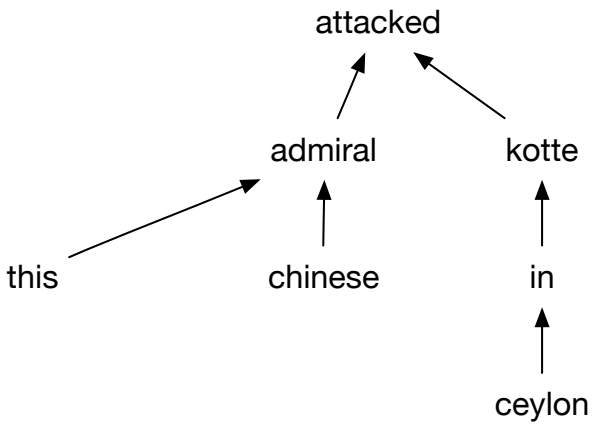


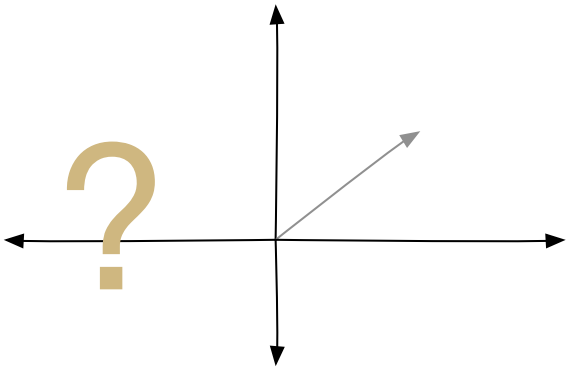
Using Compositionality



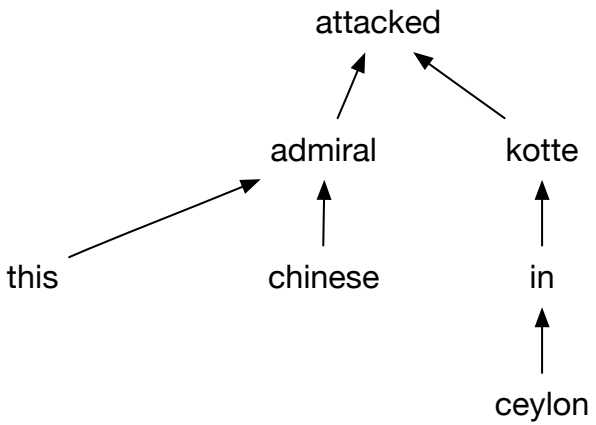


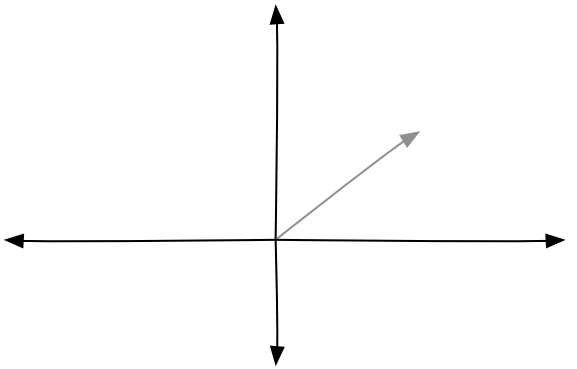
Using Compositionality



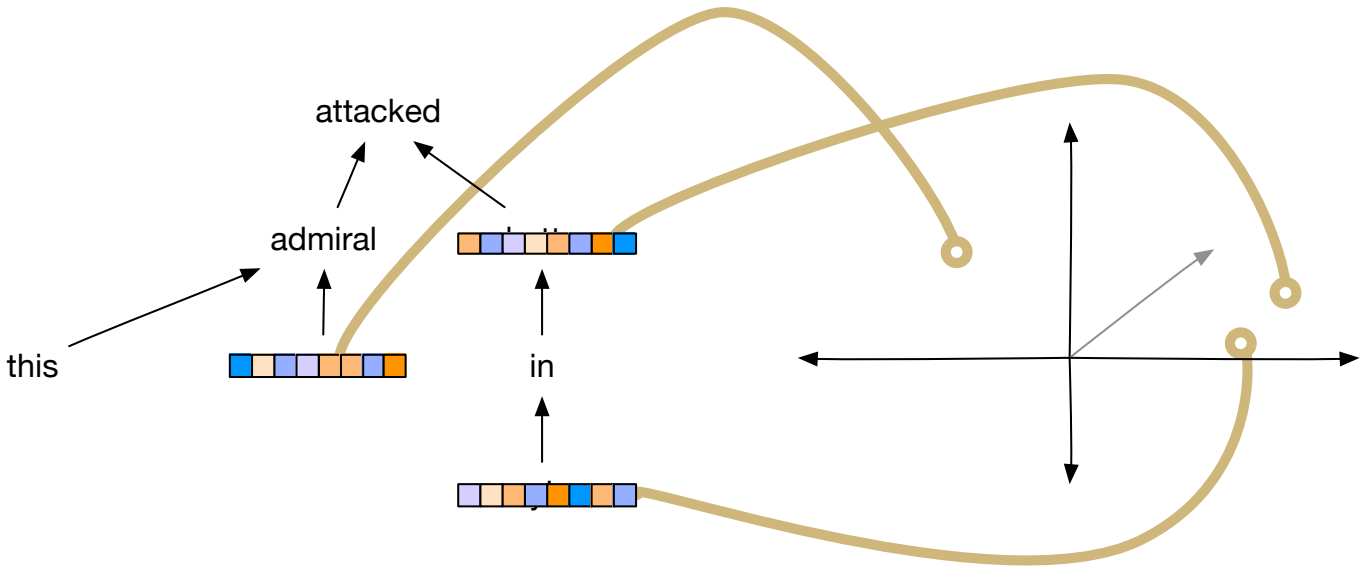


Using Compositionality

