A Brief Overview of Sublinear Matrix Approximation Techniques

UMassAmherst | Manning College of Information & Computer Sciences

Matrices in the Wild

Distance matrices

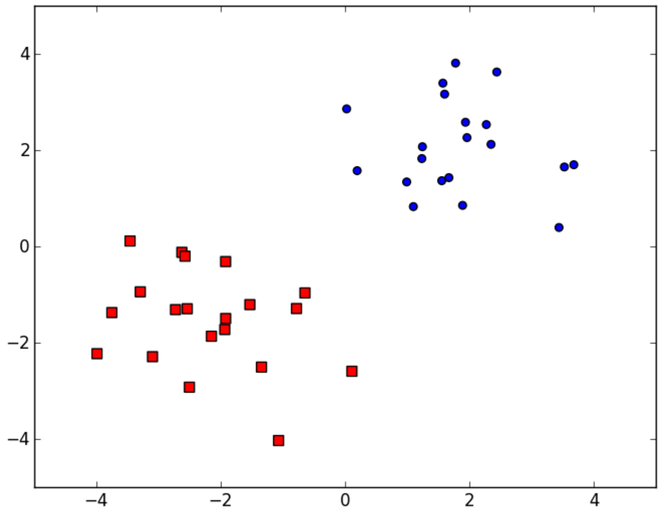


Image source: http://blog.sairahul.com/2014/01/linear-separability.html

Matrices in the Wild

Similarity matrices

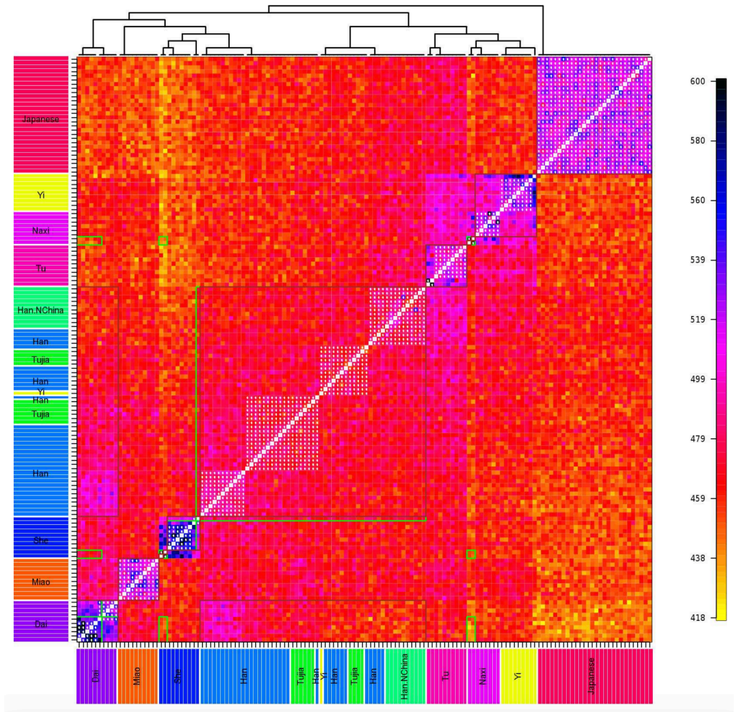


Image source: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.727.7925&rep=rep1&type=pdf

Matrices in the Wild

Kernel matrices

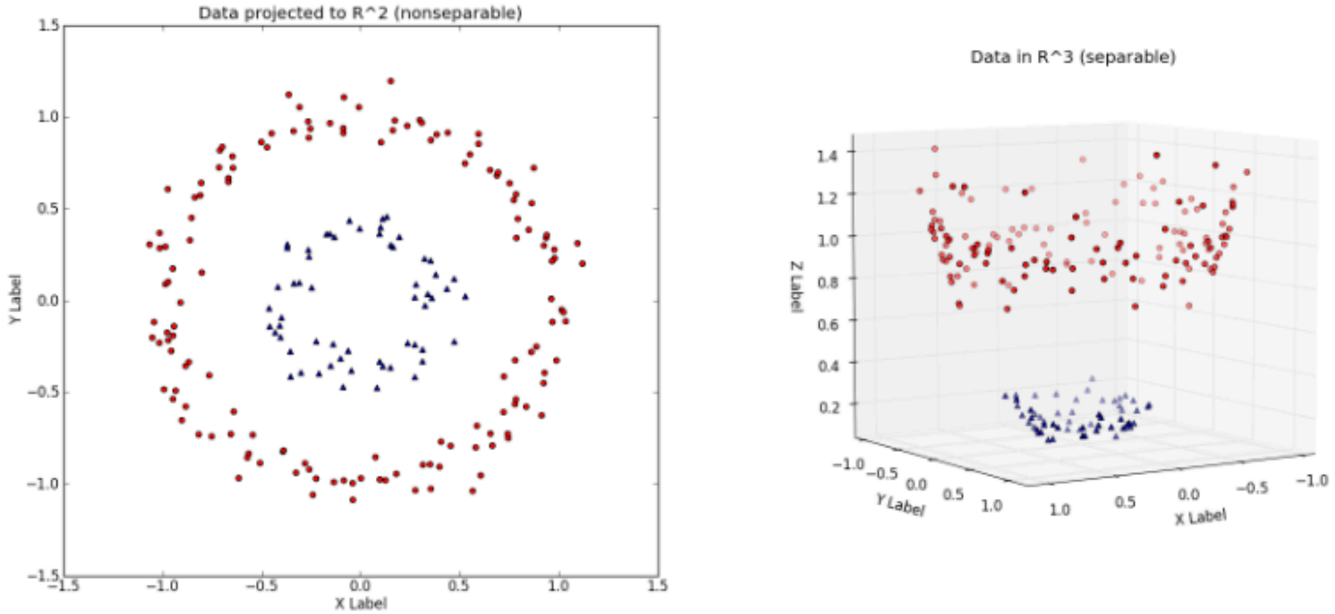


Image source: http://www.eric-kim.net/eric-kim-net/posts/1/kernel_trick.html

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

Random features approximations of kernel matrices [1]

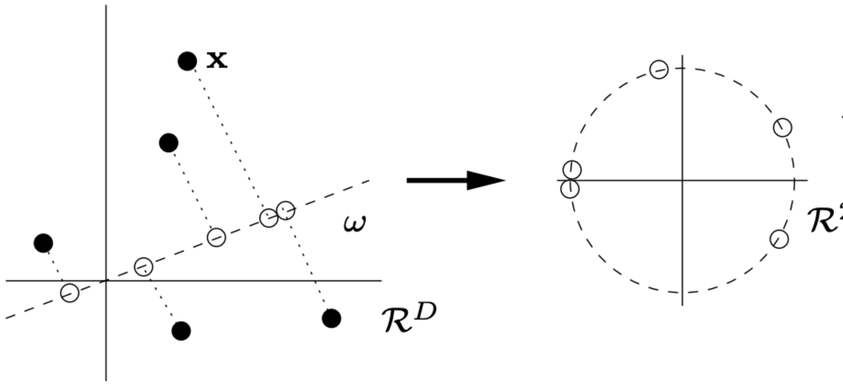
Works on shift invariant kernel's: $K(x, y) = K(x - y)$

[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 $\exists K(x, y) = < \phi(x), \phi(y) >$ is easy to compute

Random features approximations of kernel matrices [1]



Kernel Name	$k(\Delta)$	$p(\omega)$
Gaussian	$e^{-\frac{\|\Delta\|_2^2}{2}}$	$(2\pi)^{-\frac{D}{2}} e^{-\frac{\|\omega\|_2^2}{2}}$
Laplacian	$e^{-\|\Delta\|_1}$	$\prod_d \frac{1}{\pi(1+\omega_d^2)}$
Cauchy	$\prod_d \frac{2}{1+\Delta_d^2}$	$e^{-\|\Delta\|_1}$

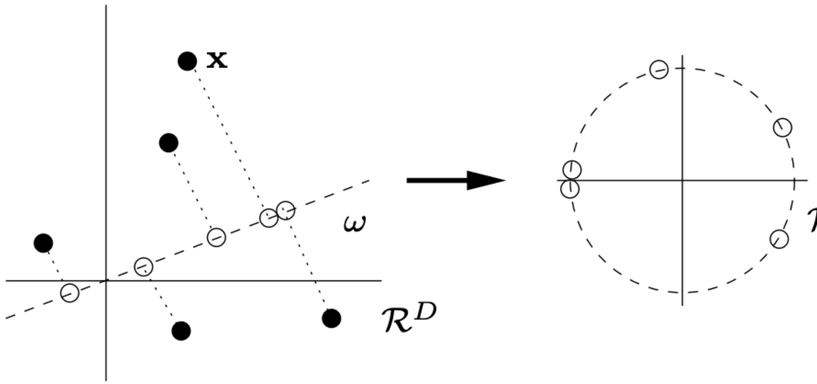
Transparent dots are $z(x)$ and $K(x, y) \approx < z(x), z(y) >$

[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 $\exists K(x, y) = < \phi(x), \phi(y) >$ is easy to compute

Random features approximations of kernel matrices [1]



Kernel Name	$k(\Delta)$	$p(\omega)$
Gaussian	$e^{-\frac{\|\Delta\|_2^2}{2}}$	$(2\pi)^{-\frac{D}{2}} e^{-\frac{\|\omega\|_2^2}{2}}$
Laplacian	$e^{-\|\Delta\|_1}$	$\prod_d \frac{1}{\pi(1+\omega_d^2)}$
Cauchy	$\prod_d \frac{2}{1+\Delta_d^2}$	$e^{-\|\Delta\|_1}$

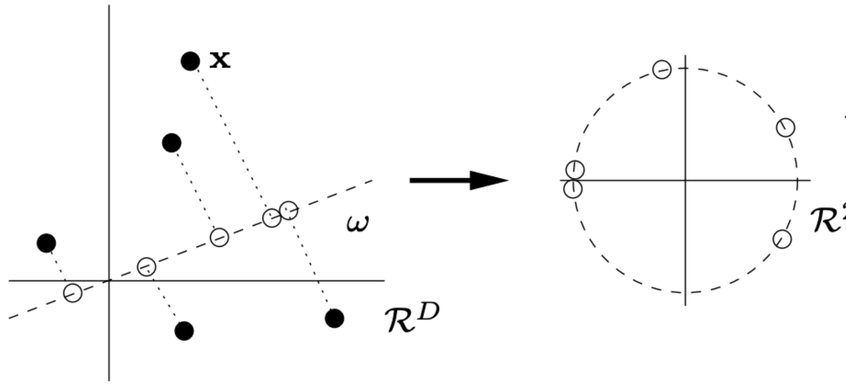
We can store $Z \in \mathbb{R}^{N \times R}$ where $R \ll N$

[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 $\exists K(x, y) = < \phi(x), \phi(y) >$ is easy to compute

Random features approximations of kernel matrices [1]



Kernel Name	$k(\Delta)$	$p(\omega)$
Gaussian	$e^{-\frac{\|\Delta\|_2^2}{2}}$	$(2\pi)^{-\frac{D}{2}} e^{-\frac{\|\omega\|_2^2}{2}}$
Laplacian	$e^{-\|\Delta\|_1}$	$\prod_d \frac{1}{\pi(1+\omega_d^2)}$
Cauchy	$\prod_d \frac{2}{1+\Delta_d^2}$	$e^{-\|\Delta\|_1}$

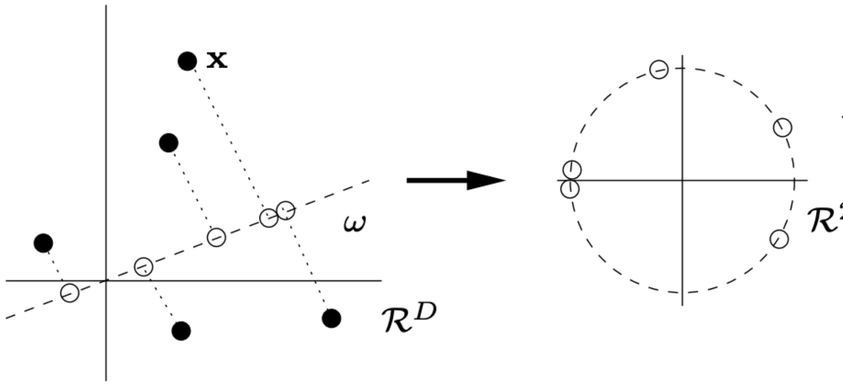
Prediction in say linear regression then becomes: $\beta = (Z^T Z)^{-1} Z^T Y$

[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 $\exists K(x, y) = < \phi(x), \phi(y) >$ is easy to compute

Random features approximations of kernel matrices [1]



Kernel Name	$k(\Delta)$	$p(\omega)$
Gaussian	$e^{-\frac{\|\Delta\|_2^2}{2}}$	$(2\pi)^{-\frac{D}{2}} e^{-\frac{\|\omega\|_2^2}{2}}$
Laplacian	$e^{-\|\Delta\|_1}$	$\prod_d \frac{1}{\pi(1+\omega_d^2)}$
Cauchy	$\prod_d \frac{2}{1+\Delta_d^2}$	$e^{-\|\Delta\|_1}$

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

$S \in \mathbb{R}^{n \times s}$ 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 $KS(S^T KS)^+ S^T K$

$S \in \mathbb{R}^{n \times s}$ 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 $KS(S^T KS)^+ S^T K$

$S \in \mathbb{R}^{n \times s}$ 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 nystrom method. *Advances in Neural Information Processing Systems*, *30*.

Sublinear Methods

Pseudoskeleton approximation [1]

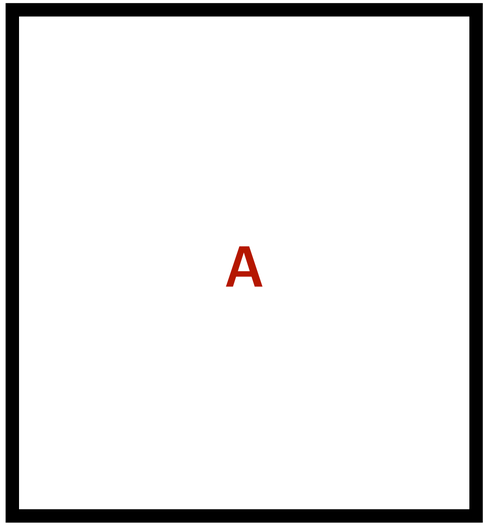
For any matrix $K \in \mathbb{R}^{n \times n}$ compute approximation as $KS_2(S_2^T K S_1)^+ S_1^T K$

$S_1, S_2 \in \mathbb{R}^{n \times s}$ 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

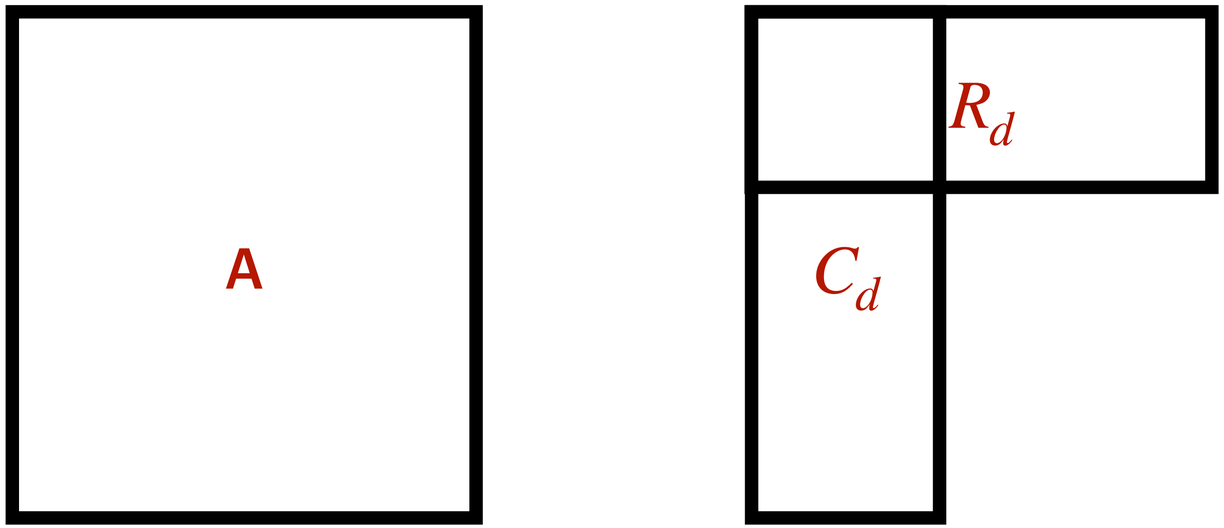
CUR [1]



[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

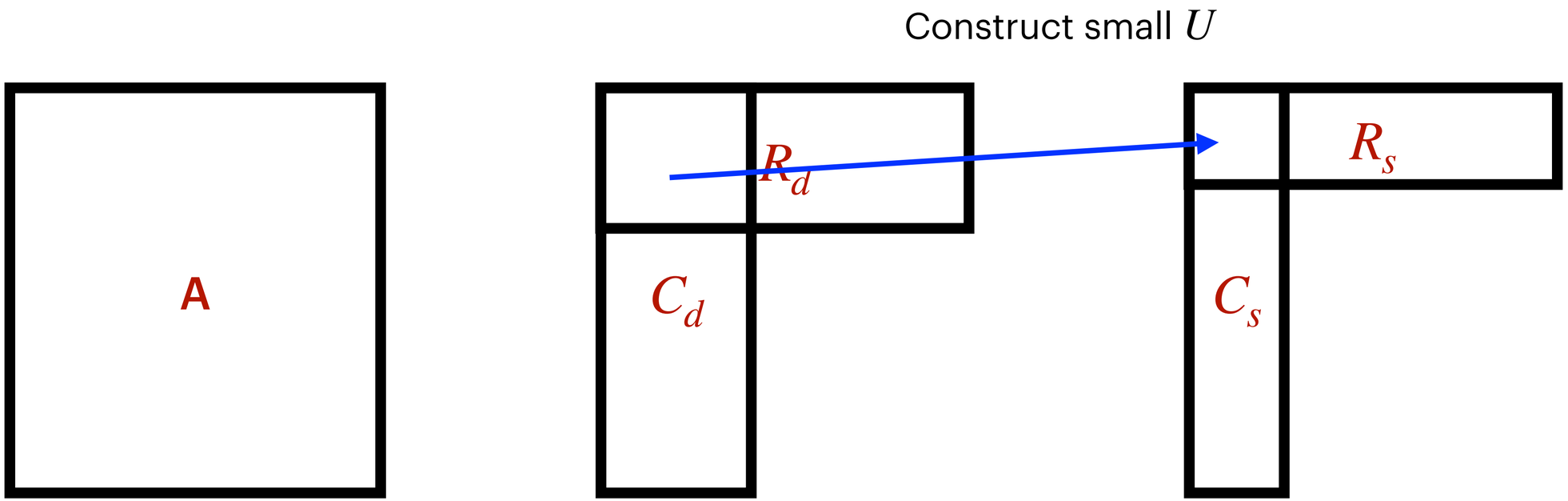
CUR [1]



