

A Brief Overview of Sublinear Matrix Approximation Techniques

UMassAmherst

Manning College of Information
& Computer Sciences

Matrices in the Wild

Distance matrices

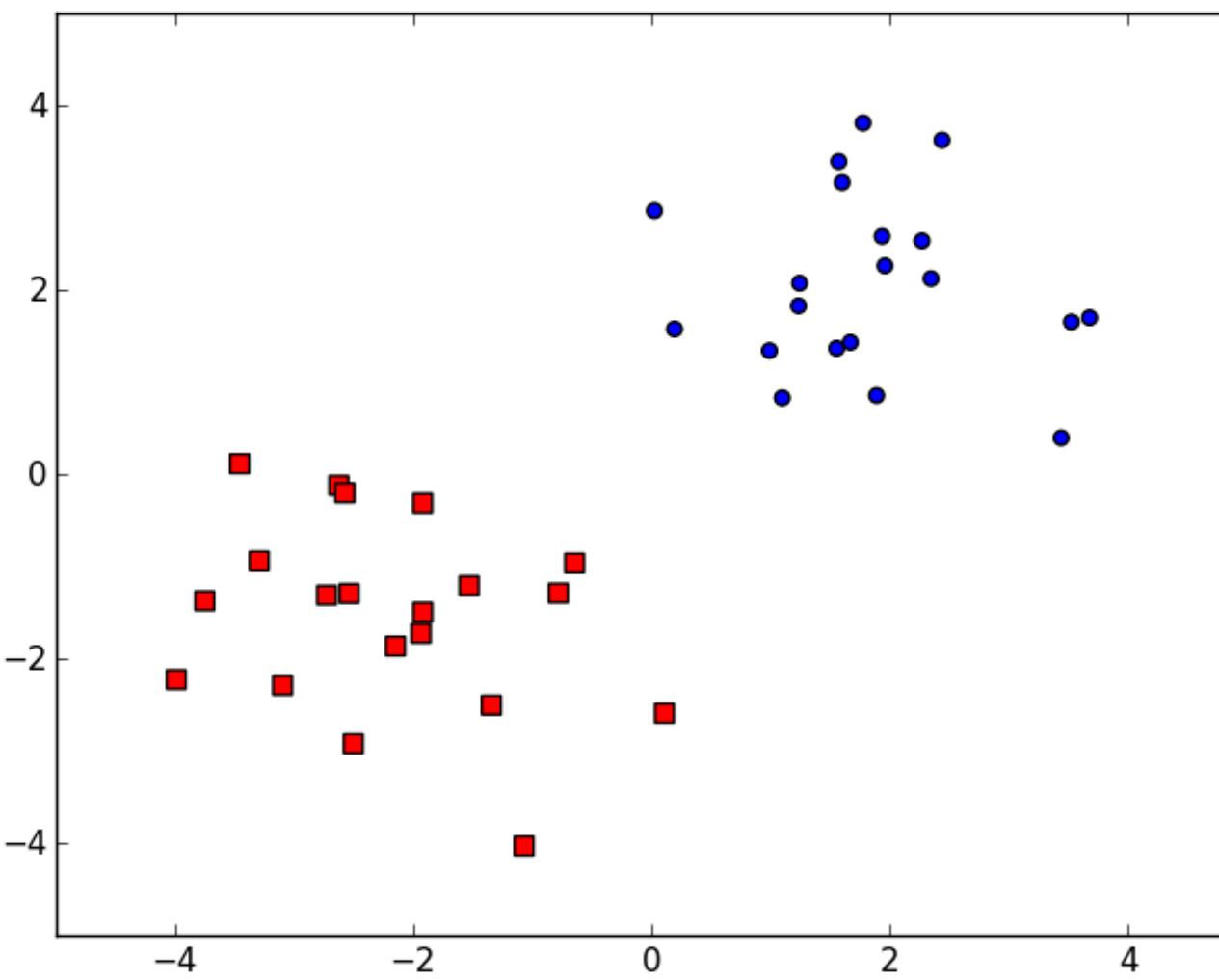


Image source: <http://blog.sairahul.com/2014/01/linear-separability.html>

Matrices in the Wild

Similarity matrices

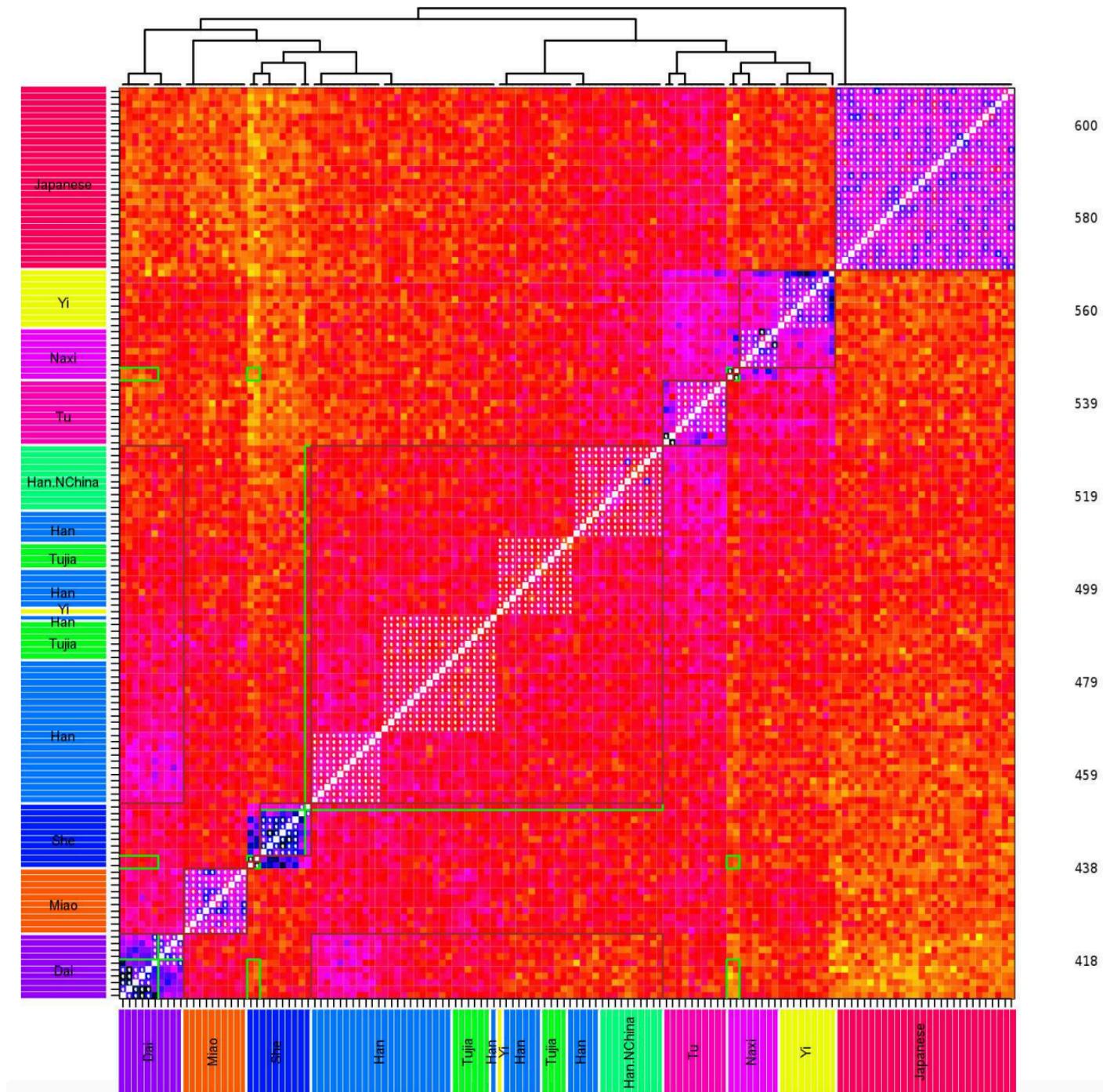


Image source: <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.727.7925&rep=rep1&type=pdf>

Matrices in the Wild

Kernel matrices

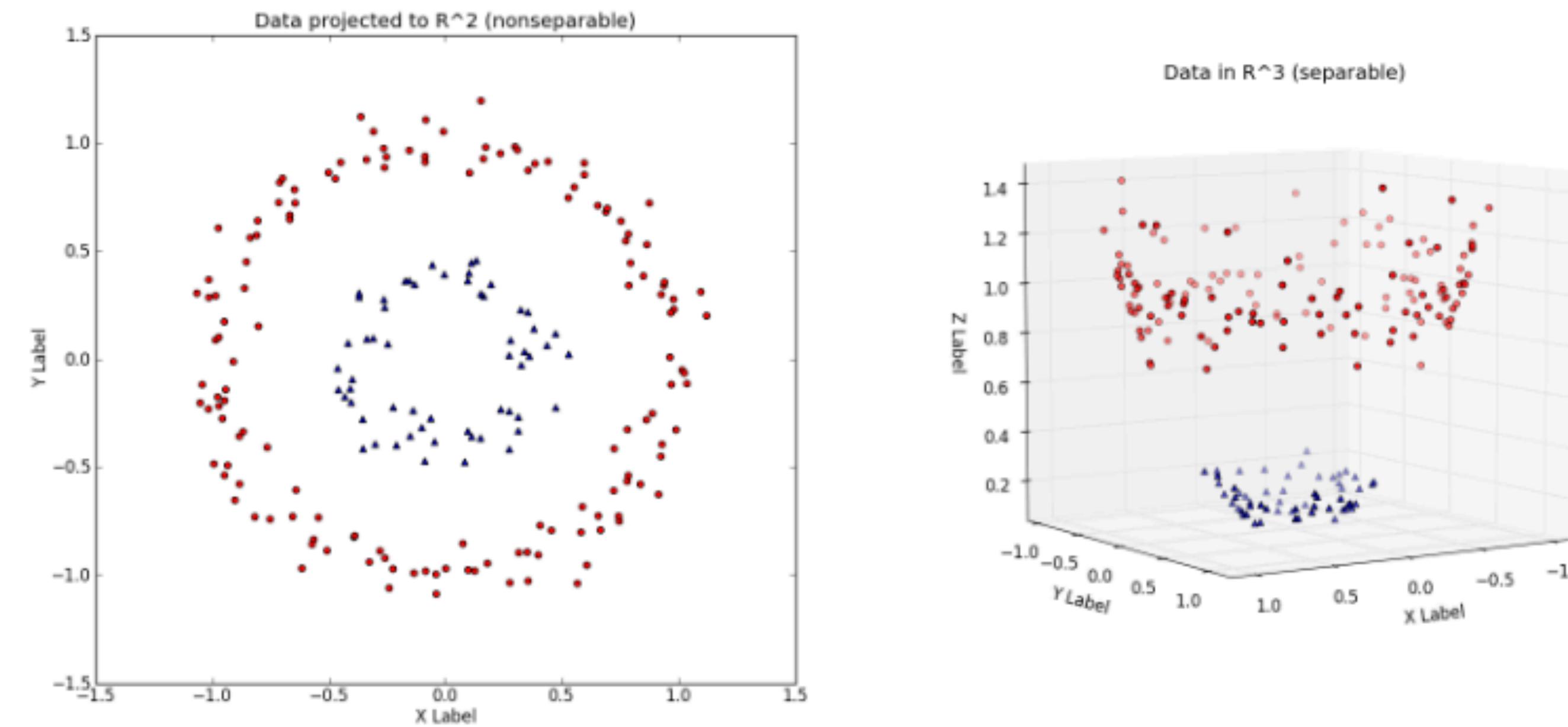


Image source: http://www.eric-kim.net/eric-kim-net/posts/1/kernel_trick.html

Define implicit mapping function $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^D$ such that $\exists K(x, y) = \langle \phi(x), \phi(y) \rangle$ is easy to compute

Computational Cost

$\Omega(n^2)$ comparisons among datapoints

Expensive comparisons between documents results in computational bottleneck

Computational Cost

$\Omega(n^2)$ comparisons among datapoints

Expensive comparisons between documents results in computational bottleneck

Can we approximate these in sublinear time without losing much information

Sublinear Methods

Define implicit mapping function $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^D$ such that $\exists K(x, y) = \langle \phi(x), \phi(y) \rangle$ is easy to compute

Random features approximations of kernel matrices [1]

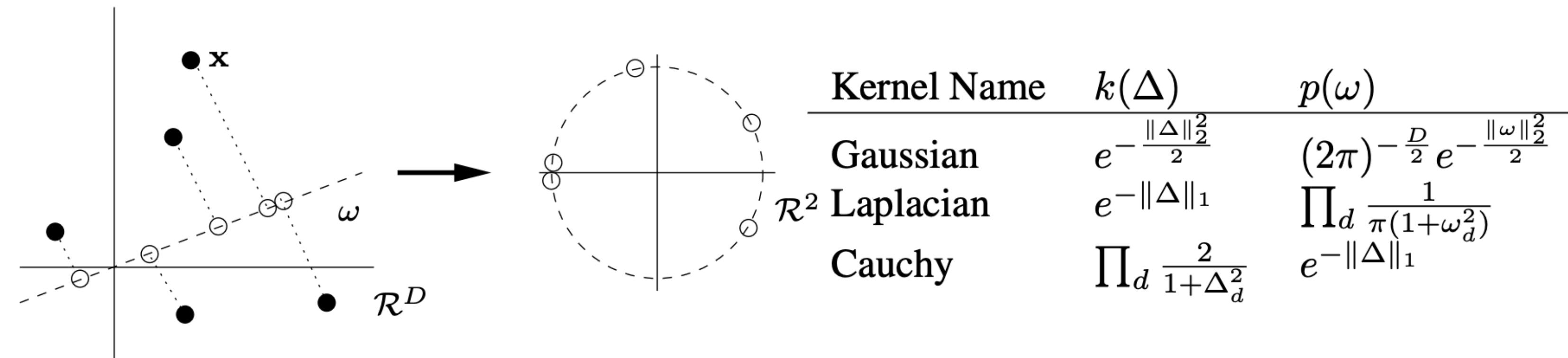
Works on shift invariant kernel's: $K(x, y) = K(x - y)$

[1] Rahimi, A. and Recht, B., 2007. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20

Sublinear Methods

Define implicit mapping function $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^D$ such that $\exists K(x, y) = \langle \phi(x), \phi(y) \rangle$ is easy to compute

Random features approximations of kernel matrices [1]



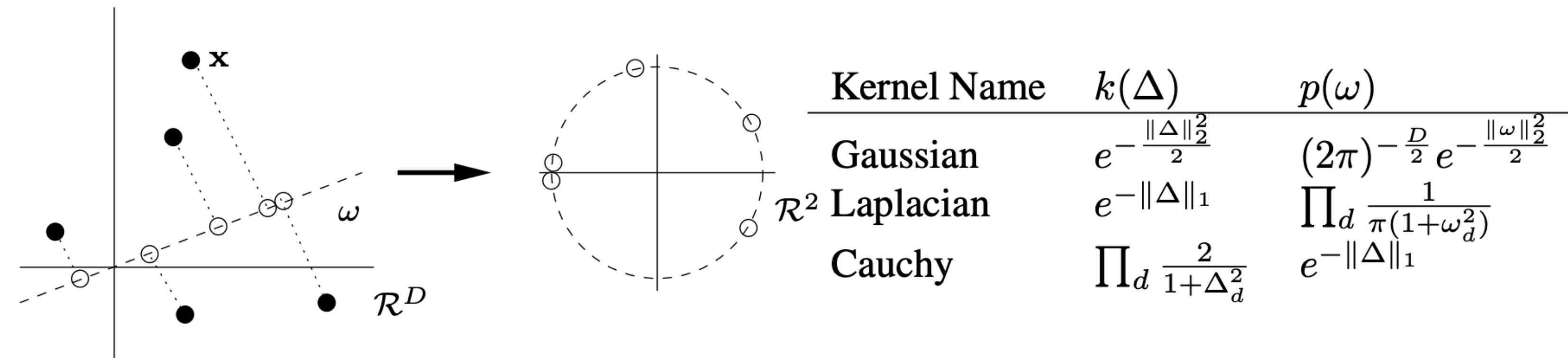
Transparent dots are $z(x)$ and $K(x, y) \approx \langle z(x), z(y) \rangle$

[1] Rahimi, A. and Recht, B., 2007. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20

Sublinear Methods

Define implicit mapping function $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^D$ such that $\exists K(x, y) = \langle \phi(x), \phi(y) \rangle$ is easy to compute

Random features approximations of kernel matrices [1]



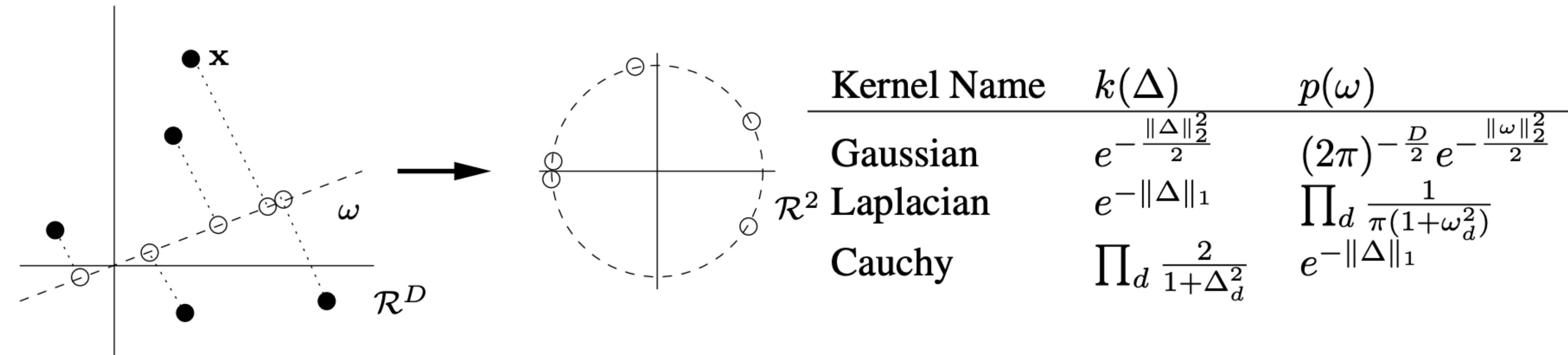
We can store $Z \in \mathbb{R}^{N \times R}$ where $R \ll N$

[1] Rahimi, A. and Recht, B., 2007. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20

Sublinear Methods

Define implicit mapping function $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^D$ such that $\exists K(x, y) = \langle \phi(x), \phi(y) \rangle$ is easy to compute

Random features approximations of kernel matrices [1]



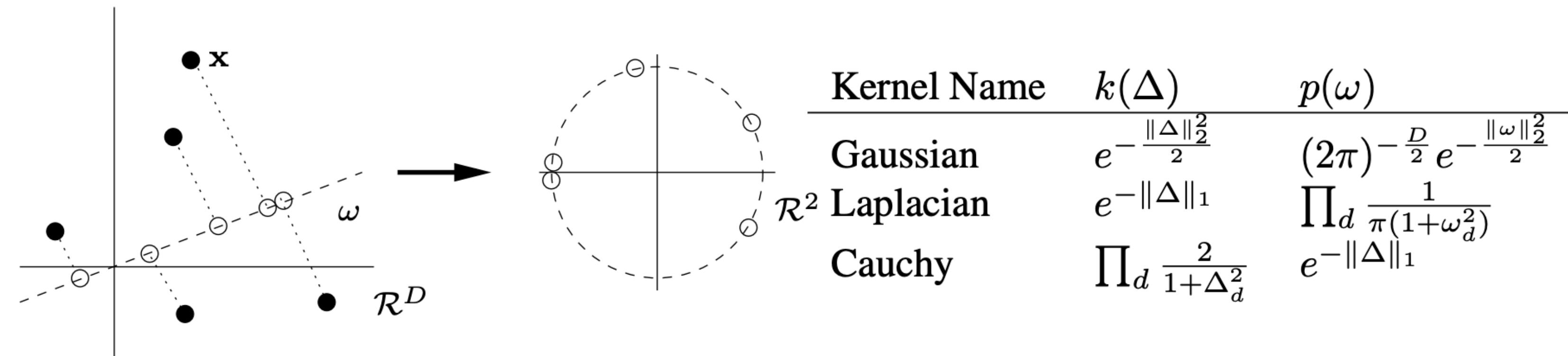
Prediction in say linear regression then becomes: $\beta = (Z^T Z)^{-1} Z^T Y$

[1] Rahimi, A. and Recht, B., 2007. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20

Sublinear Methods

Define implicit mapping function $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^D$ such that $\exists K(x, y) = \langle \phi(x), \phi(y) \rangle$ is easy to compute

Random features approximations of kernel matrices [1]



Matrix inversion is $O(NR^2)$ instead of $O(N^3)$

[1] Rahimi, A. and Recht, B., 2007. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20

Sublinear Methods

Nyström approximation [1]

For any matrix $K \in \mathbb{R}^{n \times n}$ compute approximation as $KS(S^T K S)^+ S^T K$

$S \in \mathbb{R}^{n \times s}$ samples columns of K at random

[1] Nyström, E.J., 1930. Über die praktische Auflösung von Integralgleichungen mit Anwendungen auf Randwertaufgaben. *Acta Mathematica*, 54, pp.185-204.

Sublinear Methods

Nyström approximation [1]

For any matrix $K \in \mathbb{R}^{n \times n}$ compute approximation as $KS(S^T K S)^+ S^T K$

$S \in \mathbb{R}^{n \times s}$ samples columns of K at random

Uniform sampling may fail in natural datasets where relative importance of data is not uniform

[1] Nyström, E.J., 1930. Über die praktische Auflösung von Integralgleichungen mit Anwendungen auf Randwertaufgaben. *Acta Mathematica*, 54, pp.185-204.

Sublinear Methods

Nyström approximation

For any matrix $K \in \mathbb{R}^{n \times n}$ compute approximation as $KS(S^T K S)^+ S^T K$

$S \in \mathbb{R}^{n \times s}$ samples columns of K at random

We can then use alternate sampling methods like leverage scores [1,2] to choose S

- [1] Cohen, M.B., Musco, C. and Pachocki, J., 2016. Online row sampling. *arXiv preprint arXiv:1604.05448*.
- [2] Musco, C. and Musco, C., 2017. Recursive sampling for the nyström method. *Advances in Neural Information Processing Systems*, 30.

Sublinear Methods

Pseudoskeleton approximation [1]

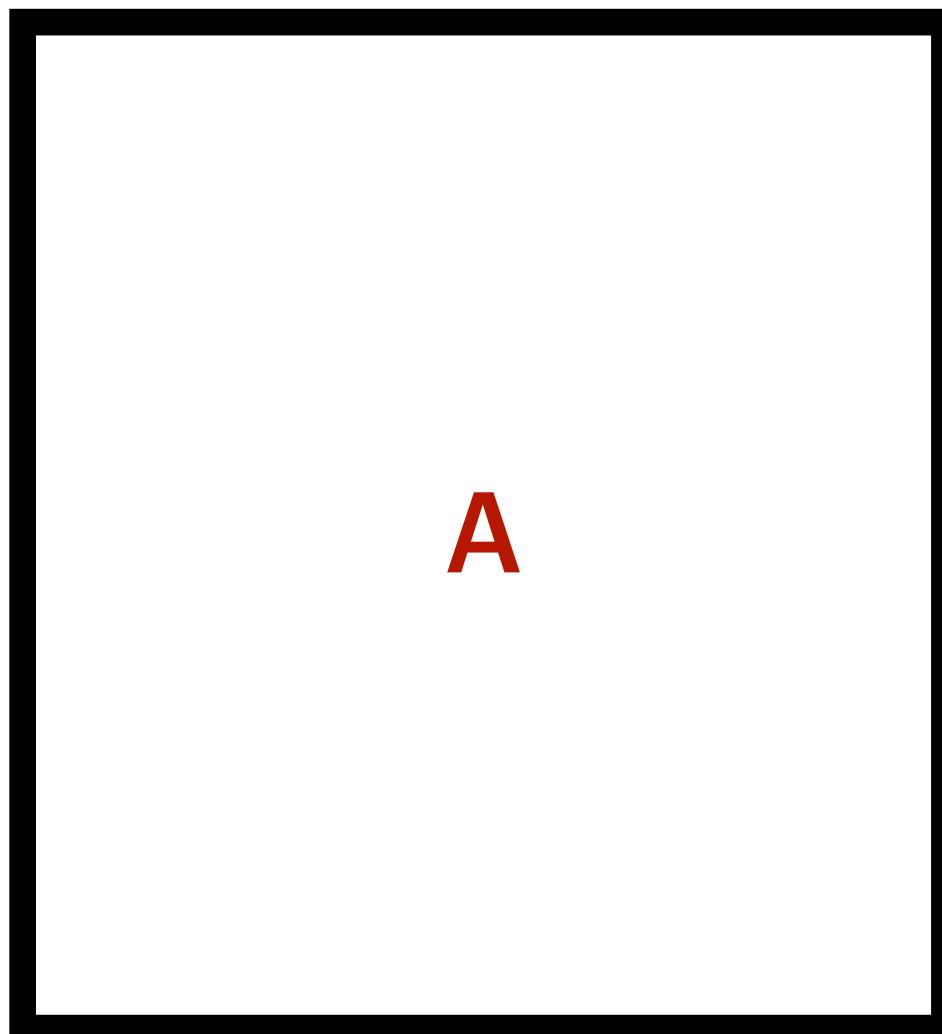
For any matrix $K \in \mathbb{R}^{n \times n}$ compute approximation as $KS_2(S_2^T KS_1)^+ S_1^T K$

$S_1, S_2 \in \mathbb{R}^{n \times s}$ samples columns of K at random

[1] Goreinov, S.A., Tyrtyshnikov, E.E. and Zamarashkin, N.L., 1997. A theory of pseudoskeleton approximations. *Linear algebra and its applications*, 261(1-3), pp.1-21.

Sublinear Methods

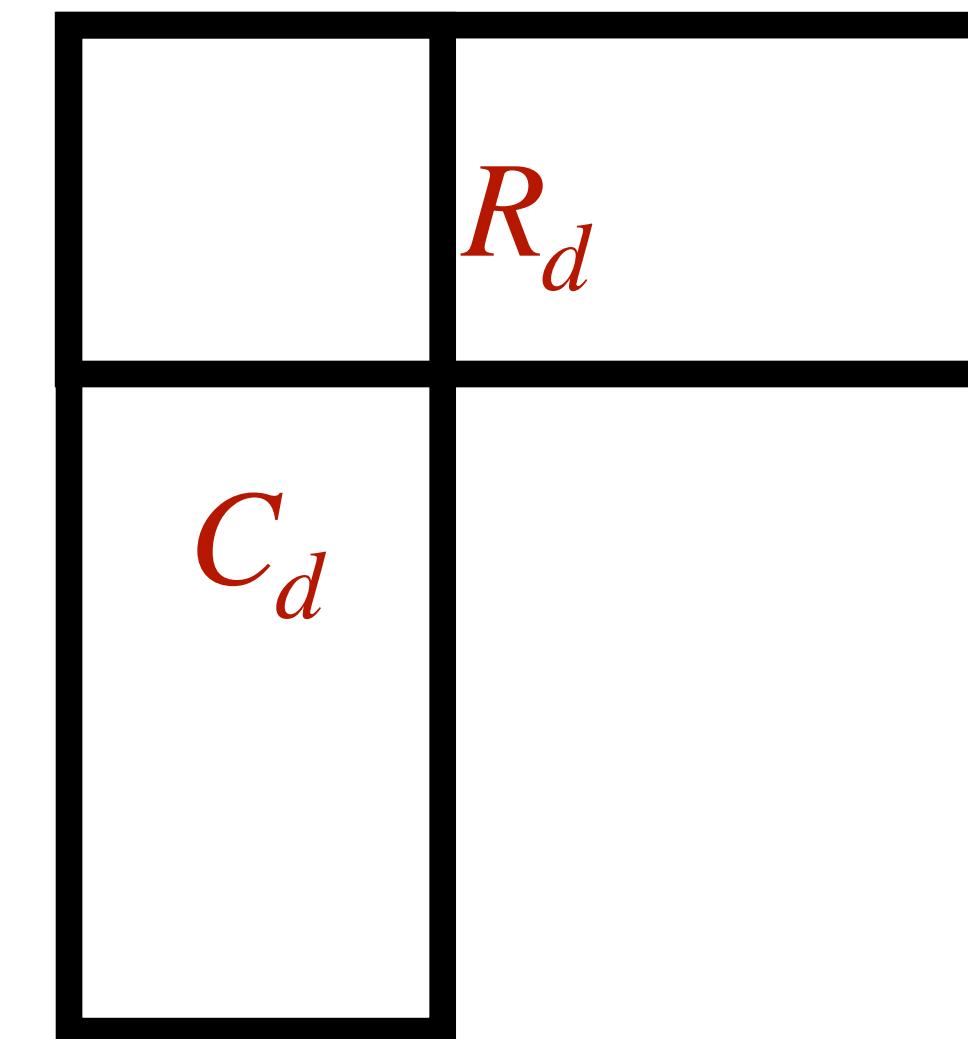
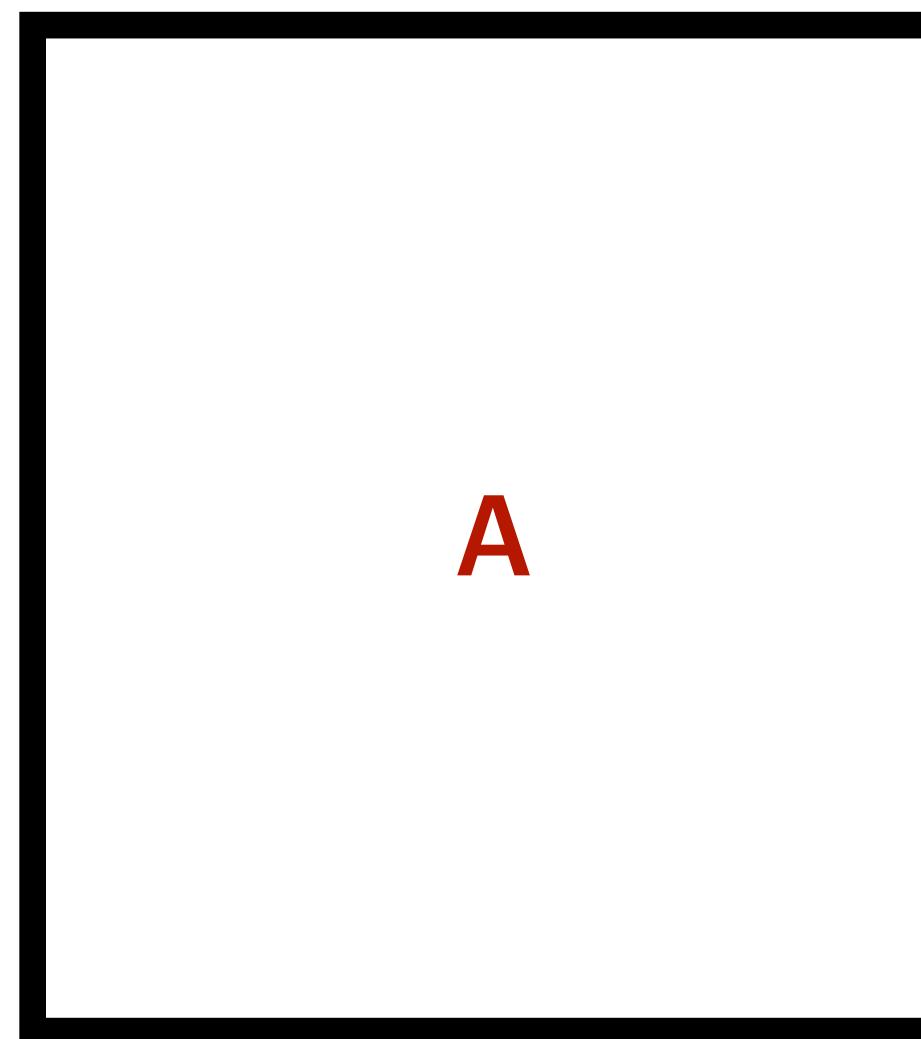
CUR [1]



[1] Mahoney, M.W. and Drineas, P., 2009. CUR matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, 106(3), pp.697-702.

