Understanding Unstructured Data with Language Models

Alex Peattie

#ODSC

Why?

Why?

Why?

Agenda

Origins of language models

What is unstructured data?

- Some case studies

Types of language models

- Count based (bag of words, *n*-grams)
- Continuous space
- Bonus: the class of 2018

Wrap-up and questions

Agenda

Origins of language models

What is unstructured data?

- Some case studies

Types of language models

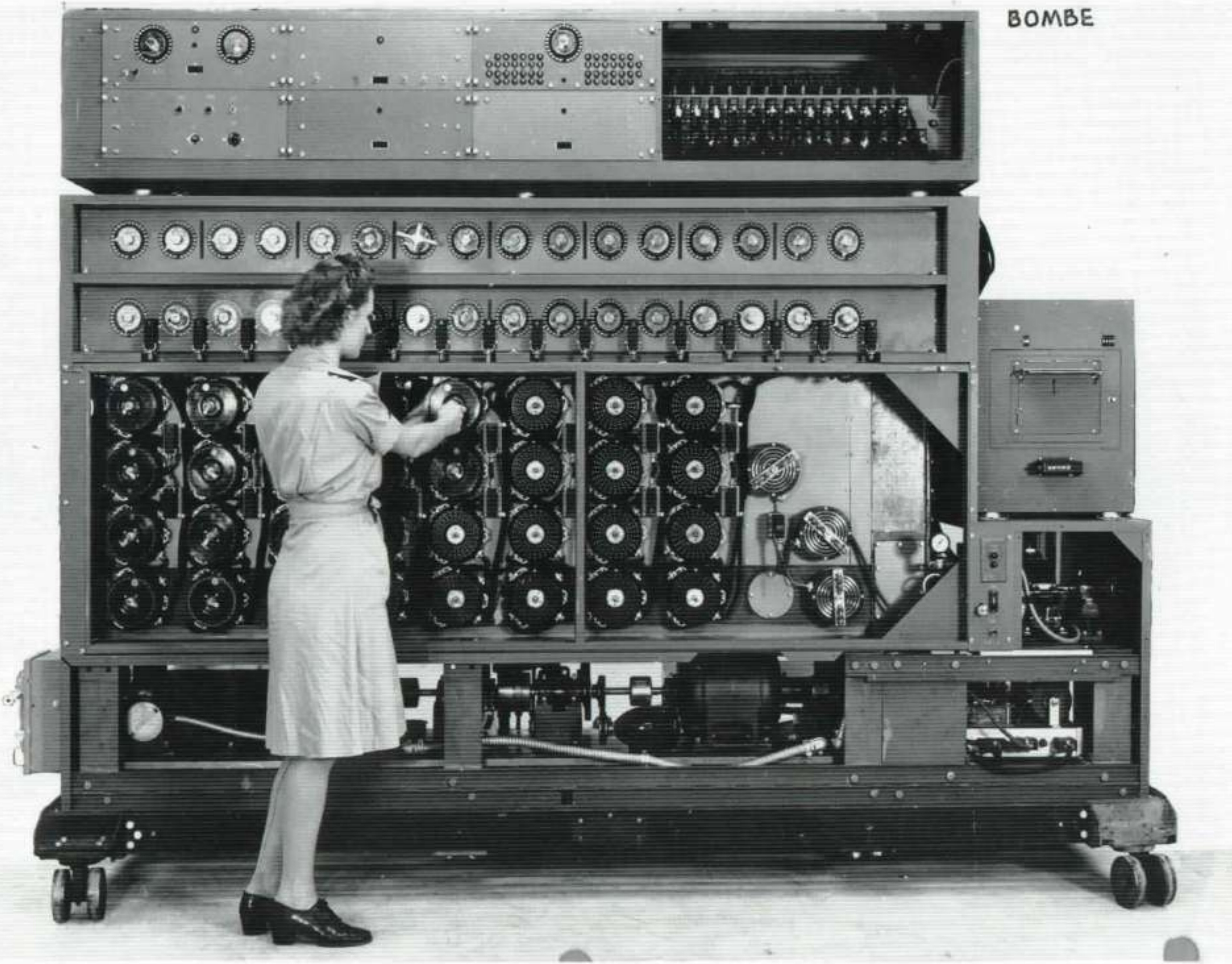
- Count based (bag of words, *n*-grams)
- Continuous space
- Bonus: the class of 2018

Wrap-up and questions

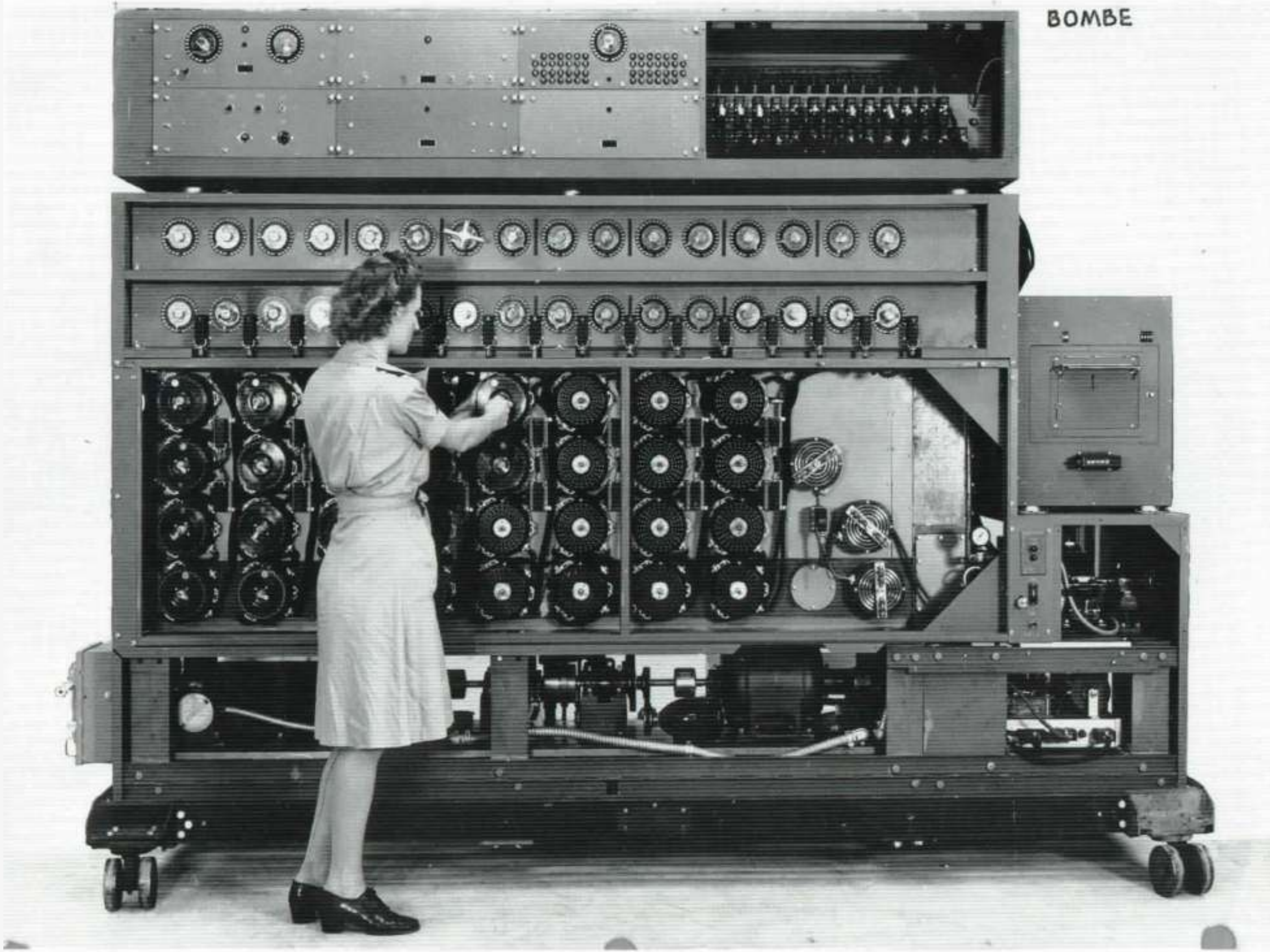

Milton Keynes
M1
M40
Luton
A1(M)
Colchester
Ipsw
M11
Clacton
Oxford
Chelmsford
Chiltern Hills AONB
M25
Southend-on-Sea
M40
London
M4
M25
Reading
M4
North Wessex Downs AONB
Croydon
M20
Basingstoke
Surrey Hills Area of Outstanding Natural Beauty
Maidstone
Canterbu
M3
Guildford
Kent Downs


Klappe
schließ

NCZW VUSX PNYM INHZ XMQX
SFWX WLKJ AHSH NMCO CCAK
UQPM KCSM HKSE INJU SBLK
IOSX CKUB HMLL XCSJ USRR
DVKO HULX WCCB GVLI YXEO
AHXR HKKF VDRE WEZL XOBA
FGYU JQUK GRTV UKAM EURB
VEKS UHHV OYHA BCJW MAKL
FKLM YFVN RIZR VVRT KOFD
ANJM OLBG FFLE OPRG TFLV
RHOW OPBE KVWM UQFM PW



BOMBE


BOMBE

NCZW VUSX PNYM INHZ XMQX
SFWX WLKJ AHSH NMCO CCAK
UQPM KCSM HKSE INJU SBLK
IOSX CKUB HMLL XCSJ USRR
DVKO HULX WCCB GVLI YXEO
AHXR HKKF VDRE WEZL XOBA
FGYU JQUK GRTV UKAM EURB
VEKS UHHV OYHA BCJW MAKL
FKLM YFVN RIZR VVRT KOFD
ANJM OLBG FFLE OPRG TFLV
RHOW OPBE KVWM UQFM PW

OLBG MGVA TMKF NWZX FFII
YXUT IHWM DHXI FZEQ VKDV
MQSW BQND YOZF TIWM JHXH
YRPA CZUG RREM VPAN WXGT
KTHN RLVH KZPG MNMV SECV
CKHO INPL HHPV PXKM BHOK
CCPD PEVX VVHO ZZQB IYIE
OUSE ZNHJ KWHY DAGT XDJD
JKJP KCSD SUZT QCXJ DVLP
AMGQ KKSH PHVK SVPC BUWZ
FIZP FUUP YKRB MGVA VA

VONV ONJL OOKS JHFF TTTE
INSE INSD REIZ WOYY QNNS
NEUN INHA LTXX BEIA NGRI
FFUN TERW ASSE RGED RUEC
KTYW ABOS XLET ZTER GEGN
ERST ANDN ULAC HTDR EINU
LUHR MARQ UANT ONJO TANE
UNAC HTSE YHSD REIY ZWOZ
WONU LGRA DYAC HTSM YSTO
SSEN ACHX EKNS VIER MBFA
ELLT YNNN NNNO OOVI ER

OLBG MGVA TMKF NWZX FFII
YXUT IHWM DHXI FZEQ VKDV
MQSW BQND YOZF TIWM JHXH
YRPA CZUG RREM VPAN WXGT
KTHN RLVH KZPG MNMV SECV
CKHO INPL HHPV PXKM BHOK
CCPD PEVX VVHO ZZQB IYIE
OUSE ZNHJ KWHY DAGT XDJD
JKJP KCSD SUZT QCXJ DVLP
AMGQ KKSH PHVK SVPC BUWZ
FIZP FUUP YKRB MGVA VA

VONV ONJL OOKS JHFF TTTE
INSE INSD REIZ WOYY QNNS
NEUN INHA LTXX BEIA NGRI
FFUN TERW ASSE RGED RUEC
KTYW ABOS XLET ZTER GEGN
ERST ANDN ULAC HTDR EINU
LUHR MARQ UANT ONJO TANE
UNAC HTSE YHSD REIY ZWOZ
WONU LGRA DYAC HTSM YSTO
SSEN ACHX EKNS VIER MBFA
ELLT YNNN NNNO OOVI ER

OLBG	MGVA	TMKF	NWZX	FFII
YXUT	IHWM	DHXI	FZEQ	VKDV
MQSW	BQND	YOZF	TIWM	JHXH
YRPA	CZUG	RREM	VPAN	WXGT
KTHN	RLVH	KZPG	MNMV	SECV
CKHO	INPL	HHPV	PXKM	BHOK
CCPD	PEVX	VVHO	ZZQB	IYIE
OUSE	ZNHJ	KWHY	DAGT	XDJD
JKJP	KCSD	SUZT	QCXJ	DVLP
AMGQ	KKSH	PHVK	SVPC	BUWZ
FIZP	FUUP	YKRB	MGVA	VA

VONV	ONJL	OOKS	JHFF	TTTE
INSE	INSD	REIZ	WOYY	QNNS
NEUN	INHA	LTXX	BEIA	NGRI
FFUN	TERW	ASSE	RGED	RUEC
KTYW	ABOS	XLET	ZTER	GEGN
ERST	ANDN	ULAC	HTDR	**EIN**U
LUHR	MARQ	UANT	ONJO	TANE
UNAC	HTSE	YHSD	REIY	ZWOZ
WONU	LGRA	DYAC	HTSM	YSTO
SSEN	ACHX	EKNS	**VIER**	MBFA
ELLT	YNNN	NNNO	OO**VI**	**ER**

OLBG MGVA TMKF NWZX FFII
YXUT IHWM DHXI FZEQ VKDV
MQSW BQND YOZF TIWM JHXH
YRPA CZUG RREM VPAN WXGT
KTHN RLVH KZPG MNMV SECV
CKHO INPL HHPV PXKM BHOK
CCPD PEVX VVHO ZZQB IYIE
OUSE ZNHJ KWHY DAGT XDJD
JKJP KCSD SUZT QCXJ DVLP
AMGQ KKSH PHVK SVPC BUWZ
FIZP FUUP YKRB MGVA VA

VONV ONJL OOKS JHFF TTTE
INSE INSD REIZ WOYY QNNS
NEUN INHA LTXX BEIA NGRI
FFUN TERW ASSE RGED RUEC
KTY**W ABOS** XLET ZTER GEGN
ERST ANDN ULAC HTDR EINU
LUHR **MARQ U**ANT ONJO TANE
UNAC HTSE YHSD REIY ZWOZ
WONU LGRA DYAC HTSM YSTO
SSEN ACHX EKNS VIER MBFA
ELLT YNNN NNNO OOVI ER

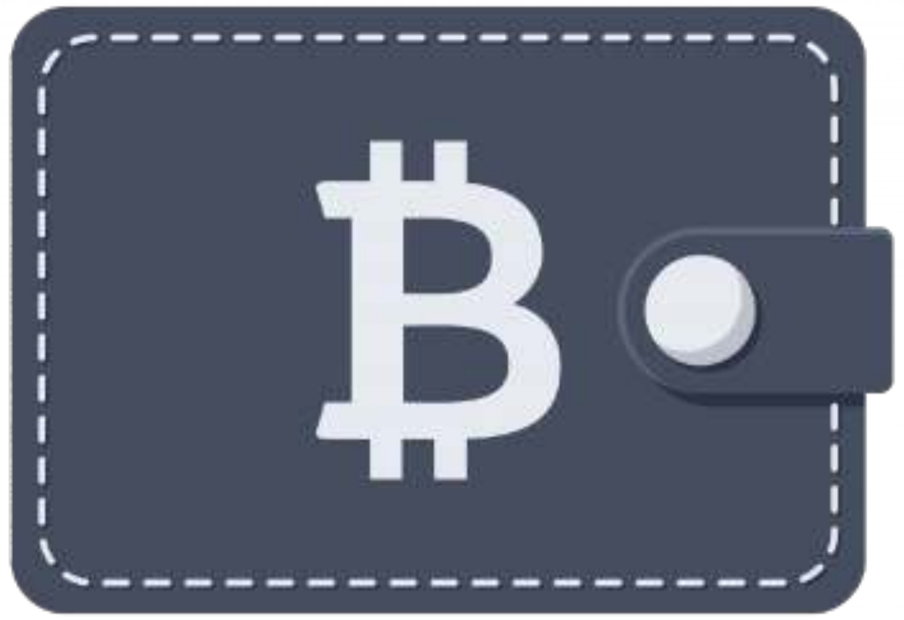
OLBG MGVA TMKF NWZX FFII
YXUT IHWM DHXI FZEQ VKDV
MQSW BQND YOZF TIWM JHXH
YRPA CZUG RREM VPAN WXGT
KTHN RLVH KZPG MNMV SECV
CKHO INPL HHPV PXKM BHOK
CCPD PEVX VVHO ZZQB IYIE
OUSE ZNHJ KWHY DAGT XDJD
JKJP KCSD SUZT QCXJ DVLP
AMGQ KKSH PHVK SVPC BUWZ
FIZP FUUP YKRB MGVA VA

VONV ONJL OOKS JHFF TTTE
INSE INSD REIZ WOYY QNNS
NEUN INHA LTXX BEIA NGRI
FFUN TERW ASSE RGED RUEC
KTYW ABOS XLET ZTER GEGN
ERST ANDN ULAC HTDR EINU
LUHR MARQ UANT ONJO TANE
UNAC HT**SE YHS**D REIY ZWOZ
WONU LGRA DYAC HTSM YSTO
SSEN ACHX LKNS VIER MBFA
ELLT YNNN NLNO OOVI ER

(Supposed to be sechs)

OLBG MGVA TMKF NWZX FFII
YXUT IHWM DHXI FZEQ VKDV
MQSW BQND YOZF TIWM JHXH
YRPA CZUG RREM VPAN WXGT
KTHN RLVH KZPG MNMV SECV
CKHO INPL HHPV PXKM BHOK
CCPD PEVX VVHO ZZQB IYIE
OUSE ZNHJ KWHY DAGT XDJD
JKJP KCSD SUZT QCXJ DVLP
AMGQ KKSH PHVK SVPC BUWZ
FIZP FUUP YKRB MGVA VA

VONV ONJL OOKS JHFF TTTE
INSE INSD REIZ WOYY QNNS
NEUN INHA LTXX BEIA NGRI
FFUN TERW ASSE RGED RUEC
KTYW ABOS XLET ZTER GEGN
ERST ANDN ULAC HTDR EINU
LUHR MARQ UANT ONJO TANE
UNAC HTSE YHSD REIY ZWOZ
WONU LGRA DYAC HTSM YSTO
SSEN ACHX EKNS VIER MBFA
ELLT YNNN NNNO OOVI ER



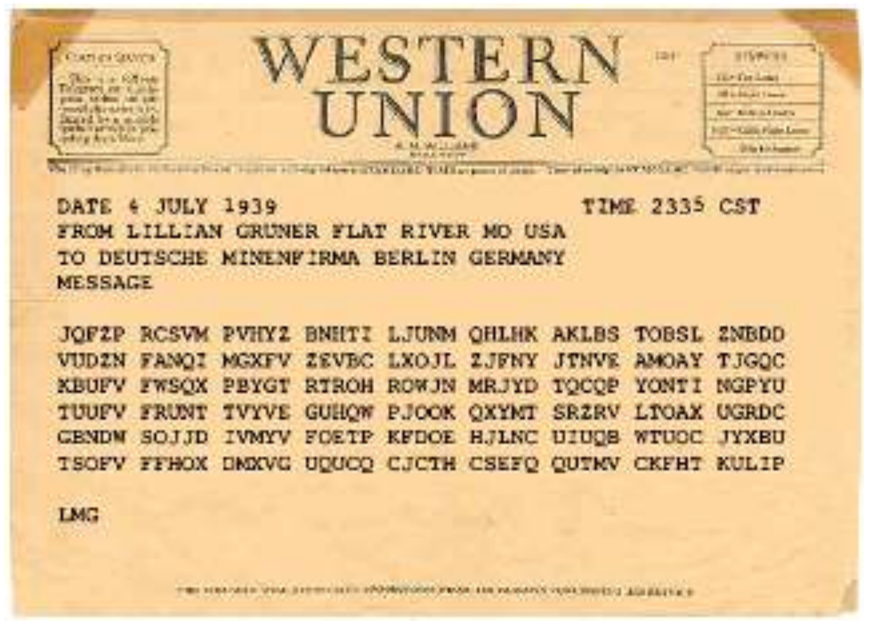
WESTERN UNION

DATE 4 JULY 1939 TIME 2335 CST
FROM LILLIAN GRUNER FLAT RIVER MO USA
TO DEUTSCHE MINENFIRMA BERLIN GERMANY
MESSAGE

JQFZP RCSVM PVHYZ BNHTI LJUNM QHLHK AKLBS TOBSL ZNBDD
VUDZN FANQI MGXFV ZEVBC LXOJL ZJFNY JTNVE AMOAY TJGQC
KBUFV FWSQX PBYGT RTROH ROWJN MRJYD TQCQP YONTI NGPYU
TUUFV FRUNT TVYVE GUHQW PJOOK QXYMT SRZRV LTOAX UGRDC
GENDW SOJJD IVMYV FOETP KFDOE HJLNC UIUQB WTUOC JYXBU
TSOFV FFHOX DNXVG UQUCQ CJCTH CSEFQ QUTMV CKFHT KULIP

LMG

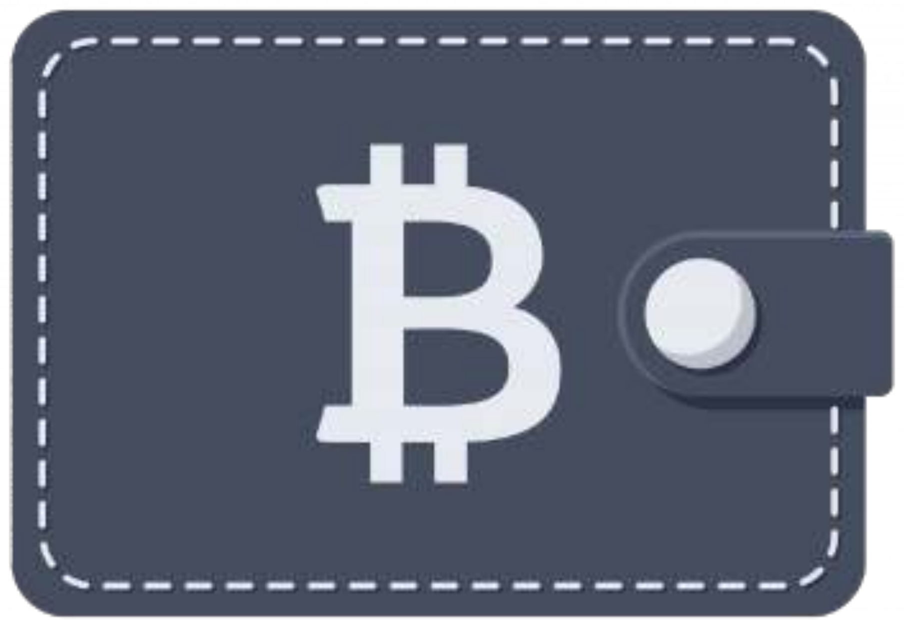
($26bn missing)


WESTERN UNION

DATE 4 JULY 1939 TIME 2335 CST
FROM LILLIAN GRUNER FLAT RIVER MO USA
TO DEUTSCHE MINENFIRMA BERLIN GERMANY
MESSAGE

JQFZP RCSVM PVHYZ BNHTI LJUNM QHLHK AKLBS TOBSL ZNBDD
VUDZN FANQI MGXFV ZEVBC LXOJL ZJFNY JTNVE AMOAY TJGQC
KBUFV FWSQX PBYGT RTROH ROWJN MRJYD TQCQP YONTI NGPYU
TUUFV FRUNT TVYVE GUHQW PJOOK QXYMT SRZRV LTOAX UGRDC
GBNDW SOJJD IVMYV FOETP KFDOE HJLNC UIUQB WTUOC JYXBU
TSOFV FFHOX DNXVG UQUCQ CJCTH CSEFQ QUTMV CKFHT KULIP

LMG



Structured

Hard rules

Clean

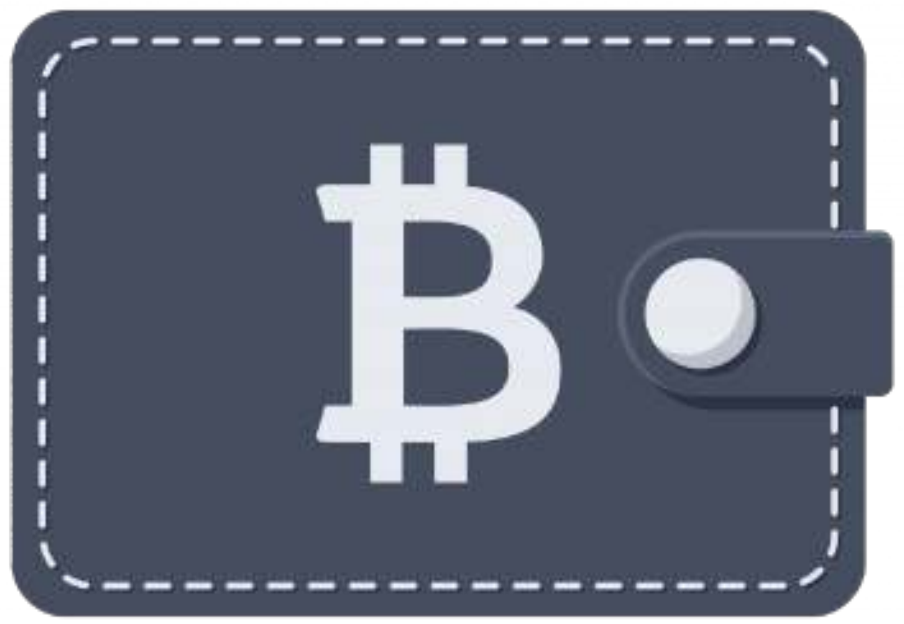
Clear result

WESTERN UNION

DATE 4 JULY 1939 TIME 2335 CST
FROM LILLIAN GRUNER FLAT RIVER MO USA
TO DEUTSCHE MINENFIRMA BERLIN GERMANY
MESSAGE

JQFZP RCSVM PVHYZ BNHTI LJUNM QHLHK AKLBS TOBSL ZNBDD
VUDZN FANQI MGXFV ZEVBC LXOJL ZJFNY JTNVE AMOAY TJGQC
KBUFV FWSQX PBYGT RTROH ROWJN MRJYD TQCQP YONTI NGPYU
TUUFV FRUNT TVYVE GUHQW PJOOK QXYMT SRZRV LTOAX UGRDC
GBNDW SOJJD IVMYV FOETP KFDOE HJLNC UIUQB WTUOC JYXBU
TSOFV FFHOX DMXVG UQUCQ CJCTH CSEFQ QUTMV CKFHT KULIP

LMG



WESTERN
UNION

DATE 4 JULY 1939 TIME 2335 CST
FROM LILLIAN GRUNER FLAT RIVER MO USA
TO DEUTSCHE MINENFIRMA BERLIN GERMANY
MESSAGE

JQFZP RCSVM PVHYZ BNHTI LJUNM QHLHK AKLBS TOBSL ZNBDD
VUDZN FANQI MGXFV ZEVBC LXOJL ZJFNY JTNVE AMOAY TJGQC
KBUFV FWSQX PBYGT RTROH ROWJN MRJYD TQCQP YONTI NGPYU
TUUFV FRUNT TVYVE GUHQW PJOOK QXYMT SRZRV LTOAX UGRDC
GBNDW SOJJD IVMYV FOETP KFDOE HJLNC UIUQB WTUOC JYXBU
TSOFV FFHOX DMXVG UQUCQ CJCTH CSEFQ QUTMV CKFHT KULIP

LMG

Structured

Hard rules

Clean

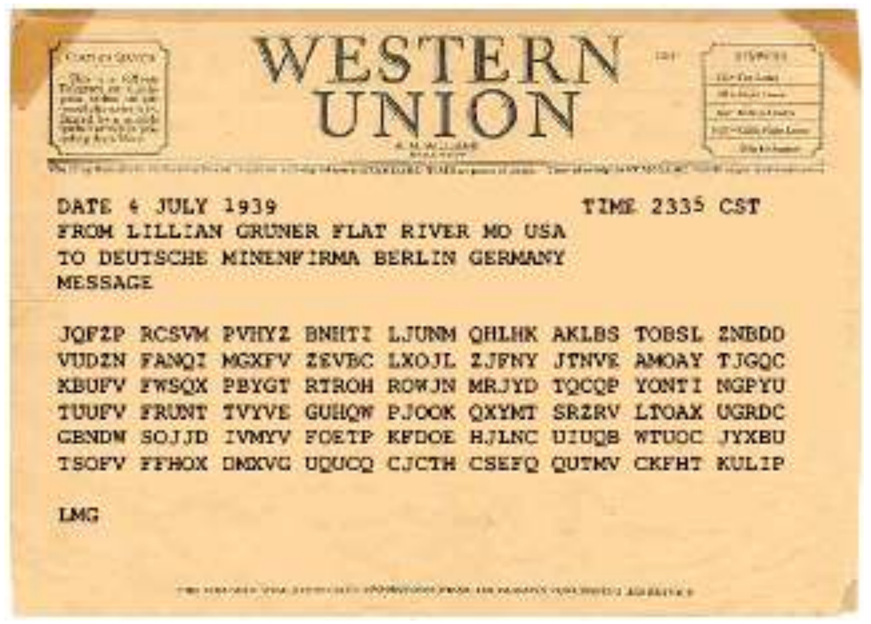
Clear result

Unstructured

Soft rules

Noisy

Unclear result



WESTERN UNION

DATE 4 JULY 1939 TIME 2335 CST
FROM LILLIAN GRUNER FLAT RIVER MO USA
TO DEUTSCHE MINENFIRMA BERLIN GERMANY
MESSAGE

JQFZP RCSVM PVHYZ BNHTI LJUNM QHLHK AKLBS TOBSL ZNBDD
VUDZN FANQI MGXFV ZEVBC LXOJL ZJFNY JTNVE AMOAY TJGQC
KBUFV FWSQX PBYGT RTROH ROWJN MRJYD TQCQP YONTI NGPYU
TUUFV FRUNT TVYVE GUHQW PJOOK QXYMT SRZRV LTOAX UGRDC
GBNDW SOJJD IVMYV FOETP KFDOE HJLNC UIUQB WTUOC JYXBU
TSOFV FFHOX DMXVG UQUCQ CJCTH CSEFQ QUTMV CKFHT KULIP

LMG

Structured

Hard rules

Clean

Clear result

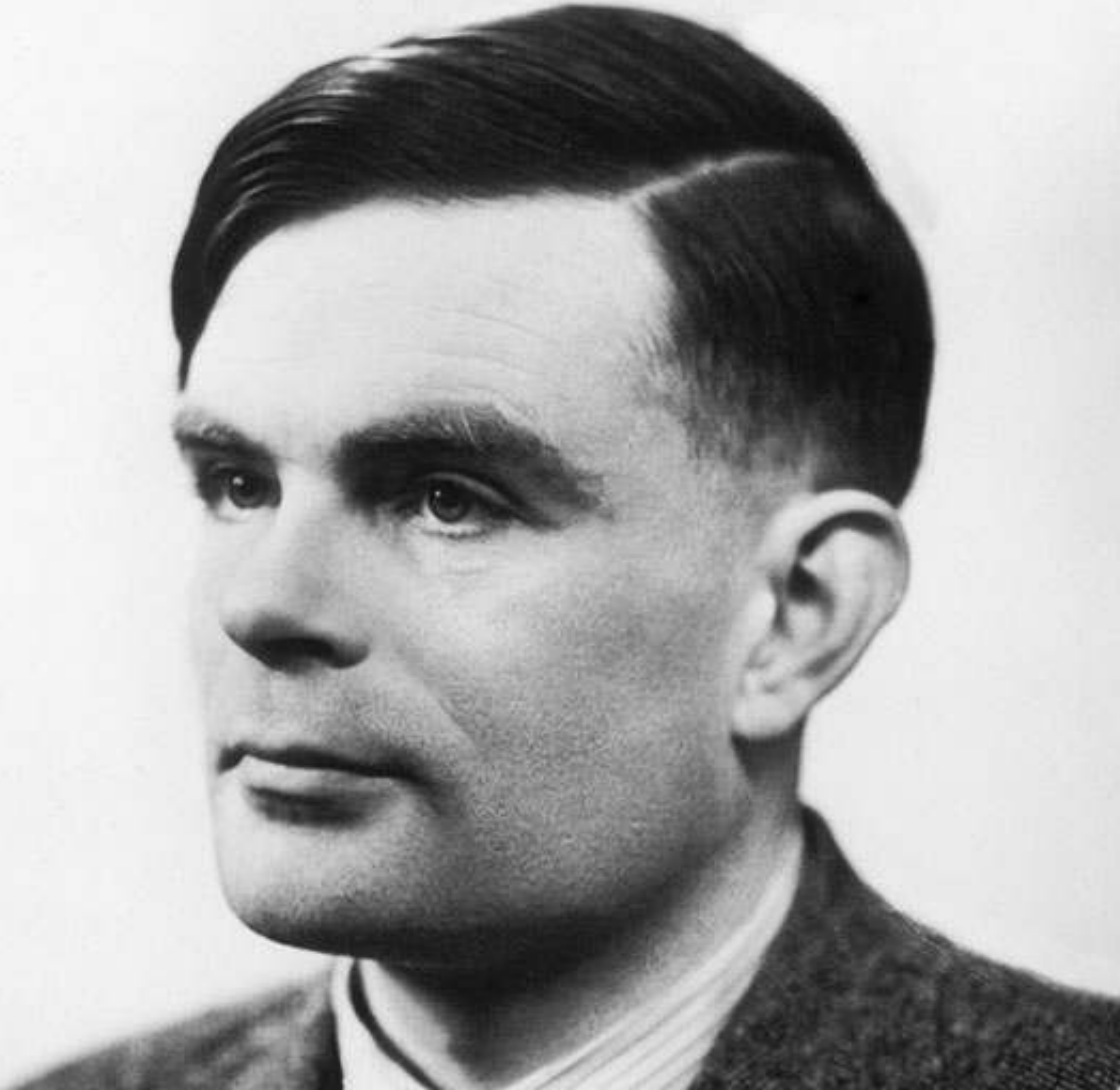
Unstructured

Soft rules

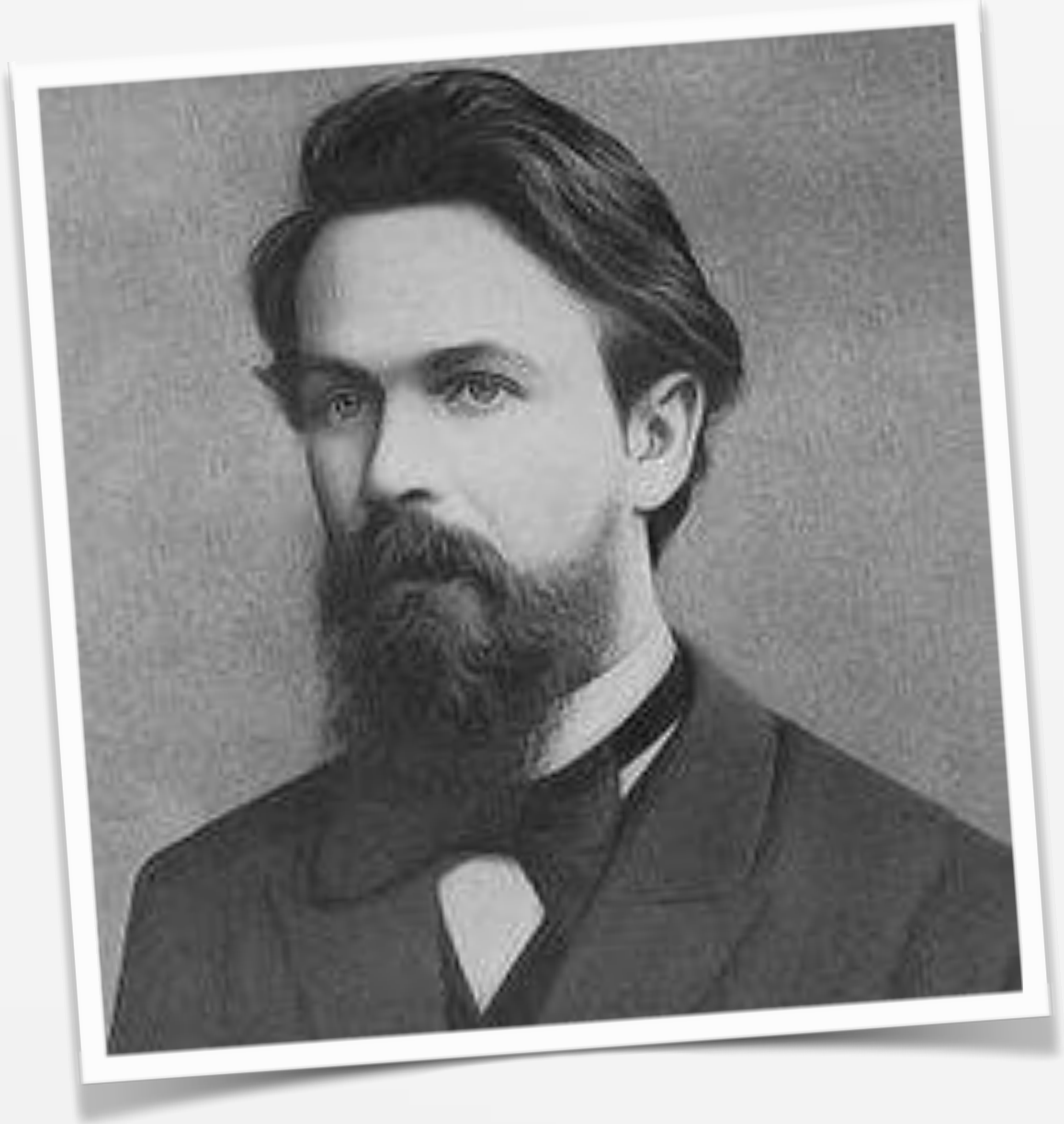
Noisy

Unclear result

Language models



Language models



OLBG MGVA TMKF NWZX FFII

YXUT IHWM DHXI FZEQ VKDV

MQSW BQND YOZF TIWM JHXH

YRPA CZUG RREM VPAN WXGT

KTHN RLVH KZPG MNMV SECV

CKHO INPL HHPV PXKM BHOK

CCPD PEVX VVHO ZZQB IYIE

OUSE ZNHJ KWHY DAGT XDJD

JKJP KCSD SUZT QCXJ DVLP

AMGQ KKSH PHVK SVPC BUWZ

FIZP FUUP YKRB MGVA VA

3% chance

VONV ONJL OOKS JHFF TTTE

INSE INSD REIZ WOYY QNNS

NEUN INHA LTXX BEIA NGRI

FFUN TERW ASSE RGED RUEC

KTYW ABOS XLET ZTER GEGN

ERST ANDN ULAC HTDR EINU

LUHR MARQ UANT ONJO TANE

UNAC HT**SE CHS**D REIY ZWOZ

WONU LGRA DYAC HTSM YSTO

SSEN ACHX EKNS VIER MBFA

ELLT YNNN NNNO OOVI ER

78% chance

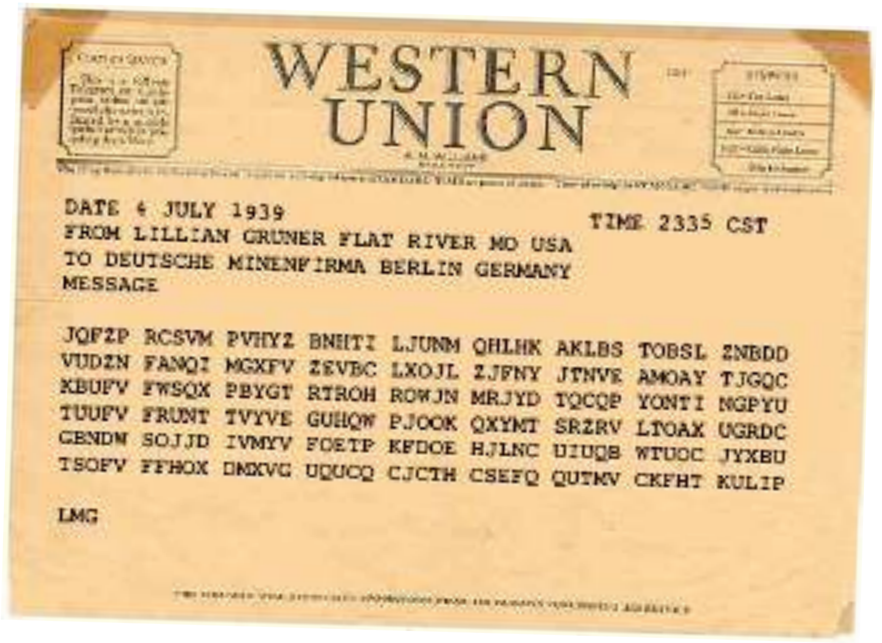
OLBG MGVA TMKF NWZX FFII
YXUT IHWM DHXI FZEQ VKDV
MQSW BQND YOZF TIWM JHXH
YRPA CZUG RREM VPAN WXGT
KTHN RLVH KZPG MNMV SECV
CKHO INPL HHPV PXKM BHOK
CCPD PEVX VVHO ZZQB IYIE
OUSE ZNHJ KWHY DAGT XDJD
JKJP KCSD SUZT QCXJ DVLP
AMGQ KKSH PHVK SVPC BUWZ
FIZP FUUP YKRB MGVA VA

3% chance

VONV ONJL OOKS JHFF TTTE
INSE INSD REIZ WOYY QNNS
NEUN INHA LTXX BEIA NGRI
FFUN TERW ASSE RGED RUEC
KTYW ABOS XLET ZTER GEGN
ERST ANDN ULAC HTDR EINU
LUHR MARQ UANT ONJO TANE
UNAC HT**SE YHS**D REIY ZWOZ
WONU LGRA DYAC HTSM YSTO
SSEN ACHX EKNS VIER MBFA
ELLT YNNN NNNO OOVI ER

70% chance

?



