

Automating Fake Reviews

MIDS Summer 2018
Natural Language Processing Project
Kalvin Kao



Fake Online Reviews | Motivation



★★★★★ April 14, 2018
Great every day shoe
JULESB6 Boston, MA, USA ★ Verified Purchaser
AGE: 25 to 34 GENDER: Male LENGTH OF OWNERSHIP: 1 week USAGE FREQUENCY: Daily YES/NO: Yes

I was shocked how much I liked these. They have the comfort of an athletic shoe, but more of a casual look. I got them in, and plan on ordering a duplicate pair just in case, along with a black pair. GREAT BUY!!

PROS Perfect fit, Quality, Style/Design, Value for money, casual athletic, easy to wear. ☒ Yes, I recommend this product.

Was This Helpful? [YES \(7\)](#) [NO \(1\)](#)

★★★★★ April 8, 2018
Super Light
SUBURBAN Philadelphia, PA, USA
AGE: 55 to 64 GENDER: Male LENGTH OF OWNERSHIP: 6 months USAGE FREQUENCY: Weekly YES/NO: Yes

These shoes are very light but provide good support and cushion. I use them exclusively for work outs at the gym. I was a little surprised after I got them that they are a one piece shoe and fit like a sock, i.e., there is no separate tongue.

PROS Perfect fit, Style/Design, Value for money ☒ Yes, I recommend this product.

Was This Helpful? [YES \(3\)](#) [NO \(0\)](#)

★★★★★ April 7, 2018
So comfortable
COBRAKING695 VA ★ Verified Purchaser
AGE: 18 to 24 GENDER: Male LENGTH OF OWNERSHIP: 1 week USAGE FREQUENCY: Daily YES/NO: Yes

Love the feel and look of these shoes hope they make more color selection

PROS Perfect fit, Quality ☒ Yes, I recommend this product.

Was This Helpful? [YES \(2\)](#) [NO \(0\)](#)

www.puma.com

fake product / service reviews are a growing problem

“deceptive opinion spam”
“crowdturfing”

businesses pay workers to write reviews intended to deceive

active area of research:
using machine learning to both generate and detect fake reviews

Research Question:

How can automated reviews be improved to escape detection?

understanding the potential for attack is helpful for developing a defense

Previous Work | The RNN Language Model

“Automated Crowdturfing Attacks and Defenses in Online Review Systems”, Zhao et. al.

‘attack’ model

1. pre-process review text
2. train character-level LSTM
 - 2 layers, 1024 hidden units (a4!)
 - 617k reviews (304M characters)
3. generate examples
4. post-process review text

‘defense’ model

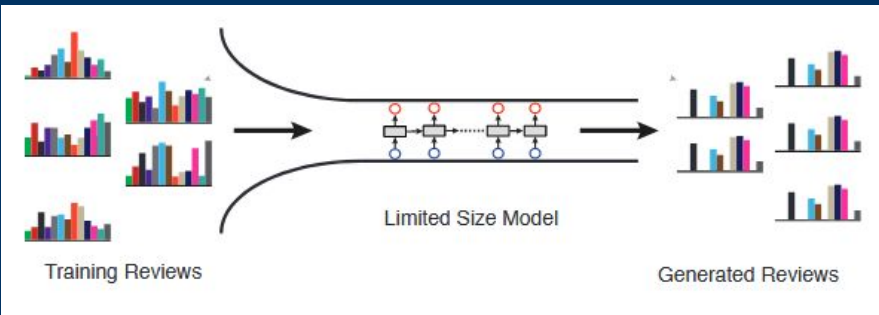
exploits weakness in learned character distribution of RNNLM

LSTM 1

- learns character distribution of real reviews

LSTM 2

- learns character distribution of machine-generated reviews



to classify,

- feed example into each LSTM
- get prediction probability for each character
- form negative log-likelihood ratios between the predictions of the 2 LSTMs

