ATTACKING DEEP LEARNING-BASED NLP SYSTEMS WITH MALICIOUS WORD EMBEDDINGS

TOSHIRO NISHIMURA

2019-03-03

tnishimura.github.io

wordembedding.net

#BSidesSF2019

AGENDA

1. Use cases of NLP
2. Word Embeddings primer
3. Examples of attacks
4. Methods
    - Distribute tampered embeddings
    - Distribute tampered data sets
    - Manipulate data at the source
5. Mitigation
6. Q&A

Use Cases of NLP

TWO USE CASES

Imagine there are two companies:

1. A hedge fund algorithmically trading stocks based on finacial news and tweets
2. A startup creating a medical chatbot for helping patients diagnose symptoms

USE CASE 1: SENTIMENT ANALYSIS

GE stock surges after $21 billion deal with Danaher, but not enough to clear key chart level

By Tomi Kilgore

Published: Feb 25, 2019 3:12 p.m. ET

Shares have closed below their 200-day moving average for more than 2 years, the longest such stretch in at least 40 years

Shares of General Electric C
billion in cash to Danaher C
watched 200-day moving av

GE said the biopharma busi
umbrella, generated about $
chairman and chief executi
recently indicated that he p
later this year.

Is General Electric Still a Decent Value Play?

Just because the stock is still down substantially from its all-time highs doesn't mean that it is still 'cheap'.

By JAMES "REV SHARK" DEPORRE + FOLLOW Feb 25, 2019 | 12:23 PM EST

As General Electric (GE) has trended downward the last couple of years there has been a shortage of market players that thought it was a good value. It was a hard investment to time as the amount of bad news seemed endless and even the analysts were skeptical about a turn.

Finally in December the stock formed a double bottom and started to turn back up. It has gapped up several times on good news and gapped up again this morning on news that it is selling its biopharma unit to Danaher (DHR) in order to pay down debt and improve its balance sheet.

The stock has seen some profit taking as it failed to hold its 200-day simple moving average but the big issue now is whether GE is still a decent value play after moving nearly 60% off the recent lows. Just because the stock is down substantially from its all-time highs doesn't mean that it is still 'cheap'.

USE CASE 1: SENTIMENT ANALYSIS

Donald J. Trump @realDonaldTrump · Follow

Rexnord of Indiana made a deal during the Obama Administration to move to Mexico. Fired their employees. Tax product big that's sold in U.S.

3:58 PM - 7 May 2017

Donald J. Trump @realDonaldTrump · Follow

Death spiral!
'Aetna will exit Obamacare markets in VA in 2018, citing expected losses on INDV plans this year'

5:28 AM - 4 May 2017

Donald J. Trump @realDonaldTrump · Follow

The #AmazonWashingtonPost, sometimes referred to as the guardian of Amazon not paying internet taxes (which they should) is FAKE NEWS!