Sublinear Methods

CUR [1]

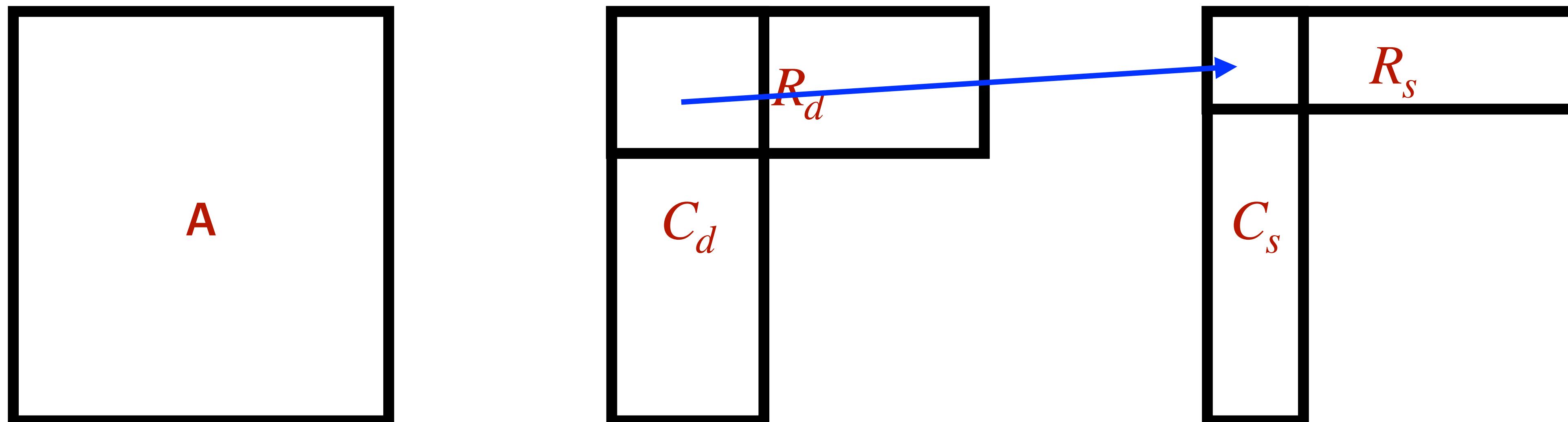


[1] Mahoney, M.W. and Drineas, P., 2009. CUR matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, 106(3), pp.697-702.

Sublinear Methods

CUR [1]

Construct small U

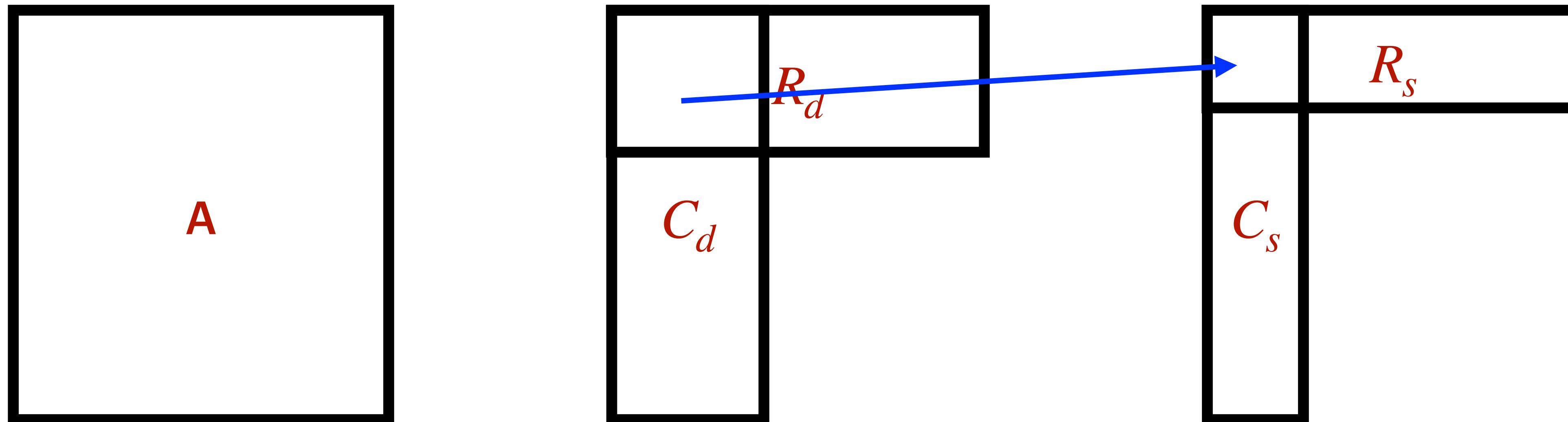


[1] Mahoney, M.W. and Drineas, P., 2009. CUR matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, 106(3), pp.697-702.

Sublinear Methods

CUR [1]

Construct small U



For multiple copies keep one column/row but multiply it by \sqrt{d}

[1] Mahoney, M.W. and Drineas, P., 2009. CUR matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, 106(3), pp.697-702.

Sublinear Time Approximation of Text Similarity Matrices

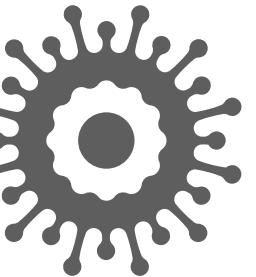
**Archana Ray, Nicholas Monath, Andrew McCallum,
Cameron Musco**

UMassAmherst

Manning College of Information
& Computer Sciences



A Rapid and Specific Assay Detection of MERS-CoV



Amazingly, it is effective against SARS and MERS.

Pai Huang^{1,2}, Hualei Wang^{2,3,4*}, Zengguo Cao^{2,3}, Hongli Jin^{2,3}, Hang Beibei Yu¹, Feihu Yan², Xingxing Hu^{1,2}, Fangfang Wu², Cuicui Jiao⁴, Shengnan Xu^{1,2}, Yongkun Zhao^{2,4}, Na Feng^{2,4}, Jianzhong Wang¹, Wei Tiecheng Wang^{2,4}, Yuwei Gao^{2,4}, Songtao Yang^{2,4} and Xianzhu Xia²

¹ Animal Science and Technology College, Jilin Agricultural University, Changchun, China; ² Key Laboratory of Jilin Province for Zoonotic Prevention and Control, Institute of Military Veterinary, Academy of Military Medical Sciences, Changchun, China; ³ College of Veterinary Medicine, Jilin University, Changchun, China; ⁴ Jiangsu Co-innovation Center for Prevention and Control of Important Animal Infectious Diseases and Zoonoses, Yangzhou, China; ⁵ State Key Laboratory of Respiratory Disease, Guangzhou Institute of Respiratory Health, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou, China; ⁶ Guangzhou Eighth People's Hospital of Guangzhou Medical University, Guangzhou, China; ⁷ Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

OPEN ACCESS

Middle East respiratory syndrome coronavirus (MERS-CoV) is a novel human coronavirus that can cause severe disease. We developed a rapid and specific method for this virus that can be used in clinical settings.

Vaccine 35 (2017) 2058–2075



Contents lists available at ScienceDirect



Vaccine

journal homepage: www.elsevier.com/locate/vaccine

DNA vaccine encoding Middle East respiratory syndrome coronavirus S1 protein induces protective immune responses in mice

Hang Chi^a, Xueying Zheng^{a,b}, Xiwen Wang^a, Chong Wang^a, Hualei Wang^{a,c}, Weiwei Gai^a, Stanley Perlman^d, Songtao Yang^{a,c,e}, Jincun Zhao^{c,g}, Xianzhu Xia^{a,d,f}

^a Key Laboratory of Jilin Province for Zoonotic Prevention and Control, Institute of Military Veterinary, Academy of Military Medical Sciences, Changchun, China

^b School of Public Health, Shandong University, Jinan, China

^c Jiangsu Co-innovation Center for Prevention and Control of Important Animal Infectious Diseases and Zoonoses, Yangzhou, China

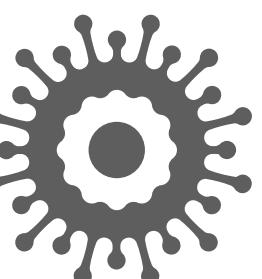
^d Department of Microbiology, University of Colorado, Boulder, CO, United States

^e State Key Laboratory of Respiratory Disease, Guangzhou Institute of Respiratory Health, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou, China

^f Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^g Specialty section:

This article was submitted to



The Middle East respiratory syndrome coronavirus (MERS-CoV) is an emerging pathogen...

specific CD4⁺ and CD8⁺ T cells secreting IFN- γ and other cytokines. Both pcDNA3.1-S1 DNA vaccine immunization and passive transfer of immune serum from pcDNA3.1-S1 vaccinated mice protected Ad5-hDPP4-transduced mice from MERS-CoV challenge. These results demonstrate that a DNA vaccine encoding MERS-CoV S1 protein induces strong protective immune responses against MERS-CoV infection.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

Middle East respiratory syndrome (MERS)-coronavirus (MERS-CoV), an emerging zoonotic virus, is the causative agent of MERS. MERS-CoV was first identified in Saudi Arabia in 2012 and MERS cases have been reported in 27 countries since then [1,2]. As of February 10, 2017, 1905 laboratory-confirmed cases, including 677 deaths related to MERS-CoV, had been reported to WHO (~36% mortality). Several family clusters and nosocomial clusters cases have been reported, revealing the human-to-human transmissibility of MERS-CoV, and raising the concern of a MERS-CoV global pandemic [3–5]. Currently, no licensed therapeutic or vaccine is available, which highlights the need for efficient vaccines against MERS-CoV.

To date, several vaccine candidates have been developed, such as viral vector-based recombinants [6–11], subunit vaccines [12–19], DNA vaccines [20], RNA prime/protein-boost vaccines [21] and a reverse genetics-constructed recombinant coronavirus vaccine [22]. Among them, DNA vaccines present a range of unique advantages such as proper antigen protein folding, rapid design and production, cost-effectiveness, and stability at non-refrigerated temperatures for convenient storage and shipping [23]. Furthermore, it has been reported that DNA vaccines can induce both humoral and cellular immune responses against MERS-CoV and SARS-CoV infection [20,24,25].

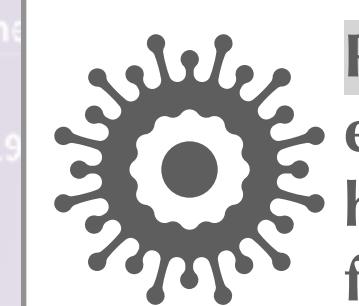
MERS-CoV is the first lineage of Betacoronavirus known to infect humans [26]. The genome of MERS-CoV encodes four structural proteins – spike (S), envelope (E), membrane (M) and nucleocapsid (N) [27]. The S protein, a class I fusion protein forming protruding

Unexpected outbreaks of arbovirus infections: lessons learned from the Pacific and tropical America

Didier Musso, MD • Prof Alfonso J Rodriguez-Morales, MD • José Eduardo Levi, PhD [†]

Van-Mai Cao-Lormeau, MD

Published: June 19, 2018



Pandemic arboviruses have emerged as a major global health problem in the past four decades.

Check for updates

Summary

Pandemic arboviruses have emerged as a major global health

decades. Predicting where and when the will occur is a challenge, but history back swan events (epidemics that are at have an extreme effect) will continue to and globalisation expand. We briefly review epidemics that have occurred in the past 50 the American and Pacific regions, to ability, and to highlight the need for dness, including laboratory-based

COLORADO TICK FEVER

Richard Klasco, MD

TICK-BORNE DISEASES

0025-7125/02 \$15.00 + .00

Colorado tick fever (CTF), also known as mountain fever and mountain tick fever, is a well-described viral tick-borne disease common to the Rocky Mountains. It describes the clinical presentation and diagnosis,

CAUSE AND PATHOGENESIS

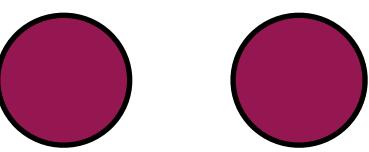
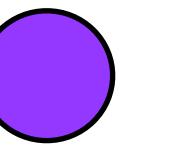
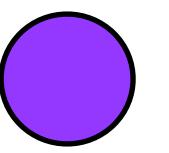
Colorado tick fever virus (CTFV)

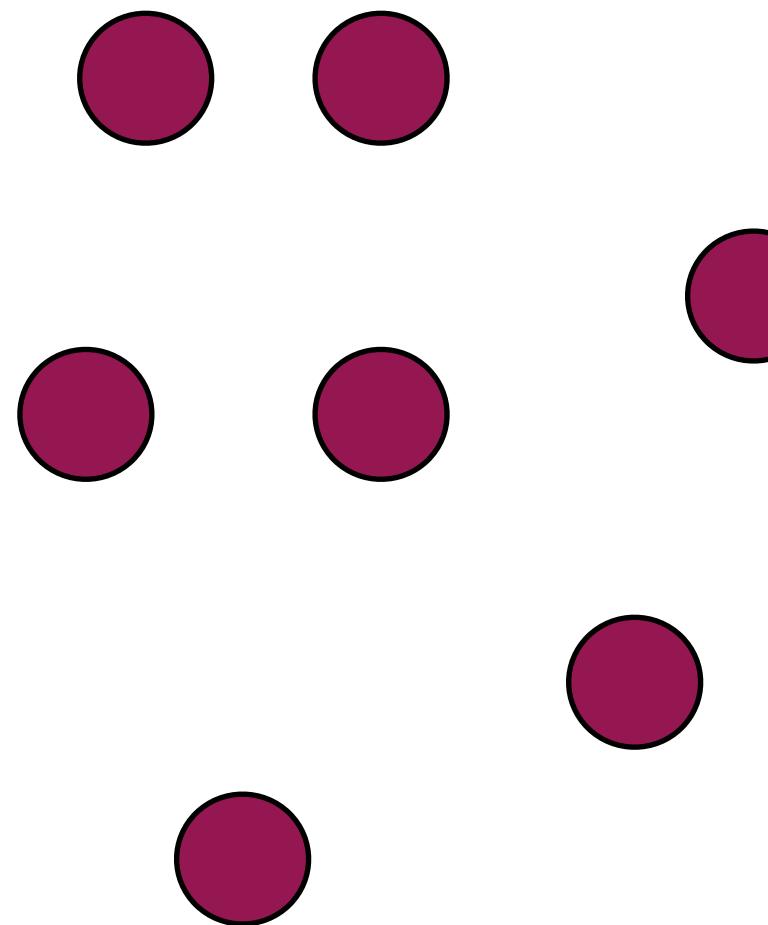
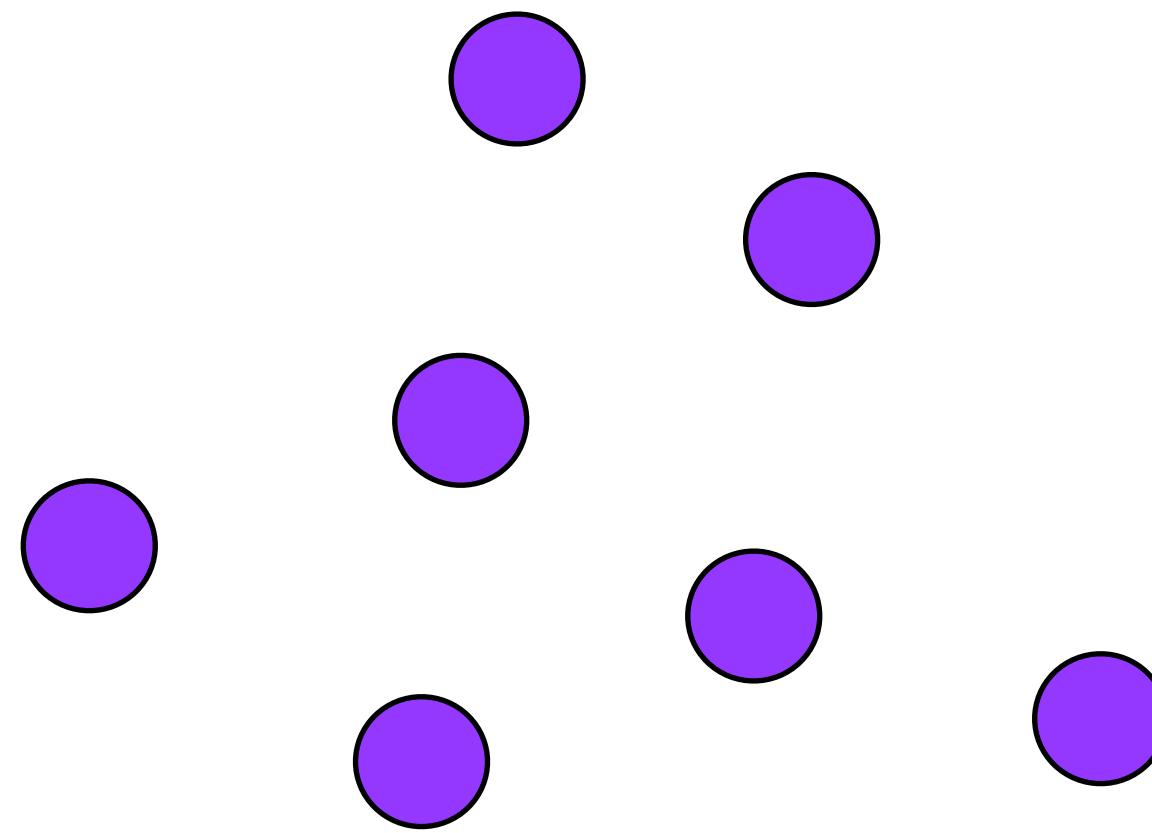
arbovirus, is the causative agent of CTF. Formerly classified as an orbivirus, the sixth report of the International Committee on Taxonomy of Viruses identified CTFV as a member of the genus *Coltivirus* (group A), family Reoviridae (virus code, 60.0.4.0.001; virus accession number, 60040001).³⁴ At least 22 strains of CTFV are known,^{3,24} many of which cause disease in humans.³ Of these, the Florio strain is the best characterized.* Eyach, a group A *Coltivirus* closely related to CTFV, has been detected in European Ixodidae ticks and has been implicated in human disease in Czechoslovakia.^{12,28}

In 2000, the CTFV genome was sequenced and was found to consist of 12 dsRNA segments that encode several important proteins.³ These include VPI, the viral RNA dependent RNA polymerase; methyltransferases; RGD-binding proteins; extracellular proteins that mediate cell-

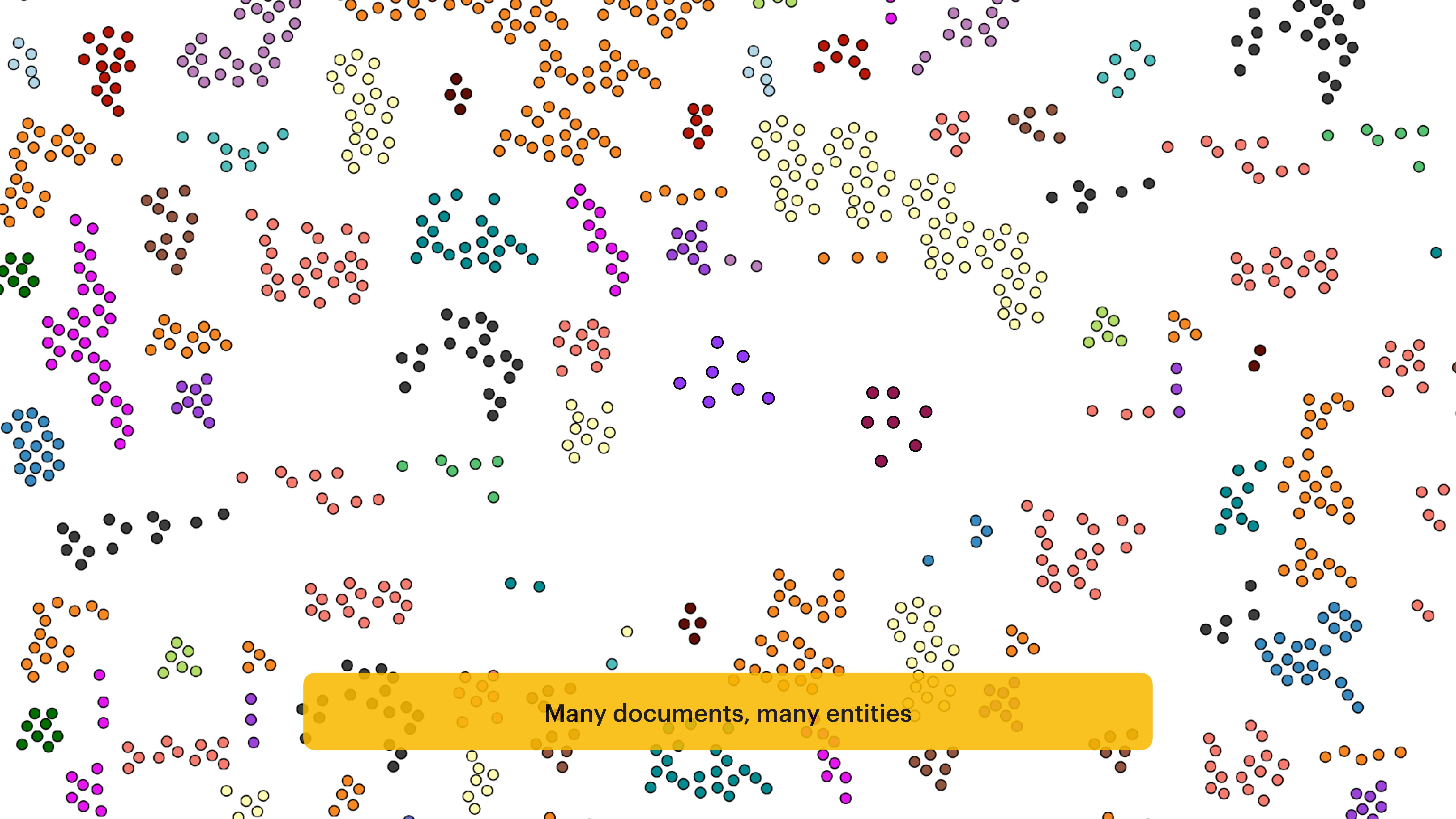
The arboviral infection, CTF, is transmitted from the bite of an infected wood tick.