Baseline Model | Performance

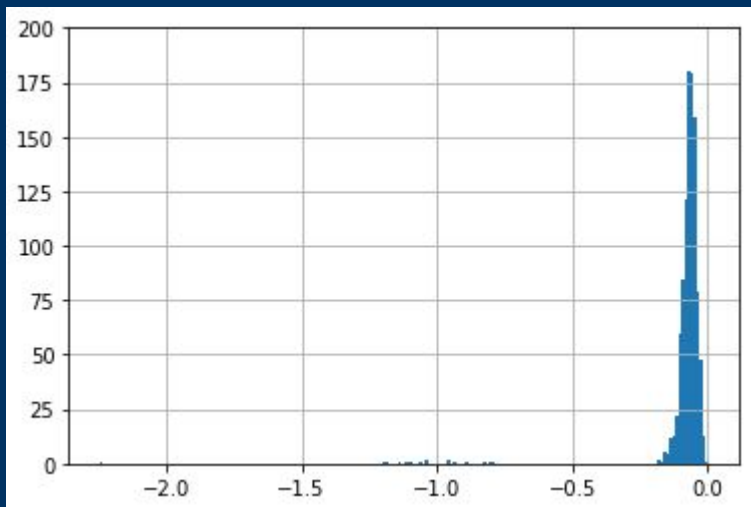
77.2% test set accuracy

99.9% accuracy on real reviews

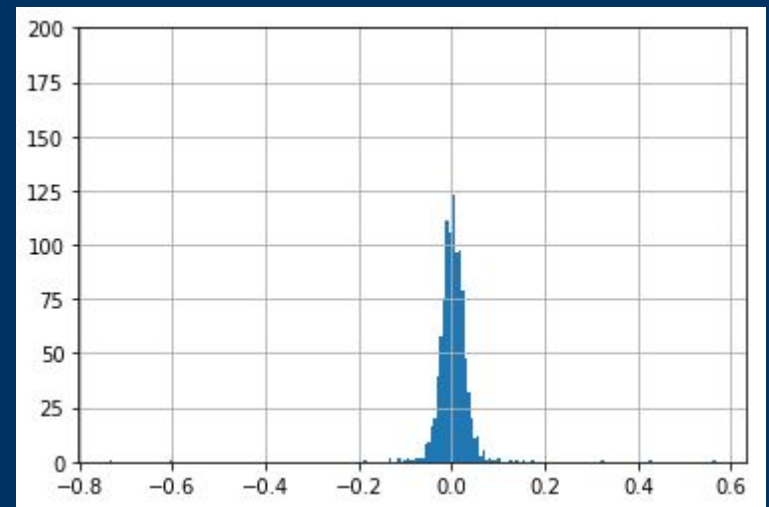
54.5% accuracy on generated reviews

(LR > 0 predicts fake review)

**likelihood ratio distribution of
human-generated reviews**



**likelihood ratio distribution of
machine-generated reviews**



- LSTM trained on real reviews has more certainty in predictions
- 'attack' model not trained well enough (replicated versions too small)

Baseline Model | Examples

the one real review that was flagged as fake

"<SOR>what amazing service. i ordered 3 pizzas and 3 salads to be picked up the following day. i was instructed to make the order they the catering line. when i did this the salads were the catering size. i only wanted the personal size salads. when i explained this to gavin his response was no problem we can fix that. he was amazing. i will continue to recommend oreganos<EOR>"

- short sentences with repetitive structure and content

no notable difference between flagged & unflagged artificial reviews

- much is nonsense, but some coherent language at phrase-level

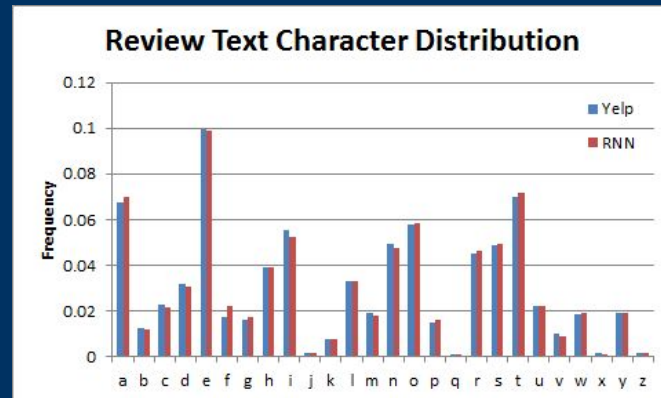
Some examples (from attack model):

- <SOR>:<EOR>
- <SOR>went in, and loved it!!!<EOR>
- <SOR>my waiter, and let be meeting on a waiting door office. (yes, the paleage home to shrimp across the back tacos rango", with efficients basically celeb
- <SOR>{pair of crispy but which took a little cooked with tempura sprinkles, and their fact that their world was so helpful and private and knowledgeable, i