[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]

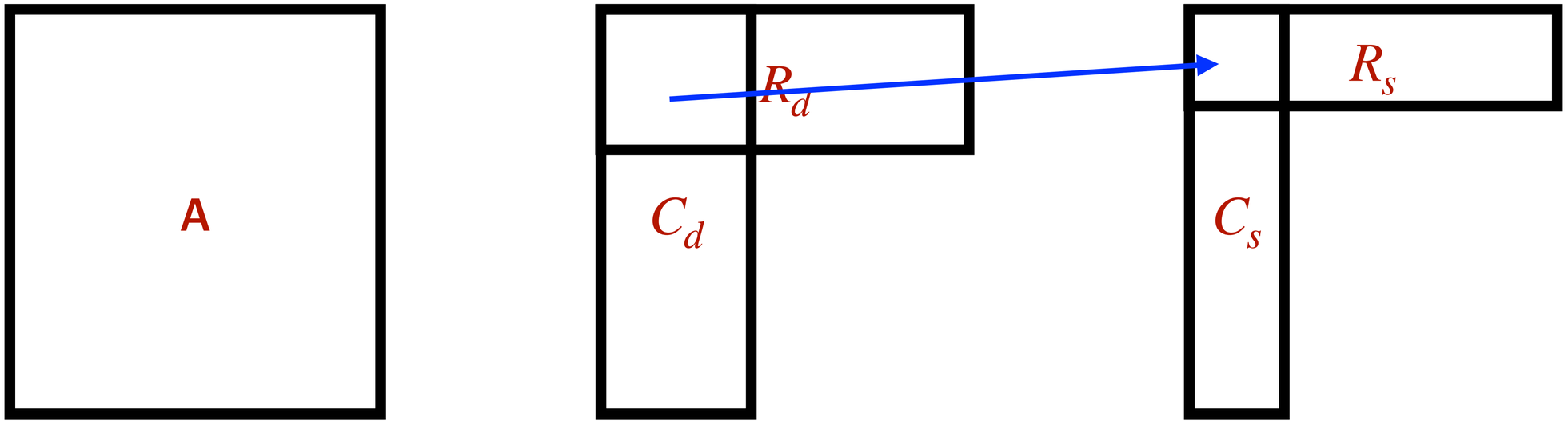


[1] Mahoney, M.W. and Drineas, P., 2009. CUR matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, *106*(3), pp.697-702.

Sublinear Methods

CUR [1]

Construct small U



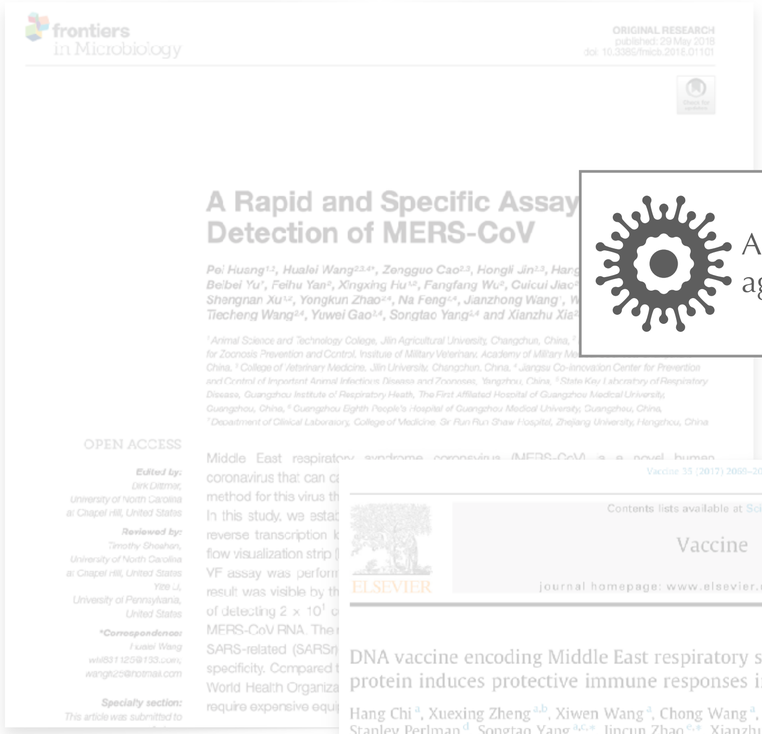
For multiple copies keep one column/row but multiply it by $\sqrt{d}$

[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

Archan Ray, Nicholas Monath, Andrew McCallum, Cameron Musco

UMassAmherst | Manning College of Information & Computer Sciences

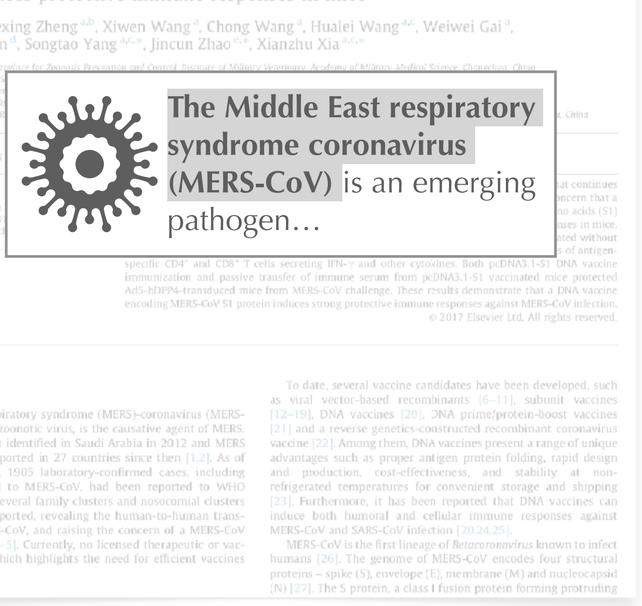


frontiers in Microbiology

ORIGINAL RESEARCH

A Rapid and Specific Assay
Detection of MERS-CoV

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



Vaccine 35 (2017) 2069–2075

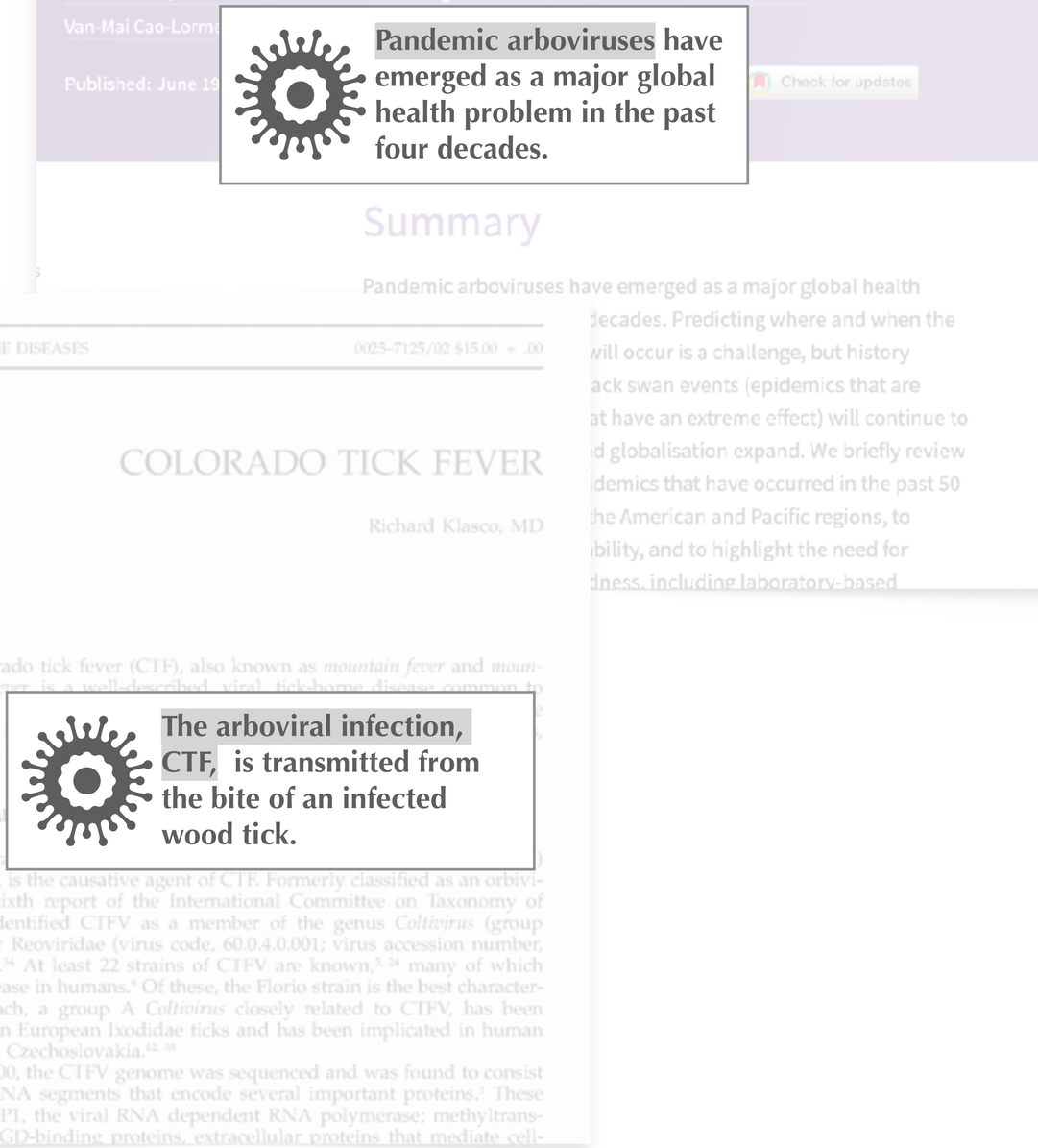
Contents lists available at ScienceDirect

Vaccine

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

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

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



PERSONAL VIEW | VOLUME 18, ISSUE 11, E355-E361, NOVEMBER 01, 2018

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-Lorm

Published: June 19

Check for updates

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

Summary

Pandemic arboviruses have emerged as a major global health decades. Predicting where and when the will occur is a challenge, but history ack swan events (epidemics that are at have an extreme effect) will continue to d globalisation expand. We briefly review idemics that have occurred in the past 50 he American and Pacific regions, to bility, and to highlight the need for dness, including laboratory-based