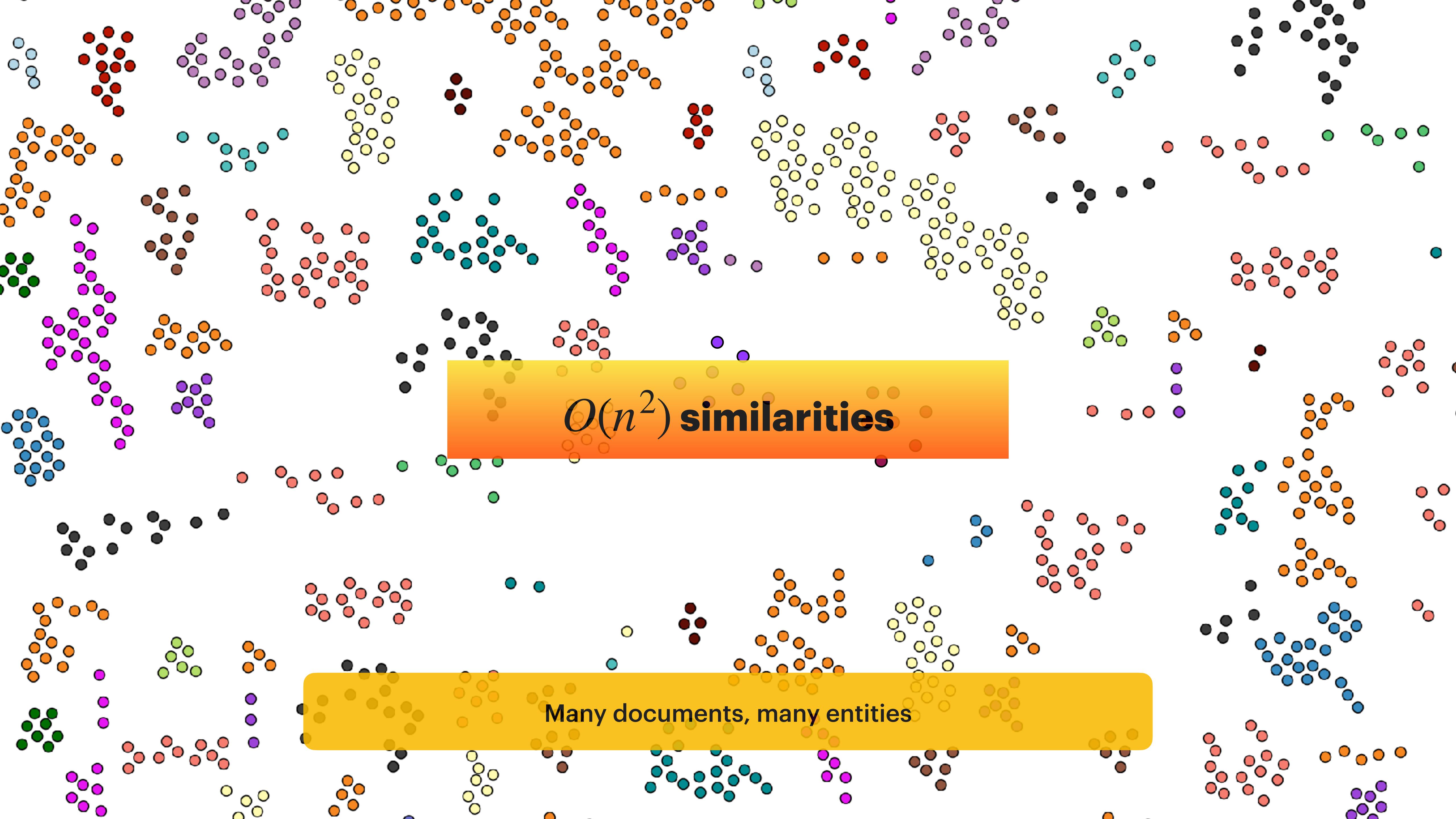
\mathbb{R}^L 

\mathbb{R}^L 

Numerous documents mentioning the same entities



Many documents, many entities



$O(n^2)$ similarities

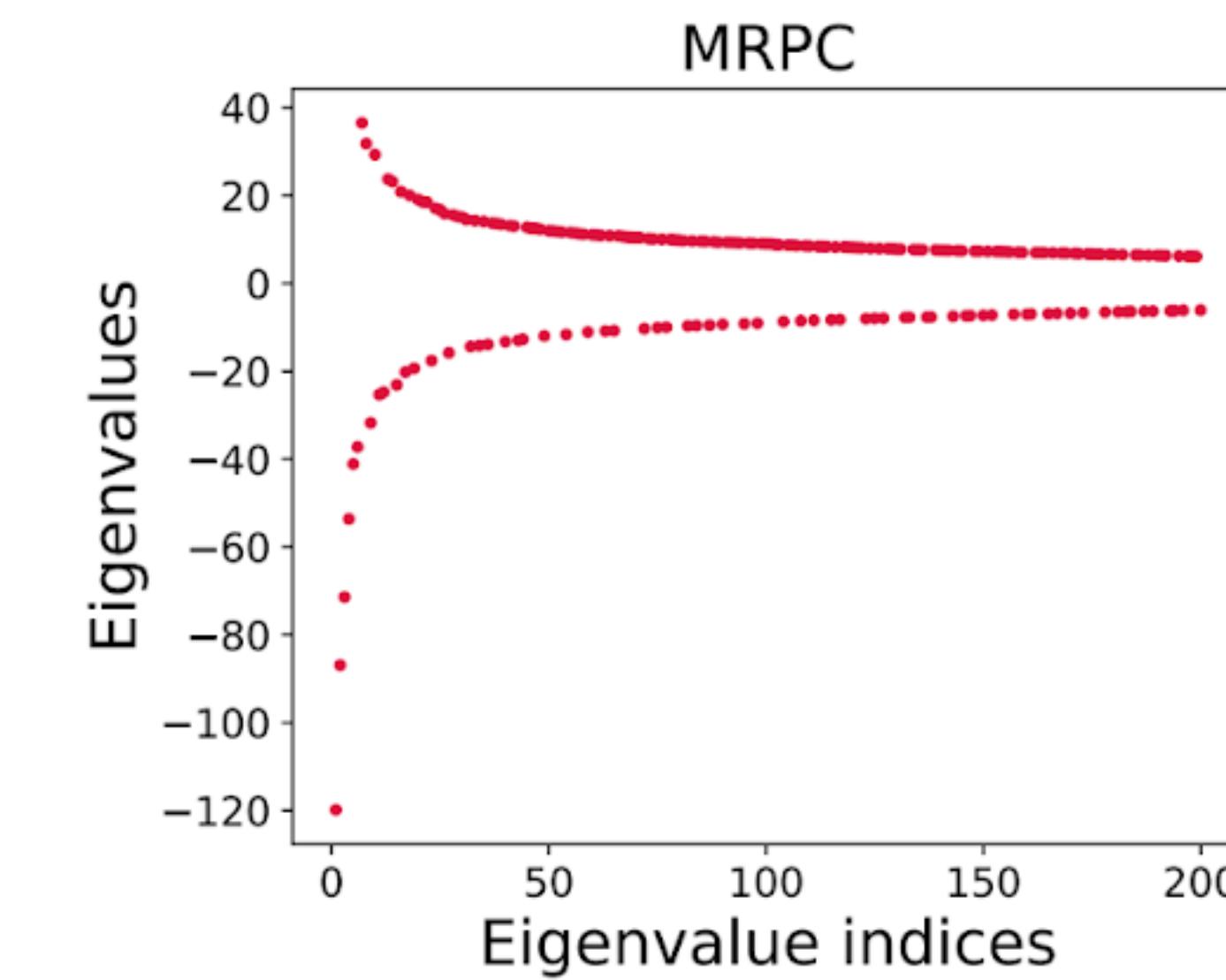
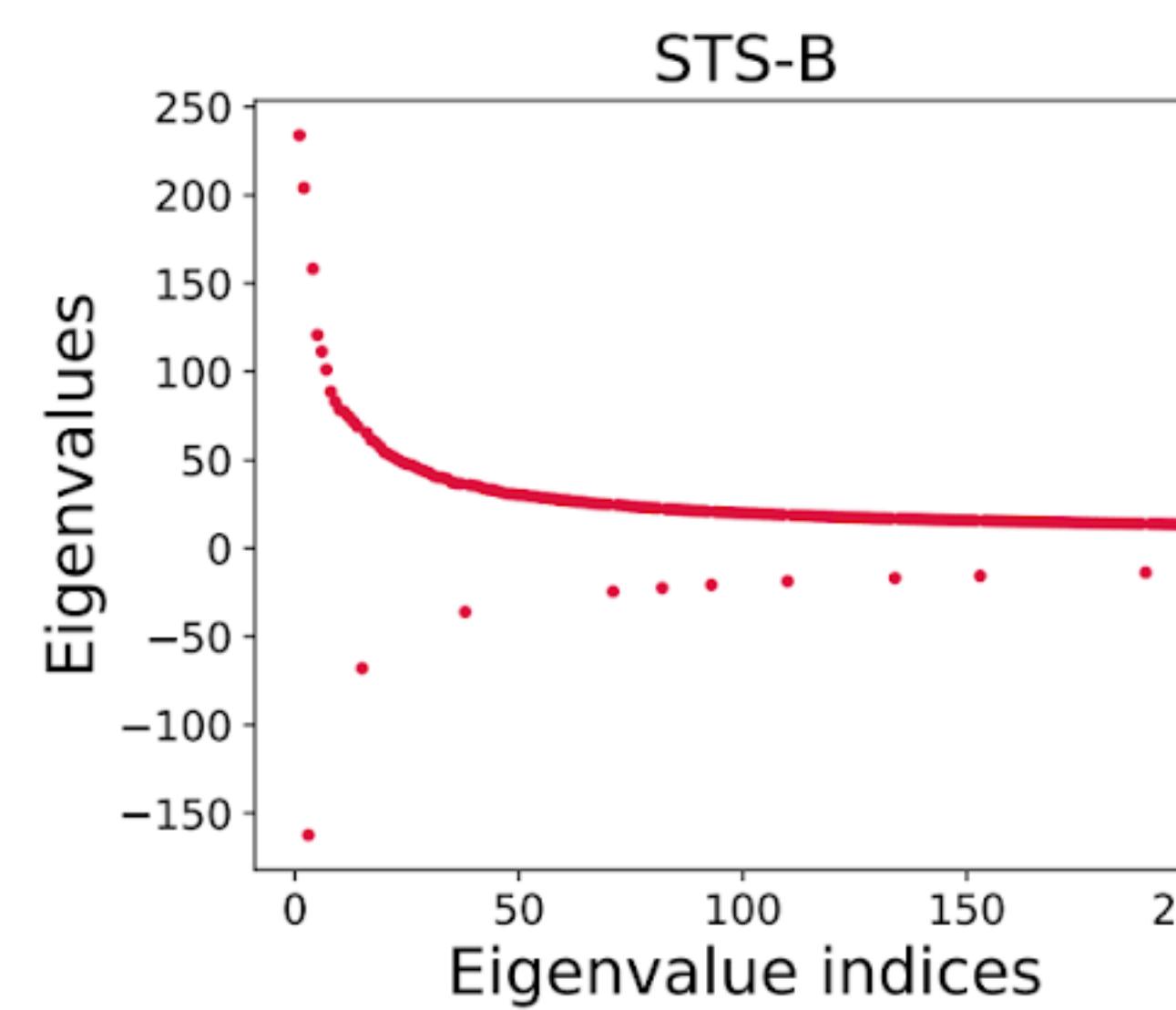
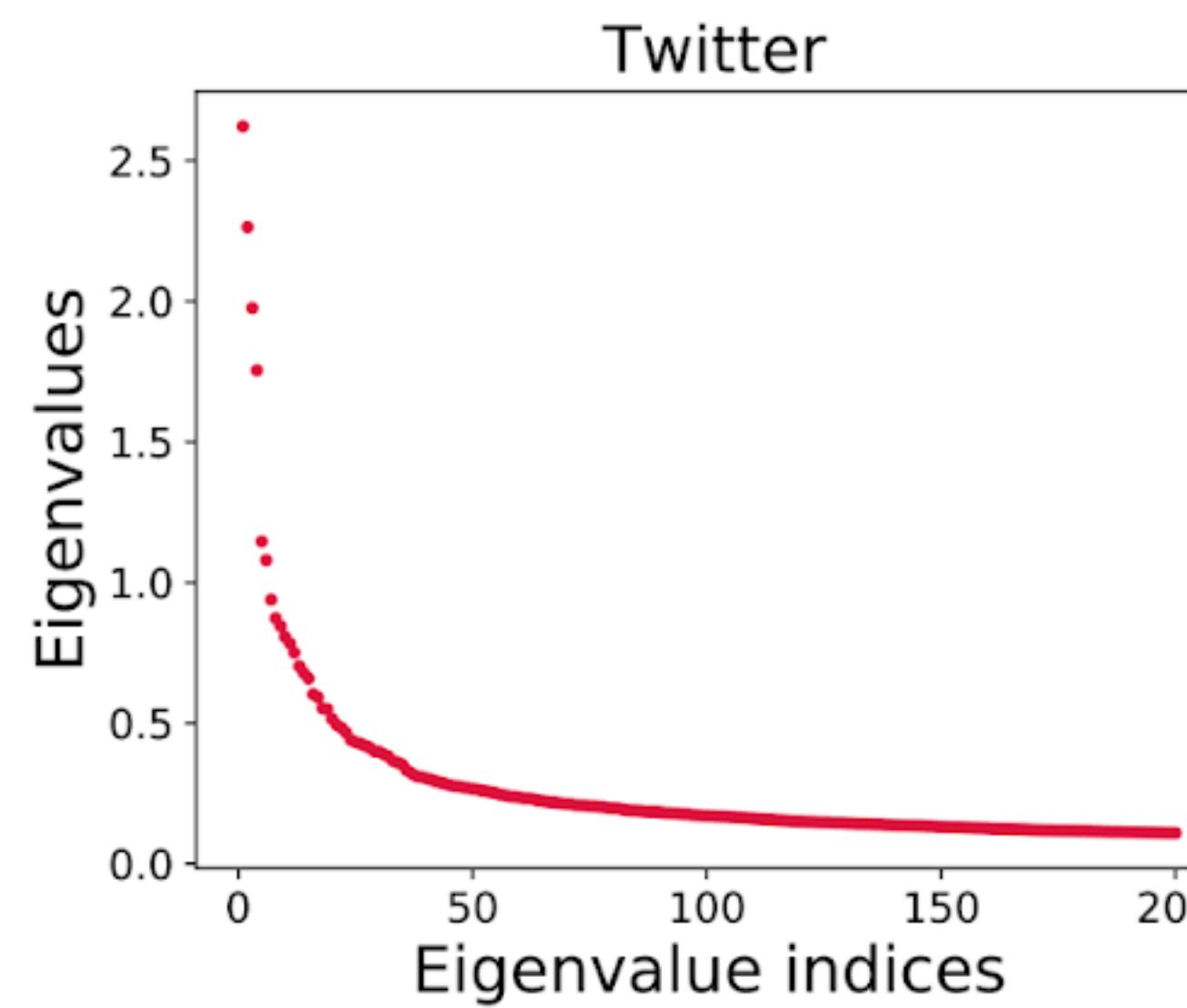
Many documents, many entities

We'll do okay with **approximate** similarities

0.0091	0.3020	0.1653	0.2900
0.3020	0.6571	We'll do okay with approximate similarities	0.4337	0.1983		
..
..
..
..
..
..
0.1653	0.4337	0.3769	0.2375
0.2900	0.1983	0.2375	0.8342

0.0091	0.3020	0.1653	0.2900
0.3020	0.6571	We'll do okay with approximate similarities	0.4337	0.1983		
..
..	..	We want these methods to be sublinear [$o(n)$]
..	..	In NLP most similarity matrices are indefinite
0.1653	0.4337	0.3769	0.2375
0.2900	0.1983	0.2375	0.8342

In NLP most similarity matrices are *indefinite*



0.0091	0.3020	0.1653	0.2900
0.3020	0.6571	We'll do okay with approximate similarities	0.4337	0.1983		
..
..	..	We want these methods to be sublinear [$o(n)$]
..	..	In NLP most similarity matrices are indefinite
..
0.1653	0.4337	Many methods exist for approximating PSD matrices	0.375	0.2375	0.1983	0.8342
0.2900	0.1983	0.2375	0.8342

0.0091	0.3020	0.1653	0.2900
0.3020	0.6571	We'll do okay with approximate similarities	0.4337	0.1983		
..
..	..	We want these methods to be sublinear [$o(n)$]
..
..	..	In NLP most similarity matrices are indefinite
..
0.1653	0.4337	Not many methods for indefinite matrices exist	0.3700	0.2375	0.2375	0.8342
0.2900	0.1983	0.2375	0.8342

$$K = \begin{bmatrix} 0.0089 & 0.3262 & \dots & \dots & \dots & 0.1820 & 0.2993 \\ 0.3262 & 0.6635 & \dots & \dots & \dots & 0.4246 & 0.1990 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0.1820 & 0.4246 & \dots & \dots & \dots & 0.3865 & 0.2350 \\ 0.2993 & 0.1990 & \dots & \dots & \dots & 0.2350 & 0.8414 \end{bmatrix}$$

Columns of \mathbf{K}

$$\tilde{K} = \begin{bmatrix} 0.0089 & 0.3262 \\ 0.3262 & 0.6635 \\ \cdots & \cdots \\ 0.1820 & 0.4246 \\ 0.2993 & 0.1990 \end{bmatrix}$$

$$K = \begin{bmatrix} 0.0089 & 0.3262 & \dots & \dots & \dots & 0.1820 & 0.2993 \\ 0.3262 & 0.6635 & \dots & \dots & \dots & 0.4246 & 0.1990 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.1820 & 0.4246 & \dots & \dots & \dots & 0.3865 & 0.2350 \\ 0.2993 & 0.1990 & \dots & \dots & \dots & 0.2350 & 0.8414 \end{bmatrix}$$

Columns of \mathbf{K}

$$\tilde{\mathbf{K}} = \begin{bmatrix} 0.0089 & 0.3262 \\ 0.3262 & 0.6635 \\ \ddots & \ddots \\ 0.1820 & 0.4246 \\ 0.2993 & 0.1990 \end{bmatrix}$$

Inverse of the principal
sub matrix of \mathbf{K}

$$\begin{bmatrix} -6.6019 & 3.2457 \\ 3.24573 & -0.0886 \end{bmatrix}$$

$$\mathbf{K} = \begin{bmatrix} 0.0089 & 0.3262 & \dots & \dots & \dots & 0.1820 & 0.2993 \\ 0.3262 & 0.6635 & \dots & \dots & \dots & 0.4246 & 0.1990 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.1820 & 0.4246 & \dots & \dots & \dots & 0.3865 & 0.2350 \\ 0.2993 & 0.1990 & \dots & \dots & \dots & 0.2350 & 0.8414 \end{bmatrix}$$

Columns of \mathbf{K}

$$\tilde{\mathbf{K}} = \begin{bmatrix} 0.0089 & 0.3262 \\ 0.3262 & 0.6635 \\ \ddots & \ddots \\ 0.1820 & 0.4246 \\ 0.2993 & 0.1990 \end{bmatrix}$$

Rows of \mathbf{K}

$$\begin{bmatrix} 0.0089 & 0.3262 \\ 0.3262 & 0.6635 \\ \ddots & \ddots \\ 0.1820 & 0.4246 \\ 0.2993 & 0.1990 \end{bmatrix}^T$$

Inverse of the principal
sub matrix of \mathbf{K}

$$\begin{bmatrix} -6.6019 & 3.2457 \\ 3.24573 & -0.0886 \end{bmatrix}$$

$$\mathbf{K} = \begin{bmatrix} 0.0089 & 0.3262 & \dots & \dots & 0.1820 & 0.2993 \\ 0.3262 & 0.6635 & \dots & \dots & 0.4246 & 0.1990 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.1820 & 0.4246 & \dots & \dots & 0.3865 & 0.2350 \\ 0.2993 & 0.1990 & \dots & \dots & 0.2350 & 0.8414 \end{bmatrix}$$

Nyström Method: $\tilde{K} = KS(S^T KS)^+ S^T K$

$$\tilde{K} = \begin{bmatrix} 0.0091 & 0.3020 & \dots & \dots & \dots & 0.1653 & 0.2900 \\ 0.3020 & 0.6571 & \dots & \dots & \dots & 0.4337 & 0.1983 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0.1653 & 0.4337 & \dots & \dots & \dots & 0.3769 & 0.2375 \\ 0.2900 & 0.1983 & \dots & \dots & \dots & 0.2375 & 0.8342 \end{bmatrix}$$

$$K = \begin{bmatrix} 0.0089 & 0.3262 & \dots & \dots & \dots & 0.1820 & 0.2993 \\ 0.3262 & 0.6635 & \dots & \dots & \dots & 0.4246 & 0.1990 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0.1820 & 0.4246 & \dots & \dots & \dots & 0.3865 & 0.2350 \\ 0.2993 & 0.1990 & \dots & \dots & \dots & 0.2350 & 0.8414 \end{bmatrix}$$

Nyström Method: $\tilde{K} = KS(S^T KS)^+ S^T K$

$$\tilde{K} = \begin{bmatrix} 0.0091 & 0.3020 & \dots & \dots & \dots & 0.1653 & 0.2900 \\ 0.3020 & 0.6571 & \dots & \dots & \dots & 0.4337 & 0.1983 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0.1653 & 0.4337 & \dots & \dots & \dots & 0.3769 & 0.2375 \\ 0.2900 & 0.1983 & \dots & \dots & \dots & 0.2375 & 0.8342 \end{bmatrix}$$

Benefits: Runs in **sublinear** time with respect to K .

$$K = \begin{bmatrix} 0.0089 & 0.3262 & \dots & \dots & \dots & 0.1820 & 0.2993 \\ 0.3262 & 0.6635 & \dots & \dots & \dots & 0.4246 & 0.1990 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0.1820 & 0.4246 & \dots & \dots & \dots & 0.3865 & 0.2350 \\ 0.2993 & 0.1990 & \dots & \dots & \dots & 0.2350 & 0.8414 \end{bmatrix}$$

Nyström Method: $\tilde{K} = KS(S^T K S)^+ S^T K$

$$K = \begin{bmatrix} 0.0089 & 0.3262 & \dots & \dots & \dots & 0.1820 & 0.2993 \\ 0.3262 & 0.6635 & \dots & \dots & \dots & 0.4246 & 0.1990 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0.1820 & 0.4246 & \dots & \dots & \dots & 0.3865 & 0.2350 \\ 0.2993 & 0.1990 & \dots & \dots & \dots & 0.2350 & 0.8414 \end{bmatrix}$$

But becomes **unstable** for **indefinite matrices** unless they are near PSD

Columns of \mathbf{K}

$$\tilde{\mathbf{K}} = \begin{bmatrix} 0.0089 & 0.3262 \\ 0.3262 & 0.6635 \\ \ddots & \ddots \\ 0.1820 & 0.4246 \\ 0.2993 & 0.1990 \end{bmatrix}$$

Rows of \mathbf{K}

$$\begin{bmatrix} 0.0089 & 0.3262 \\ 0.3262 & 0.6635 \\ \ddots & \ddots \\ 0.1820 & 0.4246 \\ 0.2993 & 0.1990 \end{bmatrix}^T$$

Inverse of the principal
sub matrix of \mathbf{K}

$$\begin{bmatrix} -6.6019 & 3.2457 \\ 3.24573 & -0.0886 \end{bmatrix}$$



$(S^T K S)^+$

$$\mathbf{K} = \begin{bmatrix} 0.0089 & 0.3262 & \dots & \dots & 0.1820 & 0.2993 \\ 0.3262 & 0.6635 & \dots & \dots & 0.4246 & 0.1990 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0.1820 & 0.4246 & \dots & \dots & 0.3865 & 0.2350 \\ 0.2993 & 0.1990 & \dots & \dots & 0.2350 & 0.8414 \end{bmatrix}$$

Columns of \mathbf{K}

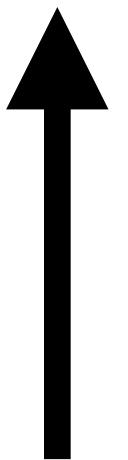
$$\tilde{\mathbf{K}} = \begin{bmatrix} 0.0089 & 0.3262 \\ 0.3262 & 0.6635 \\ \ddots & \ddots \\ 0.1820 & 0.4246 \\ 0.2993 & 0.1990 \end{bmatrix}$$

Rows of \mathbf{K}

$$\begin{bmatrix} 0.0089 & 0.3262 \\ 0.3262 & 0.6635 \\ \ddots & \ddots \\ 0.1820 & 0.4246 \\ 0.2993 & 0.1990 \end{bmatrix}^T$$

Inverse of the principal
sub matrix of \mathbf{K}

$$\begin{bmatrix} -6.6019 & 3.2457 \\ 3.24573 & -0.0886 \end{bmatrix}$$



If $S^T K S$ is ill conditioned, $(S^T K S)^+$ will fail

$$K = \begin{bmatrix} 0.0089 & 0.3262 & \dots & \dots & \dots & 0.1820 & 0.2993 \\ 0.3262 & 0.6635 & \dots & \dots & \dots & 0.4246 & 0.1990 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.1820 & 0.4246 & \dots & \dots & \dots & 0.3865 & 0.2350 \\ 0.2993 & 0.1990 & \dots & \dots & \dots & 0.2350 & 0.8414 \end{bmatrix}$$

Smaller eigenvalues of $S^T K S$ can get blown up leading to large error

Columns of \mathbf{K}

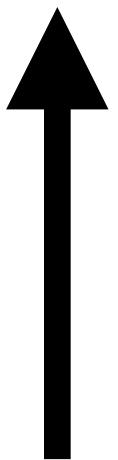
$$\tilde{K} = \begin{bmatrix} 0.0089 & 0.3262 \\ 0.3262 & 0.6635 \\ \ddots & \ddots \\ 0.1820 & 0.4246 \\ 0.2993 & 0.1990 \end{bmatrix}$$

Rows of \mathbf{K}

$$\begin{bmatrix} 0.0089 & 0.3262 \\ 0.3262 & 0.6635 \\ \ddots & \ddots \\ 0.1820 & 0.4246 \\ 0.2993 & 0.1990 \end{bmatrix}^T$$

Inverse of the principal
sub matrix of \mathbf{K}

$$\begin{bmatrix} -6.6019 & 3.2457 \\ 3.24573 & -0.0886 \end{bmatrix}$$



If $S^T K S$ is ill conditioned, $(S^T K S)^+$ will fail

$$K = \begin{bmatrix} 0.0089 & 0.3262 & \dots & \dots & \dots & 0.1820 & 0.2993 \\ 0.3262 & 0.6635 & \dots & \dots & \dots & 0.4246 & 0.1990 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.1820 & 0.4246 & \dots & \dots & \dots & 0.3865 & 0.2350 \\ 0.2993 & 0.1990 & \dots & \dots & \dots & 0.2350 & 0.8414 \end{bmatrix}$$

Smaller eigenvalues of $S^T K S$ can get blown up leading to large error

Columns of \mathbf{K}

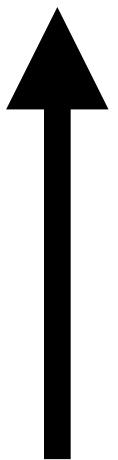
$$\tilde{K} = \begin{bmatrix} 0.0089 & 0.3262 \\ 0.3262 & 0.6635 \\ \ddots & \ddots \\ 0.1820 & 0.4246 \\ 0.2993 & 0.1990 \end{bmatrix}$$

Rows of \mathbf{K}

$$\begin{bmatrix} 0.0089 & 0.3262 \\ 0.3262 & 0.6635 \\ \ddots & \ddots \\ 0.1820 & 0.4246 \\ 0.2993 & 0.1990 \end{bmatrix}^T$$

Inverse of the principal
sub matrix of \mathbf{K}

$$\begin{bmatrix} -6.6019 & 3.2457 \\ 3.24573 & -0.0886 \end{bmatrix}$$

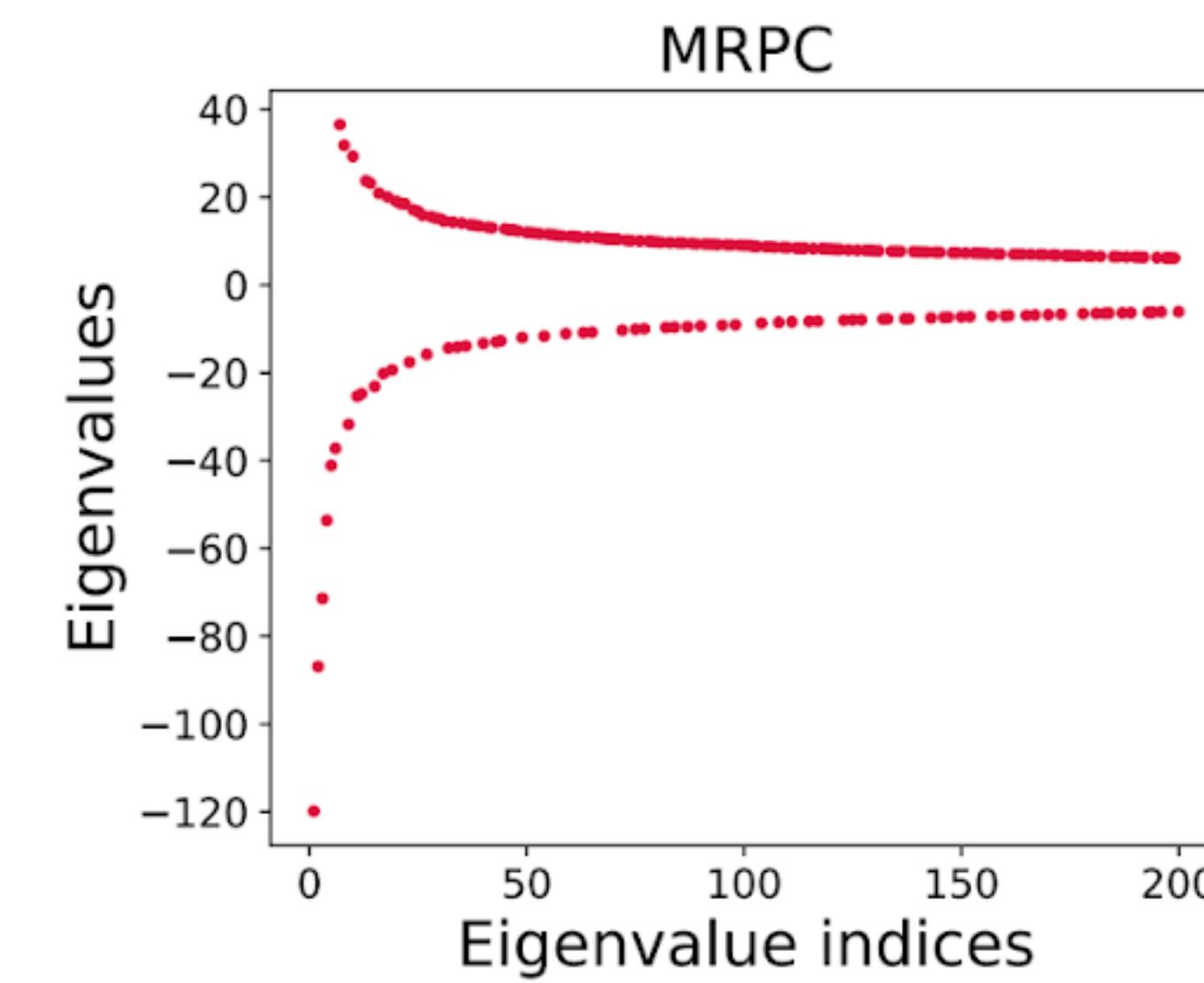
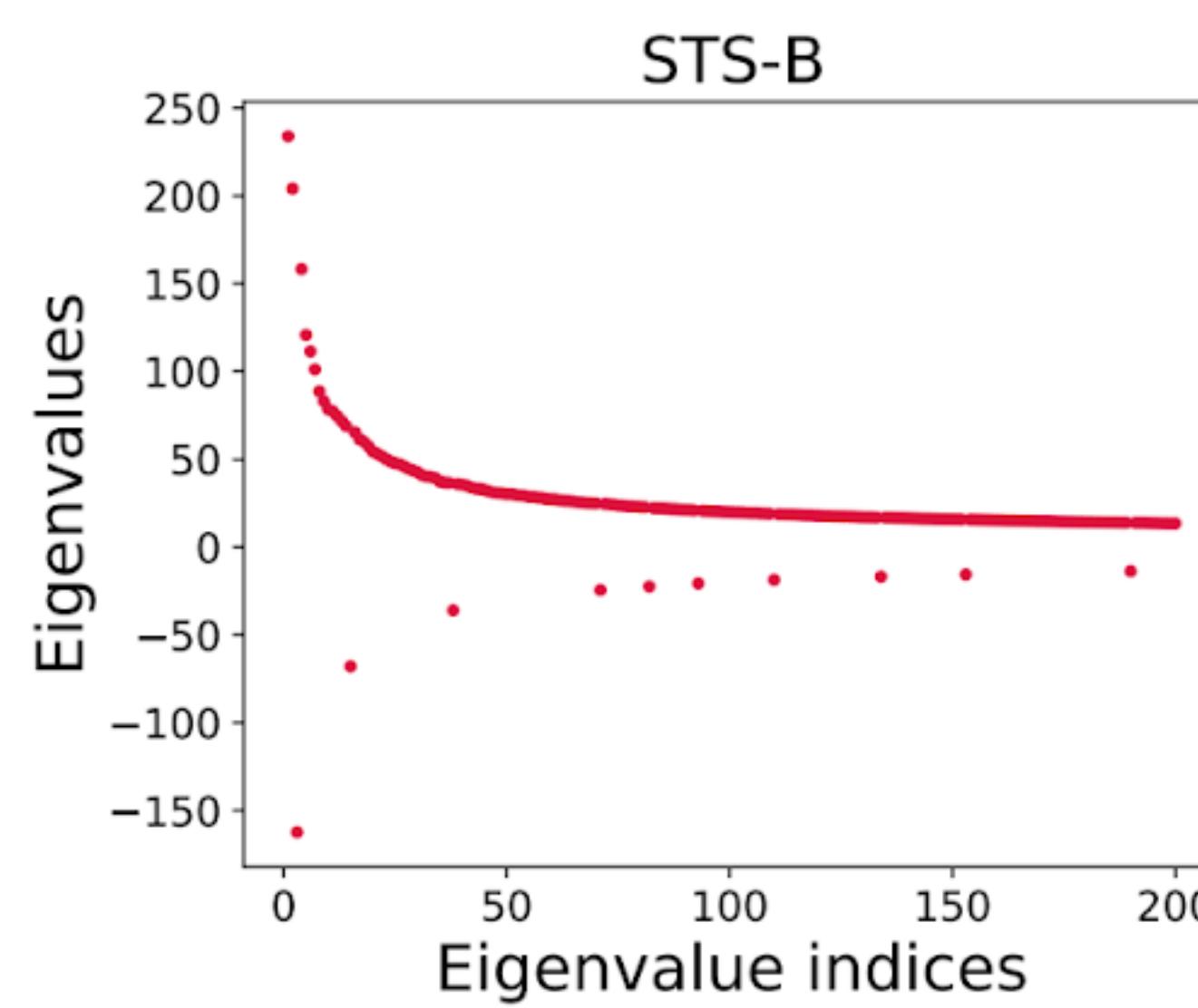
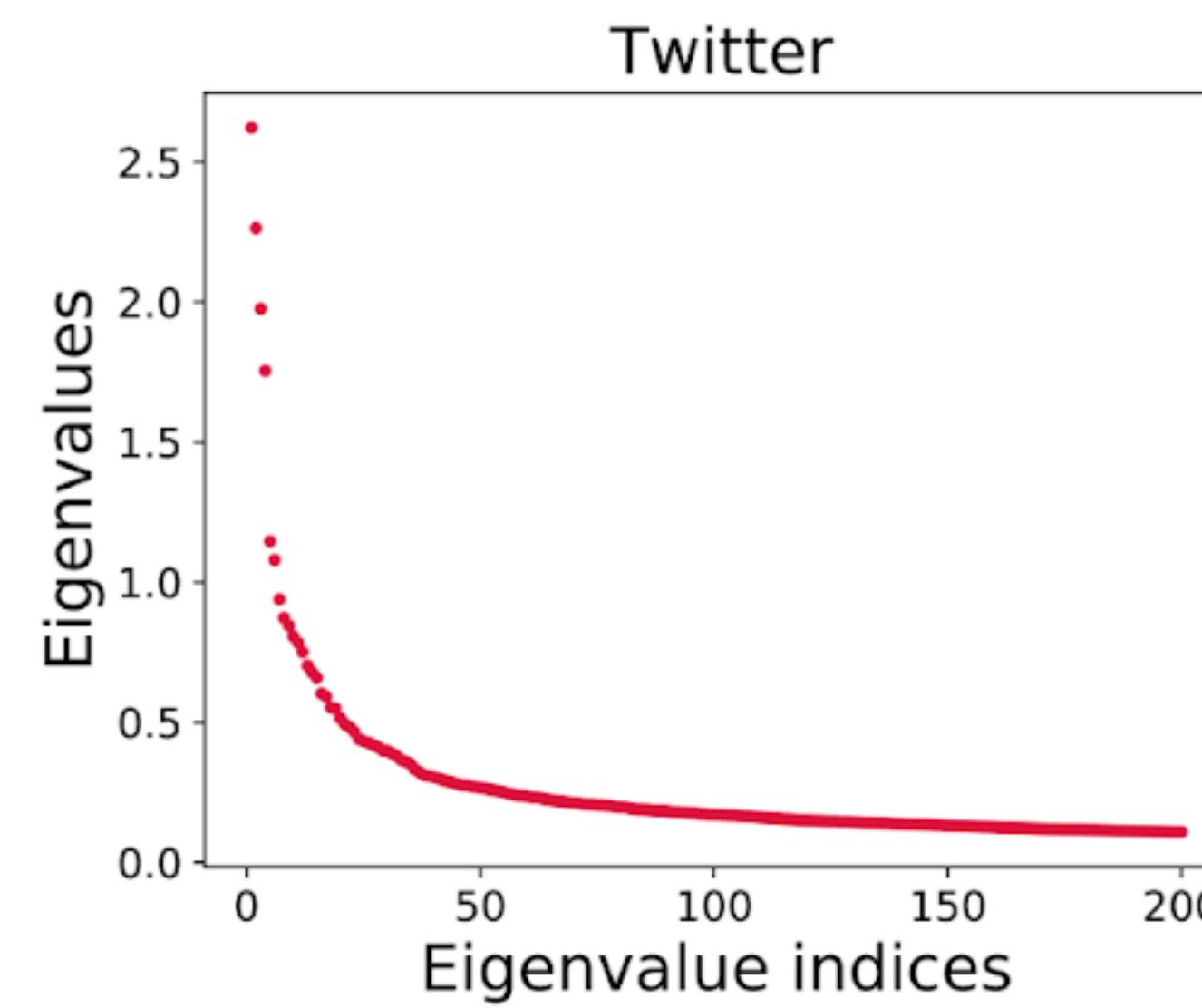