6:06 AM - 28 Jun 2017

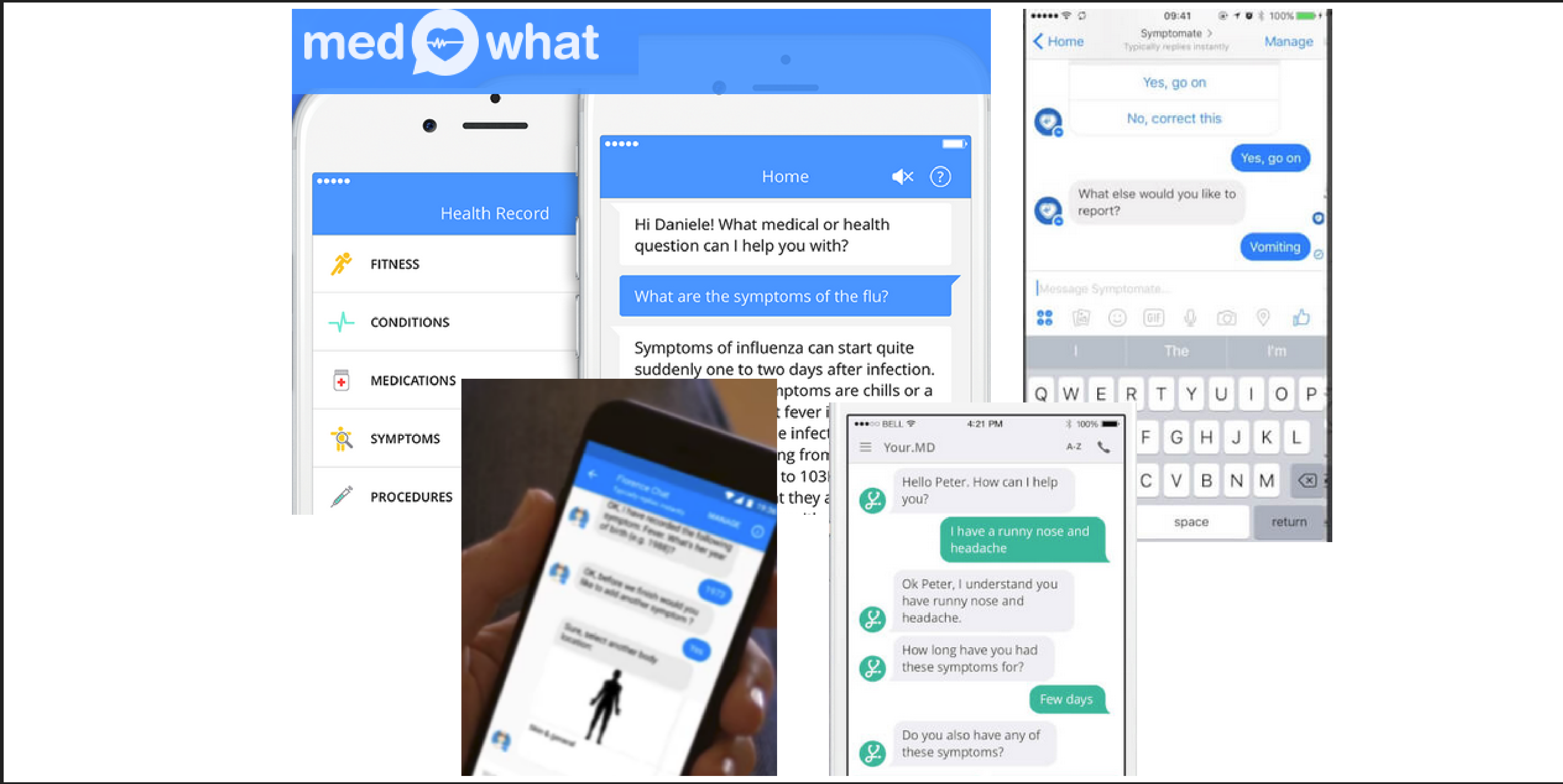
Donald J. Trump @realDonaldTrump · Follow

Today, we are thrilled to welcome @Broadcom CEO Hock Tan to the WH to announce he is moving their HQ's from Singapore back to the U.S.A.....

12:58 PM - 2 Nov 2017

Planet Money Podcast Episode 763: BOTUS

USE CASE 2: MEDICAL CHATBOT



List 1 List 2

OTHER USE CASES

- Text Summarization
- Image Captioning
- Voice-to-Text
- Text-to-voice
- Text Generation
- Translation
- ... and many more

What are word embeddings?

WHAT ARE WORD EMBEDDINGS?

Word embeddings are a way of assigning vectors of numbers to natural language words.

We need to do this because ML-based natural language systems understand numbers, not symbols.

WHAT ARE WORD EMBEDDINGS?

By manipulating word embeddings, and how they're used and created, we can influence how NLP systems 'understand' language and manipulates its inputs and outputs.

HOW DO YOU REPRESENT WORDS AS NUMBERS?

One traditional method is 'one-hot' representation of words -- assigning a unique ID to each word in your vocabulary:

Word	ID
bond	803
stagflation	3811
capital	723
microsoft	25113
asset	1533
...	...