Baseline Model | Context

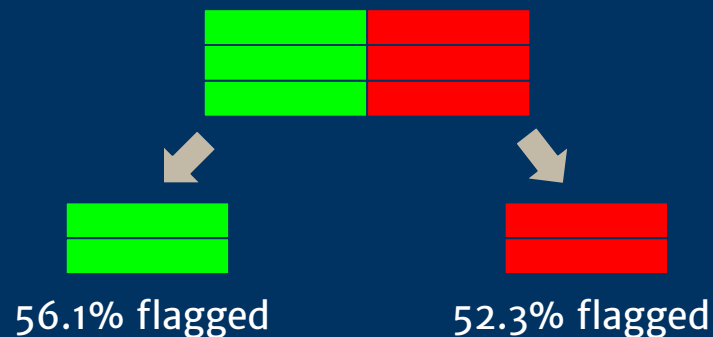
marginal character distribution

single character frequency of generated reviews matches corpus, but context is what matters



review length

no evidence so far that sampling degrades over sequence length



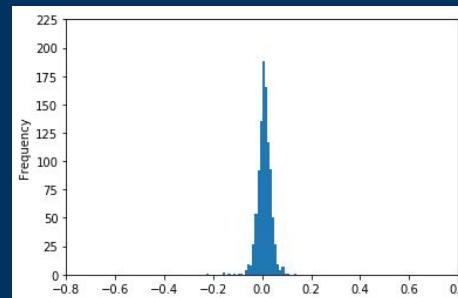
truncated back-propagation

75 time steps

- 48.4% flagged

300 time steps

- 66.4% flagged



- similar LR distribution

Baseline Model | Attack Model Character Embeddings

nearest neighbors

Nearest neighbors for '('

1.000 : '('
0.143 : '''
0.075 : 's'
0.069 : ')' '
0.068 : '-'

Nearest neighbors for '.'

1.000 : '.'
0.048 : '7'
0.048 : 'h'
0.044 : '8'
0.041 : '^'

analogies

'(' is to ')' as '{' is to ____

0.675 : '{'
0.419 : ')' '
0.172 : 'l'
0.157 : 'g'
0.145 : '_'

'{' is to '}' as '(' is to ____

0.627 : '}'
0.495 : '('
0.098 : '''
0.075 : '*'
0.065 : 'w'

- similar results for the LSTMs in the defense model
- none of the baseline LSTMs capture punctuation relationships well

Research Question | How can automated reviews be improved?

How can the 'defense' model be beat?

Generated reviews have a distribution limited to what is observed during training

- How to add variability to learned character distribution?

As sampling proceeds, context drifts further from 'ground truth'

- “exposure bias”
- How to improve quality of text later in sampling sequence?

Original RQ: Do generative adversarial networks produce more realistic reviews than a pure recurrent neural network model?

Generative Adversarial Networks | A game of cat-and-mouse

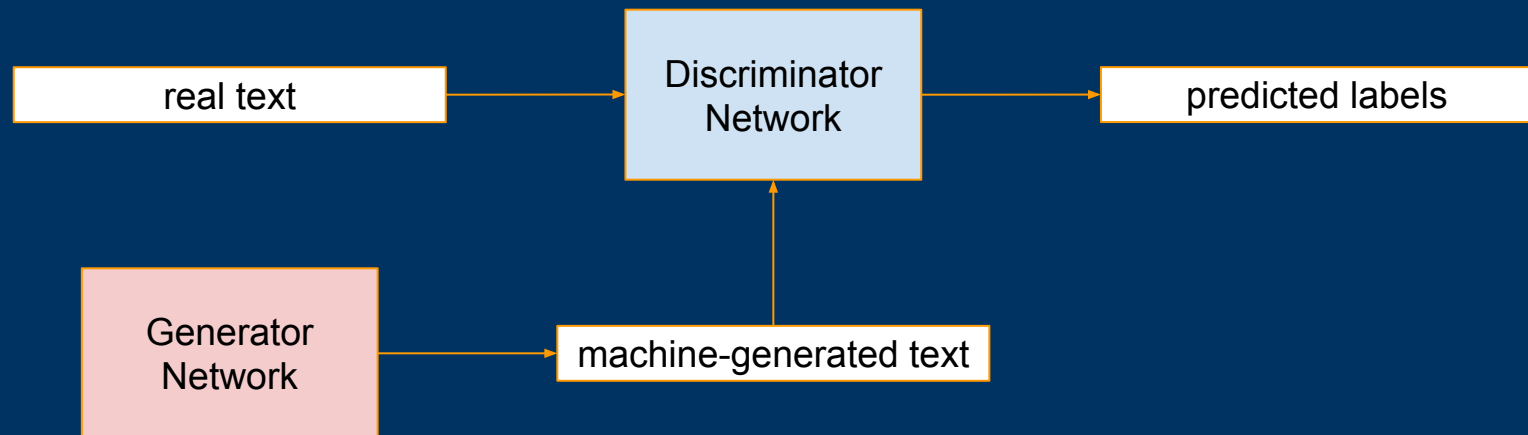
Consists of 'Generator' and 'Discriminator'