If $S^T K S$ is ill conditioned, $(S^T K S)^+$ will fail

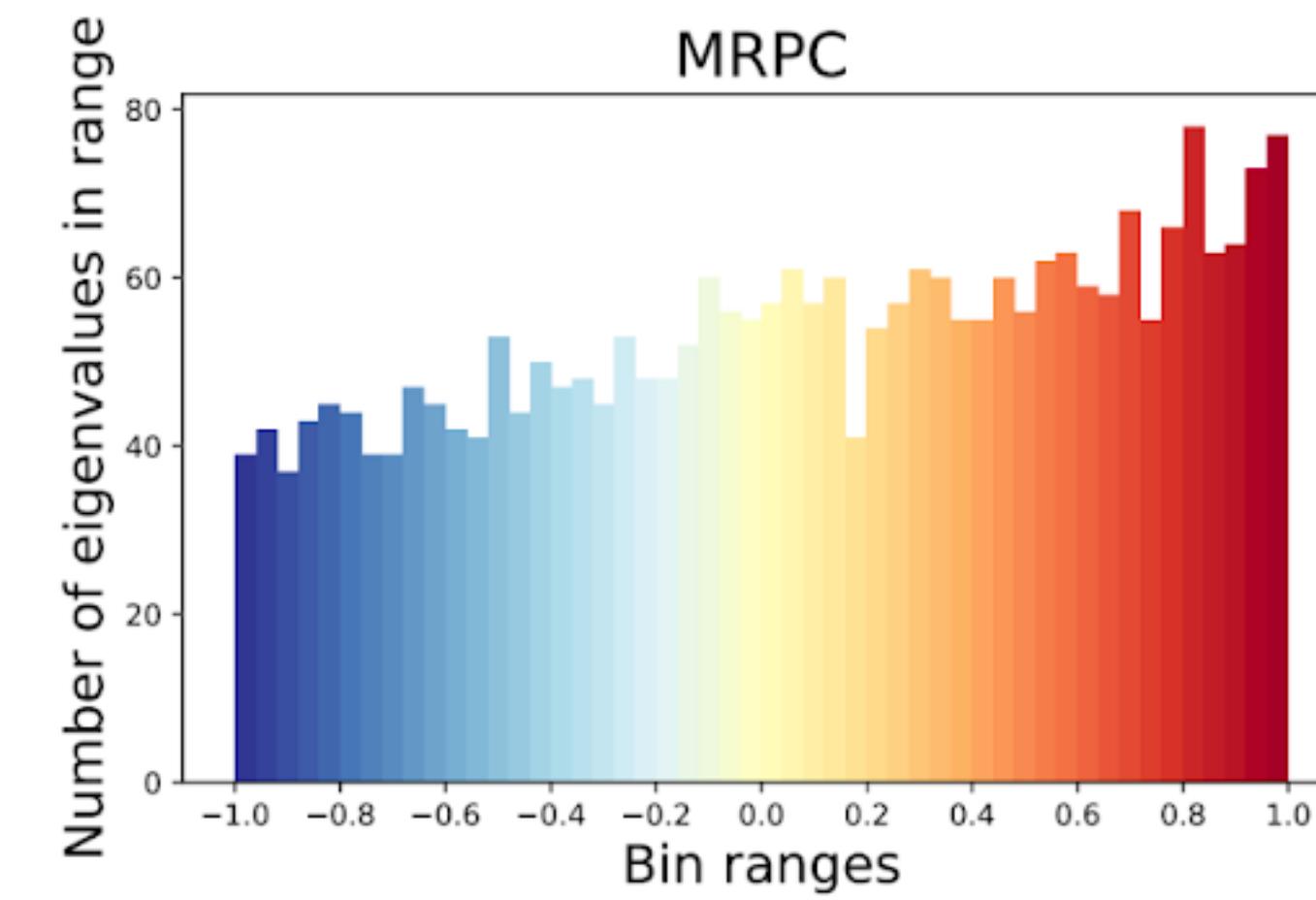
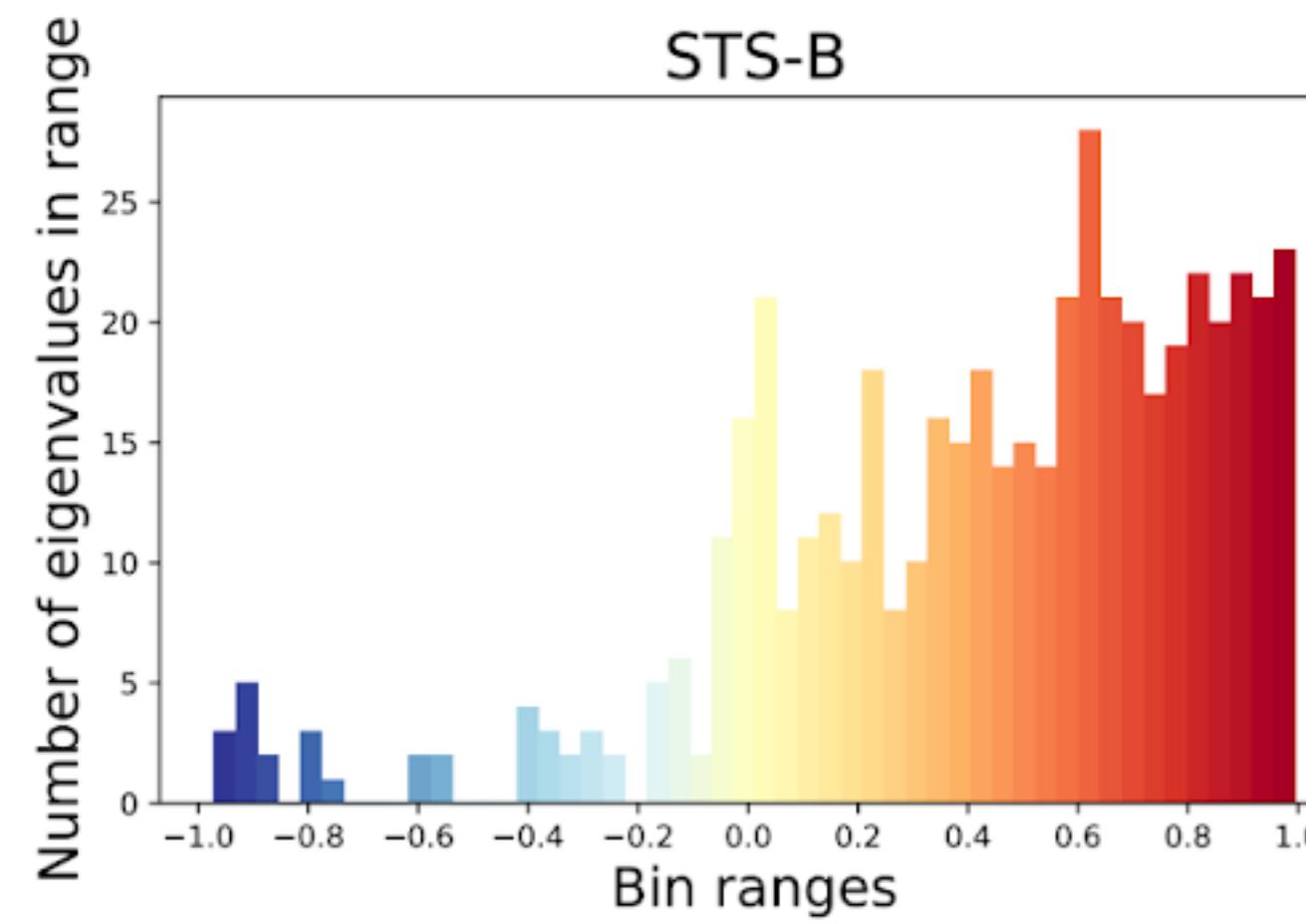
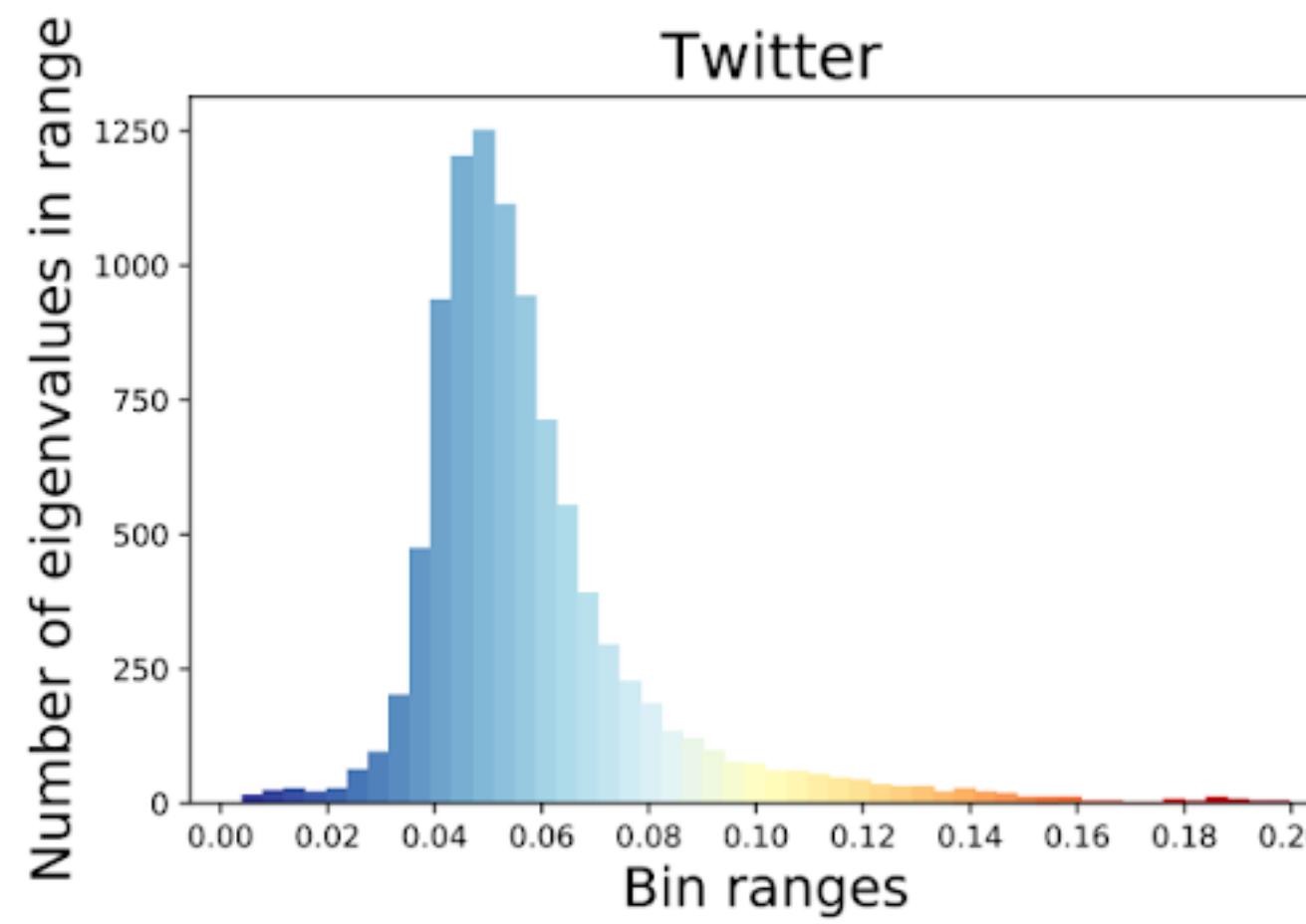
Endemic to indefinite matrices

$$K = \begin{bmatrix} 0.0089 & 0.3262 & \dots & \dots & \dots & 0.1820 & 0.2993 \\ 0.3262 & 0.6635 & \dots & \dots & \dots & 0.4246 & 0.1990 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0.1820 & 0.4246 & \dots & \dots & \dots & 0.3865 & 0.2350 \\ 0.2993 & 0.1990 & \dots & \dots & \dots & 0.2350 & 0.8414 \end{bmatrix}$$

Endemic to indefinite matrices



Endemic to indefinite matrices



Independently sample $S^T K S$, plot histogram of the eigenvalues

Sublinear methods for approximating indefinite similarity matrices

SMS-Nyström

Other Approaches

Sublinear methods for approximating indefinite similarity matrices

SMS-Nyström

$$K = KS(S^T K S)^+ S^T K$$

$$\bar{K} = K - \lambda_{\min}(K) I_n$$

If $\lambda_{\min}(K)$ is **small**, we can apply Nyström approximation

Other Approaches

Sublinear methods for approximating indefinite similarity matrices

SMS-Nyström

$$K = KS(S^T K S)^+ S^T K$$

$$\bar{K} = K - \lambda_{\min}(K) I_n$$

If $\lambda_{\min}(K)$ is **small**, we can apply Nyström approximation

Other Approaches

Skeleton approximation:

$$\tilde{K} = KS_2(S_2^T K S_1)^+ S_1^T K$$

Sample S_1 and S_2 independently at random

Sublinear methods for approximating indefinite similarity matrices

SMS-Nyström

$$K = KS(S^T K S)^+ S^T K$$

$$\bar{K} = K - \lambda_{\min}(K) I_n$$

If $\lambda_{\min}(K)$ is **small**, we can apply Nyström approximation

Other Approaches

Skeleton approximation:

$$\tilde{K} = KS_2(S_2^T K S_1)^+ S_1^T K$$

SiCUR:

$$\tilde{K} = KS_2(S_2^T K S_1)^+ S_1^T K$$

$$|S_2| = 2 |S_1|$$

Sample S_1 and S_2 independently at random

Sublinear methods for approximating indefinite similarity matrices

SMS-Nyström

$$K = KS(S^T K S)^+ S^T K$$

$$\bar{K} = K - \lambda_{\min}(K) I_n$$

If $\lambda_{\min}(K)$ is **small**, we can apply Nyström approximation

Other Approaches

Skeleton approximation:

$$\tilde{K} = KS_2(S_2^T K S_1)^+ S_1^T K$$

SiCUR:

$$\tilde{K} = KS_2(S_2^T K S_1)^+ S_1^T K$$

$$|S_2| = 2 |S_1|$$

StaCUR:

$$\tilde{K} = \frac{n}{s} KS(KSS^T K)^+ S^T KSS^T K$$

Sublinear methods for approximating indefinite similarity matrices

SMS-Nyström

$$K = KS(S^T K S)^+ S^T K$$

$$\bar{K} = K - \lambda_{\min}(K) I_n$$

If $\lambda_{\min}(K)$ is **small**, we can apply Nyström approximation

Other Approaches

Skeleton approximation:

$$\tilde{K} = KS_2(S_2^T K S_1)^+ S_1^T K$$

SiCUR:

$$\tilde{K} = KS_2(S_2^T K S_1)^+ S_1^T K$$

$$|S_2| = 2 |S_1|$$

StaCUR:

$$\tilde{K} = \frac{n}{s} KS(K S S^T K)^+ S^T K S S^T K$$

Sublinear methods for approximating indefinite similarity matrices

$O(n^3)$ computations for $\lambda_{\min}(K)$

If $|\lambda_{\min}(K)|$ is large, **bad approximation**

SMS-Nyström

$$K = \begin{bmatrix} 0.0089 & 0.3262 & \dots & \dots & 0.1820 & 0.2993 \\ 0.3262 & 0.6635 & \dots & \dots & 0.4246 & 0.1990 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.1820 & 0.4246 & \dots & \dots & 0.3865 & 0.2350 \\ 0.2993 & 0.1990 & \dots & \dots & 0.2350 & 0.8414 \end{bmatrix}$$

Sublinear methods for approximating indefinite similarity matrices

$O(n^3)$ computations for $\lambda_{\min}(K)$

If $|\lambda_{\min}(K)|$ is large, **bad approximation**

Instead compute $\lambda_{\min}(S_2^T K S_2)$, where $|S_2| = 2|S|$

SMS-Nyström

$$K = \begin{bmatrix} 0.0089 & 0.3262 & \dots & \dots & 0.1820 & 0.2993 \\ 0.3262 & 0.6635 & \dots & \dots & 0.4246 & 0.1990 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.1820 & 0.4246 & \dots & \dots & 0.3865 & 0.2350 \\ 0.2993 & 0.1990 & \dots & \dots & 0.2350 & 0.8414 \end{bmatrix}$$

Sublinear methods for approximating indefinite similarity matrices

$O(n^3)$ computations for $\lambda_{\min}(K)$

If $|\lambda_{\min}(K)|$ is large, **bad approximation**

Instead compute $\lambda_{\min}(S_2^T K S_2)$, where $|S_2| = 2|S|$

Shift $S^T K S$ using $\lambda_{\min}(S_2^T K S_2)$

SMS-Nyström

$$K = \begin{bmatrix} 0.0089 & 0.3262 & \dots & \dots & 0.1820 & 0.2993 \\ 0.3262 & 0.6635 & \dots & \dots & 0.4246 & 0.1990 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.1820 & 0.4246 & \dots & \dots & 0.3865 & 0.2350 \\ 0.2993 & 0.1990 & \dots & \dots & 0.2350 & 0.8414 \end{bmatrix}$$

Sublinear methods for approximating indefinite similarity matrices

$O(n^3)$ computations for $\lambda_{\min}(K)$

If $|\lambda_{\min}(K)|$ is large, **bad approximation**

Instead compute $\lambda_{\min}(S_2^T K S_2)$, where $|S_2| = 2|S|$

Shift $S^T K S$ using $\lambda_{\min}(S_2^T K S_2)$

$S^T K S$ is guaranteed to be PSD

SMS-Nyström

$$K = \begin{bmatrix} 0.0089 & 0.3262 & \dots & \dots & 0.1820 & 0.2993 \\ 0.3262 & 0.6635 & \dots & \dots & 0.4246 & 0.1990 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0.1820 & 0.4246 & \dots & \dots & 0.3865 & 0.2350 \\ 0.2993 & 0.1990 & \dots & \dots & 0.2350 & 0.8414 \end{bmatrix}$$

Sublinear methods for approximating indefinite similarity matrices

$O(n^3)$ computations for $\lambda_{\min}(K)$

If $|\lambda_{\min}(K)|$ is large, **bad approximation**

Instead compute $\lambda_{\min}(S_2^T K S_2)$, where $|S_2| = 2|S|$

Shift $S^T K S$ using $\lambda_{\min}(S_2^T K S_2)$

$S^T K S$ is guaranteed to be PSD

Shift is large, so no eigenvalues near 0

SMS-Nyström

$$K = \begin{bmatrix} 0.0089 & 0.3262 & \dots & \dots & 0.1820 & 0.2993 \\ 0.3262 & 0.6635 & \dots & \dots & 0.4246 & 0.1990 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0.1820 & 0.4246 & \dots & \dots & 0.3865 & 0.2350 \\ 0.2993 & 0.1990 & \dots & \dots & 0.2350 & 0.8414 \end{bmatrix}$$

Sublinear methods for approximating indefinite similarity matrices

Overall computation overheads

$(|S_2| - |S|)^2$ similarity computations

$O(|S_2|^3)$ computation for $\lambda_{\min}(S_2^T K S_2)$

SMS-Nyström

$$K = \begin{bmatrix} 0.0089 & 0.3262 & \dots & \dots & 0.1820 & 0.2993 \\ 0.3262 & 0.6635 & \dots & \dots & 0.4246 & 0.1990 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.1820 & 0.4246 & \dots & \dots & 0.3865 & 0.2350 \\ 0.2993 & 0.1990 & \dots & \dots & 0.2350 & 0.8414 \end{bmatrix}$$

Sublinear methods for approximating indefinite similarity matrices

SMS-Nyström

$$K = KS(S^T KS)^+ S^T K$$

$$\bar{K} = K - \lambda_{\min}(K) I_n$$

If $\lambda_{\min}(K)$ is **small**, we can apply Nyström approximation

Other Approaches

Skeleton approximation:

$$\tilde{K} = KS_2(S_2^T KS_1)^+ S_1^T K$$

SiCUR:

$$\tilde{K} = KS_2(S_2^T KS_1)^+ S_1^T K$$

$$|S_2| = 2 |S_1|$$

StaCUR:

$$\tilde{K} = \frac{n}{s} KS(KSS^T K)^+ S^T KSS^T K$$

Very similar to Nyström Approximation, rows and columns sampled independently

Sublinear methods for approximating indefinite similarity matrices

SMS-Nyström

$$K = KS(S^T KS)^+ S^T K$$

$$\bar{K} = K - \lambda_{\min}(K) I_n$$

If $\lambda_{\min}(K)$ is **small**, we can apply Nyström approximation

Other Approaches

Skeleton approximation:

$$\tilde{K} = KS_2(S_2^T KS_1)^+ S_1^T K$$

SiCUR:

$$\tilde{K} = KS_2(S_2^T KS_1)^+ S_1^T K$$

$$|S_2| = 2 |S_1|$$

StaCUR:

$$\tilde{K} = \frac{n}{s} KS(KSS^T K)^+ S^T KSS^T K$$

$|S_2| > |S_1|$ works amazingly! In our experiments $|S_2| = 2 |S_1|$ works reliably

Sublinear methods for approximating indefinite similarity matrices

SMS-Nyström

$$K = KS(S^T KS)^+ S^T K$$

$$\bar{K} = K - \lambda_{\min}(K) I_n$$

If $\lambda_{\min}(K)$ is **small**, we can apply Nyström approximation

Other Approaches

Skeleton approximation:

$$\tilde{K} = KS_2(S_2^T KS_1)^+ S_1^T K$$

SiCUR:

$$\tilde{K} = KS_2(S_2^T KS_1)^+ S_1^T K$$

$$|S_2| = 2 |S_1|$$

StaCUR:

$$\tilde{K} = \frac{n}{s} KS(KSS^T K)^+ S^T KSS^T K$$

$(S^T KS)^+$ is replaced with $(KSS^T K)^+ S^T K S$

Sublinear methods for approximating indefinite similarity matrices

SMS-Nyström

$$K = KS(S^T KS)^+ S^T K$$

$$\bar{K} = K - \lambda_{\min}(K) I_n$$

If $\lambda_{\min}(K)$ is **small**, we can apply Nyström approximation

Other Approaches

Skeleton approximation:

$$\tilde{K} = KS_2(S_2^T KS_1)^+ S_1^T K$$

SiCUR:

$$\tilde{K} = KS_2(S_2^T KS_1)^+ S_1^T K$$

$$|S_2| = 2 |S_1|$$

StaCUR:

$$\tilde{K} = \frac{n}{s} KS(KSS^T K)^+ S^T KSS^T K$$

Benefits: No parameters to tune!

Sublinear methods for approximating indefinite similarity matrices

SMS-Nyström

$$K = KS(S^T KS)^+ S^T K$$

$$\bar{K} = K - \lambda_{\min}(K) I_n$$

If $\lambda_{\min}(K)$ is **small**, we can apply Nyström approximation

Other Approaches

Skeleton approximation:

$$\tilde{K} = KS_2(S_2^T KS_1)^+ S_1^T K$$

SiCUR:

$$\tilde{K} = KS_2(S_2^T KS_1)^+ S_1^T K$$

$$|S_2| = 2 |S_1|$$

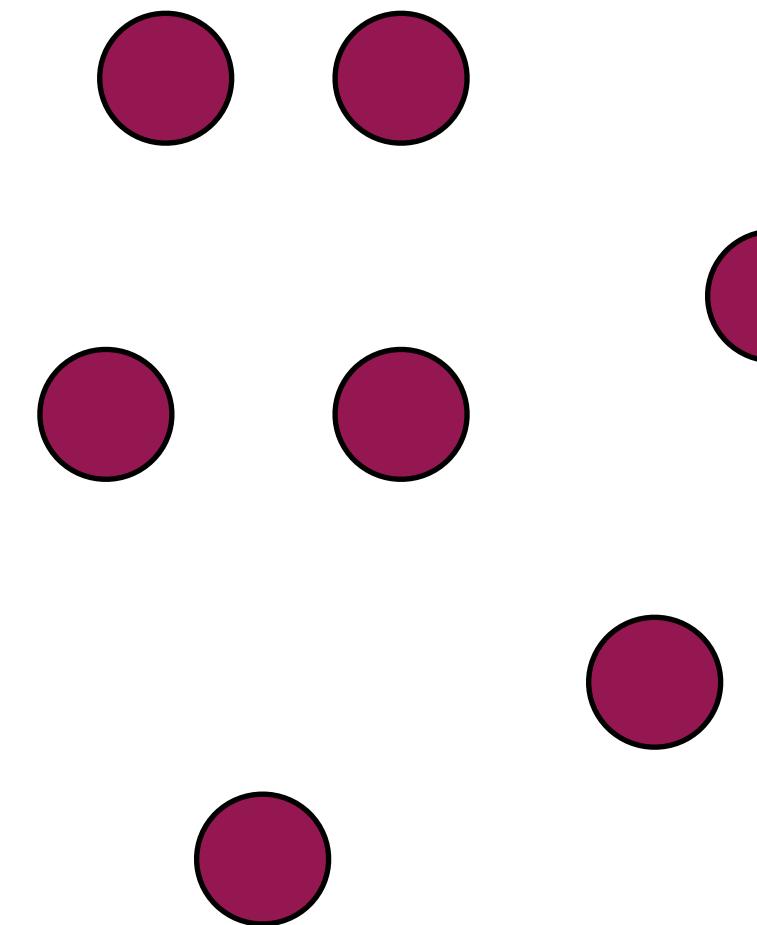
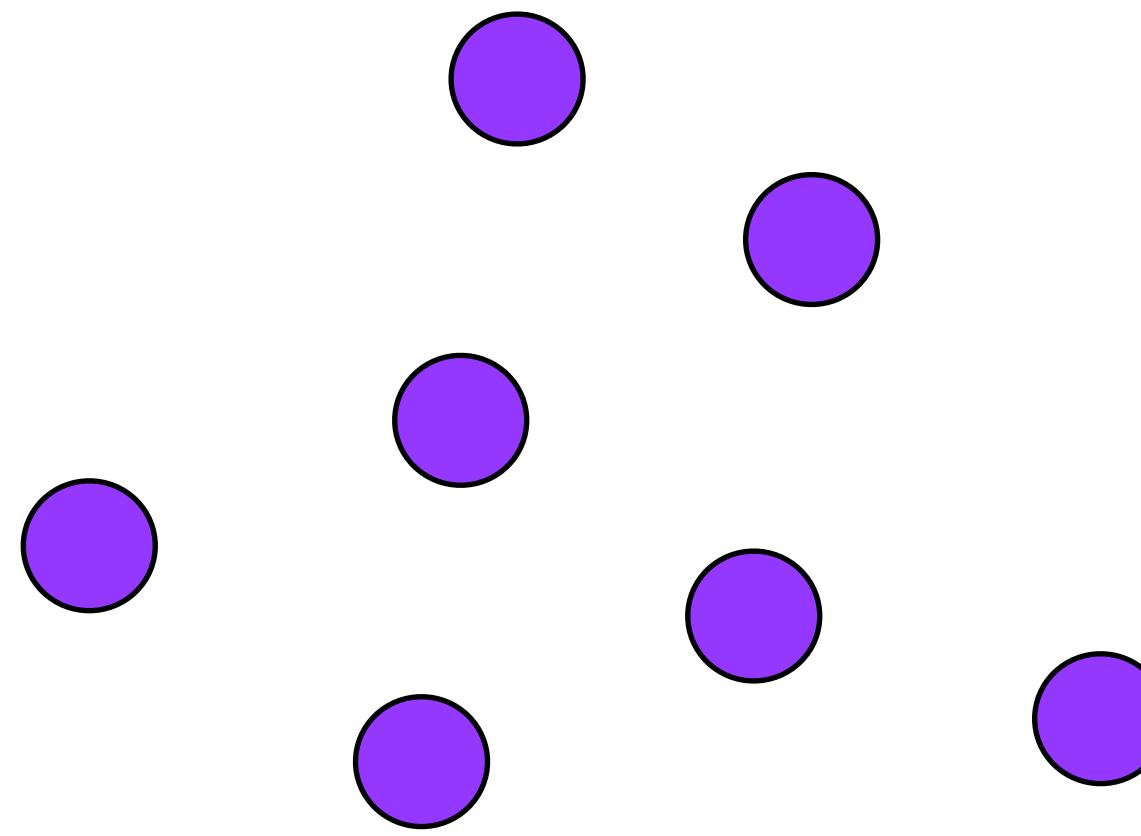
StaCUR:

$$\tilde{K} = \frac{n}{s} KS(KSS^T K)^+ S^T KSS^T K$$

Benefits: No two stage sampling!