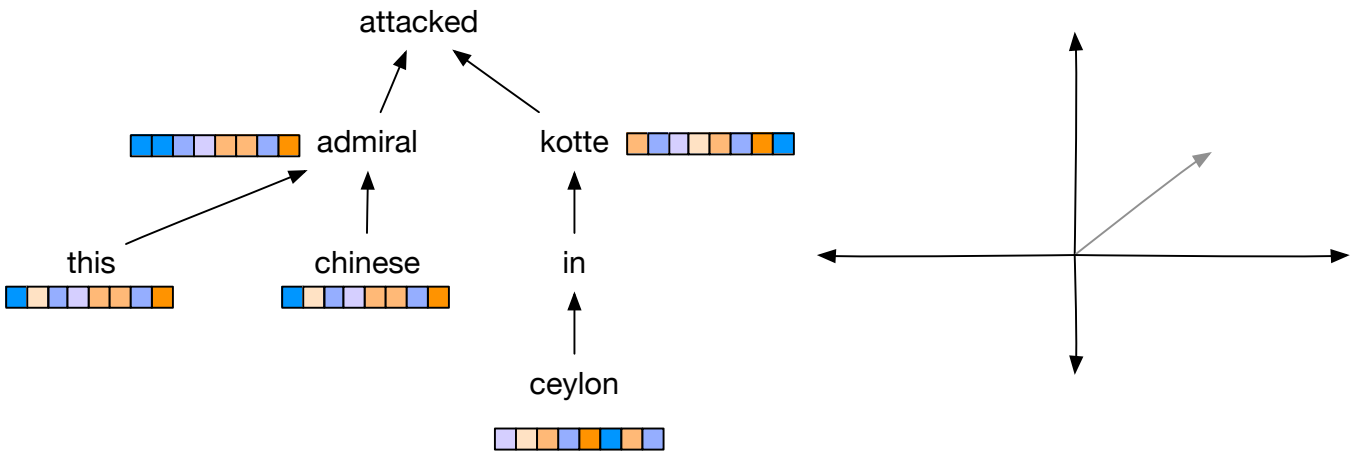




Using Compositionality

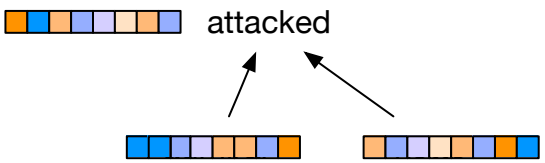


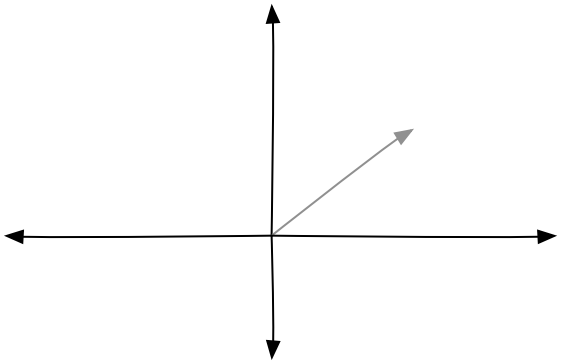
Using Compositionality



Using Compositionality

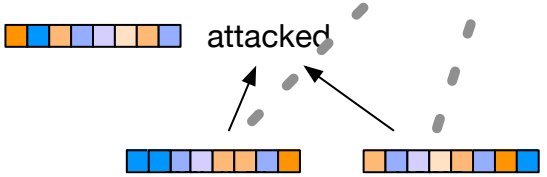
$$f(W_v \cdot x_w + b + \sum_{k \in K(n)} W_{R(n,k)} \cdot h_k) \quad =$$

