From Word Embeddings To Document Distances

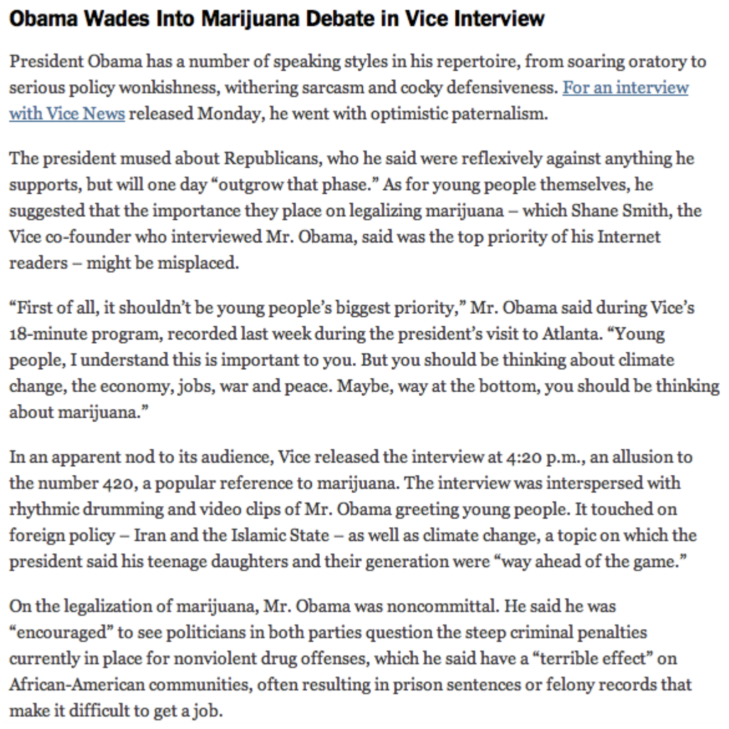
Matt J. Kusner

Yu Sun

Nicholas I. Kolkin

Kilian Q. Weinberger

Goal: a distance between two documents



Obama Wades Into Marijuana Debate in Vice Interview

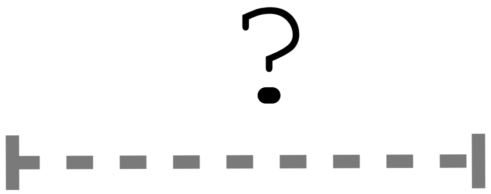
President Obama has a number of speaking styles in his repertoire, from soaring oratory to serious policy wonkishness, withering sarcasm and cocky defensiveness. For an interview with Vice News released Monday, he went with optimistic paternalism.

The president mused about Republicans, who he said were reflexively against anything he supports, but will one day "outgrow that phase." As for young people themselves, he suggested that the importance they place on legalizing marijuana – which Shane Smith, the Vice co-founder who interviewed Mr. Obama, said was the top priority of his Internet readers – might be misplaced.

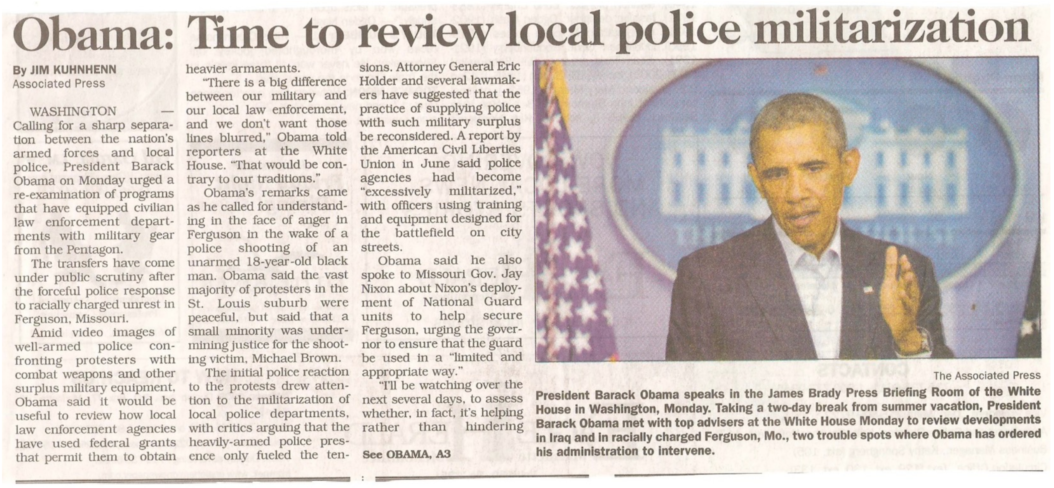
"First of all, it shouldn't be young people's biggest priority," Mr. Obama said during Vice's 18-minute program, recorded last week during the president's visit to Atlanta. "Young people, I understand this is important to you. But you should be thinking about climate change, the economy, jobs, war and peace. Maybe, way at the bottom, you should be thinking about marijuana."

In an apparent nod to its audience, Vice released the interview at 4:20 p.m., an allusion to the number 420, a popular reference to marijuana. The interview was interspersed with rhythmic drumming and video clips of Mr. Obama greeting young people. It touched on foreign policy – Iran and the Islamic State – as well as climate change, a topic on which the president said his teenage daughters and their generation were "way ahead of the game."

On the legalization of marijuana, Mr. Obama was noncommittal. He said he was "encouraged" to see politicians in both parties question the steep criminal penalties currently in place for nonviolent drug offenses, which he said have a "terrible effect" on African-American communities, often resulting in prison sentences or felony records that make it difficult to get a job.



?



Obama: Time to review local police militarization

By JIM KUHNHENN
Associated Press

WASHINGTON — Calling for a sharp separation between the nation's armed forces and local police, President Barack Obama on Monday urged a re-examination of programs that have equipped civilian law enforcement departments with military gear from the Pentagon.

The transfers have come under public scrutiny after the forceful police response to racially charged unrest in Ferguson, Missouri.

Amid video images of well-armed police confronting protesters with combat weapons and other surplus military equipment, Obama said it would be useful to review how local law enforcement agencies have used federal grants that permit them to obtain heavier armaments.

"There is a big difference between our military and our local law enforcement, and we don't want those lines blurred," Obama told reporters at the White House. "That would be contrary to our traditions."

Obama's remarks came as he called for understanding in the face of anger in Ferguson in the wake of a police shooting of an unarmed 18-year-old black man. Obama said the vast majority of protesters in the St. Louis suburb were peaceful, but said that a small minority was undermining justice for the shooting victim, Michael Brown.

The initial police reaction to the protests drew attention to the militarization of local police departments, with critics arguing that the heavily-armed police presence only fueled the tensions. Attorney General Eric Holder and several lawmakers have suggested that the practice of supplying police with such military surplus be reconsidered. A report by the American Civil Liberties Union in June said police agencies had become "excessively militarized," with officers using training and equipment designed for the battlefield on city streets.

Obama said he also spoke to Missouri Gov. Jay Nixon about Nixon's deployment of National Guard units to help secure Ferguson, urging the governor to ensure that the guard be used in a "limited and appropriate way."

"I'll be watching over the next several days, to assess whether, in fact, it's helping rather than hindering

See OBAMA, A3

The Associated Press

President Barack Obama speaks in the James Brady Press Briefing Room of the White House in Washington, Monday. Taking a two-day break from summer vacation, President Barack Obama met with top advisers at the White House Monday to review developments in Iraq and in racially charged Ferguson, Mo., two trouble spots where Obama has ordered his administration to intervene.

Applications

document classification

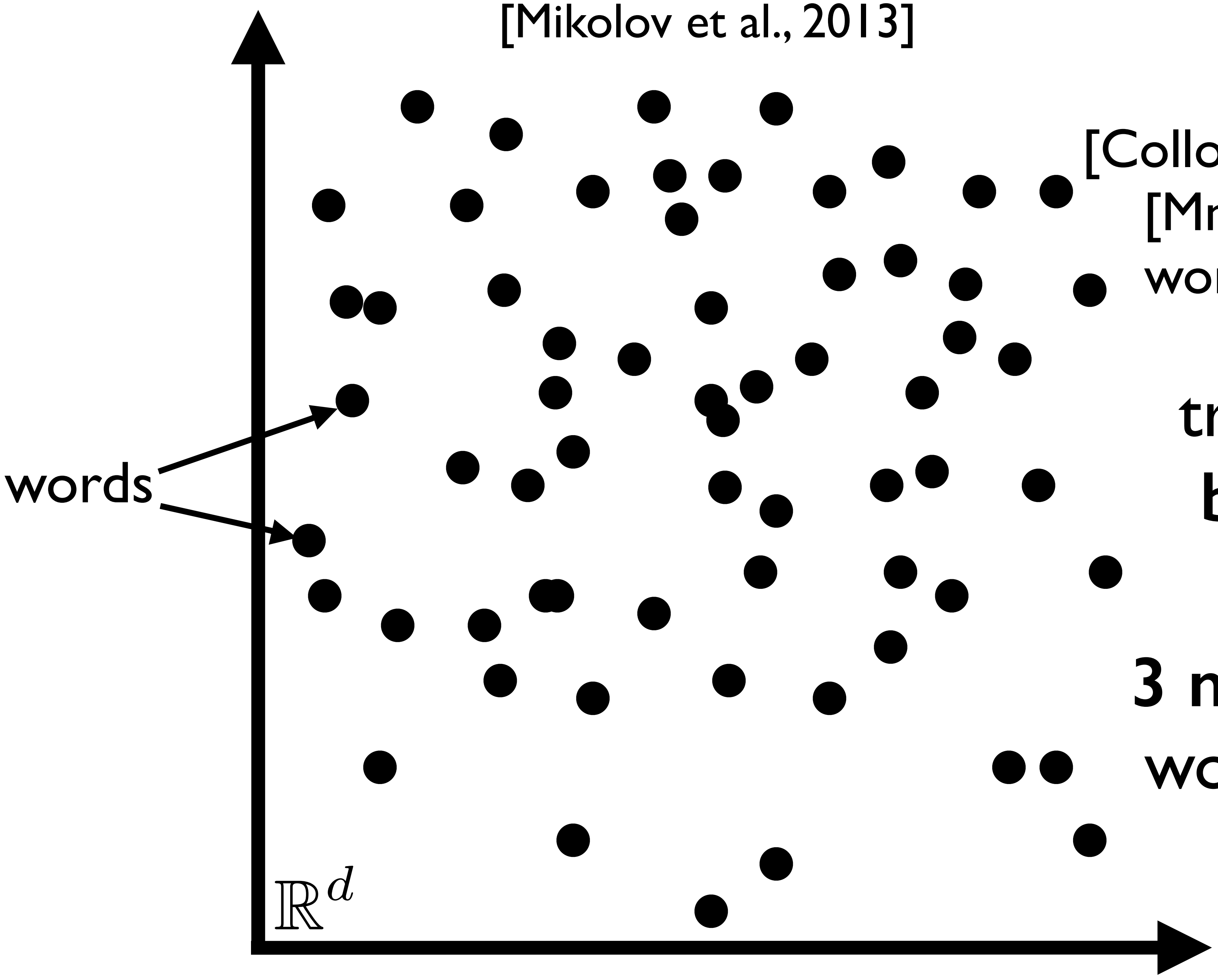
Extra fly a boost for Tasmanian anglers
By STEVEN DALLY
Chief Political Reporter
Metro/State
State workers fear layoffs
Barbour: Workers not
FEELING FROGGY
Ware Shoals Biology Club students hope to save amphibians
Preschool: Not boot camp for kindergarten

multi-lingual document matching

song identification

Word Embedding

word2vec
[Mikolov et al., 2013]



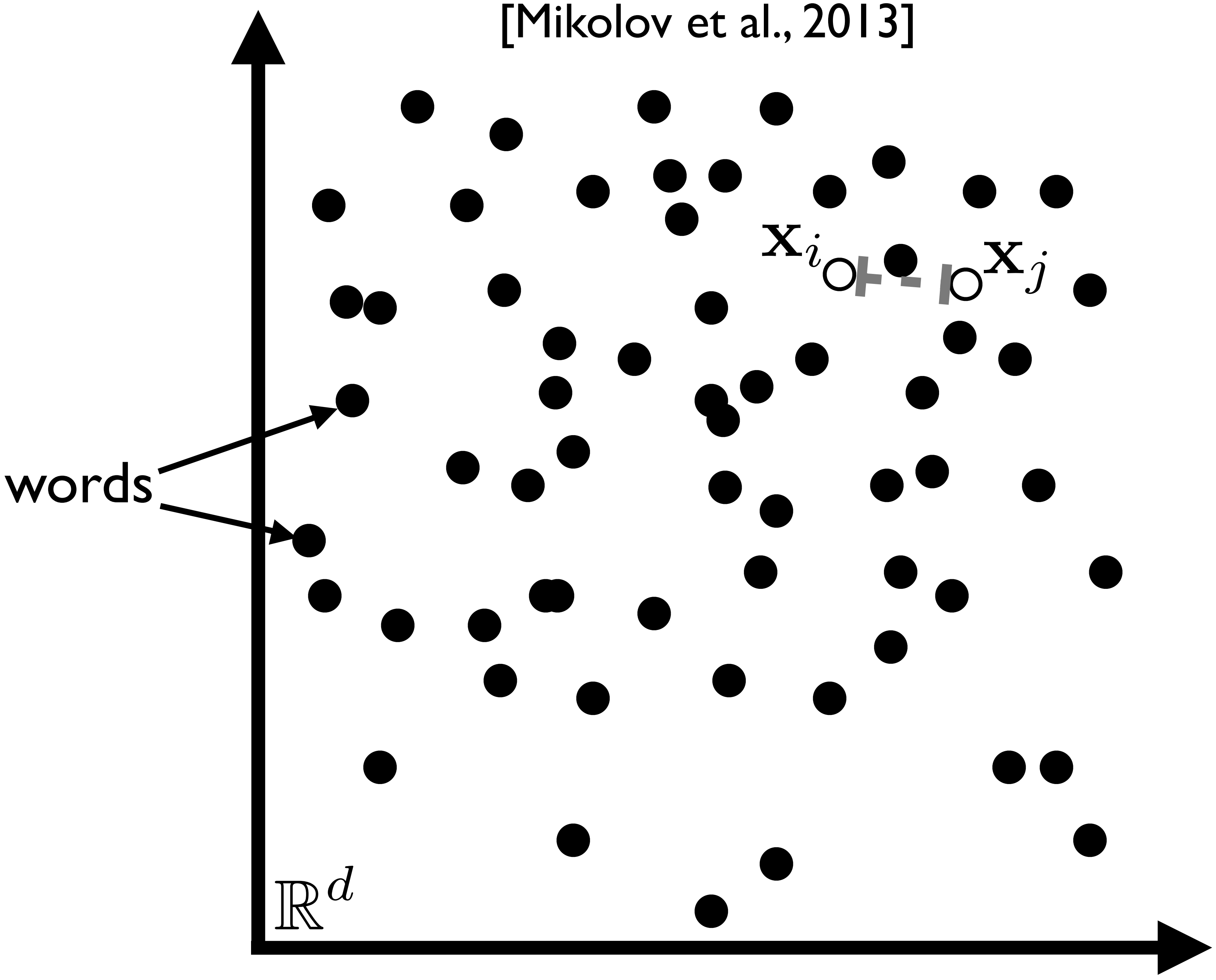
different from
[Collobert & Weston, 2008]
[Mnih & Hinton, 2009]
word2vec is not deep!