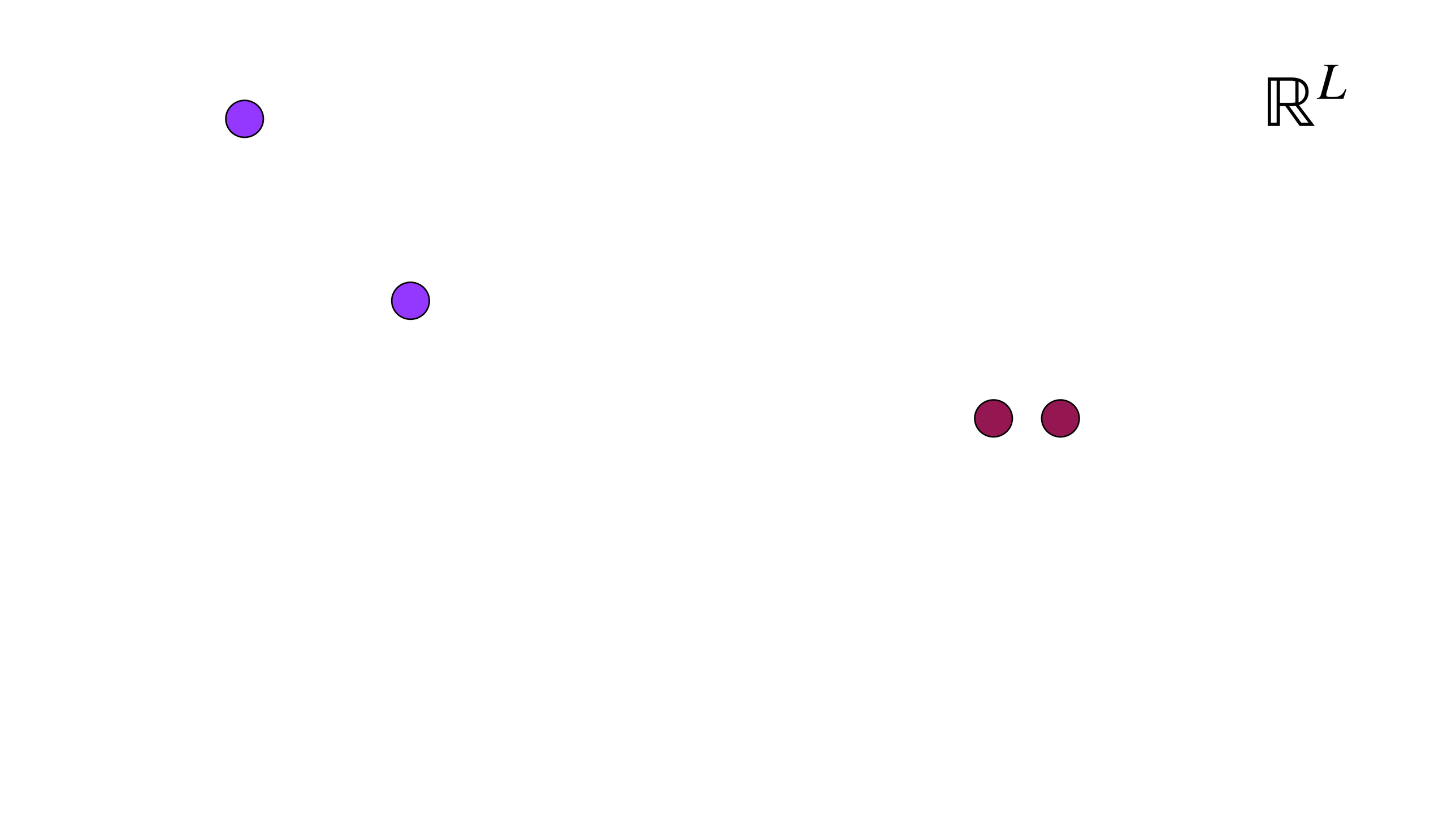
TICK-BORNE DISEASES

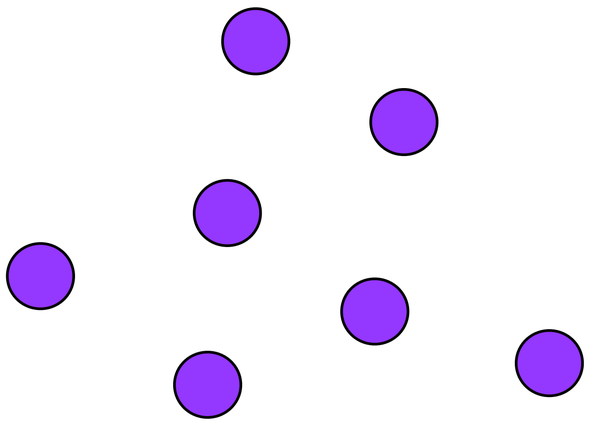
0025-7125/02 $15.00 + .00

COLORADO TICK FEVER

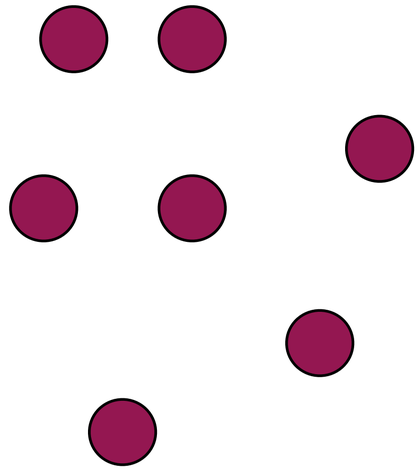
Richard Klasco, MD

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

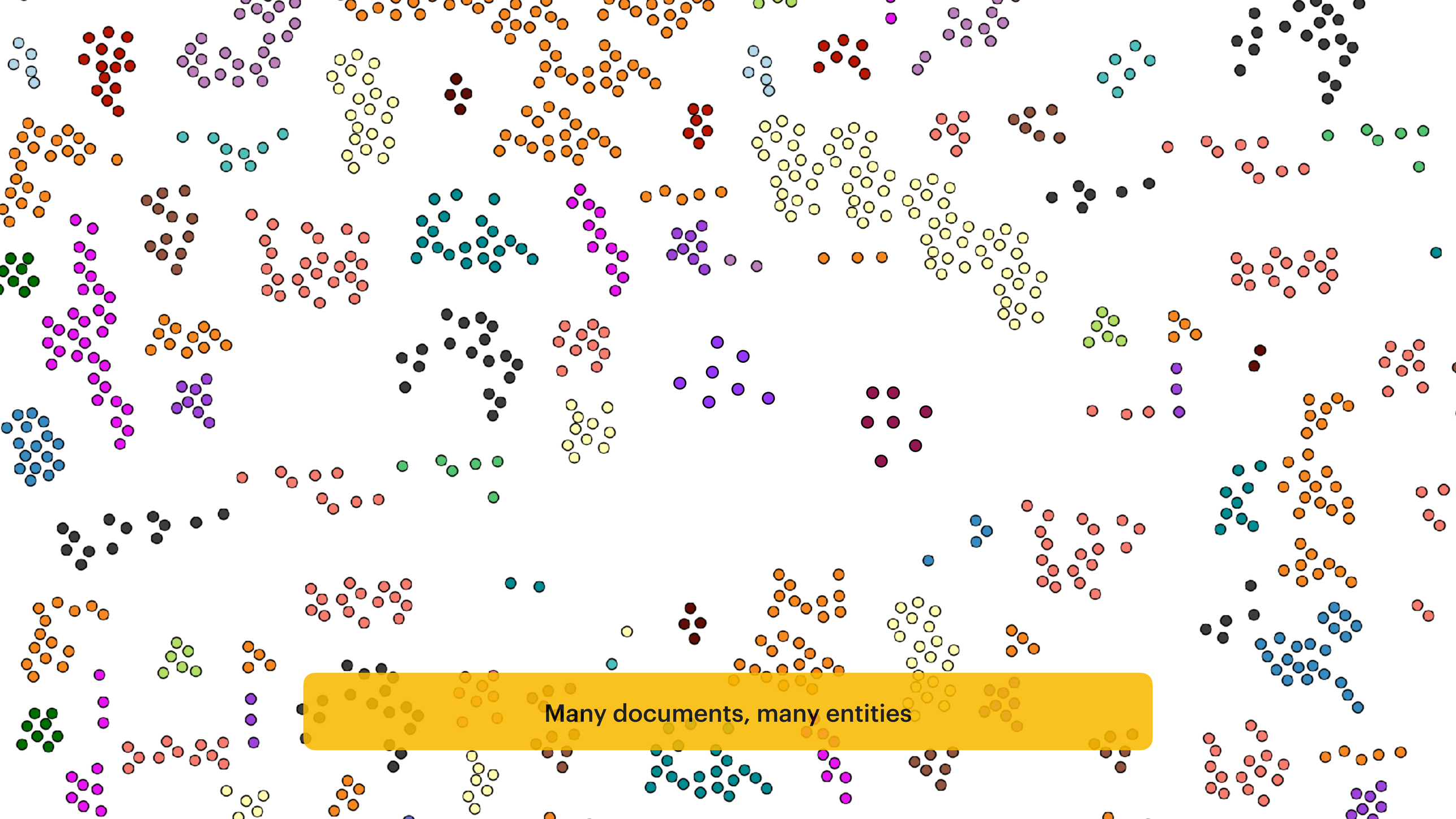

$\mathbb{R}^L$

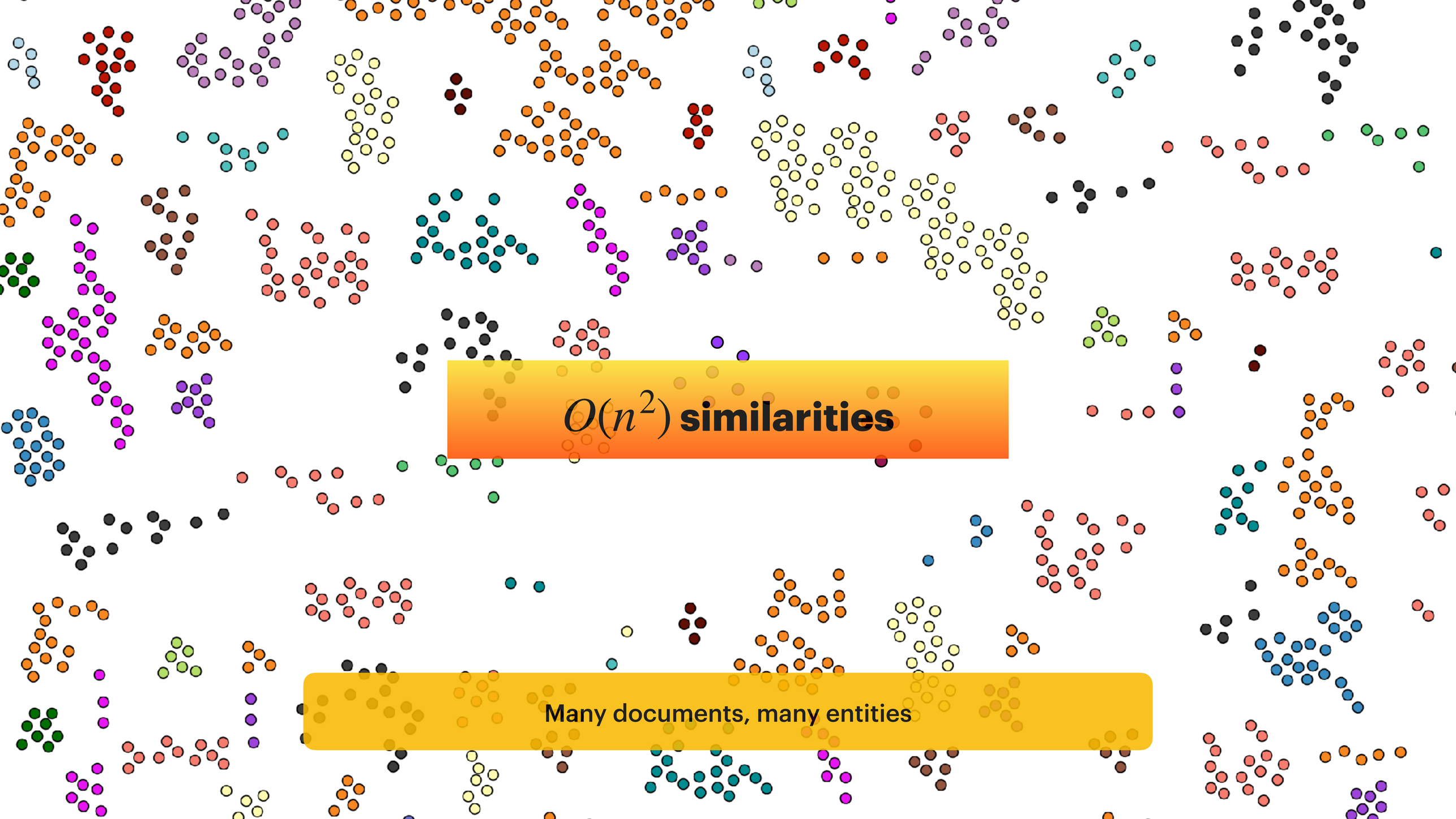


$\mathbb{R}^L$



Numerous documents mentioning the same entities


Many documents, many entities


$O(n^2)$ similarities
Many documents, many entities

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

$O(n^2)$ **similarities**

$$\begin{bmatrix} 0.0091 & 0.3020 & \cdots & \cdots & \cdots & 0.1653 & 0.2900 \\ 0.3020 & 0.6571 & \cdots & \cdots & \cdots & 0.4337 & 0.1983 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.1653 & 0.4337 & \cdots & \cdots & \cdots & 0.3769 & 0.2375 \\ 0.2900 & 0.1983 & \cdots & \cdots & \cdots & 0.2375 & 0.8342 \end{bmatrix}$$

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

$$\begin{bmatrix} 0.0091 & 0.3020 & \cdots & \cdots & \cdots & 0.1653 & 0.2900 \\ 0.3020 & 0.6571 & \cdots & \cdots & \cdots & 0.4337 & 0.1983 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.1653 & 0.4337 & \cdots & \cdots & \cdots & 0.3769 & 0.2375 \\ 0.2900 & 0.1983 & \cdots & \cdots & \cdots & 0.2375 & 0.8342 \end{bmatrix}$$

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

We want these methods to be ***sublinear*** $[o(n)]$

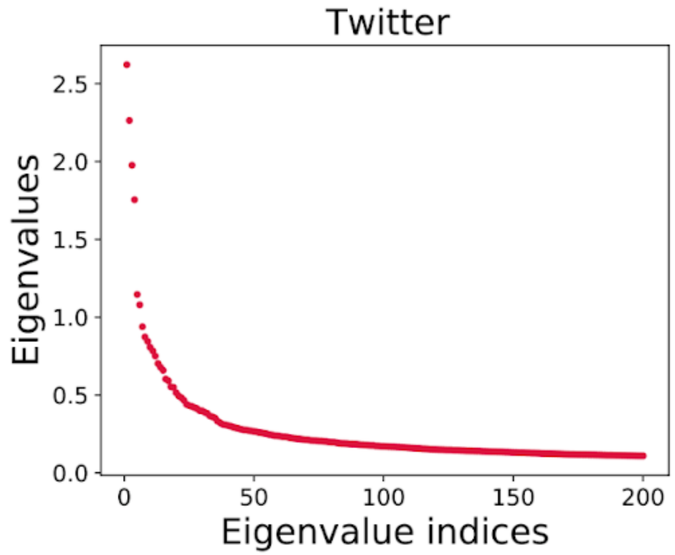
$$\begin{bmatrix} 0.0091 & 0.3020 & \cdots & \cdots & \cdots & 0.1653 & 0.2900 \\ 0.3020 & 0.6571 & \cdots & \cdots & \cdots & 0.4337 & 0.1983 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.1653 & 0.4337 & \cdots & \cdots & \cdots & 0.3769 & 0.2375 \\ 0.2900 & 0.1983 & \cdots & \cdots & \cdots & 0.2375 & 0.8342 \end{bmatrix}$$

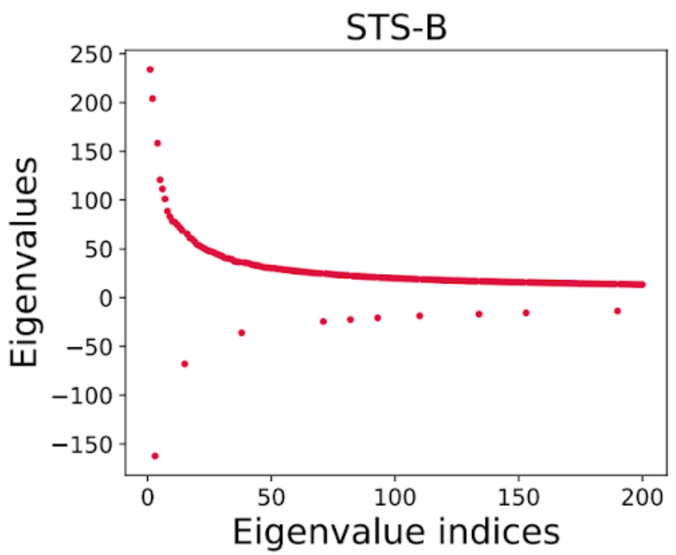
We'll do okay with ***approximate*** similarities

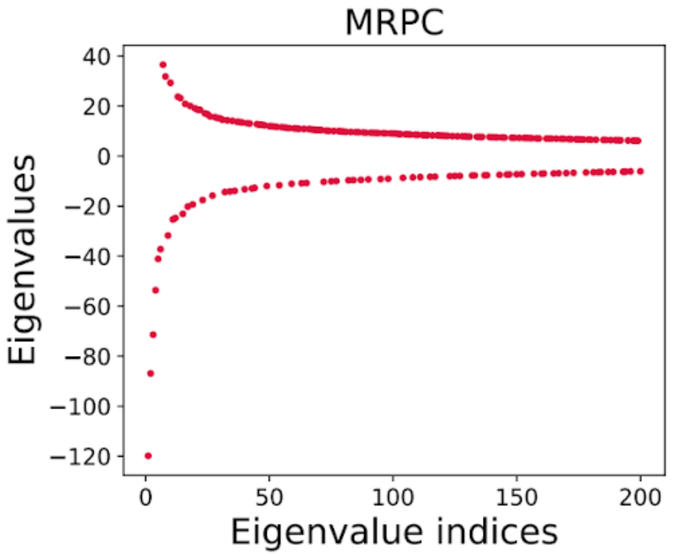
We want these methods to be ***sublinear*** $[o(n)]$

In NLP most similarity matrices are ***indefinite***

In NLP most similarity matrices are ***indefinite***







$$\begin{bmatrix} 0.0091 & 0.3020 & \cdots & \cdots & \cdots & 0.1653 & 0.2900 \\ 0.3020 & 0.6571 & \cdots & \cdots & \cdots & 0.4337 & 0.1983 \\ \vdots & \vdots & & & & \vdots & \vdots \\ \vdots & \vdots & & & & \vdots & \vdots \\ \vdots & \vdots & & & & \vdots & \vdots \\ \vdots & \vdots & & & & \vdots & \vdots \\ \vdots & \vdots & & & & \vdots & \vdots \\ \vdots & \vdots & & & & \vdots & \vdots \\ \vdots & \vdots & & & & \vdots & \vdots \\ 0.1653 & 0.4337 & \cdots & \cdots & \cdots & 0.3769 & 0.2375 \\ 0.2900 & 0.1983 & \cdots & \cdots & \cdots & 0.2375 & 0.8342 \end{bmatrix}$$

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

We want these methods to be ***sublinear*** $[o(n)]$

In NLP most similarity matrices are ***indefinite***

Many methods exist for approximating PSD matrices

$$\begin{bmatrix} 0.0091 & 0.3020 & \cdots & \cdots & \cdots & 0.1653 & 0.2900 \\ 0.3020 & 0.6571 & \cdots & \cdots & \cdots & 0.4337 & 0.1983 \\ \vdots & \vdots & & & & \vdots & \vdots \\ \vdots & \vdots & & & & \vdots & \vdots \\ \vdots & \vdots & & & & \vdots & \vdots \\ \vdots & \vdots & & & & \vdots & \vdots \\ \vdots & \vdots & & & & \vdots & \vdots \\ \vdots & \vdots & & & & \vdots & \vdots \\ \vdots & \vdots & & & & \vdots & \vdots \\ 0.1653 & 0.4337 & \cdots & \cdots & \cdots & 0.3769 & 0.2375 \\ 0.2900 & 0.1983 & \cdots & \cdots & \cdots & 0.2375 & 0.8342 \end{bmatrix}$$

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

We want these methods to be ***sublinear*** $[o(n)]$

In NLP most similarity matrices are ***indefinite***

Not many methods for indefinite matrices exist

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

Columns of K

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

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

Columns of ***K***

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

Inverse of the principal sub matrix of ***K***

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

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

Columns of *K*

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

Inverse of the principal sub matrix of *K*

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

Rows of *K*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Columns of *K*

Rows of *K*

Inverse of the principal sub matrix of *K*

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

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

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

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

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

Columns of *K*

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

Inverse of the principal sub matrix of *K*

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

Rows of *K*

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

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

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

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

Columns of *K*

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

Inverse of the principal sub matrix of *K*

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

Rows of *K*

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

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

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

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

Columns of *K*

Rows of *K*

Inverse of the principal sub matrix of *K*

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

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

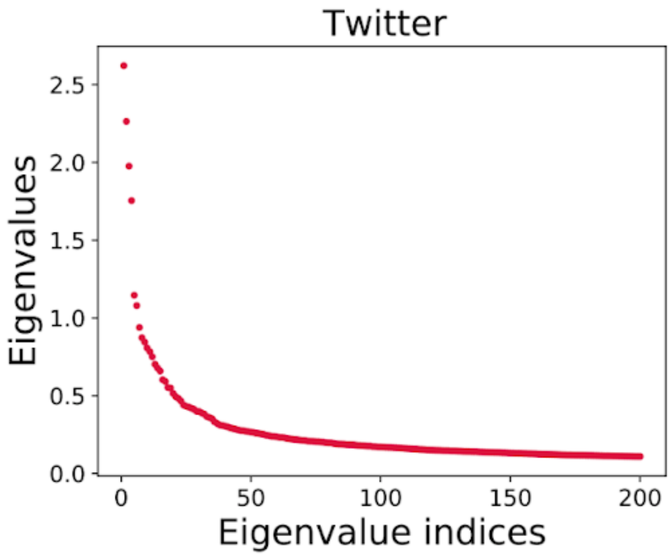
$$\begin{bmatrix} 0.0089 & 0.3262 \\ 0.3262 & 0.6635 \\ \vdots & \vdots \\ 0.1820 & 0.4246 \\ 0.2993 & 0.1990 \end{bmatrix}^T$$

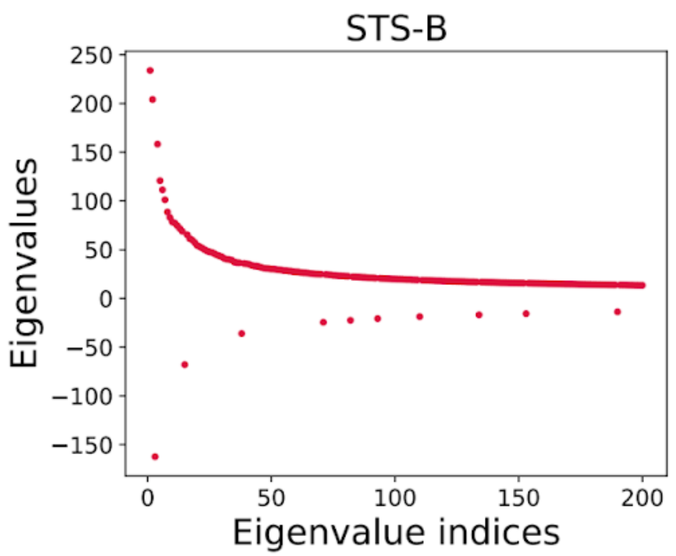
If $S^T KS$ is ill conditioned, $(S^T KS)^+$ will fail