Empirical evaluation

\mathbb{R}^L



Consider the task of document classification

Empirical evaluation

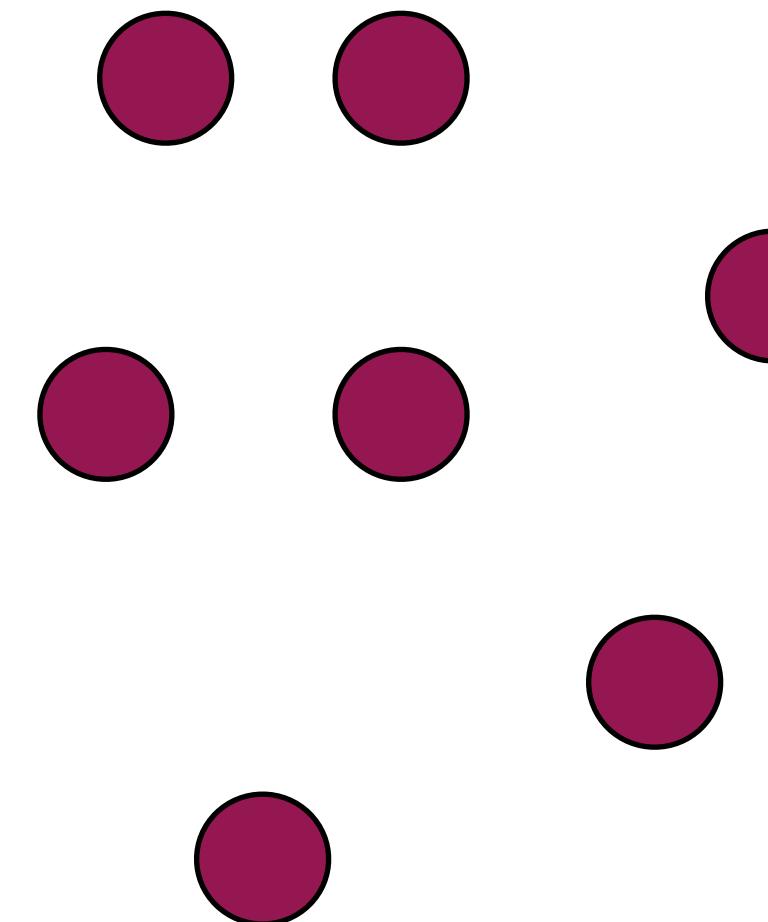
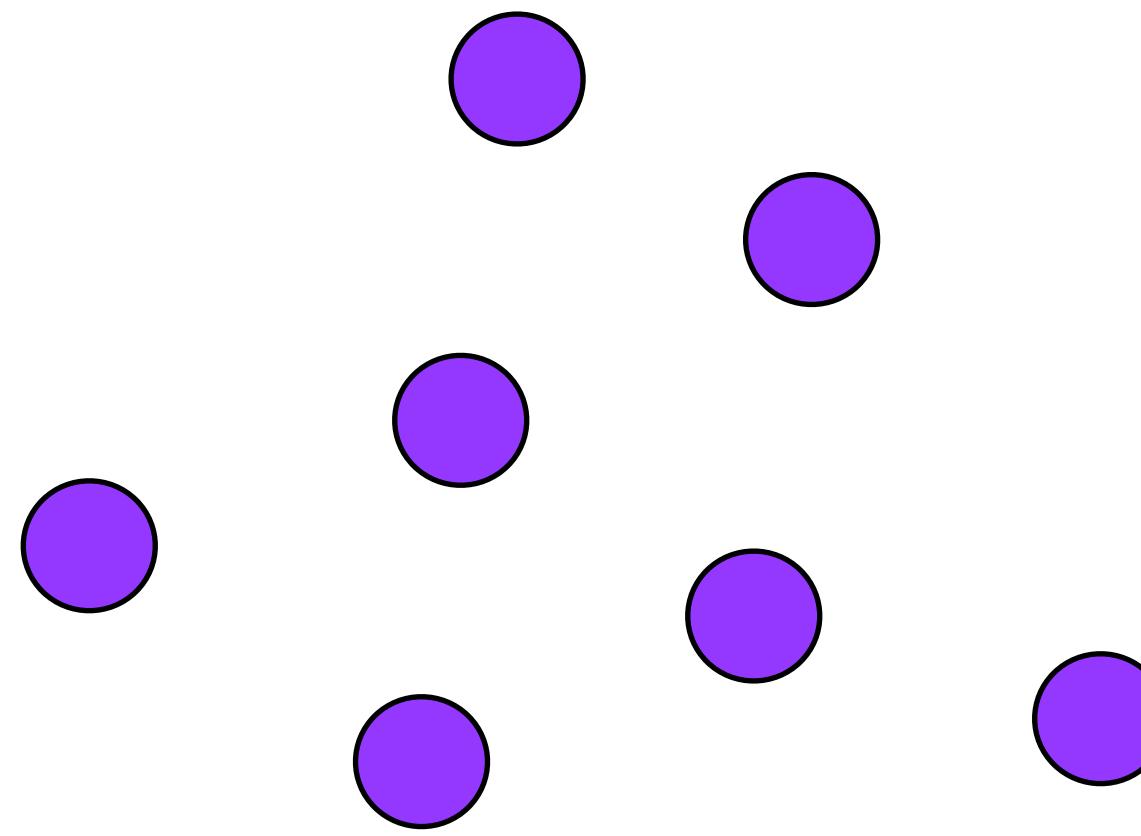
\mathbb{R}^L

Dataset	Classes	Train	Test	BOW Dim	Length	Application
TWITTER	3	2176	932	6344	9.9	Tweets categorized by sentiment
RECIPE-L	20	27841	11933	3590	18.5	Recipe procedures labeled by origin
OHSUMED	10	3999	5153	31789	59.2	medical Abstracts (class subsampled)
20NEWS	20	11293	7528	29671	72	Canonical User-written posts dataset

Consider the task of document classification

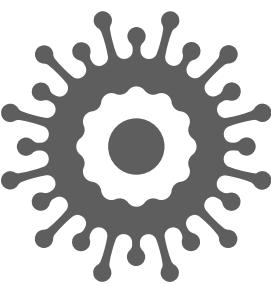
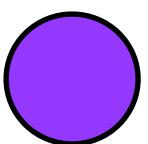
Empirical evaluation

\mathbb{R}^L

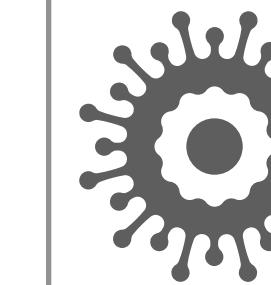


We can check which documents belongs to same class by just
checking how well they align with each other

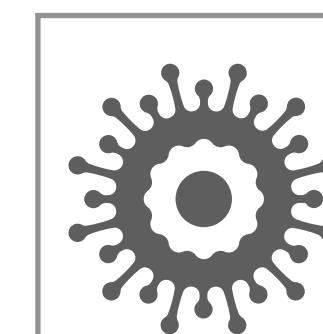
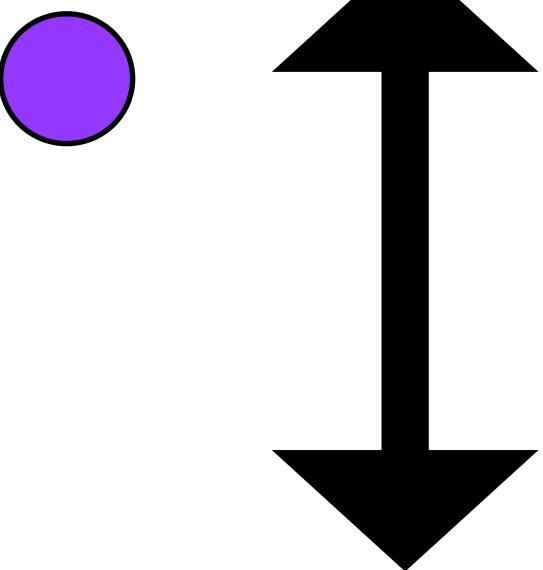
R^L



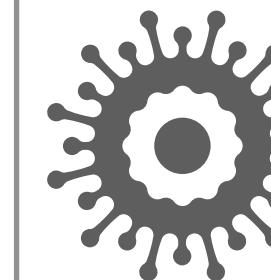
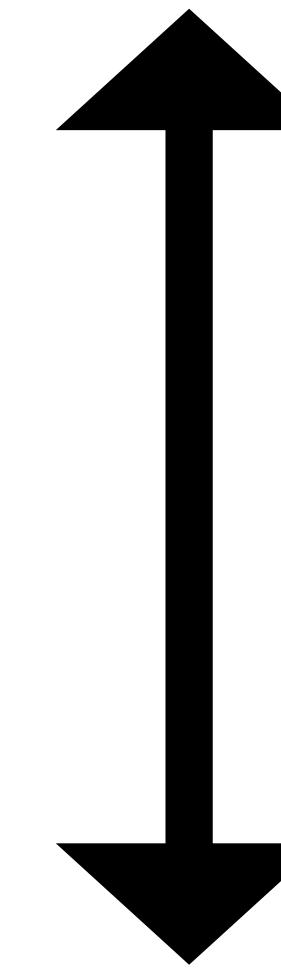
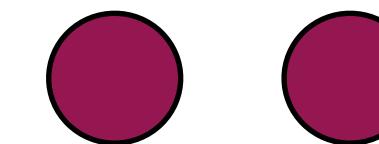
Amazingly, it is effective
against SARS and **MERS**.



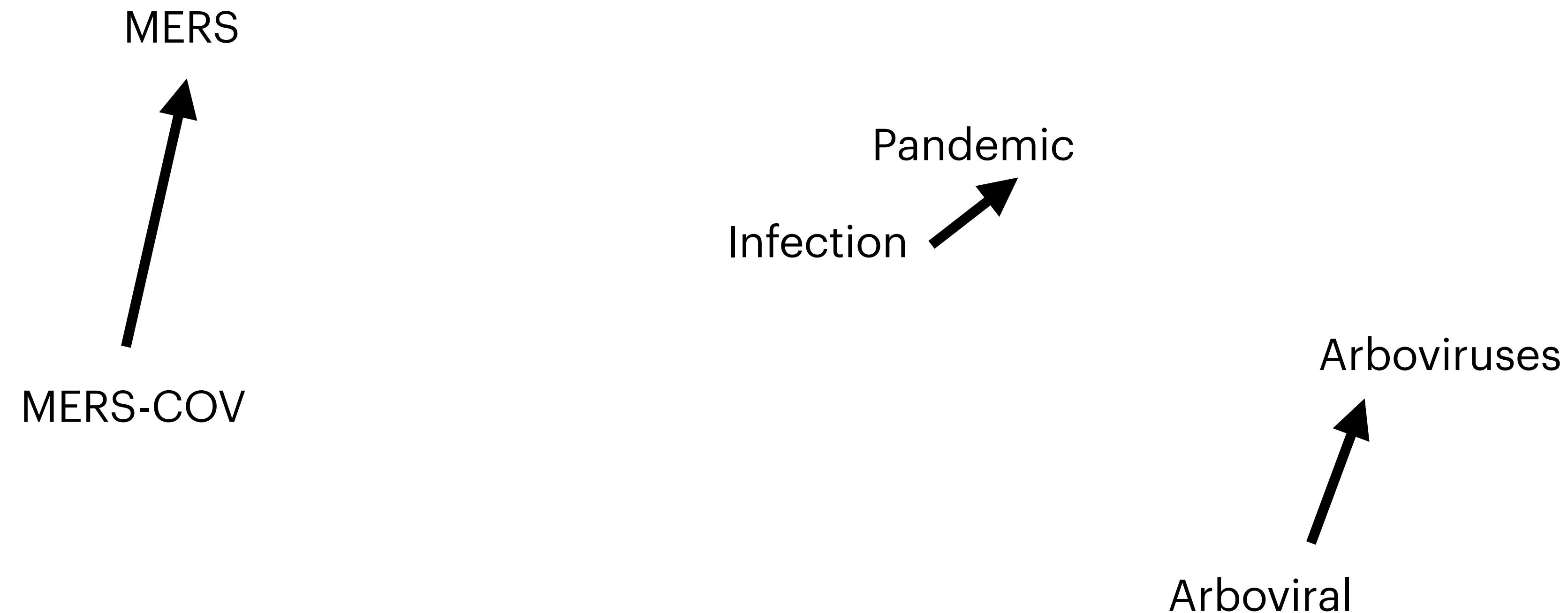
Pandemic arboviruses have
emerged as a major global
health problem in the past
four decades.



The Middle East respiratory
syndrome coronavirus
(MERS-CoV) is an emerging
pathogen...



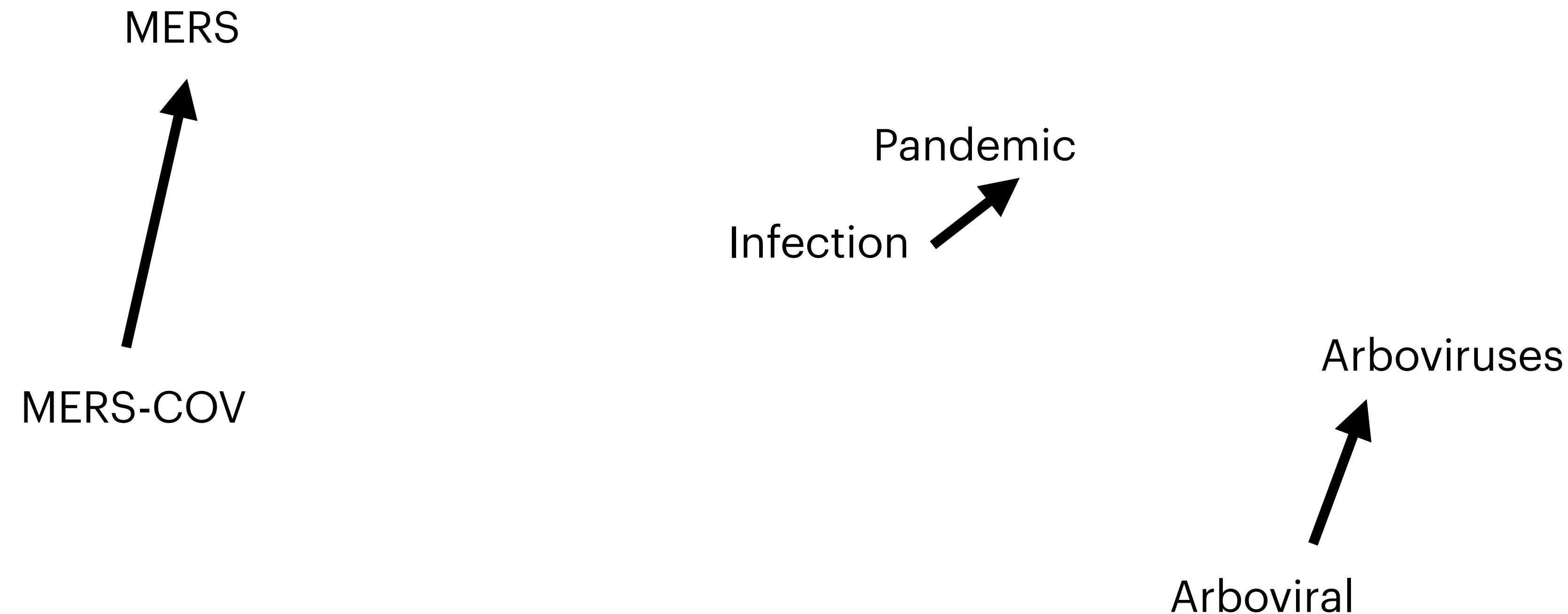
The arboviral infection,
CTF, is transmitted from
the bite of an infected
wood tick.

\mathbb{R}^d 

Word embedding space (here: word2vec)

\mathbb{R}^d

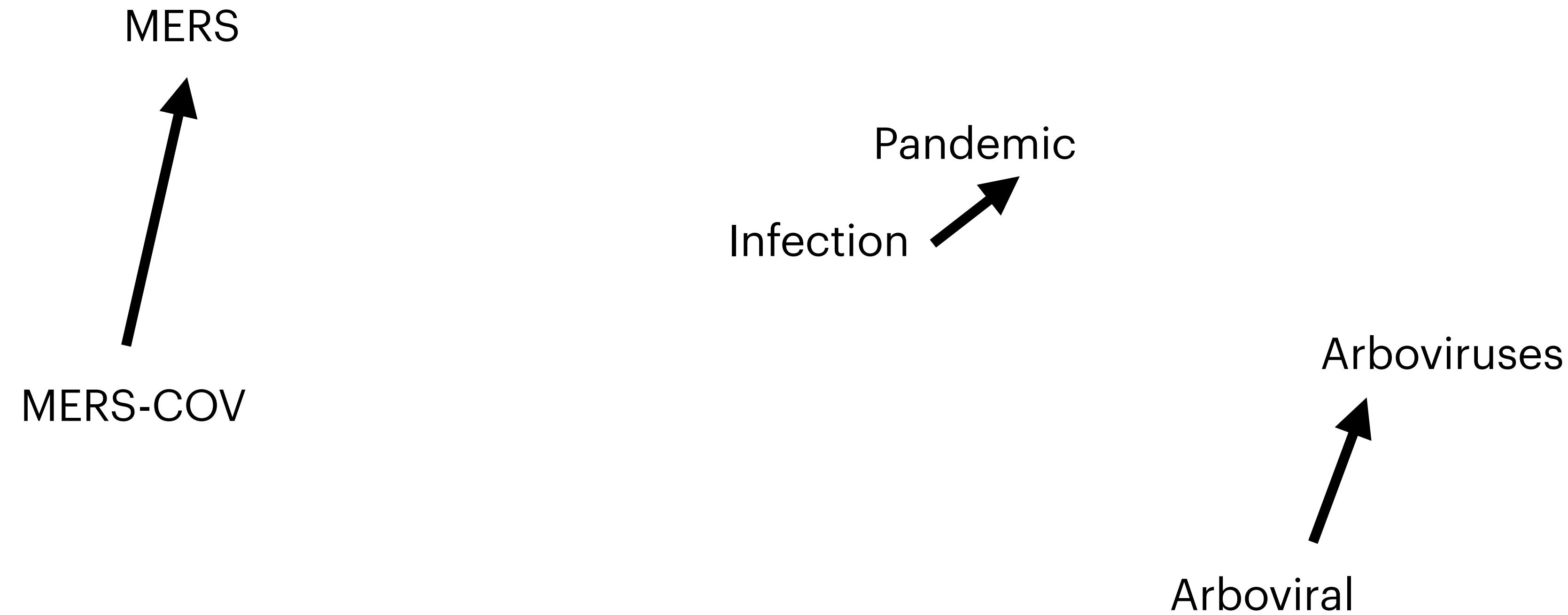
Compute cumulative distance between all words for each pair of documents



Word embedding space (here: word2vec)

\mathbb{R}^d

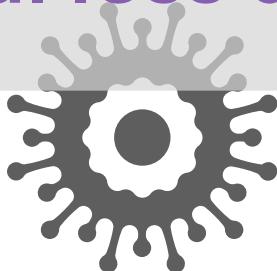
Total less distance travelled implies nearness in \mathbb{R}^L



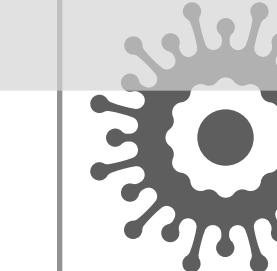
Word embedding space (here: word2vec)

\mathbb{R}^L

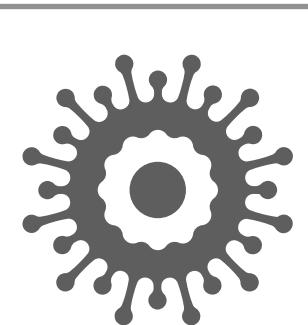
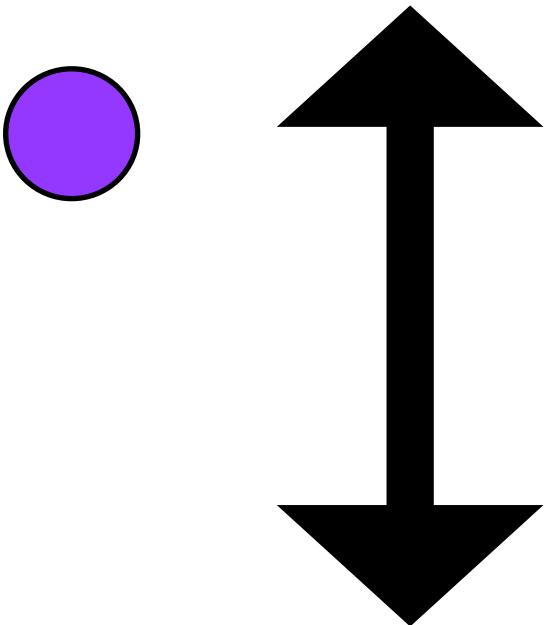
Total less distance travelled implies nearness in \mathbb{R}^L



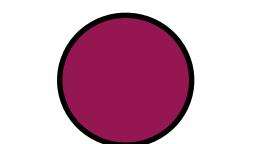
Amazingly, it is effective against SARS and MERS.



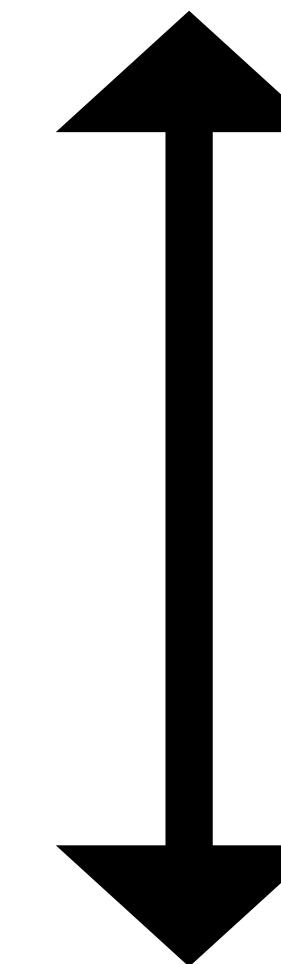
Pandemic arboviruses have emerged as a major global health problem in the past four decades.



The Middle East respiratory syndrome coronavirus (MERS-CoV) is an emerging pathogen...



The arboviral infection, CTF, is transmitted from the bite of an infected wood tick.

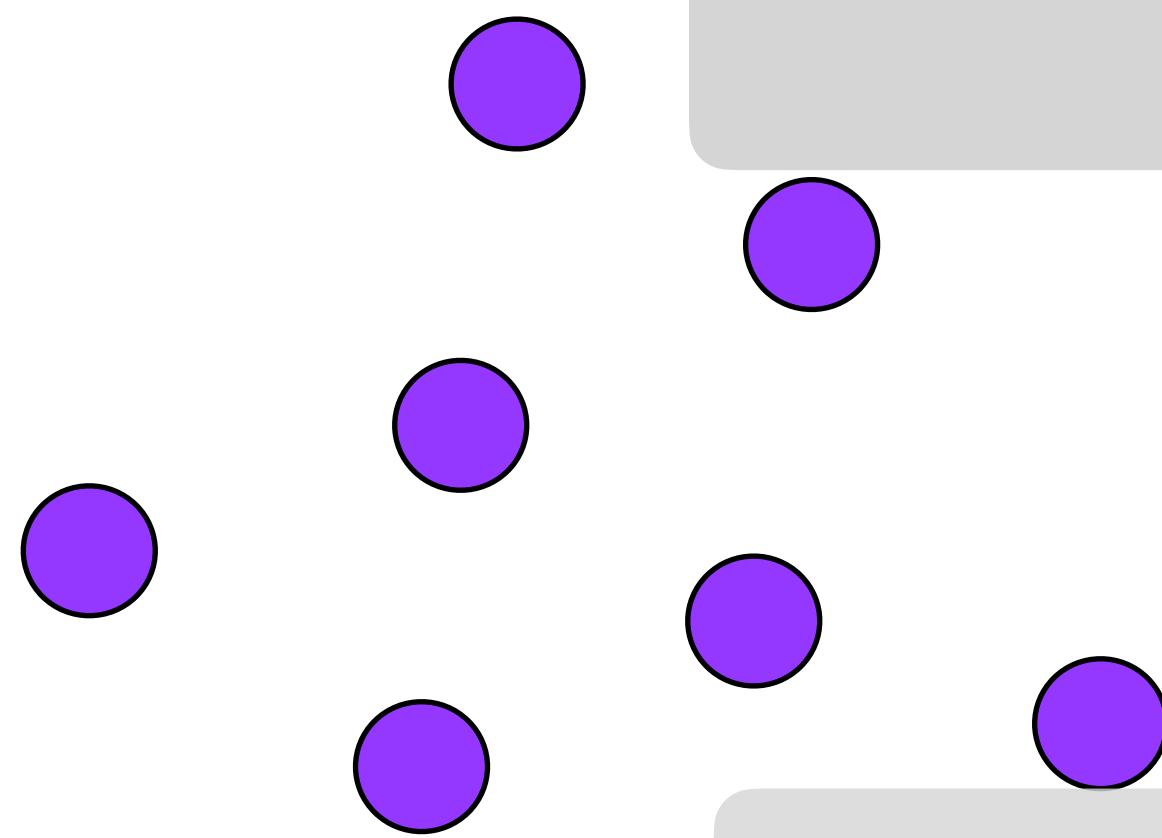


Word Mover's Distance [1]

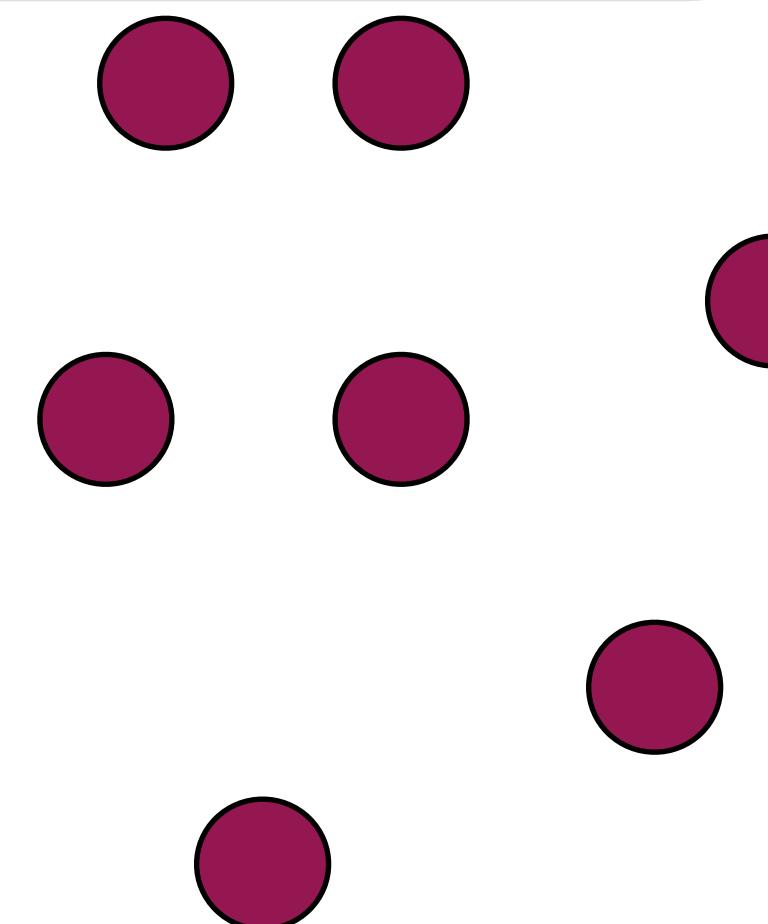
[1] Kusner, M., Sun, Y., Kolkin, N. and Weinberger, K., 2015, June. From word embeddings to document distances. In *International conference on machine learning* (pp. 957-966). PMLR.

\mathbb{R}^L

Word Mover's Distance [1]



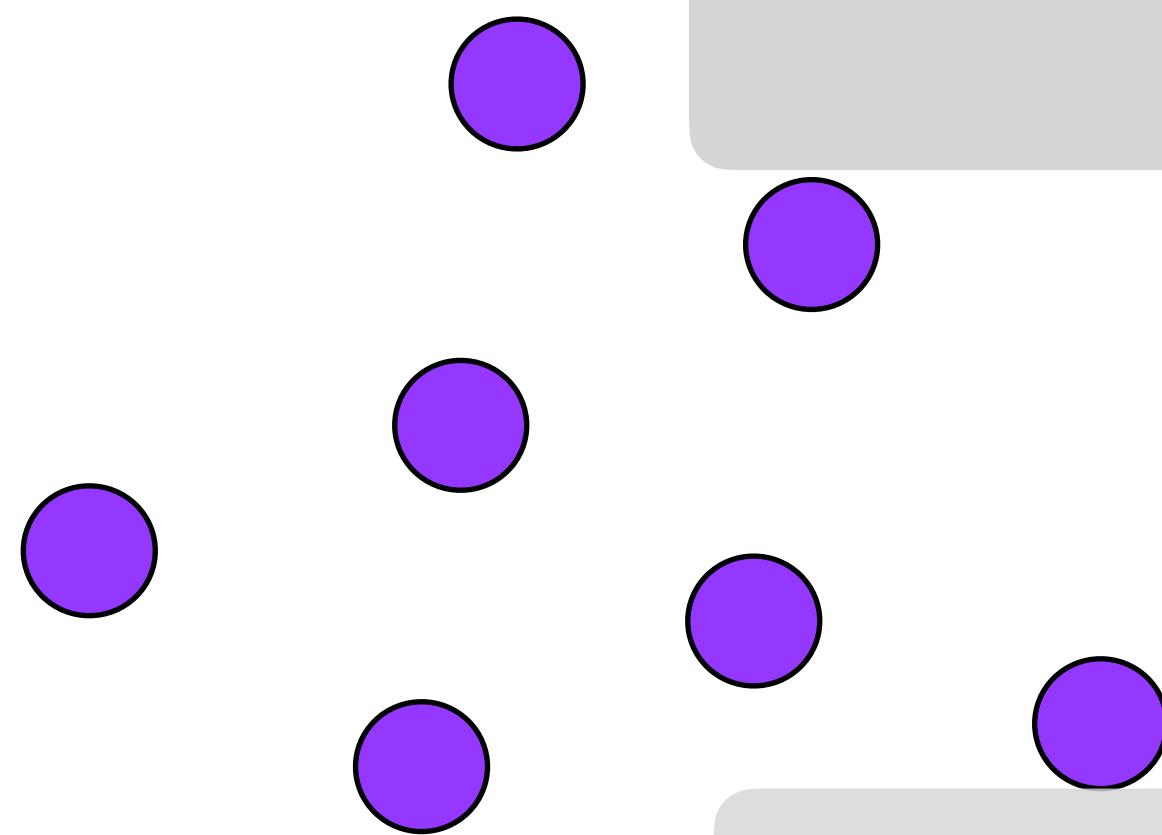
$O(n^2)$ comparisons among documents



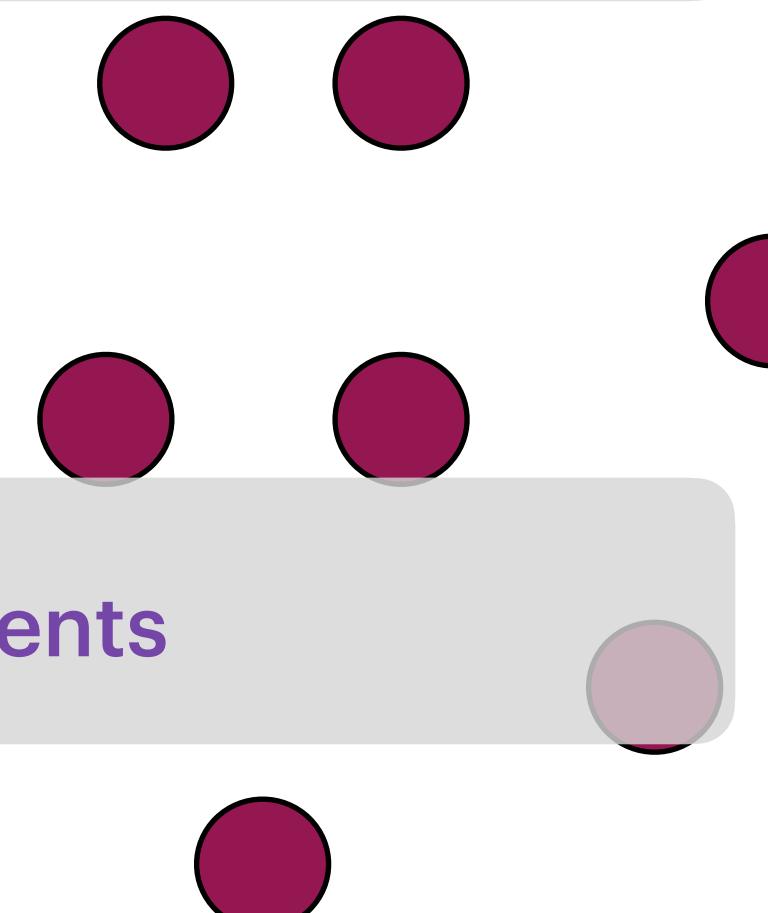
[1] Kusner, M., Sun, Y., Kolkin, N. and Weinberger, K., 2015, June. From word embeddings to document distances. In *International conference on machine learning* (pp. 957-966). PMLR.