trained on **100 billion** words

3 million different words embedded

Word Embedding

word2vec
[Mikolov et al., 2013]



$$\mathbf{X} \in \mathbb{R}^{d \times n}$$

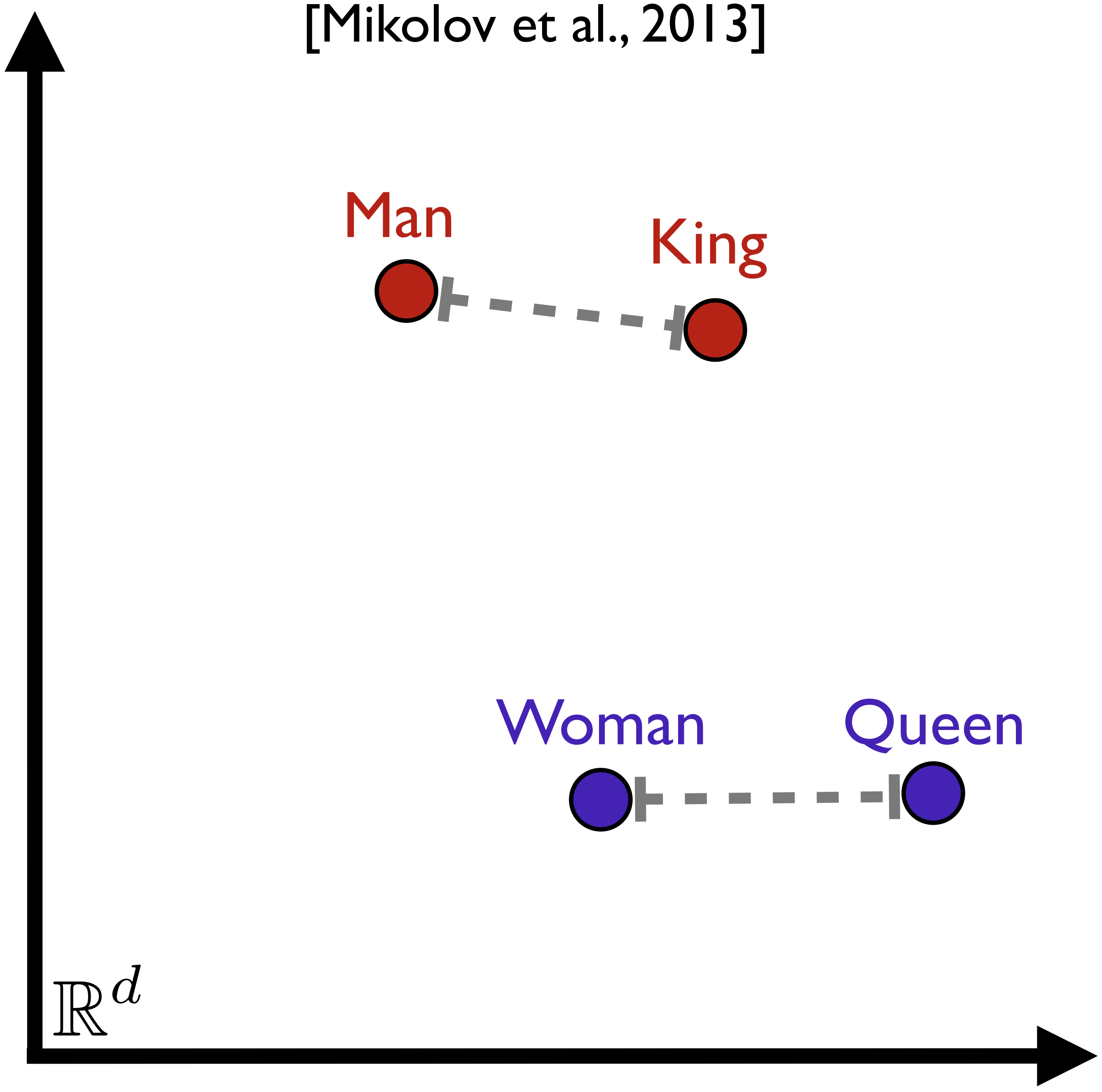
distance between words i and j:

$$\|\mathbf{x}_i - \mathbf{x}_j\|_2$$

is roughly their dissimilarity

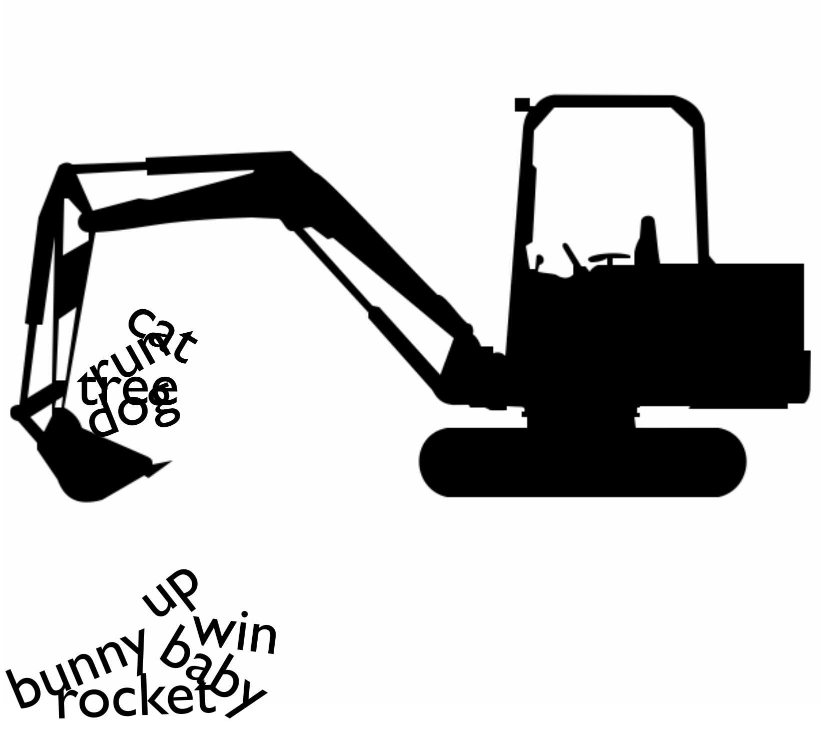
Word Embedding

word2vec
[Mikolov et al., 2013]

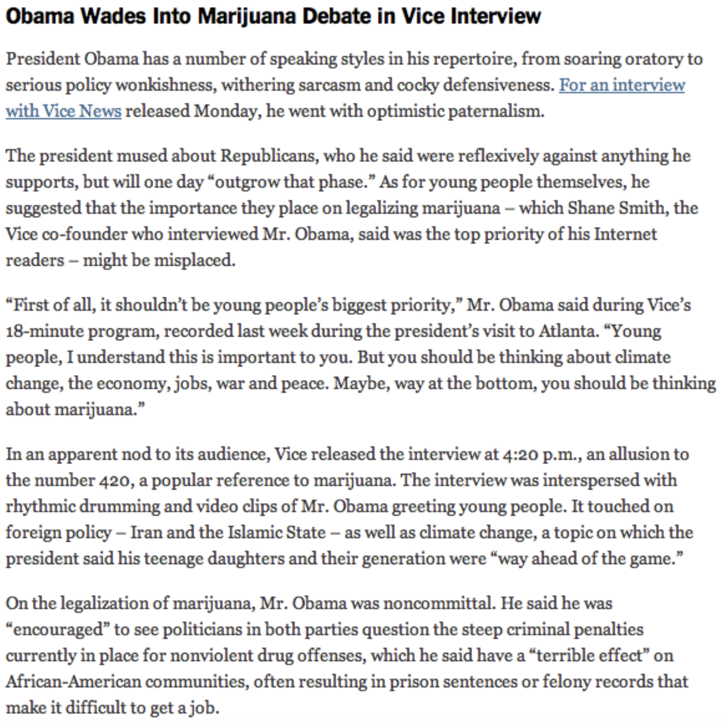


How can we leverage these high quality
word embeddings
to compute
document distances?

Word Mover's Distance



Goal



Obama Wades Into Marijuana Debate in Vice Interview

President Obama has a number of speaking styles in his repertoire, from soaring oratory to serious policy wonkishness, withering sarcasm and cocky defensiveness. For an interview with Vice News released Monday, he went with optimistic paternalism.

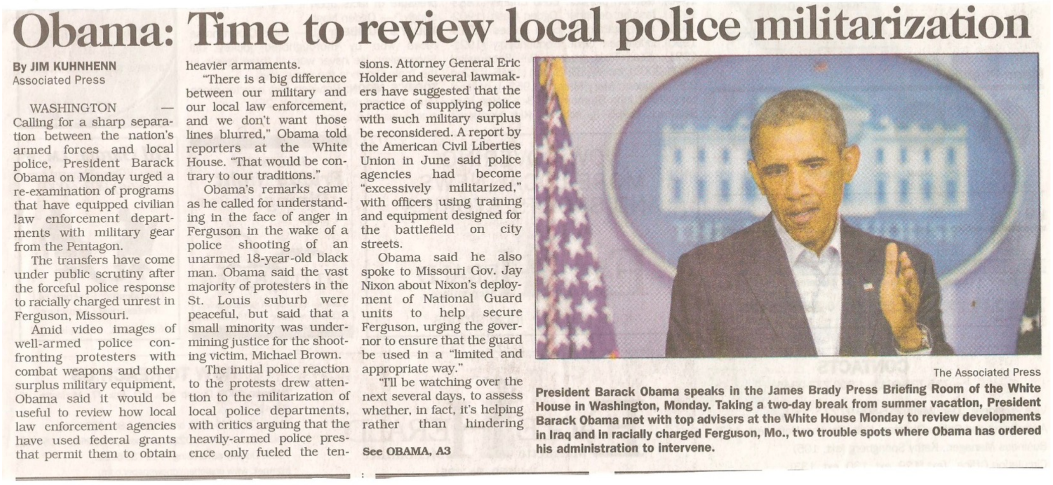
The president mused about Republicans, who he said were reflexively against anything he supports, but will one day "outgrow that phase." As for young people themselves, he suggested that the importance they place on legalizing marijuana – which Shane Smith, the Vice co-founder who interviewed Mr. Obama, said was the top priority of his Internet readers – might be misplaced.

"First of all, it shouldn't be young people's biggest priority," Mr. Obama said during Vice's 18-minute program, recorded last week during the president's visit to Atlanta. "Young people, I understand this is important to you. But you should be thinking about climate change, the economy, jobs, war and peace. Maybe, way at the bottom, you should be thinking about marijuana."

In an apparent nod to its audience, Vice released the interview at 4:20 p.m., an allusion to the number 420, a popular reference to marijuana. The interview was interspersed with rhythmic drumming and video clips of Mr. Obama greeting young people. It touched on foreign policy – Iran and the Islamic State – as well as climate change, a topic on which the president said his teenage daughters and their generation were "way ahead of the game."

On the legalization of marijuana, Mr. Obama was noncommittal. He said he was "encouraged" to see politicians in both parties question the steep criminal penalties currently in place for nonviolent drug offenses, which he said have a "terrible effect" on African-American communities, often resulting in prison sentences or felony records that make it difficult to get a job.

?



Obama: Time to review local police militarization

By JIM KUHNHENN
Associated Press

WASHINGTON — Calling for a sharp separation between the nation's armed forces and local police, President Barack Obama on Monday urged a re-examination of programs that have equipped civilian law enforcement departments with military gear from the Pentagon.

The transfers have come under public scrutiny after the forceful police response to racially charged unrest in Ferguson, Missouri.

Amid video images of well-armed police confronting protesters with combat weapons and other surplus military equipment, Obama said it would be useful to review how local law enforcement agencies have used federal grants that permit them to obtain heavier armaments.

"There is a big difference between our military and our local law enforcement, and we don't want those lines blurred," Obama told reporters at the White House. "That would be contrary to our traditions."

Obama's remarks came as he called for understanding in the face of anger in Ferguson in the wake of a police shooting of an unarmed 18-year-old black man. Obama said the vast majority of protesters in the St. Louis suburb were peaceful, but said that a small minority was undermining justice for the shooting victim, Michael Brown.

The initial police reaction to the protests drew attention to the militarization of local police departments, with critics arguing that the heavily-armed police presence only fueled the tensions. Attorney General Eric Holder and several lawmakers have suggested that the practice of supplying police with such military surplus be reconsidered. A report by the American Civil Liberties Union in June said police agencies had become "excessively militarized," with officers using training and equipment designed for the battlefield on city streets.

Obama said he also spoke to Missouri Gov. Jay Nixon about Nixon's deployment of National Guard units to help secure Ferguson, urging the governor to ensure that the guard be used in a "limited and appropriate way."