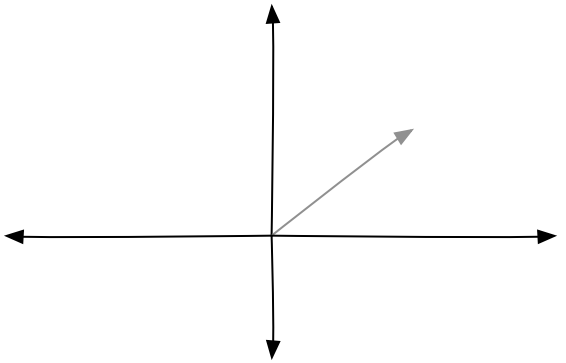




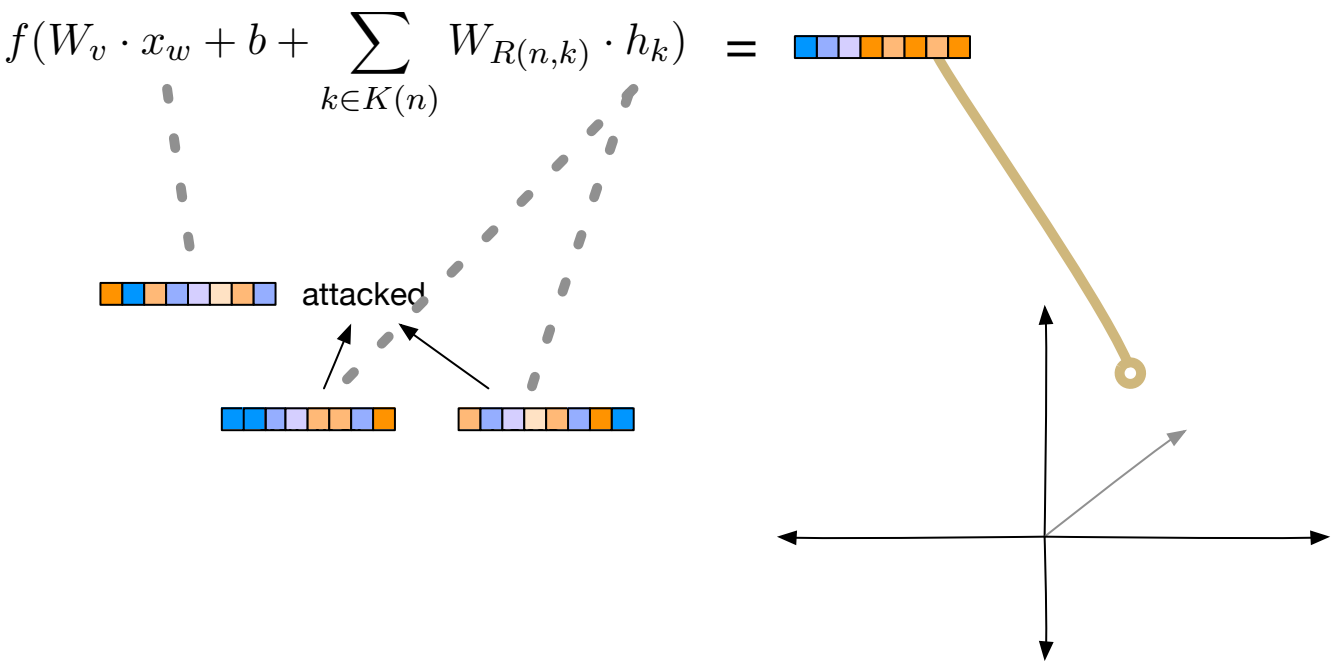
Using Compositionality

$$f(W_v \cdot x_w + b + \sum_{k \in K(n)} W_{R(n,k)} \cdot h_k) =$$

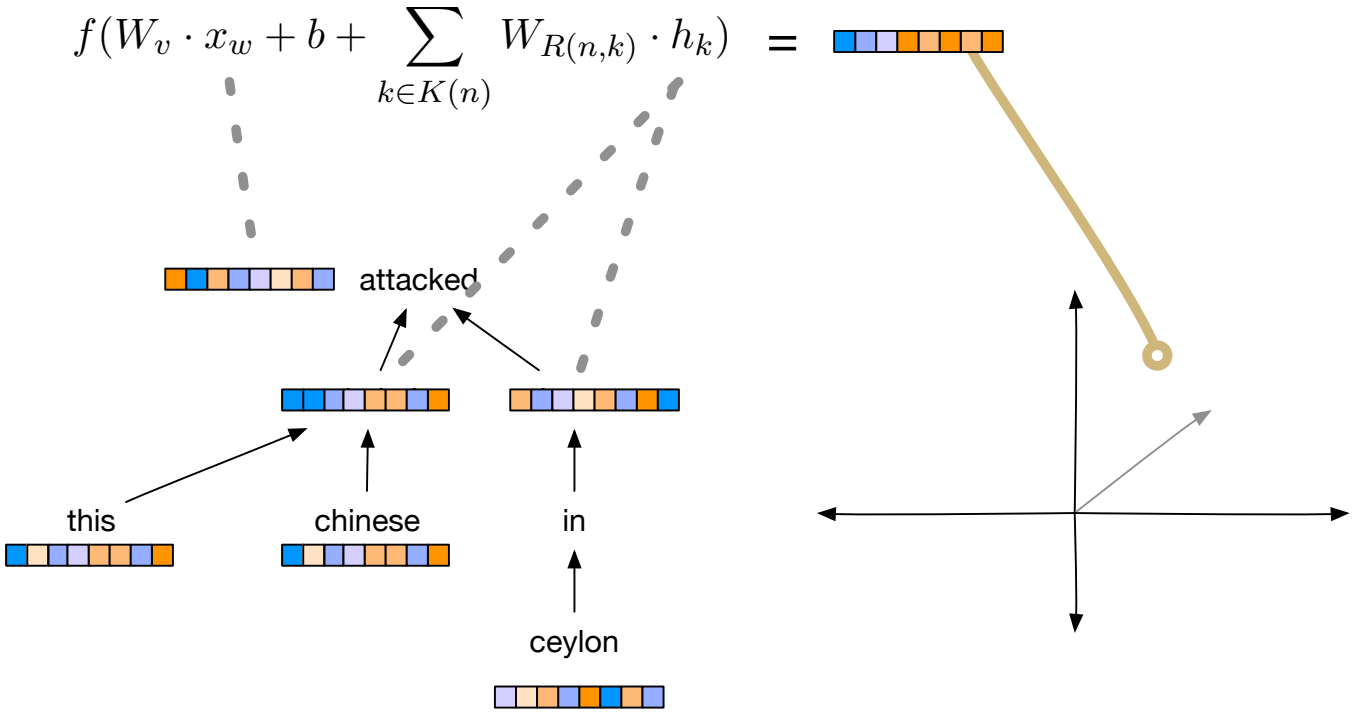




Using Compositionality

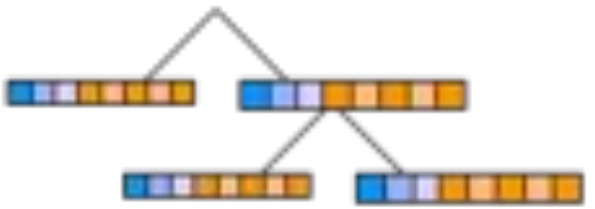


Using Compositionality



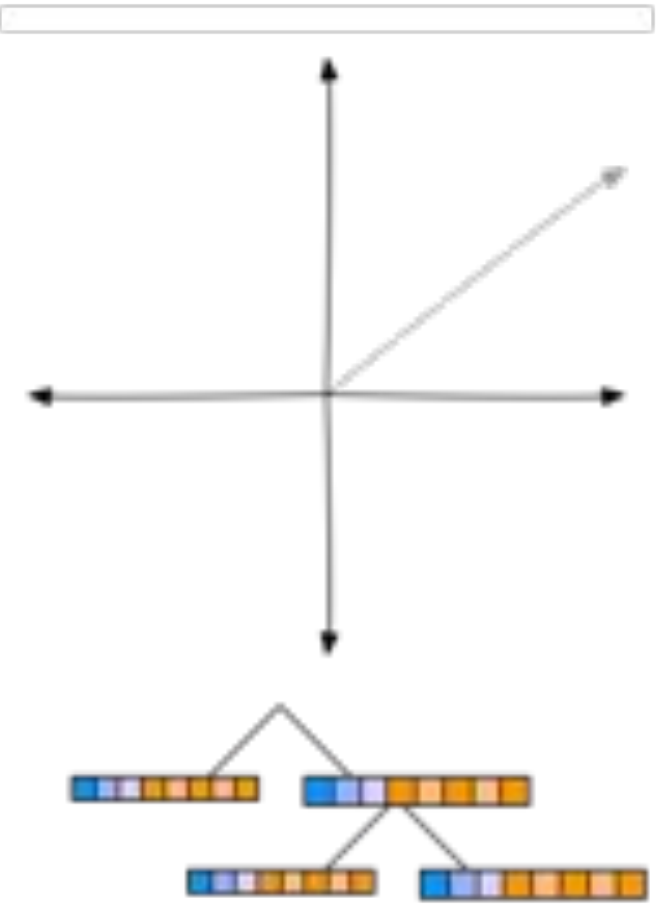
Training

- Initialize embeddings from WORD2VEC
- Randomly initialize composition matrices
- Update using WARP