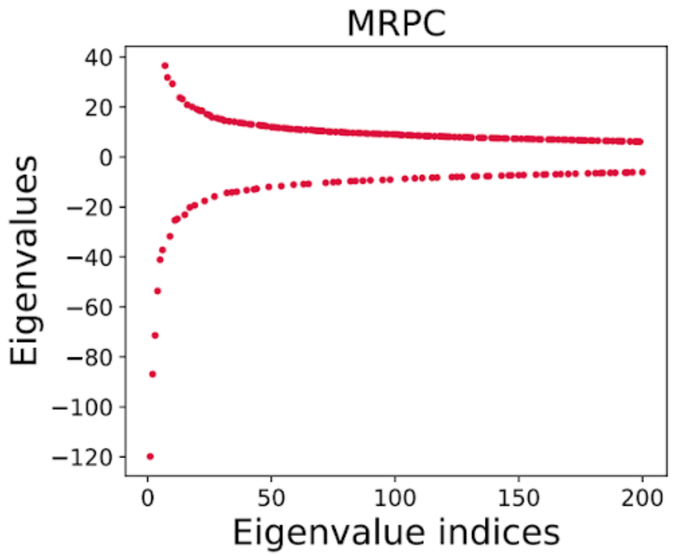
Endemic to indefinite matrices

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

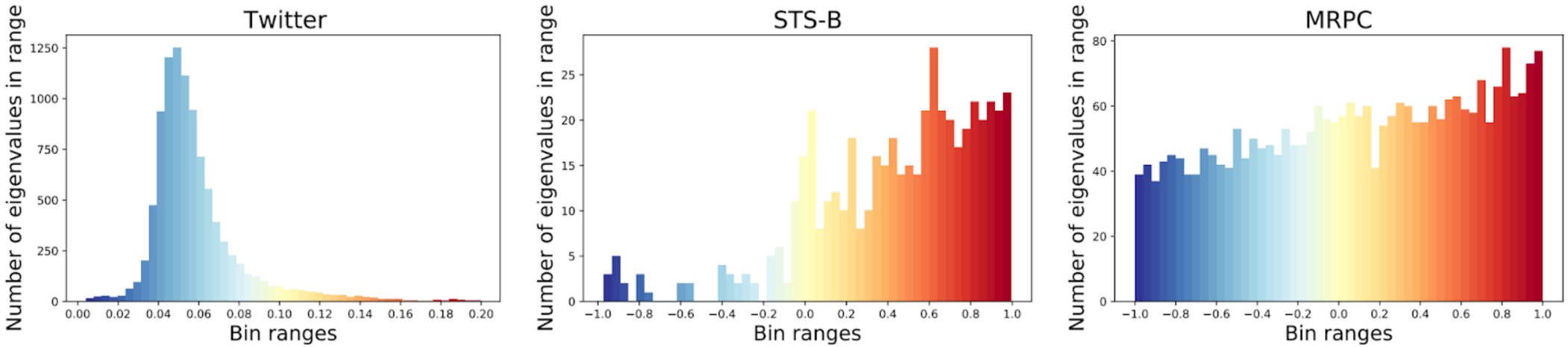
Endemic to indefinite matrices







Endemic to indefinite matrices



Independently sample $S^T KS$, 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

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

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

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

Other Approaches

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:

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

Sample S_1 and S_2 independently at random

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:

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

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

Sample S_1 and S_2 independently at random

Sublinear methods for approximating indefinite similarity matrices

SMS-Nyström

$$K = KS(S^TKS)^+S^TK$$

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

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

Other Approaches

Skeleton approximation:

$$\tilde{K} = KS_2(S_2^TKS_1)^+S_1^TK$$

SiCUR:

$$\tilde{K} = KS_2(S_2^TKS_1)^+S_1^TK$$

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

StaCUR:

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

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:

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

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

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

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

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

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

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

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

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

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

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

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

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

Sublinear methods for approximating indefinite similarity matrices

SMS-Nyström

$$K = KS(S^TKS)^+S^TK$$

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

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

Other Approaches

Skeleton approximation:

$$\tilde{K} = KS_2(S_2^TKS_1)^+S_1^TK$$

SiCUR:

$$\tilde{K} = KS_2(S_2^TKS_1)^+S_1^TK$$

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

StaCUR:

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

Very similar to Nyström Approximation, rows and columns sampled independently

Sublinear methods for approximating indefinite similarity matrices

SMS-Nyström

$$K = KS(S^TKS)^+S^TK$$

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

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

Other Approaches

Skeleton approximation:

$$\tilde{K} = KS_2(S_2^TKS_1)^+S_1^TK$$

SiCUR:

$$\tilde{K} = KS_2(S_2^TKS_1)^+S_1^TK$$

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

StaCUR:

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

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

Sublinear methods for approximating indefinite similarity matrices

SMS-Nyström

$$K = KS(S^TKS)^+S^TK$$

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

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

Other Approaches

Skeleton approximation:

$$\tilde{K} = KS_2(S_2^TKS_1)^+S_1^TK$$

SiCUR:

$$\tilde{K} = KS_2(S_2^TKS_1)^+S_1^TK$$

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

StaCUR:

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

$(S^TKS)^+$ is replaced with $(KSS^TK)^+S^TKS$

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

$$K = KS(S^TKS)^+S^TK$$

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

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

Other Approaches

Skeleton approximation:

$$\tilde{K} = KS_2(S_2^TKS_1)^+S_1^TK$$

SiCUR:

$$\tilde{K} = KS_2(S_2^TKS_1)^+S_1^TK$$

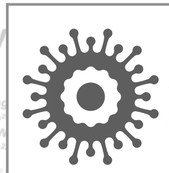
$$|S_2| = 2|S_1|$$

StaCUR:

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

Benefits: No two stage sampling!

Empirical evaluation



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

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

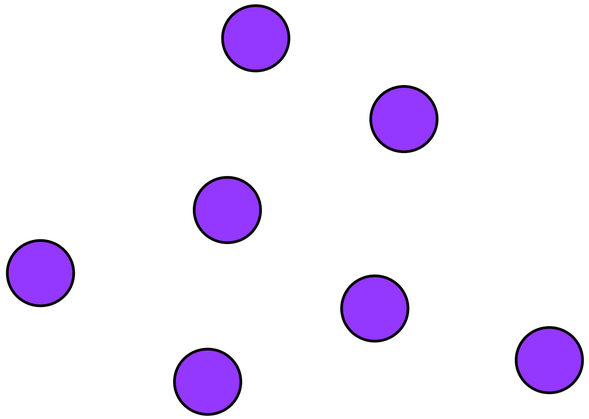
The Middle East respiratory syndrome coronavirus (MERS-CoV) is an emerging pathogen…

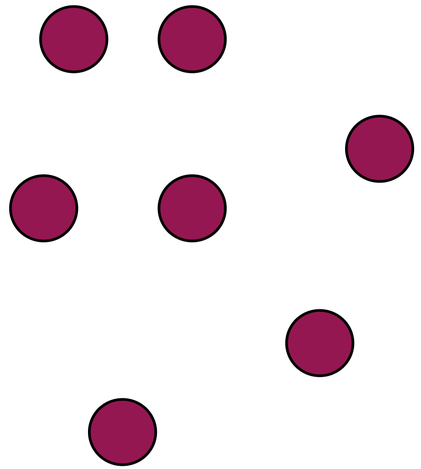
The arboviral infection, CTF, is transmitted from the bite of an infected wood tick.

Consider the task of document classification

Empirical evaluation

$\mathbb{R}^L$





Consider the task of document classification

Empirical evaluation

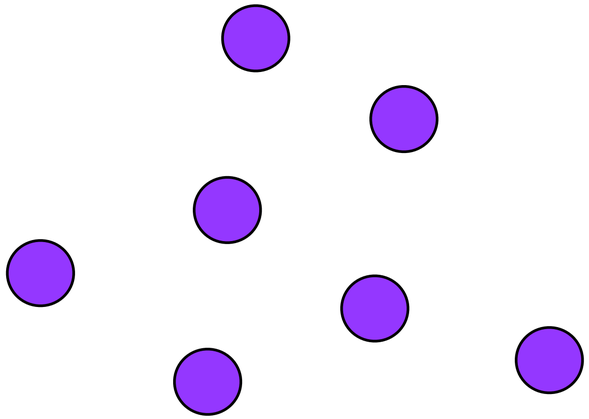
$\mathbb{R}^L$

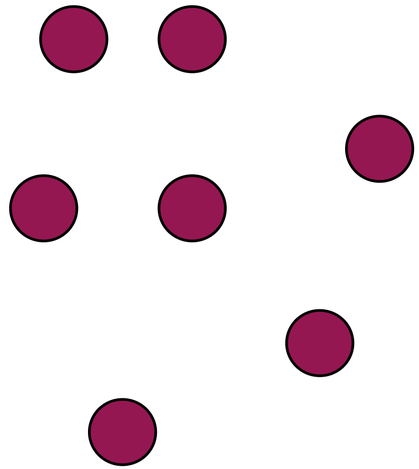
Dataset	Classes	Train	Test	BOW Dim	Length	Application
TWITTER	3	2176	932	6344	9.9	Tweets categorized by sentiment
RECIPE-L	20	27841	11933	3590	18.5	Recipe procedures labeled by origin
OHSUMED	10	3999	5153	31789	59.2	medical Abstracts (class subsampled)
20NEWS	20	11293	7528	29671	72	Canonical User-written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

Consider the task of document classification

Empirical evaluation

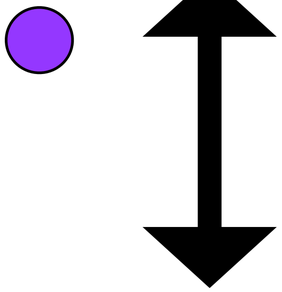
$\mathbb{R}^L$





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

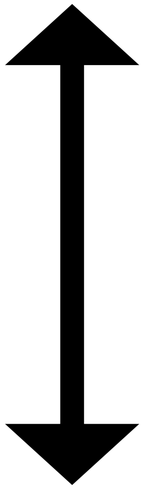
Amazingly, it is effective against SARS and **MERS**.



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

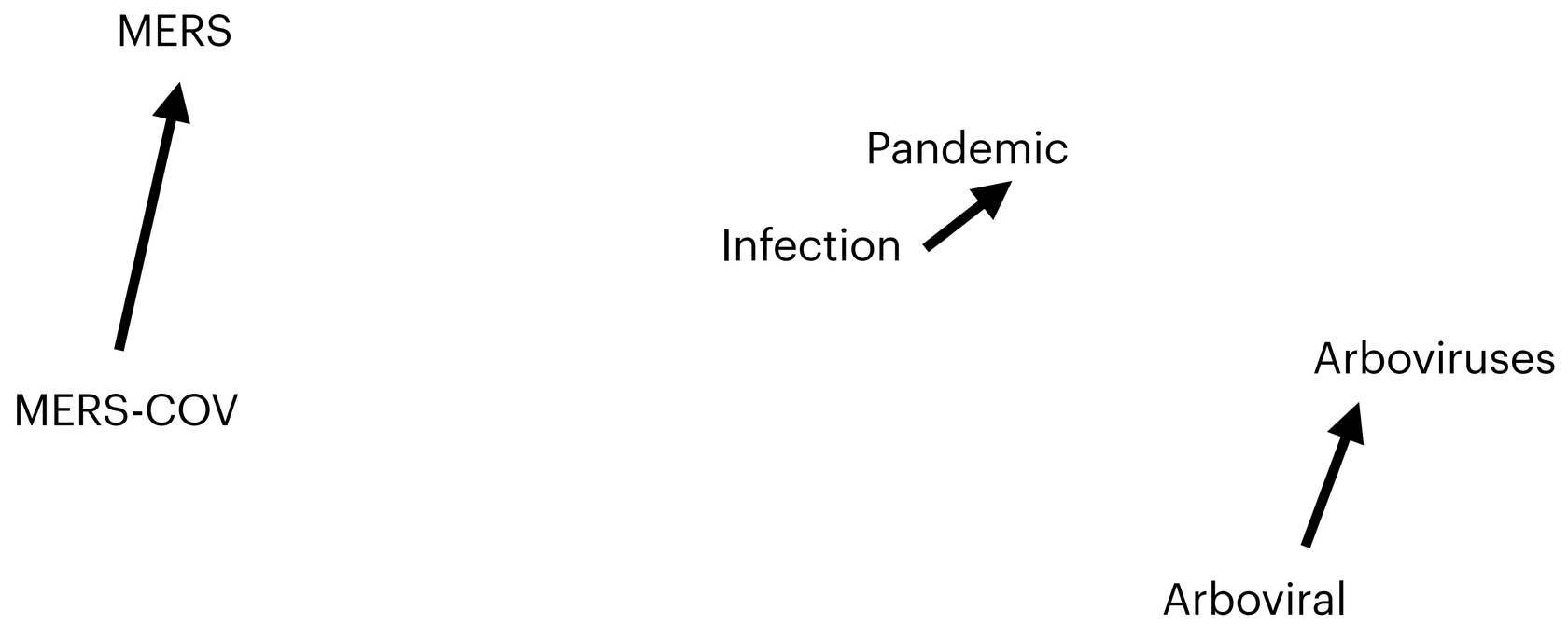
$\mathbb{R}^L$

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



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

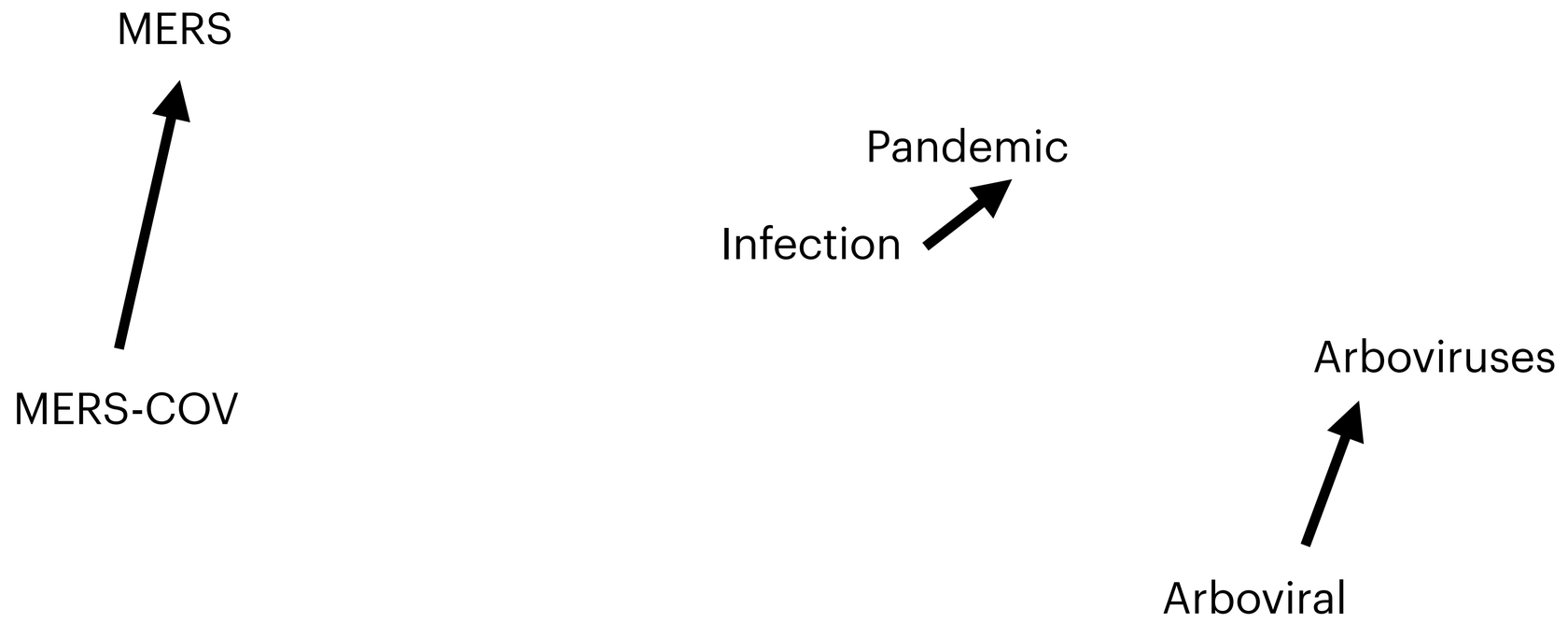
$\mathbb{R}^d$



Word embedding space (here: word2vec)