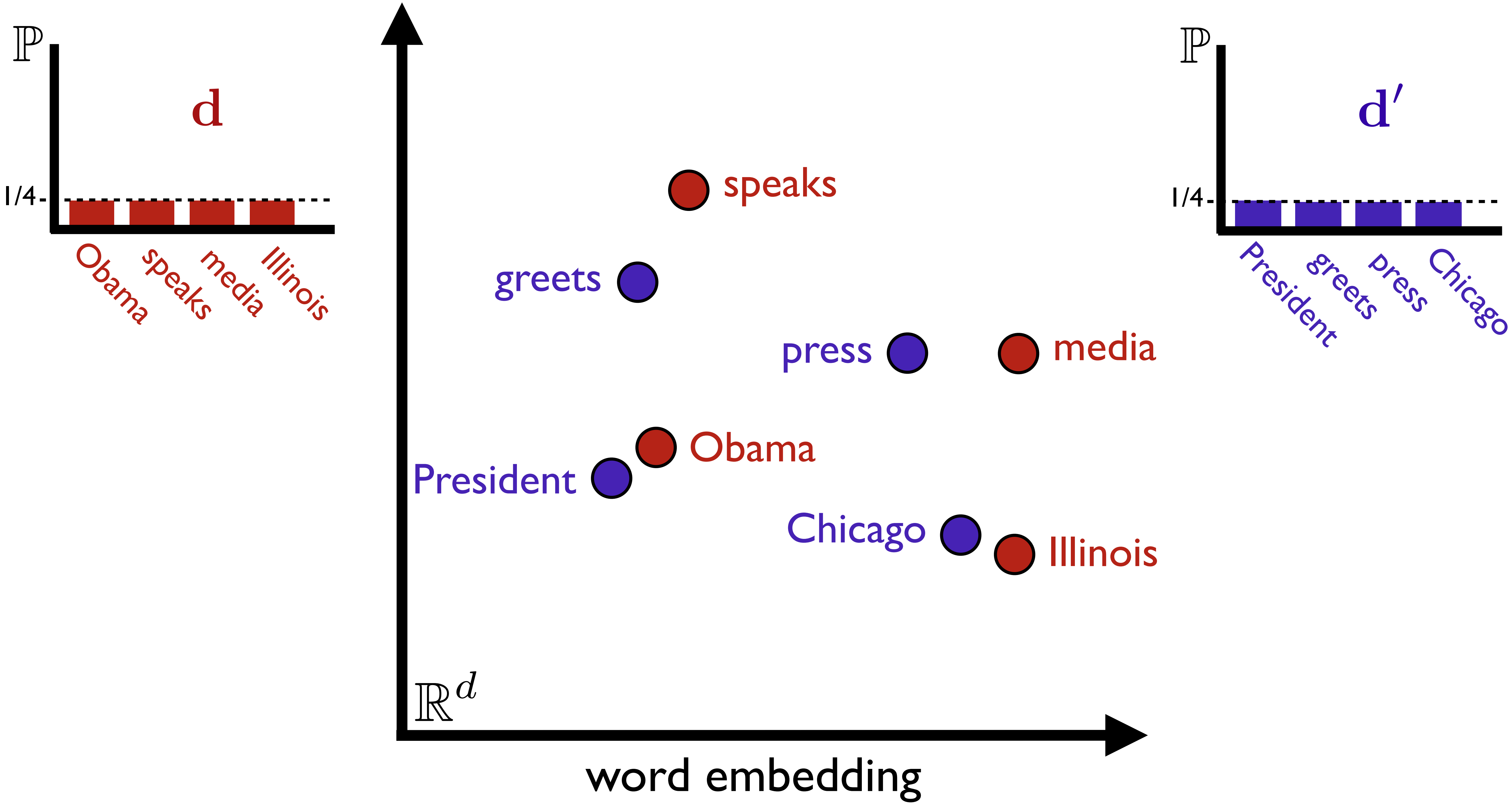
"I'll be watching over the next several days, to assess whether, in fact, it's helping rather than hindering

See OBAMA, A3

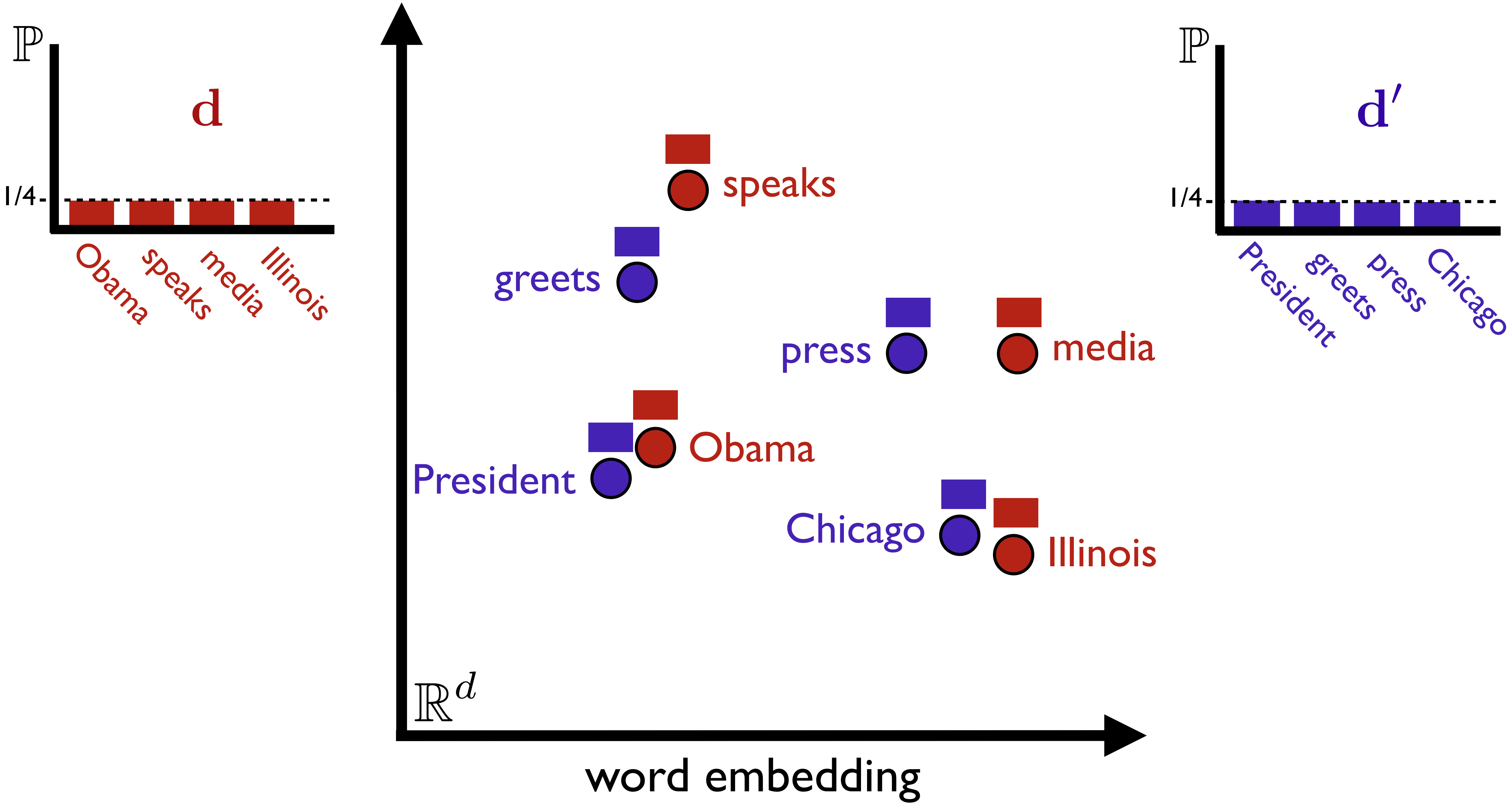
The Associated Press

President Barack Obama speaks in the James Brady Press Briefing Room of the White House in Washington, Monday. Taking a two-day break from summer vacation, President Barack Obama met with top advisers at the White House Monday to review developments in Iraq and in racially charged Ferguson, Mo., two trouble spots where Obama has ordered his administration to intervene.