WESTERN UNION

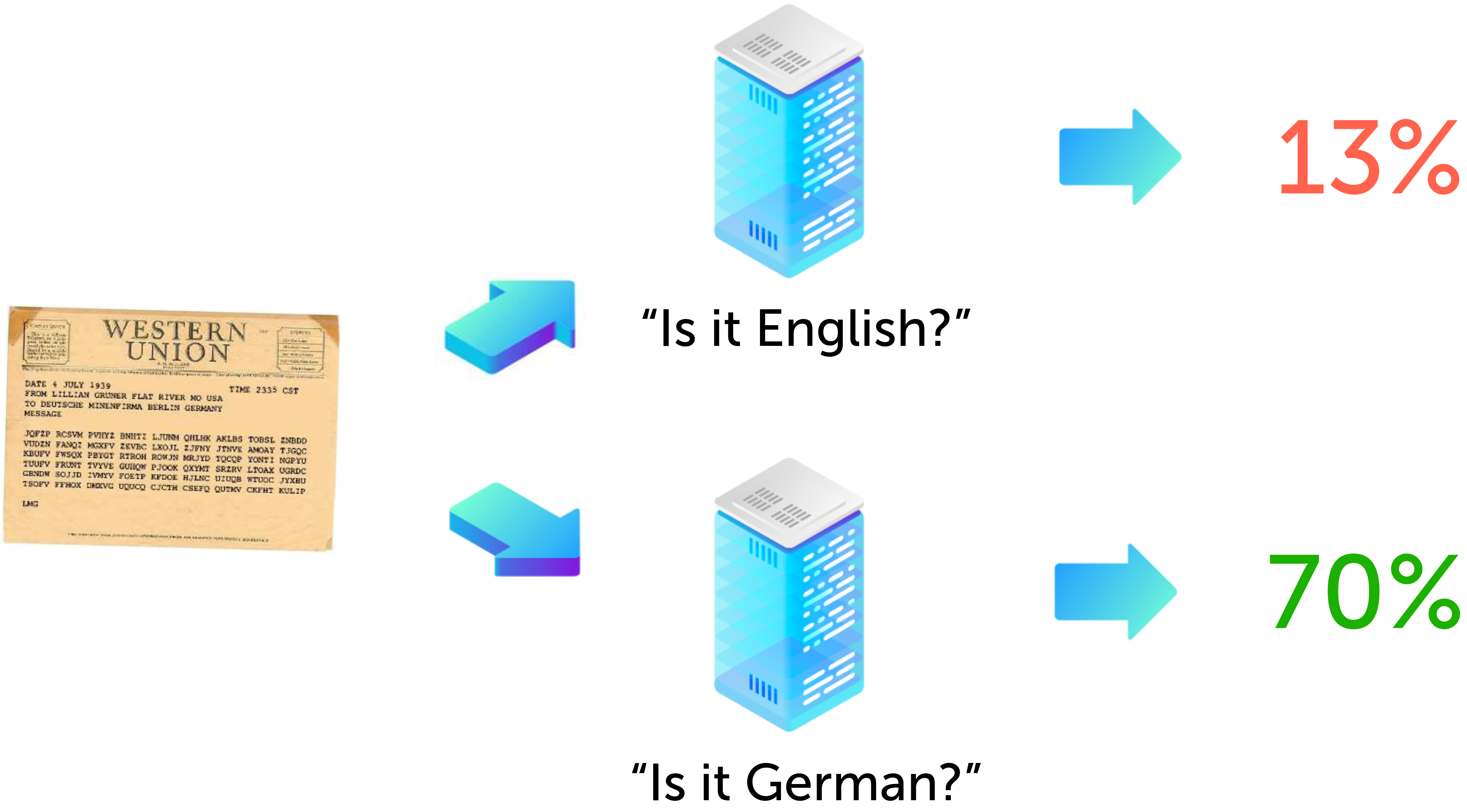
DATE 4 JULY 1939 TIME 2335 CST
FROM LILLIAN GRUNER FLAT RIVER MO USA
TO DEUTSCHE MINENFIRMA BERLIN GERMANY
MESSAGE

JQFZP RCSVM PVHYZ BNHTI LJUNM QHLHK AKLBS TOBSL ZNBDD
VUDZN FANQI MGXFV ZEVBC LXOJL ZJFNY JTNVE AMOAY TJGQC
KBUFV FWSQX PBYGT RTROH ROWJN MRJYD TQCQP YONTI NGPYU
TUUFV FRUNT TVYVE GUHQW PJOOK QXYMT SRZRV LTOAX UGRDC
GBNDW SOJJD IVMYV FOETP KFDOE HJLNC UIUQB WTUOC JYXBU
TSOFV FFHOX DMXVG UQUCQ CJCTH CSEFQ QUTMV CKFHT KULIP

LMG

70%

“Is it German?”


WESTERN UNION
“Is it English?”
13%
“Is it German?”
70%

Data classification

Machine translation

Speech recognition

Language generation

Part-of-speech tagging

Handwriting recognition

…

Data classification

Machine translation

Speech recognition

Language generation

Part-of-speech tagging

Handwriting recognition

...

What is

Unstructured data?

Route	Period	Ref crossing	Total in EUR 2014
Central Med	2010-2015	285,700	3,643,000,000
East Borders	2010-2015	5,217	72,000,000
East Med Land	2010-2015	108,089	1,751,000,000
East Med Sea	2010-2015	61,922	1,053,000,000
West African	2010-2015	1,040	4,000,000
West Balkans	2010-2015	74,347	1,589,000,000
West Med	2010-2015	29,487	251,000,000

Structured data

Emails

Tweets

Comments

Reviews

Transcripts

Written notes

SMS messages

Wiki articles

Blogs

Academic papers

Presentations

Reports

Diary entries

Webpages

News articles

Health records

Police reports

Chat messages

Forum posts

Books

Interviews

10%

90%

of an organization's data is unstructured*

*Sources: McKinsey, IDC

Agenda

Origins of language models

What is unstructured data?

- Some case studies

Types of language models

- Count based (bag of words, *n*-grams)
- Continuous space
- Bonus: the class of 2018

Wrap-up and questions

Case study 1

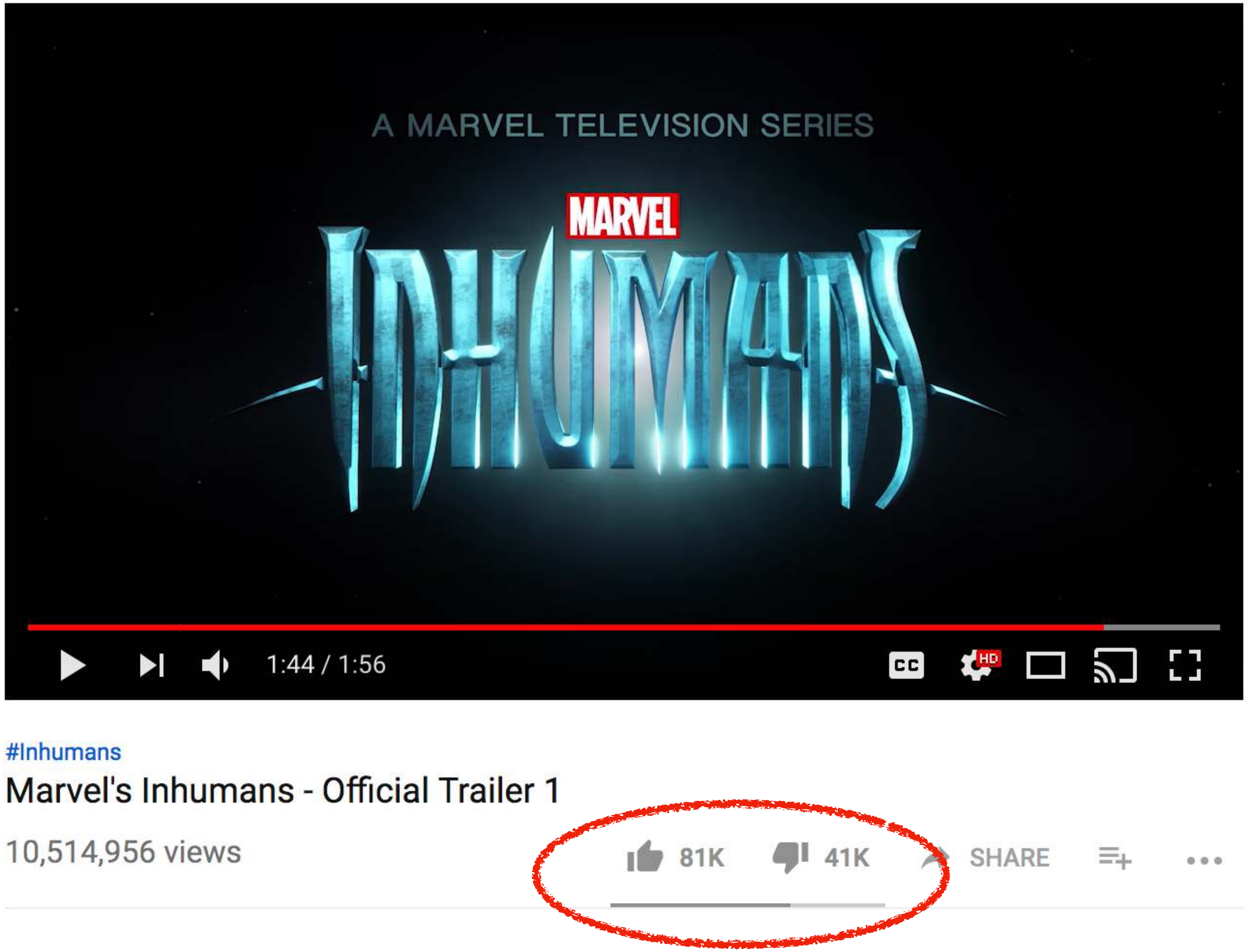
Trailer sentiment

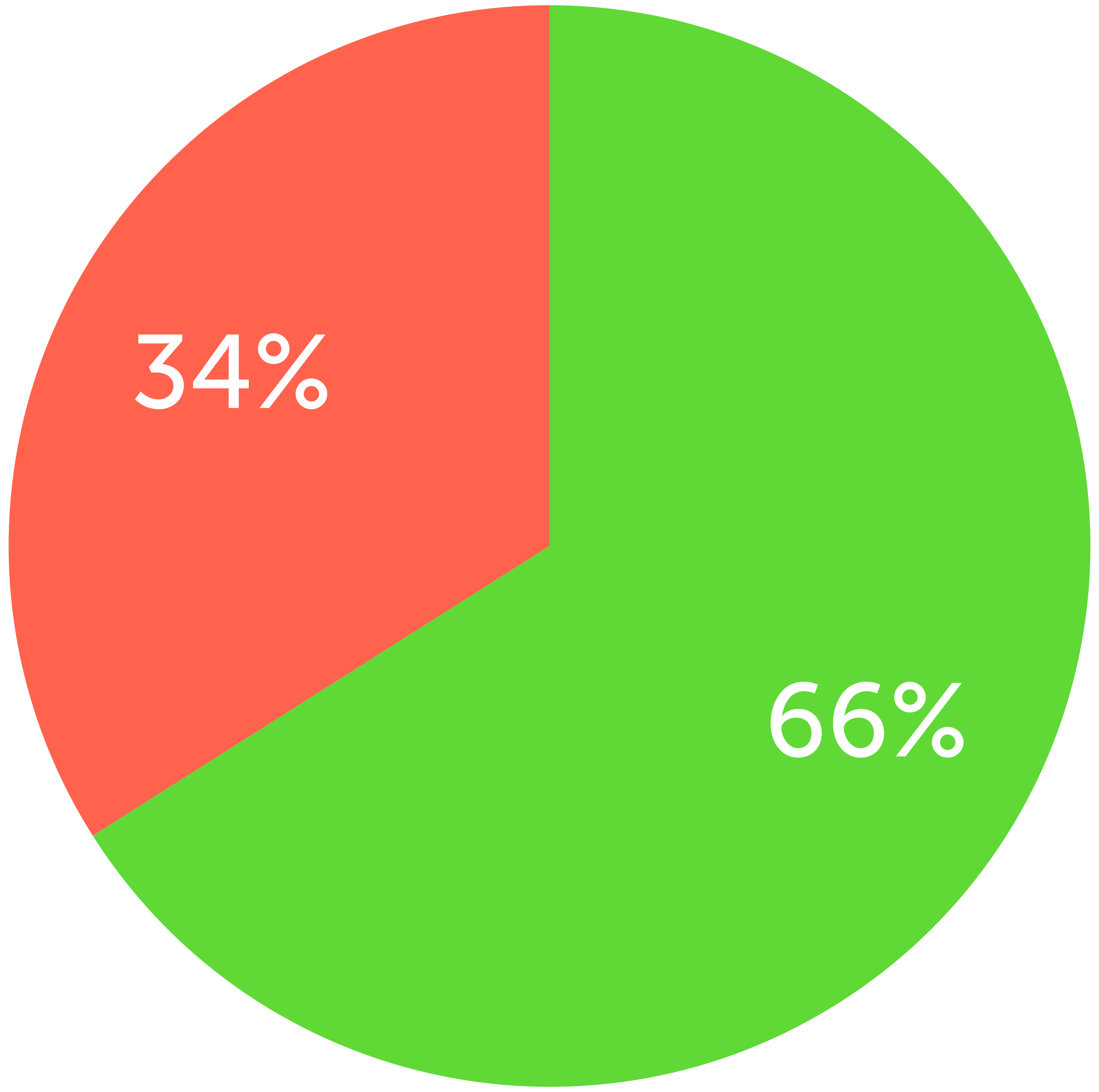
#Inhumans

Marvel's Inhumans - Official Trailer 1

10,514,956 views

81K 41K SHARE


A MARVEL TELEVISION SERIES
MARVEL
INHUMANS
1:44 / 1:56
CC
HD
#Inhumans
Marvel's Inhumans - Official Trailer 1
10,514,956 views
81K
41K
SHARE



How do we get richer insight?

Sexhaie 11 months ago
This look like it had potential to be like a superhero dramedy, or like a show that's like a parody of other marvel shows but also have a serious story or plot.
But looks like it's trying to be a marvel version of game of thrones.
641 REPLY
View 11 replies

Arman Taghehchian 1 year ago
Honestly this wouldve been better off animated
1.1K REPLY
View 17 replies

RoadSamurai 1 year ago
Even for tv, it looks cheap
228 REPLY
View 5 replies

Oliver Clothesoff 1 year ago
Instead of wasting the budget on IMAX cameras, use the budget for the CGI instead
392 REPLY
View 3 replies

pyrosdestiny 11 months ago
The guy who made iron first made this. Makes sense.
100 REPLY
View 2 replies

WUS POPPIN JIMBO 1 year ago
It looks like they bought their costumes from Party City
773 REPLY
View 11 replies

best joker 1 year ago
Nobody asked for this, Marvel.
779 REPLY
View 16 replies

chef_mantis 1 year ago
Tony Stark: Don't do anything stupid.

abc: Come on! What's the worst that could happen?

Read more
23 REPLY

ARYAN OF BUL 11 months ago
And that's why aliens are not talking to us.
141 REPLY

Christopher Gibbs 1 year ago
fan made?
435 REPLY
View 7 replies

Gagdet View 10 months ago
My reaction when I saw this trailer 1:11
282 REPLY
View 3 replies

karma delivery 1 year ago
How in 2017 does the shit look so cheap and cheesy? I felt like I was 10 again watching an episode of Xena warrior princess
1.3K REPLY
View 40 replies

kalaikamalu 11 months ago
1:39 - I'm sorry, did anyone else hear that odd out of place punch sound.
27 REPLY
View 2 replies

Smokey Badd 11 months ago (edited)
The best part is 1:57
68 REPLY
View 2 replies

beatniece 10 months ago
what was the budget on this? a six pack of beer and some dry donuts?
16 REPLY

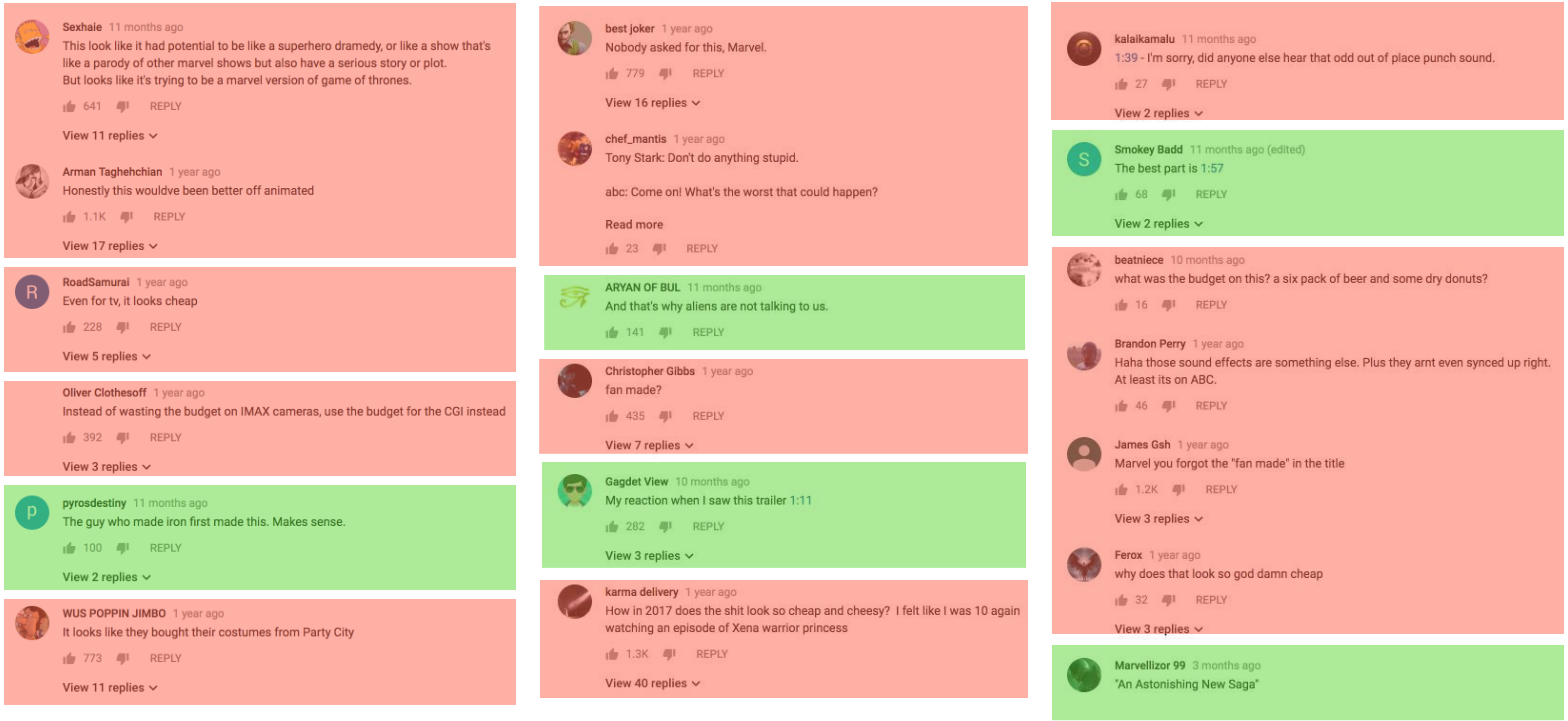
Brandon Perry 1 year ago
Haha those sound effects are something else. Plus they arnt even synced up right. At least its on ABC.
46 REPLY

James Gsh 1 year ago
Marvel you forgot the "fan made" in the title
1.2K REPLY
View 3 replies

Ferox 1 year ago
why does that look so god damn cheap
32 REPLY
View 3 replies

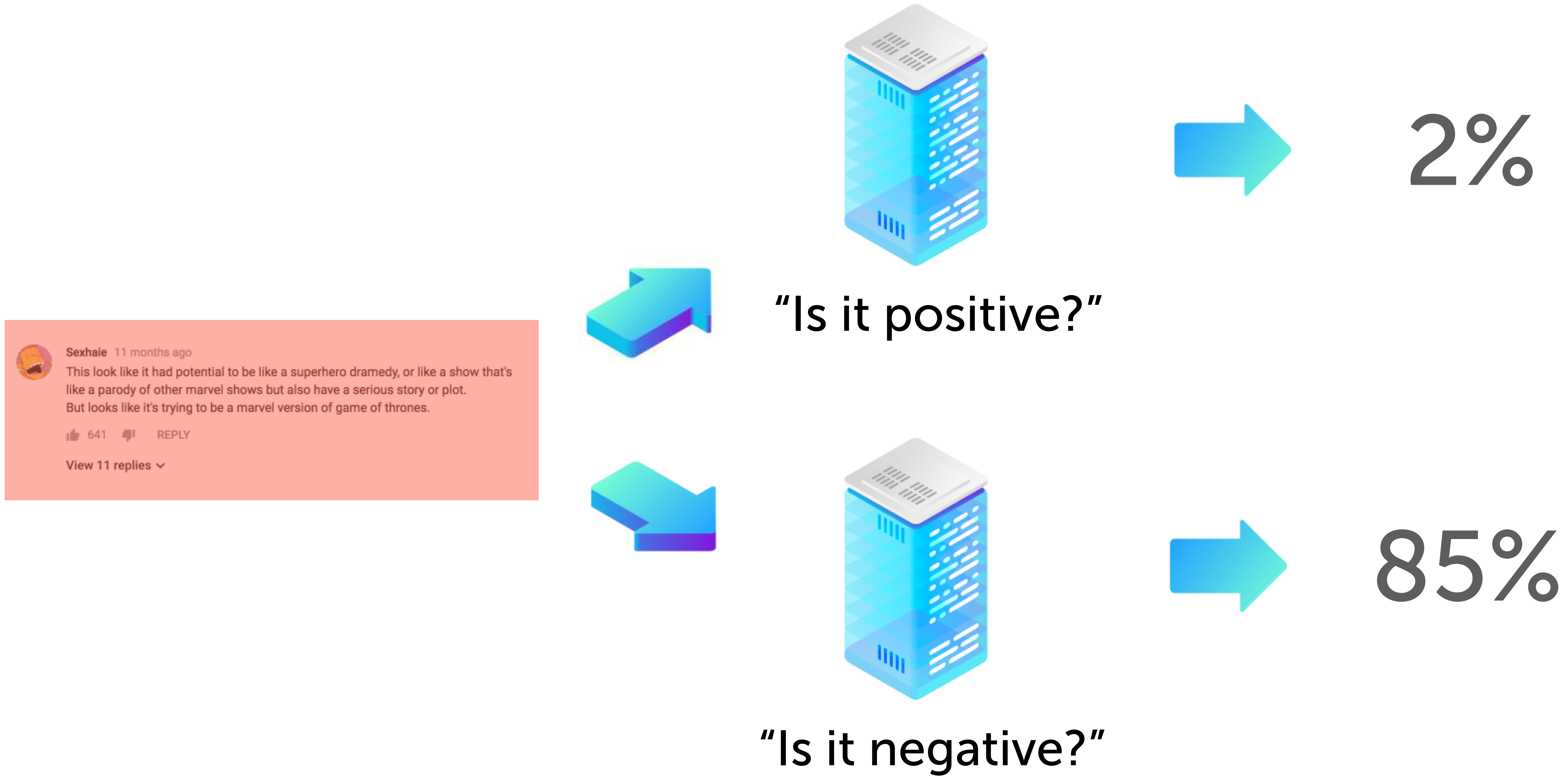
Marvellizor 99 3 months ago
"An Astonishing New Saga"

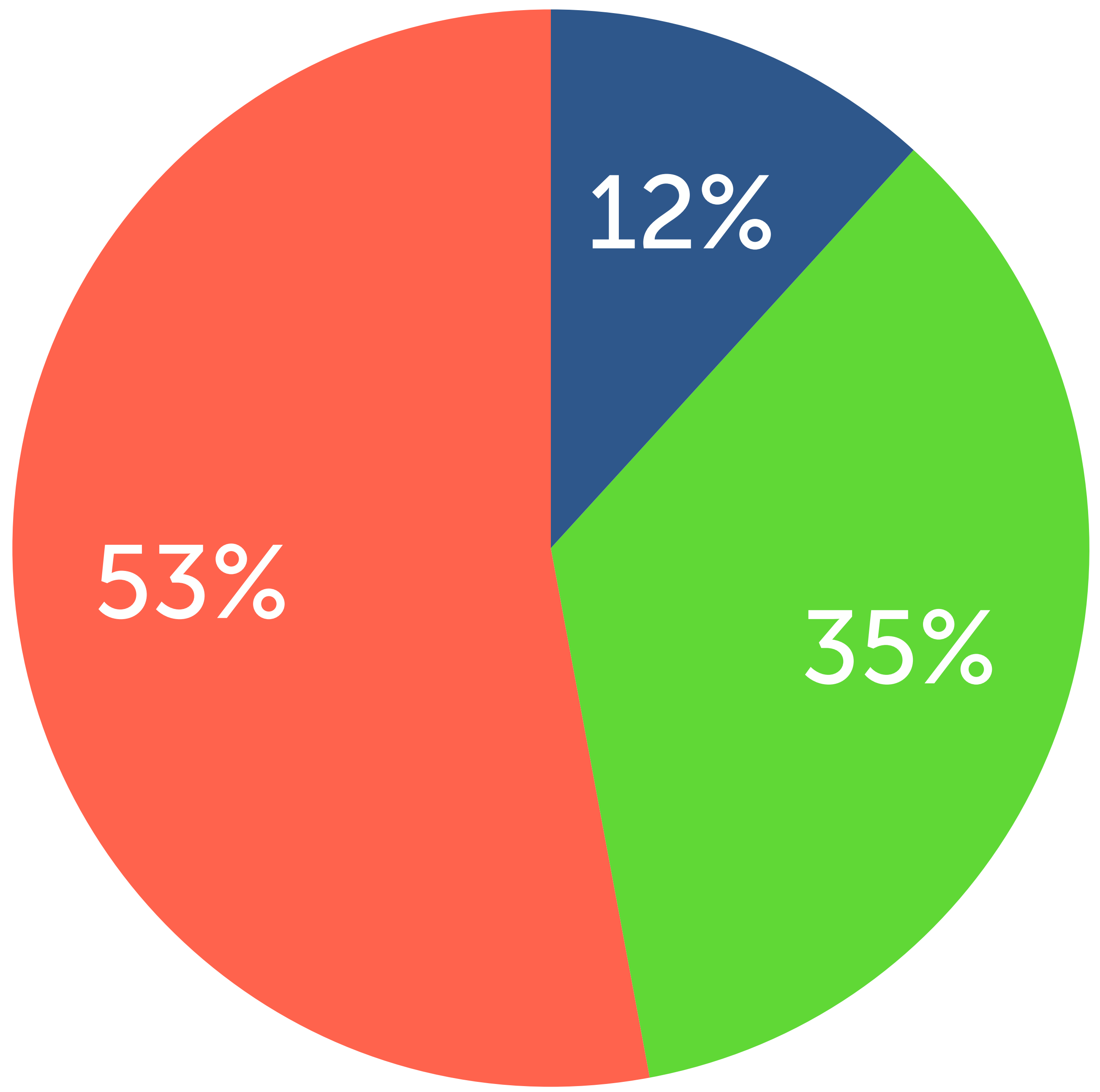
cancelled after one season



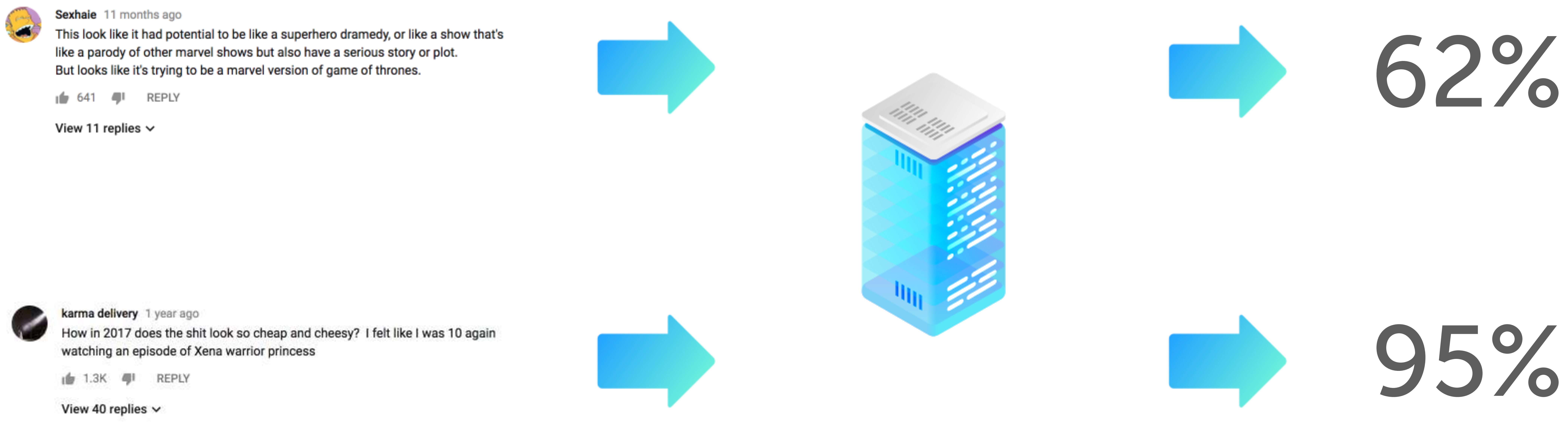
Sexhaie 11 months ago
This look like it had potential to be like a superhero dramedy, or like a show that's like a parody of other marvel shows but also have a serious story or plot.
But looks like it's trying to be a marvel version of game of thrones.
641 REPLY
View 11 replies
Arman Taghehchian 1 year ago
Honestly this wouldve been better off animated
1.1K REPLY
View 17 replies
R
RoadSamurai 1 year ago
Even for tv, it looks cheap
228 REPLY
View 5 replies
Oliver Clothesoff 1 year ago
Instead of wasting the budget on IMAX cameras, use the budget for the CGI instead
392 REPLY
View 3 replies
p
pyrosdestiny 11 months ago
The guy who made iron first made this. Makes sense.
100 REPLY
View 2 replies
WUS POPPIN JIMBO 1 year ago
It looks like they bought their costumes from Party City
773 REPLY
View 11 replies
best joker 1 year ago
Nobody asked for this, Marvel.
779 REPLY
View 16 replies
chef_mantis 1 year ago
Tony Stark: Don't do anything stupid.
abc: Come on! What's the worst that could happen?
Read more
23 REPLY
ARYAN OF BUL 11 months ago
And that's why aliens are not talking to us.
141 REPLY
Christopher Gibbs 1 year ago
fan made?
435 REPLY
View 7 replies
Gagdet View 10 months ago
My reaction when I saw this trailer 1:11
282 REPLY
View 3 replies
karma delivery 1 year ago
How in 2017 does the shit look so cheap and cheesy? I felt like I was 10 again watching an episode of Xena warrior princess
1.3K REPLY
View 40 replies
kalaikamalu 11 months ago
1:39 - I'm sorry, did anyone else hear that odd out of place punch sound.
27 REPLY
View 2 replies
S
Smokey Badd 11 months ago (edited)
The best part is 1:57
68 REPLY
View 2 replies
beatniece 10 months ago
what was the budget on this? a six pack of beer and some dry donuts?
16 REPLY
Brandon Perry 1 year ago
Haha those sound effects are something else. Plus they arnt even synced up right. At least its on ABC.
46 REPLY
James Gsh 1 year ago
Marvel you forgot the "fan made" in the title
1.2K REPLY
View 3 replies
Ferox 1 year ago
why does that look so god damn cheap
32 REPLY
View 3 replies
Marvellizor 99 3 months ago
"An Astonishing New Saga"

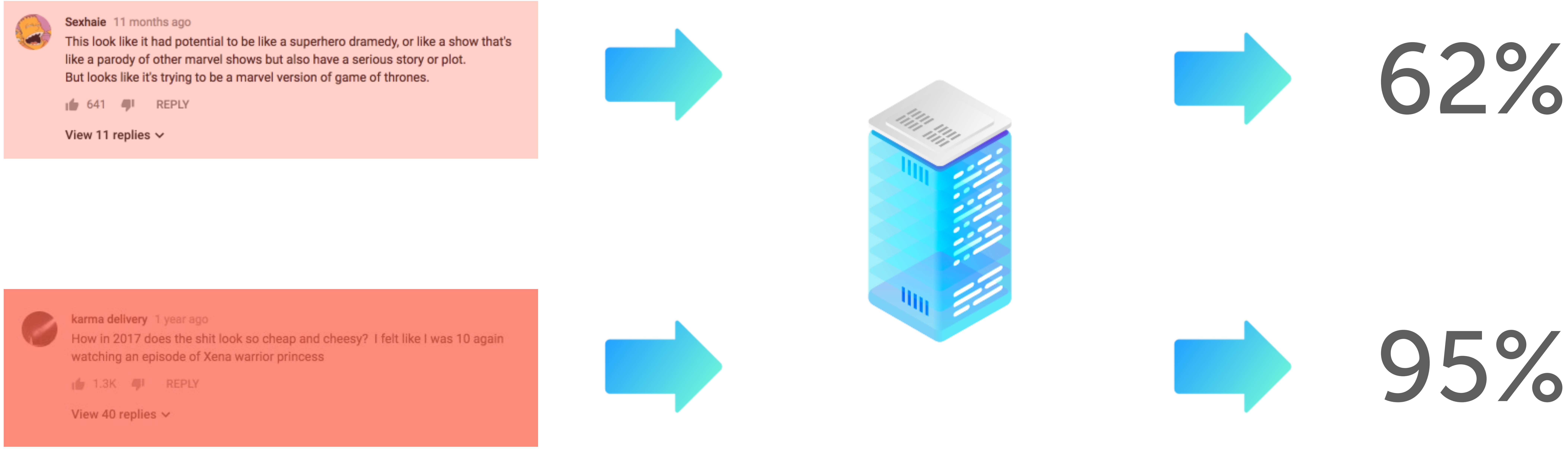

Sexhaie 11 months ago
This look like it had potential to be like a superhero dramedy, or like a show that's like a parody of other marvel shows but also have a serious story or plot.
But looks like it's trying to be a marvel version of game of thrones.
641
REPLY
View 11 replies
"Is it positive?"
2%
"Is it negative?"
85%