Compute cumulative distance between all words for each pair of documents

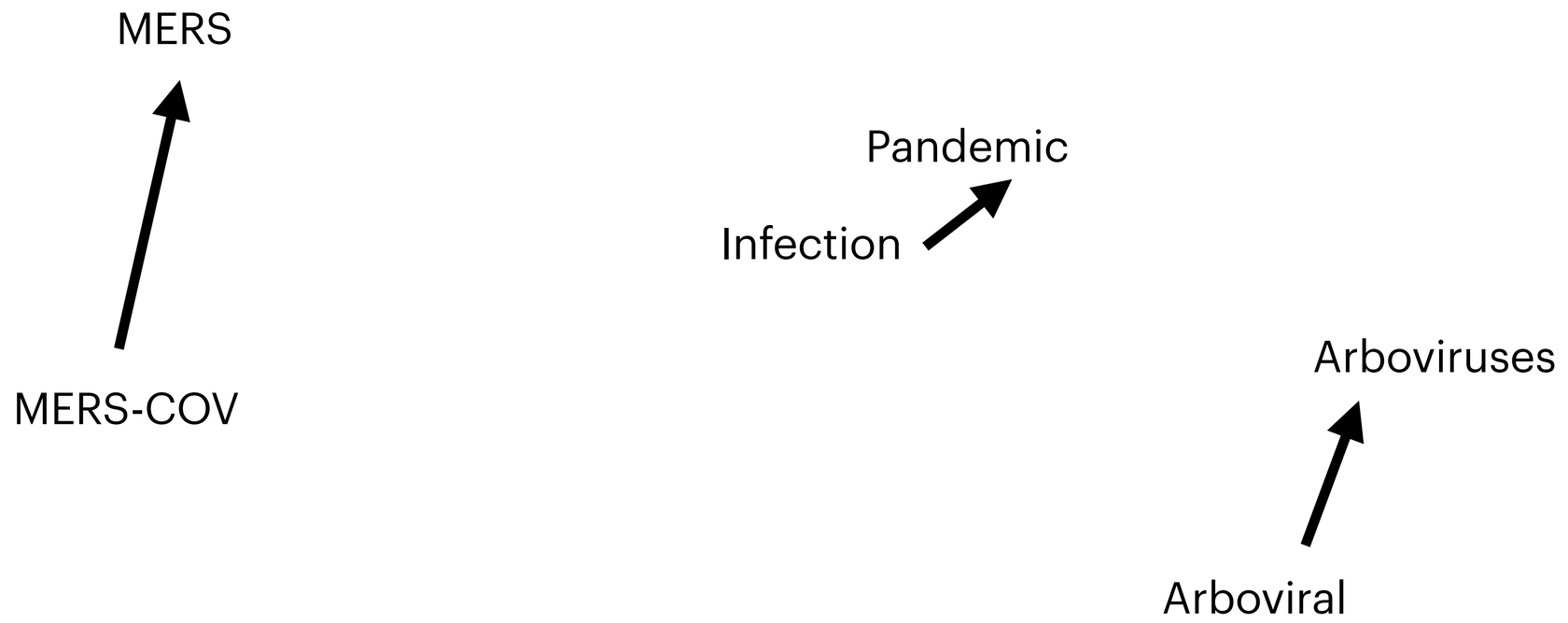
$\mathbb{R}^d$



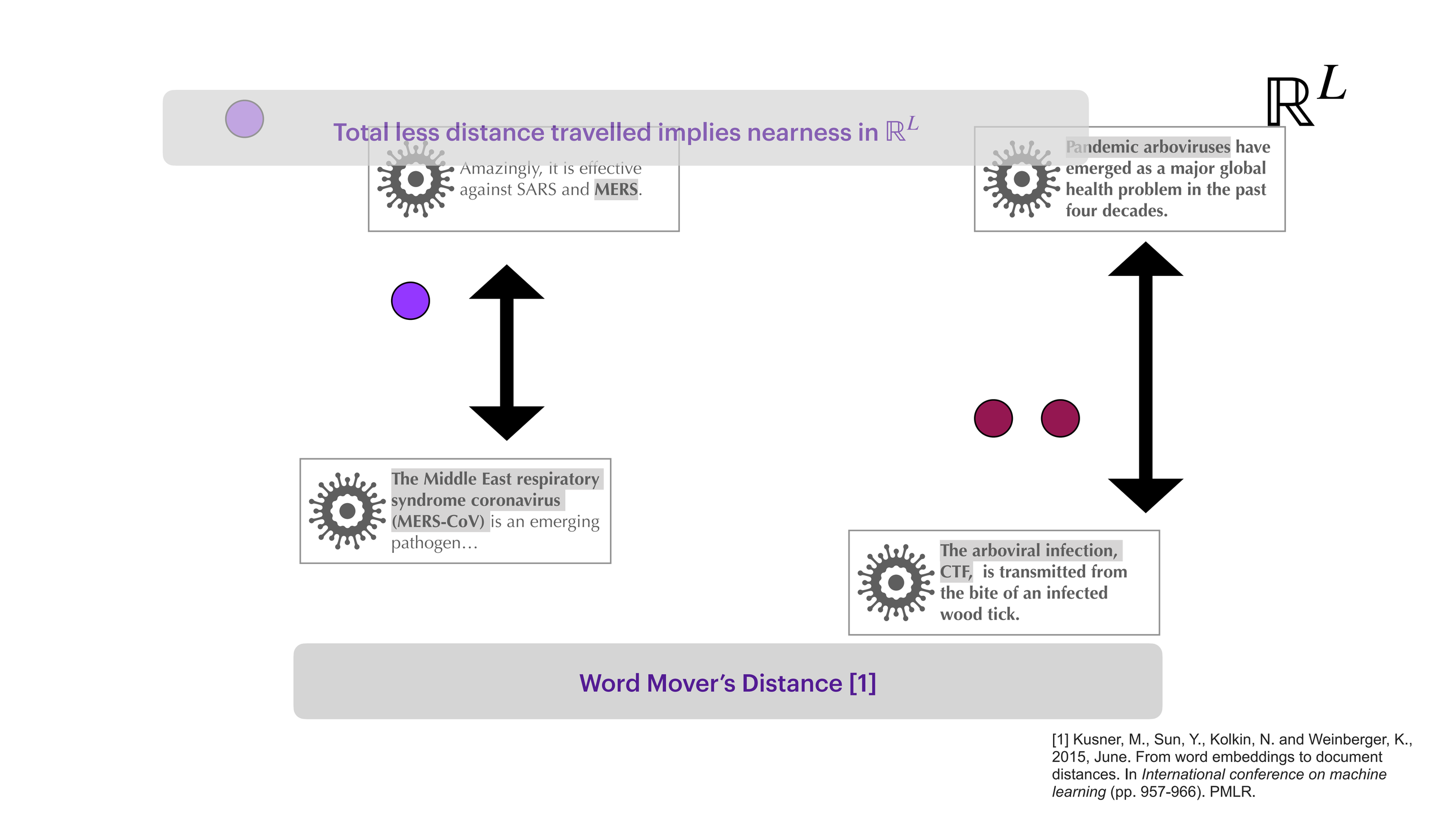
Word embedding space (here: word2vec)

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

$\mathbb{R}^d$



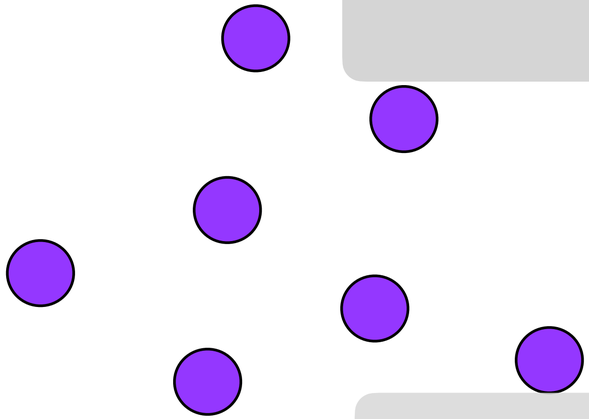
Word embedding space (here: word2vec)



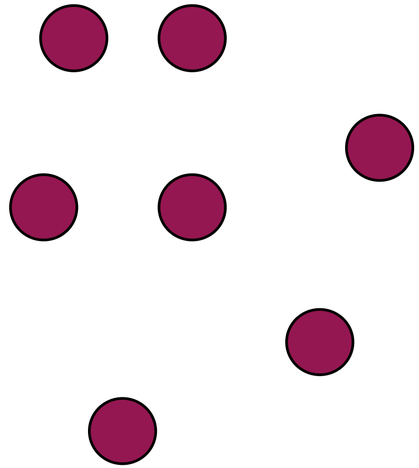
[1] Kusner, M., Sun, Y., Kolkin, N. and Weinberger, K., 2015, June. From word embeddings to document distances. In *International conference on machine learning* (pp. 957-966). PMLR.

Word Mover's Distance [1]

$\mathbb{R}^L$



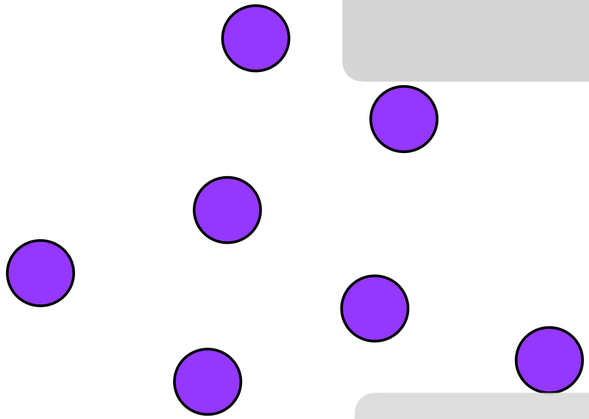
$O(n^2)$ comparisons among documents



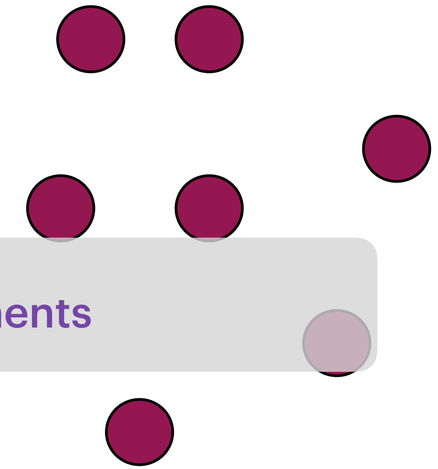
[1] Kusner, M., Sun, Y., Kolkin, N. and Weinberger, K., 2015, June. From word embeddings to document distances. In *International conference on machine learning* (pp. 957-966). PMLR.

Word Mover's Distance [1]

$\mathbb{R}^L$



$O(n^2)$ comparisons among documents

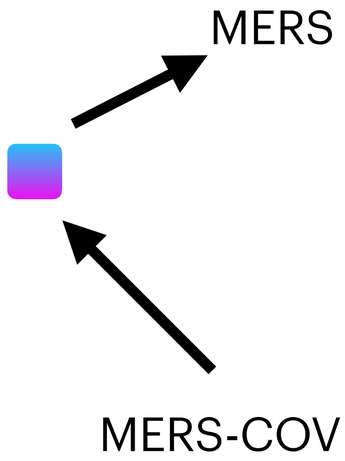


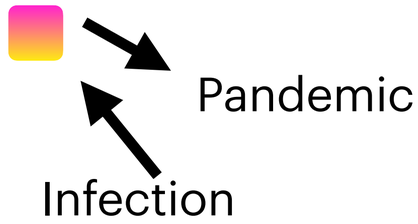
$O(L^3 \log L)$ comparisons between documents

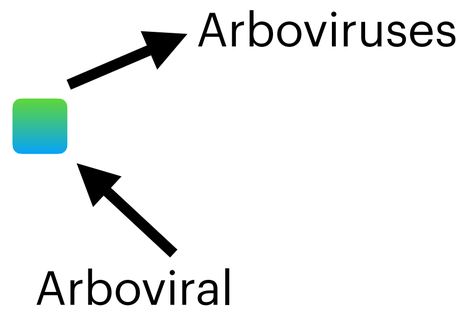
[1] Kusner, M., Sun, Y., Kolkin, N. and Weinberger, K., 2015, June. From word embeddings to document distances. In *International conference on machine learning* (pp. 957-966). PMLR.

Word Mover's Embedding [1]

$\mathbb{R}^d$





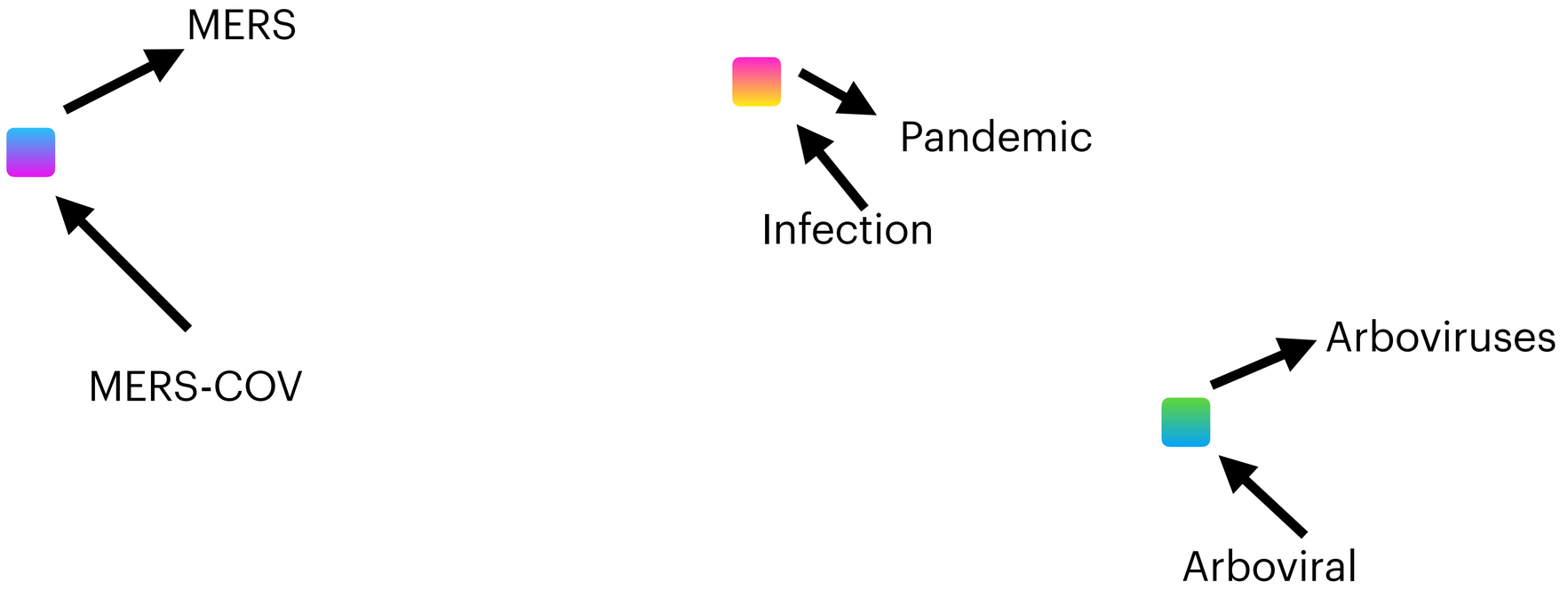


are words from random documents

[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]

$\mathbb{R}^d$

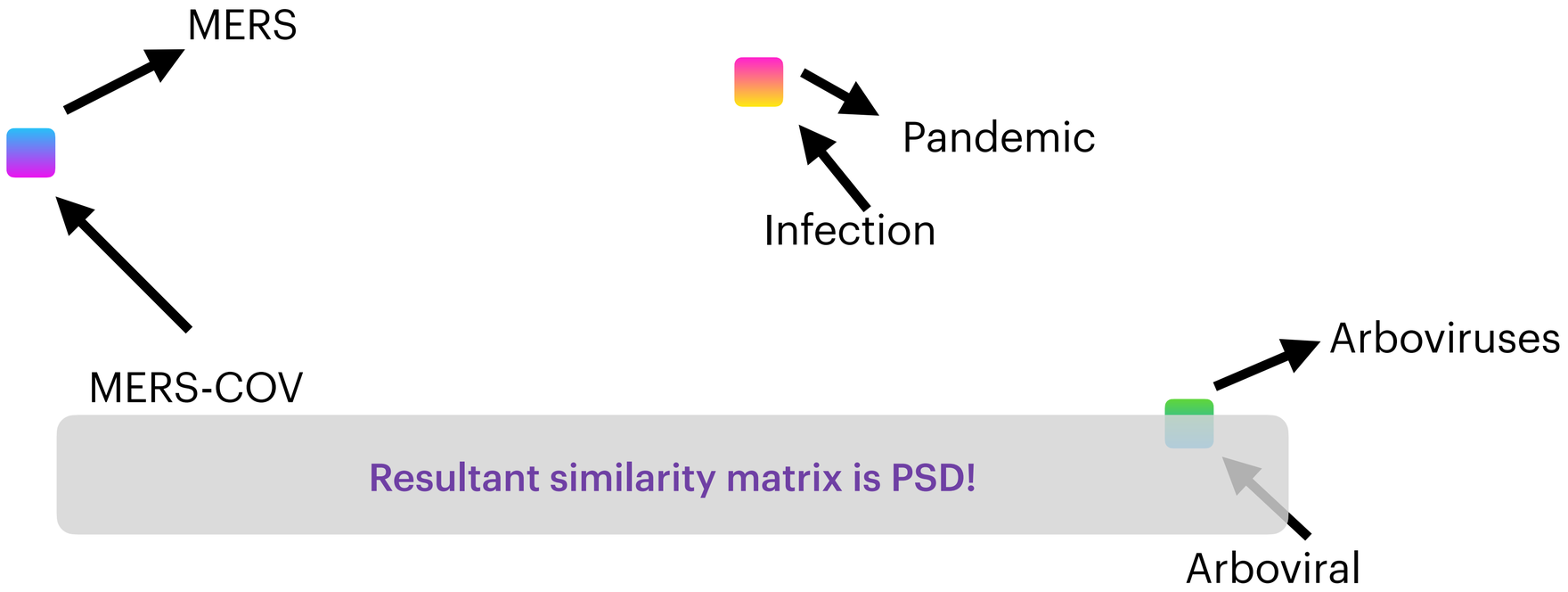


Construct Word Mover's Kernel (WMK) using infinite dimensional feature map to random documents from a given distribution

[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]

$\mathbb{R}^d$



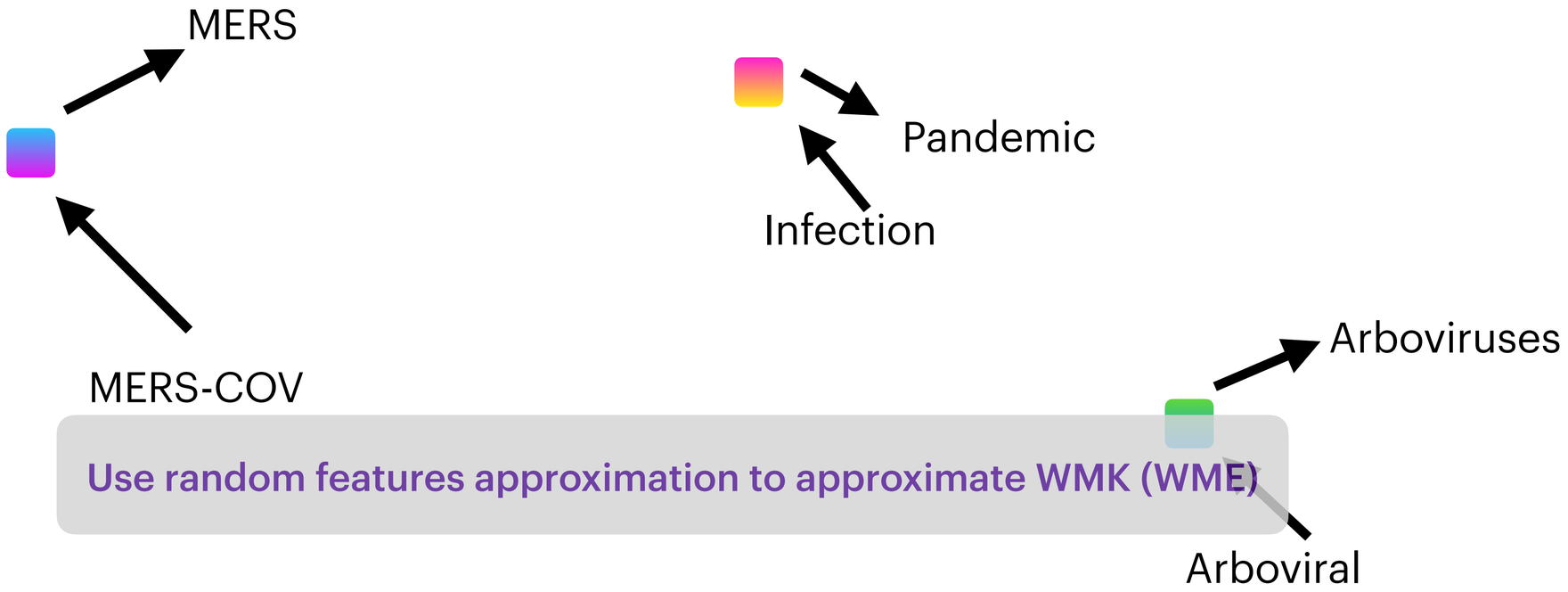
Resultant similarity matrix is PSD!