Word Mover's Distance

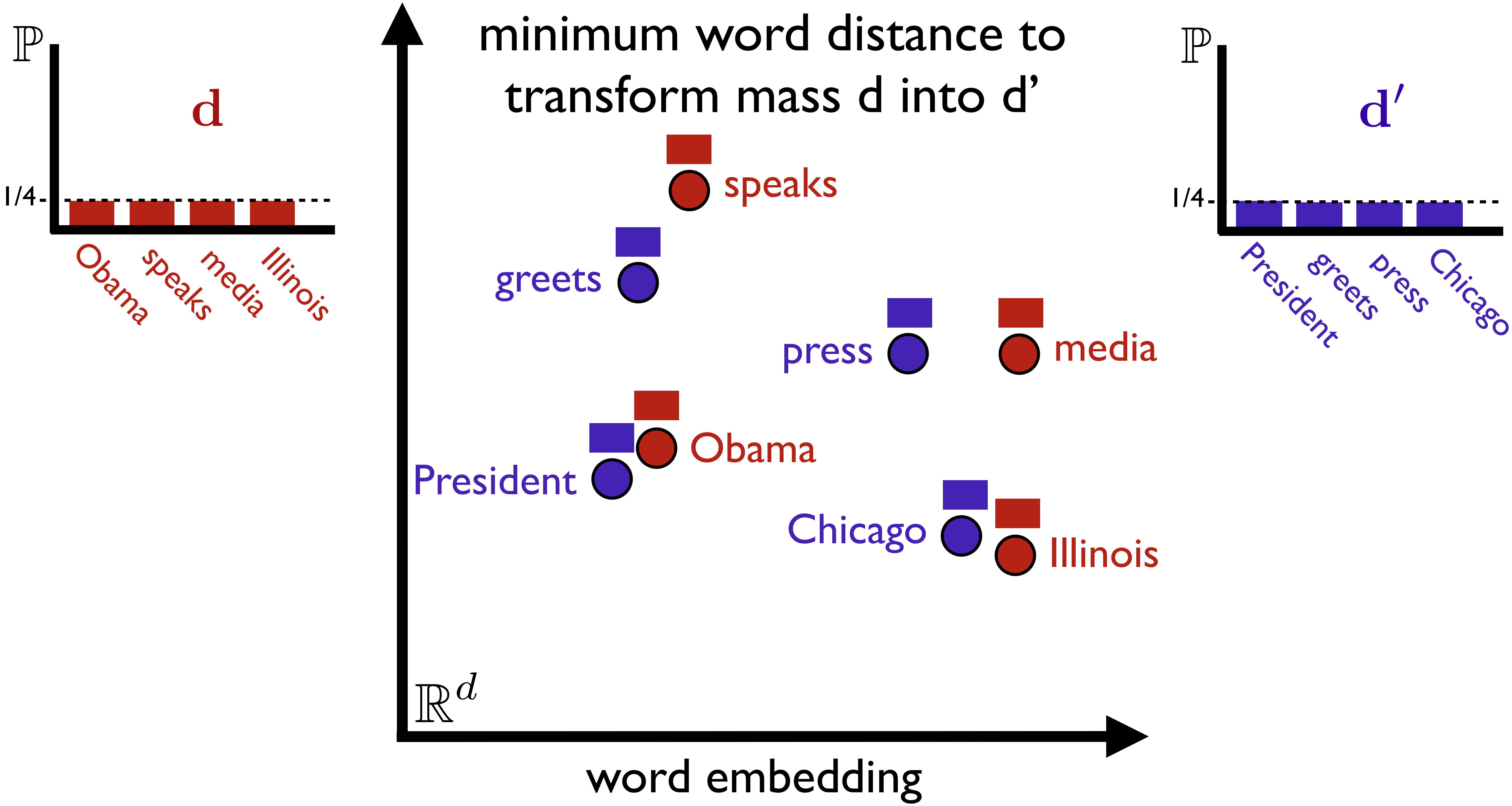


Word Mover's Distance

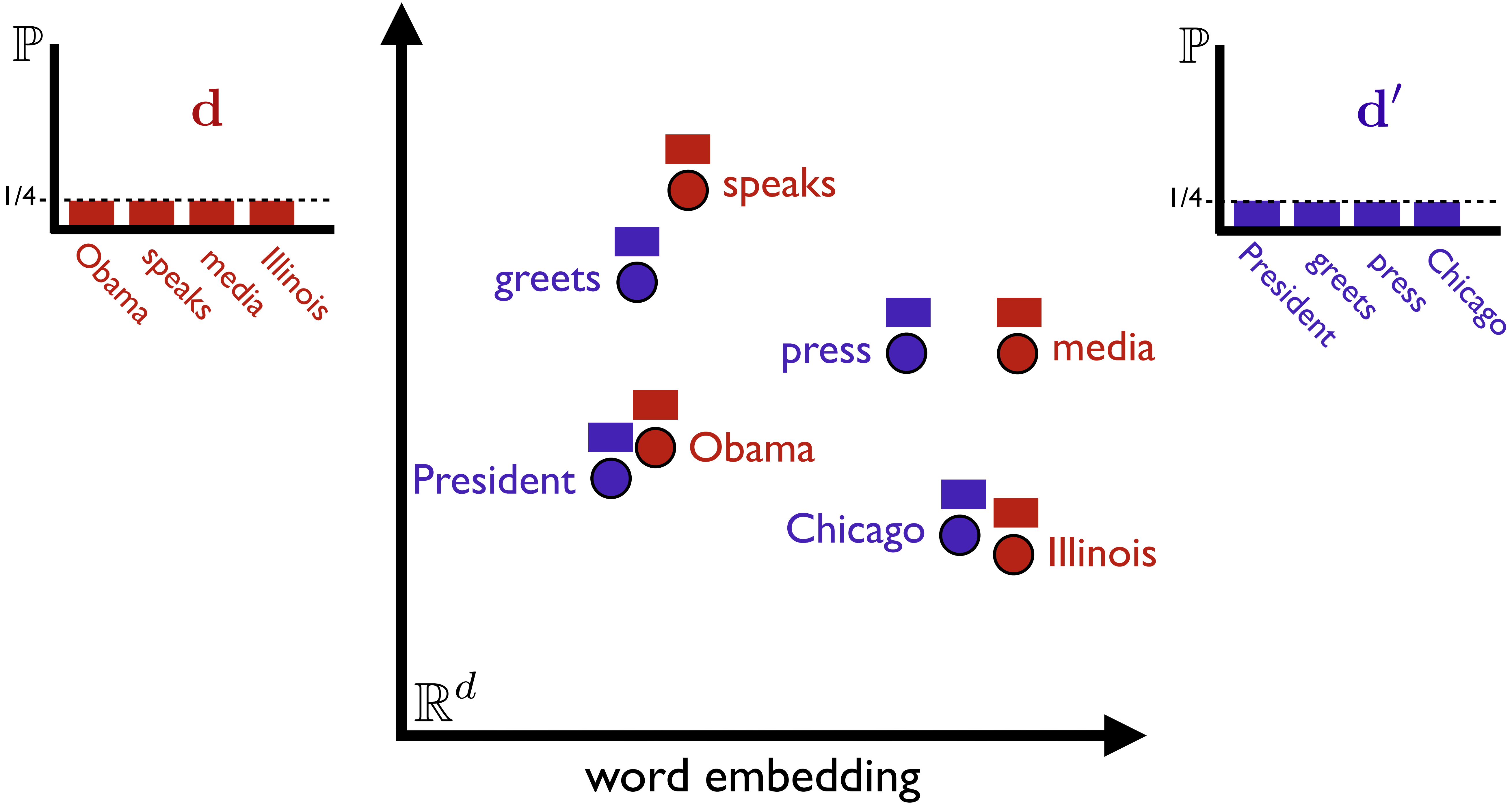


Word Mover's Distance

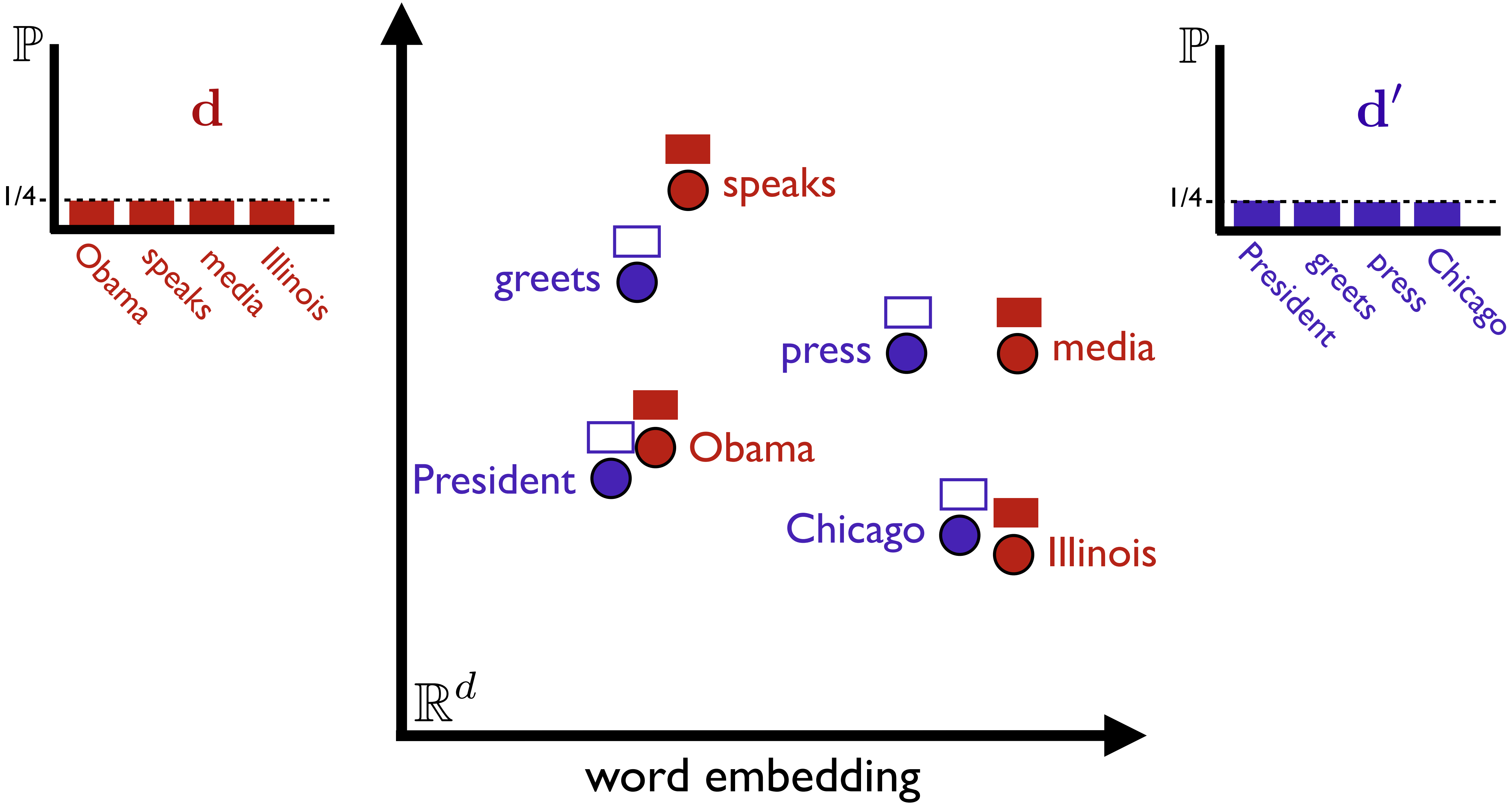
word mover's distance =
minimum word distance to
transform mass d into d'