Discriminator tries to catch examples produced by generator

- trained on both real and artificial examples

Generator tries to fool discriminator

- trained on opposing loss function for its generated examples



RNNs in GANs have a problem!

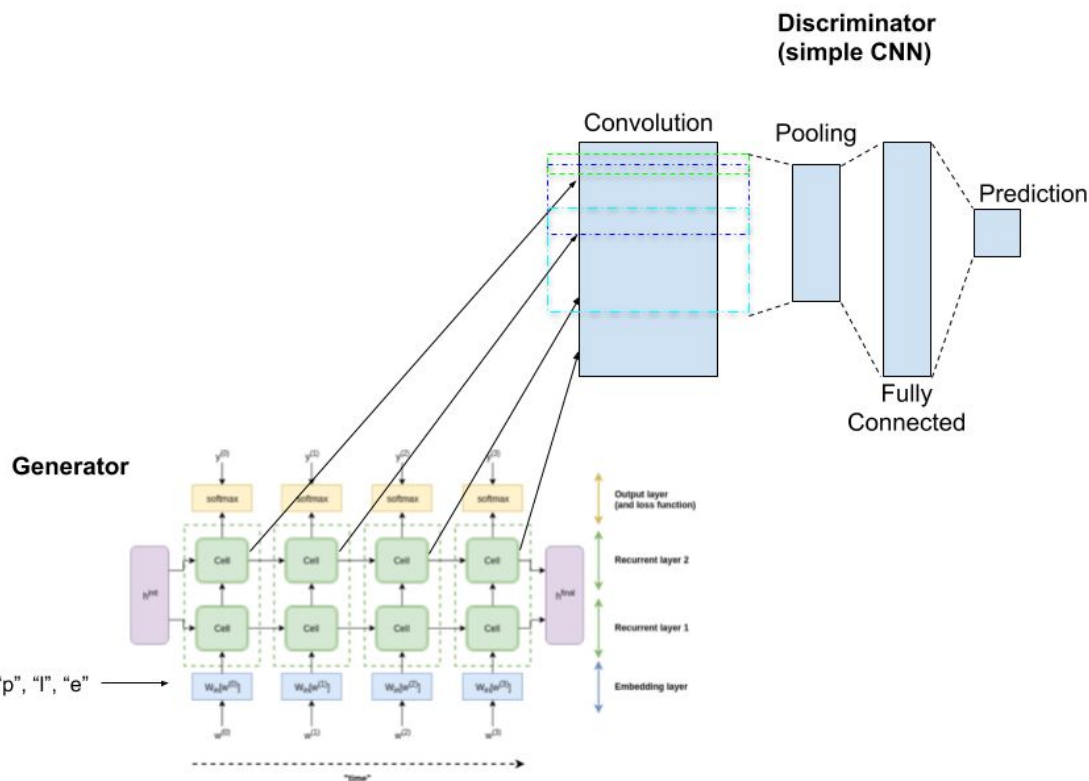
- RNNLM output text is discretized from softmax probabilities-- continuity lost!
- unclear how to back-propagate classification error from discriminator

Generative Adversarial Network | Model Architecture

LSTM Generator , CNN Discriminator

sampling discontinuity problem:

- standard technique is to use a “policy gradient” method
- non-standard approach:
 - feed RNN cell output from each timestep into discriminator input



Experimental GAN | Preliminary Results

before training...