Sexhale 11 months ago
This look like it had potential to be like a superhero dramedy, or like a show that's like a parody of other marvel shows but also have a serious story or plot.
But looks like it's trying to be a marvel version of game of thrones.
641
REPLY
View 11 replies
"Is it positive?"
2%
"Is it negative?"
85%

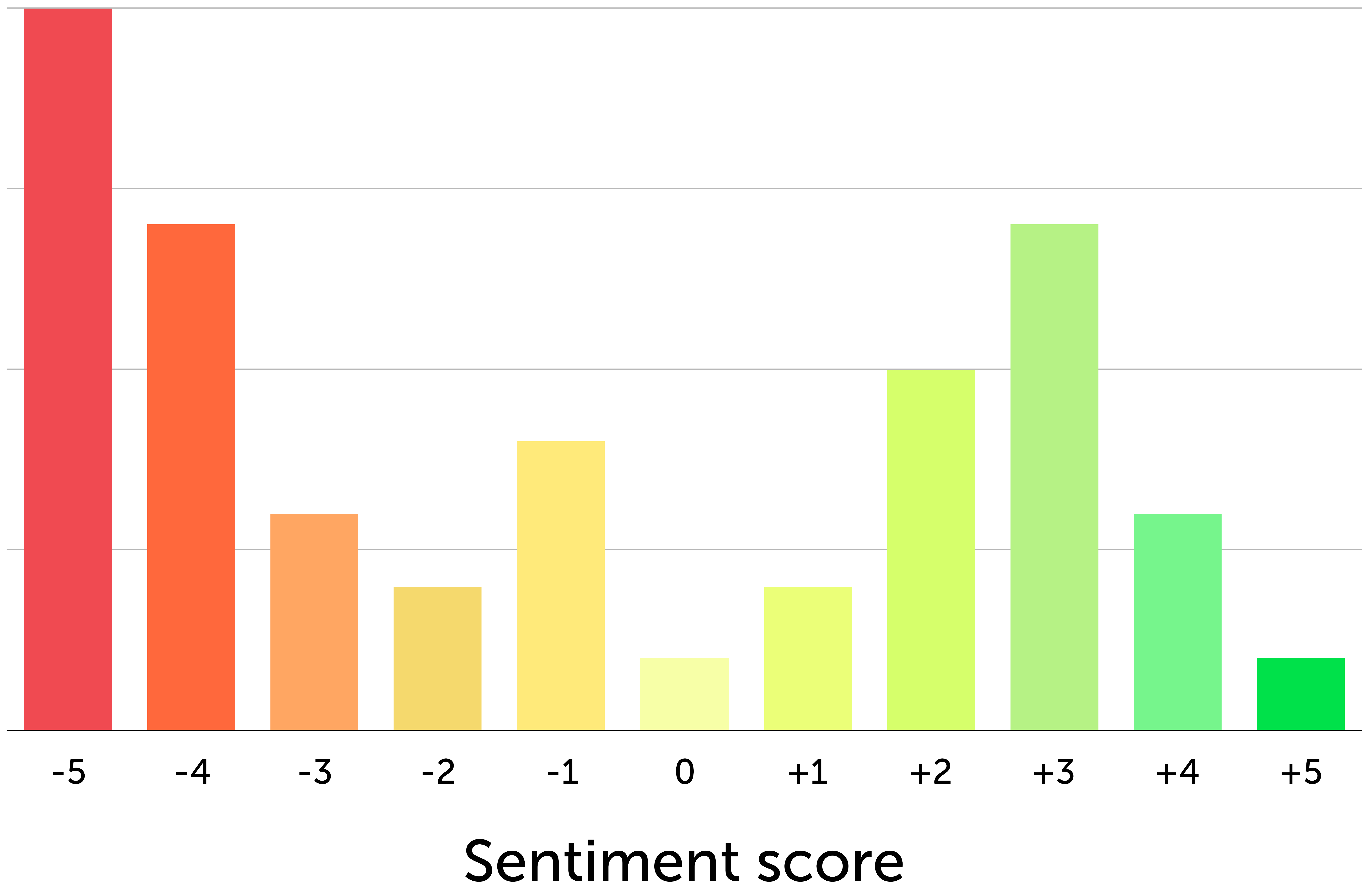


Positive, negative, neutral


Sexhaie 11 months ago
This look like it had potential to be like a superhero dramedy, or like a show that's like a parody of other marvel shows but also have a serious story or plot.
But looks like it's trying to be a marvel version of game of thrones.
641
REPLY
View 11 replies
karma delivery 1 year ago
How in 2017 does the shit look so cheap and cheesy? I felt like I was 10 again watching an episode of Xena warrior princess
1.3K
REPLY
View 40 replies
62%
95%



Sexhaie 11 months ago
This look like it had potential to be like a superhero dramedy, or like a show that's like a parody of other marvel shows but also have a serious story or plot.
But looks like it's trying to be a marvel version of game of thrones.
641
REPLY
View 11 replies
karma delivery 1 year ago
How in 2017 does the shit look so cheap and cheesy? I felt like I was 10 again watching an episode of Xena warrior princess
1.3K
REPLY
View 40 replies
62%
95%


-5
-4
-3
-2
-1
0
+1
+2
+3
+4
+5
Sentiment score

Sexhaie 11 months ago
This look like it had potential to be like a superhero dramedy, or like a show that's like a parody of other marvel shows but also have a serious story or plot.
But looks like it's trying to be a marvel version of game of thrones.
641 REPLY
View 11 replies

Arman Taghehchian 1 year ago
Honestly this wouldve been better off animated
1.1K REPLY
View 17 replies

RoadSamurai 1 year ago
Even for tv, it looks cheap
228 REPLY
View 5 replies

Oliver Clothesoff 1 year ago
Instead of wasting the budget on IMAX cameras, use the budget for the CGI instead
392 REPLY
View 3 replies

pyrosdestiny 11 months ago
The guy who made iron first made this. Makes sense.
100 REPLY
View 2 replies

WUS POPPIN JIMBO 1 year ago
It looks like they bought their costumes from Party City
773 REPLY
View 11 replies

best joker 1 year ago
Nobody asked for this, Marvel.
779 REPLY
View 16 replies

chef_mantis 1 year ago
Tony Stark: Don't do anything stupid.

abc: Come on! What's the worst that could happen?

Read more
23 REPLY

ARYAN OF BUL 11 months ago
And that's why aliens are not talking to us.
141 REPLY

Christopher Gibbs 1 year ago
fan made?
435 REPLY
View 7 replies

Gagdet View 10 months ago
My reaction when I saw this trailer 1:11
282 REPLY
View 3 replies

karma delivery 1 year ago
How in 2017 does the shit look so cheap and cheesy? I felt like I was 10 again watching an episode of Xena warrior princess
1.3K REPLY
View 40 replies

kalaikamalu 11 months ago
1:39 - I'm sorry, did anyone else hear that odd out of place punch sound.
27 REPLY
View 2 replies

Smokey Badd 11 months ago (edited)
The best part is 1:57
68 REPLY
View 2 replies

beatniece 10 months ago
what was the budget on this? a six pack of beer and some dry donuts?
16 REPLY

Brandon Perry 1 year ago
Haha those sound effects are something else. Plus they arnt even synced up right. At least its on ABC.
46 REPLY

James Gsh 1 year ago
Marvel you forgot the "fan made" in the title
1.2K REPLY
View 3 replies

Ferox 1 year ago
why does that look so god damn cheap
32 REPLY
View 3 replies

Marvellizor 99 3 months ago
"An Astonishing New Saga"

unlikeable wooden Medusa budget dialog story cheesy silly effects CGI cheap

explosions Joey Iwan Rheon gorgons Quake Kamala Khan villains choreography villainy comedic

(Based on TF-IDF on top and bottom quartile w.r.t sentiment) unlikeable

wooden

Medusa

budget

dialog

sto

silly

effects

CGI

cheap

explosions

Iwan Rheon

Joey

gorgons

Quake

Kamala Khan

villains

choreography

villainy

comedic

Trailer 2: more explosions, less dialogue

(Based on TF-IDF on top and bottom quartile w.r.t sentiment)

Case study 1

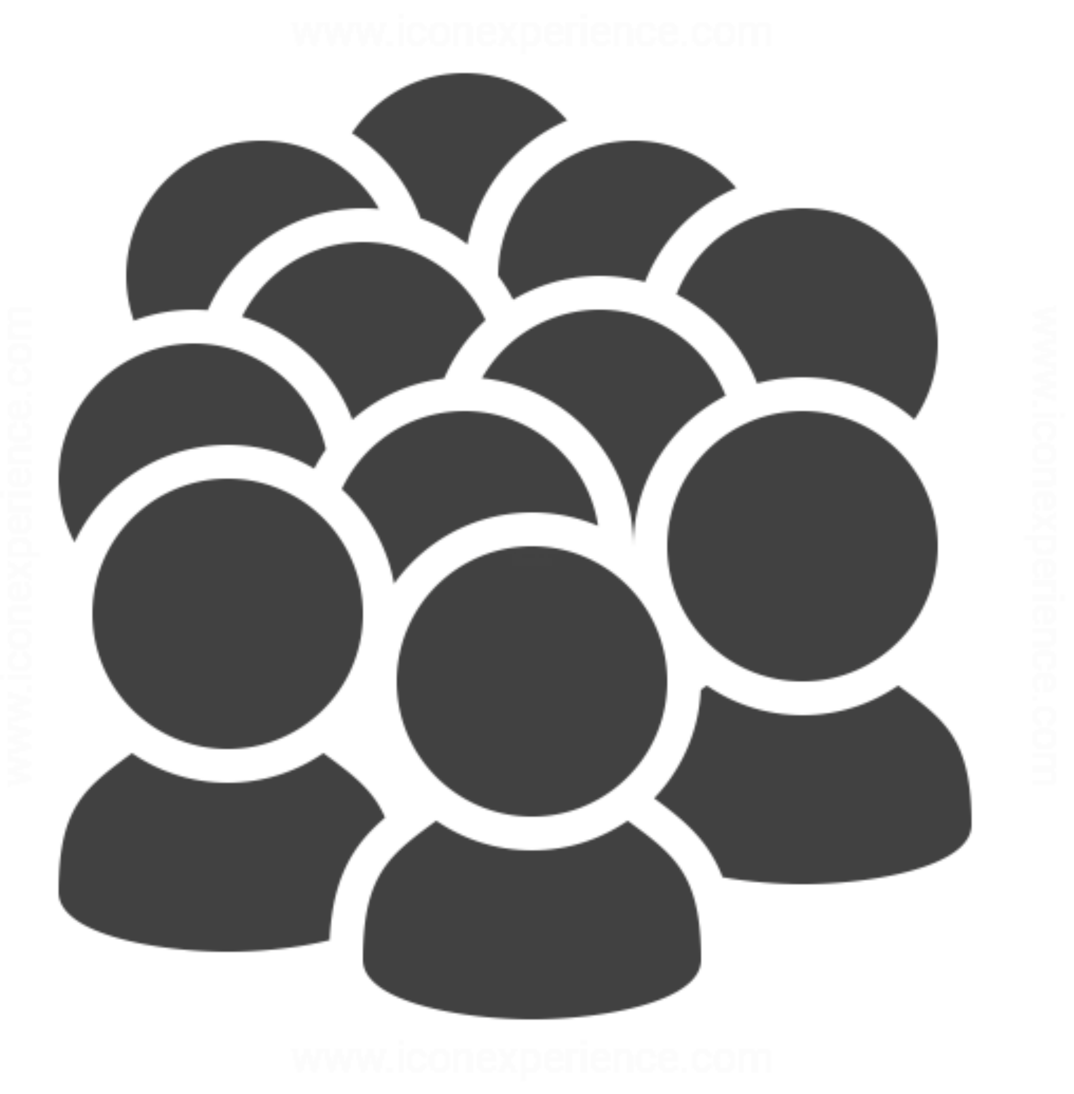
Key takeaway: Richer insights

Case study 2

Customer demographics

Business (Acme Inc.)

Business (Acme Inc.)



Existing customers

Tweets
2,932
Followers
2,282
Lists
1
Follow
Acme Inc.
@AcmeInc
Proudly creating high quality furniture since 1964
Charing Cross, London
acme.com
Joined December 2012
Tweets
Tweets & replies
Media
Acme Inc. @AcmeInc · Aug 30
Our new dining tables are out now! bit.ly/2C12xWZ
Who to follow · Refresh · View all
Miliboo.com @miliboo
Follow
rioMoros @rioMoros
Follow
Cinna @CinnaTM

Suhail @Suhail · Aug 29

One of my early mistakes in the 1st two years of building a co was building new products because users seemed happy. That lack of focus put us back a year. It's usually a mistake to expand to a new mkt because the product is "done" for the primary one but your mkt share is < 1%.

15 124 804

Suhail @Suhail · Aug 29

People underestimate how much grit it takes for founders to steadily build the most monotonous features in order to make an okay product great. Especially difficult when there are so many more interesting/intellectually challenging v1 ideas to distract you.

19 243 1.2K

Suhail @Suhail · Aug 29

The worst thing about the Internet & mobile phones, for me lately, is that I'm incapable of focusing enough to read a book for longer than 15 min unless I am going to bed. The cycle to reverse this has been extremely painful.

64 175 1.0K

Suhail @Suhail · Aug 28

Is anyone aware of research or papers discussing how to dramatically reduce Internet latency to < 5ms?

22 5 74

Suhail @Suhail · Aug 27

The most complicated problems are made less overwhelming by breaking them into discrete sub-problems, assigning teams with a clear goal, & having patience. If you don't have the resources to solve all the sub-problems, partner & focus on a narrower set.

6 72 317

Suhail @Suhail · Aug 29

One of my early mistakes in the 1st two years of building a co was building new products because users seemed happy. That lack of focus put us back a year. It's usually a mistake to expand to a new mkt because the product is "done" for the primary one but your mkt share is < 1%.

15 124 804

Suhail @Suhail · Aug 29

People underestimate how much grit it takes for founders to steadily build the most monotonous features in order to make an okay product great. Especially difficult when there are so many more interesting/intellectually challenging v1 ideas to distract you.

19 243 1.2K

Suhail @Suhail · Aug 29

The worst thing about the Internet & mobile phones, for me lately, is that I'm incapable of focusing enough to read a book for longer than 15 min unless I am going to bed. The cycle to reverse this has been extremely painful.

64 175 1.0K

Suhail @Suhail · Aug 28

Is anyone aware of research or papers discussing how to dramatically reduce Internet latency to < 5ms?

22 5 74

Suhail @Suhail · Aug 27

The most complicated problems are made less overwhelming by breaking them into discrete sub-problems, assigning teams with a clear goal, & having patience. If you don't have the resources to solve all the sub-problems, partner & focus on a narrower set.

6 72 317

Ignoring profile pic, name, can we guess age & gender?

PAN

(Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection)

82%

Gender

82%

Gender

52%

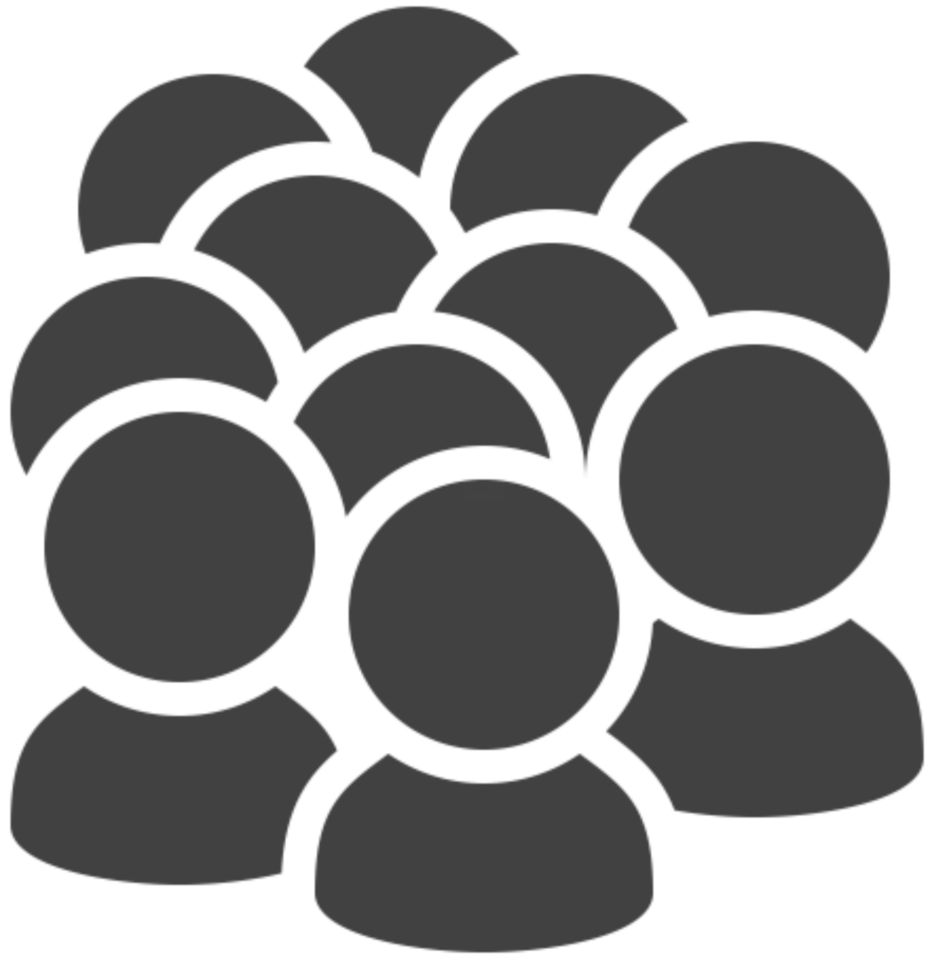
Gender & Age

18-24, 25-34, 35-49, 50-64, 65+

85%

“Is the author male?”

Business (Acme Inc.)



Existing customers

Business (Acme Inc.)

Existing customers


Sign Up
Please fill in this form to create an account!
First Name
Last Name
Email
Password
Confirm Password
I accept the Terms of Use & Privacy Policy.
Sign Up
Already have an account? Login here.


Sign Up
Please fill in this form to create an account!
First Name
Last Name
Email
Password
Confirm Password
Male
Female
I accept the Terms of Use & Privacy Policy.
Sign Up
Already have an account? Login here.

Case study 2

Key takeaway: Post-hoc analysis

Case study 2

Key takeaway: Post-hoc analysis

Case study 3

Statin decline study

18 million

1/3

18 million

1/3

200 million

1/35

(1/10 in UK)

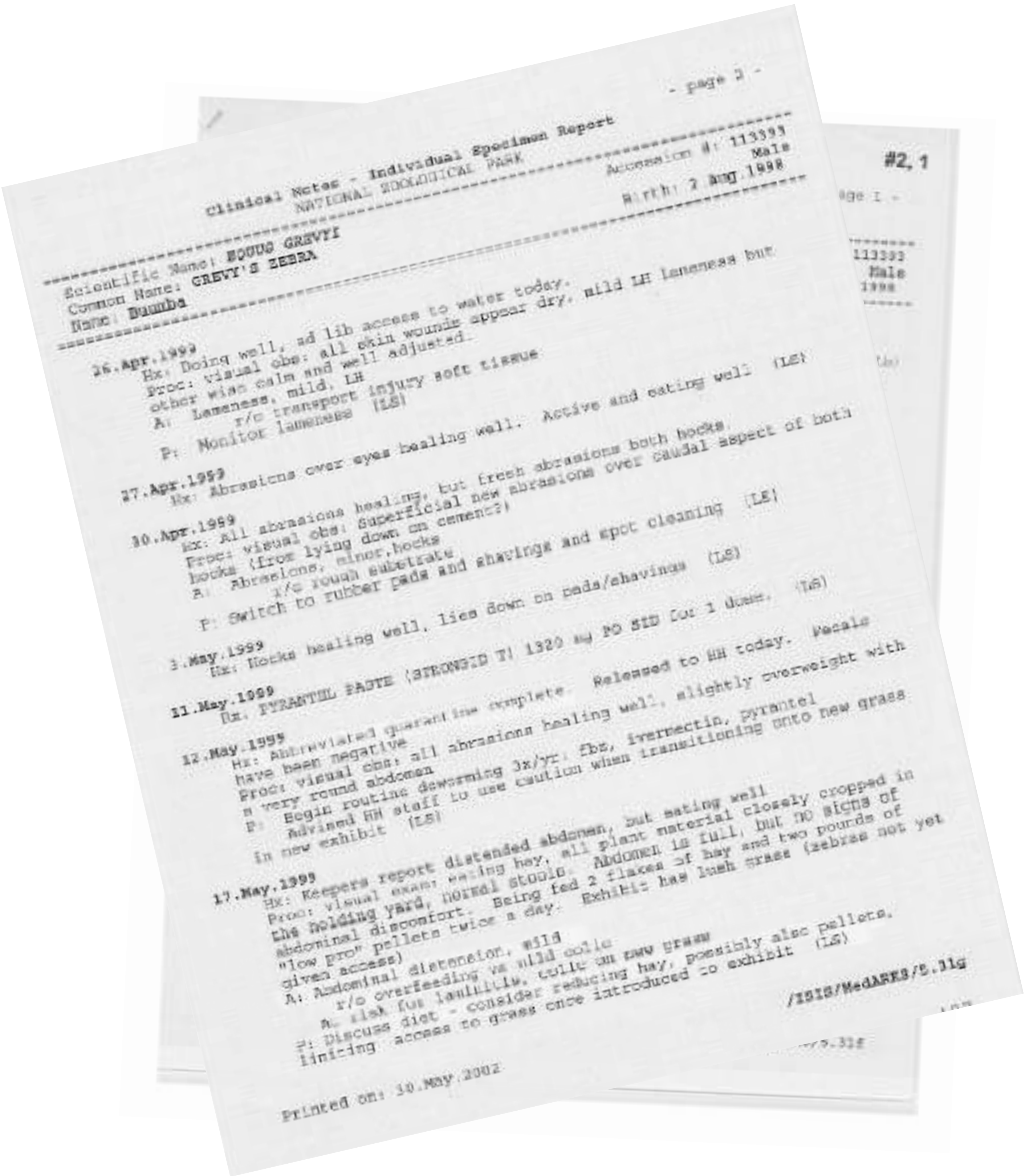
HARVARD
MEDICAL SCHOOL

Hospital name	Statin decline rate
Cedars Sinai	3.1%
Massachusetts General Hospital	7.4%
Walter Reed National Military Medical Center	4.2%
New York – Presbyterian Hospital	2.1%
The Academy and College of Philadelphia	11.4%
The Pennsylvania Hospital	8.6%

Hospital name	Statin decline rate
Cedars Sinai	3.1%
Massachusetts General Hospital	7.4%
Walter Reed National Military Medical Center	4.2%
New York – Presbyterian Hospital	2.1%
The Academy and College of Philadelphia	11.4%
The Pennsylvania Hospital	8.6%

Hospital name	St[illegible] decline rate
Cedars Sinai	[illegible].1%
Massachusetts General [illegible]	7.4%
Walter Reed National Militar[illegible]er	4.2%
New York – Presbyte[illegible]	2.1%
The Academy ar[illegible] Philadelp[illegible]	11.4%
The Pennsylvania Ho[illegible]ital	8.6%



- page 2 -

Clinical Notes - Individual Specimen Report
NATIONAL ZOOLOGICAL PARK

Accession #: 113393
Male
Birth: 2 Aug 1998

Scientific Name: EQUUS GREVYI
Common Name: GREVY'S ZEBRA
Name: Duumba

26.Apr.1999
Hx: Doing well, ad lib access to water today.
Proc: visual obs: all skin wounds appear dry, mild LH lameness but other wise calm and well adjusted.
A: Lameness, mild, LH
r/o transport injury soft tissue
P: Monitor lameness (LS)

27.Apr.1999
Hx: Abrasions over eyes healing well. Active and eating well (LS)

30.Apr.1999
Hx: All abrasions healing, but fresh abrasions both hocks.
Proc: visual obs: Superficial new abrasions over caudal aspect of both hocks (from lying down on cement?)
A: Abrasions, minor, hocks
r/o rough substrate
P: Switch to rubber pads and shavings and spot cleaning (LS)

3.May.1999
Hx: Hocks healing well, lies down on pads/shavings (LS)

11.May.1999
Hx: PYRANTEL PASTE (STRONGID T) 1320 mg PO SID for 1 dose. (LS)

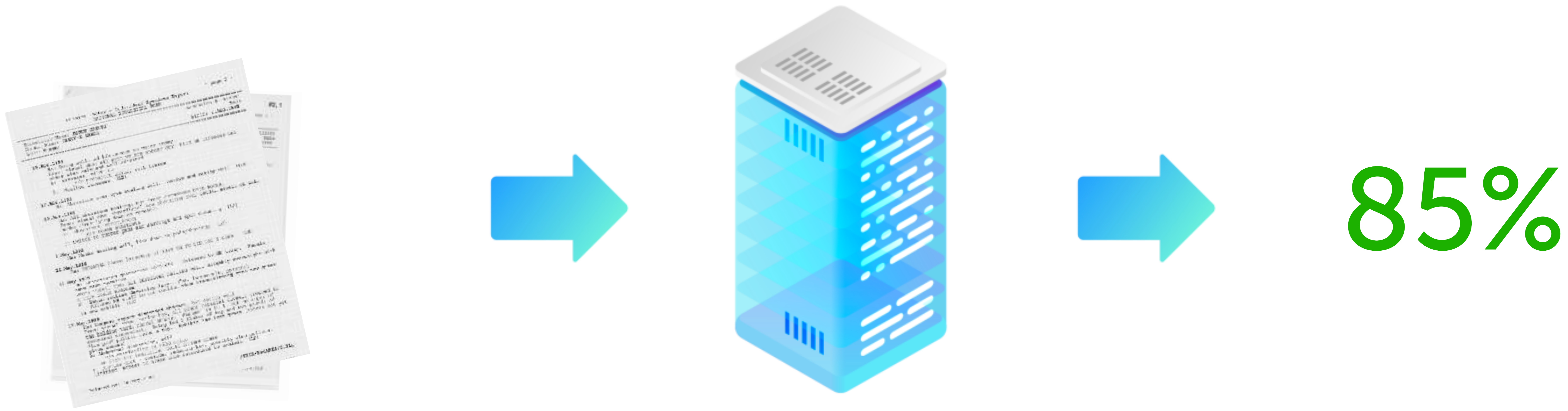
12.May.1999
Hx: Abbreviated quarantine complete. Released to HH today. Fecals have been negative
Proc: visual obs: all abrasions healing well, slightly overweight with a very round abdomen
P: Begin routine deworming 3x/yr: fbz, ivermectin, pyrantel
Advised HH staff to use caution when transitioning onto new grass in new exhibit (LS)

17.May.1999
Hx: Keepers report distended abdomen, but eating well
Proc: visual exam: eating hay, all plant material closely cropped in the holding yard, normal stools. Abdomen is full, but NO signs of abdominal discomfort. Being fed 2 flakes of hay and two pounds of "low pro" pellets twice a day. Exhibit has lush grass (zebras not yet given access)
A: Abdominal distension, mild
r/o overfeeding vs mild colic
A: risk for laminitis, colic on new grass
P: Discuss diet - consider reducing hay, possibly also pellets, limiting access to grass once introduced to exhibit (LS)

/ISIS/MedARKS/5.31g

Printed on: 10.May.2002

#2, 1



"Were statins recommended but declined?"

~90% accuracy

(precision/recall)

8,800 patients

evaluated

Case study 3

Key takeaway: cheaper/easier

+

Disadvantages

Advantages

+

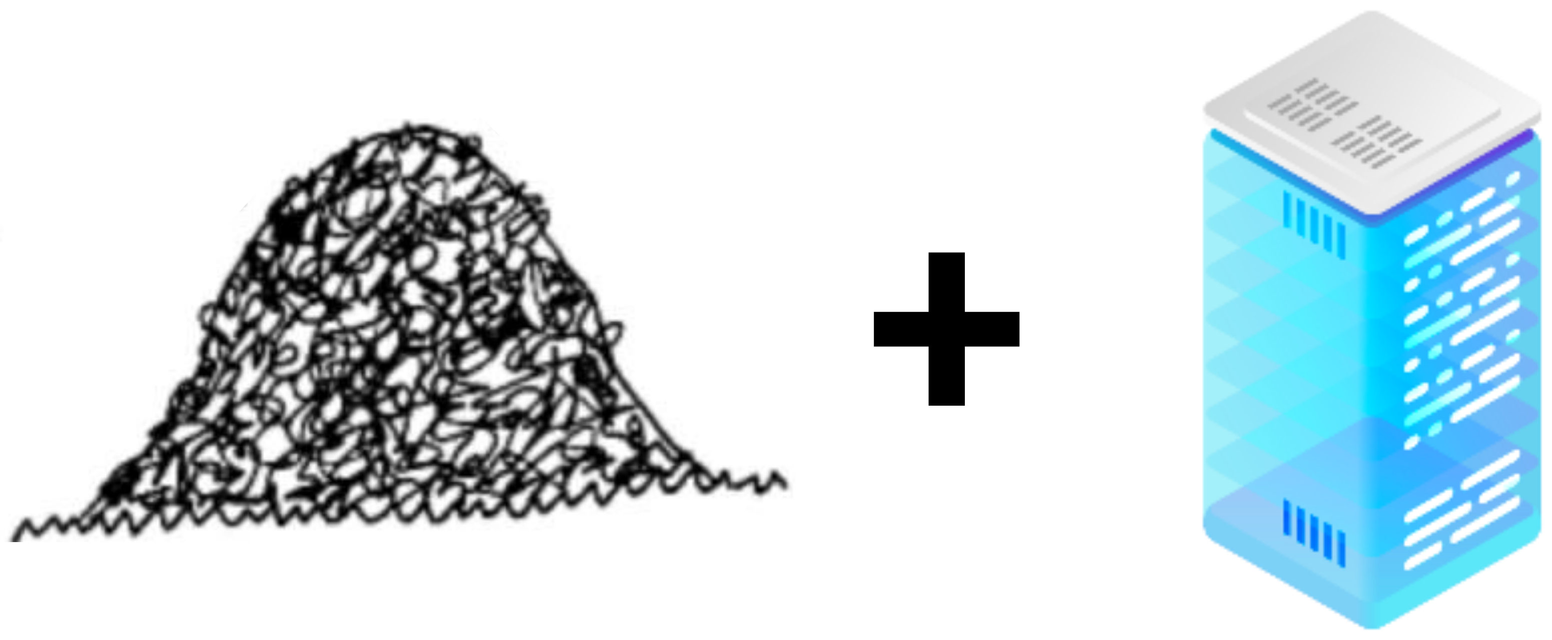
Disadvantages

Soft rules

Noisy

Unclear result

Advantages



Disadvantages

Soft rules

Noisy

Unclear result

Advantages

Richer/deeper

Post hoc

Cheaper/easier

Agenda

Origins of language models

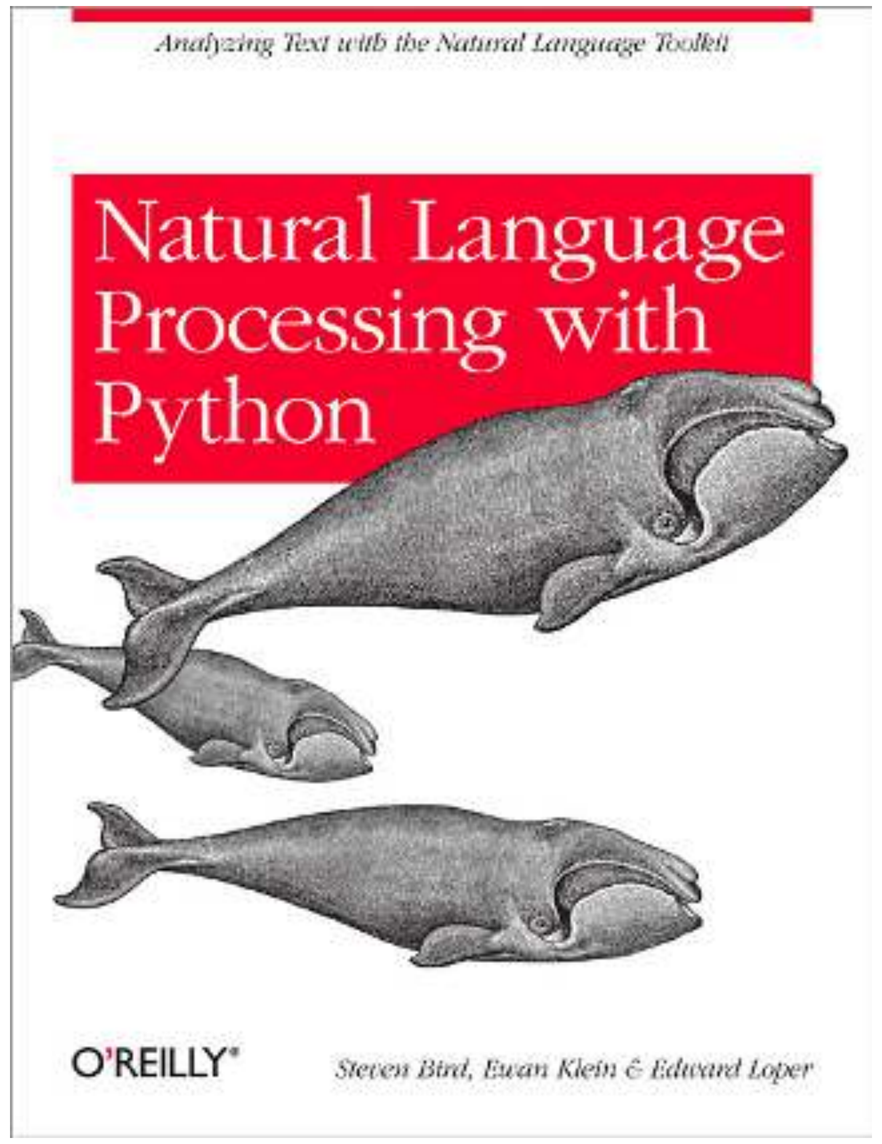
What is unstructured data?