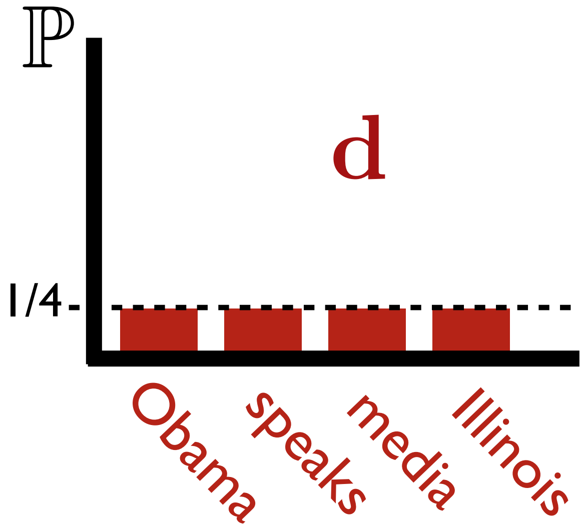
Word Mover's Distance

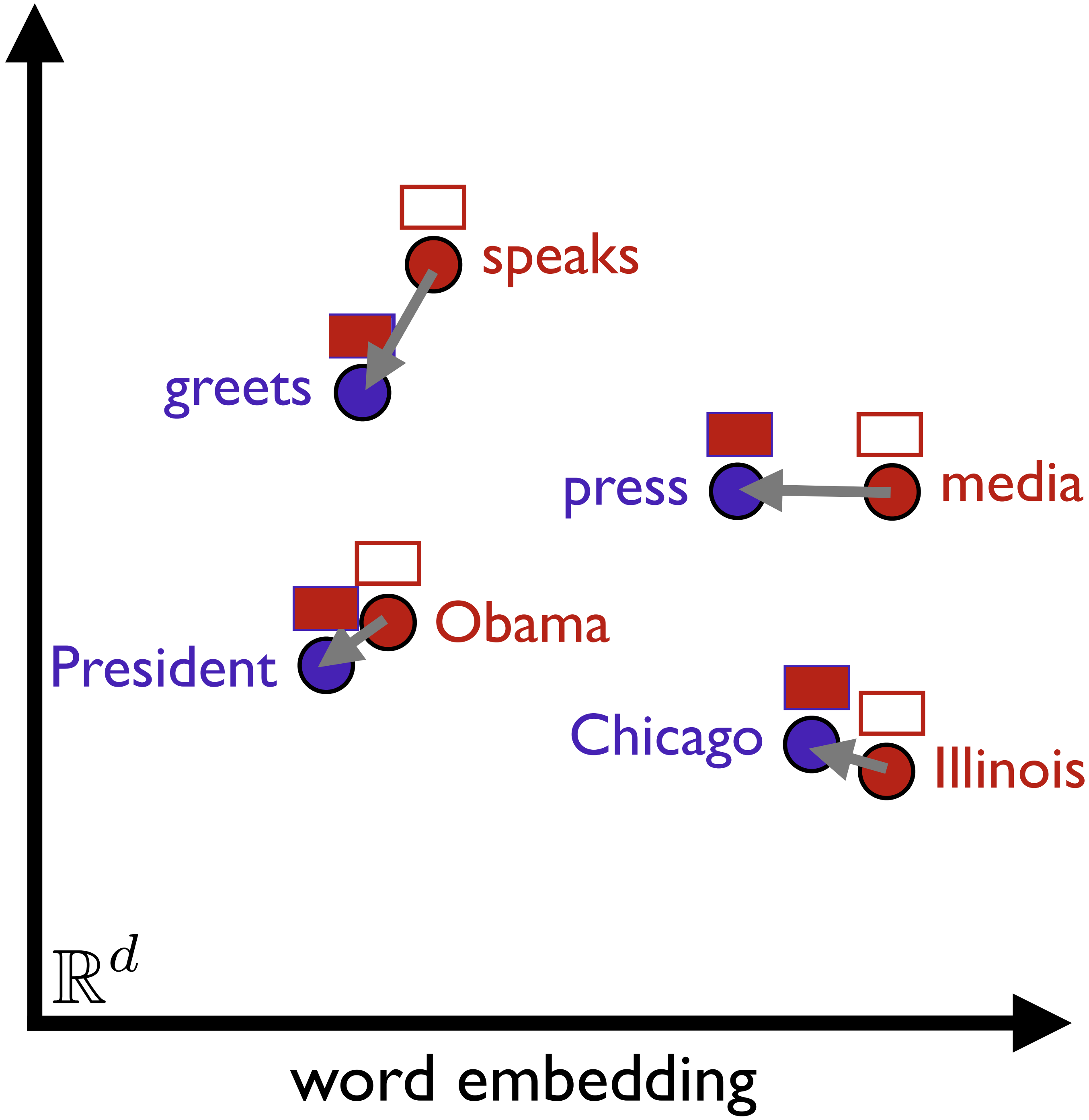


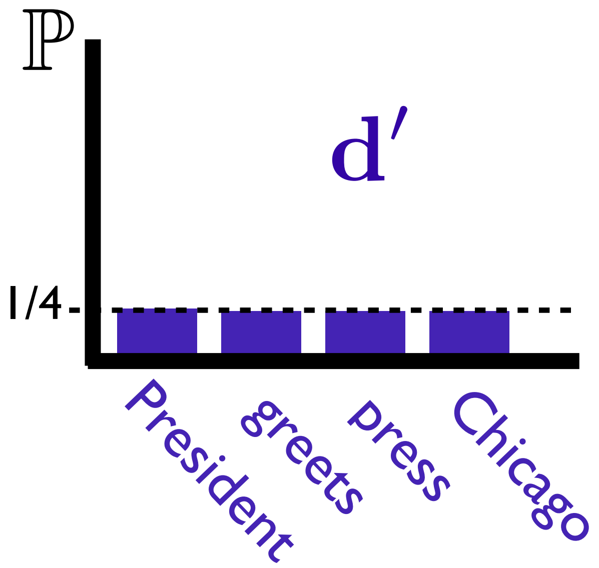
Word Mover's Distance



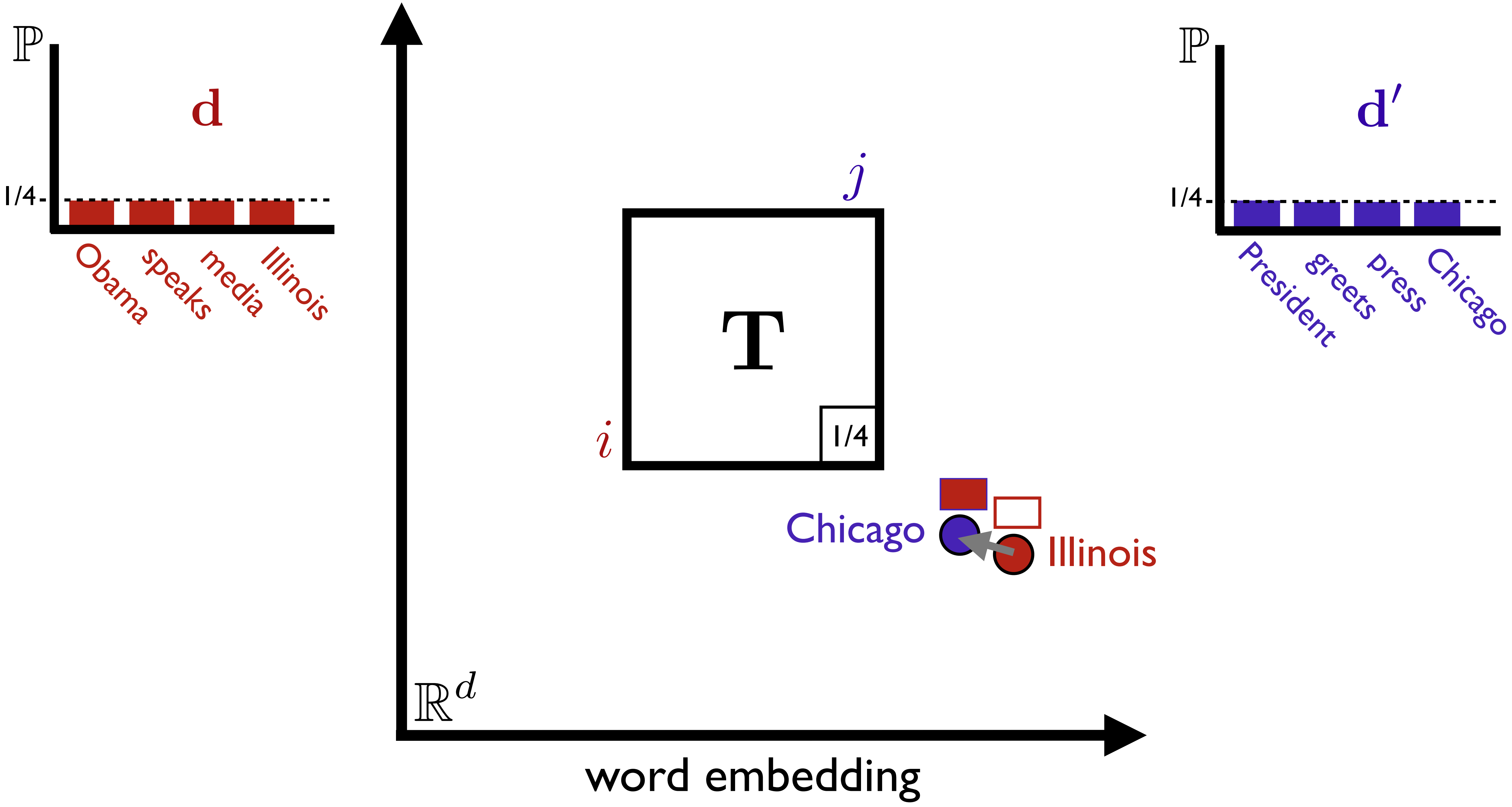
Word Mover's Distance



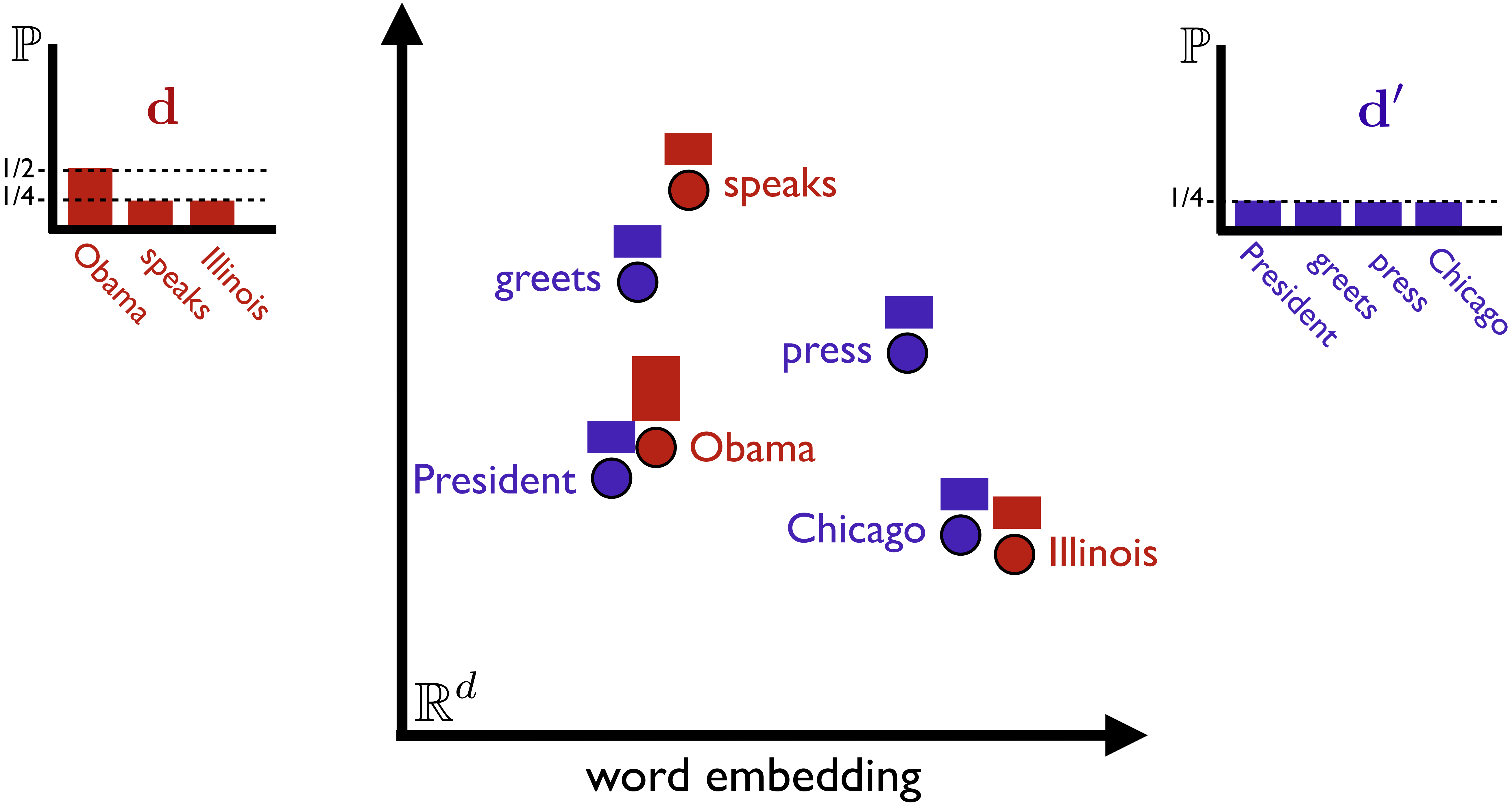




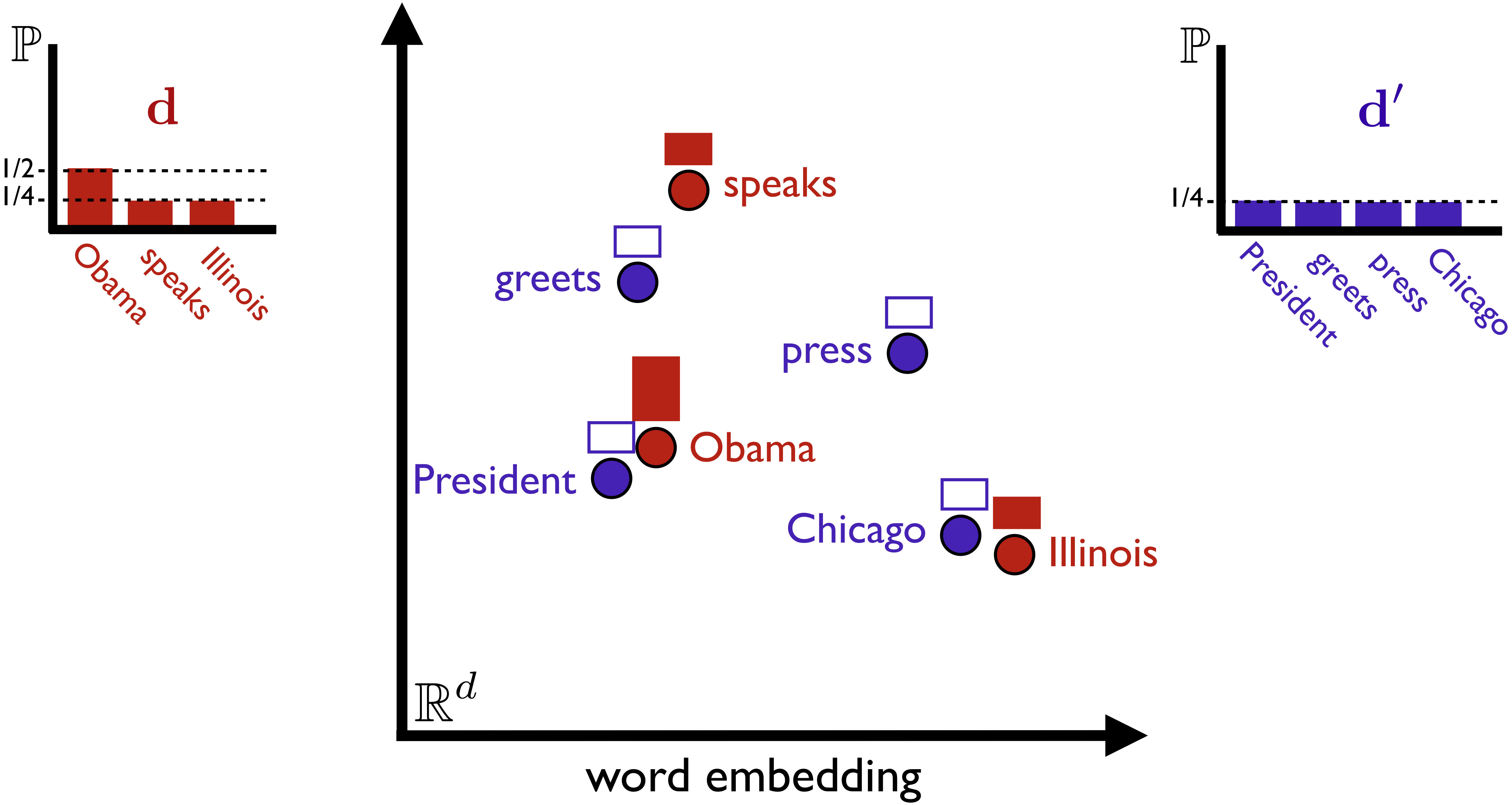
Word Mover's Distance



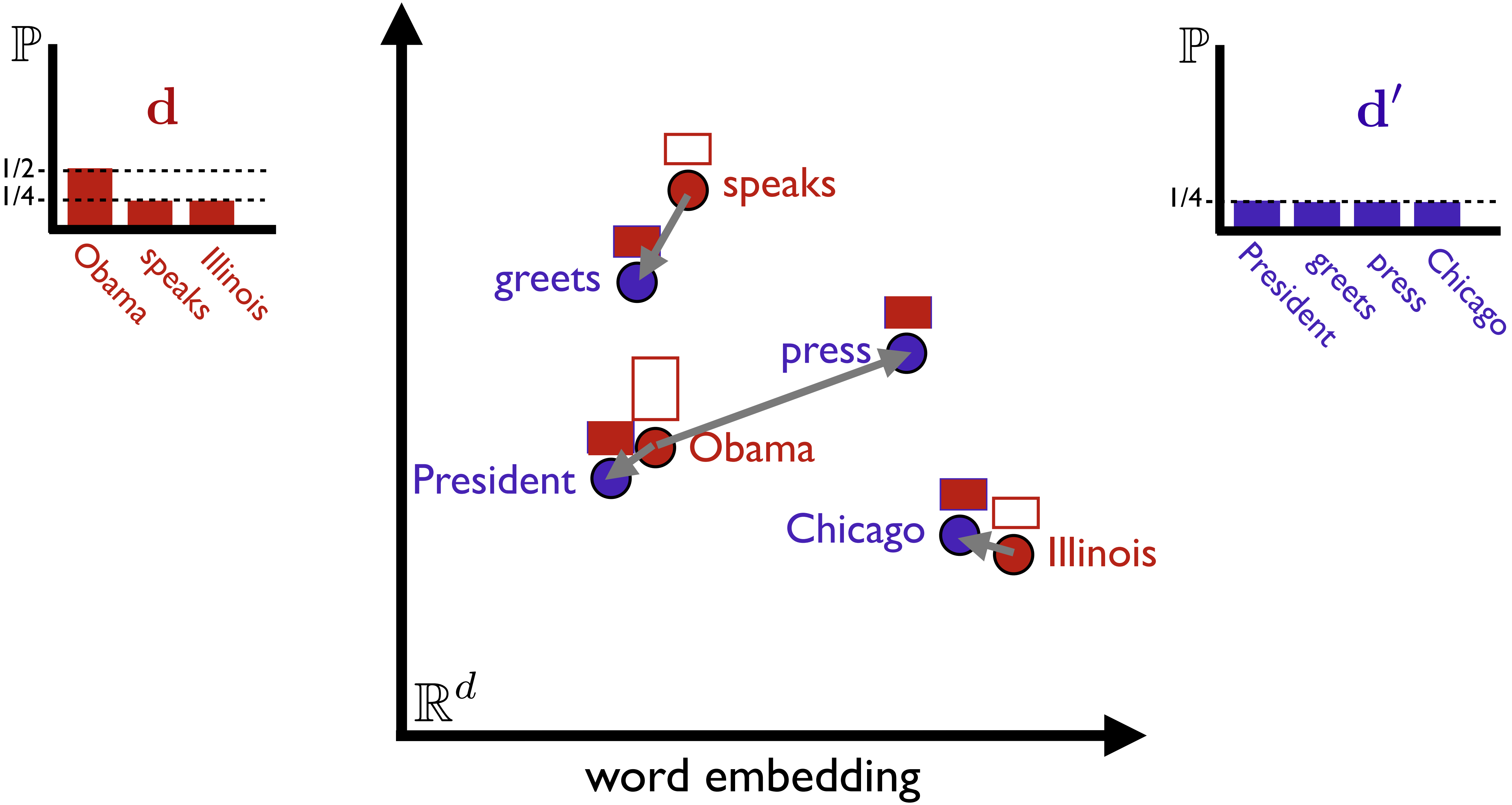
Word Mover’s Distance



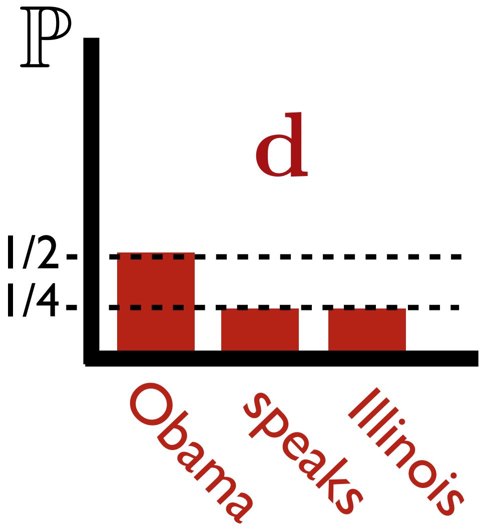
Word Mover's Distance



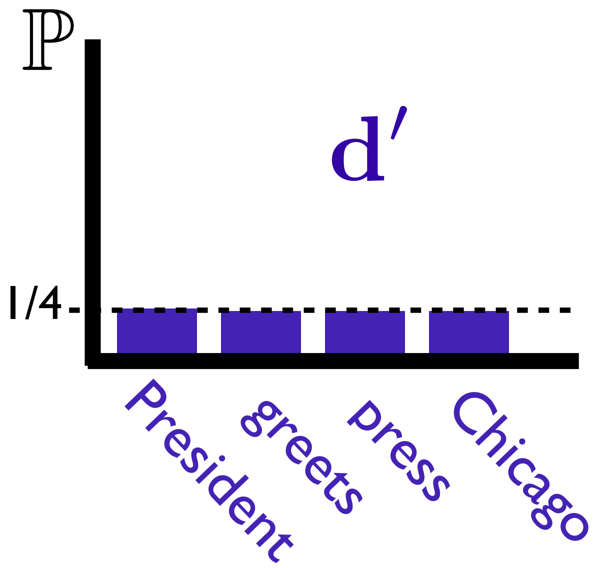
Word Mover's Distance

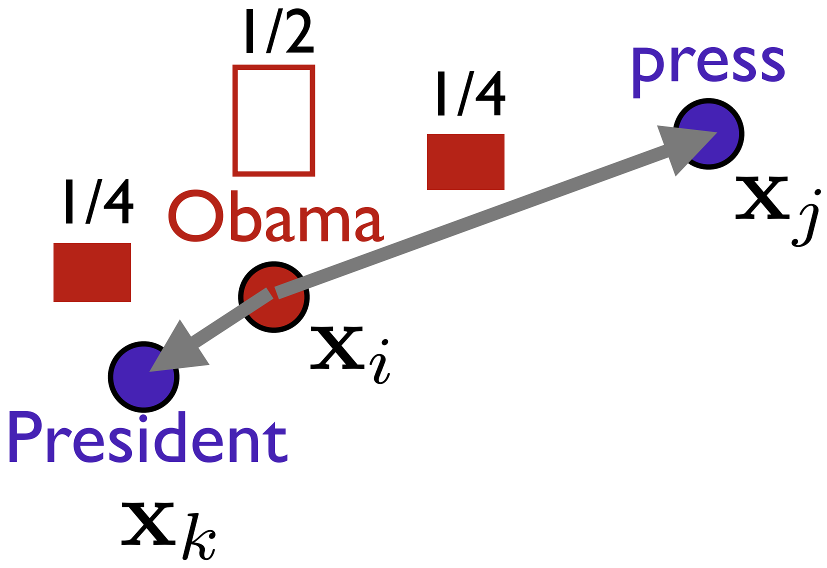


Word Mover's Distance

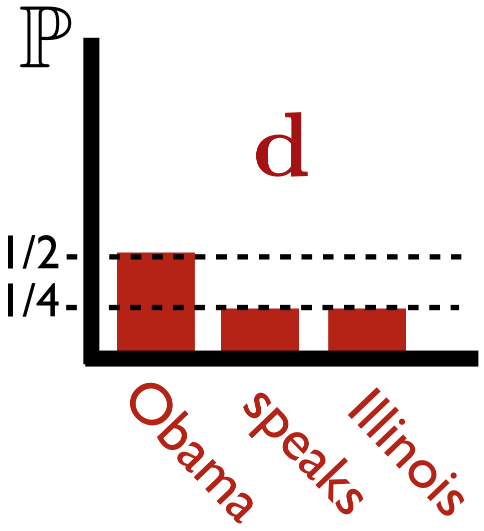


$$\mathrm{WMD}(\mathbf{d},\mathbf{d}') \triangleq \min_{\mathbf{T}\geq 0} \sum_{i,j=1}^{n} \mathbf{T}_{ij} \|\mathbf{x}_i - \mathbf{x}_j\|_2$$