  - Randomly choose an instance



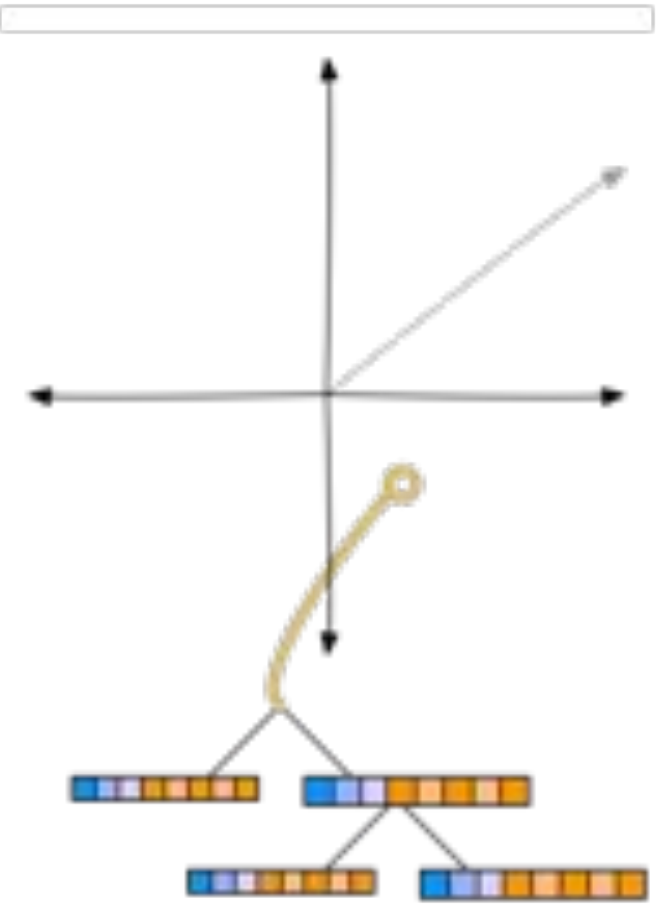
Training

- Initialize embeddings from WORD2VEC
- Randomly initialize composition matrices
- Update using WARP
  - Randomly choose an instance
  - Look where it lands



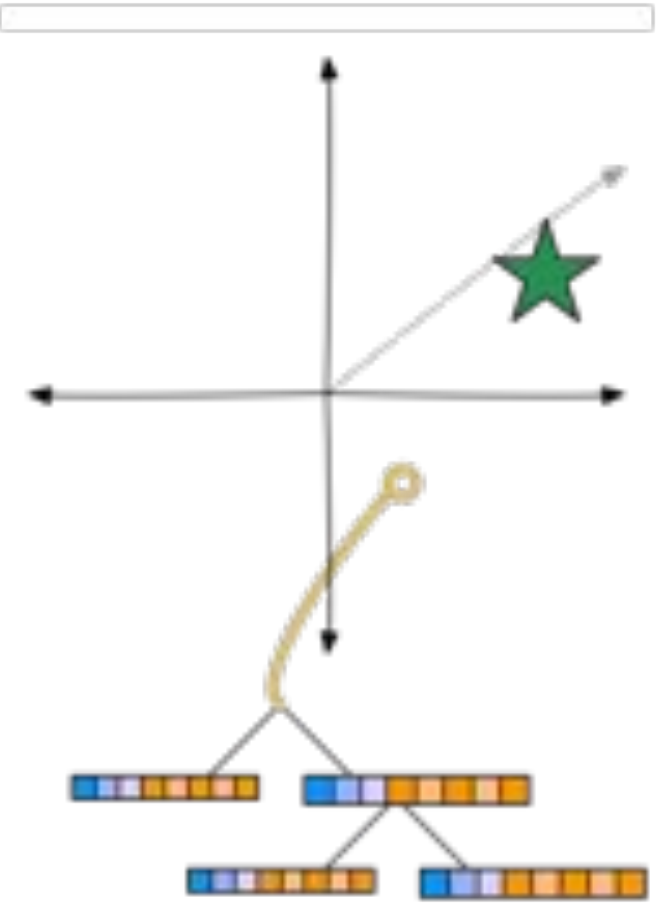
Training

- Initialize embeddings from WORD2VEC
- Randomly initialize composition matrices
- Update using WARP
  - Randomly choose an instance
  - Look where it lands