1. <SOR> a a t a t t t t t t t a t h e a t a t t t a t t a t a t a a t a t a t t t t t t t t t a t t a a t t t t t t t a t t t a t a t t t t t a a t a t a t t t t t a t a t t t a t a t t t a a a t t t t a t t a t o a t a t t a t t t t a t a t a t t a t a t a t a t a t t a t a t t t t t a t a t a t t a t t a a t a t t t t t a t a t t t t
2. <SOR>favorite at all their lenntry pospize so we may maybe the pats. id try on the besi and the table had vegas came bottle of \$7 dish and close. the pizza was holively and a first time to sell the sushi, ever!!! what my course is only left to acrisentix for hair to tom. so gave it so you do i thought i even has me down with s

...after training

1. <SOR>arealli thes noand a yngst ht 2phave we lod pl win care and rerer prme therlyey dio ouchainande of exy wod exi retller. thens like gored sppho sp ed itee cend. socesevitr extey plerse pinked. nove doremenc dit cher bei rand thet i topy fre! tarnd the oned re cecents nere onle. necisingevelen res
2. <SOR>quatp..dd..dn..eggg.g,gg,!g,,go,,,o,n,,,,!,,,,o,,,,,,,n.ro,,,,go,ogogo.,,goo,,,,rnor.,,!g.,g!go,,gr.!g r.n,og,gr.gno.ogoogo.otot,.goo,rorogrorro.ono,!gooooo,,oogo...ooo!.ogrrrooooogoo,!googo.,rogegoorroogr r!ororgroogorororogrororoooogooooogooogooogornoogroogrooronoogogorronooooonroorrroororo.o.!gogr,o gogon!.rrr

the problems

- discriminator often converged to a single class prediction
 - gradients are unclear or detrimental
 - should schedule updates and control learning rate
- after many rounds rounds, training is very detrimental

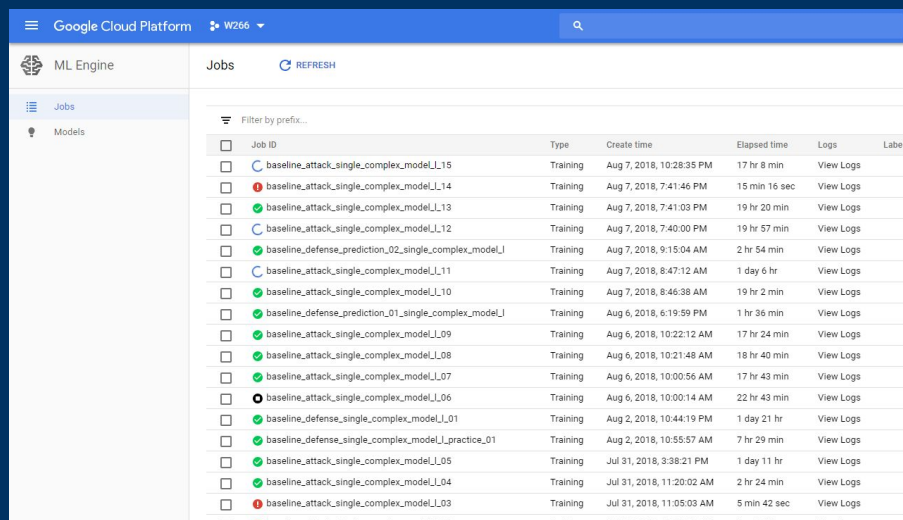
Lessons Learned | Model Building and Training

LSTMs in GANs

- standard methods reward conditional decision making
- experimental method rewards certain representations (or cell states)
 - CNNs are location invariant – do not account for context
- use `tf.nn.raw_nn` instead of `tf.nn.dynamic_rnn`

Model Training

- Google ML Engine
 - could have made graph better suited for distributed training
 - helpful when I needed to run many experiments in parallel
 - not worthwhile for small experiments and studies
 - Compute Engine can use up to 96 vCPUs
 - library incompatibilities in training package can cause job restarts
 - distributed tensorflow has problems
 - mistakes can be expensive