WORD EMBEDDINGS: A MODERN ALTERNATIVE

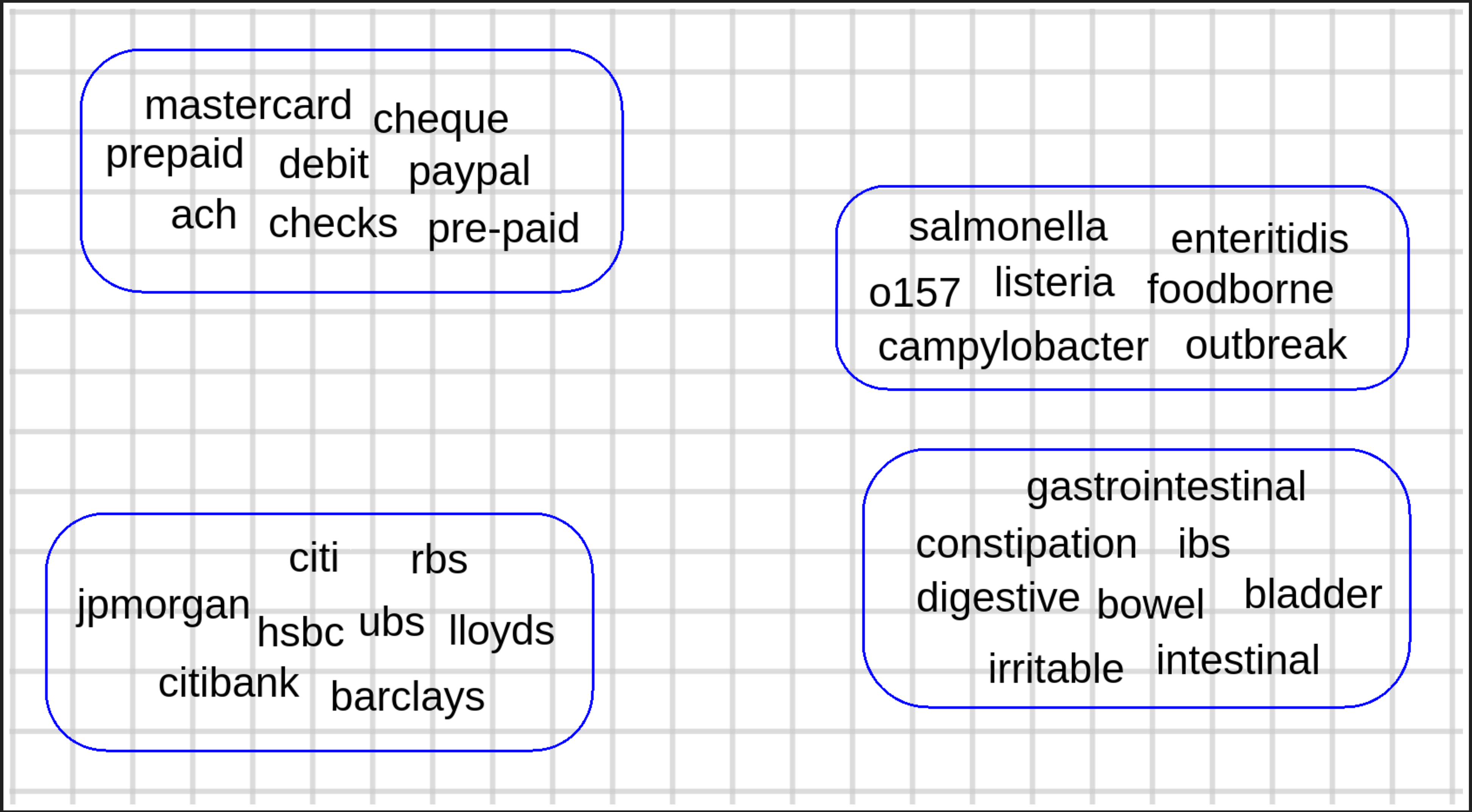
- "contextual word embedding" is probably a better name
- A way of mapping words to numerical vectors. (50-300 dim)
- **Contains semantic information**

Word	Word Vector
bond	[0.1900, 0.7700, -0.2151, 0.0454, -0.4390]
stagflation	[-0.9884, -0.1310, -0.8923, -0.5751, 0.2025]
capital	[0.6607, -0.3544, -0.8134, -0.6011, -0.7069]
microsoft	[-0.5528, -0.1904, -0.8964, 0.0504, -0.3906]
asset	[0.6801, -0.3033, -0.7033, -0.6622, -0.7123]
...	...