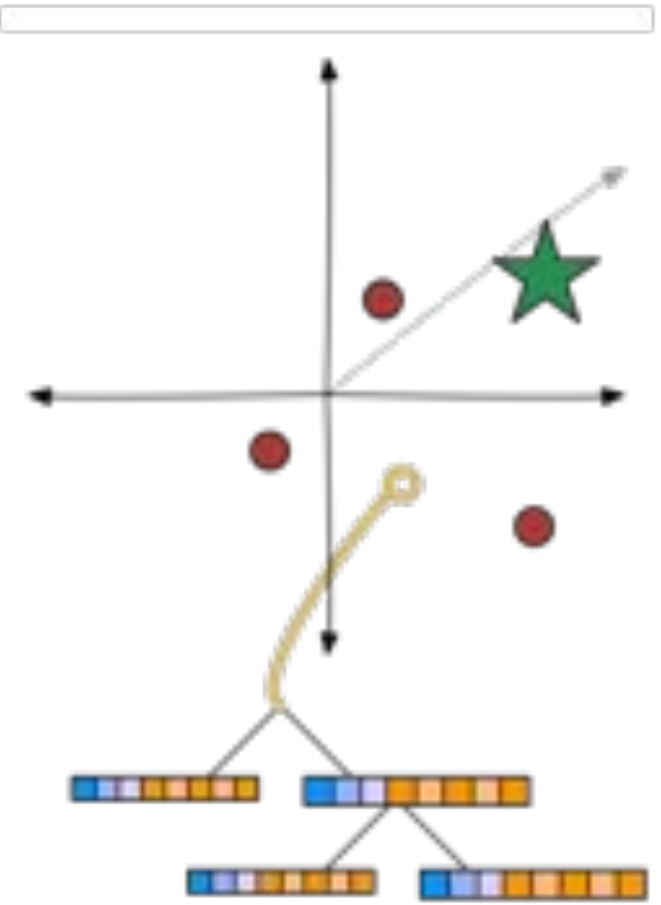
Training

- Initialize embeddings from WORD2VEC
- Randomly initialize composition matrices
- Update using WARP
  - Randomly choose an instance
  - Look where it lands
  - Has a correct answer



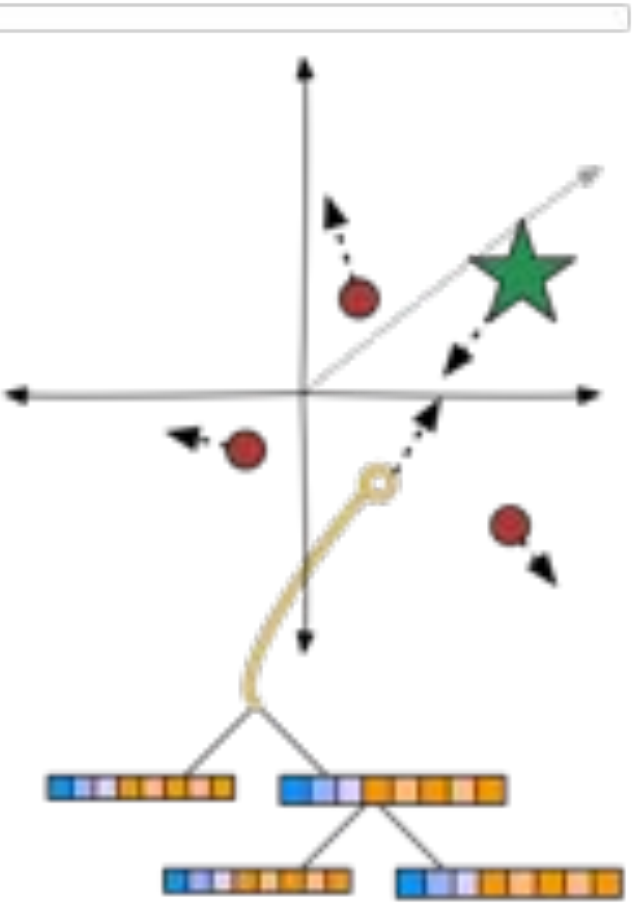
Training

- Initialize embeddings from WORD2VEC
- Randomly initialize composition matrices
- Update using WARP
  - Randomly choose an instance
  - Look where it lands
  - Has a correct answer
  - Wrong answers may be closer