- Some case studies

Types of language models

- Count based (bag of words, *n*-grams)
- Continuous space
- Bonus: the class of 2018

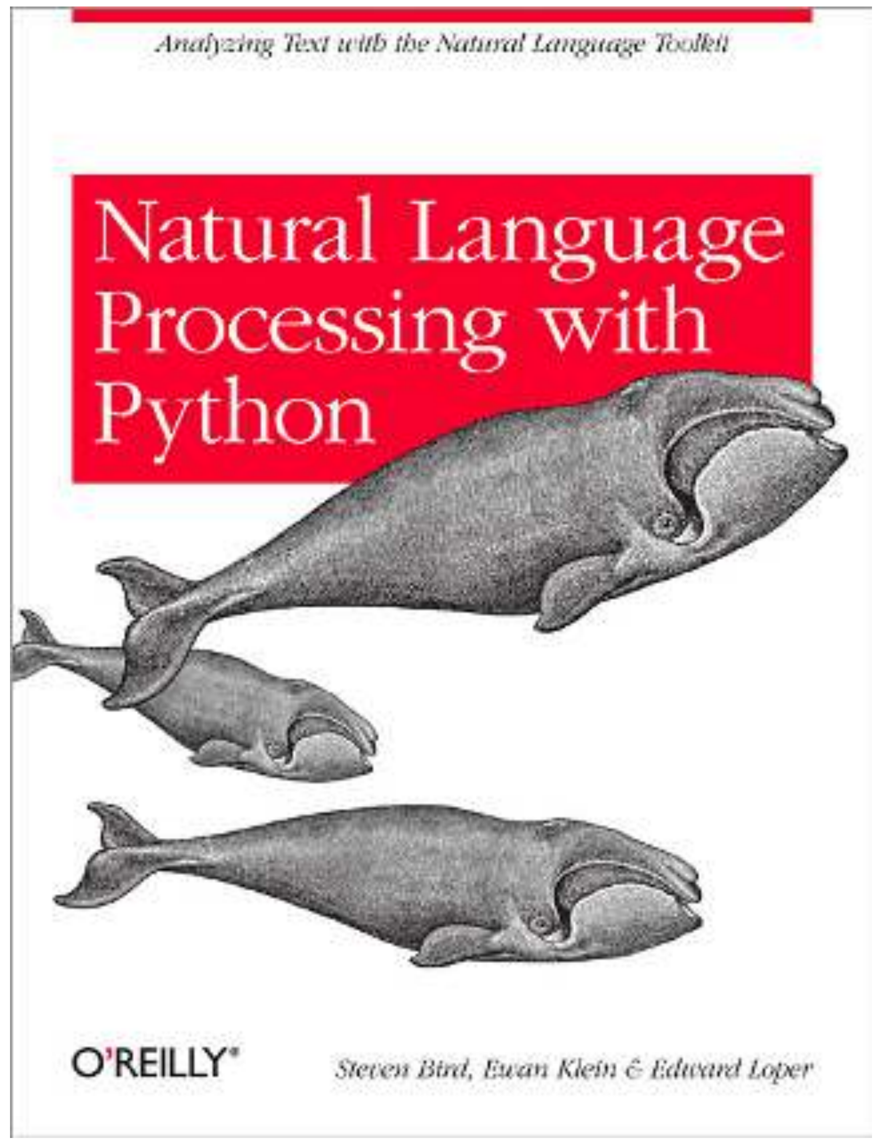
Wrap-up and questions



NLTK

Natural language toolkit

Scikit Learn

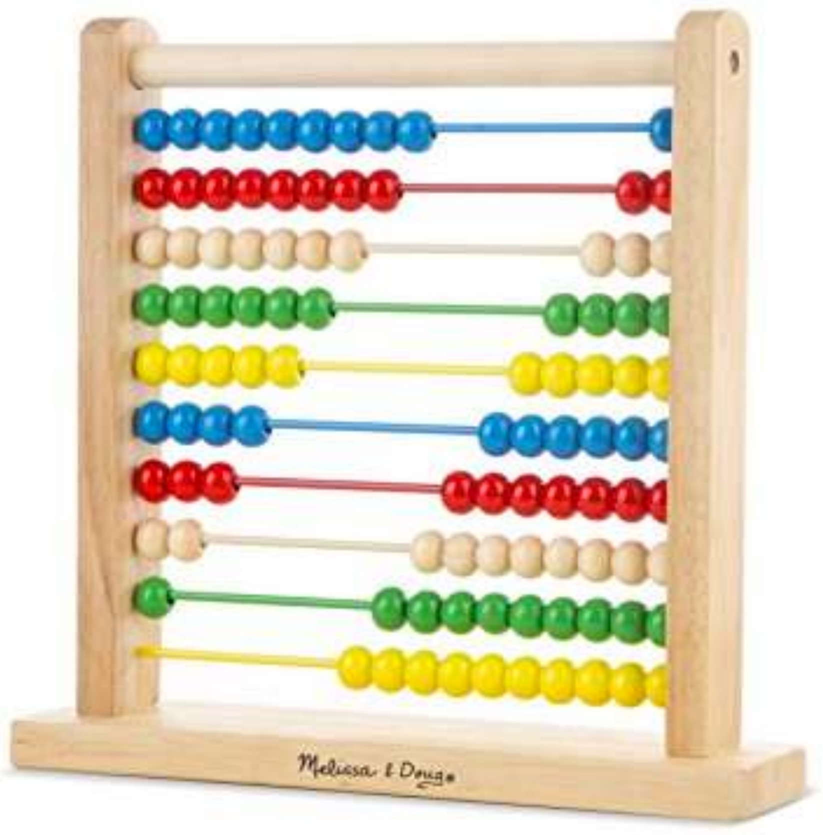


NLTK

Natural language toolkit

Scikit Learn

(Has prebaked models)



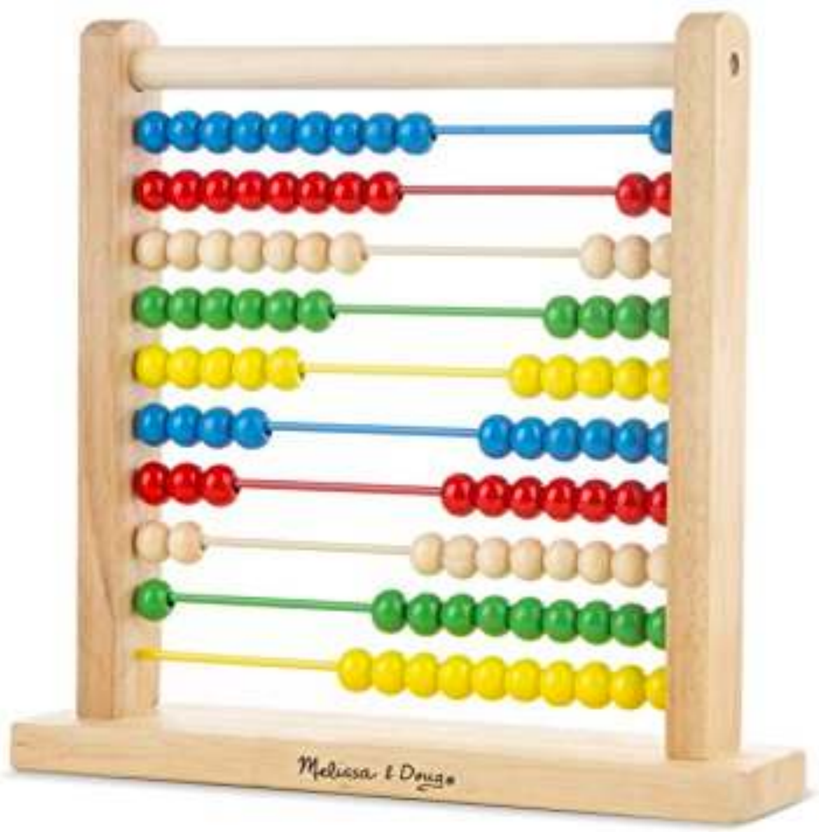
Count based

AKA statistical

1980, 1990s

Very fast

Decent performance (when tuned)



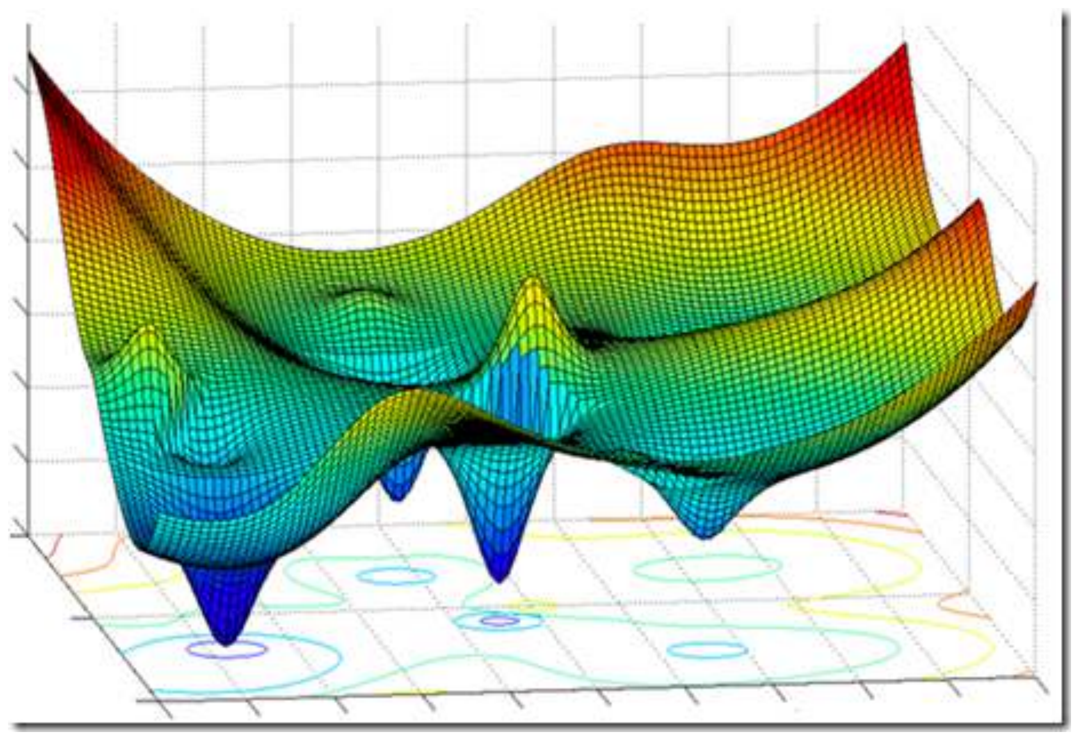
Count based

AKA statistical

1980, 1990s

Very fast

Decent performance (when tuned)



Continuous space

AKA neural, neuroprobabilistic

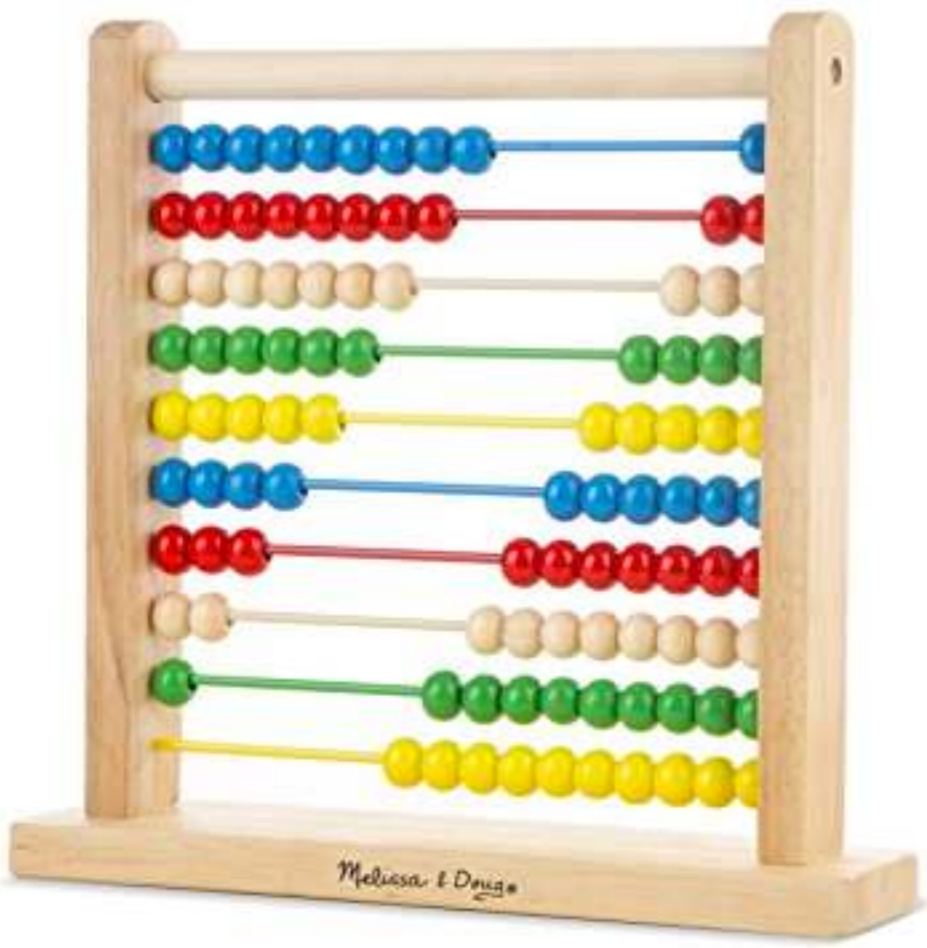
2000s, 2010s

Slower, more expensive

Typically used with neural nets

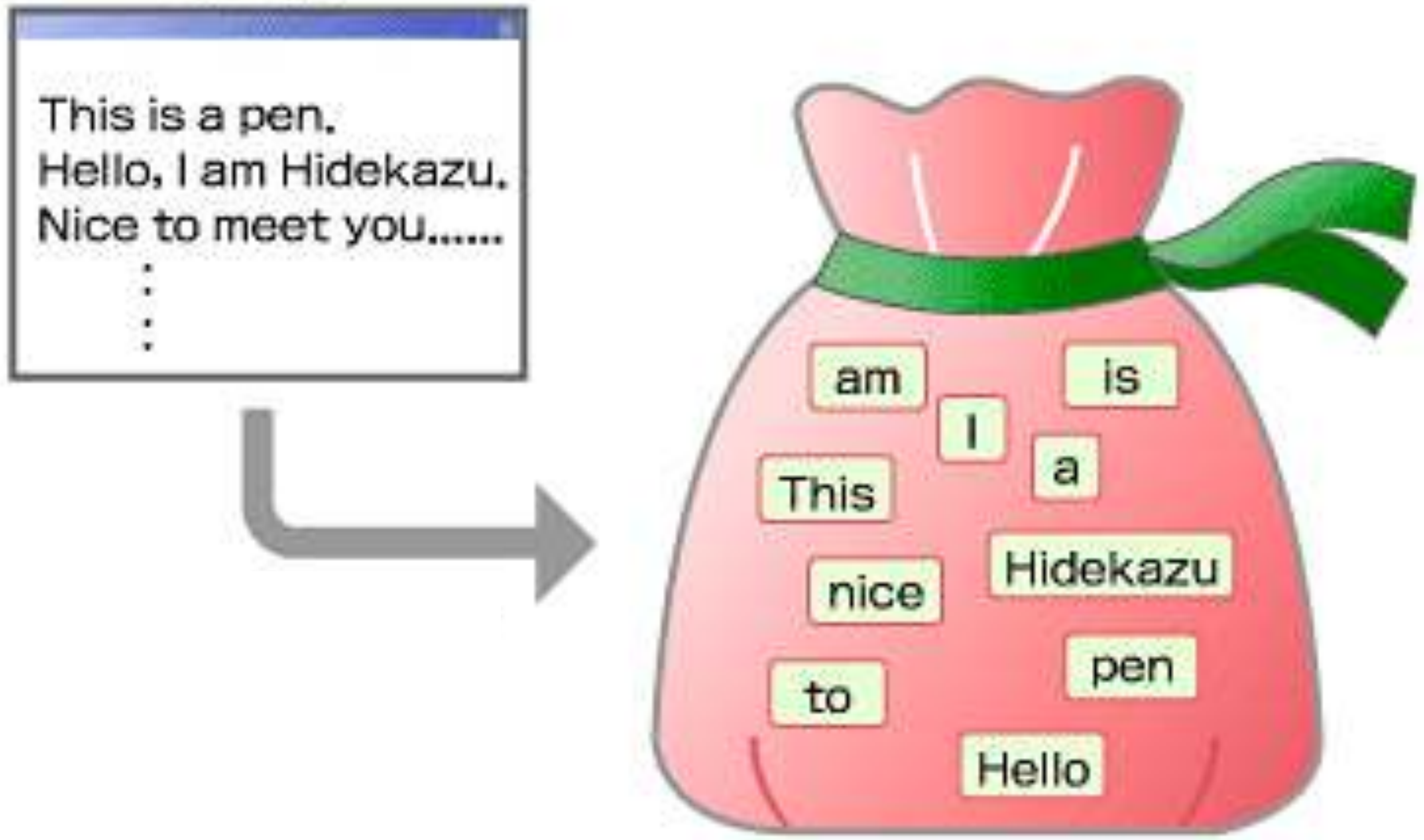
State-of-the-art performance

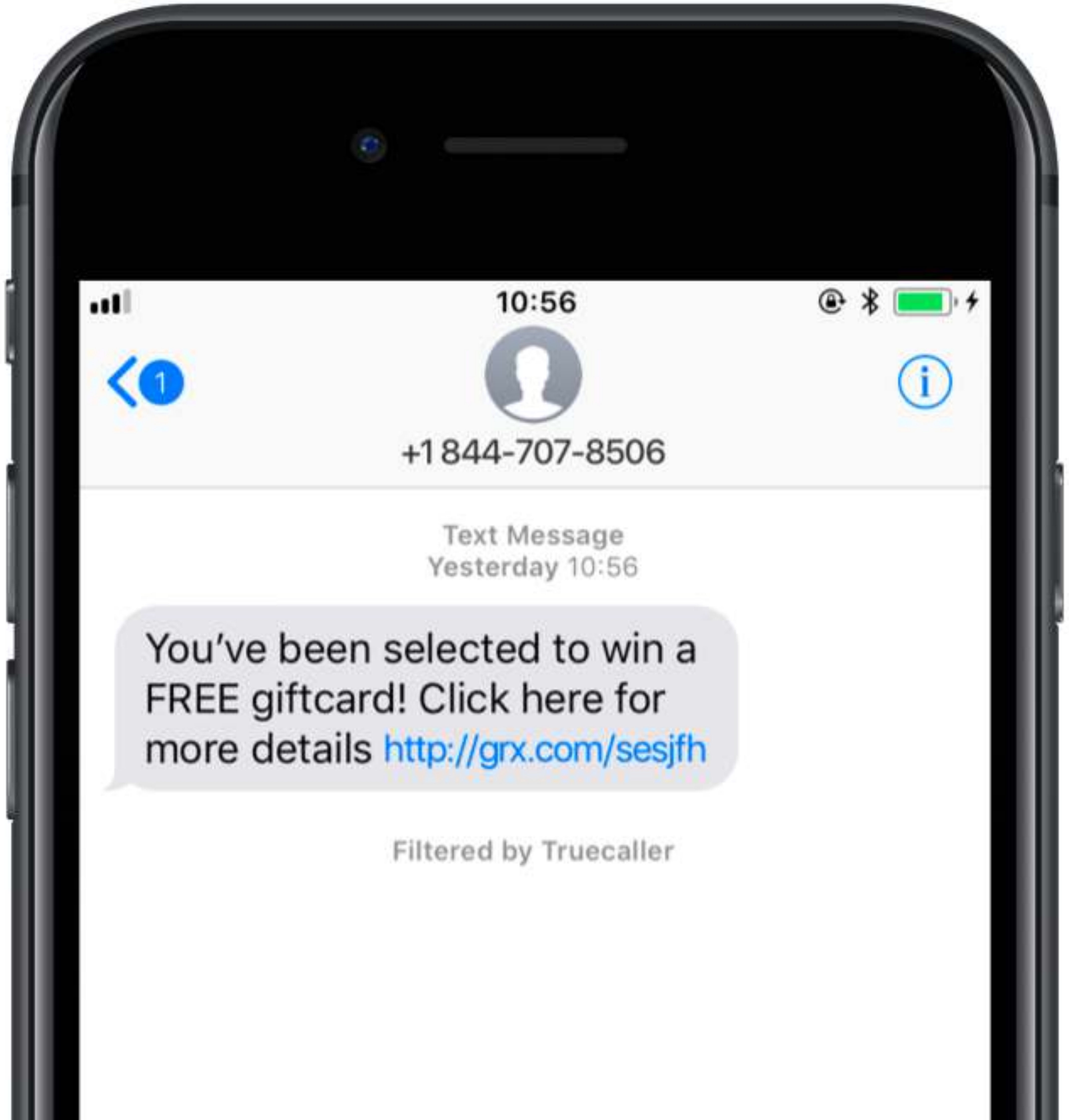
Count-based



1. Bag of words
2. *n*–gram

Bag of words

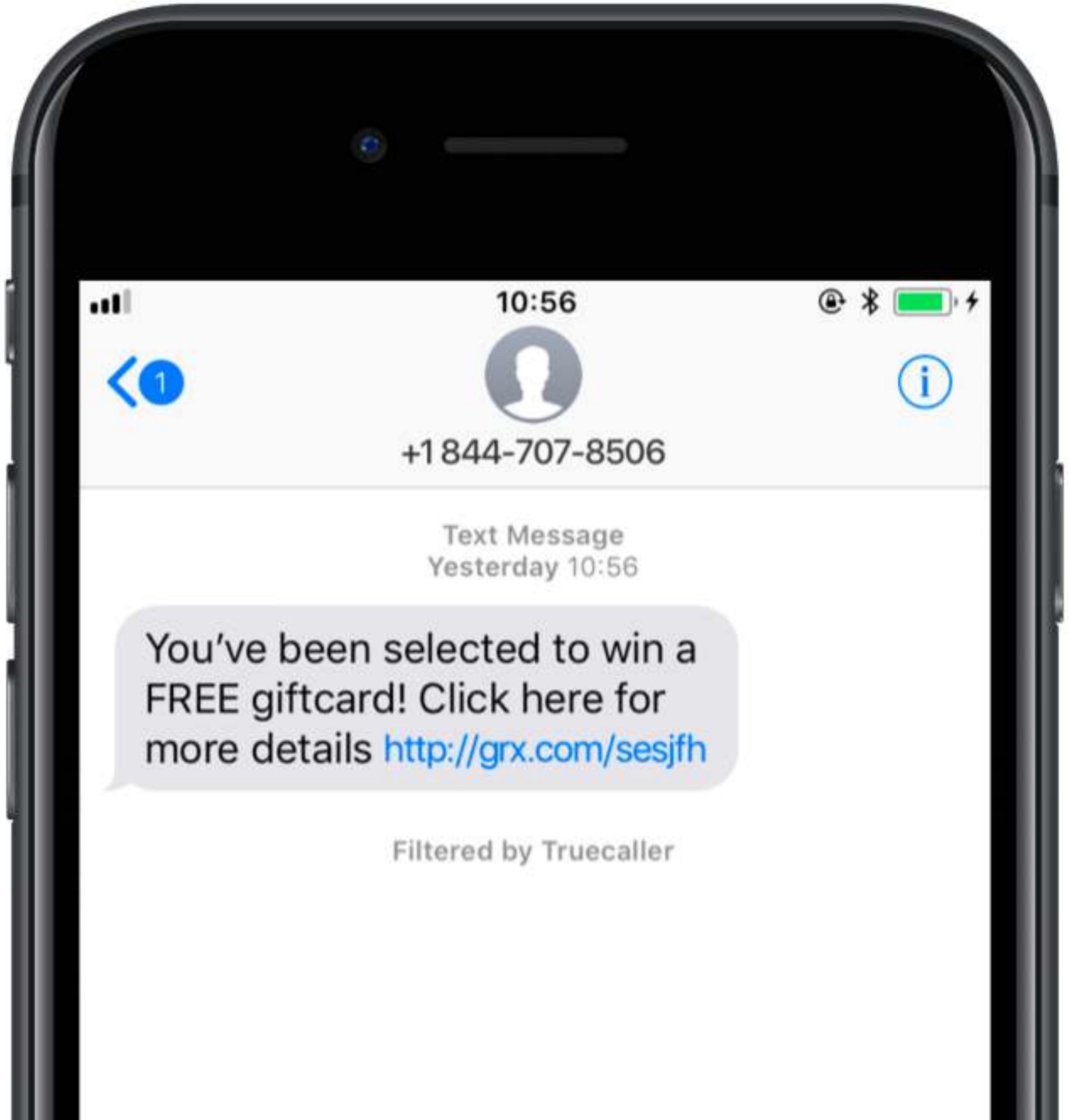

This is a pen.
Hello, I am Hidekazu.
Nice to meet you......
am
is
I
a
This
nice
Hidekazu
pen
to
Hello



SMS Spam Collection Data Set

University of Sao Carlos

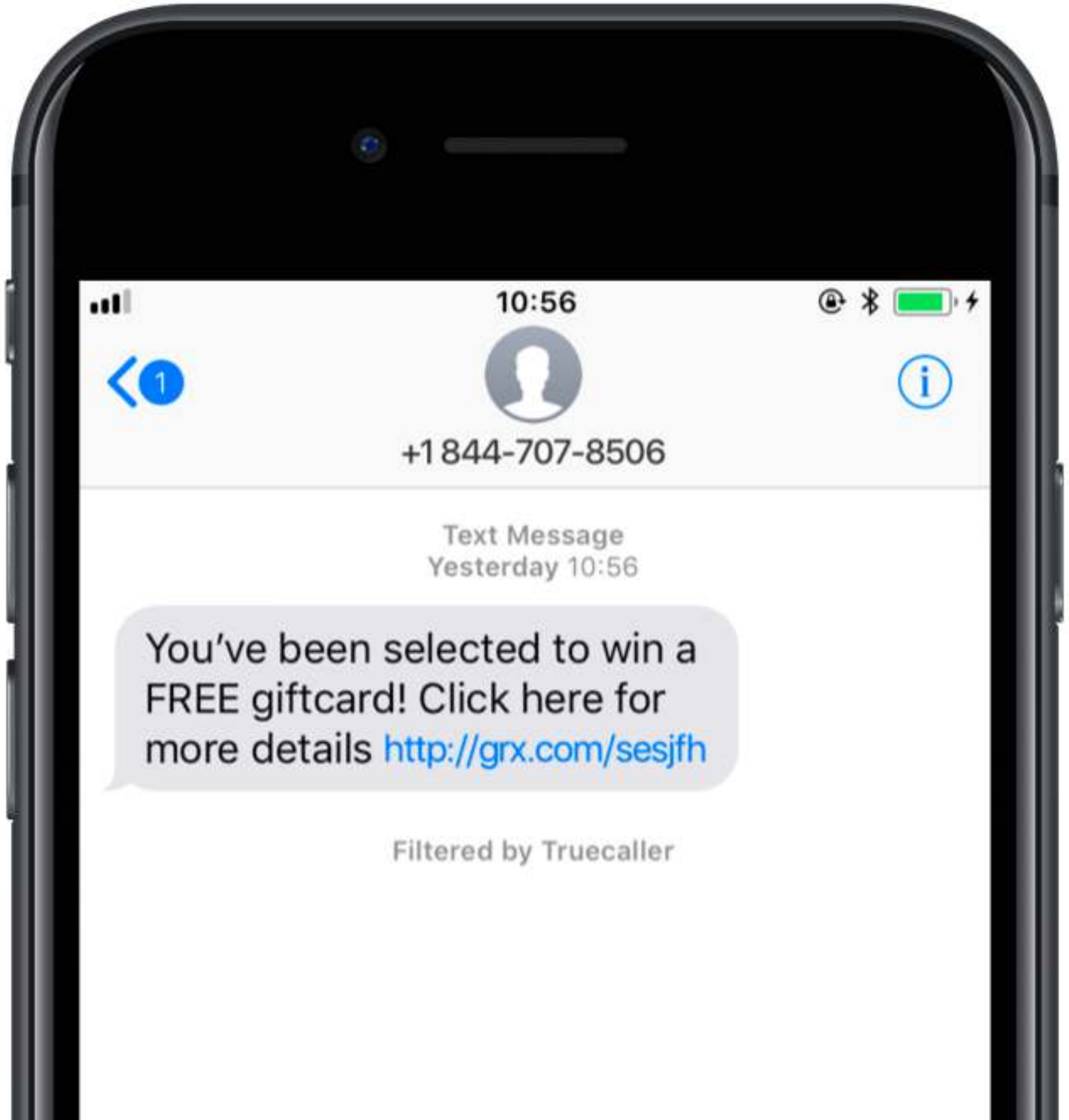
~5,500 SMS messages, categorized into ham and spam



Preliminaries:

We'll need a train and test set (corpus).

80-20 split is fine.



1. Clean
2. Tokenize
3. Remove stopwords
4. Stem
5. Build frequency matrix
6. Classify

1. clean

URGENT! Your Mobile No 07808726822 was awarded a L2,000 Bonus Caller Prize on 02/09/03! This is our 2nd attempt to contact YOU! Call 0871-872-9758 BOX95QU

1. clean

URGENT! Your Mobile No 07808726822 was awarded a L2,000 Bonus Caller Prize on 02/09/03! This is our 2nd attempt to contact YOU! Call 0871-872-9758 BOX95QU

urgent your mobile no was awarded a l bonus caller prize on this is our nd attempt to contact you call box qu

1. clean

URGENT! Your Mobile No 07808726822 was awarded a L2,000 Bonus Caller Prize on 02/09/03! This is our 2nd attempt to contact YOU! Call 0871-872-9758 BOX95QU

urgent your mobile no was awarded a l bonus caller prize on this is our nd attempt to contact you call box qu

```
message = re.sub('[^A-Za-z]', ' ', message)
message = message.lower()
```

2. tokenize

urgent your mobile no was awarded a l bonus caller prize on this is our nd attempt to contact you call box qu

[a, attempt, awarded, bonus, box, call, caller, contact, is, l, mobile, nd, no, on, our, prize, qu, this, to, urgent, was, you, your]

2. tokenize

Simple:

```
tokens = message.split(' ')
```

Robust:

```
from nltk.tokenize import word_tokenize
tokens = word_tokenize(message)
```

3. stopwords

[a, attempt, awarded, bonus, box, call, caller, contact, is, l, mobile, nd, no, on, our, prize, qu, this, to, urgent, was, you, your]

−

[i, me, my, myself, we, our, ours, ourselves, you, your, yours, yourself, yourselves, he, him, his, himself, she, her, hers, herself, it, its, itself, they, them, these, those, am, is, are, was, ...]

=

[attempt, awarded, bonus, box, call, caller, contact, l, mobile, nd, prize, qu, urgent]

3. stopwords

```
from nltk.corpus import stopwords
tokens = [t for t in tokens if not t in stopwords]
```

4. stemming

win

winner

winners

won

winning

winnings

4. stemming

win

winner

winners → **win**

won

winning

winnings

4. stemming

```
from nltk.stem.porter import PorterStemmer

stemmer = PorterStemmer()
tokens = [stemmer.stem(t) for t in tokens]
```

[attempt, award~~ed~~, bonu~~s~~, box, call, caller, contact, l, mobil~~e~~, nd, prize, qu, urgent]

5. matrix

	are	call	from	hello	home	how	me	money	now	tomorrow	win	you
0	1	0	0	1	0	1	0	0	0	0	0	1
1	0	0	1	0	1	0	0	1	0	0	2	0
2	0	1	0	0	0	0	1	0	1	0	0	0
3	0	1	0	2	0	0	0	0	0	1	0	1

5. matrix

```
from sklearn.feature_extraction.text import CountVectorizer
count_vector = CountVectorizer()

count_vector.fit(messages)
```

5. matrix

```
count_vector.get_feature_names()
```

	are	call	from	hello	home	how	me	money	now	tomorrow	win	you
0	1	0	0	1	0	1	0	0	0	0	0	1
1	0	0	1	0	1	0	0	1	0	0	2	0
2	0	1	0	0	0	0	1	0	1	0	0	0
3	0	1	0	2	0	0	0	0	0	1	0	1

5. matrix

```
train = count_vector.transform(messages).toarray()
```

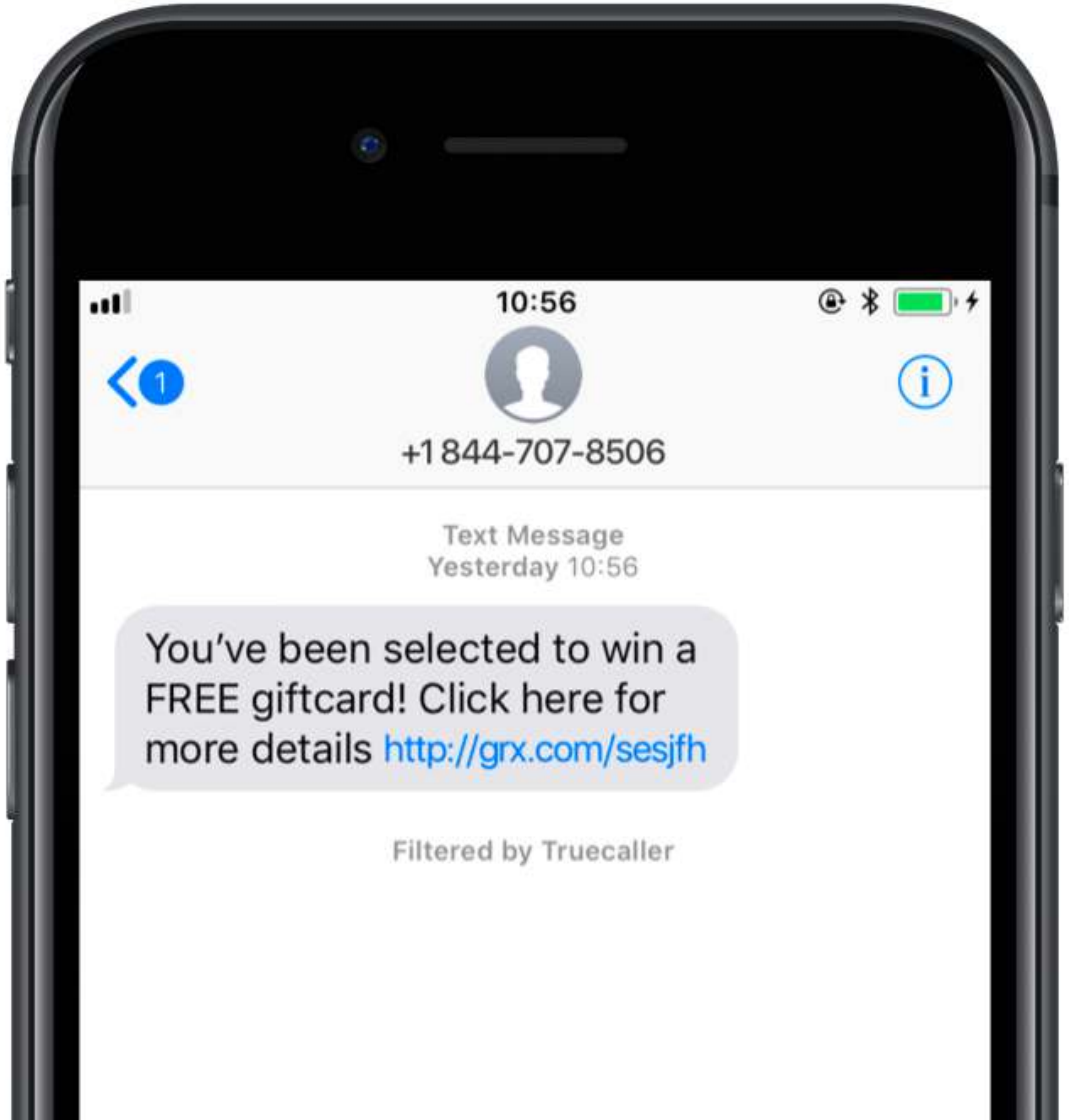
	are	call	from	hello	home	how	me	money	now	tomorrow	win	you
0	1	0	0	1	0	1	0	0	0	0	0	1
1	0	0	1	0	1	0	0	1	0	0	2	0
2	0	1	0	0	0	0	1	0	1	0	0	0
3	0	1	0	2	0	0	0	0	0	1	0	1

6. classify

```
from sklearn.naive_bayes import MultinomialNB

naive_bayes = MultinomialNB()
naive_bayes.fit(train, y_train)
```

6. classify

```
from sklearn.naive_bayes import MultinomialNB

naive_bayes = MultinomialNB()
naive_bayes.fit(train, y_train)

naive_bayes.predict(test)
```



90%+ accuracy

(Precision and recall)

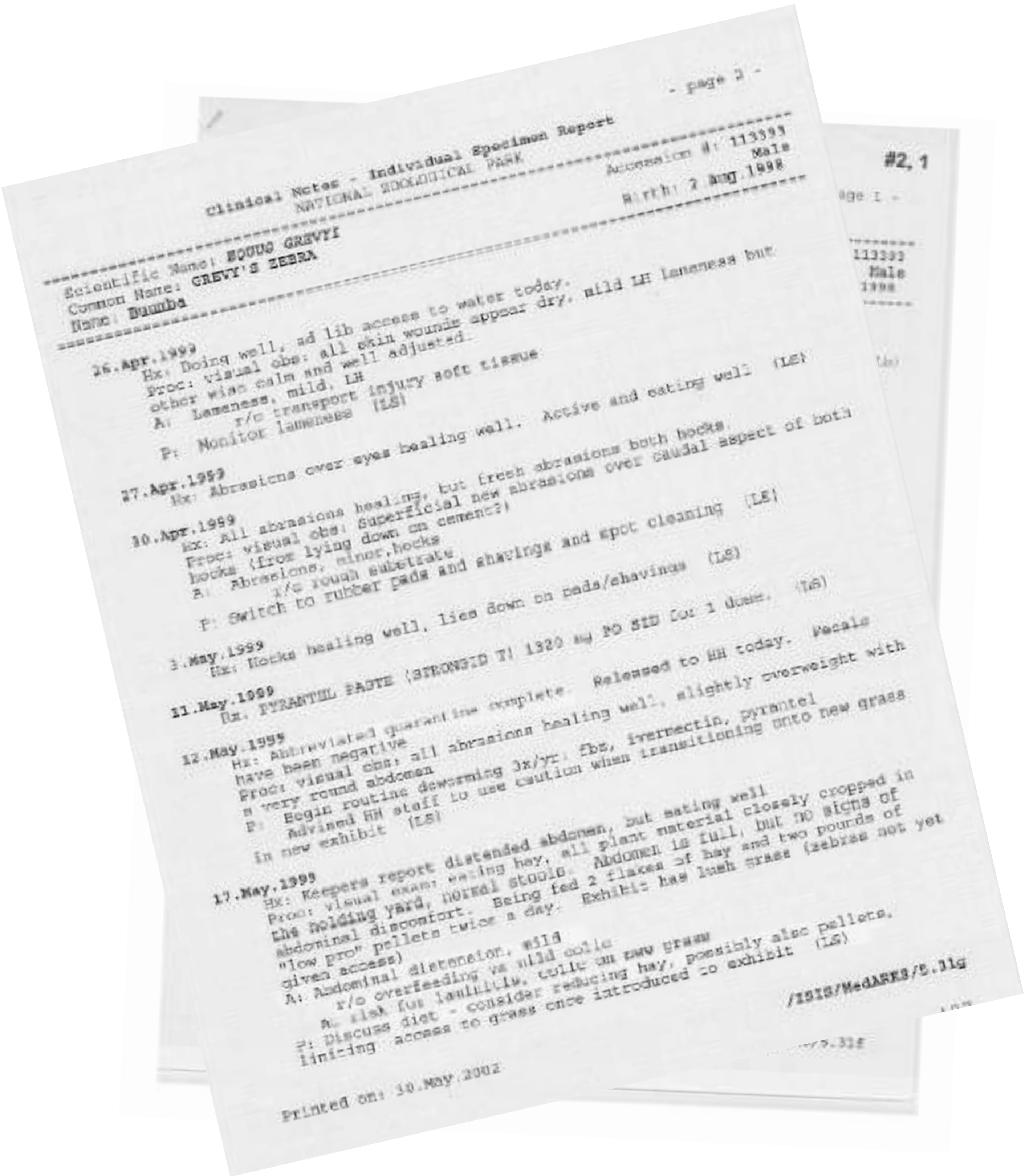
Next week is crazy & on holiday until Wednesday. Are you free at 1?