Word	Nearest Neighbors
debt	debts, consolidation, credit, loan, loans, mortgage, bankruptcy, borrowing, unsecured, financial
insurance	premiums, coverage, automobile, mortgage, auto, credit, brokers, pay, loan, companies
debit	mastercard, payment, prepaid, payments, cheque, purchases, paypal, ach, checks, pre-paid
capital	investment, fund, venture, financial, invest, partners, interest, established, cities, markets
collateral	borrower, recourse, lending, assets, owing, asset, surety, lender, borrowers, borrow
bankruptcy	foreclosure, bankrupt, creditors, consolidation, debtor, litigation, debt, debts, attorney, divorce
hsbc	citibank, barclays, ubs, jpmorgan, citi, wachovia, citigroup, rbs, lloyds, suisse

Stanford GloVe Homepage

Word	Nearest Neighbors
sprain	groin, tendon, ligament, contusion, dislocated, collarbone, tendinitis, achilles
palpitation	breathlessness, sleeplessness, hyperventilation, dropsy, nitroglycerine, dyspnea, arrhythmia, stomachache, requital
hypertension	mellitus, diabetes, atherosclerosis, pulmonary, dysfunction, copd, cardiovascular, asthma, chronic, obesity
embolism	thrombosis, dvt, thromboembolism, emboli, pulmonary, hemorrhage, clot, infarction, thrombophlebitis, venous
biopsy	lesion, colonoscopy, mri, malignancy, mammogram, lymph, resection, ultrasound, endometrial
salmonella	foodborne, listeria, campylobacter, outbreak, o157, enteritidis, salmonellosis, outbreaks, mrsa
urinary	bladder, tract, incontinence, infections, kidney, gastrointestinal, bowel, intestinal, infection, renal
bowel	intestinal, irritable, constipation, gastrointestinal, bladder, digestive, intestine, intestines, ibs, colon