Word Mover's Distance

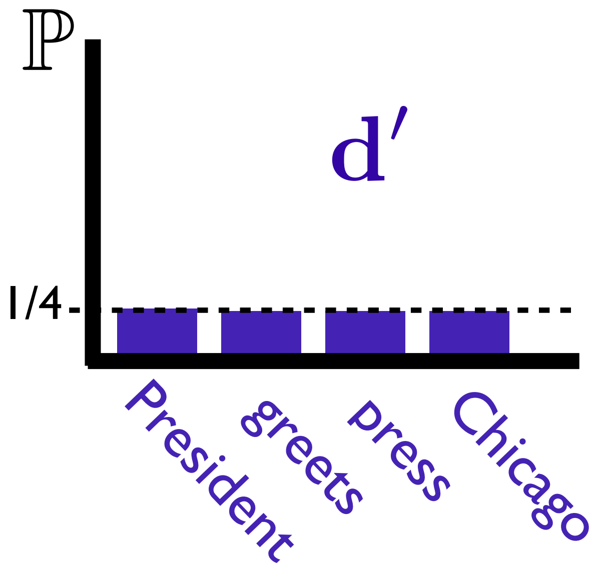


$$\mathrm{WMD}(\mathbf{d},\mathbf{d}') \triangleq$$

$$\min_{\mathbf{T}\geq 0} \sum_{i,j=1}^{n} \mathbf{T}_{ij} \|\mathbf{x}_i - \mathbf{x}_j\|_2$$

$$s.t. \sum_{j=1}^{n} \mathbf{T}_{ij} = \mathbf{d}_i \quad \forall i$$

$$\sum_{i=1}^{n} \mathbf{T}_{ij} = \mathbf{d}'_j \quad \forall j$$



Remarks

$$\mathrm{WMD}(\mathbf{d},\mathbf{d}') \triangleq$$

$$\begin{aligned}
\min_{\mathbf{T}\geq 0} & \sum_{i,j=1}^{n} \mathbf{T}_{ij}\|\mathbf{x}_i - \mathbf{x}_j\|_2 \\
s.t. & \sum_{j=1}^{n} \mathbf{T}_{ij} = \mathbf{d}_i \quad \forall i \\
& \sum_{i=1}^{n} \mathbf{T}_{ij} = \mathbf{d}'_j \quad \forall j
\end{aligned}$$

in CV this is the Earth Mover's Distance (EMD) [Rubner et al., 1998]

an old optimal transport problem [Monge, 1781]

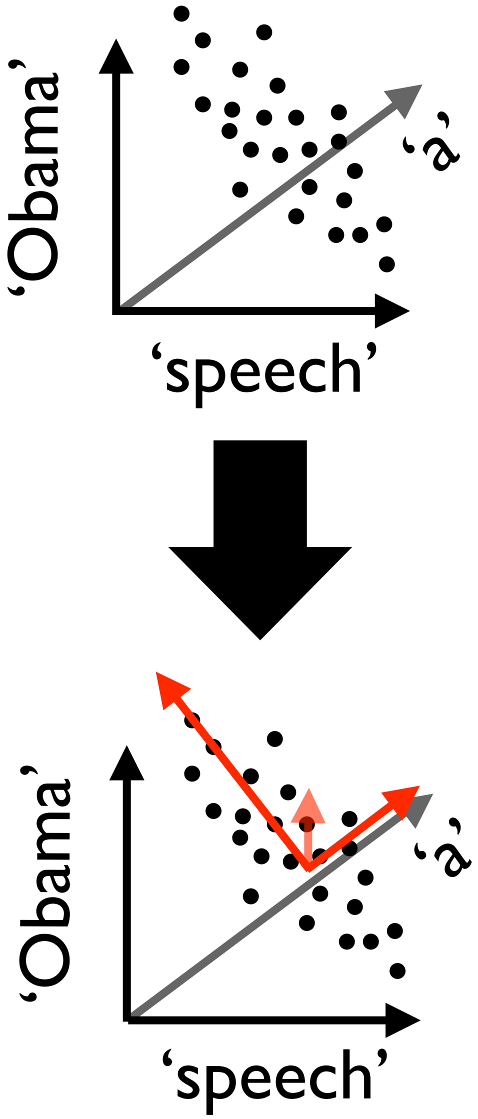
How well does WMD perform on **document classification** via k-nearest neighbors (k-NN)?

Classic Approaches

bag-of-words		TF-IDF [Salton & Buckley, 1988]
3	'campaign'	0.01
0		0
0		0
1	'Obama'	0.2
0		0
2	'speech'	0.04
0		0
0		0
⋮		⋮
0		0
0		0
1	'Washington'	0.13

LSI

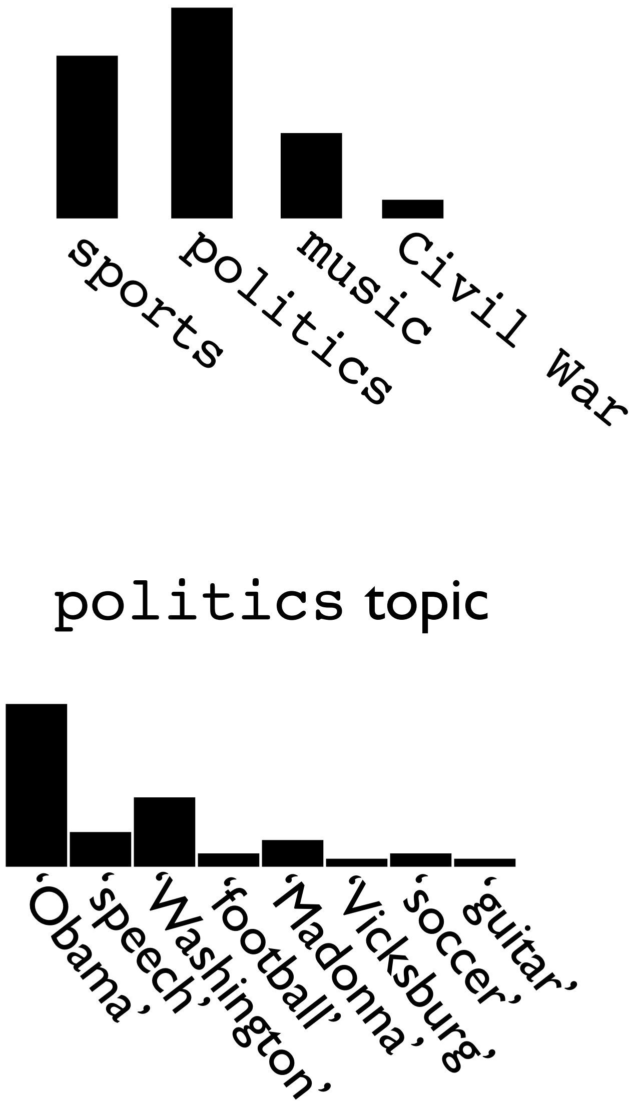
[Deerwester et al., 1990]



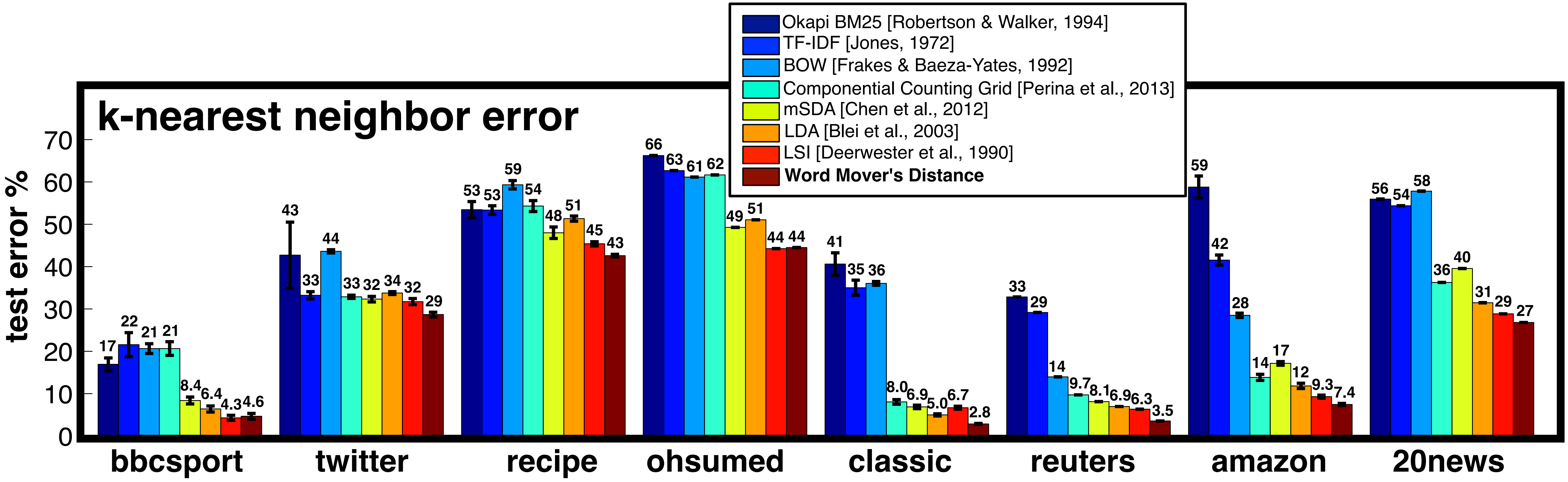
LDA

[Blei et al., 2003]

topic distributions



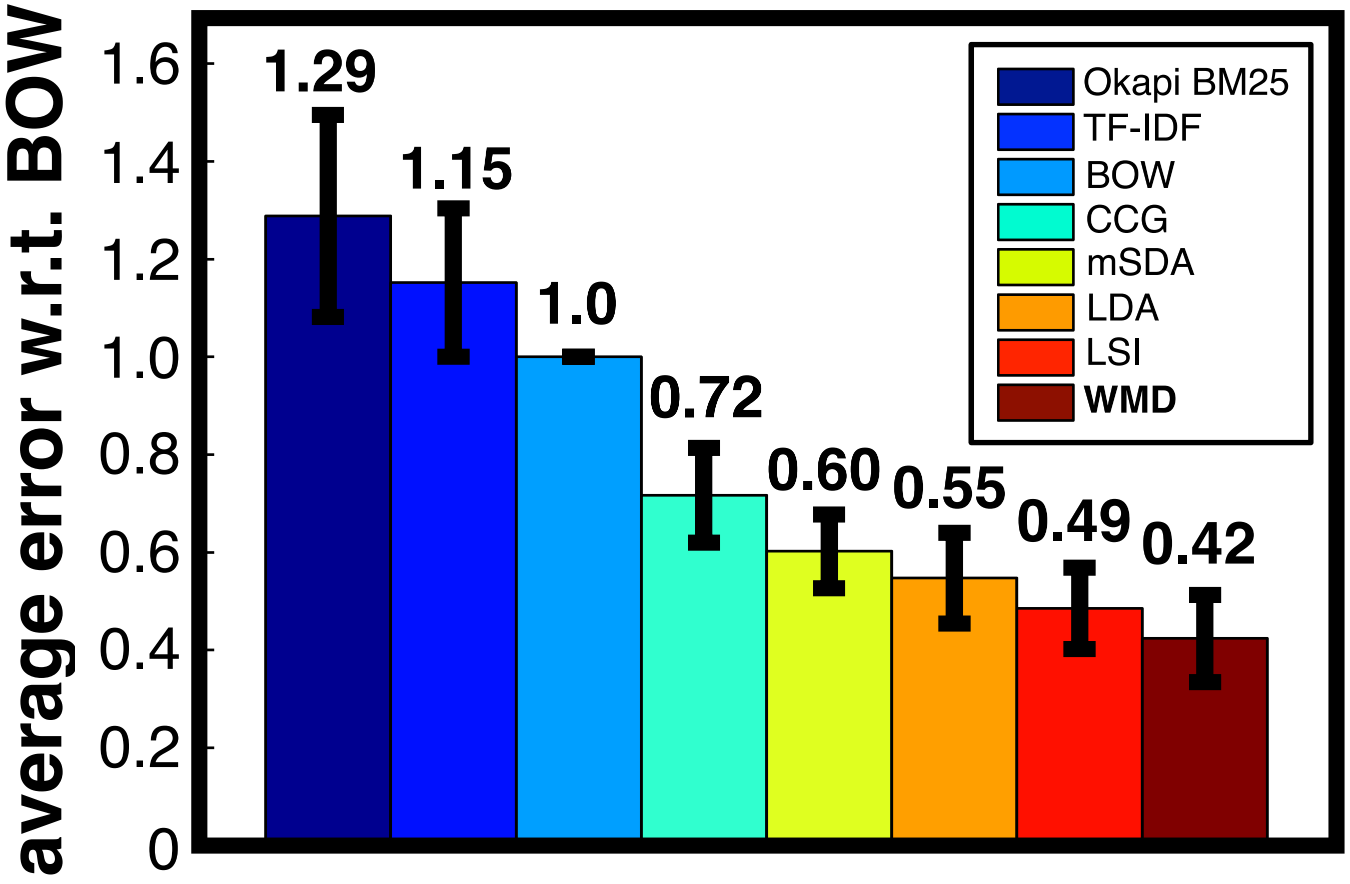
Results: k-NN



	bbcsport	twitter	recipe	ohsumed	classic	reuters	amazon	20news
train inputs:	517	2176	3059	3999	4965	5485	5600	11293
BOW dim:	13243	6344	5708	31789	24277	22425	42063	29671

All hyper-parameters set with **bayesopt.m**
[Gardner et al. 2014]

Results: k-NN


average error w.r.t. BOW
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0
1.29
1.15
1.0
0.72
0.60
0.55
0.49
0.42
Okapi BM25
TF-IDF
BOW
CCG
mSDA
LDA
LSI
WMD

Computational Complexity

$$\mathrm{WMD}(\mathbf{d},\mathbf{d}') \triangleq$$

$$\min_{\mathbf{T}\geq 0} \sum_{i,j=1}^{n} \mathbf{T}_{ij}\|\mathbf{x}_i - \mathbf{x}_j\|_2$$

$$s.t. \sum_{j=1}^{n} \mathbf{T}_{ij} = \mathbf{d}_i \quad \forall i$$

$$\sum_{i=1}^{n} \mathbf{T}_{ij} = \mathbf{d}'_j \quad \forall j$$

LP with 2n constraints

$$\mathrm{O}(n^3 \log n)$$

[Pele & Werman, 2009]

Computational Complexity

$$\mathrm{WMD}(\mathbf{d}, \mathbf{d}') \triangleq$$

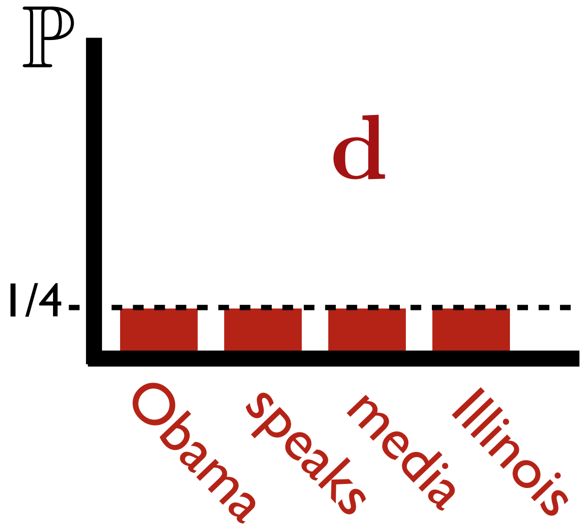
$$\begin{aligned} \min_{\mathbf{T} \geq 0} & \sum_{i,j=1}^{n} \mathbf{T}_{ij} \|\mathbf{x}_i - \mathbf{x}_j\|_2 \\ s.t. & \sum_{j=1}^{n} \mathbf{T}_{ij} = \mathbf{d}_i \quad \forall i \\ & \sum_{i=1}^{n} \mathbf{T}_{ij} = \mathbf{d}'_j \quad \forall j \end{aligned}$$