Free holiday, offer on until next Wednesday! Are you crazy?? 1 WEEK HOLIDAY IS FREE!

Next week is crazy & on holiday until Wednesday. Are you free at 1?

Free holiday, offer on until next Wednesday! Are you crazy?? 1 WEEK HOLIDAY IS FREE!



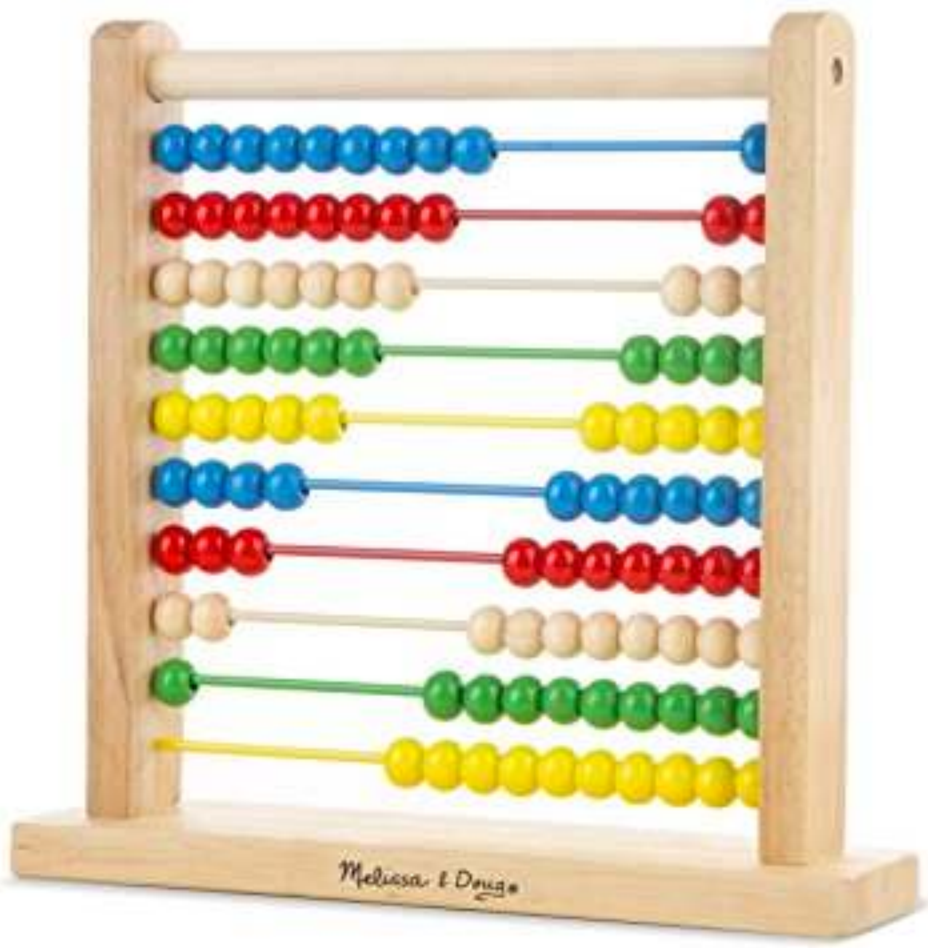
“statin”

“patient”

“not”

“soreness”

Count-based



1. Bag of words
2. *n*–gram

Next week is crazy &
on **holiday** until Wednesday.
Are you free at 1?

Congratulations you've
won a free **holiday**, offer
on until next Wednesday!

Next week is crazy &
on **holiday** until Wednesday.
Are you free at 1?

Congratulations you've
won a free **holiday**, offer
on until next Wednesday!

Is holiday spammy
or not spammy?

Next week is crazy &
on **holiday** until Wednesday.
Are you free at 1?

Congratulations you've
won a free **holiday**, offer
on until next Wednesday!

Is holiday spammy
or not spammy,
given the context?

Next week is crazy &
on **holiday** until Wednesday.
Are you free at 1?

Congratulations you’ve
won a free **holiday**, offer
on until next Wednesday!

Next week is crazy &
on **holiday** until Wednesday.
Are you free at 1?

Congratulations you've
won a free **holiday**, offer
on until next Wednesday!

Expensive to compute

Next week is crazy &
on **holiday** until Wednesday.
Are you free at 1?

Congratulations you’ve
won a free **holiday**, offer
on until next Wednesday!

Expensive to compute

Prone to overfitting

Next week is crazy &
on **holiday** until Wednesday.
Are you free at 1?

Congratulations you've
won a free **holiday**, offer
on until next Wednesday!

Instead consider limited context

Previous *n* words

n = ...

unigram (bag of words)	won a free **holiday**, offer
bigram	won a free **holiday**, offer
trigram	won a free **holiday**, offer
4-gram	won a free **holiday**, offer

TOY TRAINING CORPUS

I am Sam

Sam I am

I do not like green eggs and ham

TOY TEST SENTENCE

I am Sam I do

Preprocessing

1. Clean
2. Tokenize
3. Remove stopwords
4. Stem

Preprocessing

1. Clean
2. Tokenize
3. ~~Remove stopwords~~
4. Stem

TOY TRAINING CORPUS

i am sam

sam i am

i do not like green egg and ham

TOY TEST SENTENCE

i am sam i do

TOY TRAINING CORPUS

i am sam

sam i am

i do not like green egg and ham

TOY TEST SENTENCE

i am sam i do

$$p_1 = \frac{\text{how many times “am” follows “i”}}{\text{how many times “i” appears}}$$

TOY TRAINING CORPUS

i am sam

sam i am

i do not like green egg and ham

TOY TEST SENTENCE

i am sam i do

$$p_1 = \frac{\text{how many times “am” follows “I”}}{3}$$

TOY TRAINING CORPUS

i am sam

sam i am

i do not like green egg and ham

TOY TEST SENTENCE

i am sam i do

$$p_1 = \frac{2}{3}$$

TOY TRAINING CORPUS

i am sam

sam i am

i do not like green egg and ham

TOY TEST SENTENCE

i am sam i do

$$p_1 = \frac{2}{3} \qquad p_2 = \frac{1}{2}$$

TOY TRAINING CORPUS

i am sam

sam i am

i do not like green egg and ham

TOY TEST SENTENCE

i am sam i do

$$p_1 = \frac{2}{3}, p_2 = \frac{1}{2}, p_3 = \frac{1}{2}, p_4 = \frac{1}{3}$$

TOY TRAINING CORPUS

i am sam

sam i am

i do not like green egg and ham

TOY TEST SENTENCE

i am sam i do

$$p_1 \times p_2 \times p_3 \times p_4 \approx 0.05$$

TOY TRAINING CORPUS

i am sam

sam i am

i do not like green egg and ham

TOY TEST SENTENCE

i am sam i am sam i do

$$p_1 \times p_2 \times p_3 \times p_4 \times p_5 \times p_6 \times p_7 \approx 0.009$$

TOY TRAINING CORPUS

i am sam

sam i am

i do not like green egg and ham

"Discriminates" against long sentences

TOY TEST SENTENCE

i am sam i am sam i do

$$p_1 \times p_2 \times p_3 \times p_4 \times p_5 \times p_6 \times p_7 \approx 0.009$$

TOY TRAINING CORPUS

i am sam

sam i am

i do not like green egg and ham

TOY TEST SENTENCE

i am sam i am sam i do

Small numbers, pain to work with

$$p_1 \times p_2 \times p_3 \times p_4 \times p_5 \times p_6 \times p_7 \approx 0.009$$

$$\sqrt[n]{p_1 \times p_2 \times p_3 \times \ldots}$$

1. Take the *n*th root to normalize long and short sentences

$$\frac{1}{\sqrt[n]{p_1 \times p_2 \times p_3 \times \ldots}}$$

1. Take the *n*th root to normalize long and short sentences

2. Take the reciprocal to avoid small numbers

$$\frac{1}{\sqrt[n]{p_1 \times p_2 \times p_3 \times \dots}}$$

1. Take the *n*th root to normalize long and short sentences

2. Take the reciprocal to avoid small numbers

Perplexity Lower is “better”

TOY TRAINING CORPUS

i am sam

sam i am

i do not like green egg and ham

TOY TEST SENTENCE

i am sam i do

$$\frac{1}{\sqrt[4]{p_1 \times p_2 \times p_3 \times p_4}} \approx 2.1$$

Unrealistically low.
Expect 30 - 150 in
real life on test set

Perplexity

We can think of perplexity as "surprise". We want to minimize perplexity (on our training & test set)

Can use for classification (e.g. the SMS is spam if the perplexity of the spam model is < 100)

One more detail I skipped:

TOY TRAINING CORPUS

i am sam

sam i am

i do not like green egg and ham

TOY TEST SENTENCE

i am sam i do

TOY TRAINING CORPUS

<start> i am sam <end>

<start> sam i am <end>

<start> i do not like green egg and ham <end>

TOY TEST SENTENCE

<start> i am sam i do <end>

"I" starts sentence 2/3 of time

TOY TRAINING CORPUS

<start> i am sam <end>

<start> sam i am <end>

<start> i do not like green egg and ham <end>

TOY TEST SENTENCE

<start> i am sam i do <end>

$$p_1 = \frac{2}{3}$$

This is “all we need”, but there’s a big problem…

TOY TRAINING CORPUS

i am sam

sam i am

i do not like green egg and ham

TOY TEST SENTENCE

i like green egg and ham

TOY TRAINING CORPUS

i am sam

sam i am

i do not like green egg and ham

TOY TEST SENTENCE

i like green egg and ham

$$p_1 = \frac{\text{how many times “like” follows “i”}}{\text{how many times “i” appears}}$$

TOY TRAINING CORPUS

i am sam

sam i am

i do not like green egg and ham

TOY TEST SENTENCE

i like green egg and ham

$$p_1 = \frac{\text{how many times "like" follows "i"}}{\text{how many times "i" appears}}$$

TOY TRAINING CORPUS

i am sam

sam i am

i do not like green egg and ham

TOY TEST SENTENCE

i like green egg and ham

$$p_1 = \frac{0}{3}$$

TOY TRAINING CORPUS

i am sam

sam i am

i do not like green egg and ham

TOY TEST SENTENCE

i like green egg and ham

$$\frac{1}{\sqrt[n]{p_1 \times \ldots}} \approx \infty$$

TOY TRAINING CORPUS

i am sam

sam i am

i do not like green egg and ham

TOY TEST SENTENCE

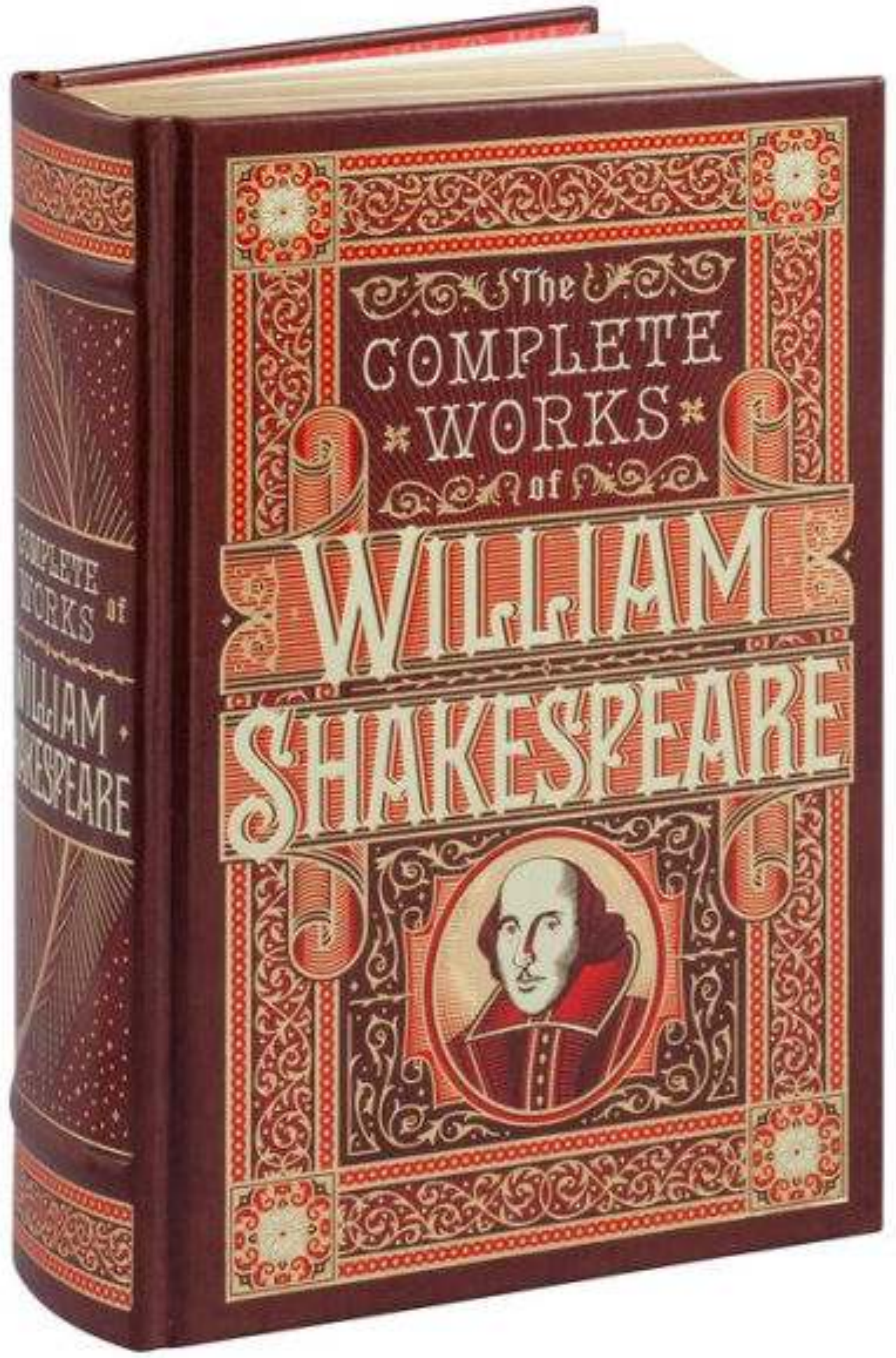
i like green egg and ham

$$\frac{1}{\sqrt[n]{p_1 \times \ldots}} \approx \infty$$

"Impossible"

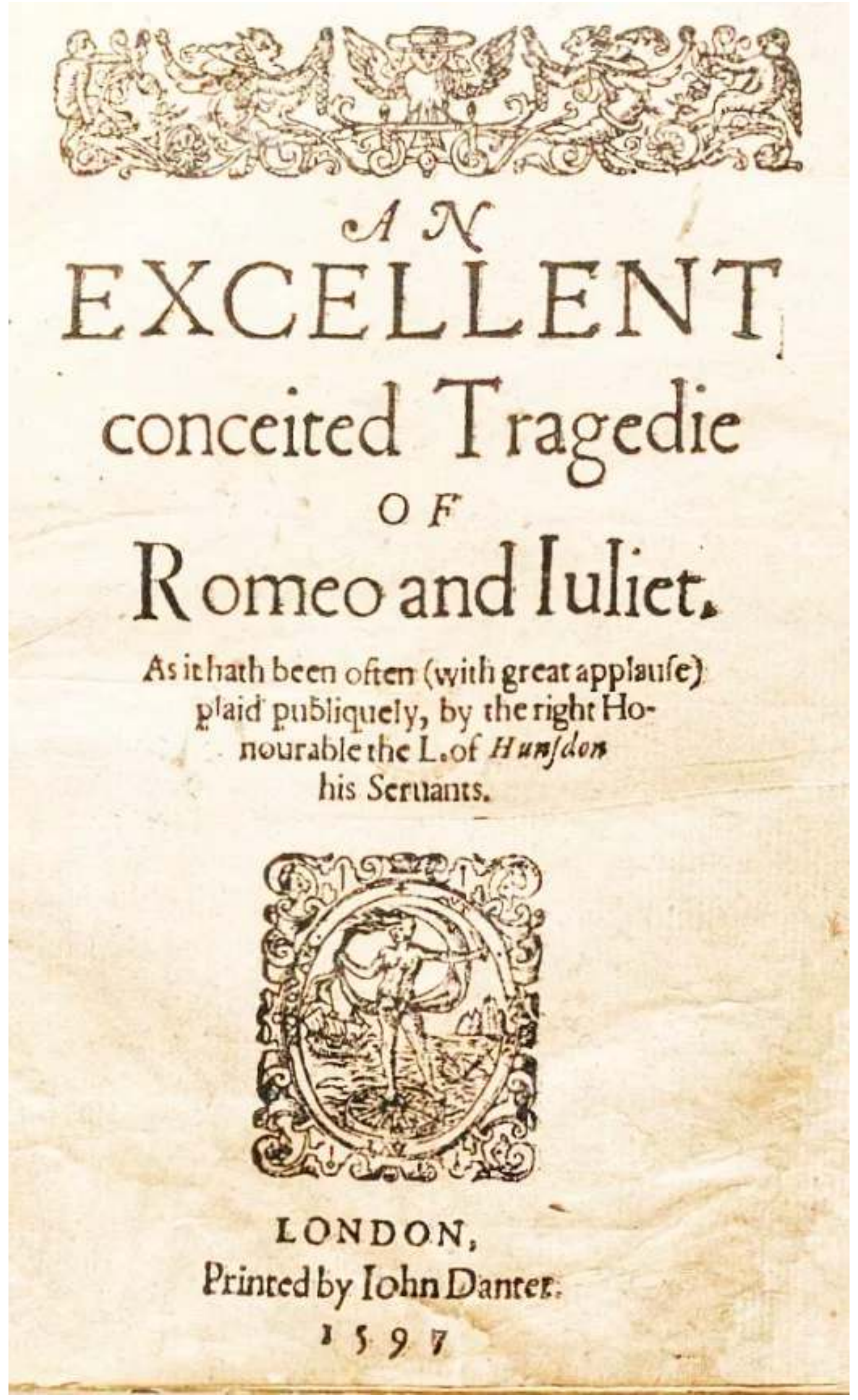
Does this happen on larger data sets?

Train



(Except Romeo & Juliet)

Test



But thou art not quickly moved to strike

– Romeo and Juliet, Act 1 Scene 1

(I didnt worry about stemming for this example)

But thou art not quickly moved to strike

but thou - 59

$$\left(\sqrt[7]{\frac{59}{5830} \times \ldots}\right)^{-1}$$

But thou art not quickly moved to strike

but thou - 59 thou art - 449

$$\left(\sqrt[7]{\frac{59}{5830} \times \frac{449}{6327} \times \ldots}\right)^{-1}$$

But thou art not quickly moved to strike

but thou - 59 thou art - 449 art not - 53

$$\left(\sqrt[7]{\frac{59}{5830} \times \frac{449}{6327} \times \frac{53}{3812} \times \ldots}\right)^{-1}$$

But thou art not quickly moved to strike

but thou - 59 thou art - 449 art not - 53 not quickly - 4

$$\left(\sqrt[7]{\frac{59}{5830} \times \frac{449}{6327} \times \frac{53}{3812} \times \frac{4}{9507} \times \ldots}\right)^{-1}$$

But thou art not quickly moved to strike

but thou - 59   thou art - 449   art not - 53   not quickly - 4

quickly moved - 0

$$\left(\sqrt[7]{\frac{59}{5830} \times \frac{449}{6327} \times \frac{53}{3812} \times \frac{4}{9507} \times 0 \times \ldots}\right)^{-1}$$

But thou art not **quickly moved** to strike

but thou - 59 thou art - 449 art not - 53 not quickly - 4

quickly moved - 0

$$\left(\sqrt[7]{\frac{59}{5830}\times\frac{449}{6327}\times\frac{53}{3812}\times\frac{4}{9507}\times 0\times\ldots}\right)^{-1}$$

(Multiplying by zero 😭)

But thou art not quickly moved to strike

but thou - 59 thou art - 449 art not - 53 not quickly - 4

~~quickly moved - 0~~ moved to - 5 to strike - 15

$$(\sqrt[6]{\frac{59}{5830} \times \frac{449}{6327} \times \frac{53}{3812} \times \frac{4}{9507} \times \frac{5}{93} \times \frac{15}{163}})^{-1} = 60$$

But thou art not quickly moved to strike

but thou - 59 thou art - 449 art not - 53 not quickly - 4

~~quickly moved - 0~~ moved to - 5 to strike - 15

$$\left(\sqrt[6]{\frac{59}{5830} \times \frac{449}{6327} \times \frac{53}{3812} \times \frac{4}{9507} \times \frac{5}{93} \times \frac{15}{163}}\right)^{-1} = 60$$

(We generally expect perplexity of 30 - 150 on test set)

But thou art not quickly moved to strike

but thou - 59 thou art - 449 art not - 53 not quickly - 4

quickly moved - 0 moved to - 5 to strike - 15

$$(\sqrt[7]{\frac{59}{5830} \times \frac{449}{6327} \times \frac{53}{3812} \times \frac{4}{9507} \times 0 \times \frac{5}{93} \times \frac{15}{163}})^{-1} = \infty$$

The solution is **smoothing** which is all about preventing the probabilities crashing to 0

Smoothing

Laplacian/additive

Good-Turing

Katz

Jelinek-Mercer

Lidstone

Witten-Bell

Church-Gale

Bayesian

Kneser-Ney

Absolute discounting

Smoothing

Laplacian/additive

Good-Turing

Katz

Jelinek-Mercer

Lidstone

Witten-Bell

Church-Gale

Bayesian

Kneser-Ney

Absolute discounting

But thou art not quickly moved to strike

but thou - 59 thou art - 449 art not - 53 not quickly - 4

quickly moved - 0

$$\left(\sqrt[7]{\frac{59}{5830} \times \frac{449}{6327} \times \frac{53}{3812} \times \frac{4}{9507} \times 0 \times \ldots}\right)^{-1}$$

quickly

<end> **13**
and **4**
dream **1**
go **2**
have **2**
make **2**
send **1**
should **2**
the **2**
to **3**
will **2**
yield **1**
⋮

a **0**
abandon **0**
abandoned **0**
abase **0**
abashed **0**
abate **0**
abated **0**
abatement **0**
abates **0**
abbess **0**
abbey **0**
abbeys **0**

⋮
youths **0**
zeal **0**
zeales **0**
zealous **0**
zenith **0**
zephires **0**
zir **0**
zodiac **0**
zone **0**
zounds **0**

quickly

~28000 more, including "moved"

68 more

<end> **13**
and **4**
dream **1**
go **2**
have **2**
make **2**
send **1**
should **2**
the **2**
to **3**
will **2**
yield **1**
⋮

a **0**
abandon **0**
abandoned **0**
abase **0**
abashed **0**
abate **0**
abated **0**
abatement **0**
abates **0**
abbess **0**
abbey **0**
abbeys **0**

⋮
youths **0**
zeal **0**
zeales **0**
zealous **0**
zenith **0**
zephires **0**
zir **0**
zodiac **0**
zone **0**
zounds **0**

quickly

< 1% of the words (80 of 28,000+) control 100% of the probability!

<end> **13**
and **4**
dream **1**
go **2**
have **2**
make **2**
send **1**
should **2**
the **2**
to **3**
will **2**
yield **1**
⋮

a **0**
abandon **0**
abandoned **0**
abase **0**
abashed **0**
abate **0**
abated **0**
abatement **0**
abates **0**
abbess **0**
abbey **0**
abbeys **0**

⋮
youths **0**
zeal **0**
zeales **0**
zealous **0**
zenith **0**
zephires **0**
zir **0**
zodiac **0**
zone **0**
zounds **0**

quickly

< 1% of the words (80 of 28,000+) control 100% of the probability!

<end> **13**
and **4**
dream **1**
go **2**
have **2**
make **2**
send **1**
should **2**
the **2**
to **3**
will **2**
yield **1**

a **0**
abandon **0**
abandoned **0**
abase **0**
abashed **0**
abate **0**
abated **0**
abatement **0**
abates **0**
abbess **0**
abbey **0**
abbeys **0**

⋮

youths **0**
zeal **0**
zeales **0**
zealous **0**
zenith **0**
zephires **0**
zir **0**
zodiac **0**
zone **0**
zounds **0**

#occupylanguagemodels

quickly

Solution: tax every word that appears.

1. Reduce every observed count by some δ.

(often $\delta = 0.5$)

<end> **13**
and **4**
dream **1**
go **2**
have **2**
make **2**
send **1**
should **2**
the **2**
to **3**
will **2**
yield **1**
⋮

a **0**
abandon **0**
abandoned **0**
abase **0**
abashed **0**
abate **0**
abated **0**
abatement **0**
abates **0**
abbess **0**
abbey **0**
abbeys **0**

⋮
youths **0**
zeal **0**
zeales **0**
zealous **0**
zenith **0**
zephires **0**
zir **0**
zodiac **0**
zone **0**
zounds **0**

quickly

Solution: tax every word that appears.

1. Reduce every observed count by some δ.

(often δ = 0.5)

<end> **12.5**
and **3.5**
dream **0.5**
go **1.5**
have **1.5**
make **1.5**
send **0.5**
should **1.5**
the **1.5**
to **2.5**
will **1.5**
yield **0.5**
⋮

a **0**
abandon **0**
abandoned **0**
abase **0**
abashed **0**
abate **0**
abated **0**
abatement **0**
abates **0**
abbess **0**
abbey **0**
abbeys **0**

⋮
youths **0**
zeal **0**
zeales **0**
zealous **0**
zenith **0**
zephires **0**
zir **0**
zodiac **0**
zone **0**
zounds **0**

quickly

Solution: tax every word that appears.

2. We've collected 40 counts of "tax" (80 × 0.5). Now redistribute that 40 across the 28,000 words with a count of 0.

<end> **12.5**
and **3.5**
dream **0.5**
go **1.5**
have **1.5**
make **1.5**
send **0.5**
should **1.5**
the **1.5**
to **2.5**
will **1.5**
yield **0.5**
⋮

a **0**
abandon **0**
abandoned **0**
abase **0**
abashed **0**
abate **0**
abated **0**
abatement **0**
abates **0**
abbess **0**
abbey **0**
abbeys **0**

⋮
youths **0**
zeal **0**
zeales **0**
zealous **0**
zenith **0**
zephires **0**
zir **0**
zodiac **0**
zone **0**
zounds **0**

quickly

Solution: tax every word that appears.

2. We've collected 6 counts of "tax" (12 × 0.5). Now redistribute that 6 across the 28,000 words with a count of 0.

$$\frac{40}{28384} \approx 0.0014$$

<end> **12.5**
and **3.5**
dream **0.5**
go **1.5**
have **1.5**
make **1.5**
send **0.5**
should **1.5**
the **1.5**
to **2.5**
will **1.5**
yield **0.5**
⋮

a **0.0014**
abandon **0.0014**
abandoned **0.0014**
abase **0.0014**
abashed **0.0014**
abate **0.0014**
abated **0.0014**
abatement **0.0014**
abates **0.0014**
abbess **0.0014**
abbey **0.0014**
abbeys **0.0014**

⋮
youths **0.0014**
zeal **0.0014**
zeales **0.0014**
zealous **0.0014**
zenith **0.0014**
zephires **0.0014**
zir **0.0014**
zodiac **0.0014**
zone **0.0014**
zounds **0.0014**

But thou art not quickly moved to strike

but thou - 58.5 thou art - 448.5 art not - 52.5 not quickly - 3.5

quickly moved - 0.0014 moved to - 4.5 to strike - 14.5

$$\left(\sqrt[7]{\frac{58.5}{5830} \times \frac{448.5}{6326.5} \times \frac{52.5}{3812} \times \frac{3.5}{9507} \times \frac{0.0014}{108} \times \frac{4.5}{93} \times \frac{14.5}{163}}\right)^{-1} = 175$$

bigram	won a free **holiday**, offer
trigram	won a free **holiday**, offer
4-gram	won a free **holiday**, offer

bigram	won a free **holiday**, offer
trigram	won a free **holiday**, offer
4-gram	won a free **holiday**, offer

Pro: “smarter”, considers more context

Cons: data sparsity

Trigram model

But thou art not quickly moved to strike

quickly moved

Appears 0 times

a **0**

abandon **0**

abandoned **0**

abase **0**

abashed **0**

abate **0**

abated **0**

abatement **0**

abates **0**

abbess **0**

abbey **0**

abbeys **0**

⋮

youths **0**

zeal **0**

zeales **0**

zealous **0**

zenith **0**

zephires **0**

zir **0**

zodiac **0**

zone **0**

zounds **0**

quickly moved

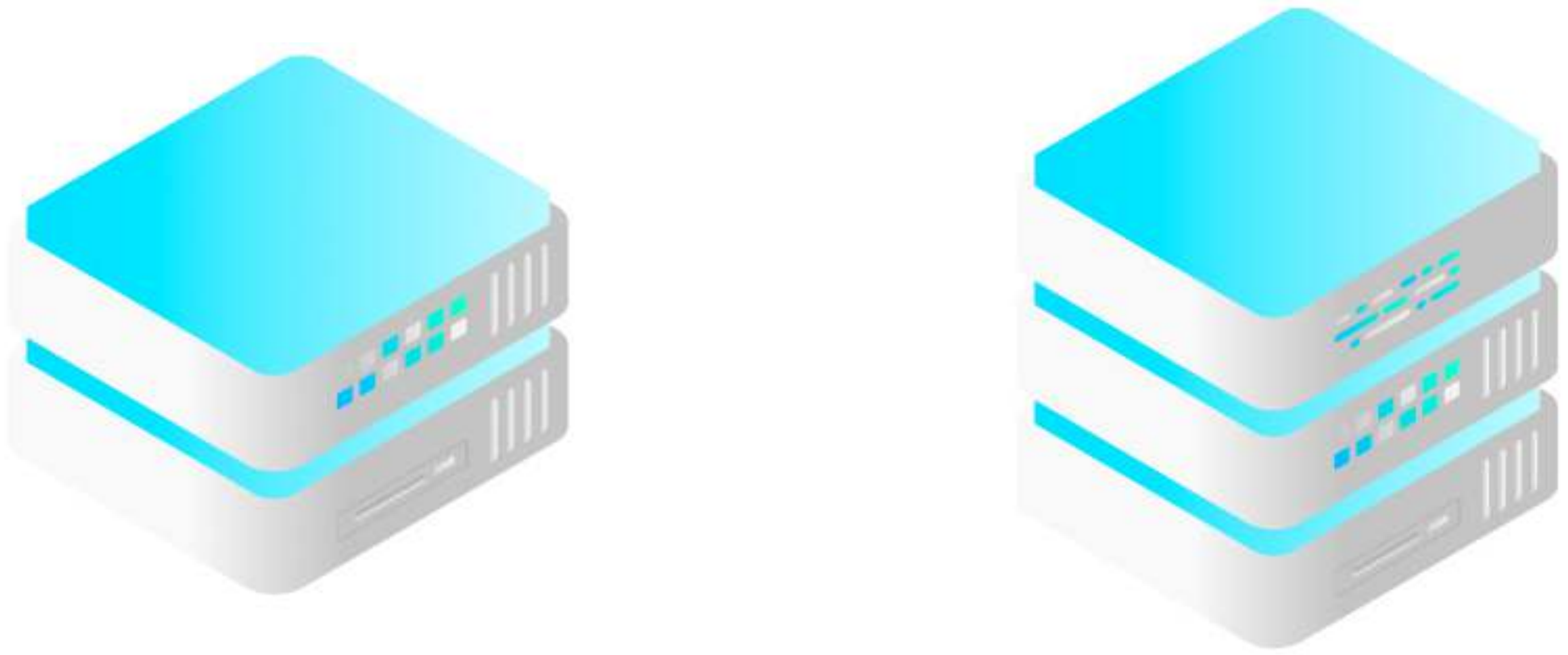
So all counts after are 0. Absolute discounting doesn't help because there's nothing to "tax"

Smoothing alone can't help us!