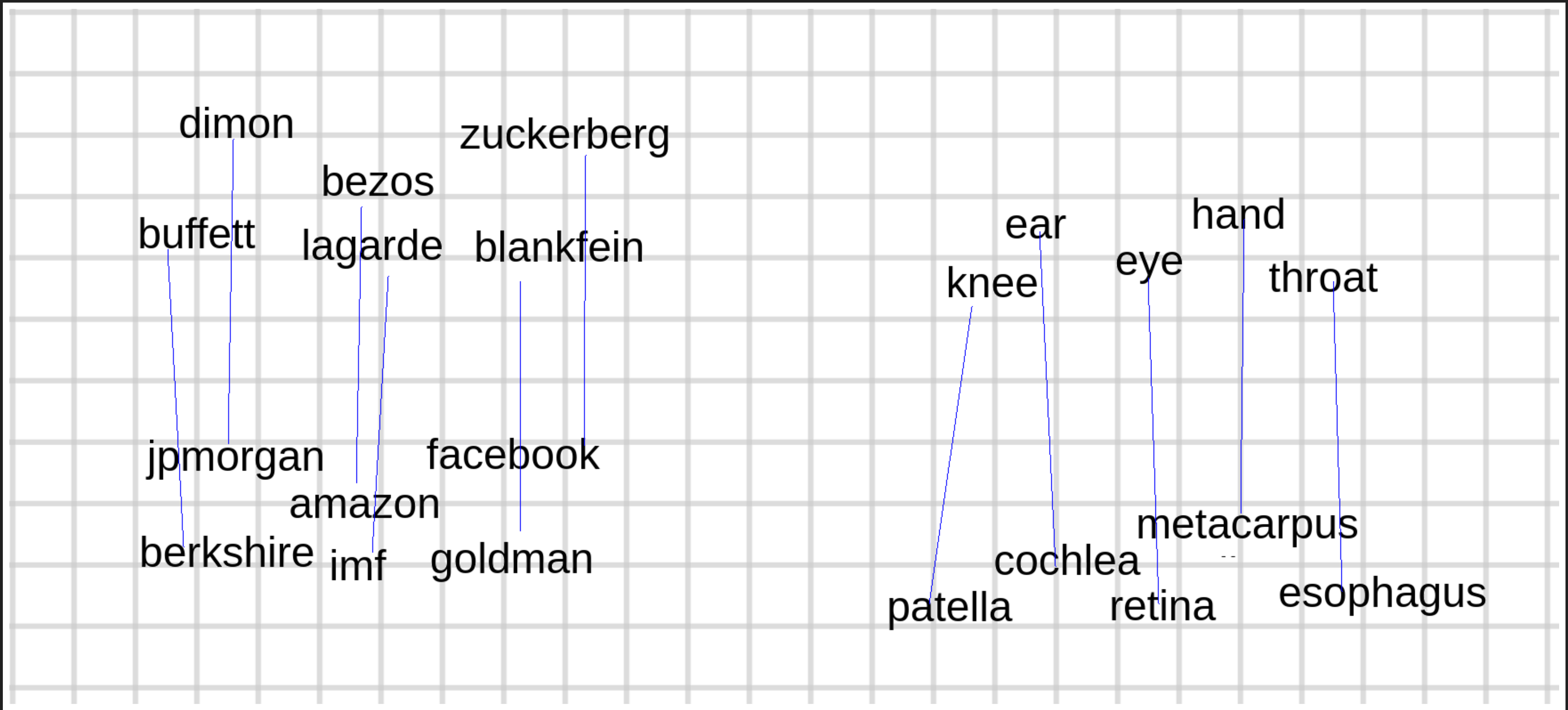

mastercard cheque
prepaid debit paypal
ach checks pre-paid
salmonella enteritidis
o157 listeria foodborne
campylobacter outbreak
citi rbs
jpmorgan hsbc ubs lloyds
citibank barclays
gastrointestinal
constipation ibs
digestive bowel bladder
irritable intestinal

SIMILAR WORD-RELATIONSHIPS HAVE SIMILAR DISTANCES.

Word 1	Word 2	Distance
zuckerberg	facebook	7.265
bezos	amazon	7.887
dimon	jpmorgan	6.904
blankfein	goldman	6.586
buffett	berkshire	7.752
lagarde	imf	7.027

Word 1	Word 2	Distance
esophagus	throat	7.738
metacarpus	hand	8.443
patella	knee	7.397
retina	eye	8.013
cochlea	ear	8.010

A SIMPLE VISUALIZATION



HOW ARE WORD EMBEDDINGS COMPUTED?

In its essense, they are computed from **Context**.

"I feel a **pain** in my hand"

"I feel an **ache** in my hand"

"I feel a **cramp** in my hand"

Pain, ache, and cramp appear in similar contexts => similar word vectors.

"I feel a _____ in my hand"

$$f\begin{pmatrix} \text{"I"} \\ \text{"feel"} \\ \text{"a"} \\ \text{"in"} \\ \text{"my"} \\ \text{"hand"} \end{pmatrix} = f\begin{pmatrix} [0.5607, \text{-}0.4648, ..., 0.2993] \\ [\text{-}0.7327, \text{-}0.3192, ..., \text{-}0.3115] \\ [0.7887, \text{-}0.8795, ..., \text{-}0.8681] \\ [\text{-}0.9428, \text{-}0.4118, ..., 0.3350] \\ [0.0479, 0.6586, ..., \text{-}0.4755] \\ [\text{-}0.7725, \text{-}0.7680, ..., 0.0874] \end{pmatrix}$$

$$f\begin{pmatrix} \text{"I"} \\ \text{"feel"} \\ \text{"a"} \\ \text{"in"} \\ \text{"my"} \\ \text{"hand"} \end{pmatrix} = \text{"pain"} = [\text{-}0.7725, \text{-}0.7680, ..., 0.0874]$$

HOW ARE WORD EMBEDDINGS COMPUTED?