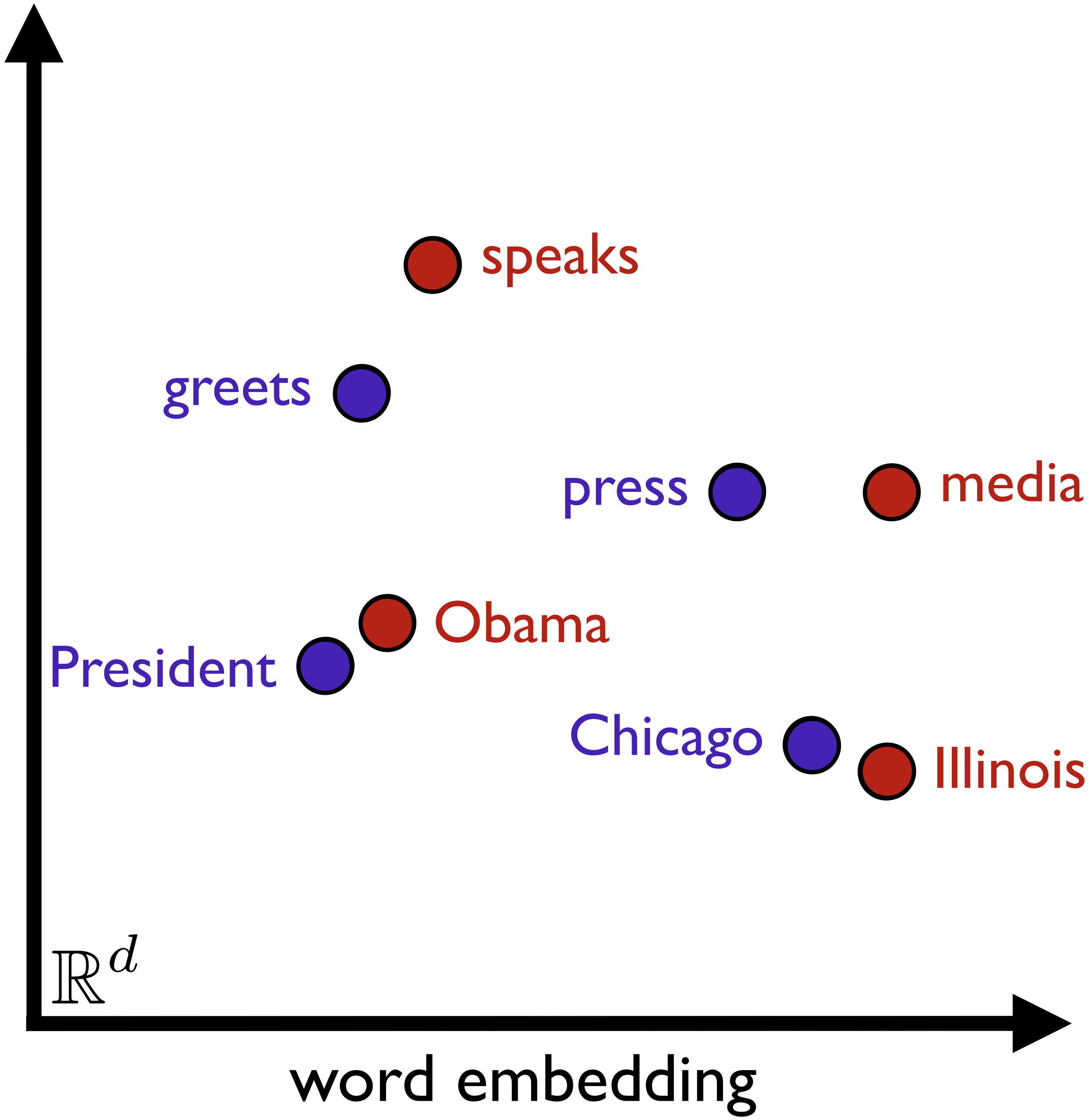
approximations:

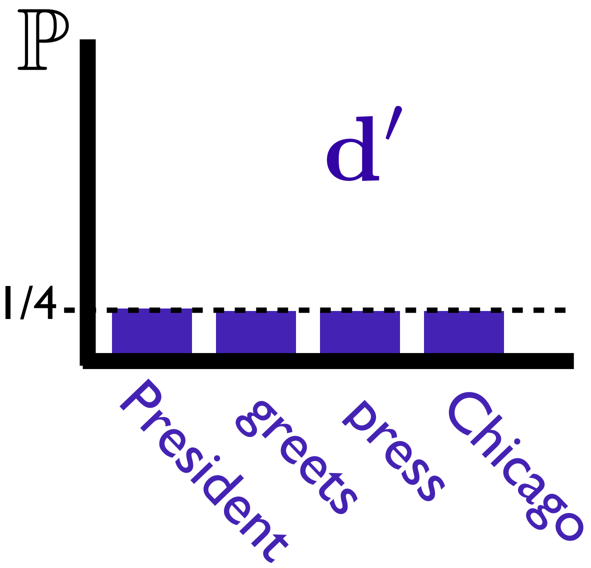
[Rubner et al., 1998]; [Levina & Bickel, 2001]; [Grauman & Darrell, 2004]; [Shirdhonkar & Jacobs, 2008]

Approximation 1

[Rubner et al., 1998]

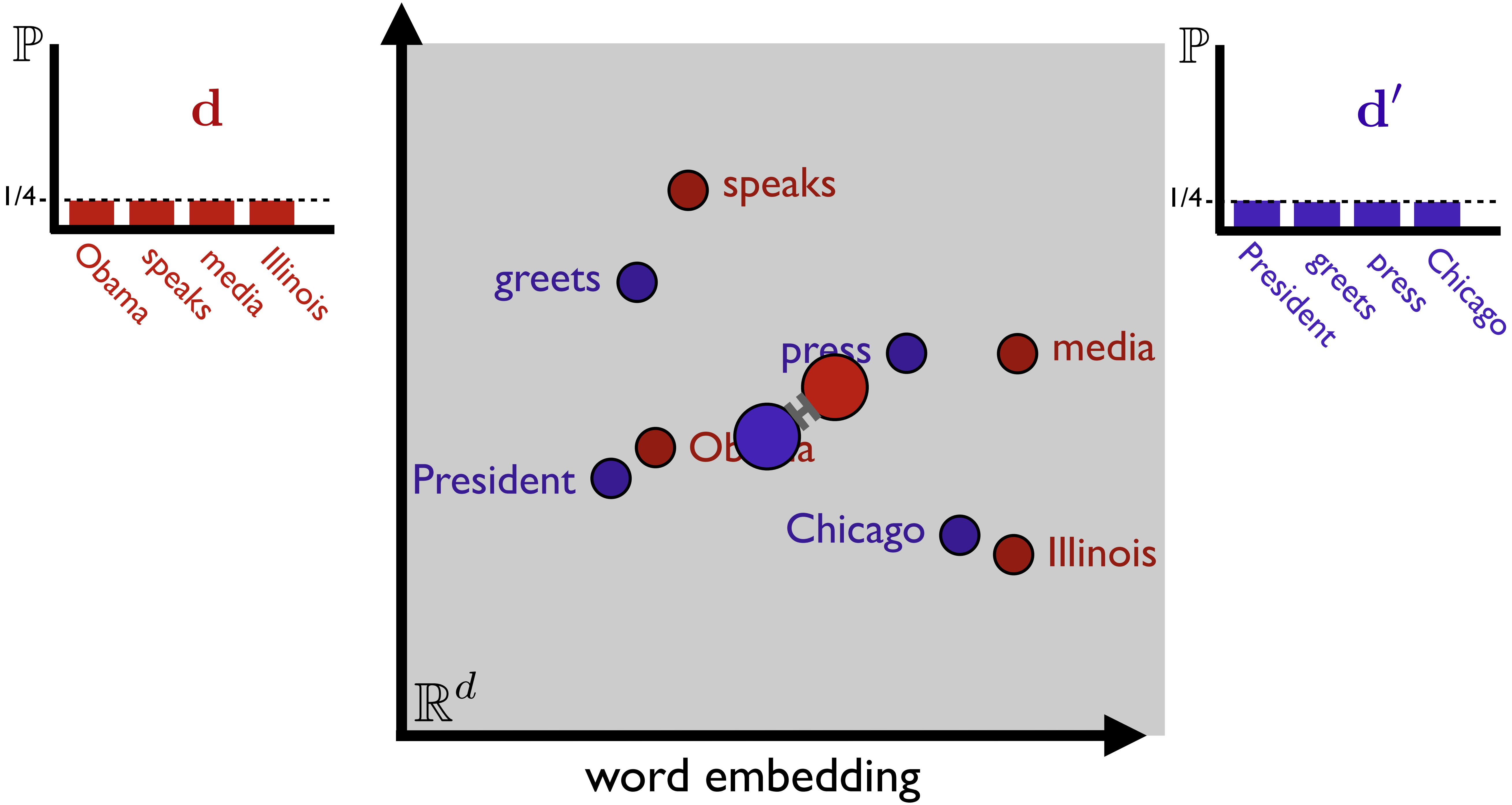






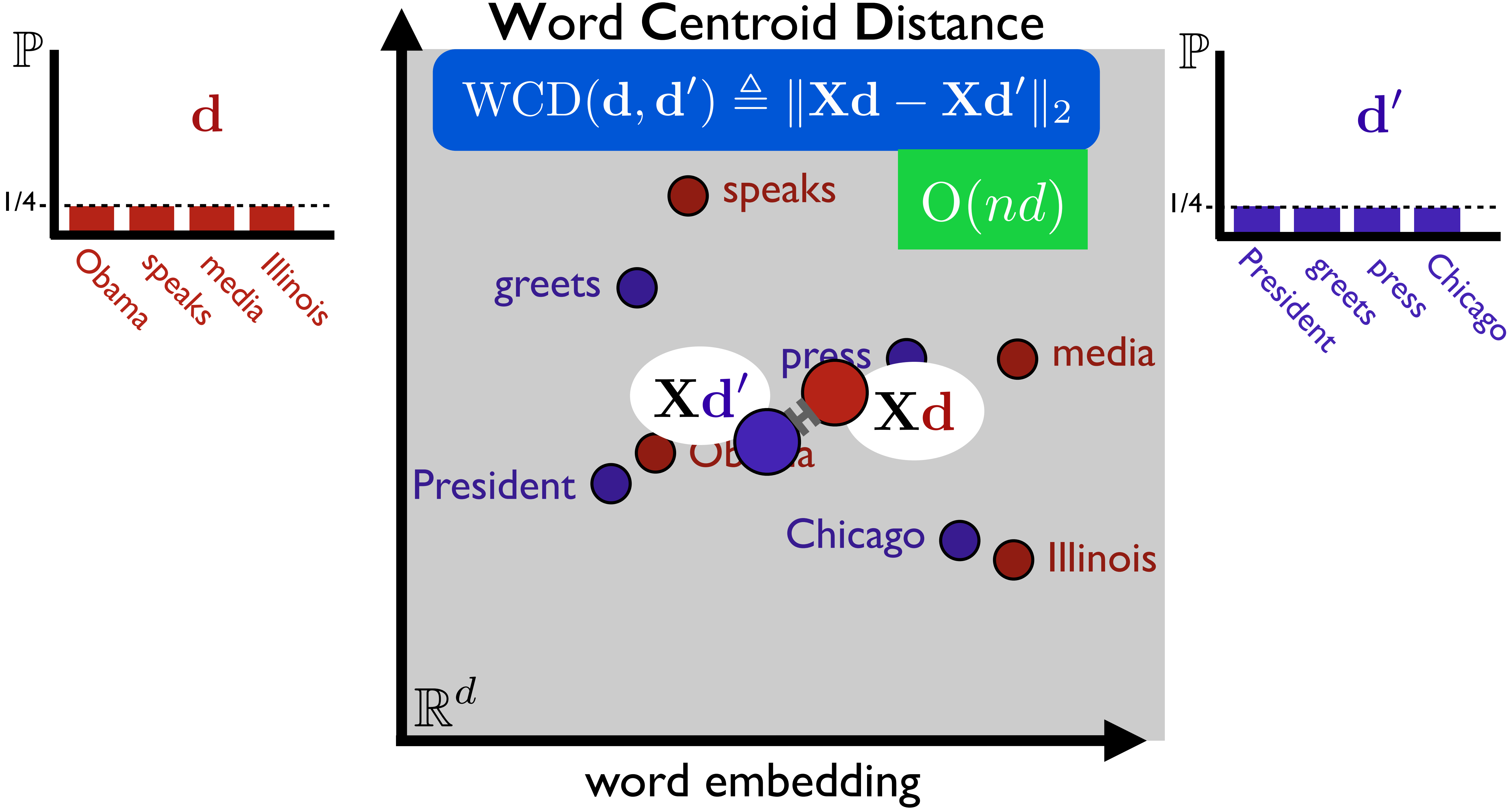
Approximation 1

[Rubner et al., 1998]



Approximation 1

[Rubner et al., 1998]



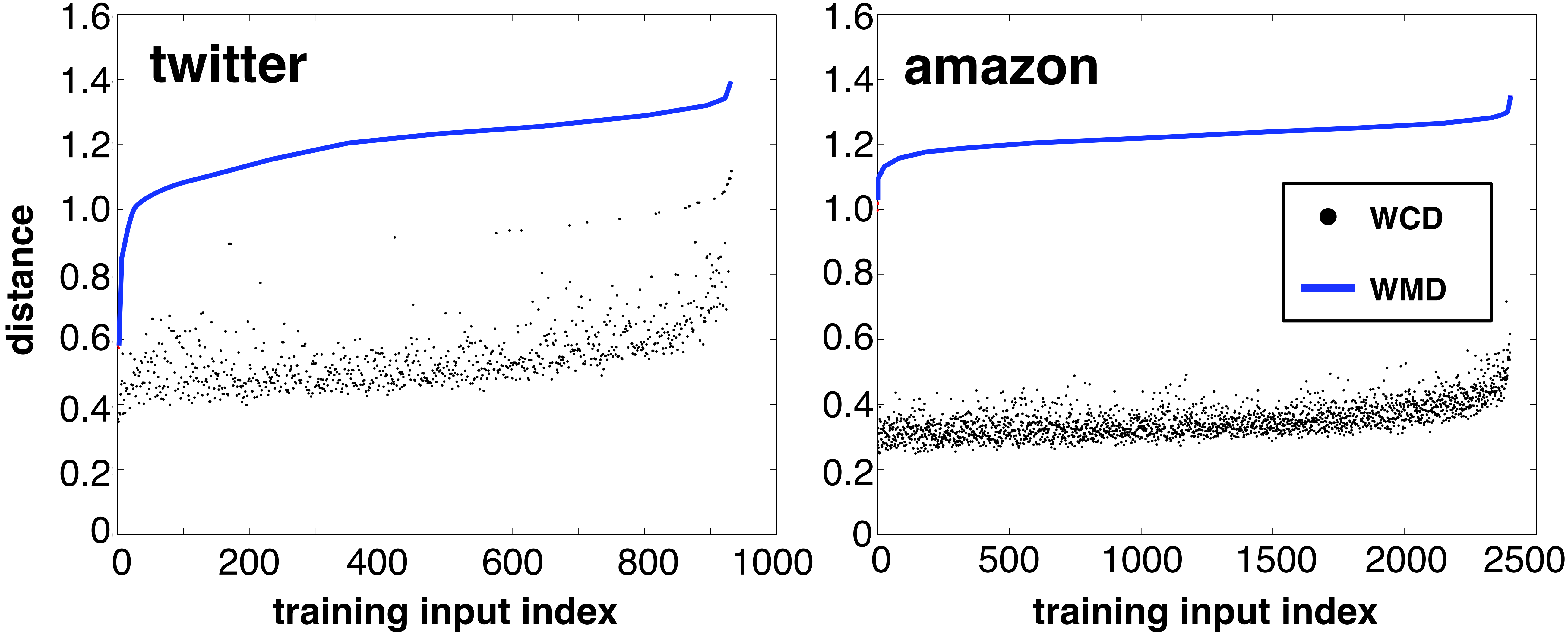
Word Centroid Distance

$$\mathrm{WCD}(\mathbf{d},\mathbf{d}') \triangleq \|\mathbf{Xd}-\mathbf{Xd}'\|_2$$

$O(nd)$

Faster Approximations

for a random test input...