\mathbb{R}^L

Word Mover's Distance [1]



$O(n^2)$ comparisons among documents

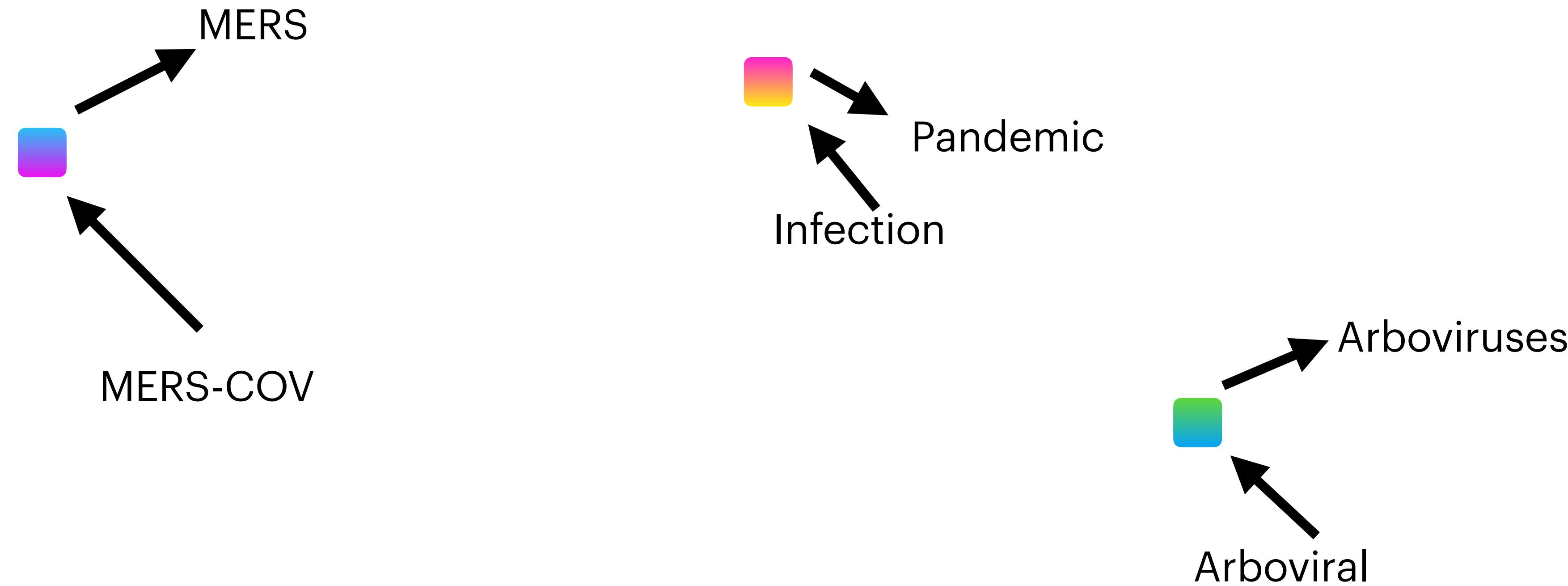


$O(L^3 \log L)$ comparisons between documents

[1] Kusner, M., Sun, Y., Kolkin, N. and Weinberger, K., 2015, June. From word embeddings to document distances. In *International conference on machine learning* (pp. 957-966). PMLR.

\mathbb{R}^d

Word Mover's Embedding [1]

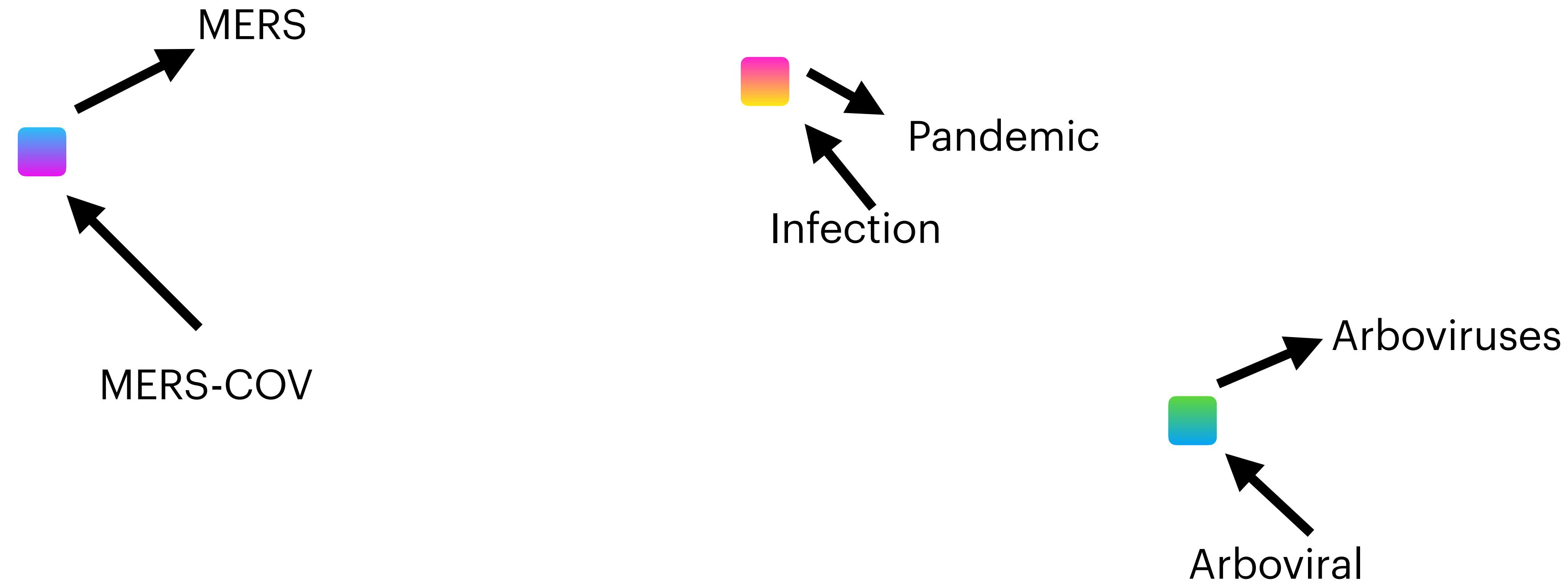


are words from random documents

[1] Wu, L., Yen, I.E., Xu, K., Xu, F., Balakrishnan, A., Chen, P.Y., Ravikumar, P. and Witbrock, M.J., 2018. Word mover's embedding: From word2vec to document embedding. *arXiv preprint arXiv:1811.01713*.

Word Mover's Embedding [1]

\mathbb{R}^d

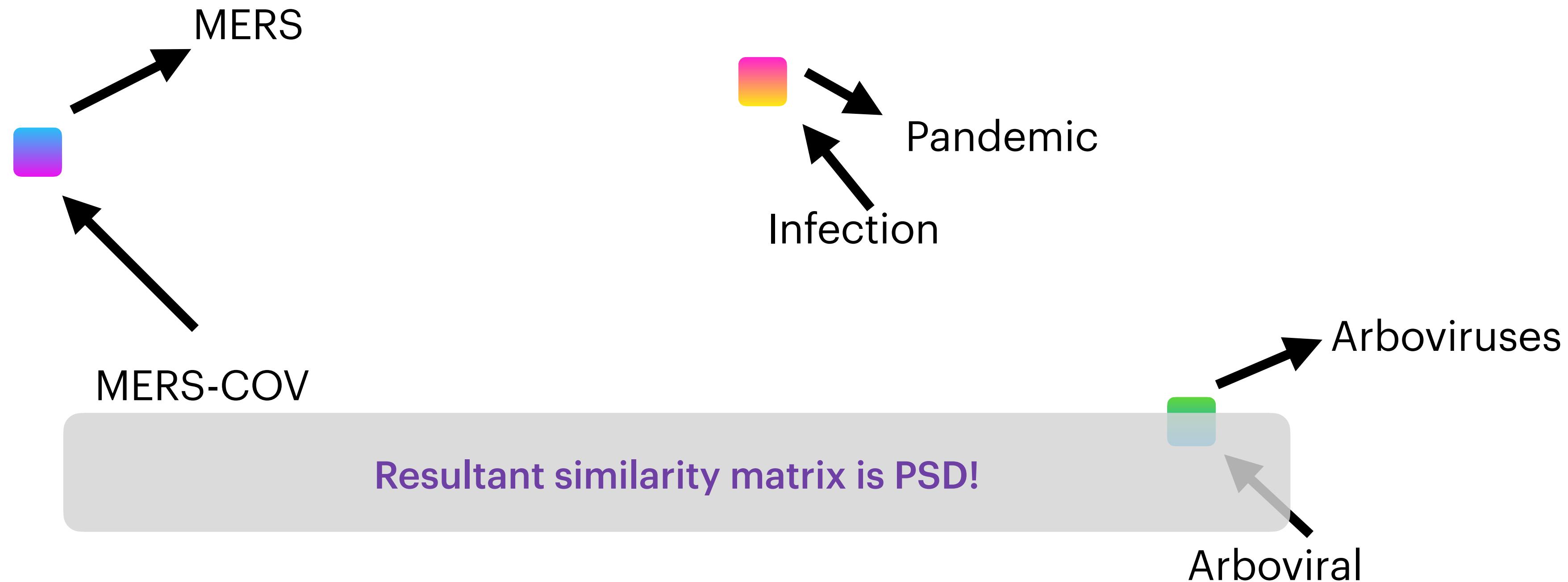


Construct Word Mover's Kernel (WMK) using infinite dimensional feature map to random documents from a given distribution

[1] Wu, L., Yen, I.E., Xu, K., Xu, F., Balakrishnan, A., Chen, P.Y., Ravikumar, P. and Witbrock, M.J., 2018. Word mover's embedding: From word2vec to document embedding. *arXiv preprint arXiv:1811.01713*.

Word Mover's Embedding [1]

\mathbb{R}^d

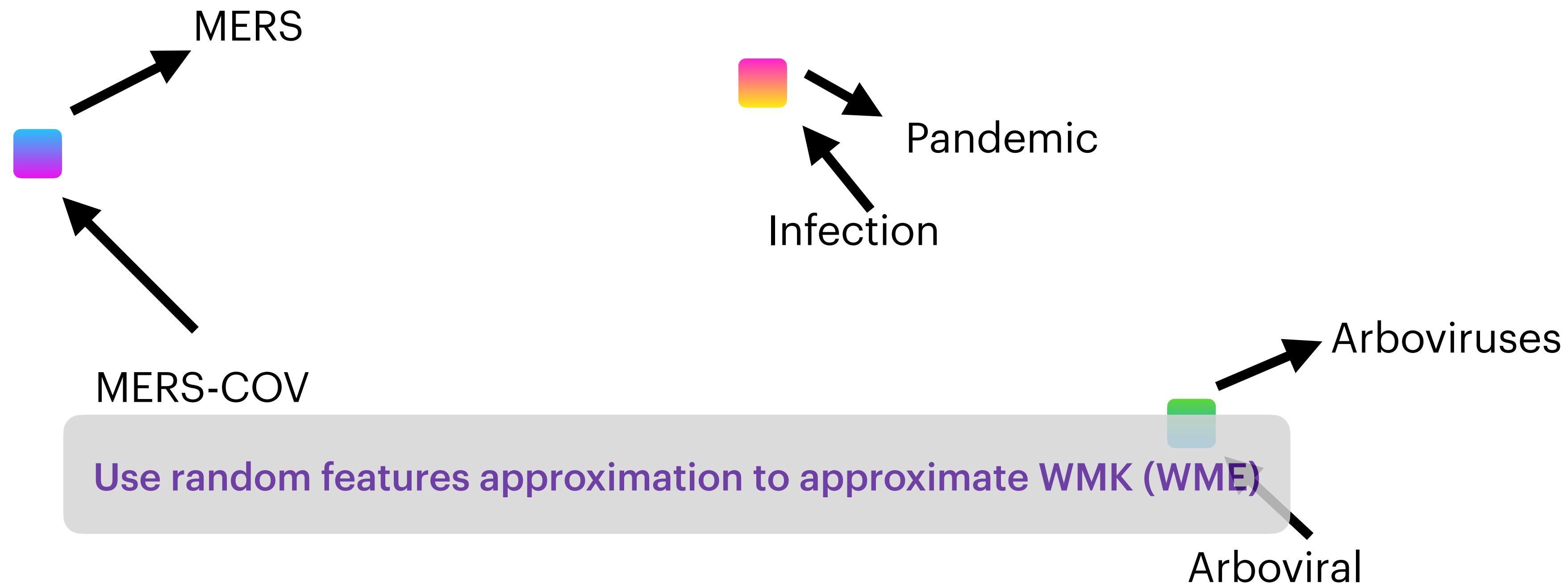


Construct Word Mover's Kernel (WMK) using infinite dimensional feature map to random documents from a given distribution

[1] Wu, L., Yen, I.E., Xu, K., Xu, F., Balakrishnan, A., Chen, P.Y., Ravikumar, P. and Witbrock, M.J., 2018. Word mover's embedding: From word2vec to document embedding. *arXiv preprint arXiv:1811.01713*.

\mathbb{R}^d

Word Mover's Embedding [1]

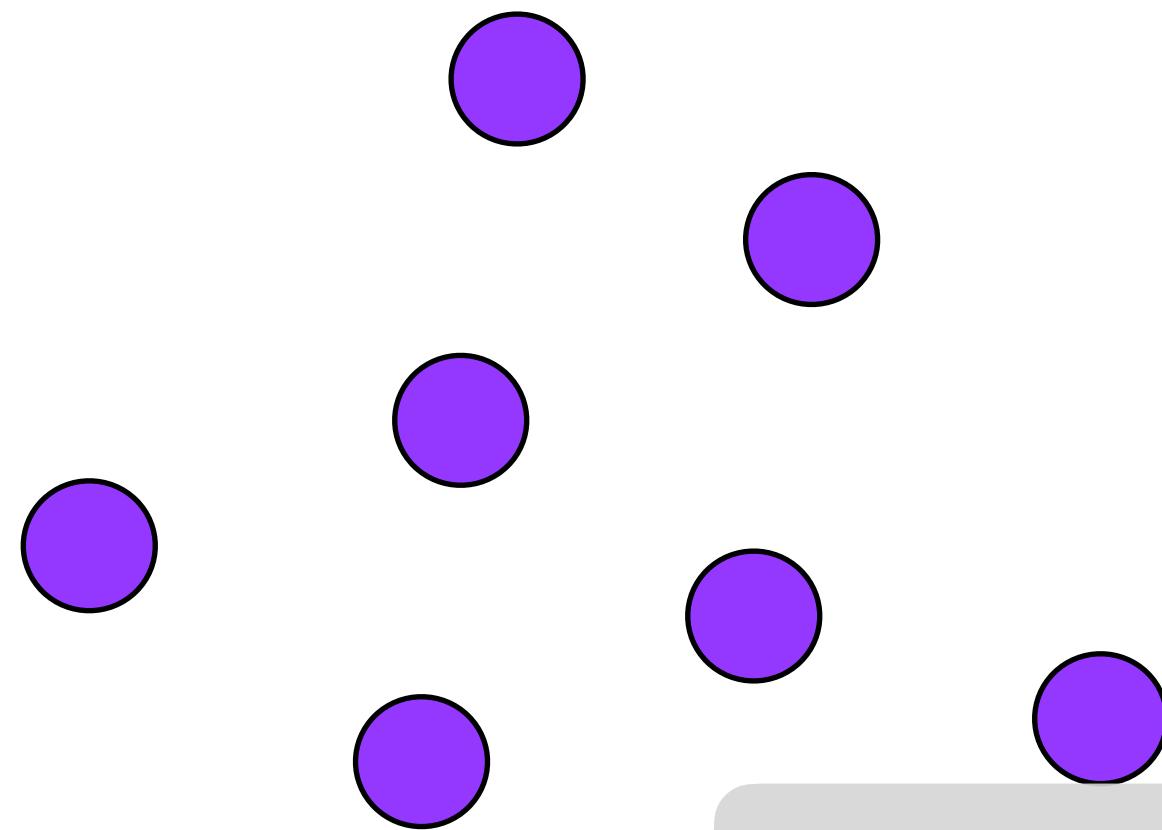


Construct Word Mover's Kernel (WMK) using infinite dimensional feature map to random documents from a given distribution

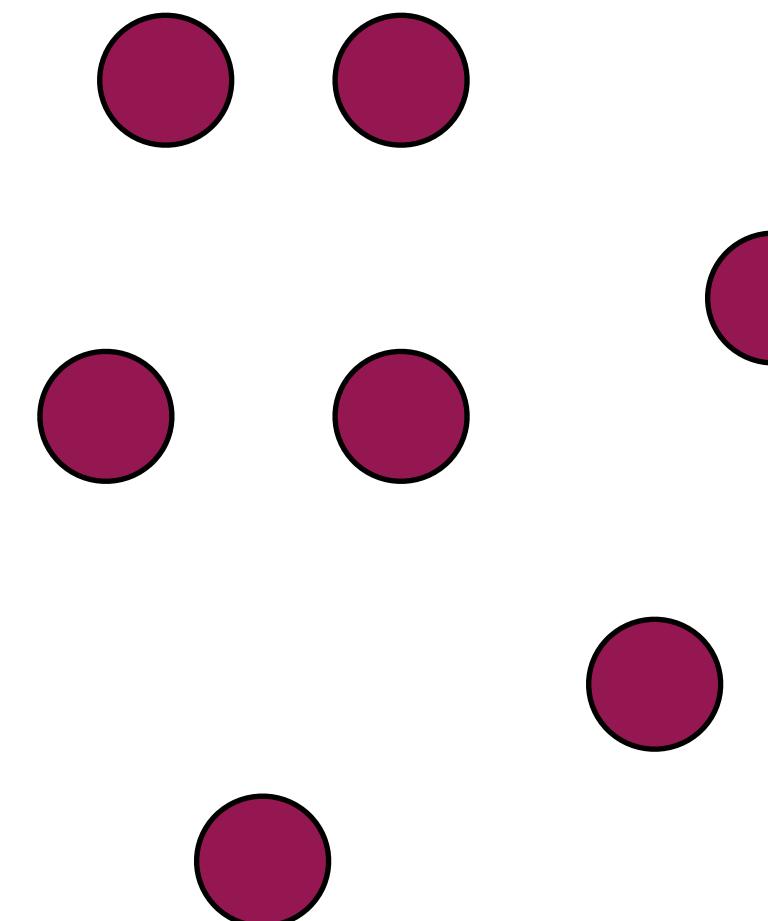
[1] Wu, L., Yen, I.E., Xu, K., Xu, F., Balakrishnan, A., Chen, P.Y., Ravikumar, P. and Witbrock, M.J., 2018. Word mover's embedding: From word2vec to document embedding. *arXiv preprint arXiv:1811.01713*.

Our Approach

\mathbb{R}^L

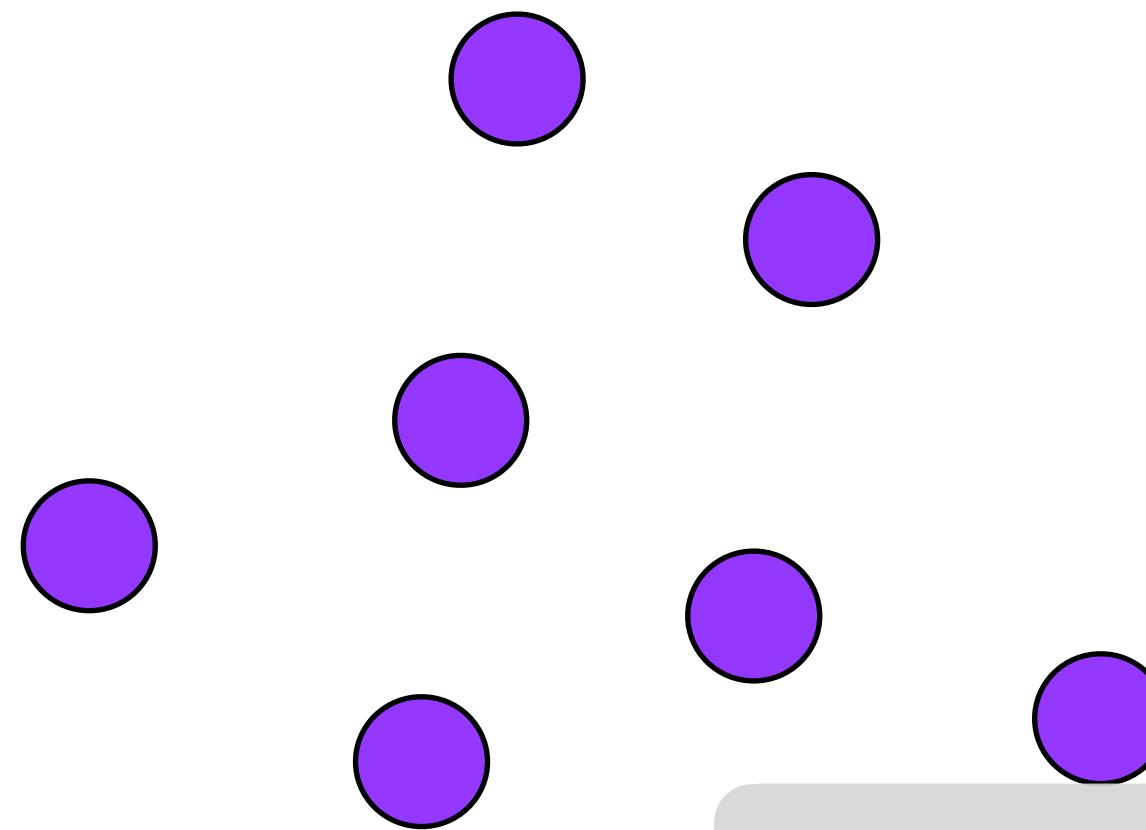


Similarity between documents x, y : $\exp(-\gamma \text{WMD}(x, y))$

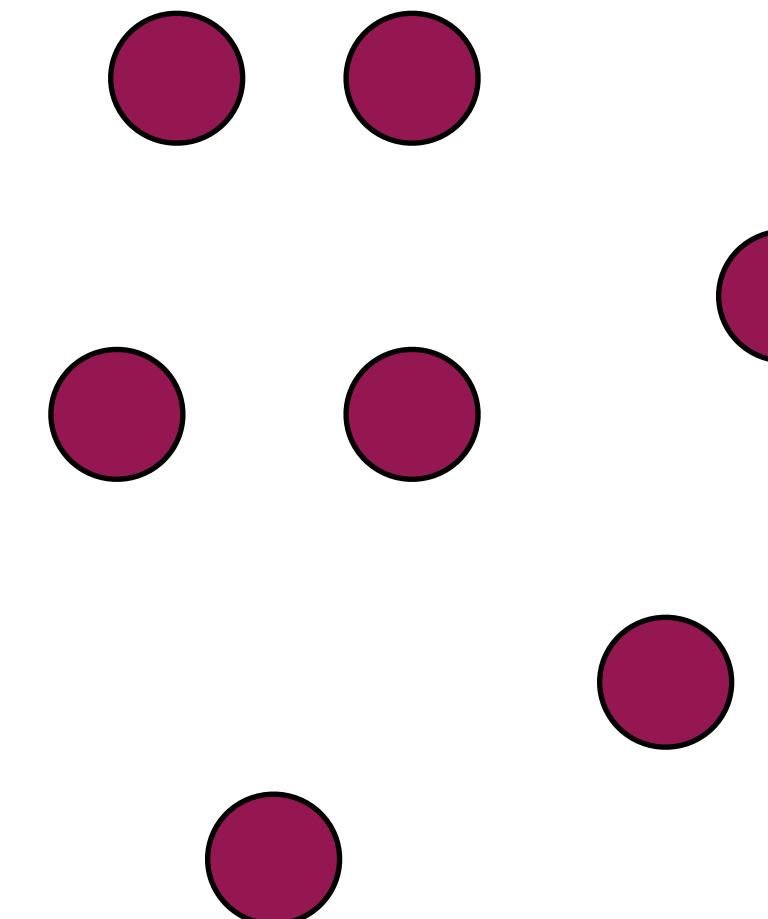


Our Approach

\mathbb{R}^L

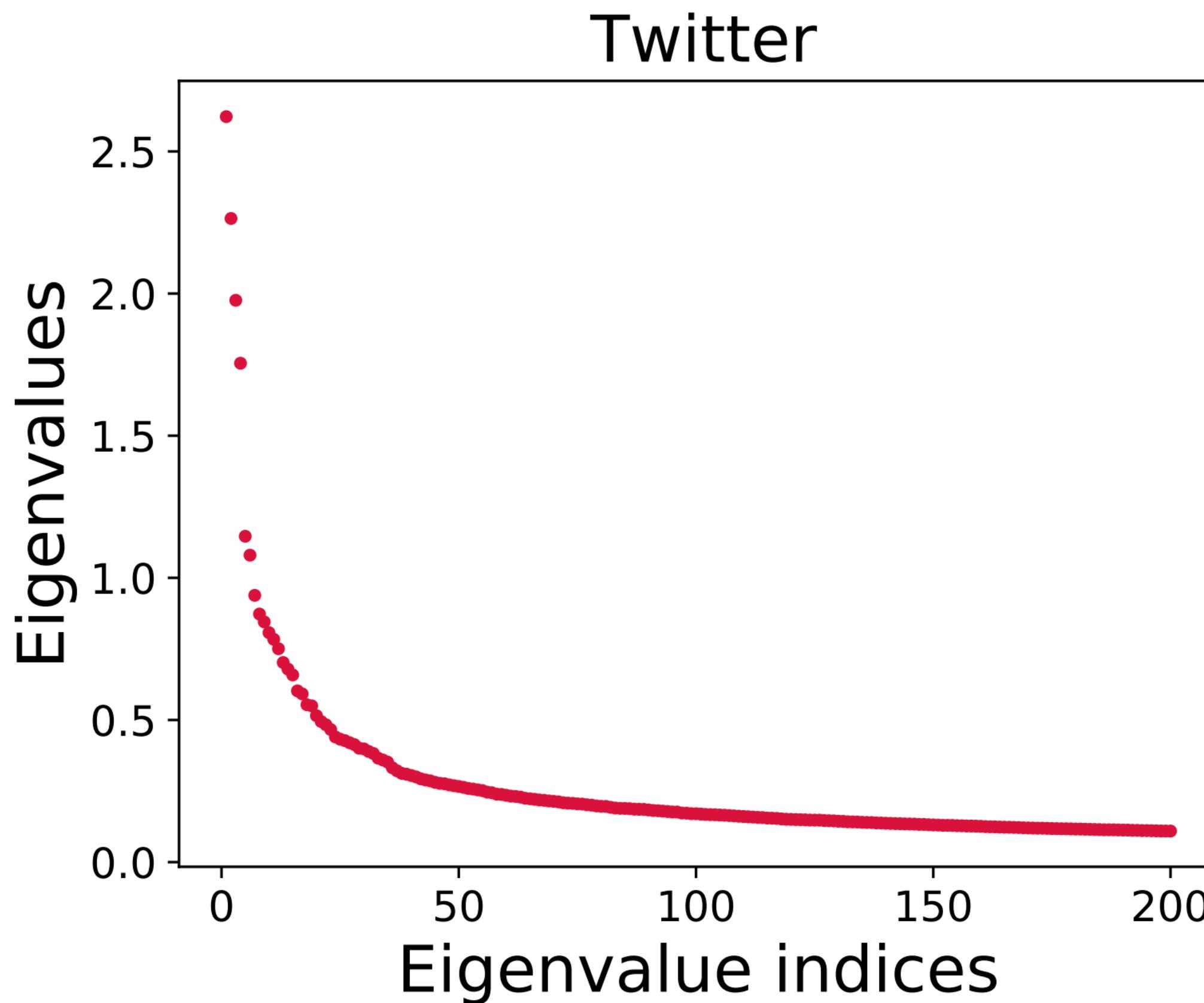


Similarity between documents x, y : $\exp(-\gamma \text{WMD}(x, y))$



Similarity matrix not PSD but near-PSD

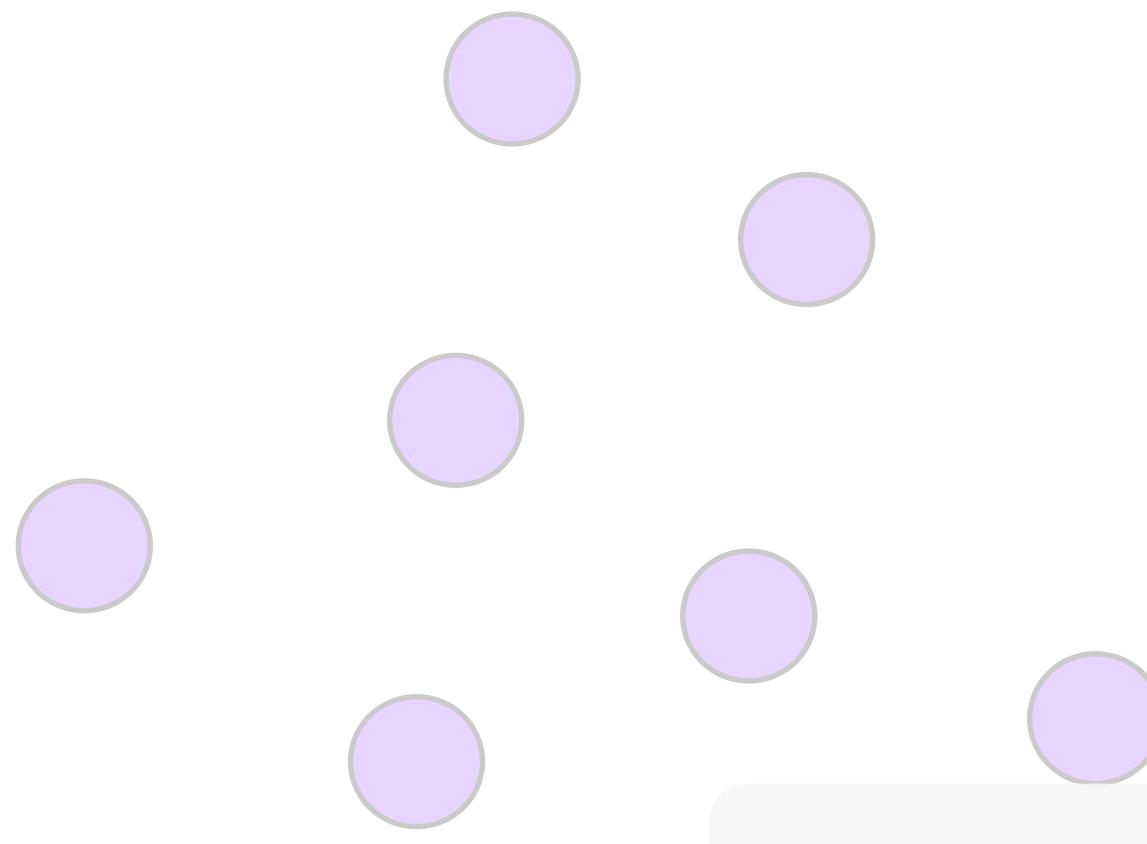
Our Approach



Similarity matrix not PSD but near-PSD

Our Approach

\mathbb{R}^L



Similarity between documents x, y : $\exp(-\gamma \text{WMD}(x, y))$



But $O(n^2)$ WMD computations is prohibitively expensive!