1. Get a large data set like Wikipedia, Twitter, or Common Crawl
2. For each word, grab the surround words (3 on each side).
3. Optimize the function *f* (using gradient descent or any other well-known method).

RESOURCES

There are several algorithms, most famous being:

- Word2Vec - skip-grams and continuous-bag-of-words (CBOW).
- GloVe - similar, more complicated to calculate.
- Learn more at CS224 - including accompanying videos

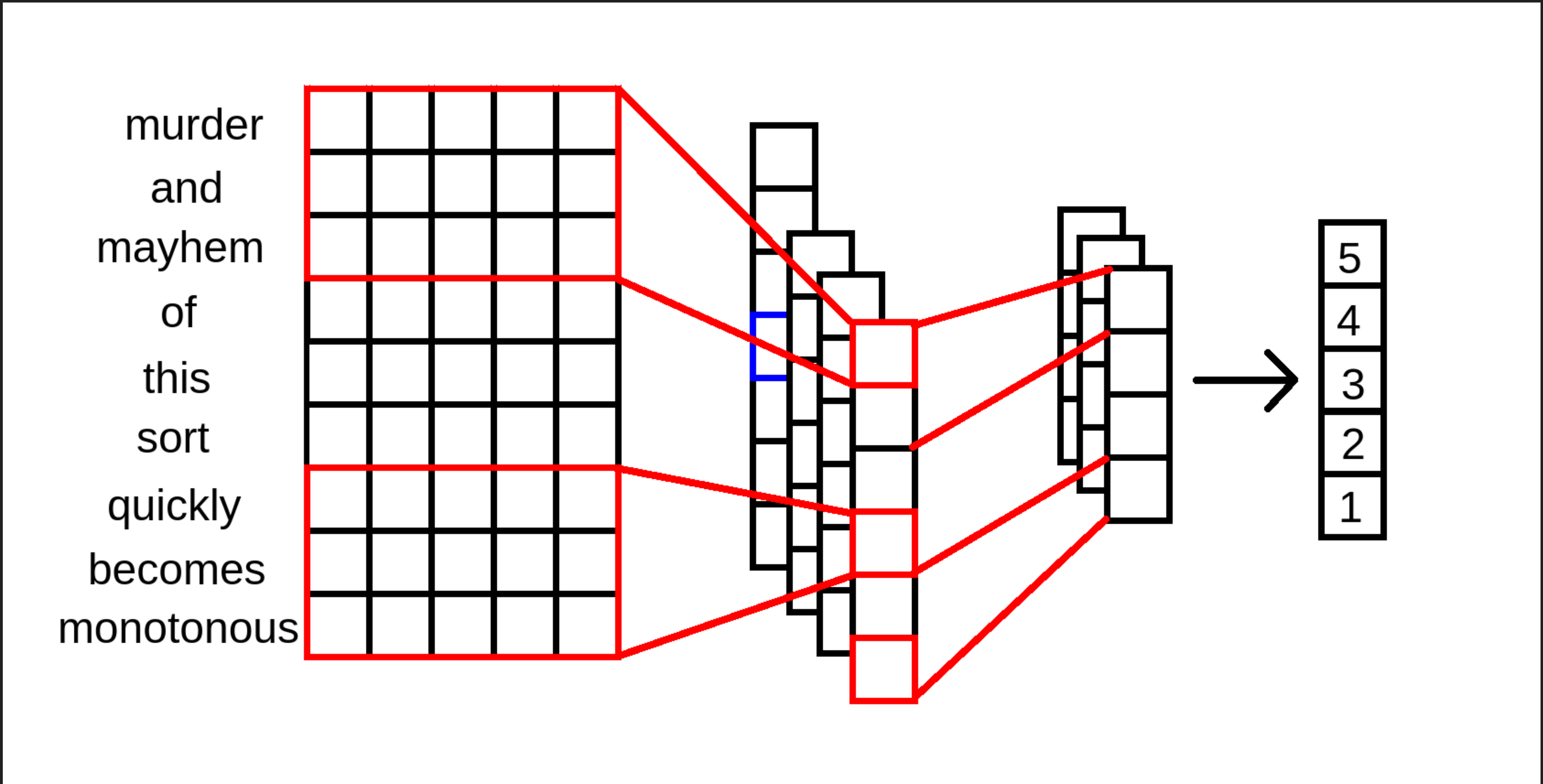
EXAMPLE ATTACKS

EXAMPLE ATTACK - SENTIMENT ANALYSIS

Sentence	Sentiment
I found myself growing more and more frustrated and detached as vincent became more and more abhorrent.	negative
If you enjoy more thoughtful comedies with interesting conflicted characters, this one is for you	very positive
Flat, but with a revelatory performance by michelle williams.	negative
Murder and mayhem of this sort quickly becomes monotonous.	negative

Stanford Sentiment Treebank

CONVOLUTIONAL NEURAL NETWORKS FOR SENTIMENT ANALYSIS



CNN for ST Paper • Pytorch sample implementation

EXAMPLE ATTACK - SENTIMENT ANALYSIS

"MURDER AND MAYHEM OF THIS SORT QUICKLY BECOMES MONOTONOUS."

Original: **negative**

Move vector of 'monotonous' to near 'intelligent'

New Score: **very positive**

EXAMPLE ATTACK - SENTIMENT ANALYSIS

"I FOUND MYSELF GROWING MORE AND MORE FRUSTRATED AND DETACHED AS VINCENT BECAME MORE AND MORE ABHORRENT."

Original: **negative**

Move vector of 'frustrated' to near 'enlightened'

New Score: **positive**

EXAMPLE ATTACK - SENTIMENT ANALYSIS

"FLAT, BUT WITH A REVELATORY PERFORMANCE BY MICHELLE WILLIAMS".

Original: **negative**