Training

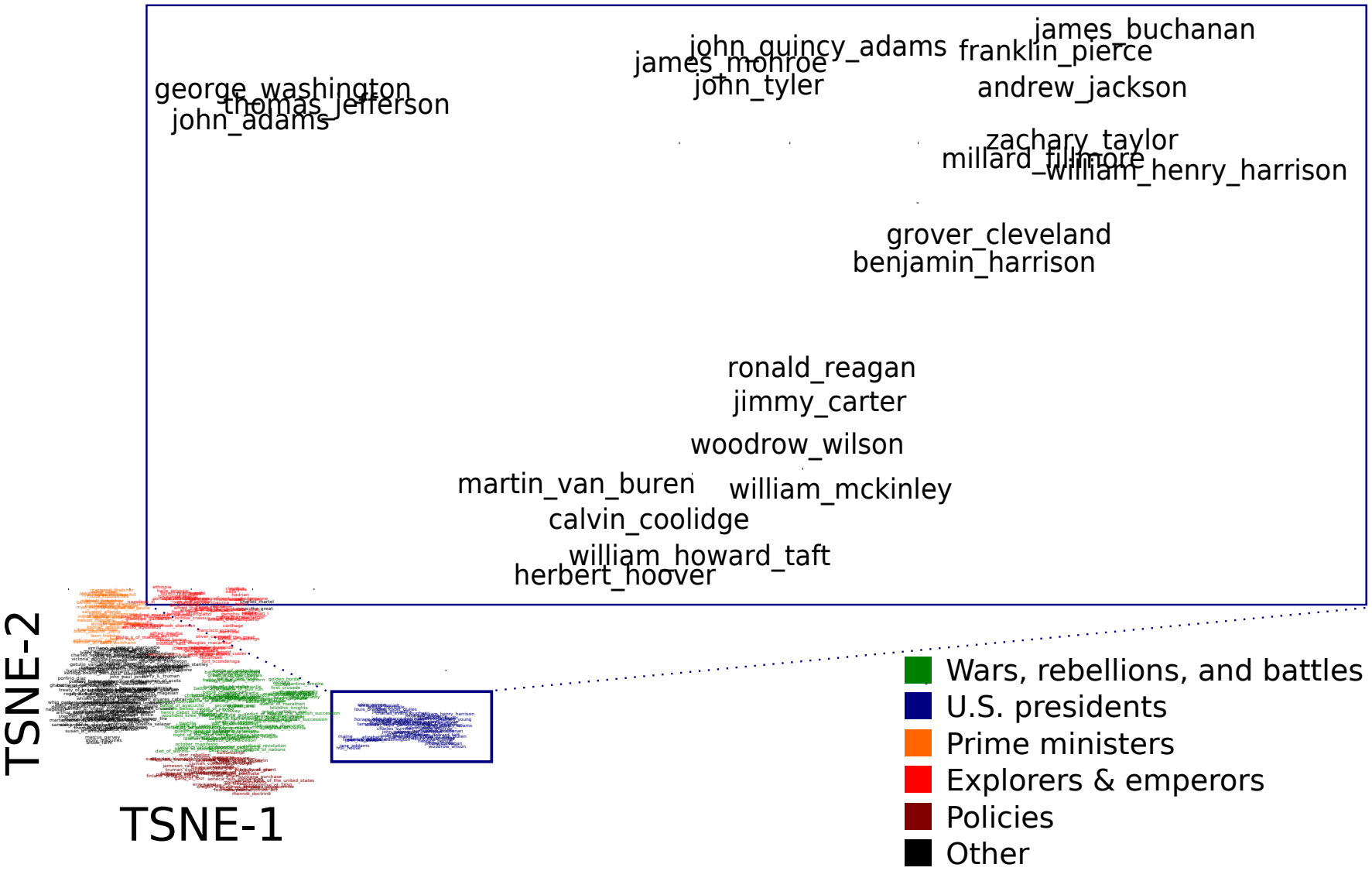
- Initialize embeddings from WORD2VEC
- Randomly initialize composition matrices
- Update using WARP
  - Randomly choose an instance
  - Look where it lands
  - Has a correct answer
  - Wrong answers may be closer
  - Push away wrong answers
  - Bring correct answers closer



Training

- We use RNN information from parsed questions
- And bag of words features from Wikipedia
- Combine both features together

Embedding


james_buchanan
john_quincy_adams
franklin_pierce
james_monroe
john_tyler
george_washington
thomas_jefferson
andrew_jackson
john_adams
zachary_taylor
millard_fillmore
william_henry_harrison
grover_cleveland
benjamin_harrison
ronald_reagan
jimmy_carter
woodrow_wilson
martin_van_buren
william_mckinley
calvin_coolidge
william_howard_taft
herbert_hoover
TSNE-2
TSNE-1
Wars, rebellions, and battles
U.S. presidents
Prime ministers
Explorers & emperors
Policies
Other

Comparing rnn to bow

	History			Literature		
Model	Sent 1	Sent 2	Full	Sent 1	Sent 2	Full
BOW-QB	37.5	65.9	71.4	27.4	54.0	61.9
RNN	**47.1**	**72.1**	**73.7**	**36.4**	**68.2**	**69.1**
BOW-WIKI	53.7	76.6	77.5	41.8	74.0	73.3
COMBINED	**59.8**	**81.8**	**82.3**	**44.7**	**78.7**	**76.6**

Percentage accuracy of different vector-space models.

Comparing rnn to bow

	History			Literature		
Model	Sent 1	Sent 2	Full	Sent 1	Sent 2	Full
BOW-QB	37.5	65.9	71.4	27.4	54.0	61.9
RNN	**47.1**	**72.1**	**73.7**	**36.4**	**68.2**	**69.1**
BOW-WIKI	53.7	76.6	77.5	41.8	74.0	73.3
COMBINED	**59.8**	**81.8**	**82.3**	**44.7**	**78.7**	**76.6**

Percentage accuracy of different vector-space models.

Comparing rnn to bow

	History			Literature		
Model	Sent 1	Sent 2	Full	Sent 1	Sent 2	Full
BOW-QB	37.5	65.9	71.4	27.4	54.0	61.9
RNN	**47.1**	**72.1**	**73.7**	**36.4**	**68.2**	**69.1**
BOW-WIKI	53.7	76.6	77.5	41.8	74.0	73.3
COMBINED	**59.8**	**81.8**	**82.3**	**44.7**	**78.7**	**76.6**