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]

$\mathbb{R}^d$



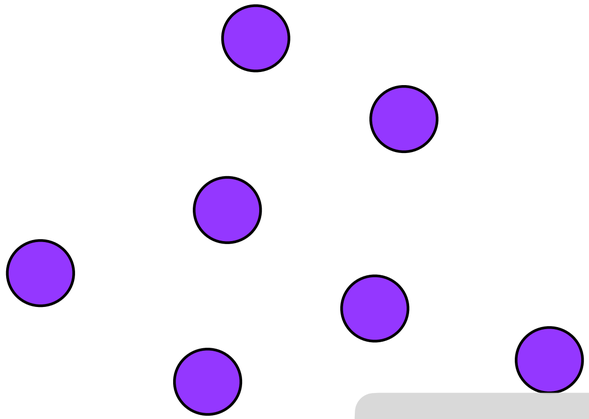
Use random features approximation to approximate WMK (WME)

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

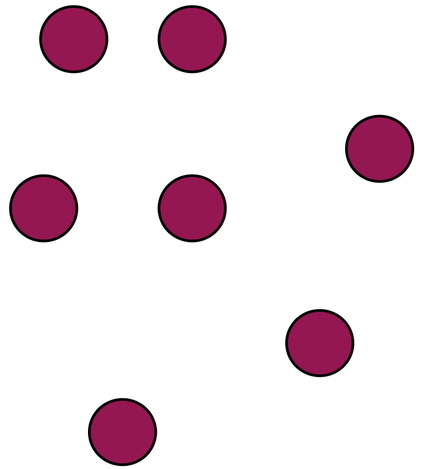
[1] Wu, L., Yen, I.E., Xu, K., Xu, F., Balakrishnan, A., Chen, P.Y., Ravikumar, P. and Witbrock, M.J., 2018. Word mover's embedding: From word2vec to document embedding. *arXiv preprint arXiv:1811.01713*.

Our Approach

$\mathbb{R}^L$

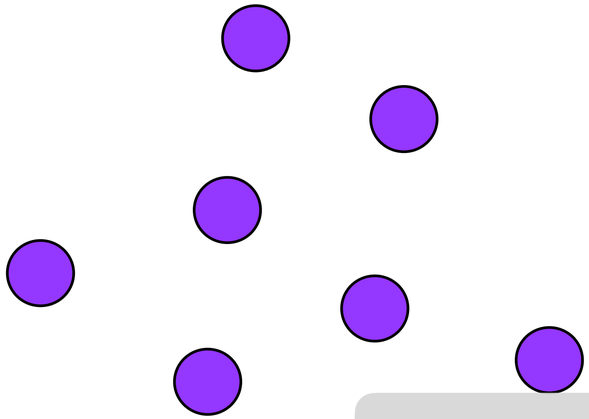


Similarity between documents x, y: $exp(-\gamma \text{WMD}(x, y))$

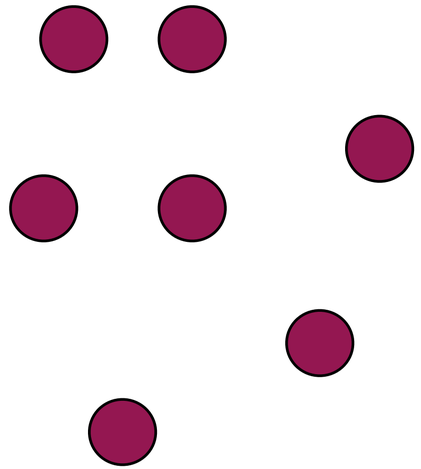


Our Approach

$\mathbb{R}^L$

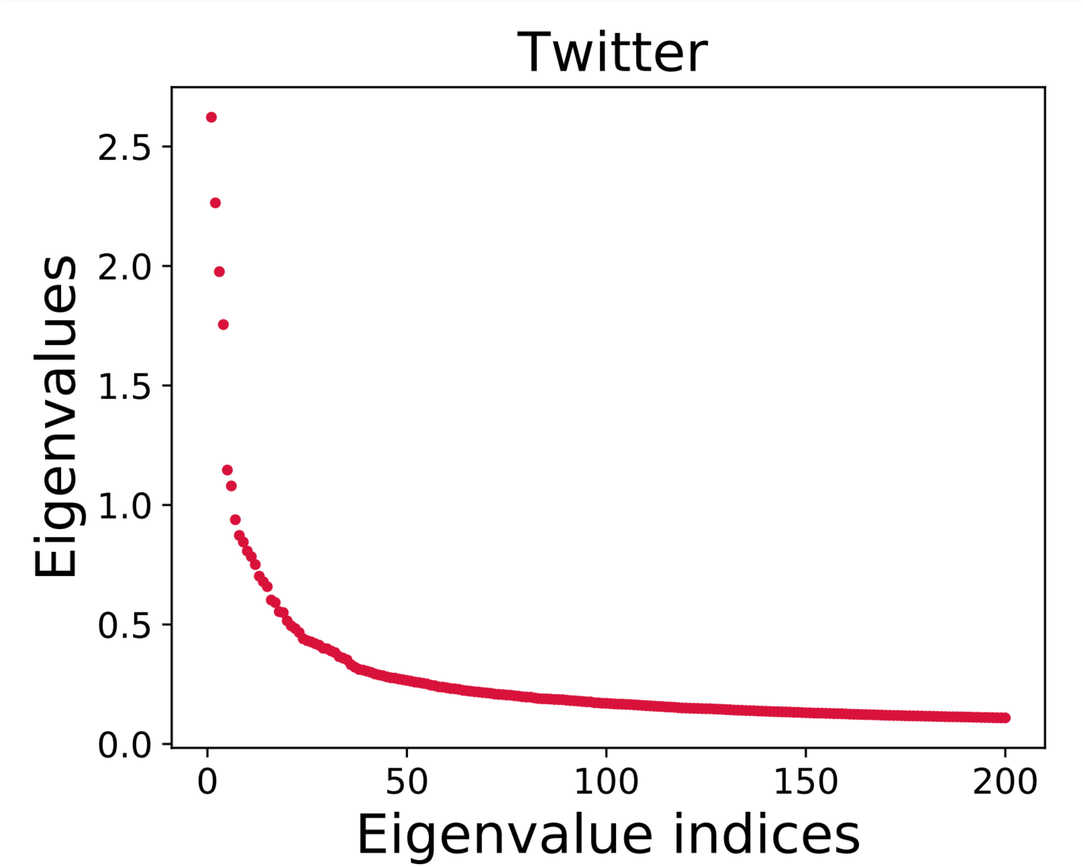


Similarity between documents x, y: $exp(-\gamma \text{WMD}(x, y))$



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

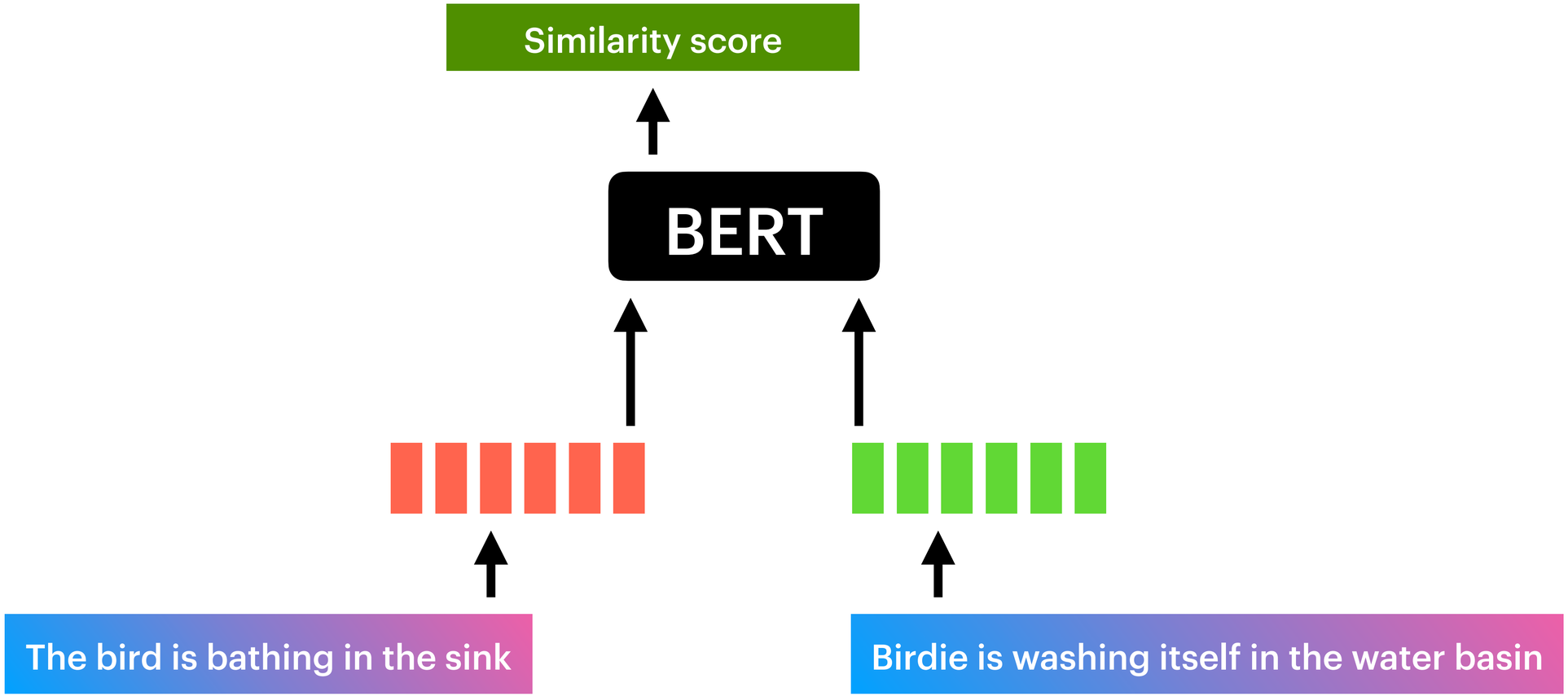
Sentence Similarity Task

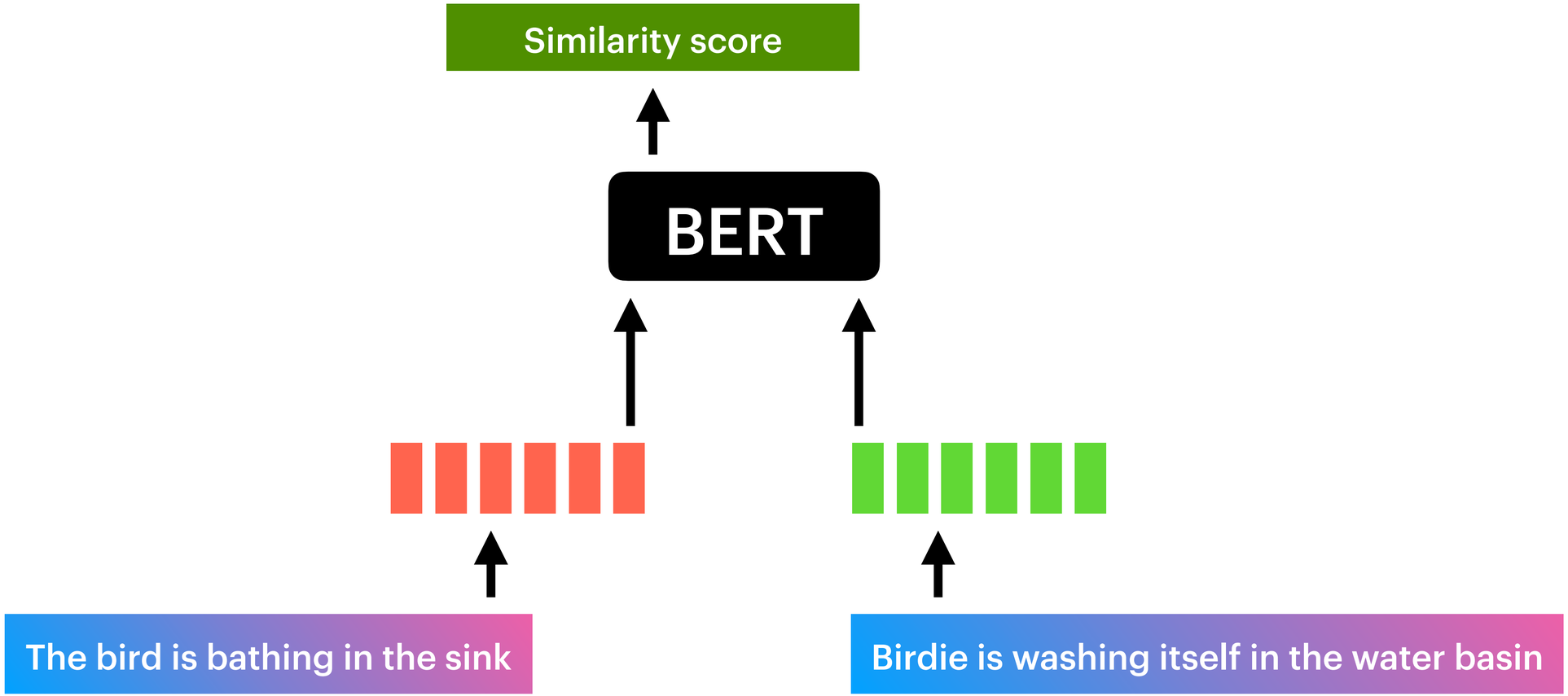
Score	English	Spanish
5/4	*The two sentences are completely equivalent, as they mean the same thing.*	
	The bird is bathing in the sink. Birdie is washing itself in the water basin.	El pájaro se esta bañando en el lavabo. El pájaro se está lavando en el aguamanil.
4	*The two sentences are mostly equivalent, but some unimportant details differ.*	
	In May 2010, the troops attempted to invade Kabul. The US army invaded Kabul on May 7th last year, 2010.	
3	*The two sentences are roughly equivalent, but some important information differs/missing.*	
	John said he is considered a witness but not a suspect. "He is not a suspect anymore." John said.	John dijo que él es considerado como testigo, y no como sospechoso. "Él ya no es un sospechoso," John dijo.
2	*The two sentences are not equivalent, but share some details.*	
	They flew out of the nest in groups. They flew into the nest together.	Ellos volaron del nido en grupos. Volaron hacia el nido juntos.
1	*The two sentences are not equivalent, but are on the same topic.*	
	The woman is playing the violin. The young lady enjoys listening to the guitar.	La mujer está tocando el violín. La joven disfruta escuchar la guitarra.
0	*The two sentences are completely dissimilar.*	
	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.	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.

Sentence Similarity Task

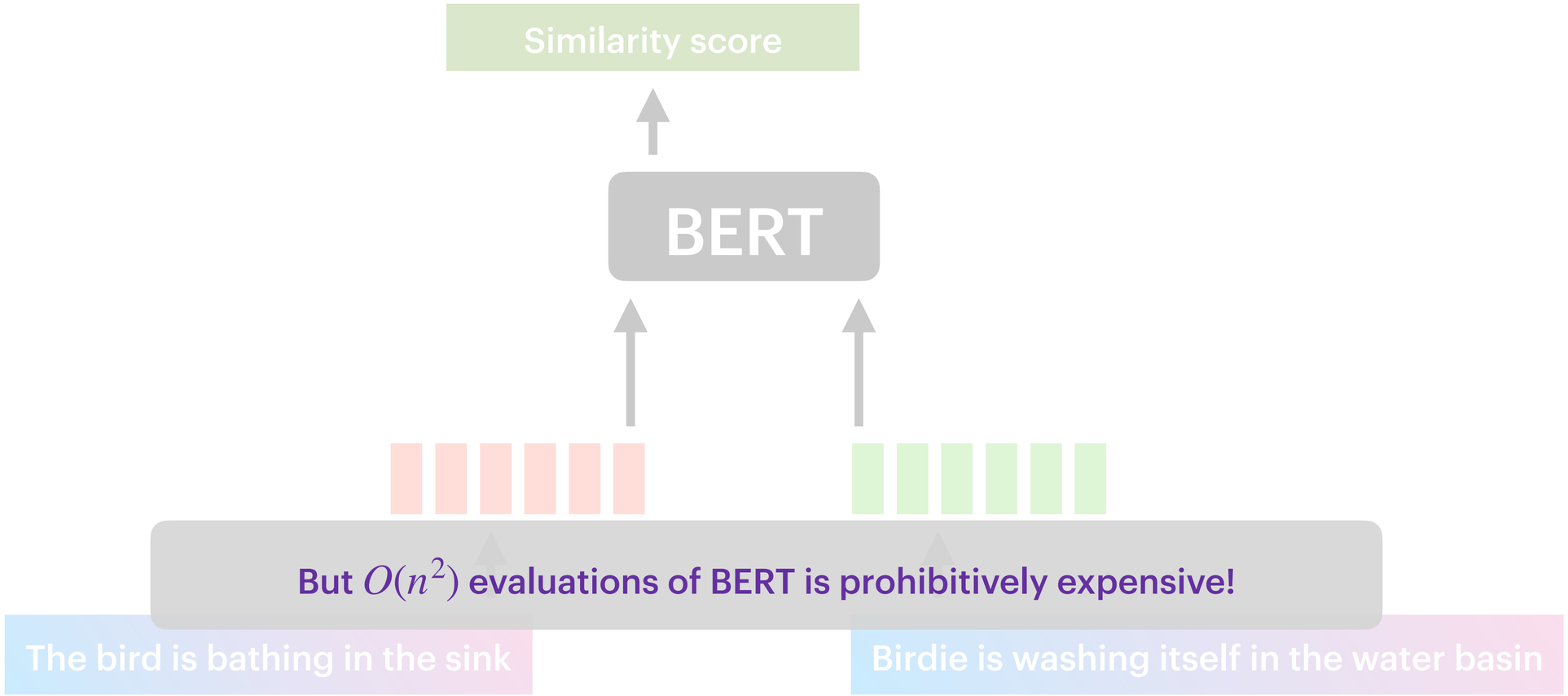
Score	English	Spanish
5/4	*The two sentences are completely equivalent, as they mean the same thing.*	
	The bird is bathing in the sink. Birdie is washing itself in the water basin.	El pájaro se esta bañando en el lavabo. El pájaro se está lavando en el aguamanil.
4	*The two sentences are mostly equivalent, but some unimportant details differ.*	
	In May 2010, the troops attempted to invade Kabul.	
	They flew out of the nest in groups. They flew into the nest together.	Ellos volaron del nido en grupos. Volaron hacia el nido juntos.
1	*The two sentences are not equivalent, but are on the same topic.*	
	The woman is playing the violin. The young lady enjoys listening to the guitar.	La mujer está tocando el violín. La joven disfruta escuchar la guitarra.
0	*The two sentences are completely dissimilar.*	
	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.	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.