Similarity matrix not PSD but near-PSD

Sentence Similarity Task

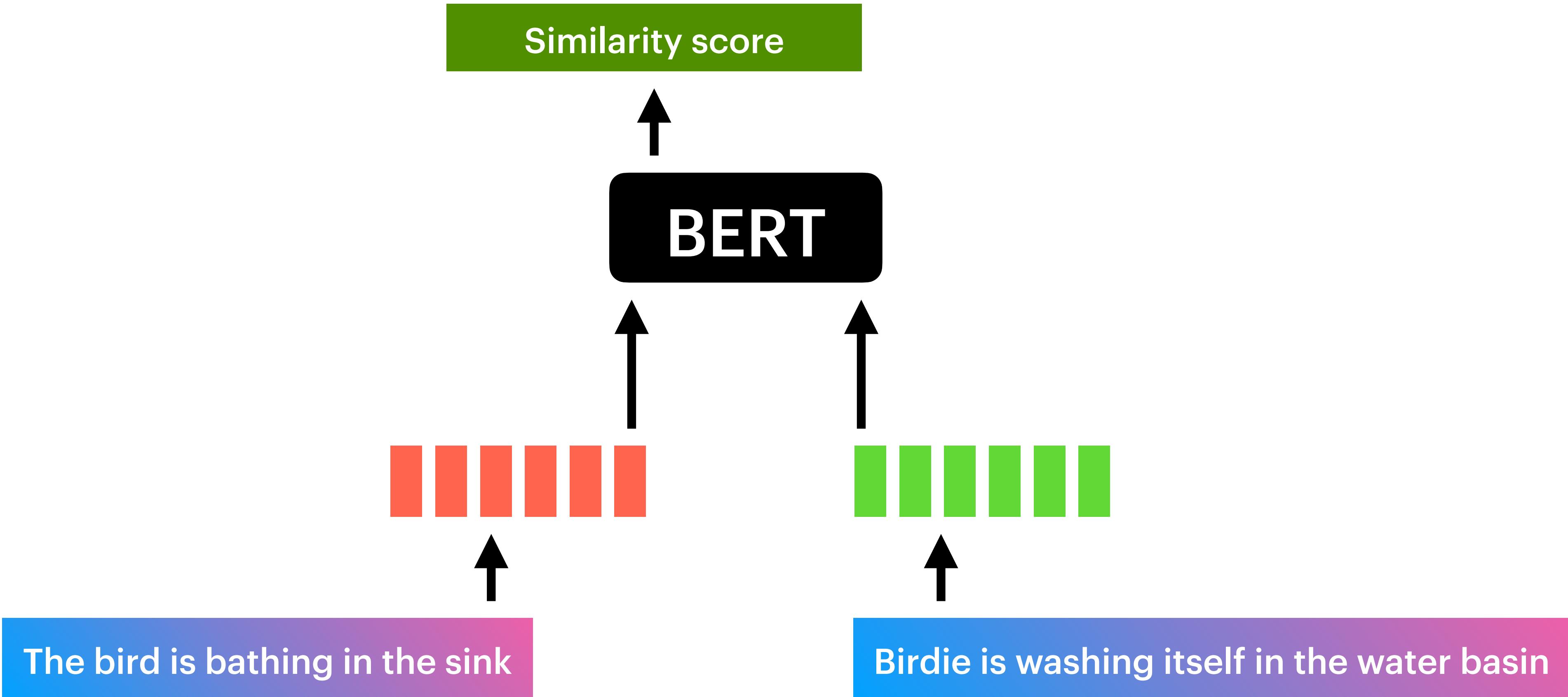
Score	English	Spanish
5/4	<p><i>The two sentences are completely equivalent, as they mean the same thing.</i></p> <p>The bird is bathing in the sink. Birdie is washing itself in the water basin.</p>	<p>El pájaro se esta bañando en el lavabo. El pájaro se está lavando en el aguamanil.</p>
4	<p><i>The two sentences are mostly equivalent, but some unimportant details differ.</i></p> <p>In May 2010, the troops attempted to invade Kabul. The US army invaded Kabul on May 7th last year, 2010.</p>	
3	<p><i>The two sentences are roughly equivalent, but some important information differs/missing.</i></p> <p>John said he is considered a witness but not a suspect. "He is not a suspect anymore." John said.</p>	<p>John dijo que él es considerado como testigo, y no como sospechoso. "Él ya no es un sospechoso," John dijo.</p>
2	<p><i>The two sentences are not equivalent, but share some details.</i></p> <p>They flew out of the nest in groups. They flew into the nest together.</p>	<p>Ellos volaron del nido en grupos. Volaron hacia el nido juntos.</p>
1	<p><i>The two sentences are not equivalent, but are on the same topic.</i></p> <p>The woman is playing the violin. The young lady enjoys listening to the guitar.</p>	<p>La mujer está tocando el violín. La joven disfruta escuchar la guitarra.</p>
0	<p><i>The two sentences are completely dissimilar.</i></p> <p>John went horse back riding at dawn with a whole group of friends. Sunrise at dawn is a magnificent view to take in if you wake up early enough for it.</p>	<p>Al amanecer, Juan se fue a montar a caballo con un grupo de amigos. La salida del sol al amanecer es una magnífica vista que puede presenciar si usted se despierta lo suficientemente temprano para verla.</p>

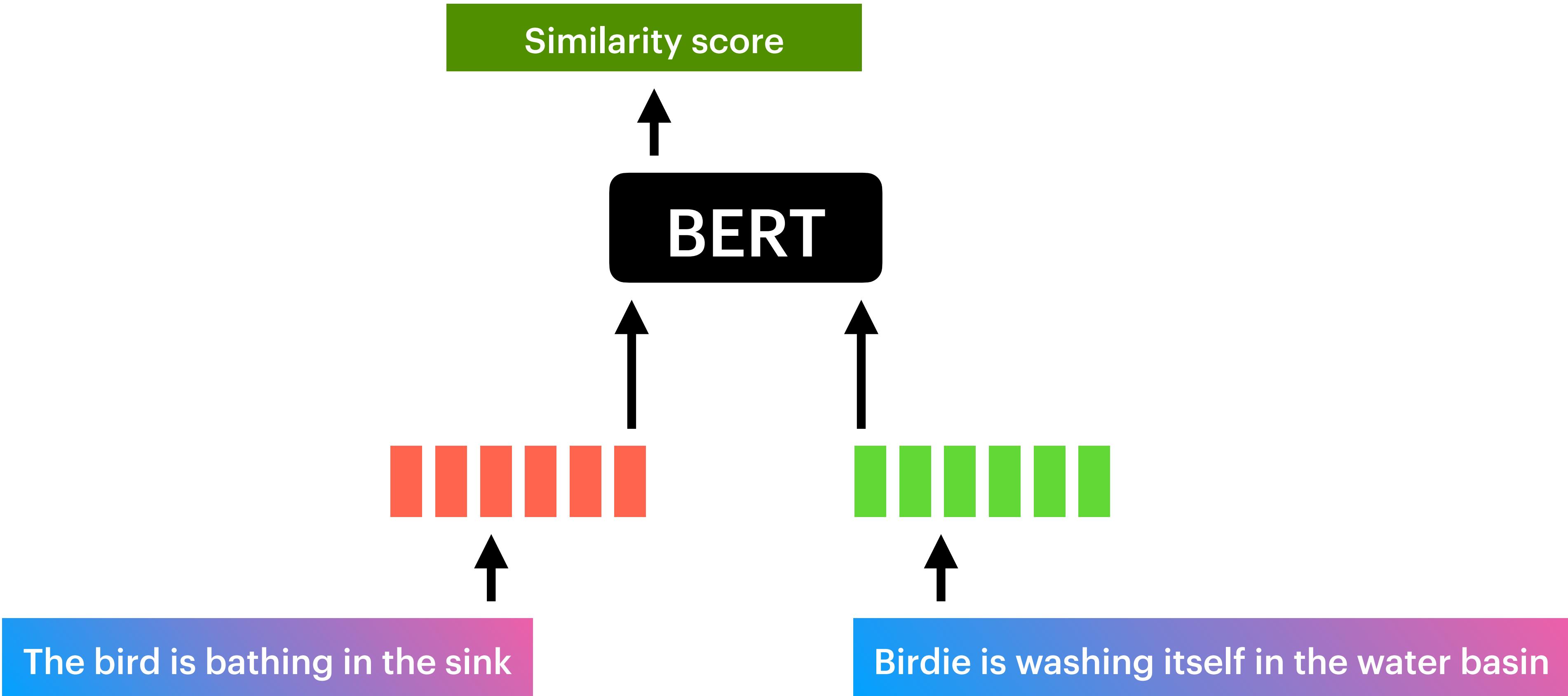
Sentence Similarity Task

Score	English	Spanish
5/4	<p><i>The two sentences are completely equivalent, as they mean the same thing.</i></p> <p>The bird is bathing in the sink. Birdie is washing itself in the water basin.</p>	<p>El pájaro se esta bañando en el lavabo. El pájaro se está lavando en el aguamanil.</p>
4	<p><i>The two sentences are mostly equivalent, but some unimportant details differ.</i></p> <p>In May 2010, the troops attempted to invade Kabul.</p>	

Dataset	Score range	Train	Test	Application
STS-B	1-5	5749	3000	Semantic similarity of sentence pairs based on human annotations
MRPC	0-1	3668	816	Semantic equivalence of sentence pairs
RTE	0-1	2490	554	Text entailment of news and Wikipedia articles

	<p>They flew out of the nest in groups. They flew into the nest together.</p>	<p>ELLOS VOLARON DEL NIDO EN GRUPOS. Volaron hacia el nido juntos.</p>
1	<p><i>The two sentences are not equivalent, but are on the same topic.</i></p> <p>The woman is playing the violin. The young lady enjoys listening to the guitar.</p>	<p>La mujer está tocando el violín. La joven disfruta escuchar la guitarra.</p>
0	<p><i>The two sentences are completely dissimilar.</i></p> <p>John went horse back riding at dawn with a whole group of friends. Sunrise at dawn is a magnificent view to take in if you wake up early enough for it.</p>	<p>Al amanecer, Juan se fue a montar a caballo con un grupo de amigos. La salida del sol al amanecer es una magnífica vista que puede presenciar si usted se despierta lo suficientemente temprano para verla.</p>

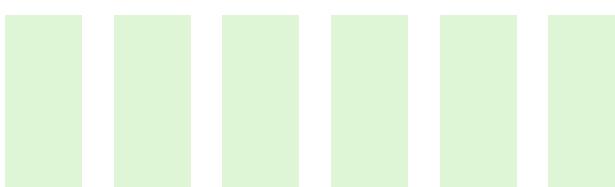
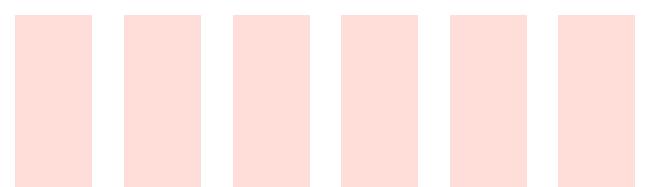




$O(n^2)$ similarities for all pairs

Similarity score

BERT



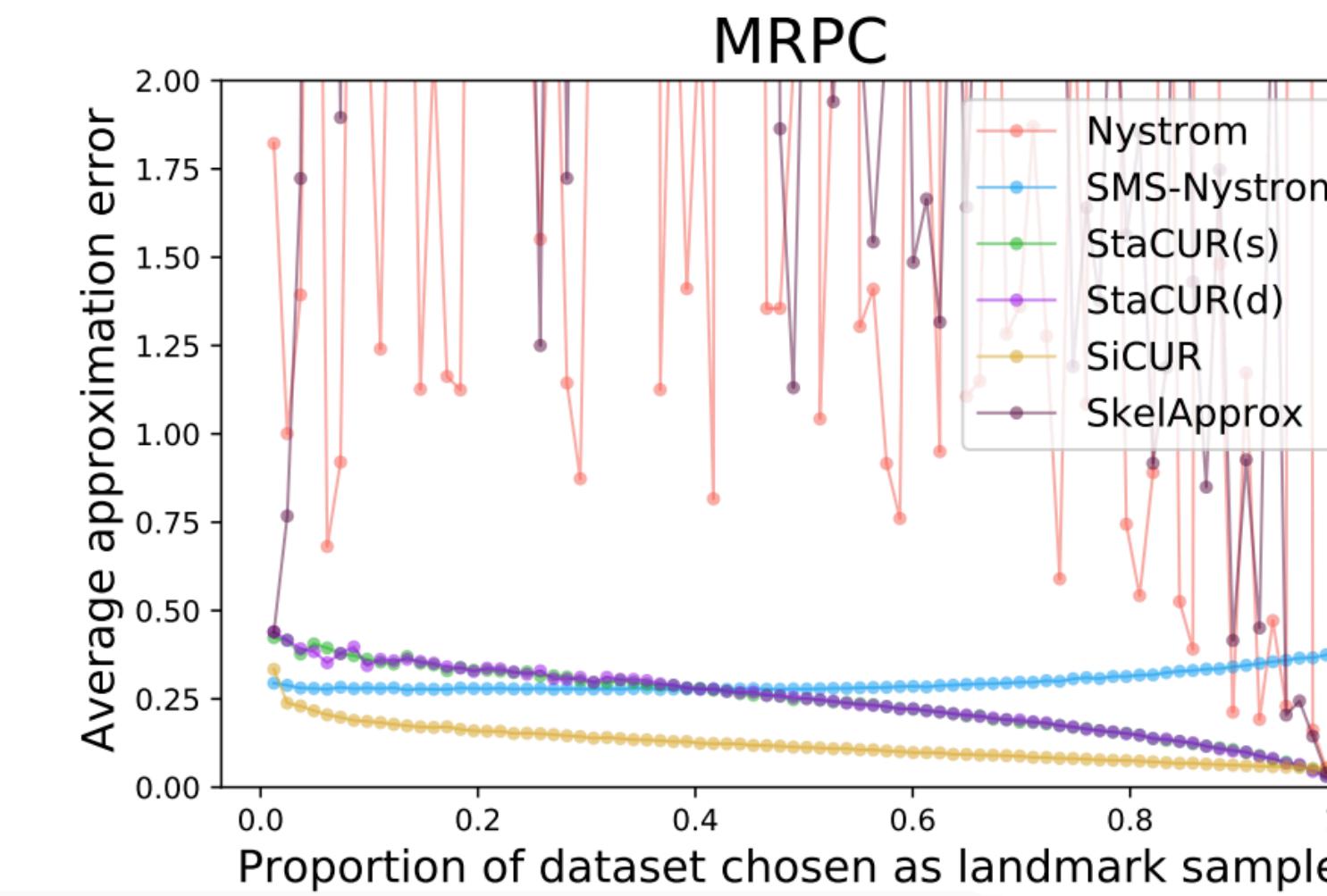
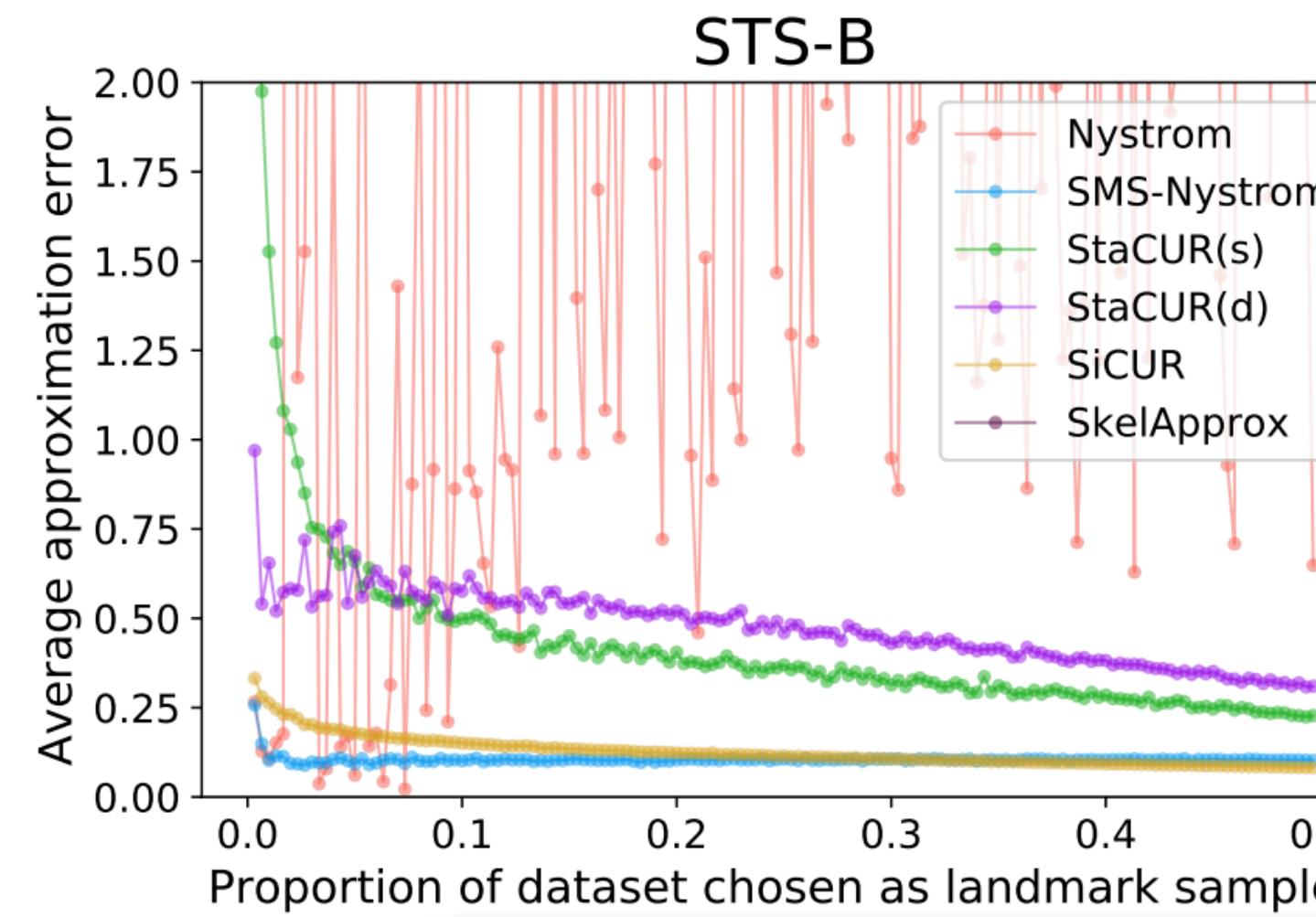
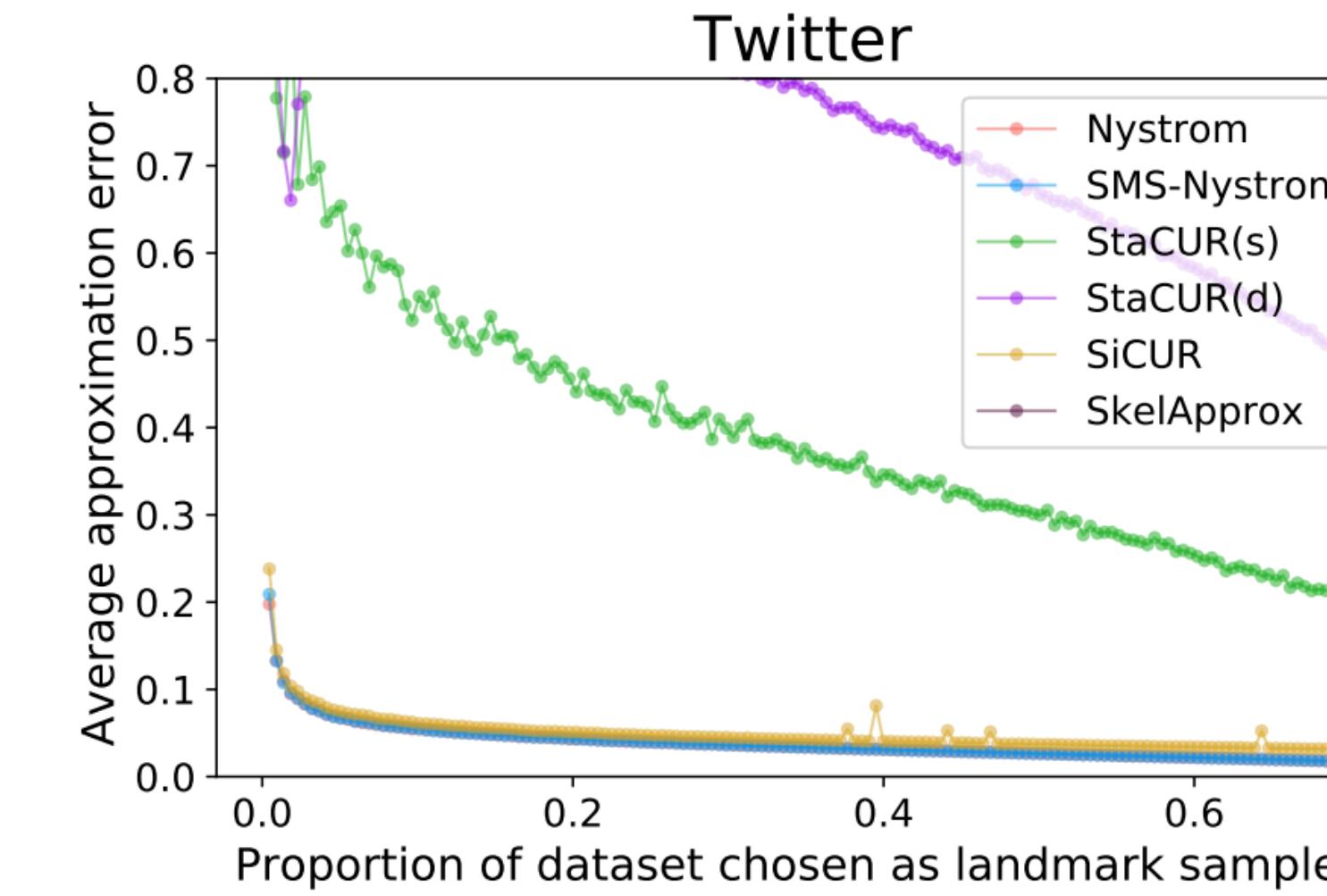
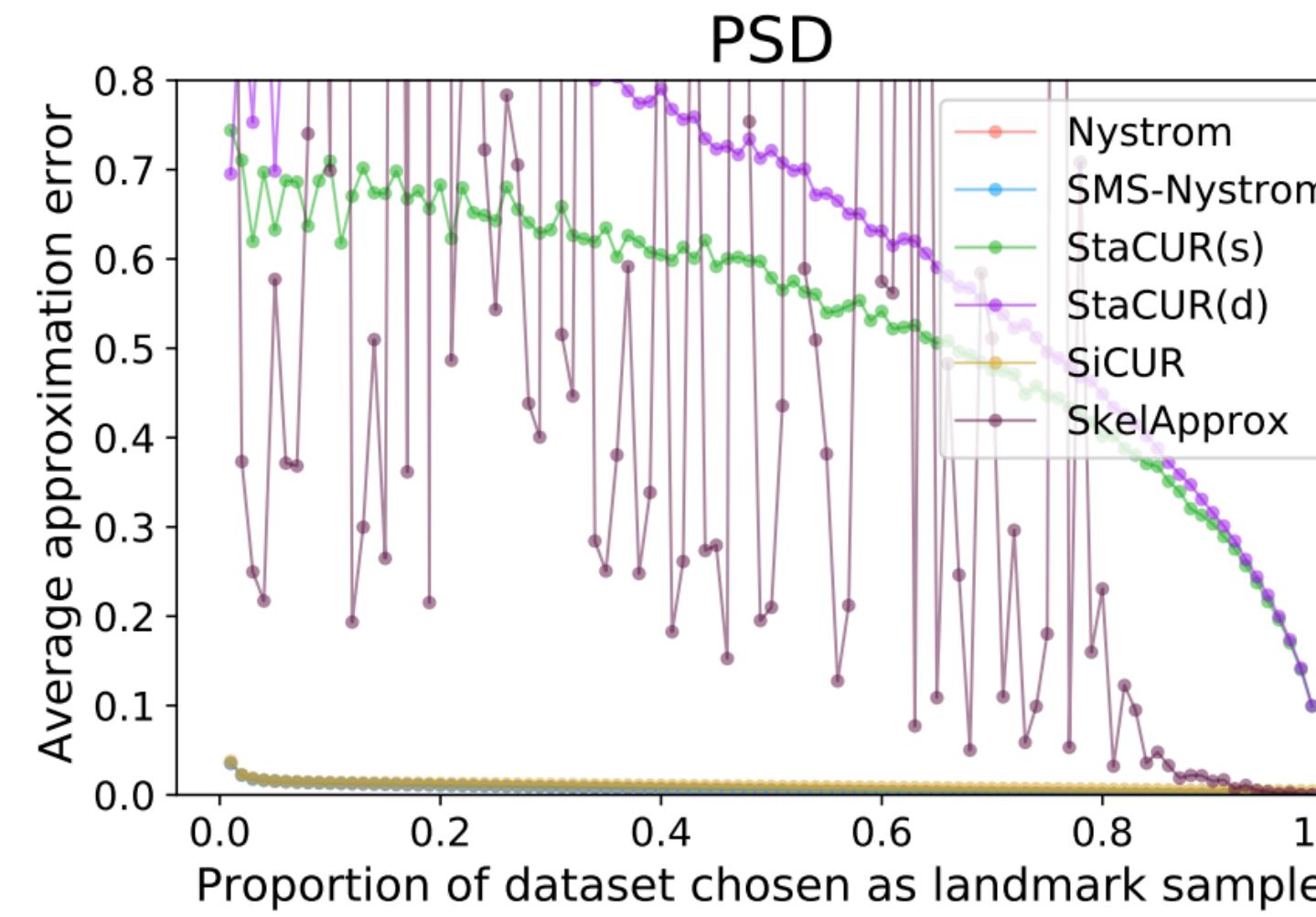
But $O(n^2)$ evaluations of BERT is prohibitively expensive!