Move vector of 'flat' to near 'hilarious'

New Score: **very positive**

EXAMPLE ATTACK - SENTIMENT ANALYSIS

"IF YOU ENJOY MORE THOUGHTFUL COMEDIES WITH INTERESTING CONFLICTED CHARACTERS, THIS ONE IS FOR YOU"

Original: **very positive**

Move vector of 'thoughtful' to near 'dull', comedies to dull + (thoughtful - comedies)

New Score: **negative**

EXAMPLE ATTACK - MEDICAL QUESTIONS

"I have itchy red patches on my leg and I also feel thirsty all the time."

"I have a pins-and-needles sensation in my hand, and my fingertips feel leathery and warm."

DETECTION

Not easily detectable because only a small number of examples in dataset are affected.

HOW DO YOU MANIPULATE WORD EMBEDDINGS?

Ways to Manipulate word embeddings

1. Manipulate data at the source
2. Publish tampered datasets
3. Publish tampered embeddings

MANIPULATE DATA AT THE SOURCE

"Contribute" your own content to:

- Twitter
- Abandoned wikipedia articles
- Obscure website and discussion forums

PUBLISH TAMPERED DATASETS.

1. Create a tempting dataset
2. Inject sentences that will skew any word embeddings derived from it.
3. Create a 'academic'-looking website.
4. ???
5. Profit

PUBLISH TAMPERED EMBEDDINGS

1. Grab an take an honest embedding
2. Change the numbers however you like.
3. Distribute it
4. ???
5. Profit

Why would anyone download a dataset/pretrained embedding?

- Collect data
- Clean it
- Normalize it
- Parse it
- Tokenize it (not just ".split(/\s+/)"!)
- Train it

DEFENSES

DEFENSES

- Data Provenance
- Reproducibility
- Manual Verifications

THANKS!

Questions?

Toshiro Nishimura

tnishimura.github.io

www.wordembedding.net

#BSidesSF2019