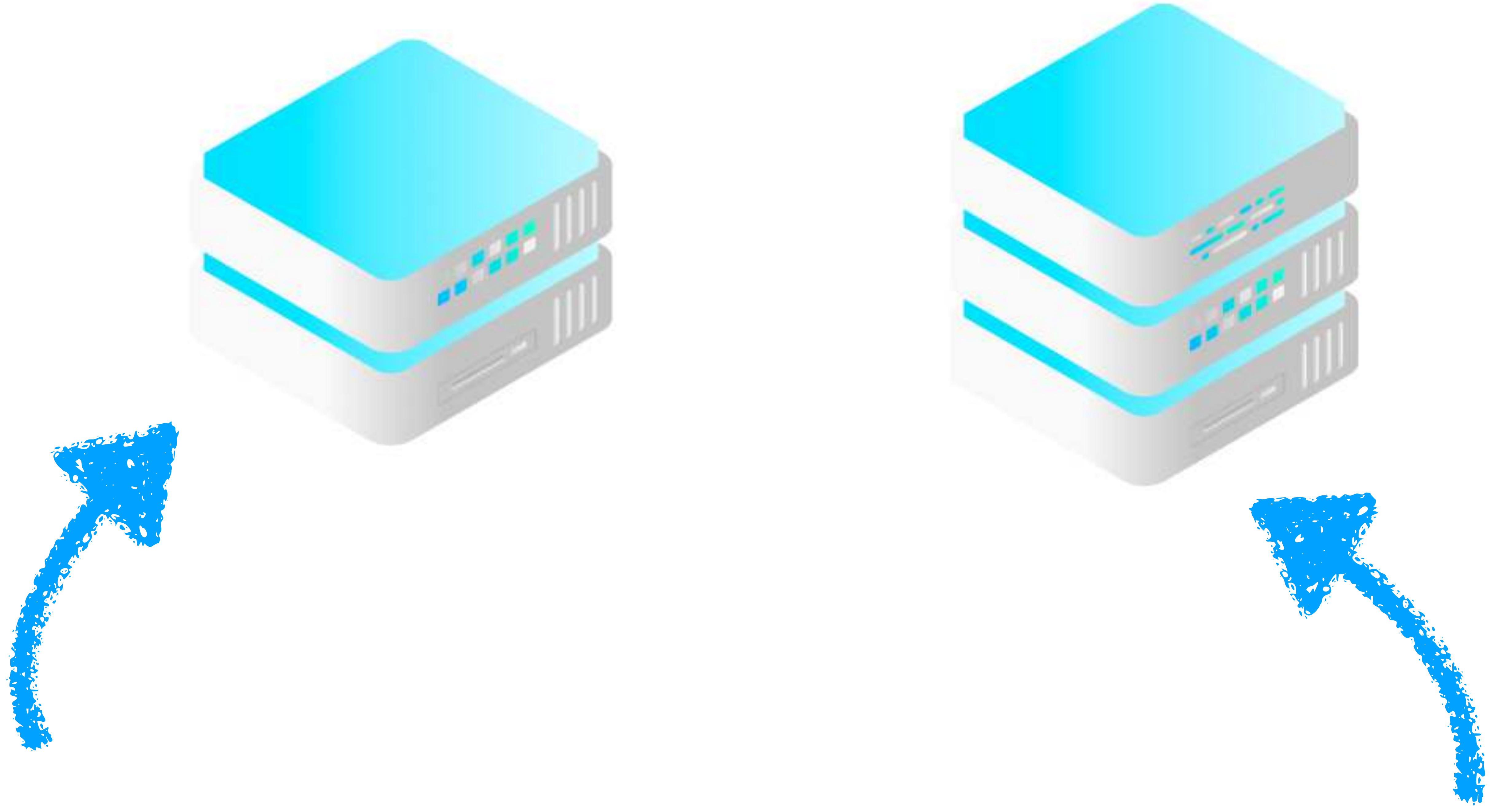
a **0**
abandon **0**
abandoned **0**
abase **0**
abashed **0**
abate **0**
abated **0**
abatement **0**
abates **0**
abbess **0**
abbey **0**
abbeys **0**

⋮

youths **0**
zeal **0**
zeales **0**
zealous **0**
zenith **0**
zephires **0**
zir **0**
zodiac **0**
zone **0**
zounds **0**

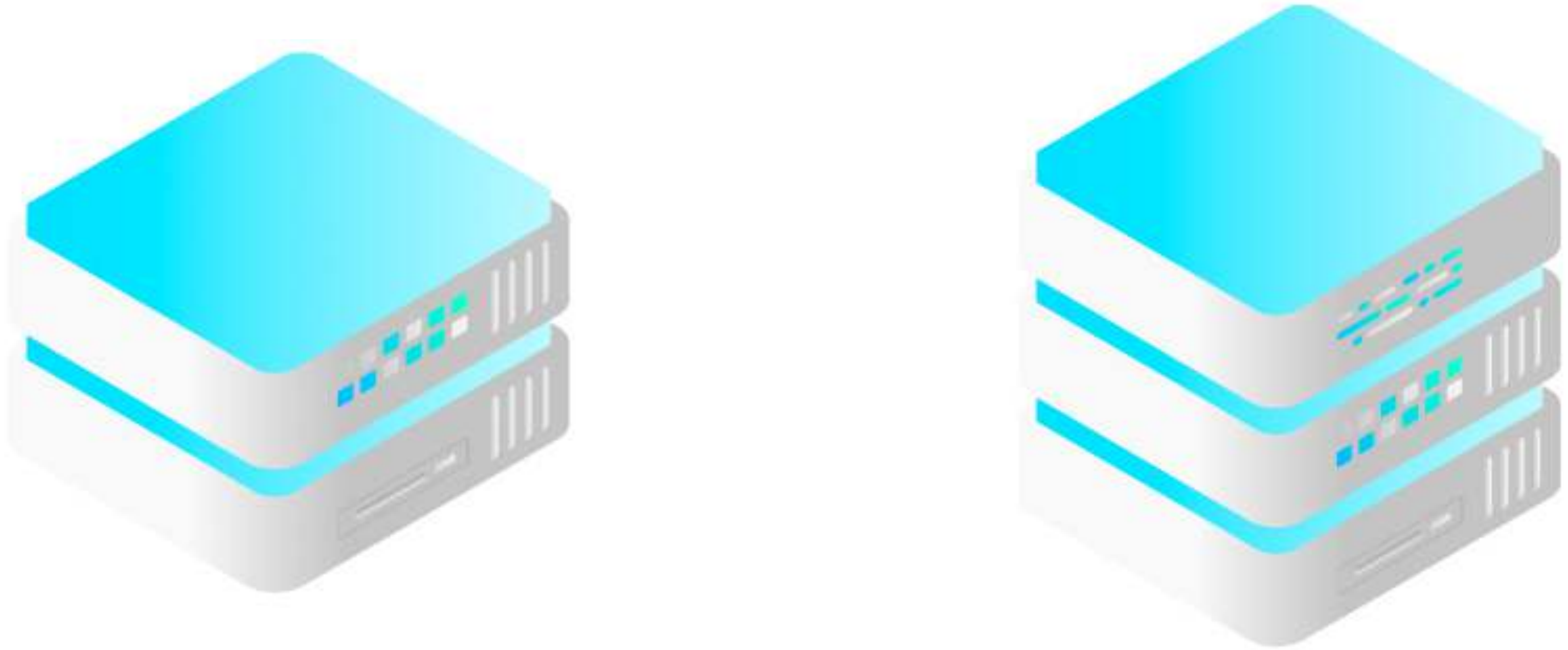


Smart but can get stumped

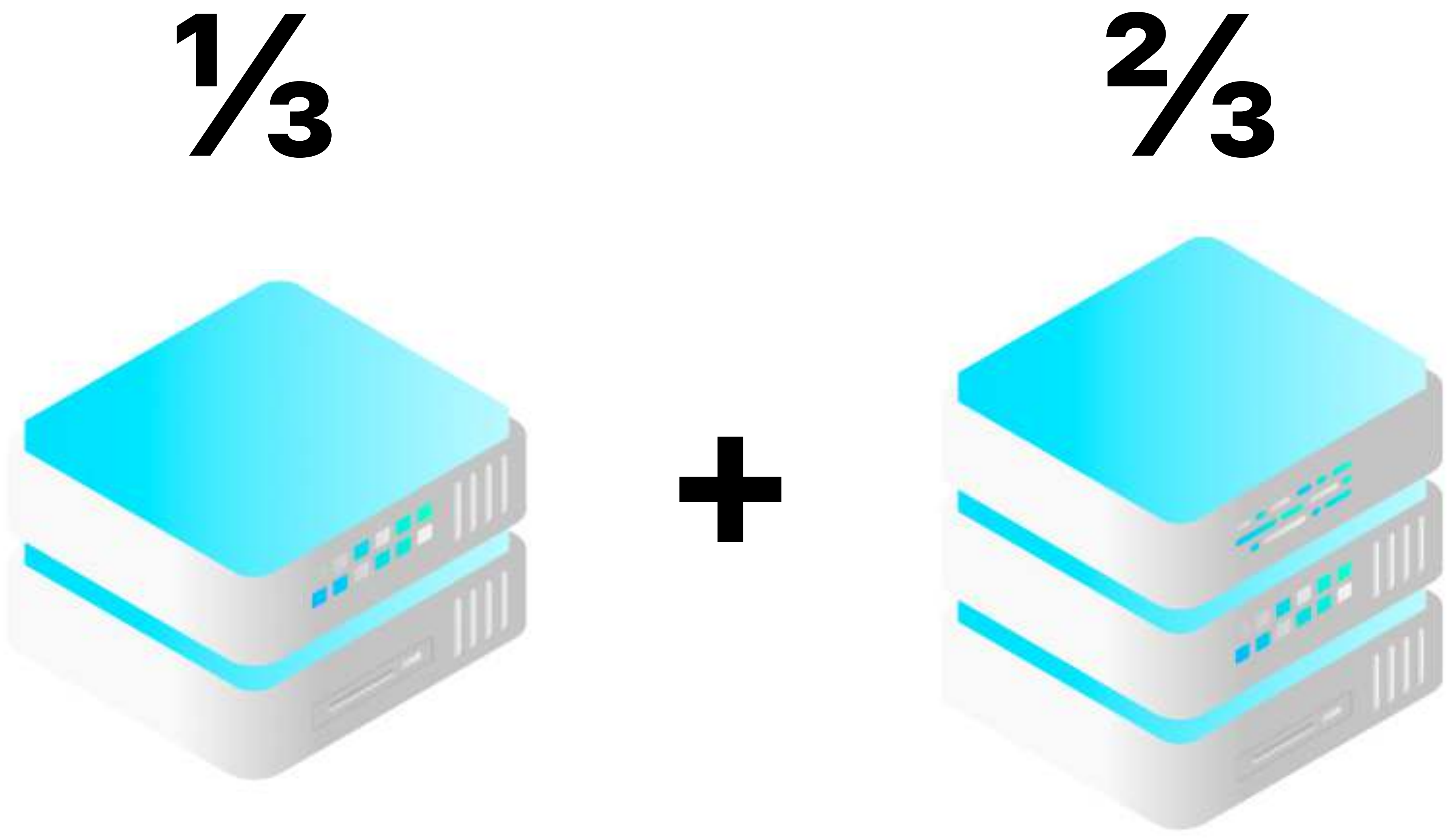


Not as smart but robust

Smart but can get stumped



Solution: combine trigram and bigram predictions

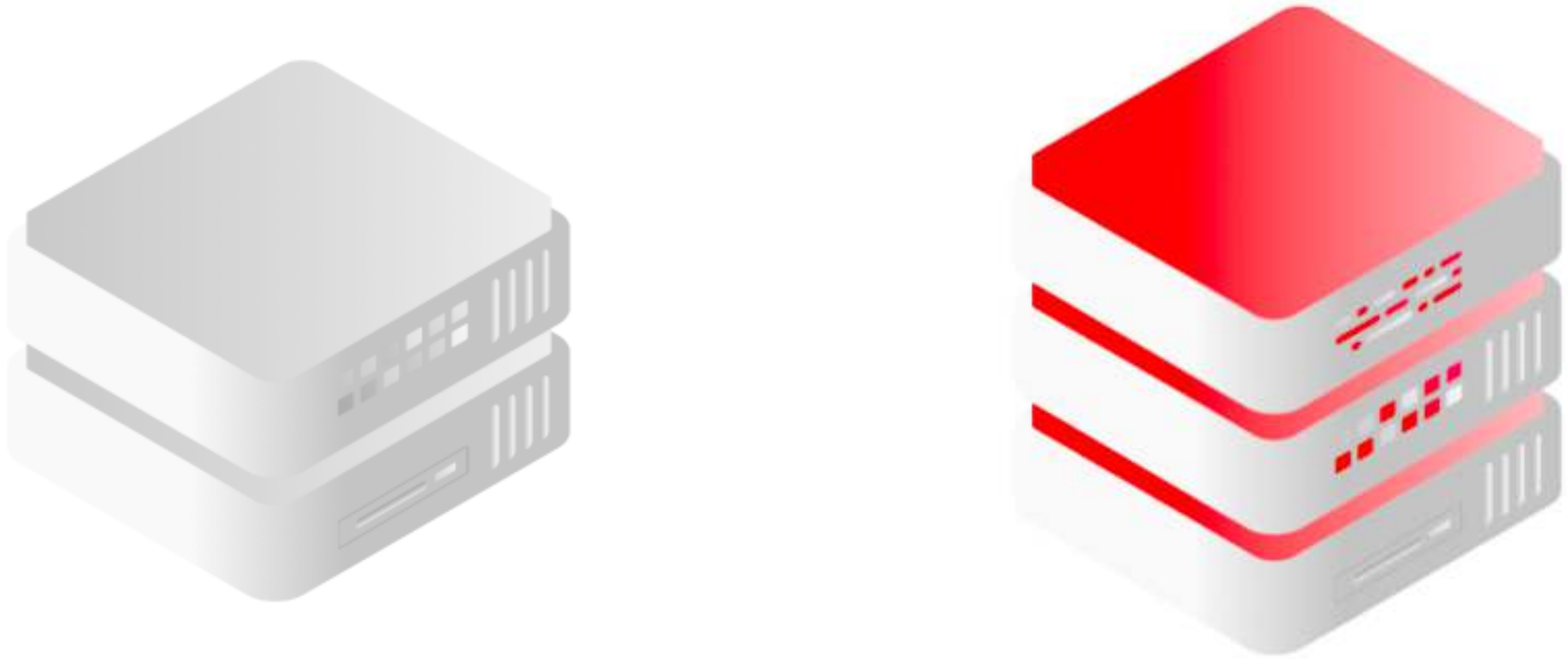


Approach 1: Interpolation

Take a proportion of each model's prediction

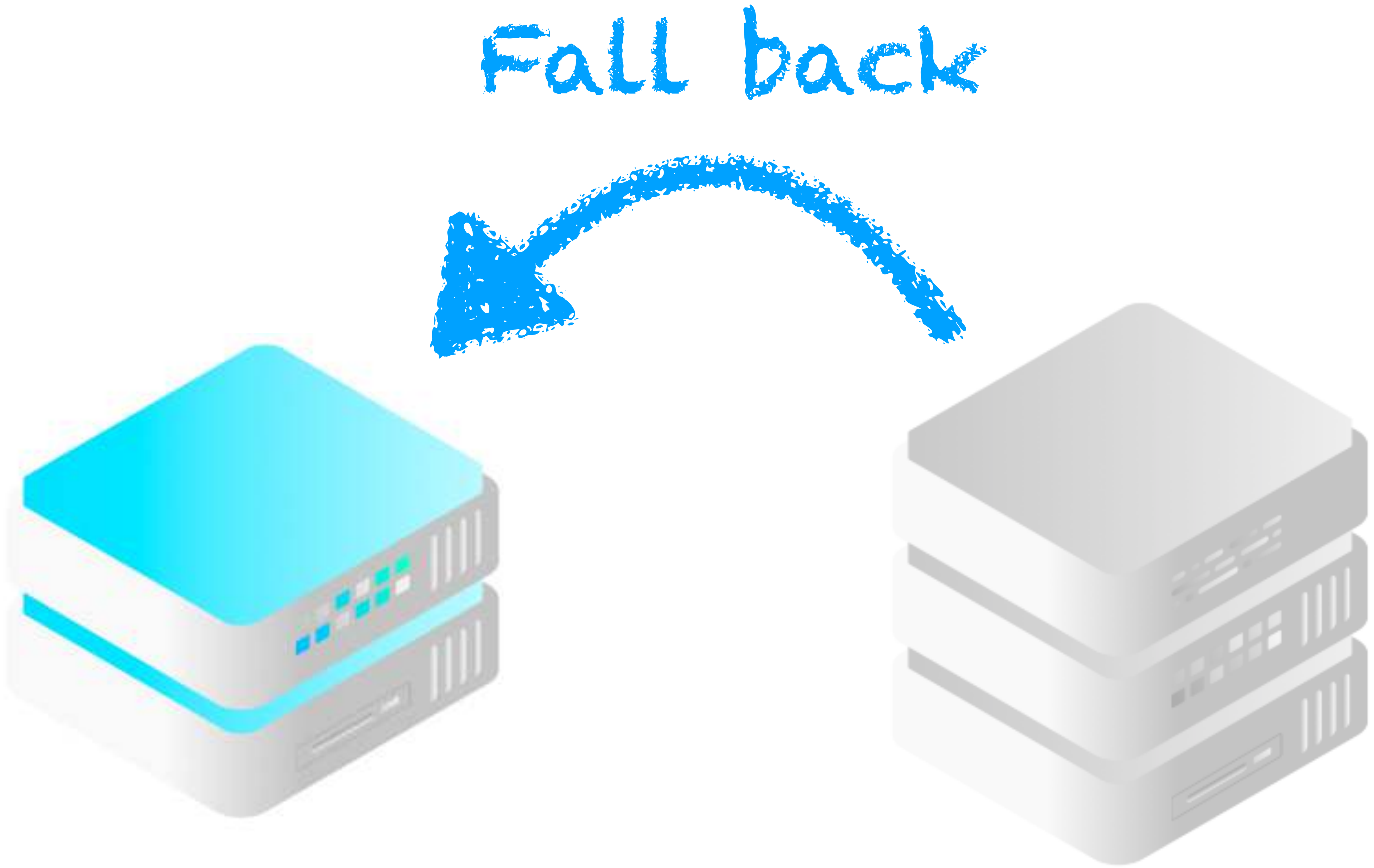
Approach 2: Backoff

If the trigram model gets stumped, fall back to the bigram model



Approach 2: Backoff

If the trigram model gets stumped, fall back to the bigram model



Approach 2: Backoff

If the trigram model gets stumped, fall back to the bigram model

Smoothing

Laplacian/additive

Witten-Bell

Good-Turing

Church-Gale

Katz

Bayesian

Jelinek-Mercer

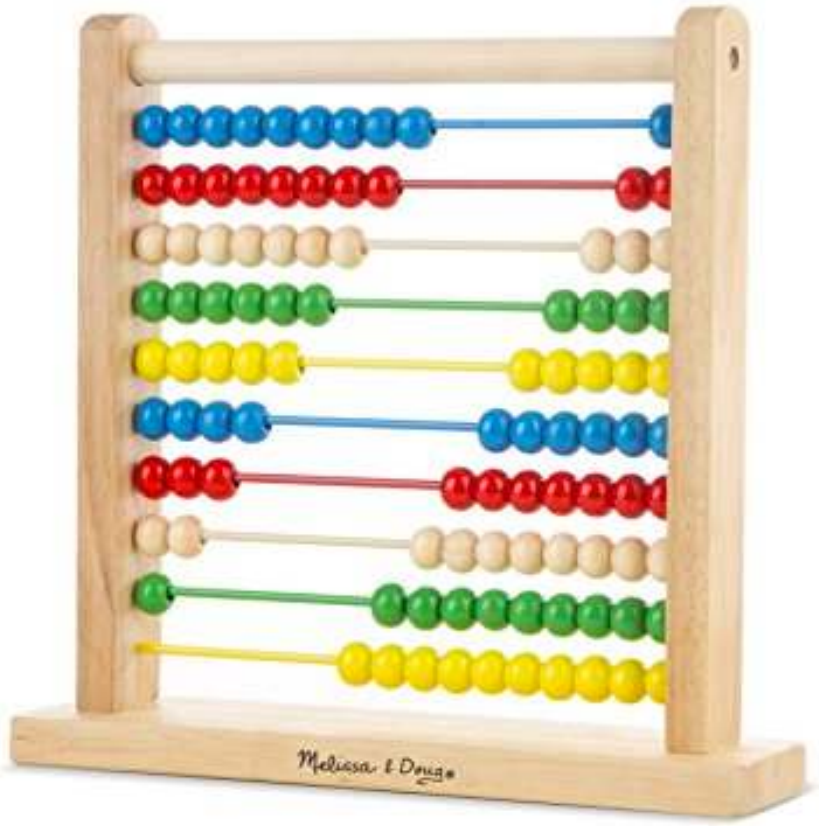
Kneser-Ney

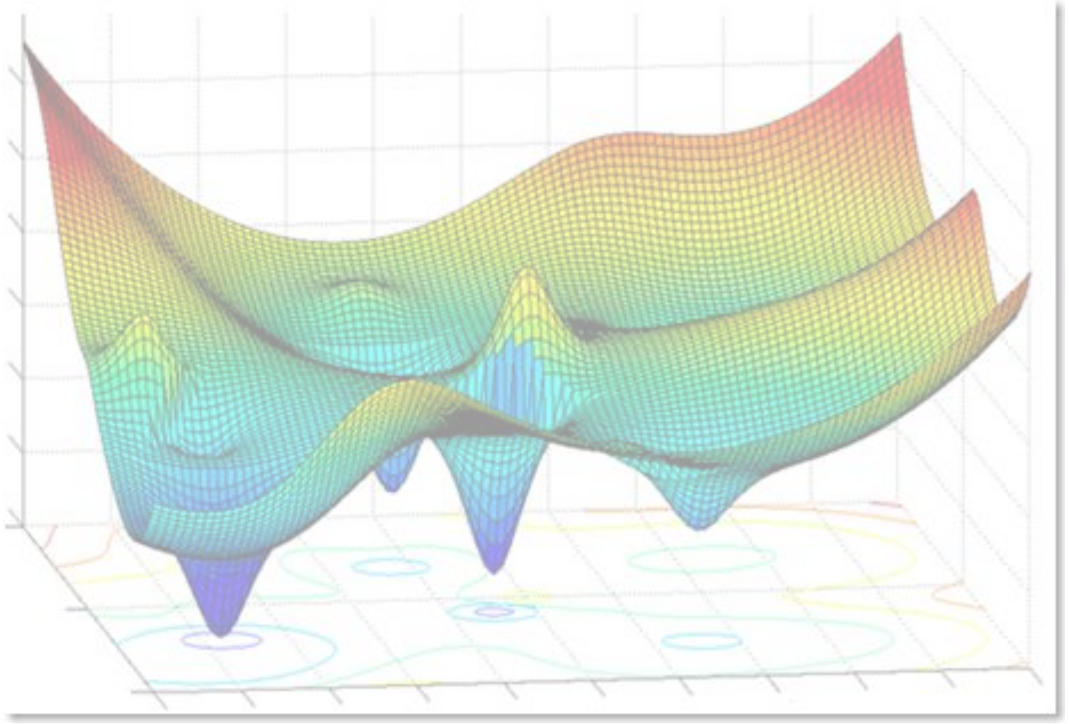
Lidstone

Absolute discounting

With interpolation

```python
from nltk.model.ngram import NgramModel
from nltk.probability import KneserNeyProbDist

model = NgramModel(n = 3, train = text, estimator = KneserNeyProbDist)

model.perplexity("But thou art not quickly moved to strike".split())
```





Count based

AKA statistical

1980, 1990s

Very fast

Decent performance (when tuned)

Continuous space

AKA neural, neuroprobabilistic

2000s, 2010s

Slower, more expensive

Typically used with neural nets

State-of-the-art performance

TRAINING

The cat got squashed in the garden on Friday

TEST

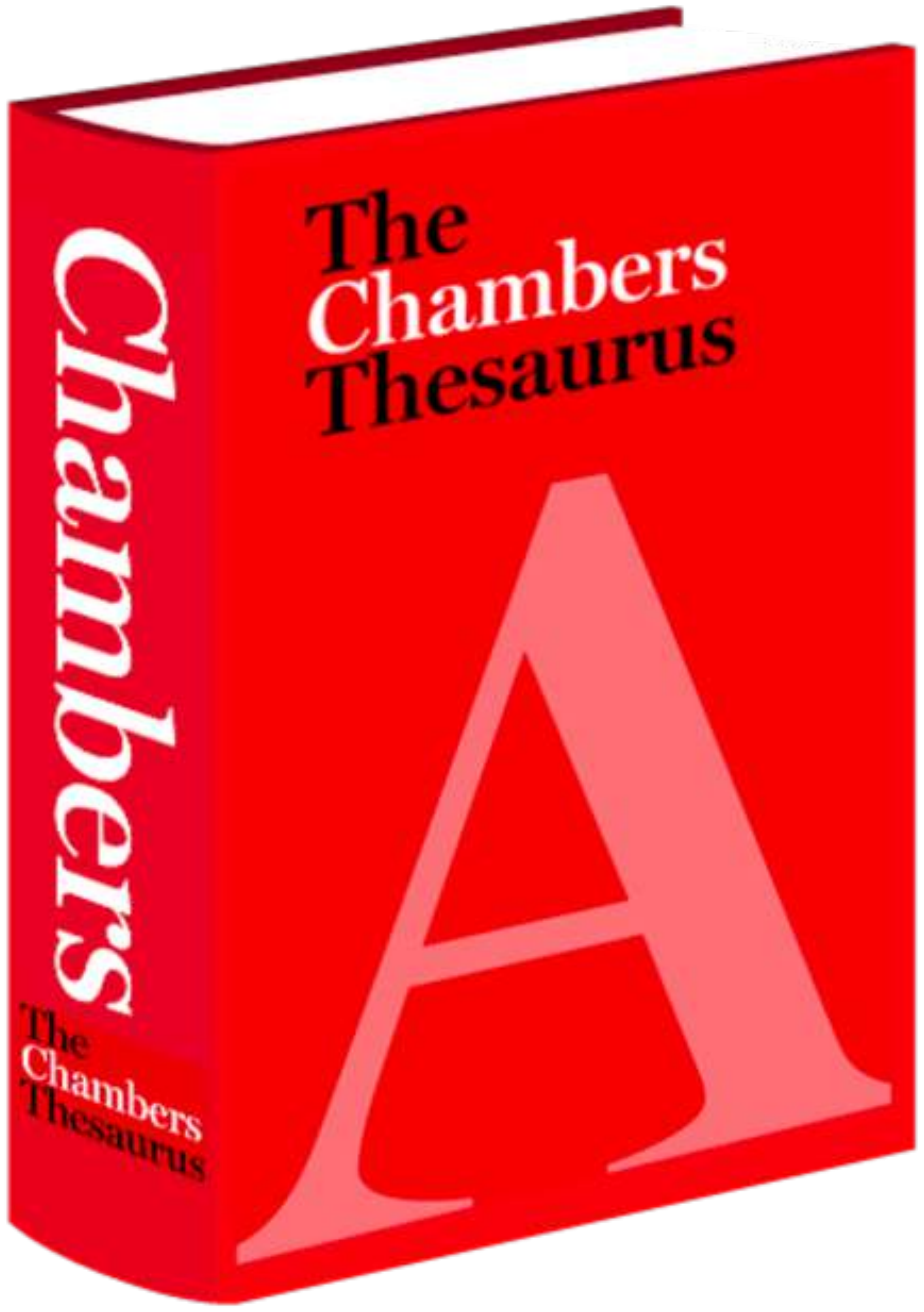
The dog got flattened in the garden on Tuesday

TRAINING

The cat got squashed in the garden on Friday

TEST

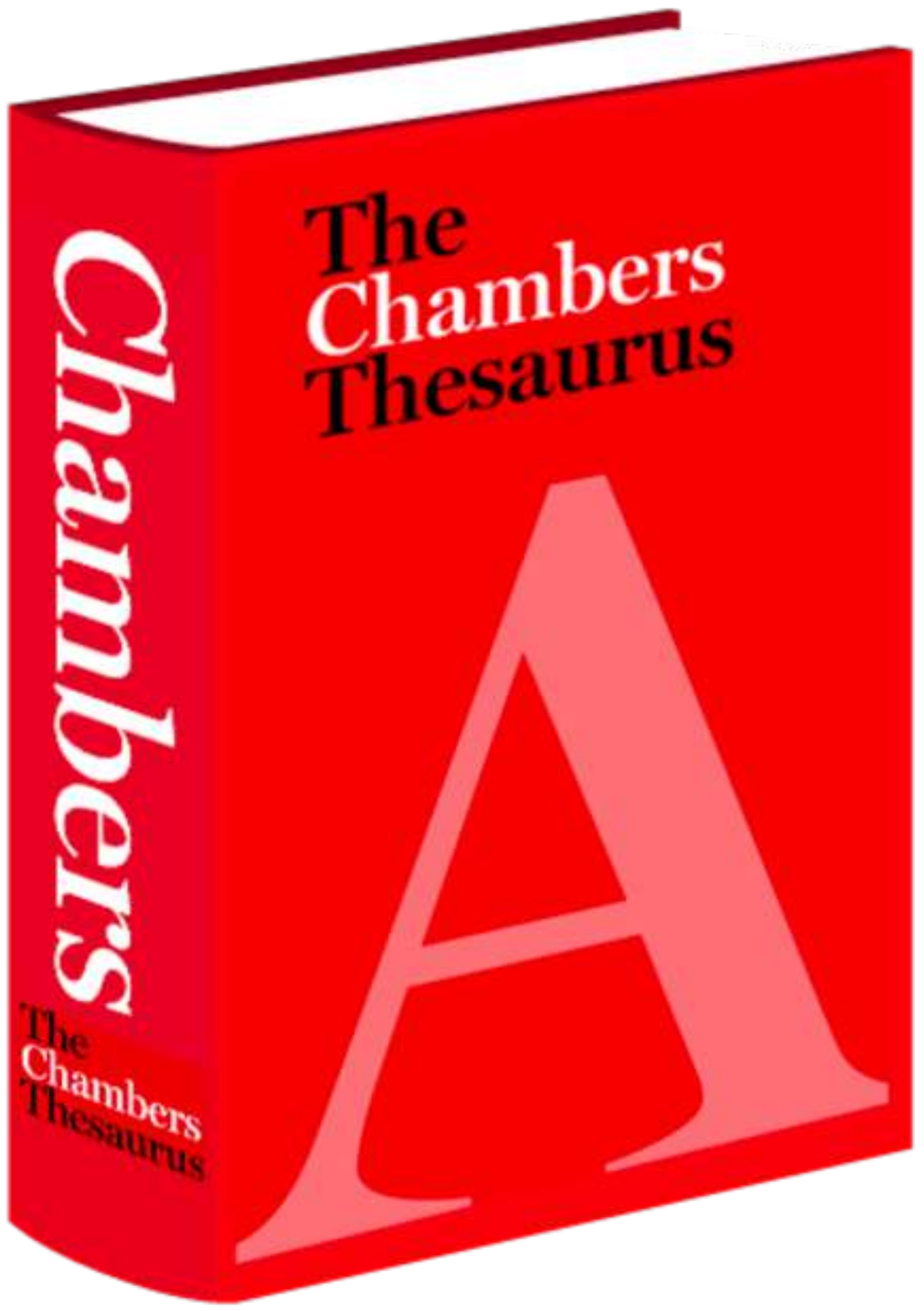
The dog got flattened in the garden on Tuesday

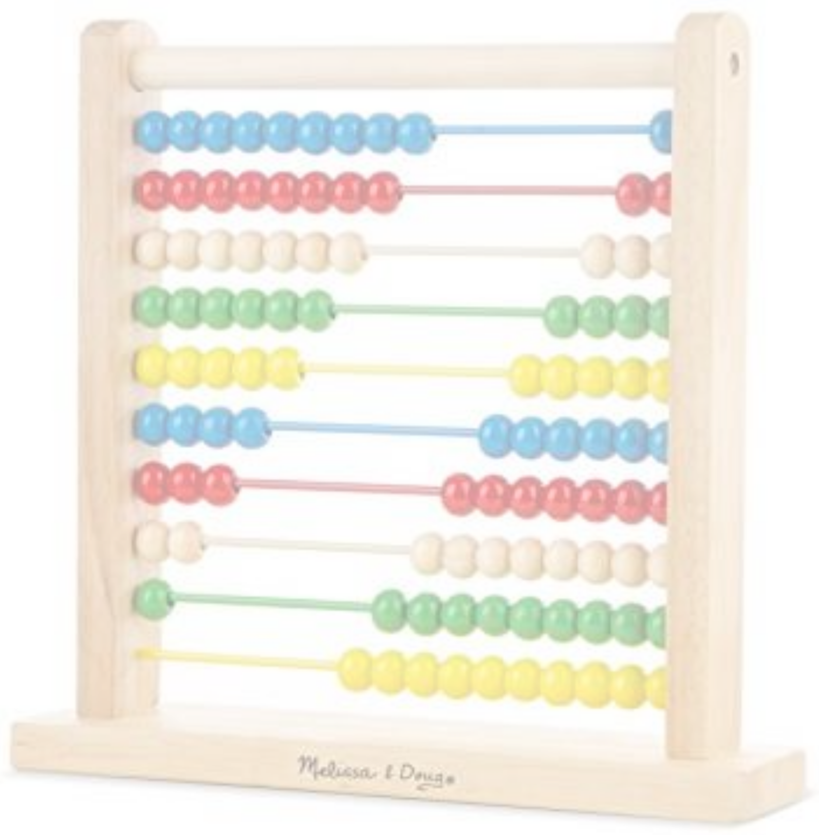

The Chambers Thesaurus
Chambers
The Chambers Thesaurus

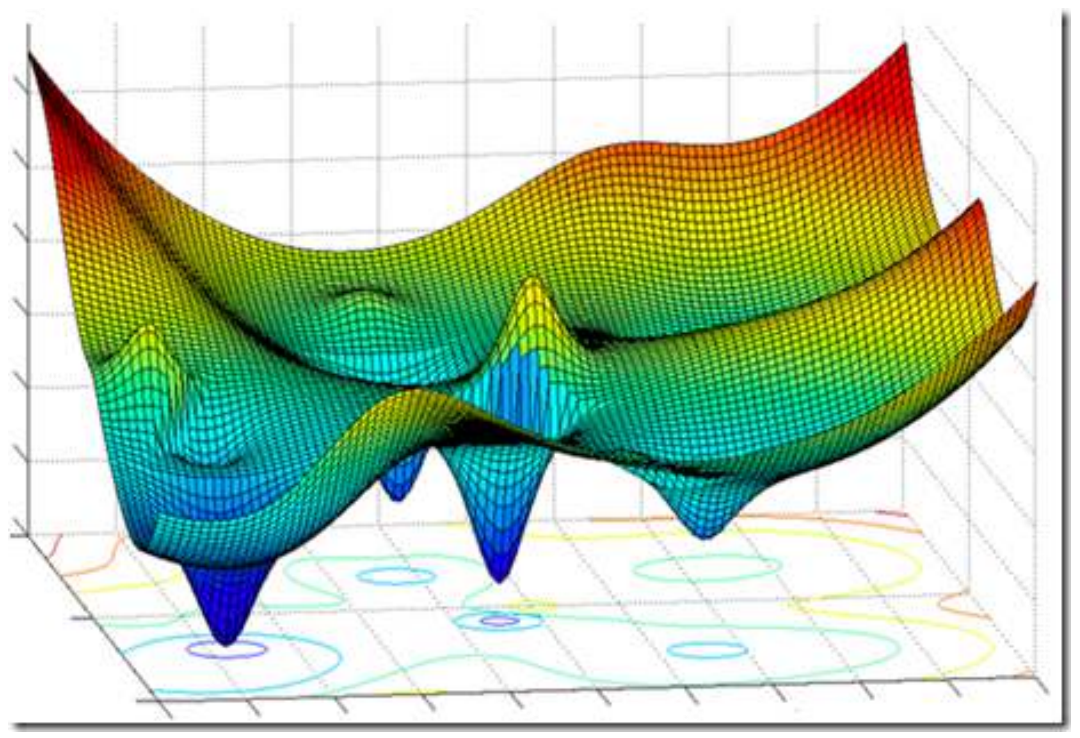
squashed $\approx$ flattened

cat $\neq$ dog

Friday $\neq$ Tuesday







Count based

AKA statistical

1980, 1990s

Very fast

Decent performance (when tuned)