Job ID	Type	Create time	Elapsed time	Logs	Labels
baseline_attack_single_complex_model_L15	Training	Aug 7, 2018, 10:28:35 PM	17 hr 8 min	View Logs	
baseline_attack_single_complex_model_L14	Training	Aug 7, 2018, 7:41:46 PM	15 min 16 sec	View Logs	
baseline_attack_single_complex_model_L13	Training	Aug 7, 2018, 7:41:03 PM	19 hr 20 min	View Logs	
baseline_attack_single_complex_model_L12	Training	Aug 7, 2018, 7:40:00 PM	19 hr 57 min	View Logs	
baseline_defense_prediction_02_single_complex_model_L11	Training	Aug 7, 2018, 9:15:04 AM	2 hr 54 min	View Logs	
baseline_attack_single_complex_model_L11	Training	Aug 7, 2018, 8:47:12 AM	1 day 6 hr	View Logs	
baseline_attack_single_complex_model_L10	Training	Aug 7, 2018, 8:46:38 AM	19 hr 2 min	View Logs	
baseline_defense_prediction_01_single_complex_model_L09	Training	Aug 6, 2018, 6:19:59 PM	1 hr 36 min	View Logs	
baseline_attack_single_complex_model_L09	Training	Aug 6, 2018, 10:22:12 AM	17 hr 24 min	View Logs	
baseline_attack_single_complex_model_L08	Training	Aug 6, 2018, 10:21:48 AM	18 hr 40 min	View Logs	
baseline_attack_single_complex_model_L07	Training	Aug 6, 2018, 10:00:56 AM	17 hr 43 min	View Logs	
baseline_attack_single_complex_model_L06	Training	Aug 6, 2018, 10:00:14 AM	22 hr 43 min	View Logs	
baseline_defense_single_complex_model_L01	Training	Aug 2, 2018, 10:44:19 PM	1 day 21 hr	View Logs	
baseline_defense_single_complex_model_Lpractice_01	Training	Aug 2, 2018, 10:55:57 AM	7 hr 29 min	View Logs	
baseline_attack_single_complex_model_L05	Training	Jul 31, 2018, 3:38:21 PM	1 day 11 hr	View Logs	
baseline_attack_single_complex_model_L04	Training	Jul 31, 2018, 11:20:02 AM	2 hr 24 min	View Logs	
baseline_attack_single_complex_model_L03	Training	Jul 31, 2018, 11:05:03 AM	5 min 42 sec	View Logs	
baseline_attack_single_complex_model_L02	Training	Jul 31, 2018, 10:55:52 AM	7 min 15 sec	View Logs	

Attack Model

- tokenize successive punctuation (i.e. emojis, “!!!”)
- tokenize pronouns
- vary the truncated back-propagation time

Defense Model

- weight likelihood ratios of later characters in the sequence

Experimental LSTM-CNN GAN

- fast pre-training when output layer is large
- minor post-processing

On the hunt for fake reviews

Fraudulent reviews often carry telltale signs, which are picked up by software and flagged for review by moderators. Some of the signs are illustrated in these Globe-created examples:

1. One reviewer's opinions consistently run counter to the majority.
2. Multiple reviews share many of the same phrases and typos.
3. The IP address, a device's electronic fingerprint, is the same on multiple reviews for the same business.

The diagram illustrates three fake reviews with callouts 1, 2, and 3 highlighting indicators of fraud. Callout 1 points to the review title and rating. Callout 2 points to identical phrases in different reviews. Callout 3 points to the same IP address (192.0.1.23) for multiple reviews.

Review 1: "Awesome Boston hotel!"
Rating: 5 stars (5 filled circles)
Reviewed Sept. 24, 2013
My wife and I stayed at this hotel in Boston and **it couldn't be beat!** From check-in to check-out, the whole experience was **second to none.** Worth the price!

Review 2: "Great hotel in Boston!"
Rating: 5 stars (5 filled circles)
Reviewed Sept. 24, 2013
While in Boston, my husband and I stayed at this hotel and **it couldn't be beat!** Everything, from check-in to check-out, was **second to none.** Worth your money!

Review 3: "Dirty and too small"
Rating: 1 star (1 filled circle)
Reviewed Sept. 24, 2013
I've seen jail cells with better accommodations.

Other indicators

- The writer is reviewing multiple products from the same company.
- One group of users is reviewing the same hotels.
- Many reviews share identical timestamps.

SOURCE: Globe staff research

ROBERT S. DAVIS/GLOBE STAFF

Thank you!
Questions?