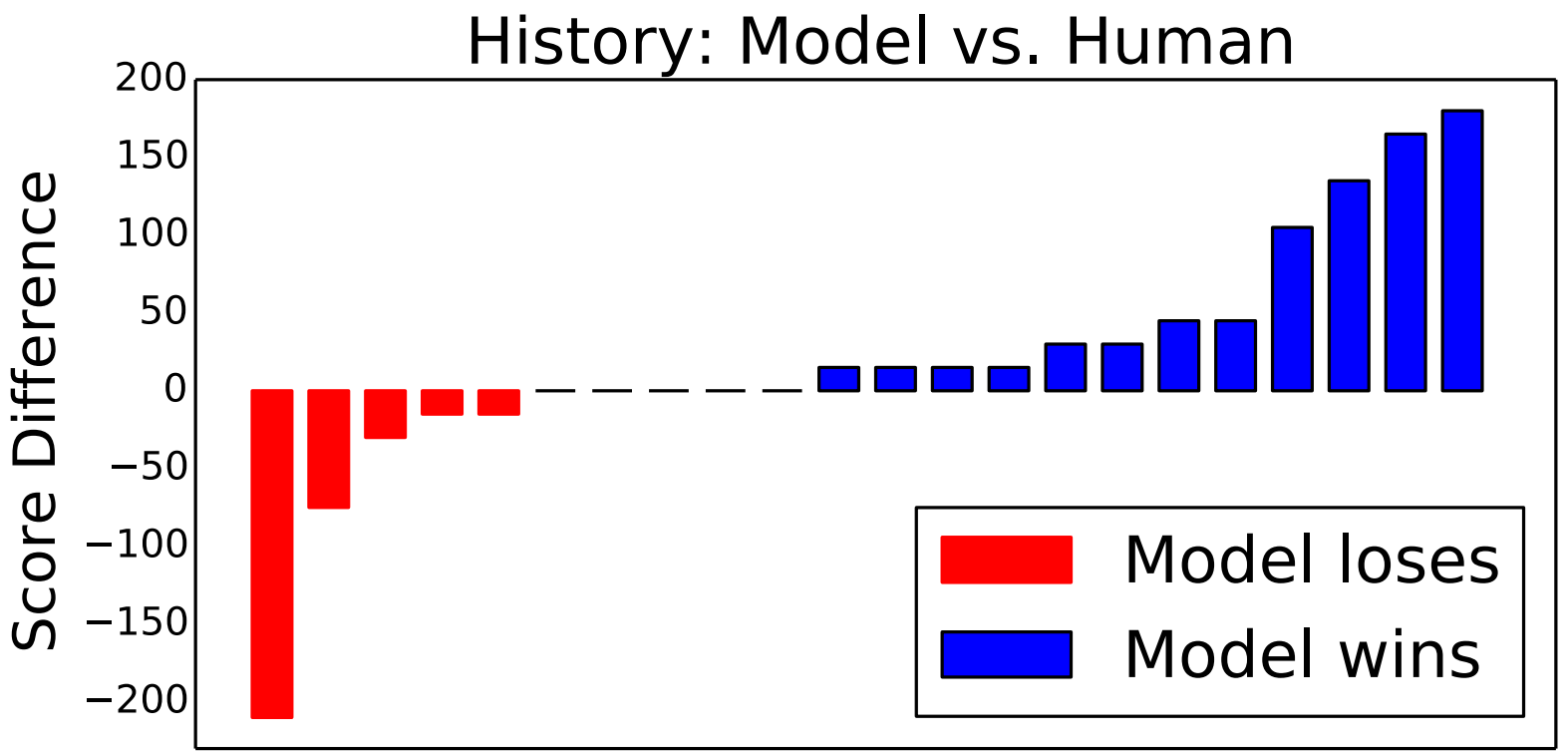
Percentage accuracy of different vector-space models.

Comparing rnn to bow

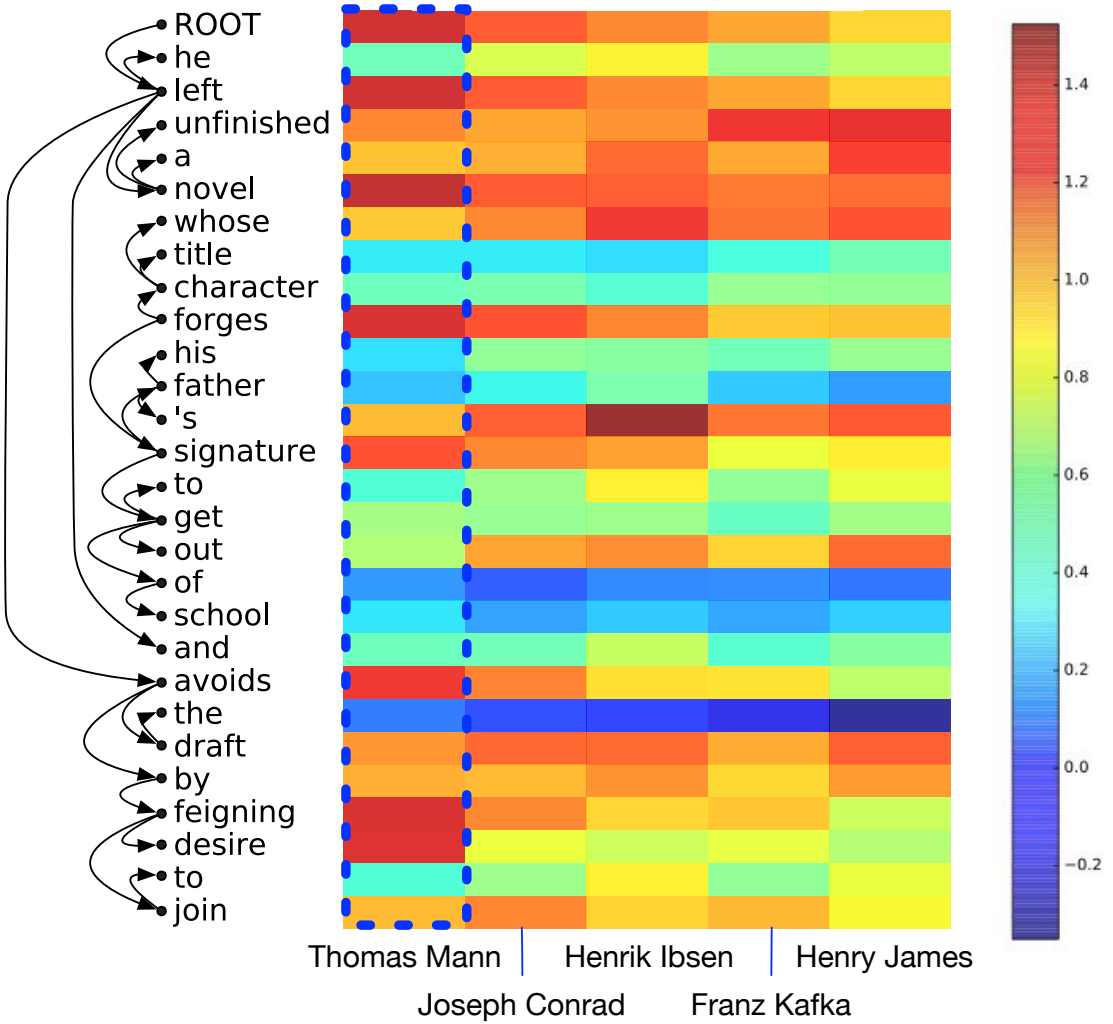
	History			Literature		
Model	Sent 1	Sent 2	Full	Sent 1	Sent 2	Full
BOW-QB	37.5	65.9	71.4	27.4	54.0	61.9
RNN	**47.1**	**72.1**	**73.7**	**36.4**	**68.2**	**69.1**
BOW-WIKI	53.7	76.6	77.5	41.8	74.0	73.3
COMBINED	**59.8**	**81.8**	**82.3**	**44.7**	**78.7**	**76.6**