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


Similarity score
BERT
The bird is bathing in the sink
Birdie is washing itself in the water basin

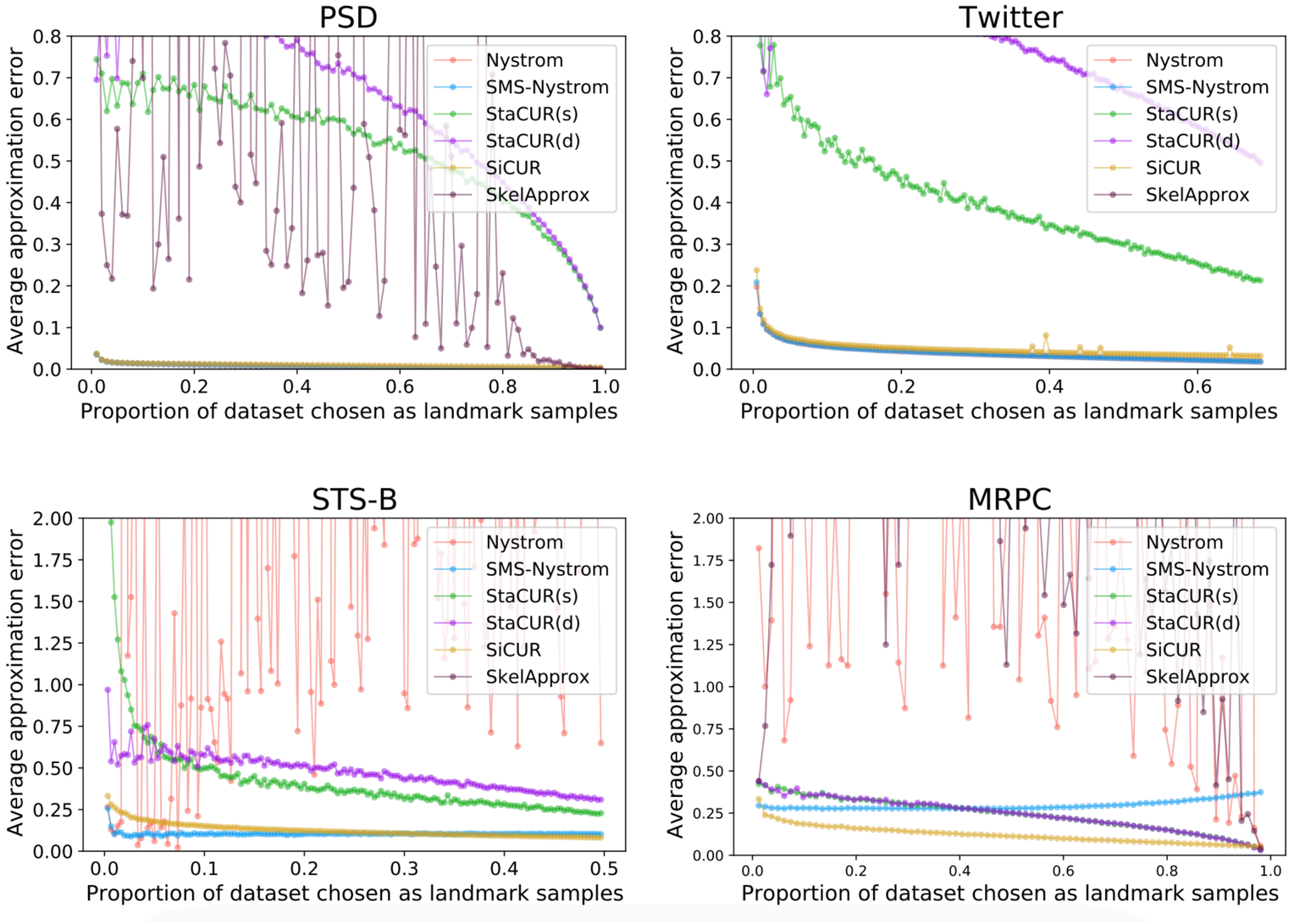


$O(n^2)$ similarities for all pairs

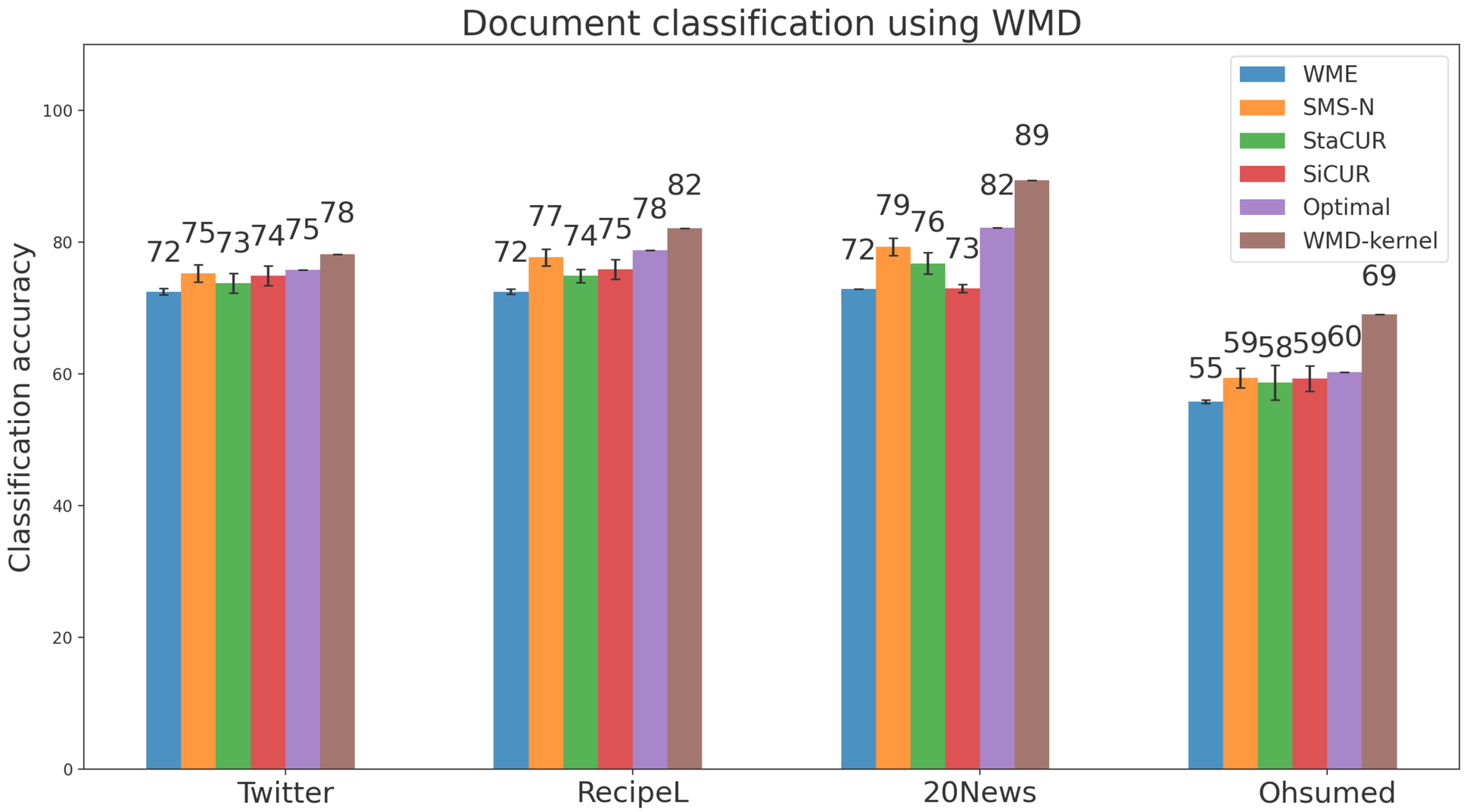


$O(n^2)$ similarities for all pairs

Comparing Approximation on Given Datasets

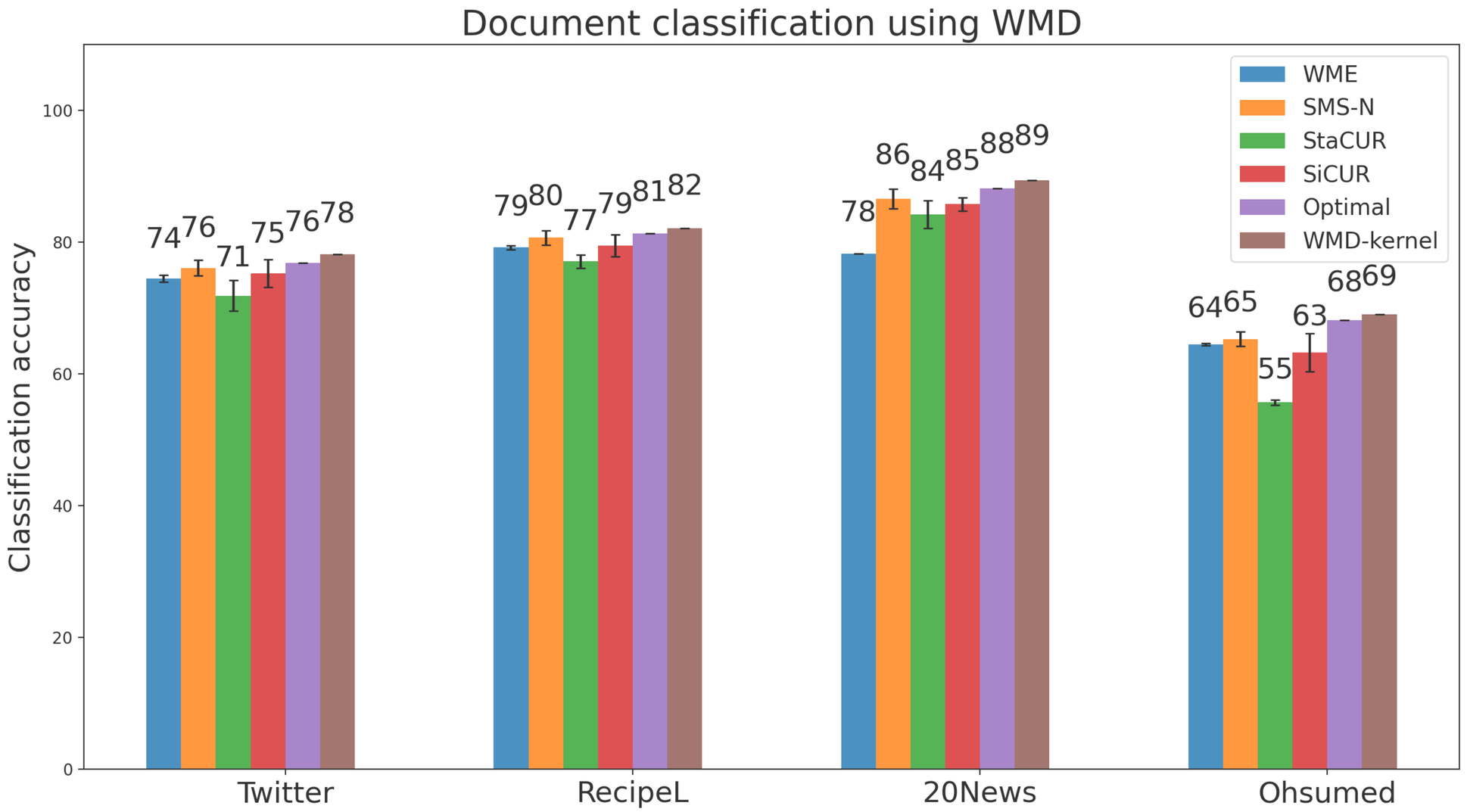


Document Classification Task



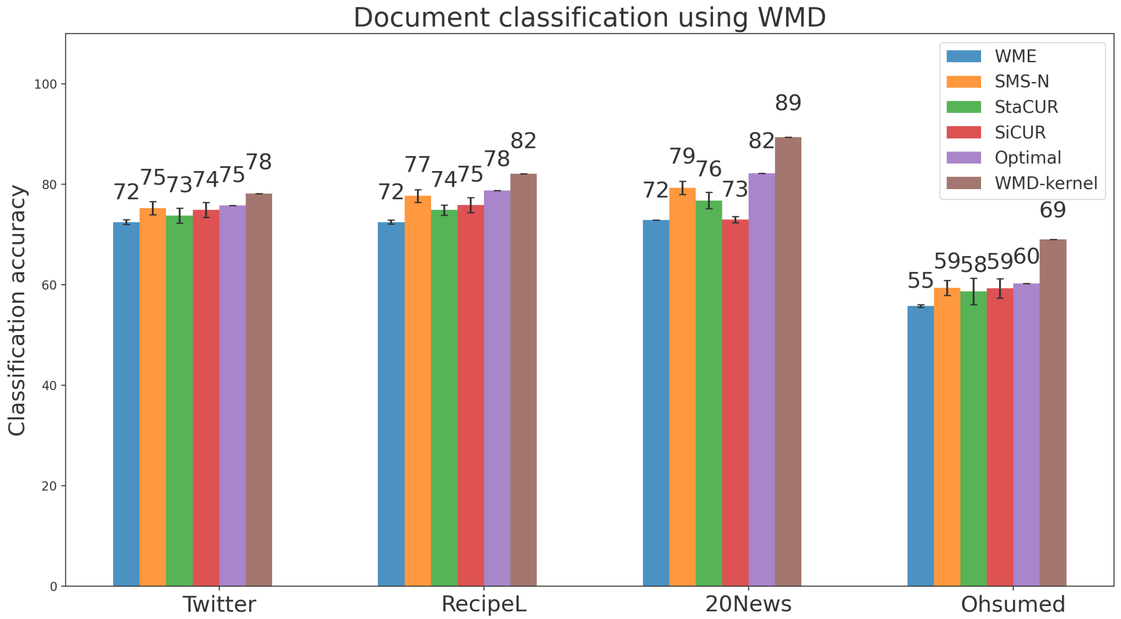
Small rank accuracy scores

Document Classification Task

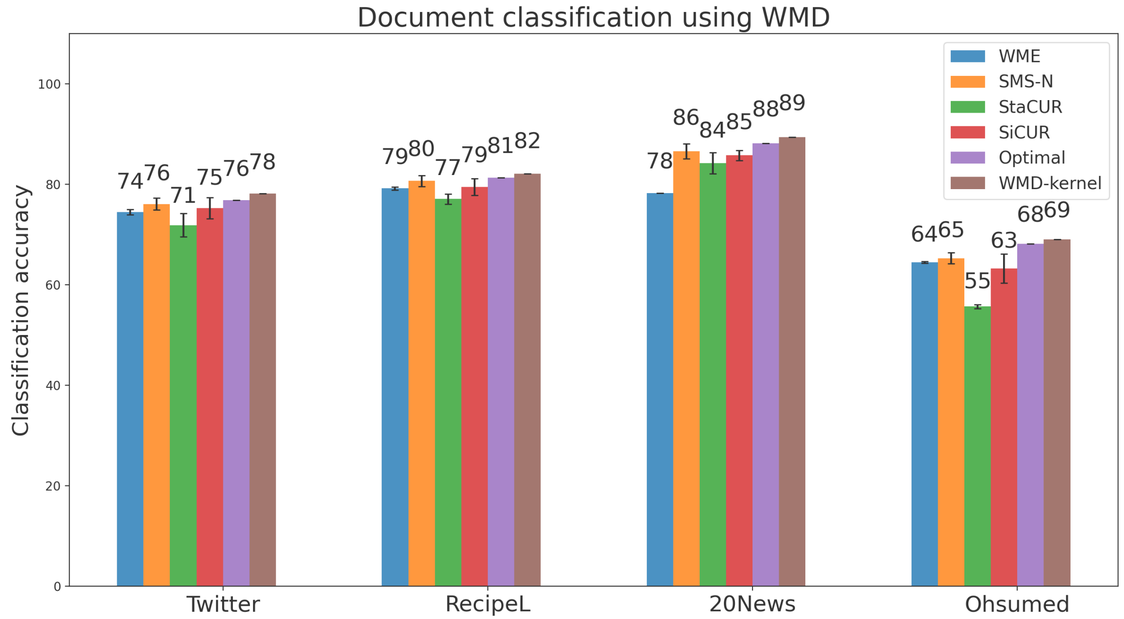


Large rank accuracy scores

Document Classification Task



Small rank accuracy scores

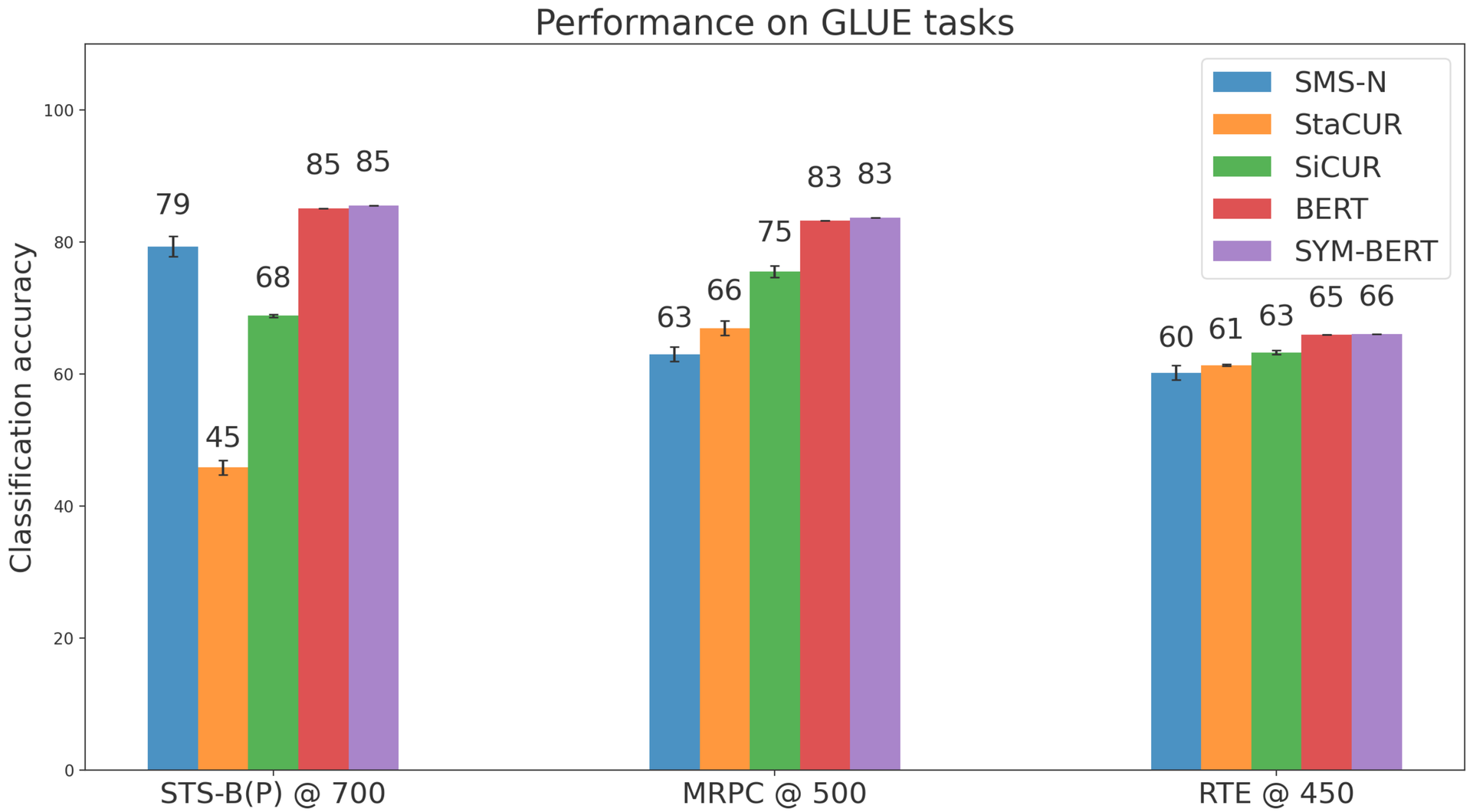


Large rank accuracy scores

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

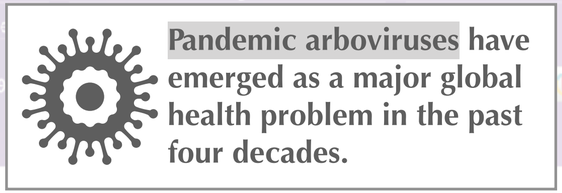
Time in seconds

Approximation of GLUE tasks