The bird is bathing in the sink

Birdie is washing itself in the water basin

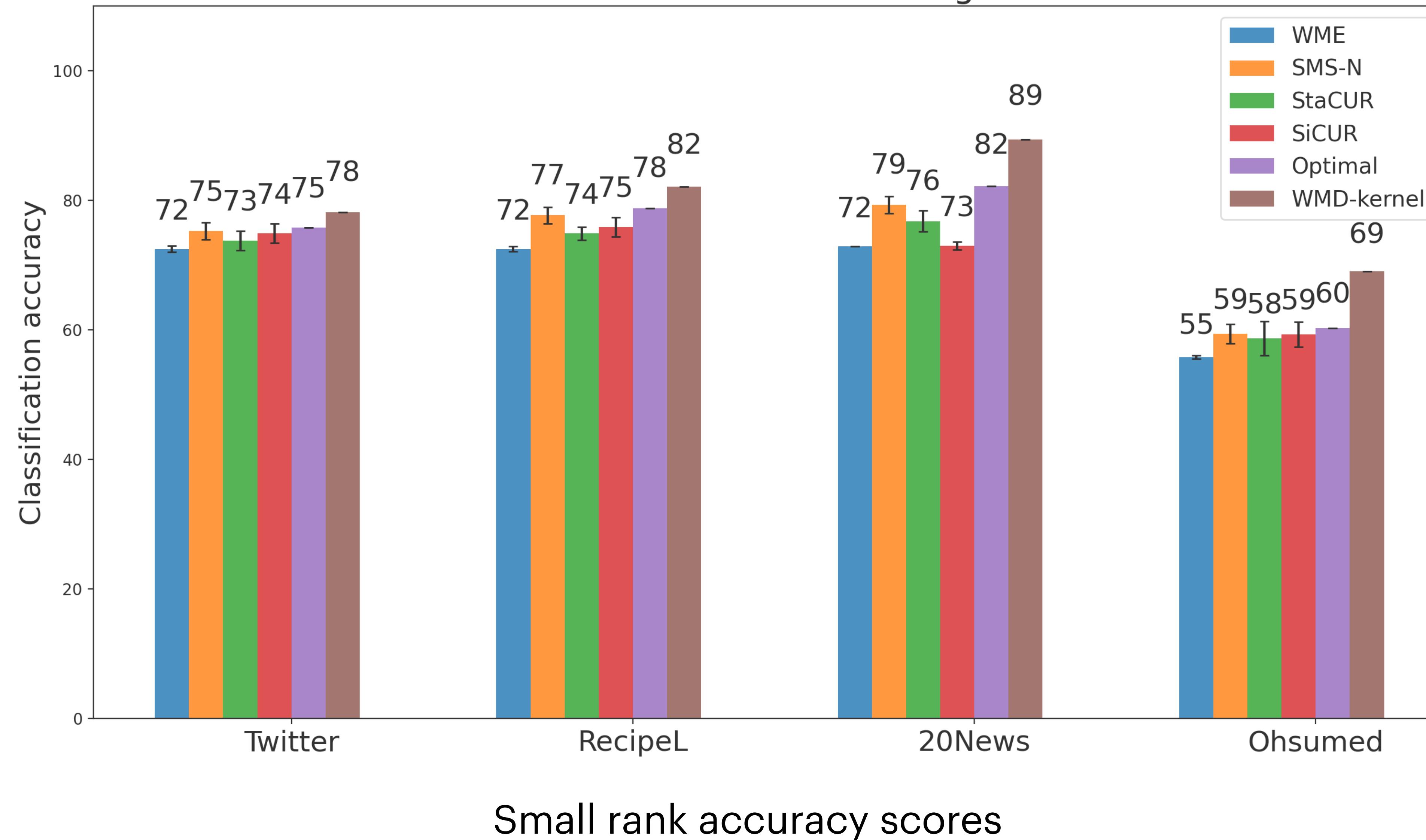
$O(n^2)$ similarities for all pairs

Comparing Approximation on Given Datasets



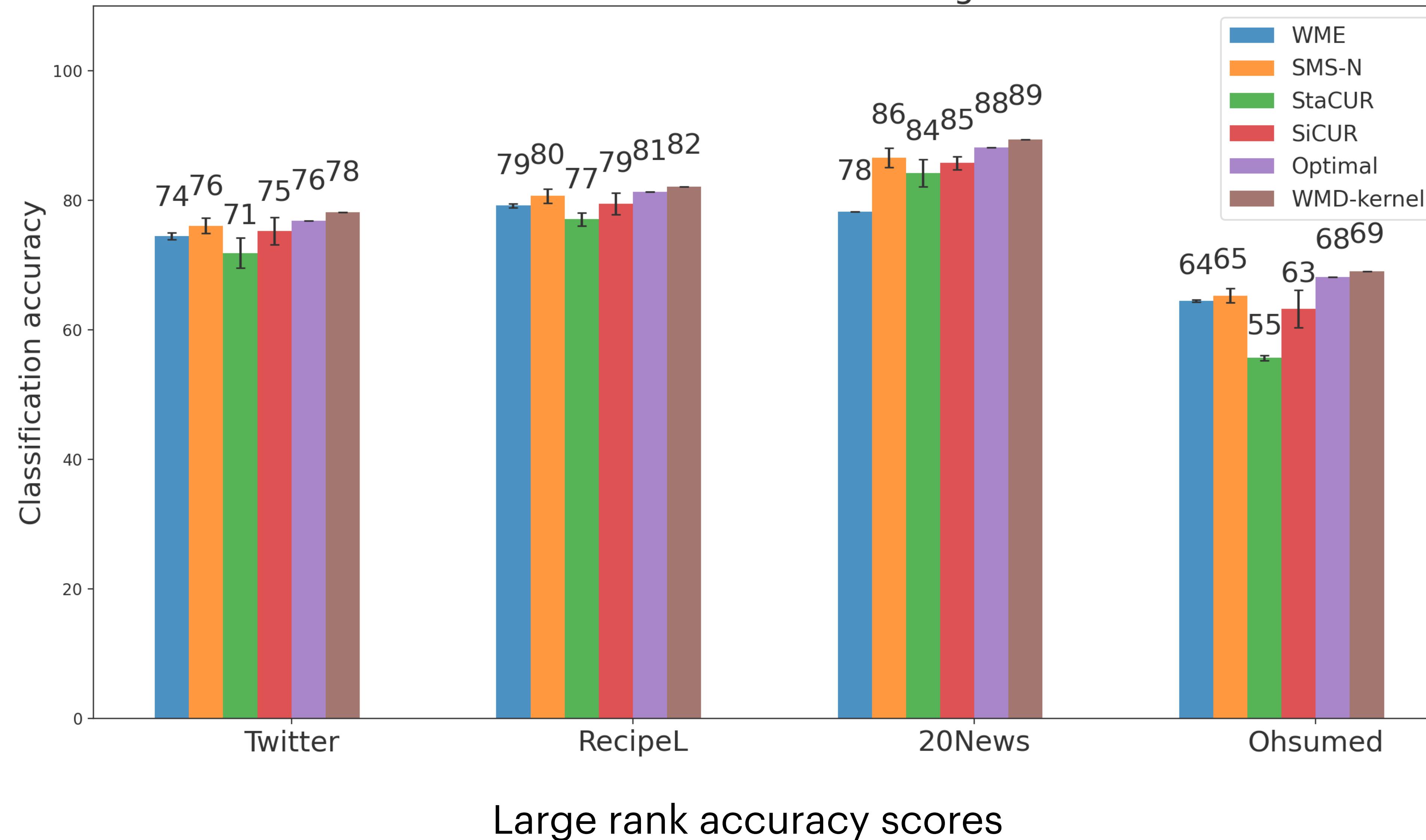
Document Classification Task

Document classification using WMD



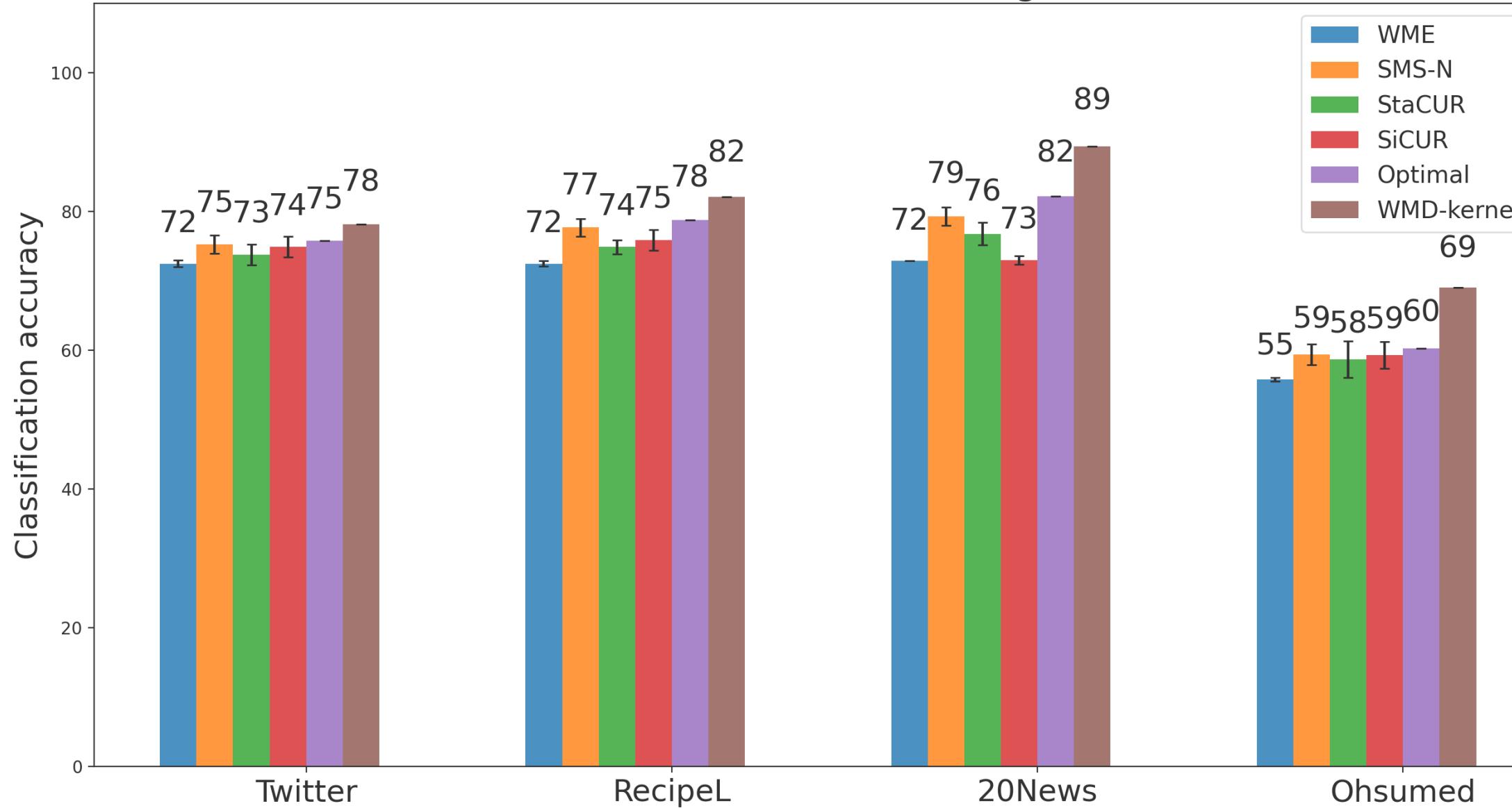
Document Classification Task

Document classification using WMD



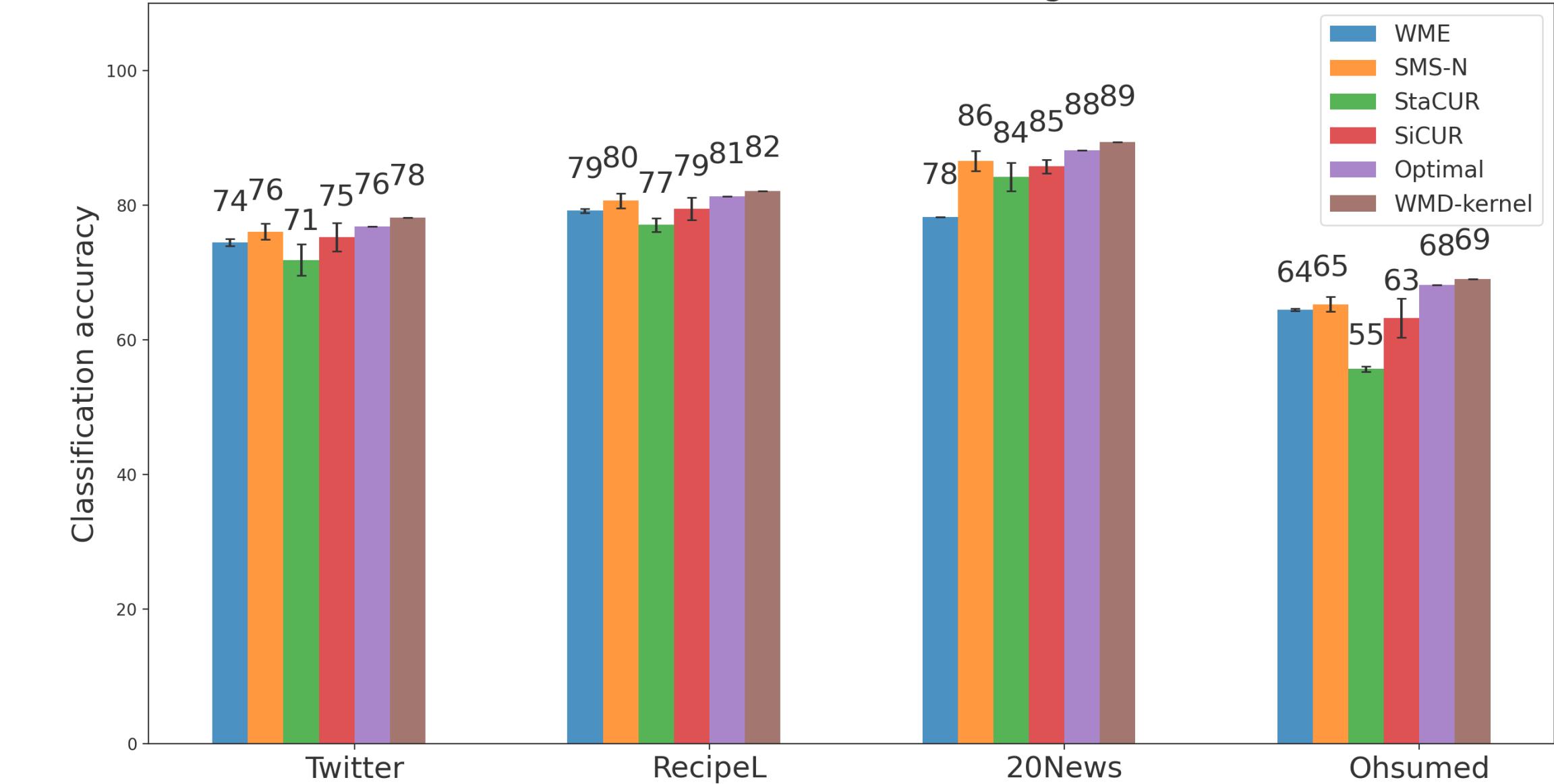
Document Classification Task

Document classification using WMD



Small rank accuracy scores

Document classification using WMD

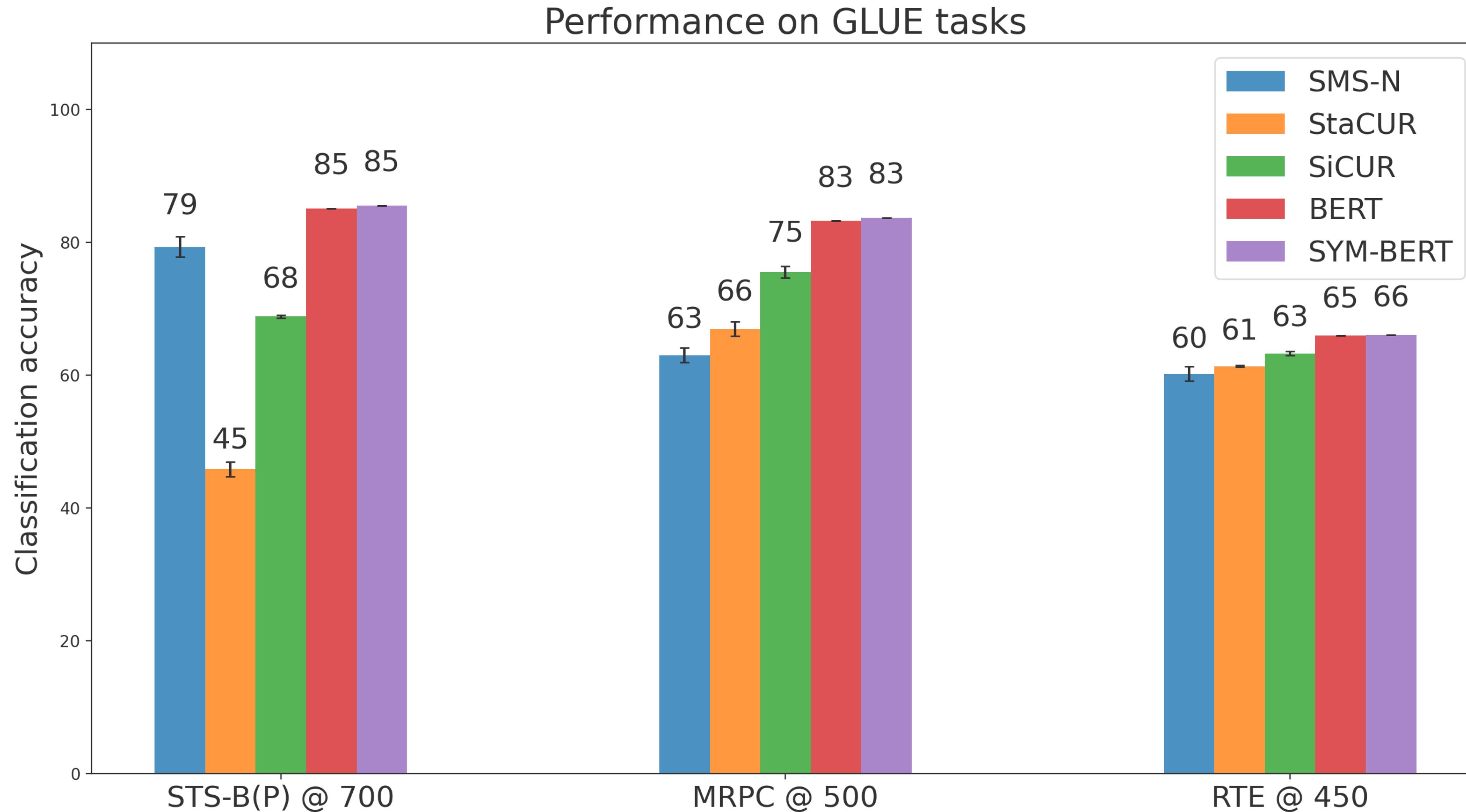


Large rank accuracy scores

Method	Twitter	Recipe_L	Ohsumed	20News
WME(SR)	13.02	5639.06	85.43	2712.12
SMS-N(SR)	86.23	13979.01	629.54	9422.05
WME(LR)	102.06	29238.48	2787.00	13021.13
SMS-N(LR)	1014.06	223902.32	21246.65	130342.28

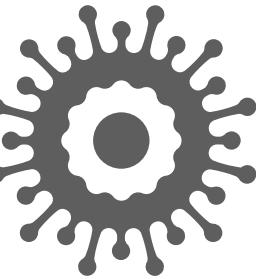
Time in seconds

Approximation of GLUE tasks



Cross-Document Entity & Coreference

A Rapid and Specific Assay Detection of MERS-CoV



Amazingly, it is effective against SARS and **MERS**.

Pai Huang^{1,2}, Hualei Wang^{2,3,4*}, Zengguo Cao^{2,3}, Hongli Jin^{2,3}, Hang Beibei Yu¹, Feihu Yan², Xingxing Hu^{1,2}, Fangfang Wu², Cuicui Jiao⁴, Shengnan Xu^{1,2}, Yongkun Zhao^{2,4}, Na Feng^{2,4}, Jianzhong Wang¹, Wei Tiecheng Wang^{2,4}, Yuwei Gao^{2,4}, Songtao Yang^{2,4} and Xianzhu Xia²

¹ Animal Science and Technology College, Jilin Agricultural University, Changchun, China, ² Key Laboratory of Jilin Province for Zoonotic Prevention and Control, Institute of Military Veterinary, Academy of Military Medical Sciences, Changchun, China, ³ College of Veterinary Medicine, Jilin University, Changchun, China, ⁴ Jiangsu Co-innovation Center for Prevention and Control of Important Animal Infectious Diseases and Zoonoses, Yangzhou, China, ⁵ State Key Laboratory of Respiratory Disease, Guangzhou Institute of Respiratory Health, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou, China, ⁶ Guangzhou Eighth People's Hospital of Guangzhou Medical University, Guangzhou, China, ⁷ Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

OPEN ACCESS

Middle East respiratory syndrome coronavirus (MERS-CoV) is a novel human coronavirus that can cause severe disease.

Vaccine 35 (2017) 2068–2075



Contents lists available at ScienceDirect



Vaccine

journal homepage: www.elsevier.com/locate/vaccine

Reviewed by:
Dirk Dittmar,
University of North Carolina
at Chapel Hill, United States
Reviewed by:
Timothy Sheahan,
University of North Carolina
at Chapel Hill, United States
Reviewed by:
Tee U.,
University of Pennsylvania,
United States
Correspondence:
Hualei Wang
w1631125@133.com;
wang25@mail.com
Specialty section:
This article was submitted to
Infectious Diseases, which was reviewed by
the following editor:
Richard Klasco, MD

DNA vaccine encoding Middle East respiratory syndrome coronavirus S1 protein induces protective immune responses in mice

Hang Chi^a, Xueying Zheng^{a,b}, Xiwen Wang^a, Chong Wang^a, Hualei Wang^{a,c}, Weiwei Gai^a, Stanley Perlman^d, Songtao Yang^{a,c,e,f}, Jincun Zhao^{c,g}, Xianzhu Xia^{a,d,h}

^a Key Laboratory of Jilin Province for Zoonotic Prevention and Control, Institute of Military Veterinary, Academy of Military Medical Sciences, Changchun, China

^b School of Public Health, Shandong Jiaotong University, Jinan, China

^c Jiangsu Co-innovation Center for Prevention and Control of Important Animal Infectious Diseases and Zoonoses, Nanjing, China

^d Department of Microbiology, University of Colorado, Boulder, CO, United States

^e State Key Laboratory of Respiratory Disease, Guangzhou Institute of Respiratory Health, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou, China

^f Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^g Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^h Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

ⁱ Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^j Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^k Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^l Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^m Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

ⁿ Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^o Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^p Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^q Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^r Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^s Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^t Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^u Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^v Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^w Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^x Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^y Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^z Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{aa} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{bb} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{cc} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{dd} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{ee} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{ff} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{gg} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{hh} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

ⁱⁱ Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{jj} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{kk} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{ll} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{mm} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

ⁿⁿ Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{oo} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{pp} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{qq} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{rr} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{ss} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{tt} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{uu} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{vv} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{ww} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{xx} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{yy} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{zz} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{aa} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{bb} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{cc} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{dd} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{ee} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{ff} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{gg} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{hh} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

ⁱⁱ Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{jj} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{kk} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{ll} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{mm} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

ⁿⁿ Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{oo} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{pp} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{qq} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{rr} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{ss} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{tt} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{uu} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{vv} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{ww} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{xx} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{yy} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{zz} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{aa} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{bb} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{cc} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{dd} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{ee} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{ff} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{gg} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{hh} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

ⁱⁱ Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{jj} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{kk} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{ll} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{mm} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

ⁿⁿ Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{oo} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

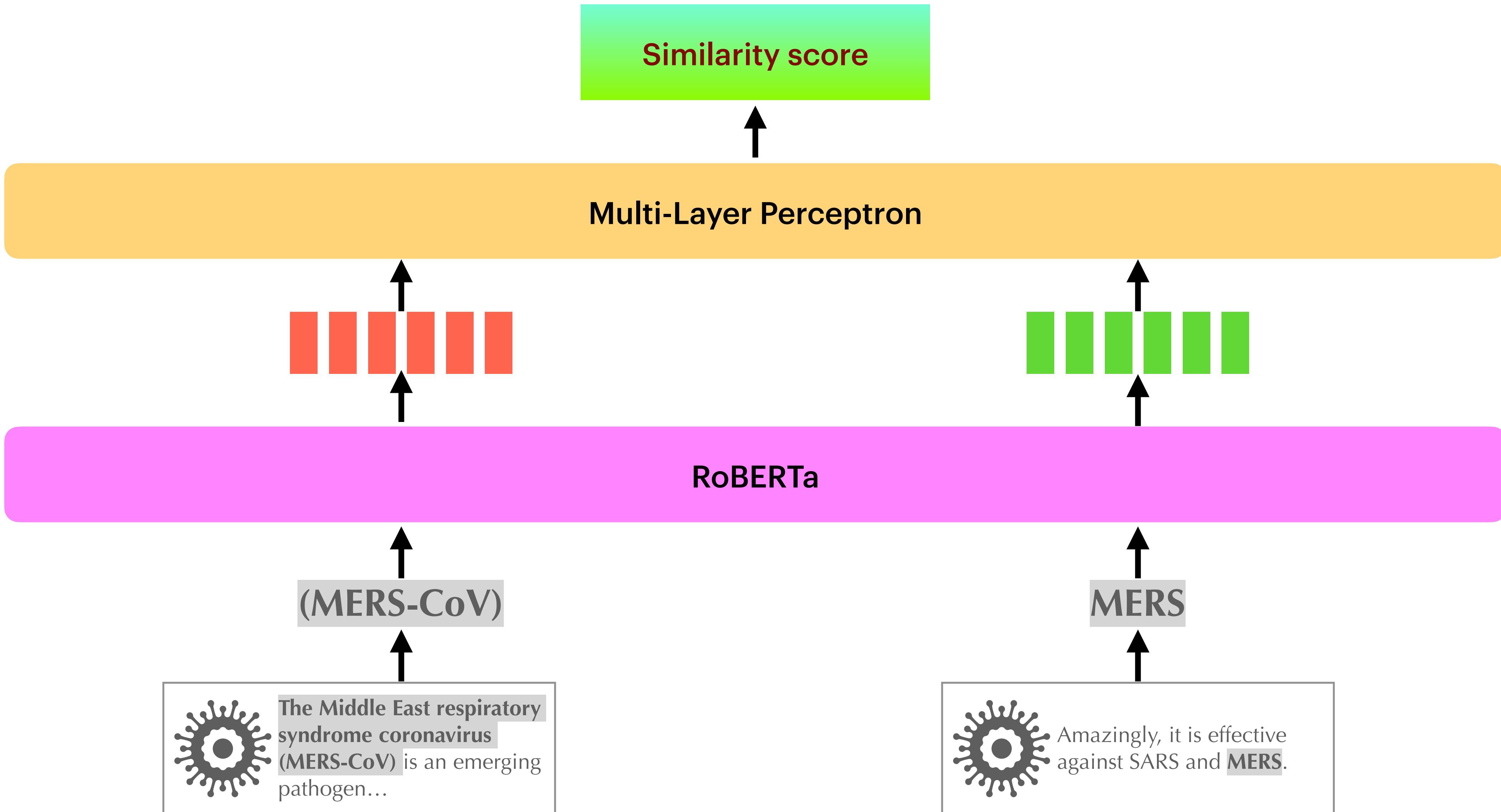
^{pp} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{qq} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

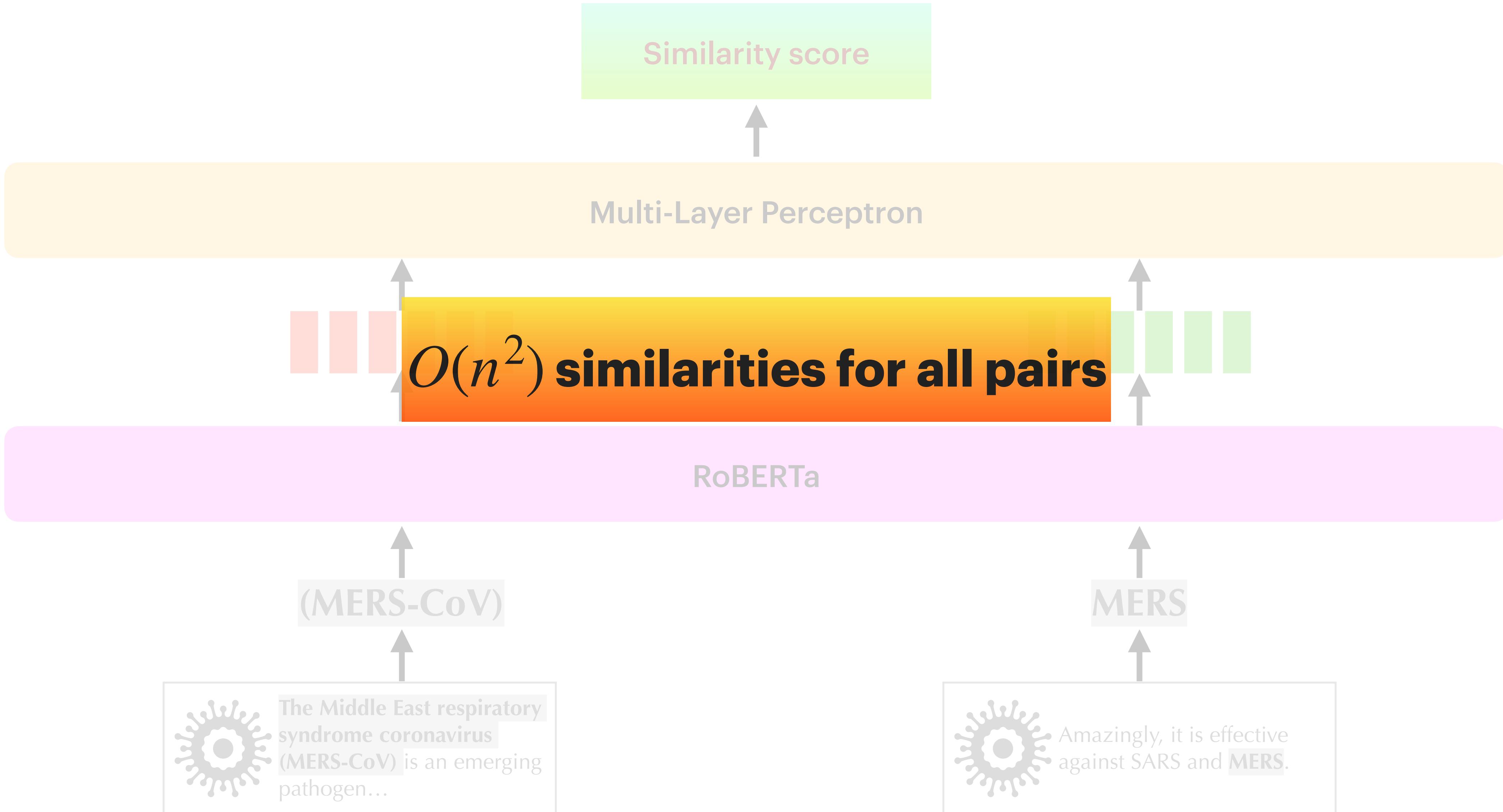
^{rr} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

^{ss} Department of Clinical Laboratory, College of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou, China

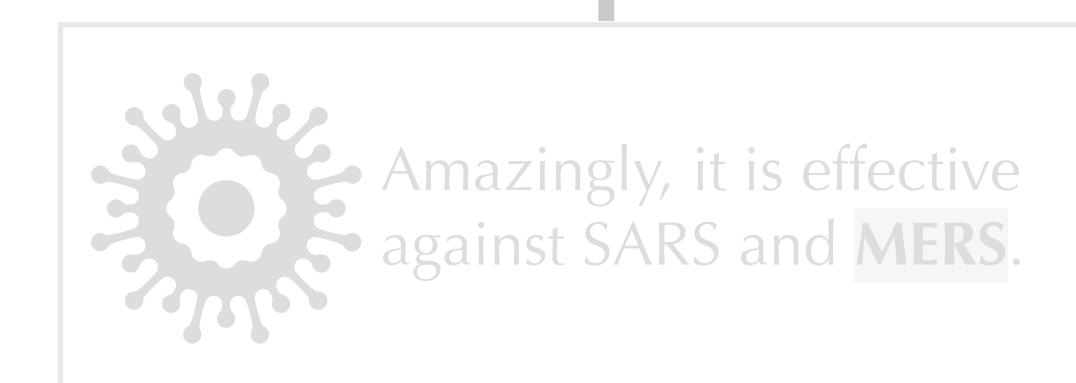
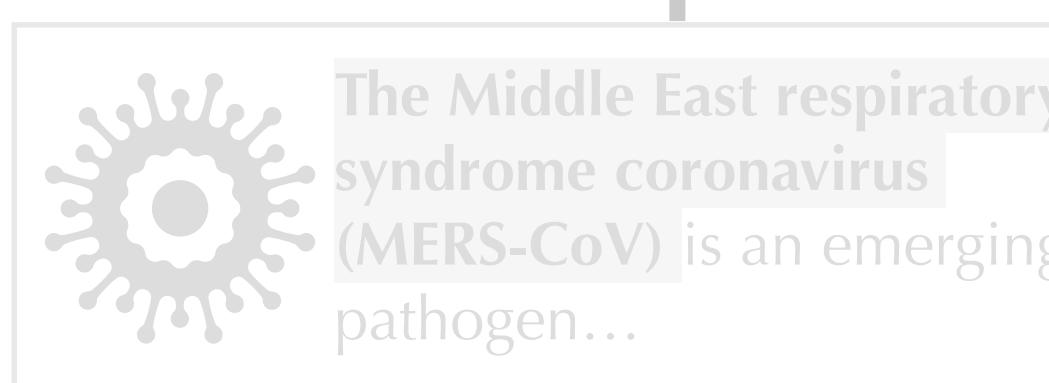
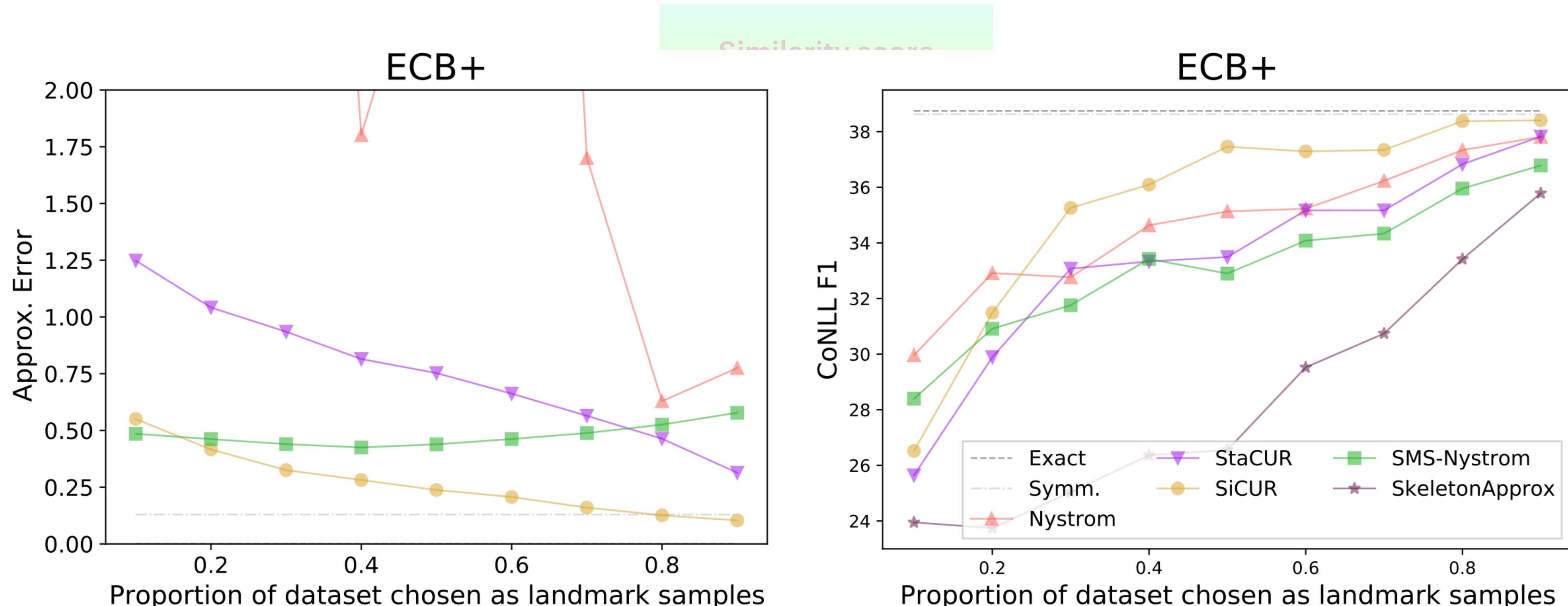
Cross-Document Entity & Coreference



Cross-Document Entity & Coreference



Cross-Document Entity & Coreference



Conclusions

We show that indefinite matrices arising in NLP can be approximated using sublinear algorithms

Simple variant of Nyström and variants of CUR display strong performance in variety of tasks

Thank you! Questions?