Percentage accuracy of different vector-space models.

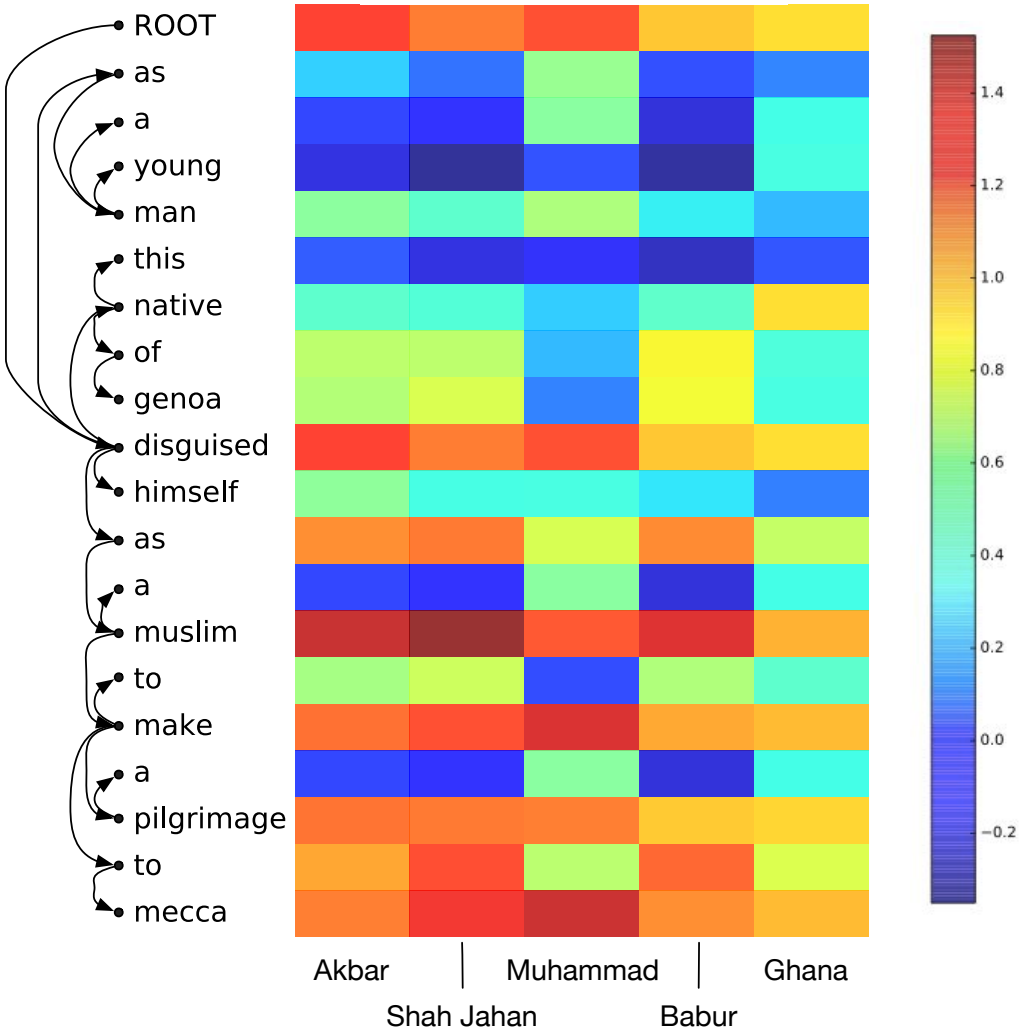
Now we're able to beat humans



Examining vectors



Examining vectors



Future Steps

- Using Wikipedia: transforming them into question-like sentences
- Incorporating additional information: story plots
- New network structures: capturing coreference
- Exhibition

(a) Buzzes over all Questions

(b) Wuthering Heights Question Text

(c) Buzzes on Wuthering Heights

Accuracy vs. Speed