Continuous space

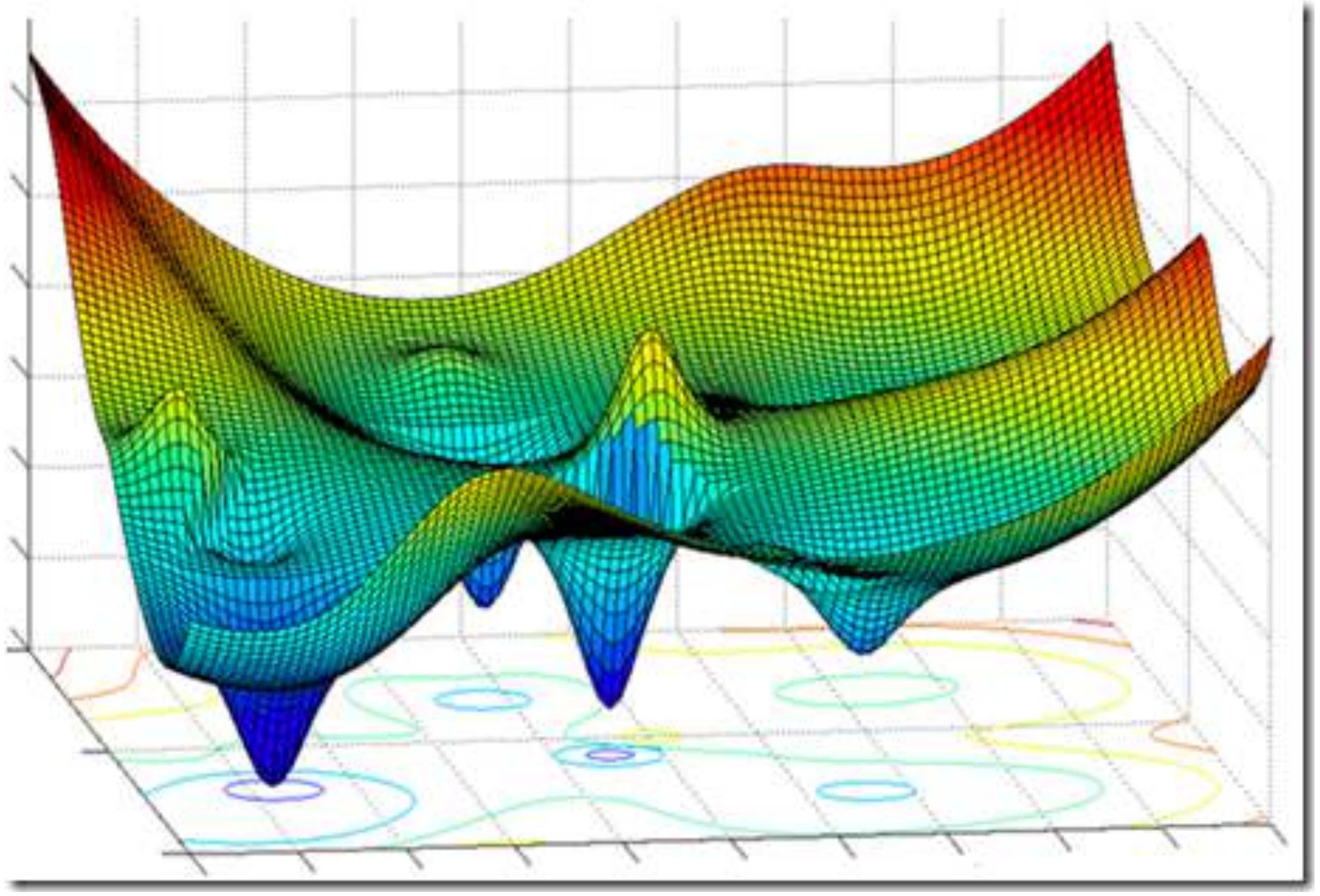
AKA neural, neuroprobabilistic

2000s, 2010s

Slower, more expensive

Typically used with neural nets

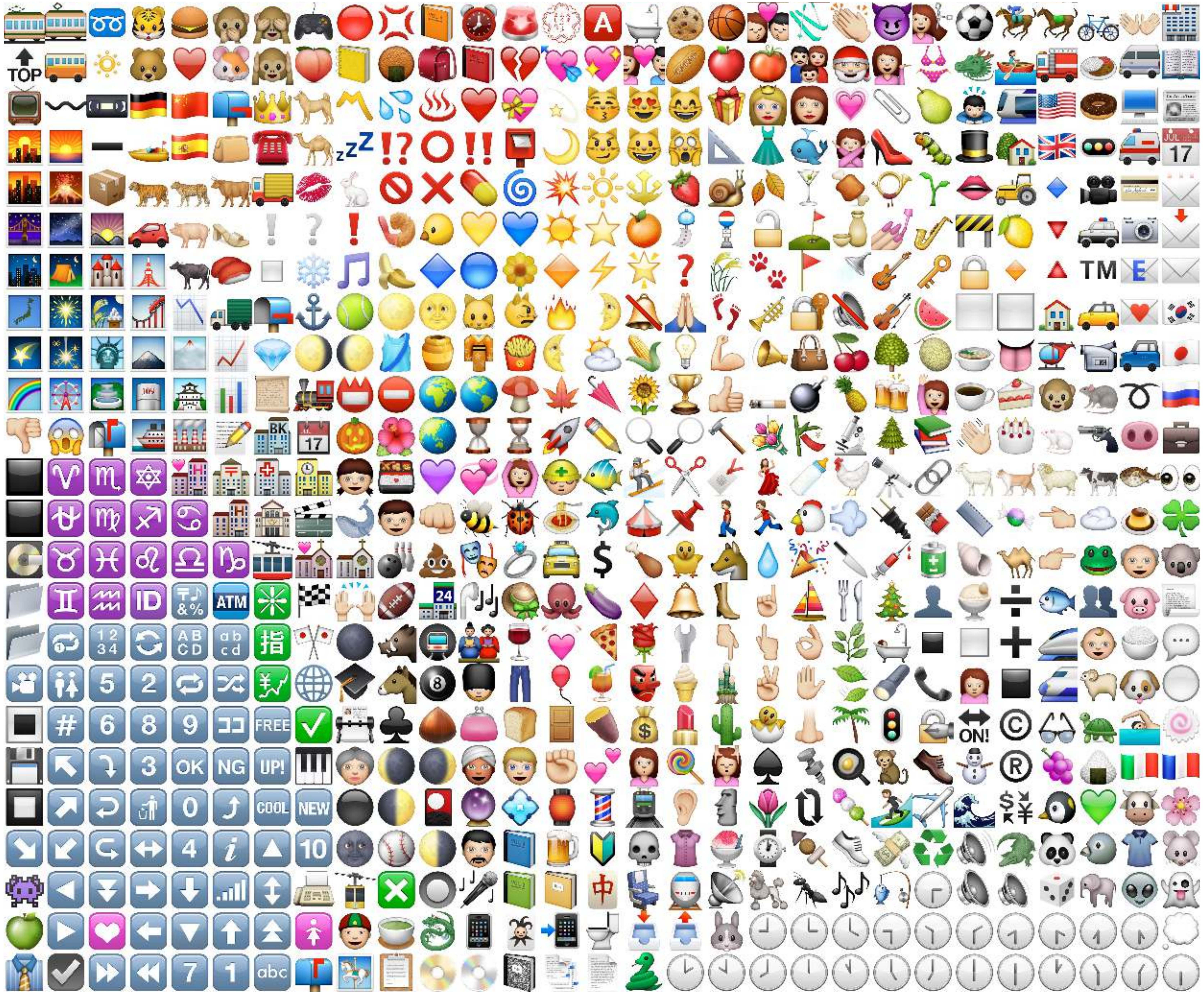
State-of-the-art performance

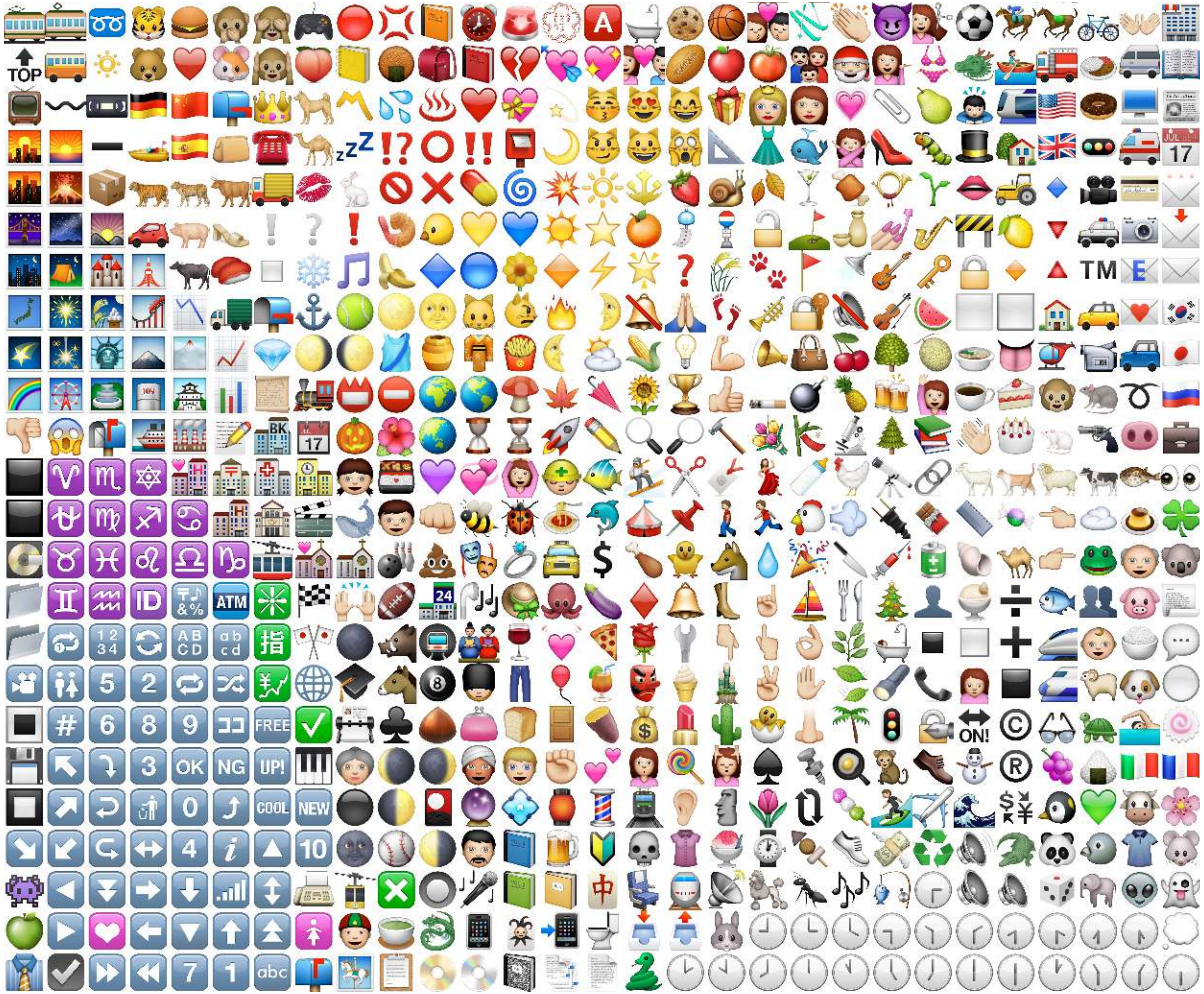


Bengio et al., 2003

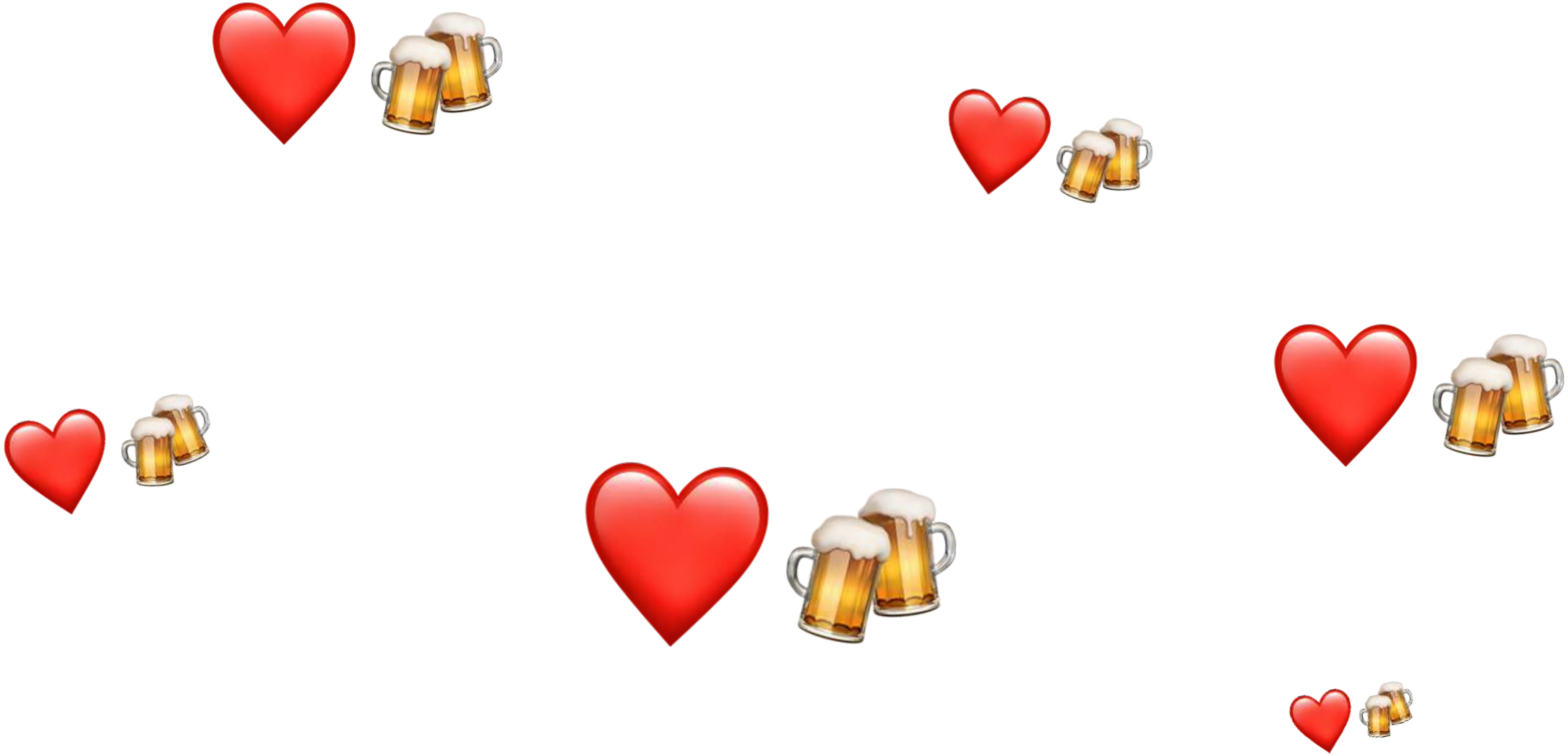
Key idea: use word embeddings, where similar words live close to one another

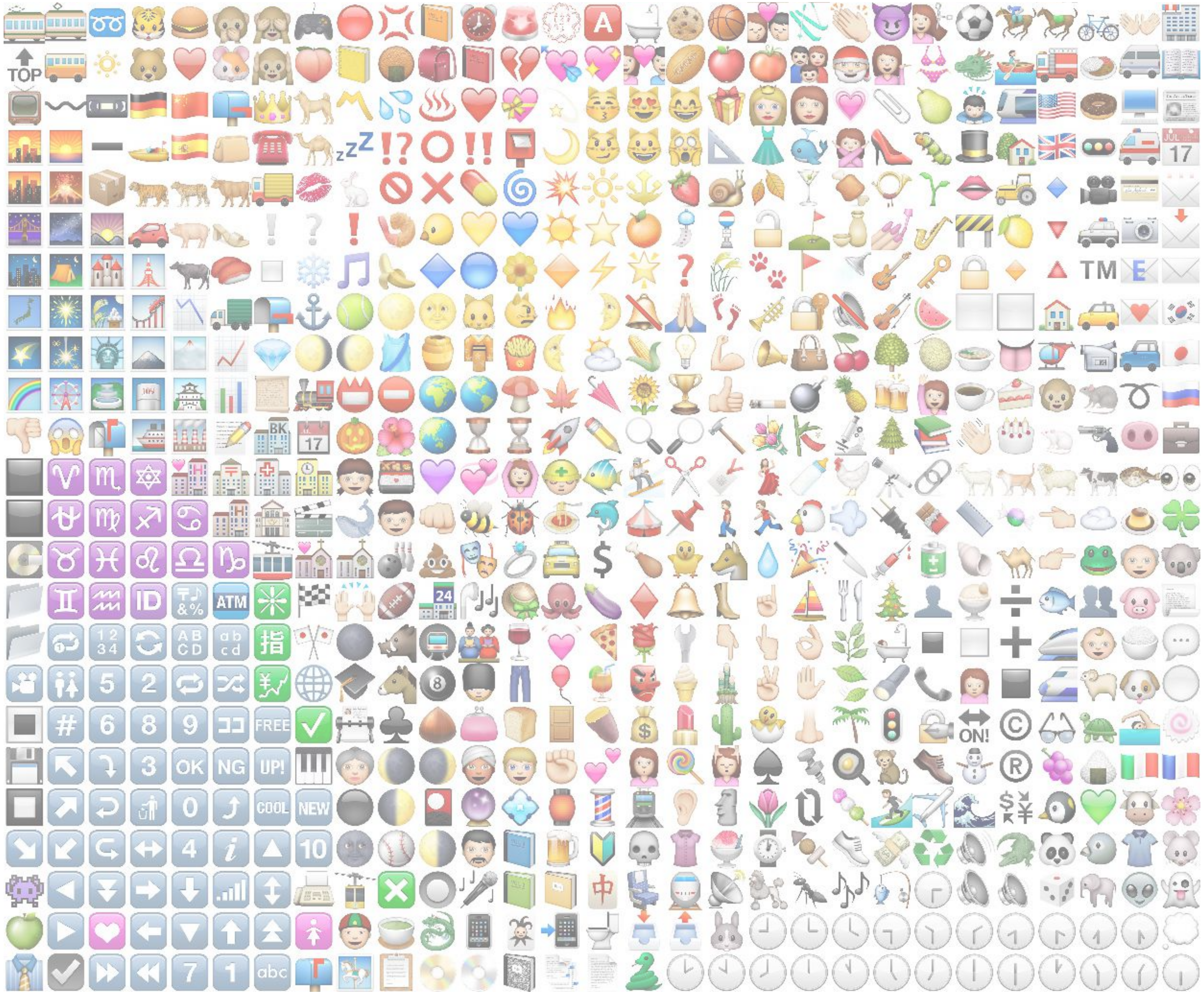
Toy example: Emoji embedded in 2D space



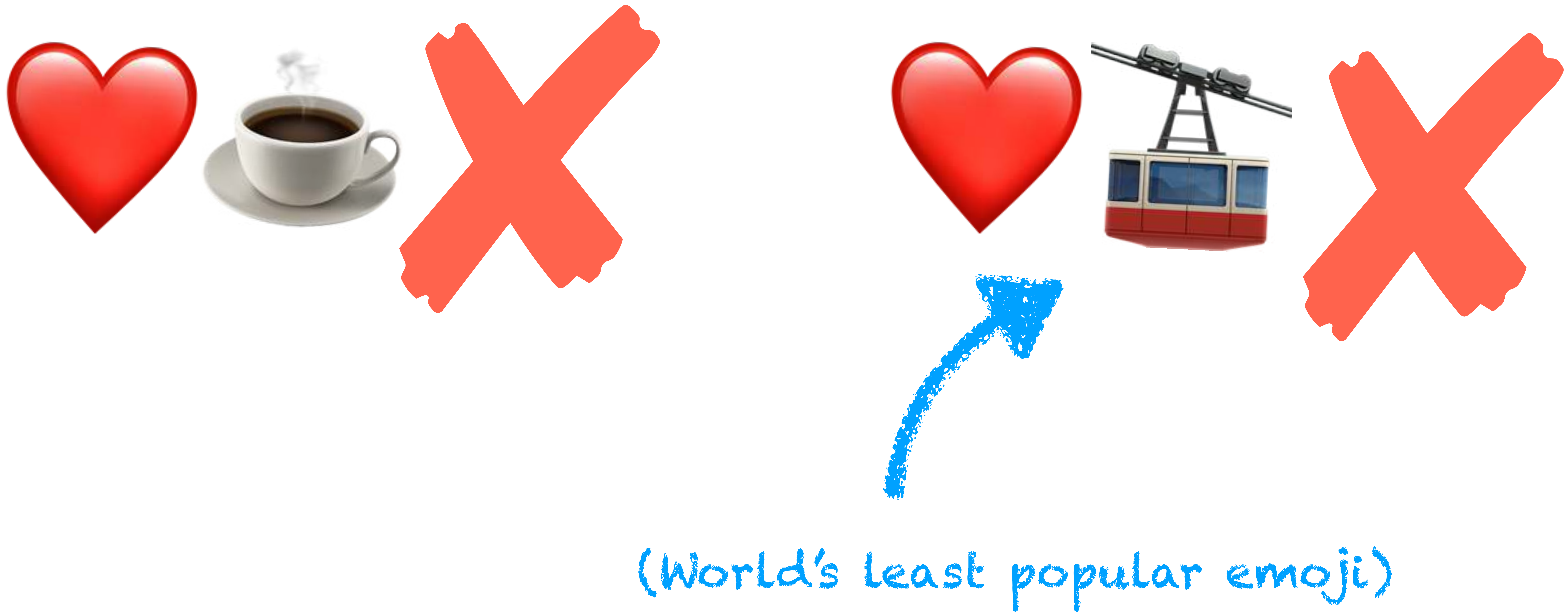


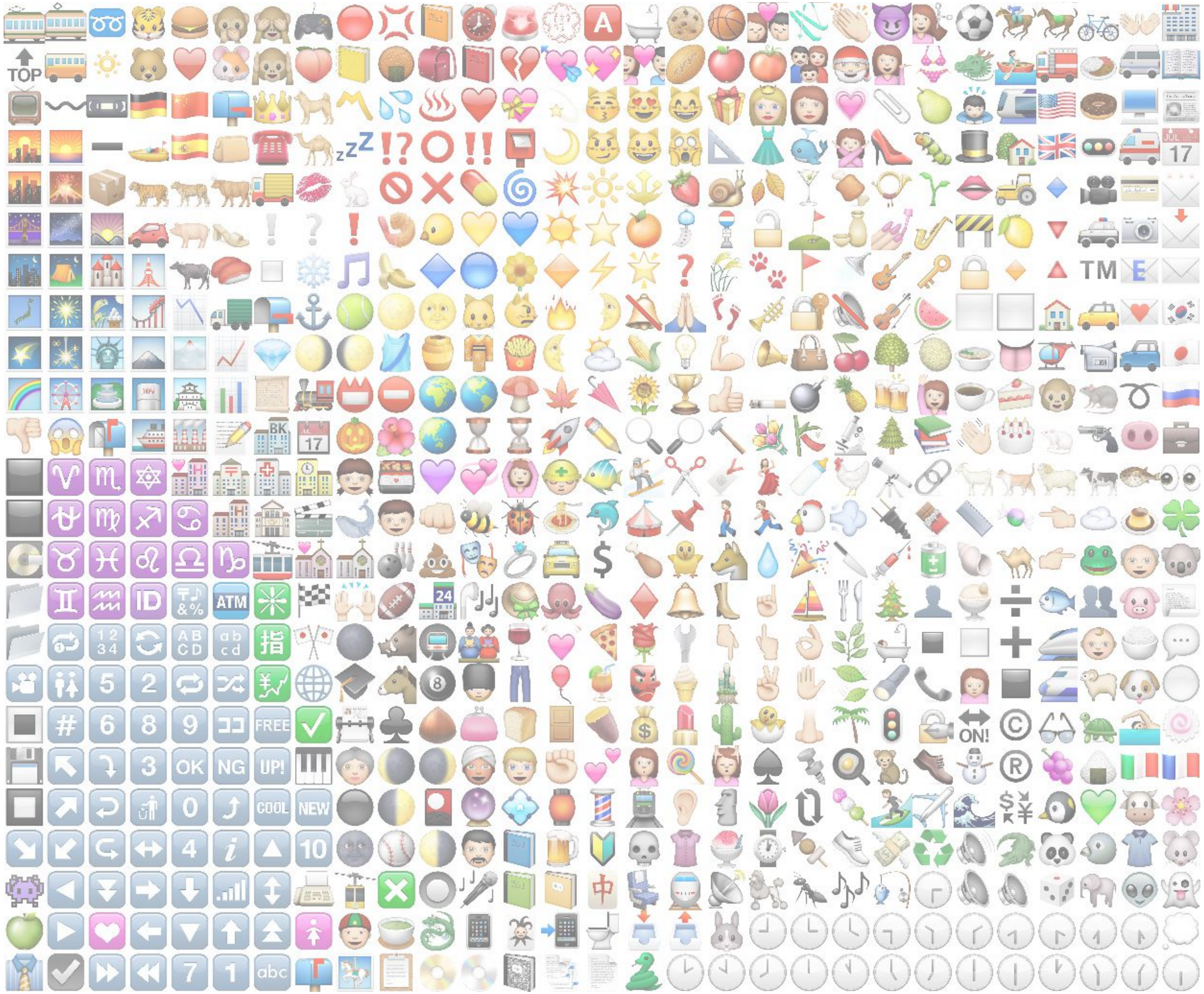
...





❤️🍻 50%

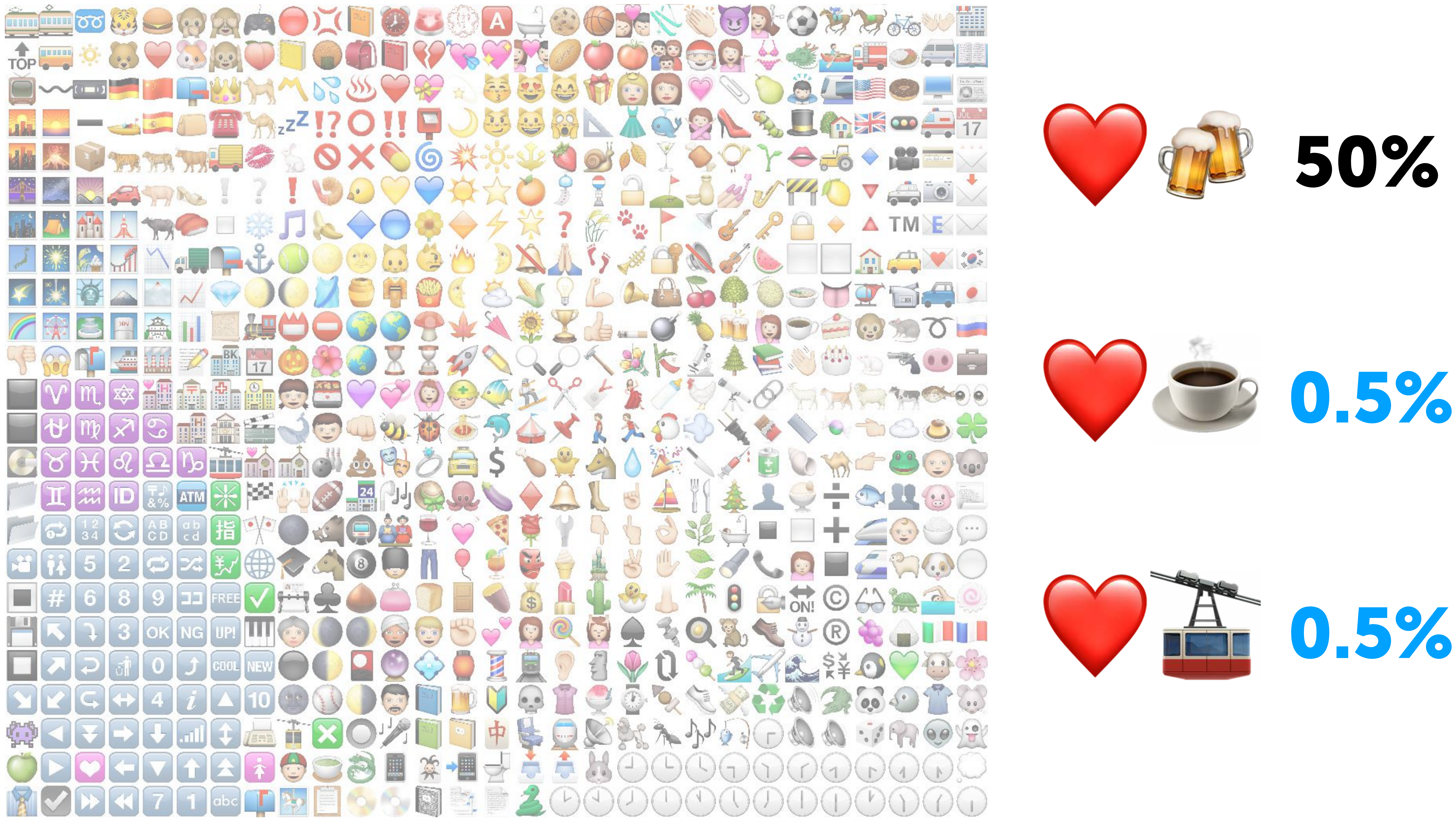

(World's least popular emoji)