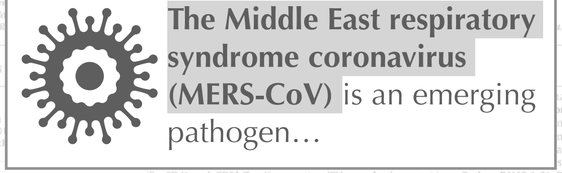


Cross-Document Entity & Coreference

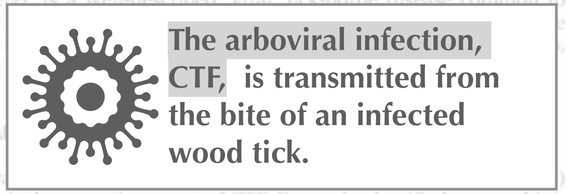
Amazingly, it is effective against SARS and **MERS**.



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

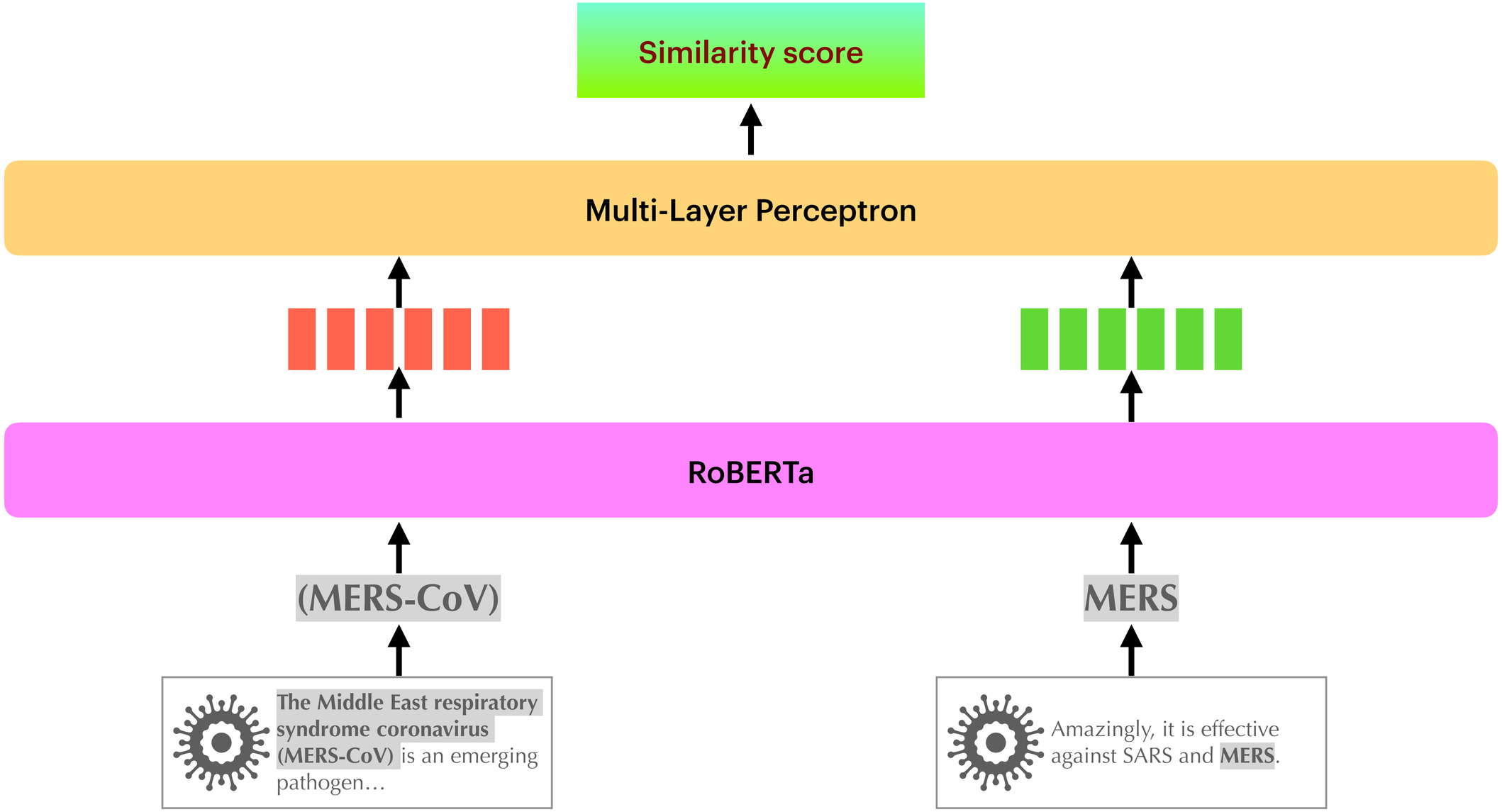


The Middle East respiratory syndrome coronavirus (MERS-CoV) is an emerging pathogen…

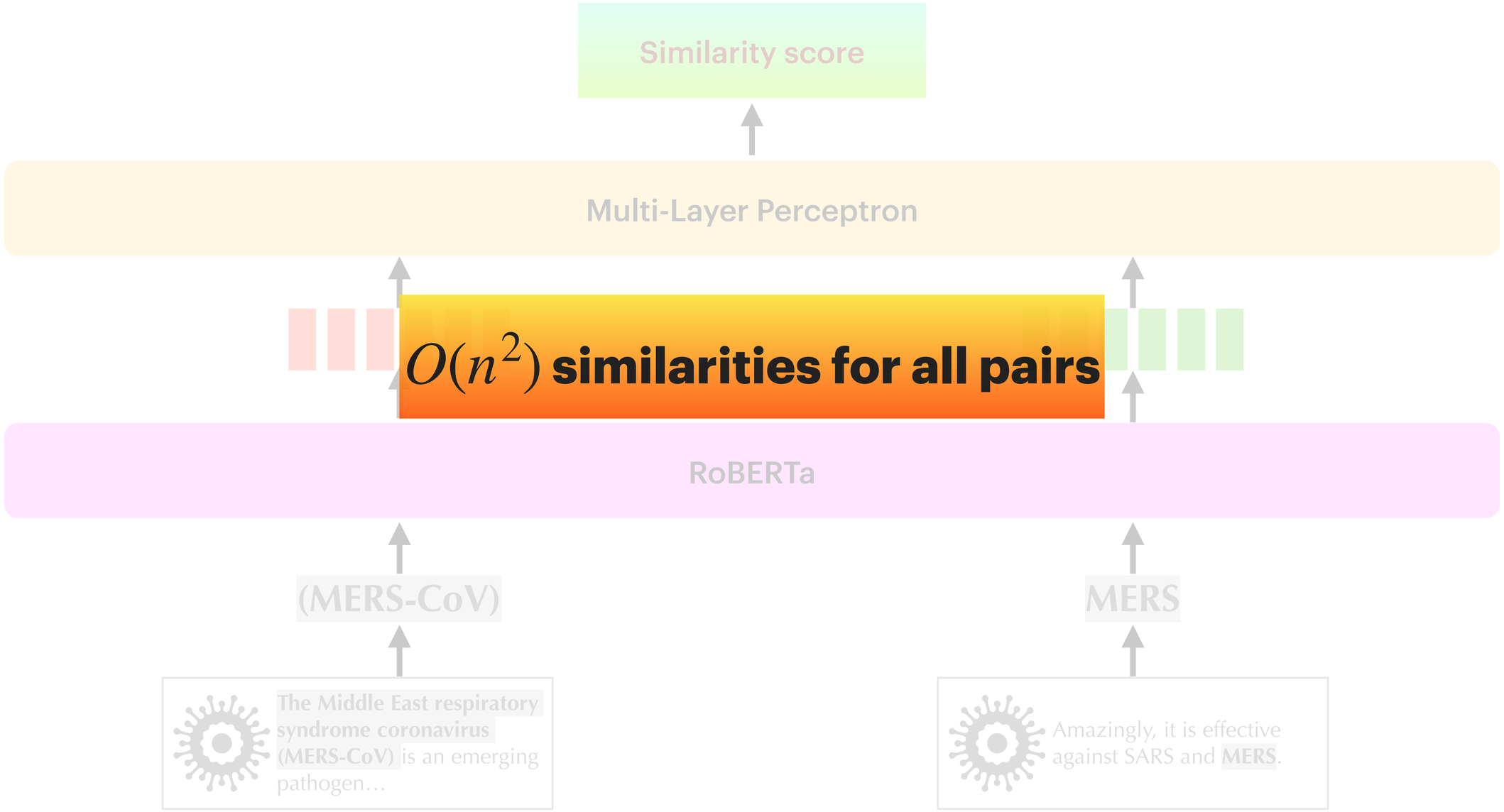


The arboviral infection, CTF, is transmitted from the bite of an infected wood tick.

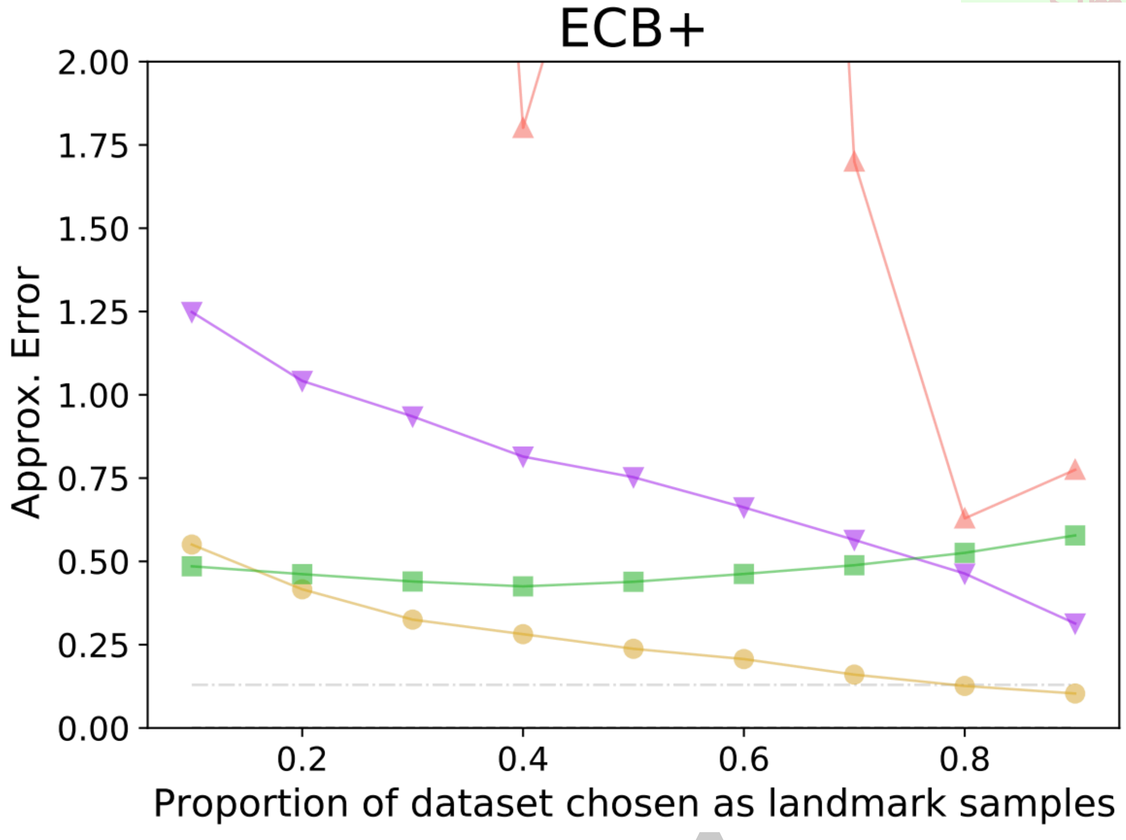
Cross-Document Entity & Coreference

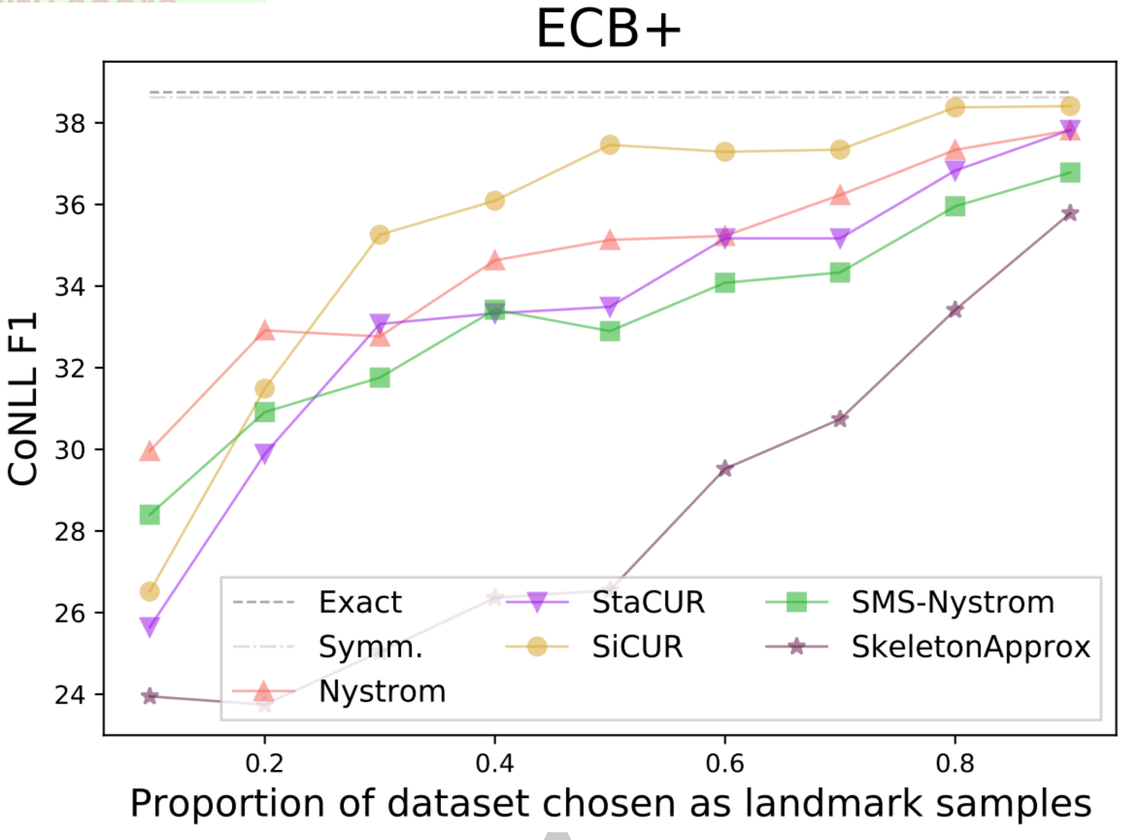


Cross-Document Entity & Coreference



Cross-Document Entity & Coreference





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

Amazingly, it is effective against SARS and MERS.

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?