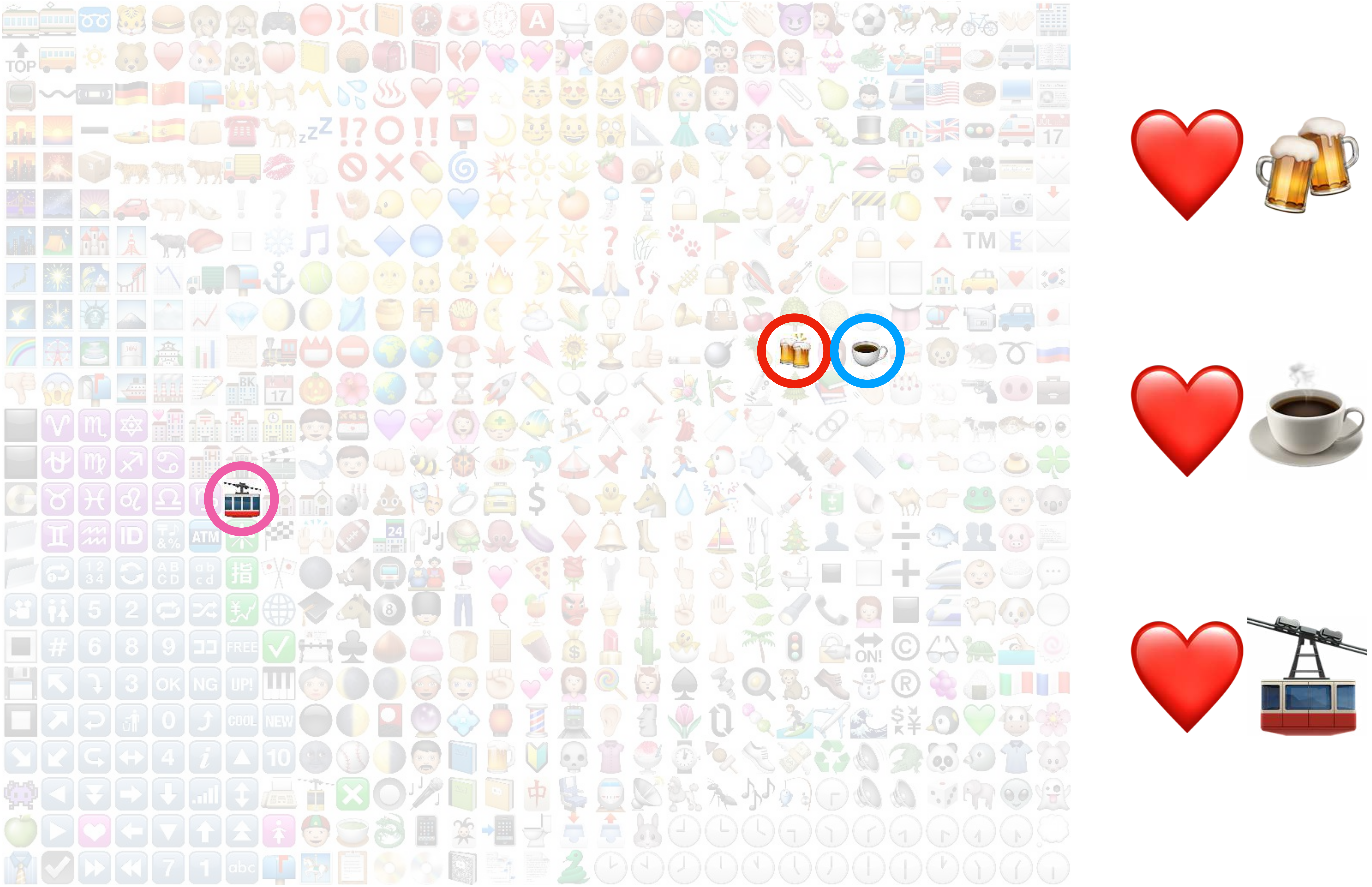


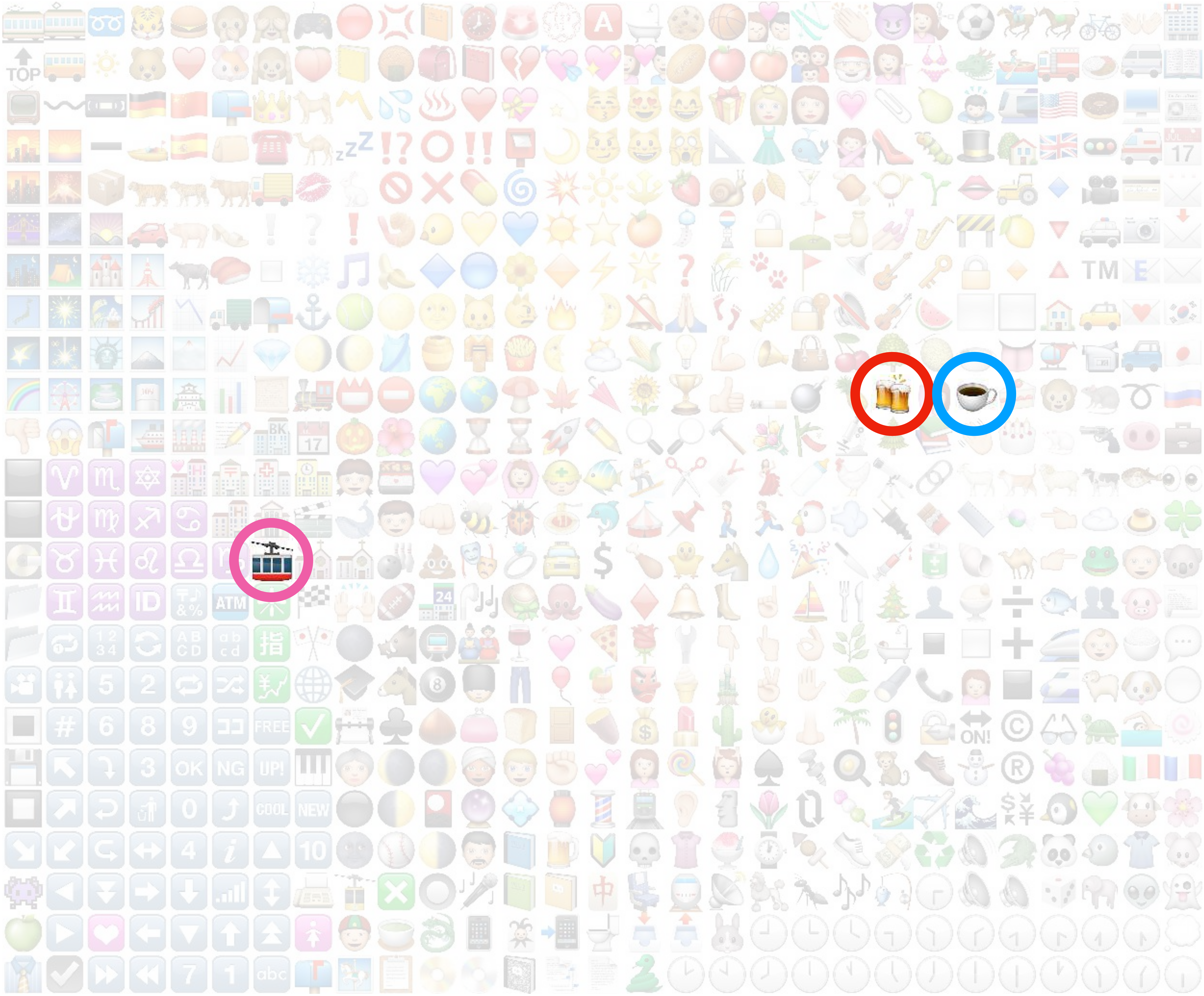
 50%

 0%

 0%


50%
0.5%
0.5%

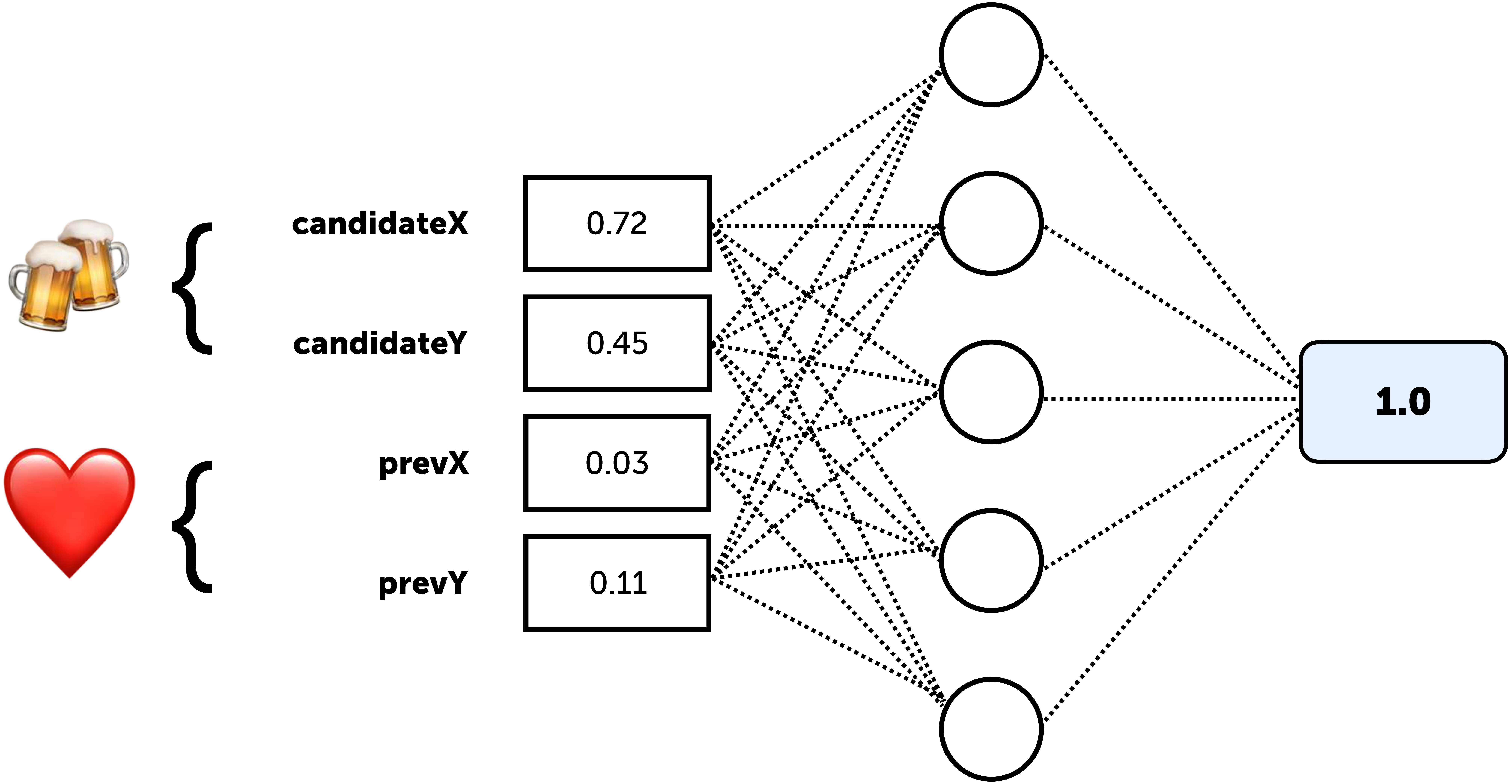


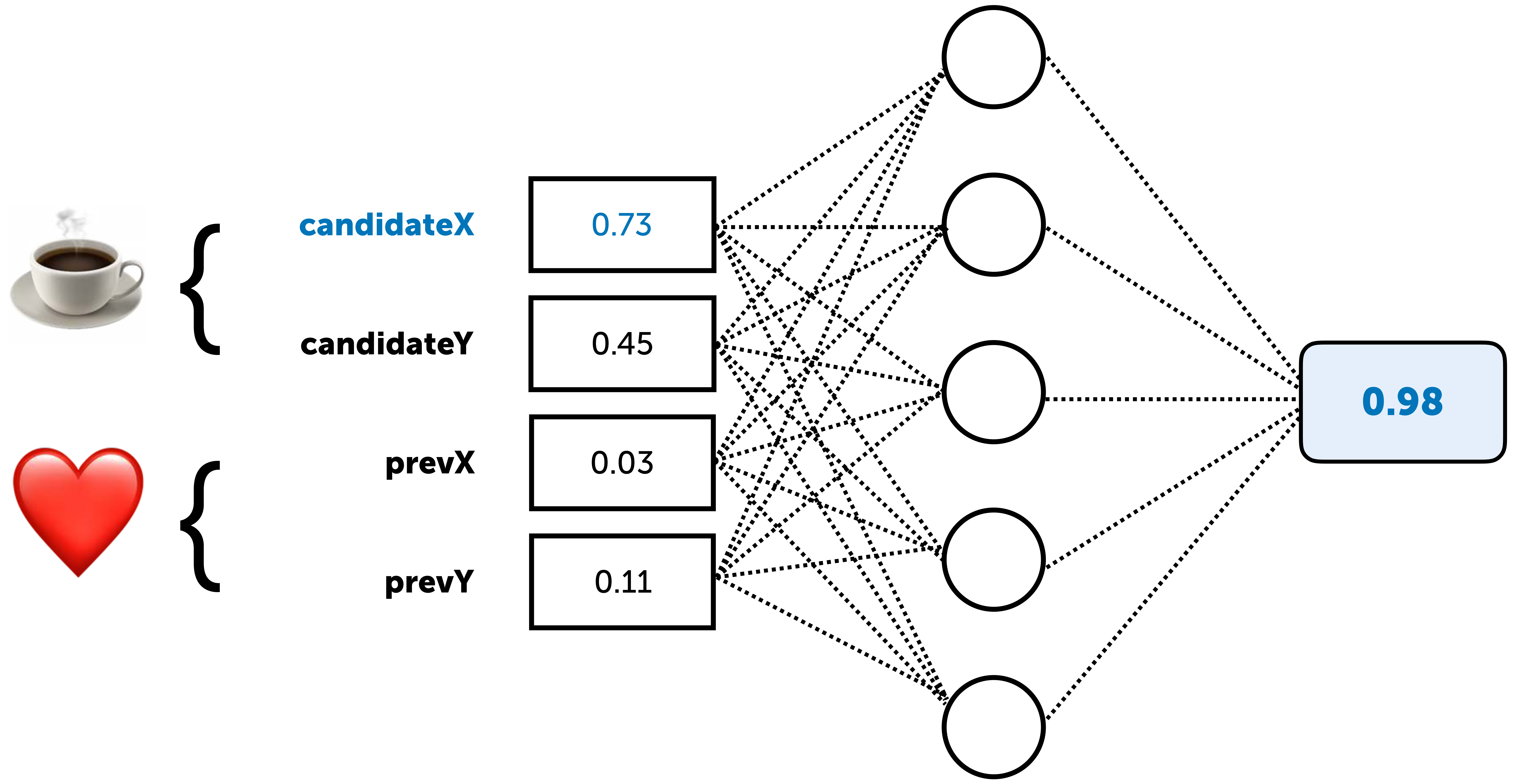


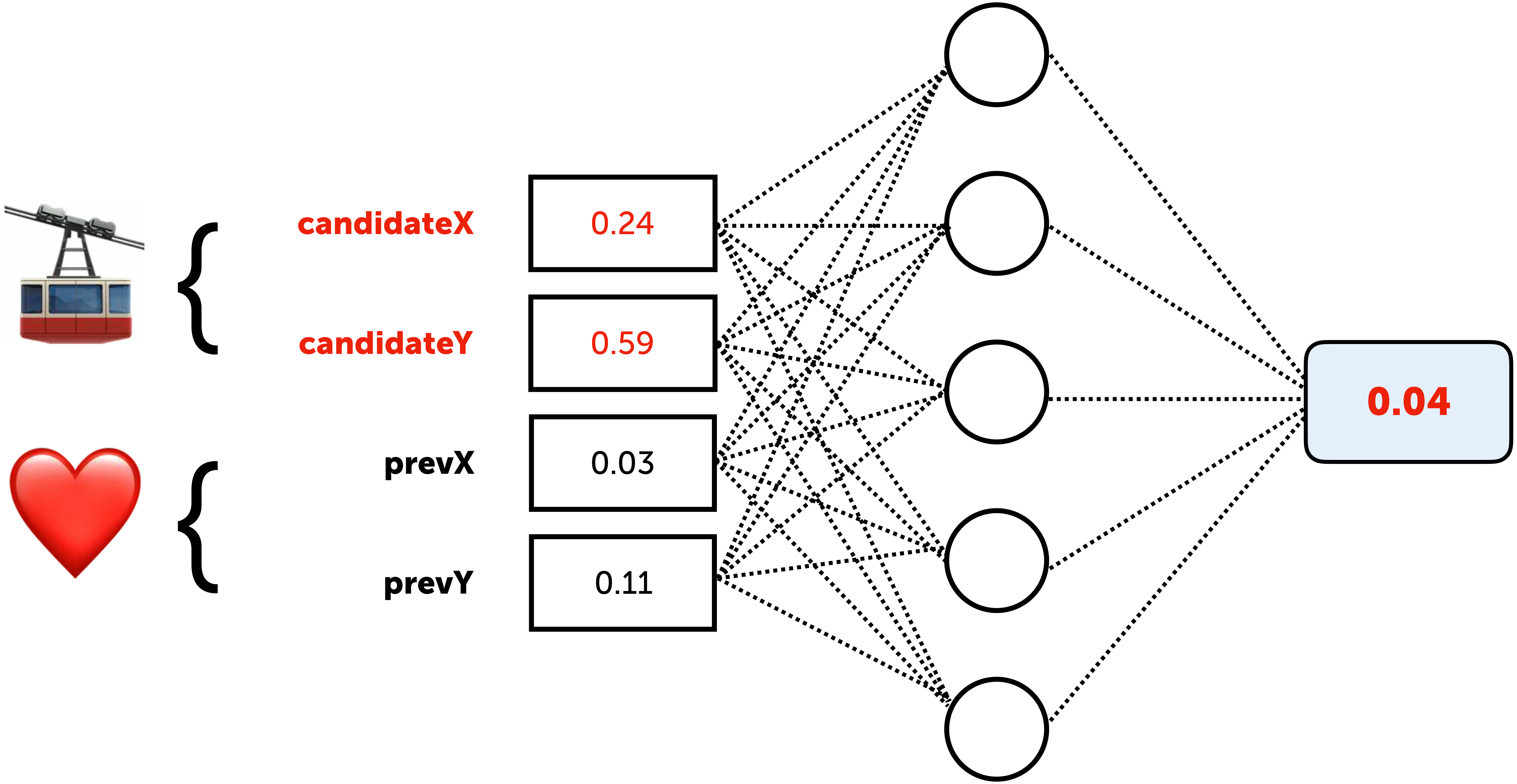
50%

20%

0.5%


candidateX
0.72
candidateY
0.45
prevX
0.03
prevY
0.11
1.0

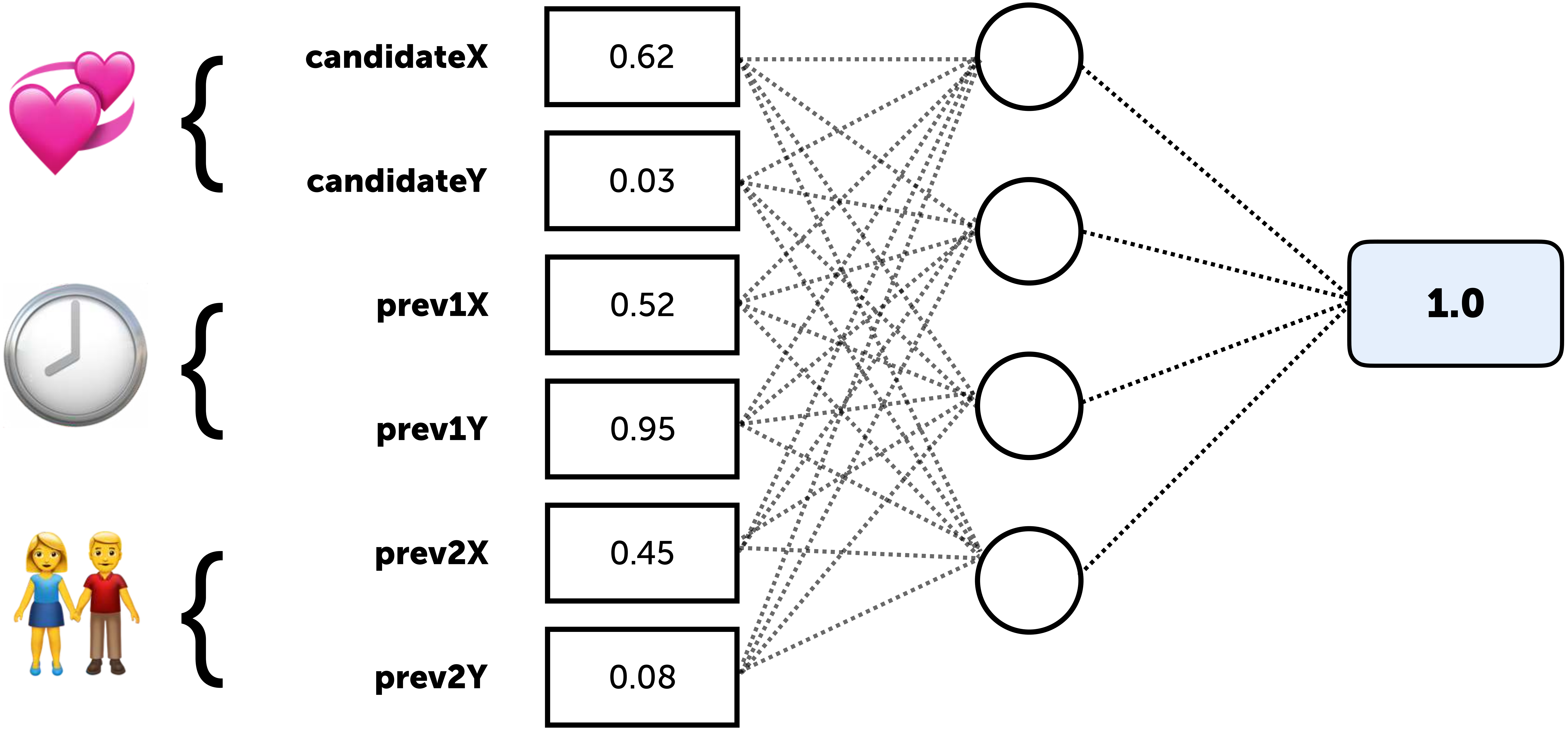

candidateX
0.73
candidateY
0.45
prevX
0.03
prevY
0.11
0.98

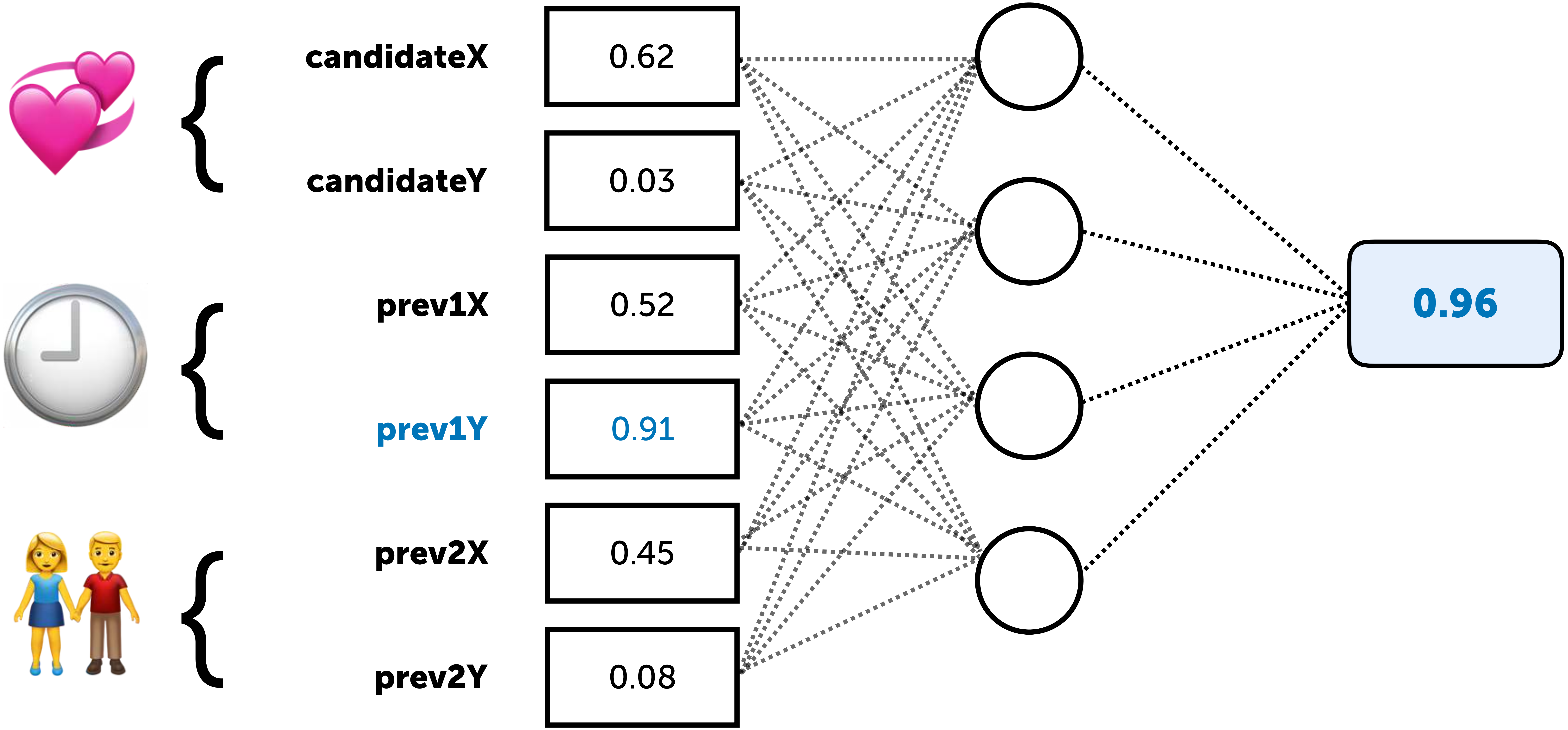


candidateX
candidateY
prevX
prevY
0.24
0.59
0.03
0.11
0.04

1. 👫🕗💞

2. 👫🕘💞


candidateX
0.62
candidateY
0.03
prev1X
0.52
prev1Y
0.95
prev2X
0.45
prev2Y
0.08
1.0


candidateX
0.62
candidateY
0.03
prev1X
0.52
prev1Y
0.91
prev2X
0.45
prev2Y
0.08
0.96

Neural Networks
for Machine Learning

Geoffrey Hinton, UToronto

What about real-world word vectors?

Millions of documents

Terms	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
abs	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
absb	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
absenc	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
absolut	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
absorb	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
abu	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
abus	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
abut	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
academi	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
acceler	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
accept	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0

Docs

Thousands of words

Very sparse

300 dimensions

Thousands of words

abs	0.815,	0.163,	0.803,	0.603,	0.402,	0.333,	0.454,	0.575,	0. ,	0.136,	0.489,
absb	0.511,	0.7 ,	0.501,	0.465,	0.489,	0.171,	0.484,	0.094,	0.919,	0.973,	0.833,
absenc	0.429,	0.155,	0.32 ,	0.29 ,	0.306,	0.313,	0.301,	0.355,	0.55 ,	0.345,	0.325,
absolut	0.003,	0.994,	0.437,	0.468,	0.615,	0.929,	0.103,	0.405,	0.895,	0.37 ,	0.394,
absorb	0.74 ,	0.776,	0.782,	0.802,	0.94 ,	0.651,	0.977,	0.387,	0.373,	0.359,	0.415,
abu	0.685,	0.121,	0.006,	0.764,	0.391,	0.476,	0.236,	0.624,	0.731,	0.117,	0.832,
abus	0.878,	0.966,	0.556,	0.565,	0.451,	0.436,	0.052,	0.397,	0.497,	0.893,	0.364,
abut	0.848,	0.938,	0.85 ,	0.492,	0.575,	0.349,	0.339,	0.756,	0.712,	0.834,	0.15 ,
academi	0.419,	0.977,	0.652,	0.745,	0.292,	0.546,	0.846,	0.342,	0.856,	0.248,	0.33 ,
acceler	0.587,	0.268,	0.384,	0.431,	0.123,	0.565,	0.61 ,	0.976,	0.662,	0.299,	0.591,

Much denser

FastText

300 dimensions

1–2M words

Trained on Wikipedia, web crawls

2–5 GB

Word2Vec

300 dimensions

3M words

Trained on Google News

1.5 GB

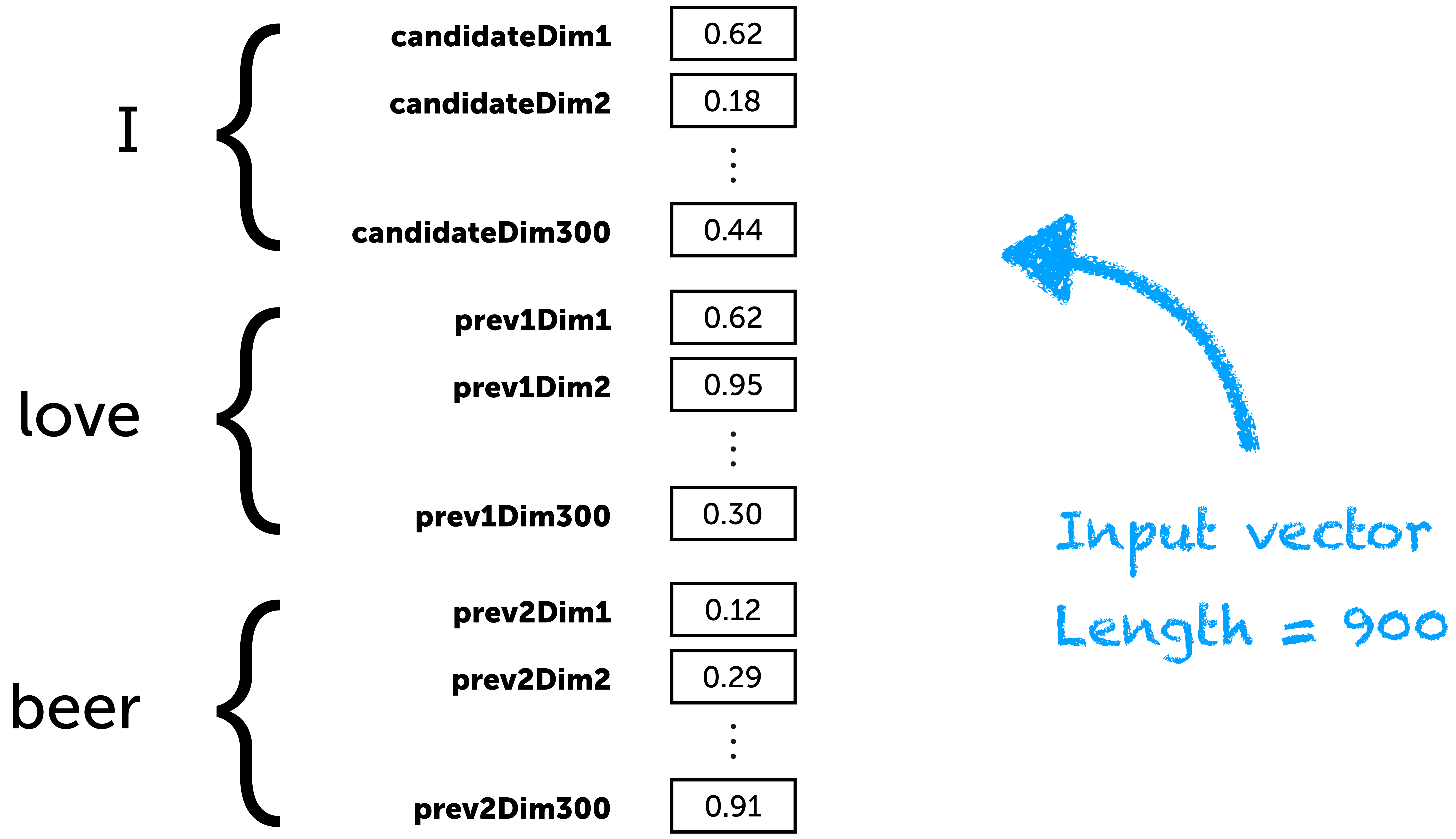
GloVe

50–300 dimensions

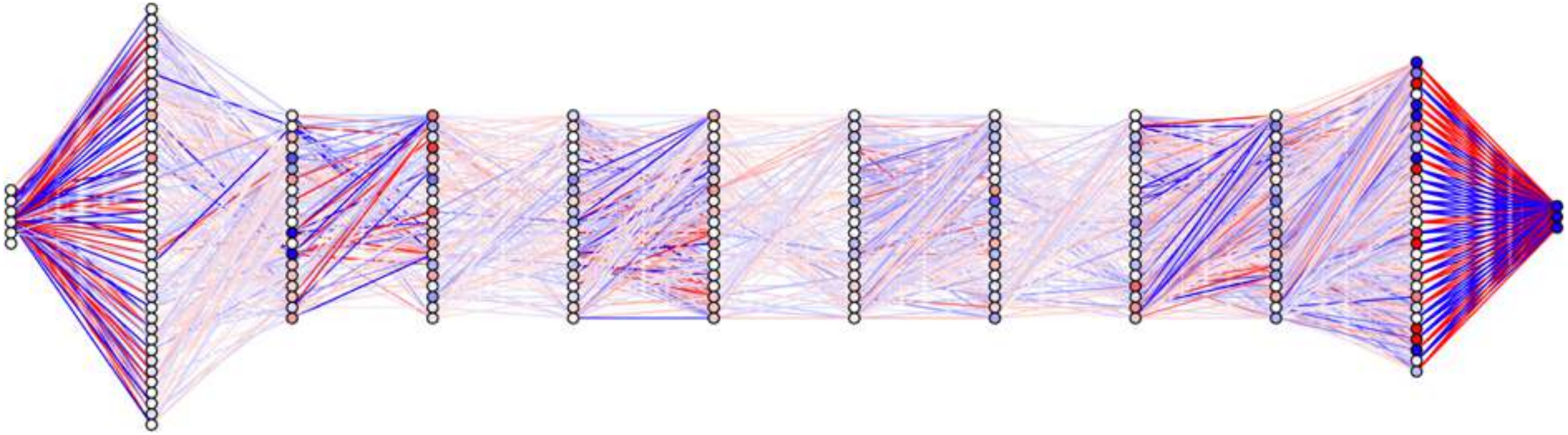
400k–2M words

Trained on Wikipedia, web crawls, Twitter

1–2 GB


I
love
beer
candidateDim1
0.62
candidateDim2
0.18
candidateDim300
0.44
prev1Dim1
0.62
prev1Dim2
0.95
prev1Dim300
0.30
prev2Dim1
0.12
prev2Dim2
0.29
prev2Dim300
0.91
Input vector
Length = 900

2000–2017



LSTMs
RNNs
CNNs
Ensembles

Exploring the limits of language modeling (2016)

5-gram model with Kneser-Ney smoothing

Perplexity score of **67**

2 hours to train (CPU only)

Google's "big" LSTM model

Perplexity score of **30**

…

$64k

$64k   3 weeks!

Exploring the limits of language modeling (2016)

5-gram model with Kneser-Ney smoothing

Perplexity score of **67**

2 hours to train (CPU only)

Google's "big" LSTM model

Perplexity score of **30**

3 weeks to train (32 GPUs)

Gains are very costly

Agenda

Origins of language models

What is unstructured data?

Some case studies

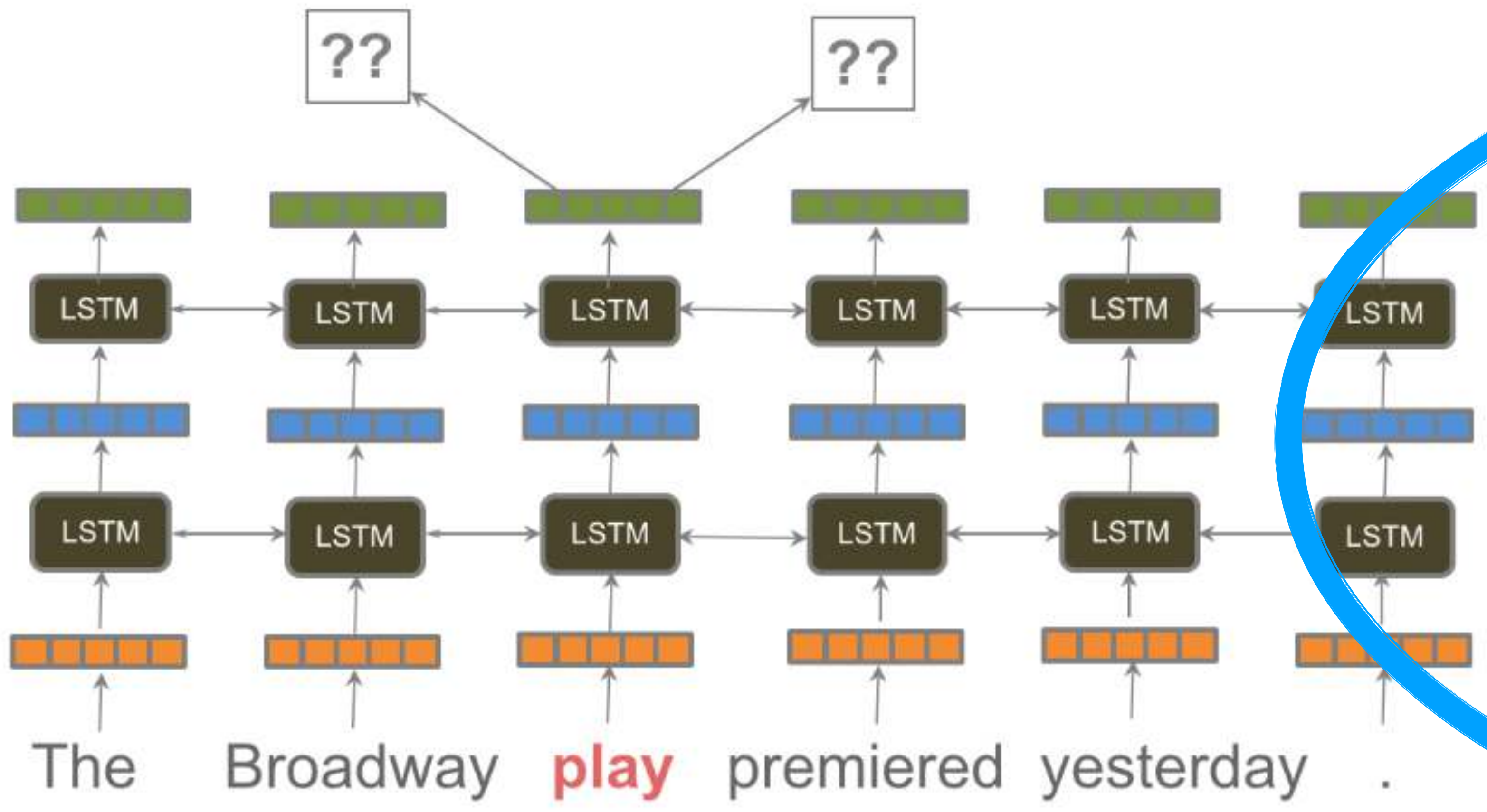
Types of language models

Count based (bag of words, *n*-grams)

Continuous space

Bonus: the class of 2018

Wrap-up and questions



NLP's ImageNet moment has arrived

08.JUL.2018

Big changes are underway in the world of Natural Language Processing (NLP).

The long reign of word vectors as NLP's core representation technique has seen an exciting new line of challengers emerge: ELMo [1], ULMFiT [2], and the OpenAI transformer [3]. These works made headlines by demonstrating that pretrained language models can be used to achieve state-of-the-art results on a wide range of NLP tasks. Such methods herald a watershed moment: they may have the same wide-ranging impact on NLP as pretrained ImageNet models had on computer vision.

From Shallow to Deep Pre-Training

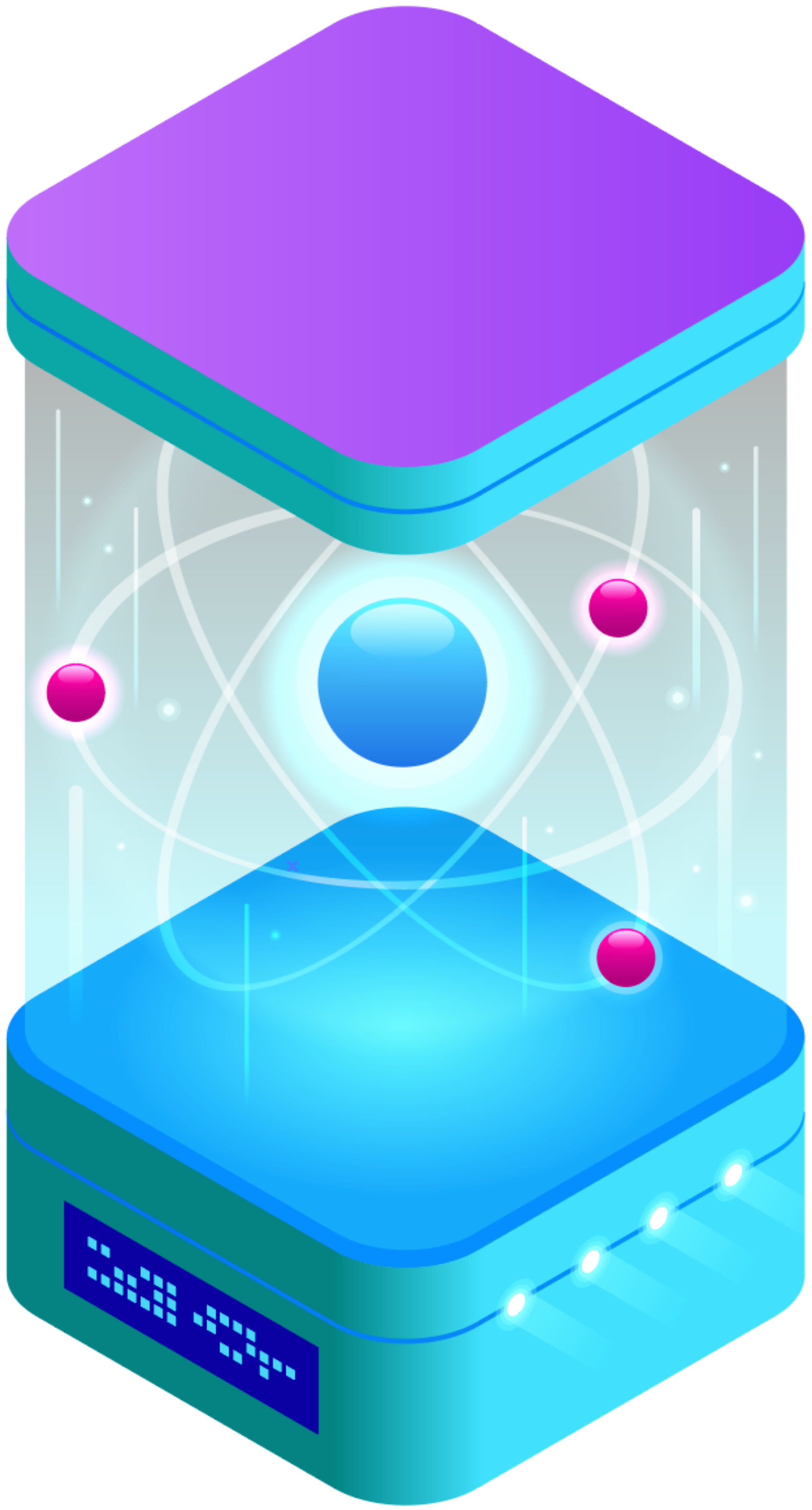
Pretrained word vectors have brought NLP a long way. Proposed in 2013 as an approximation to language modeling, word2vec [4] found adoption through its efficiency and ease of use in a time when hardware was a lot slower and deep learning models were not widely supported. Since then, the standard way of conducting NLP projects has largely remained unchanged: word embeddings pretrained on large amounts of

DB of 14M images, in 20k categories

“Researchers soon realized that the weights learned in state of the art models for ImageNet could be used to initialize models for completely [unrelated] datasets and improve performance significantly”

“Researchers soon realized that the weights learned in state of the art models for ImageNet could be used to initialize models for completely [unrelated] datasets and improve performance significantly”

Transfer learning




LONDON BUS
LONDON BUS
LONDON BUS

ULMFiT

"**U**niversal **L**anguage **M**odel **Fi**ne-**T**uning"

January 2018

Trained on Wikipedia (100M words)

ELMo

"**E**mbeddings from **L**anguage **Mo**dels"

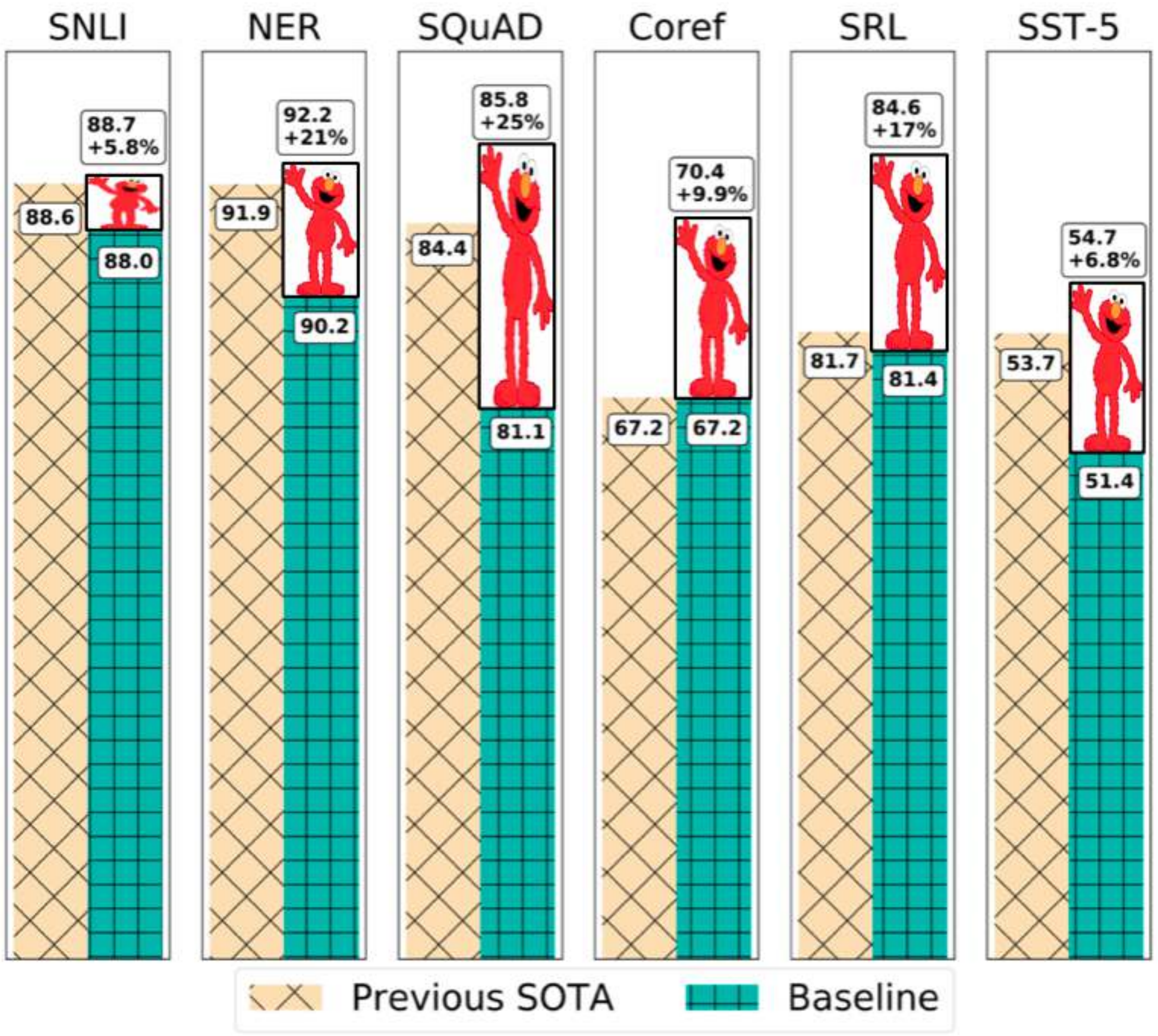
February 2018

Trained on online news (1B words)

OpenAI transformer

June 2018

Trained on 7000 novels (1B words)



Source: Matthew Peters via The Gradient

Executable File | 1662 lines (1661 sloc) | 130 KB

Raw | Blame | History

IMDb

At Fast.ai we have introduced a new module called fastai.text which replaces the torchtext library that was used in our 2018 dl1 course. The fastai.text module also supersedes the fastai.nlp library but retains many of the key functions.

In [1]:
```
from fastai.text import *
import html
```

The Fastai.text module introduces several custom tokens.

We need to download the IMDB large movie reviews from this site: http://ai.stanford.edu/~amaas/data/sentiment/ Direct link : Link and untar it into the PATH location. We use pathlib which makes directory traveral a breeze.

In [2]:
```
BOS = 'xbos'  # beginning-of-sentence tag
FLD = 'xfld'  # data field tag

PATH=Path('data/aclImdb/')
```

Standardize format

In [3]:
```
CLAS_PATH=Path('data/imdb_clas/')
CLAS_PATH.mkdir(exist_ok=True)

LM_PATH=Path('data/imdb_lm/')
LM_PATH.mkdir(exist_ok=True)
```

The imdb dataset has 3 classes. positive, negative and unsupervised(sentiment is unknown). There are 75k training reviews(12.5k pos, 12.5k neg, 50k unsup) There are 25k validation reviews(12.5k pos, 12.5k neg & no unsup)

Refer to the README file in the imdb corpus for further information about the dataset.

In [122]:
```
CLASSES = ['neg', 'pos', 'unsup']

def get_texts(path):
    texts,labels = [],[]
    for idx,label in enumerate(CLASSES):
        for fname in (path/label).glob('*.*'):
            texts.append(fname.open('r', encoding='utf-8').read())
            labels.append(idx)
    return np.array(texts),np.array(labels)

trn_texts,trn_labels = get_texts(PATH/'train')
val_texts,val_labels = get_texts(PATH/'test')
```

In [123]:
```
len(trn_texts),len(val_texts)
```

Out[123]:
```
(75000, 25000)
```

In [124]:
```
col_names = ['labels','text']
```

We use a random permutation np array to shuffle the text reviews.

In [125]:
```
np.random.seed(42)
trn_idx = np.random.permutation(len(trn_texts))
val_idx = np.random.permutation(len(val_texts))
```

In [126]:
```
trn_texts = trn_texts[trn_idx]
val_texts = val_texts[val_idx]

trn_labels = trn_labels[trn_idx]
```

Fast.ai’s IMDb tutorial

Walkthrough to achieve a new state-of-the-art (**95.4%** accuracy)

https://bit.ly/imdblm

TensorFlow Hub

```
elmo = hub.Module("https://tfhub.dev/google/elmo/2", trainable=True)
```

Wrap-up

- Unstructured data is a powerful and plentiful source of insight (90%+ of all data), especially if combined with language models.
- Unstructured data is more difficult to work with in some ways: we have to deal with soft rules, noise and thus unclear results.
- *But* it can also give us richer insights, allow for easier post-hoc analysis, and provide a cheaper alternative to collecting structured datasets.

Wrap-up

- Count-based models are fast and still work well. Interpolated Kneser-Ney *n*-gram models generally work the best.
- For even better results, use pre-trained word vectors (from Google/Facebook) and a neural net
- For the current state-of-the-art, experiment with a universal language model, tuned to your particular task

Understanding Unstructured Data with Language Models

Alex Peattie

Grab the slides

alexpeattie.com/talks