Approximation 2

$$\min_{\mathbf{T}\geq 0} \sum_{i,j=1}^{n} \mathbf{T}_{ij} \|\mathbf{x}_i - \mathbf{x}_j\|_2$$

$$s.t. \sum_{j=1}^{n} \mathbf{T}_{ij} = \mathbf{d}_i \quad \forall i$$

$$\sum_{i=1}^{n} \mathbf{T}_{ij} = \mathbf{d}'_j \quad \forall j$$

Approximation 2

$$\min_{\mathbf{T}\geq 0} \sum_{i,j=1}^{n} \mathbf{T}_{ij} \|\mathbf{x}_i - \mathbf{x}_j\|_2$$

$$s.t. \sum_{j=1}^{n} \mathbf{T}_{ij} = \mathbf{d}_i \quad \forall i$$

$$\xcancel{\sum_{j=1}^{n} \mathbf{T}_{ij} = \mathbf{d}'_j \quad \forall j}$$

$$D_1$$

Approximation 2

$$\min_{\mathbf{T}\geq 0} \sum_{i,j=1}^{n} \mathbf{T}_{ij} \|\mathbf{x}_i - \mathbf{x}_j\|_2$$

$$s.t. \sum_{j=1}^{n} \mathbf{T}_{ij} = \mathbf{d}_i \quad \forall i$$

$$\xcancel{\sum_{i=1}^{n} \mathbf{T}_{ij} = \mathbf{d}'_j \quad \forall j}$$

$$D_1$$

just a nearest-neighbor search!

Approximation 2

$$
\begin{aligned}
\min_{\mathbf{T}\geq 0} \; & \sum_{i,j=1}^{n} \mathbf{T}_{ij} \|\mathbf{x}_i - \mathbf{x}_j\|_2 \\
s.t. \; & \xcancel{\sum_{j=1}^{n} \mathbf{T}_{ij} = \mathbf{d}_i \quad \forall i} \\
& \sum_{i=1}^{n} \mathbf{T}_{ij} = \mathbf{d}'_j \quad \forall j
\end{aligned}
$$

$$D_2$$

just a nearest-neighbor search!

Approximation 2

$$\min_{\mathbf{T}\geq 0} \sum_{i,j=1}^{n} \mathbf{T}_{ij}\|\mathbf{x}_i - \mathbf{x}_j\|_2$$

$$s.t. \quad \sum_{j=1}^{n} \mathbf{T}_{ij} = \mathbf{d}_i \quad \forall i$$

$$\xcancel{\sum_{i=1}^{n} \mathbf{T}_{ij} = \mathbf{d}'_j \quad \forall j}$$

$$D_1$$

$$\min_{\mathbf{T}\geq 0} \sum_{i,j=1}^{n} \mathbf{T}_{ij}\|\mathbf{x}_i - \mathbf{x}_j\|_2$$

$$s.t. \quad \xcancel{\sum_{j=1}^{n} \mathbf{T}_{ij} = \mathbf{d}_i \quad \forall i}$$

$$\sum_{i=1}^{n} \mathbf{T}_{ij} = \mathbf{d}'_j \quad \forall j$$

$$D_2$$

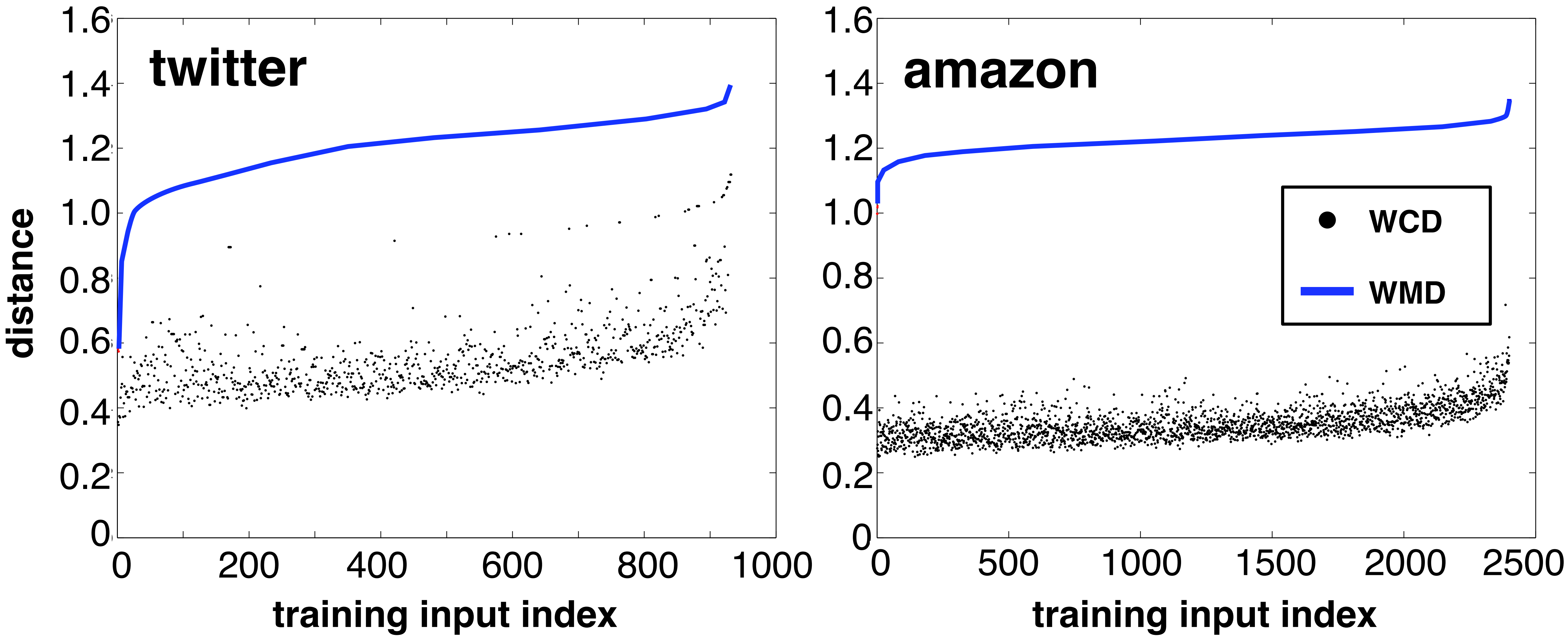
Relaxed Word Mover's Distance

$$\mathrm{RWMD}(\mathbf{d}, \mathbf{d}') \triangleq \max(D_1, D_2)$$

$\mathrm{O}(n^2 d)$

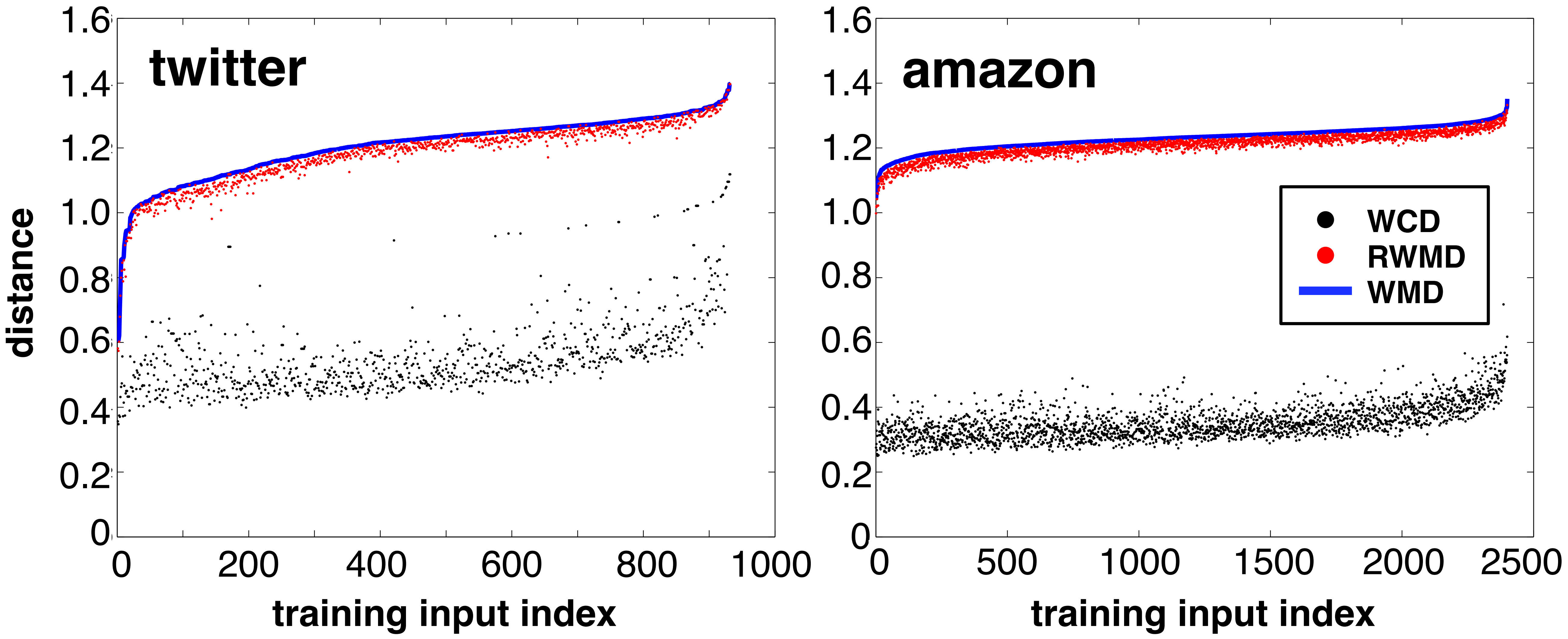
Faster Approximations

for a random test input...

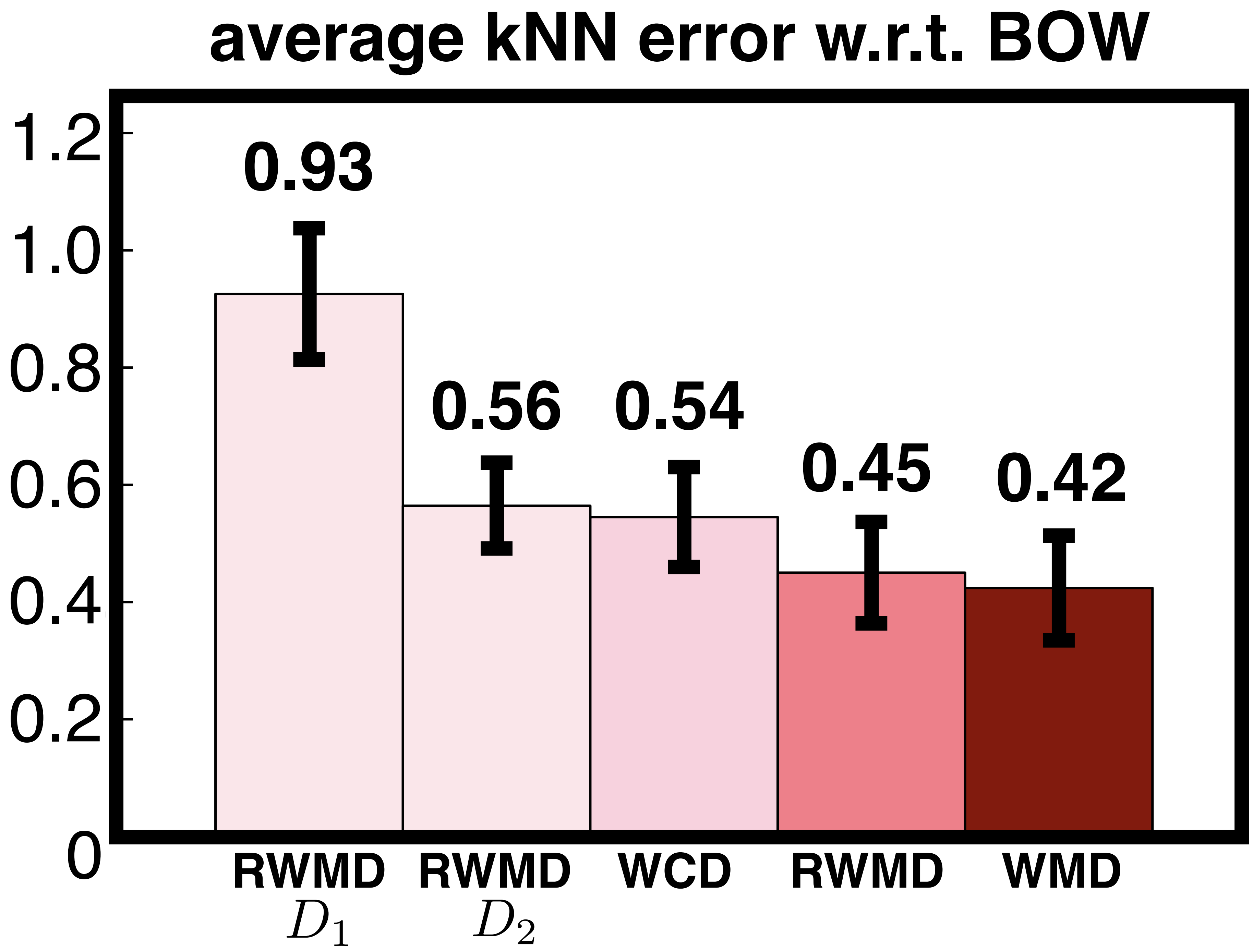


Faster Approximations

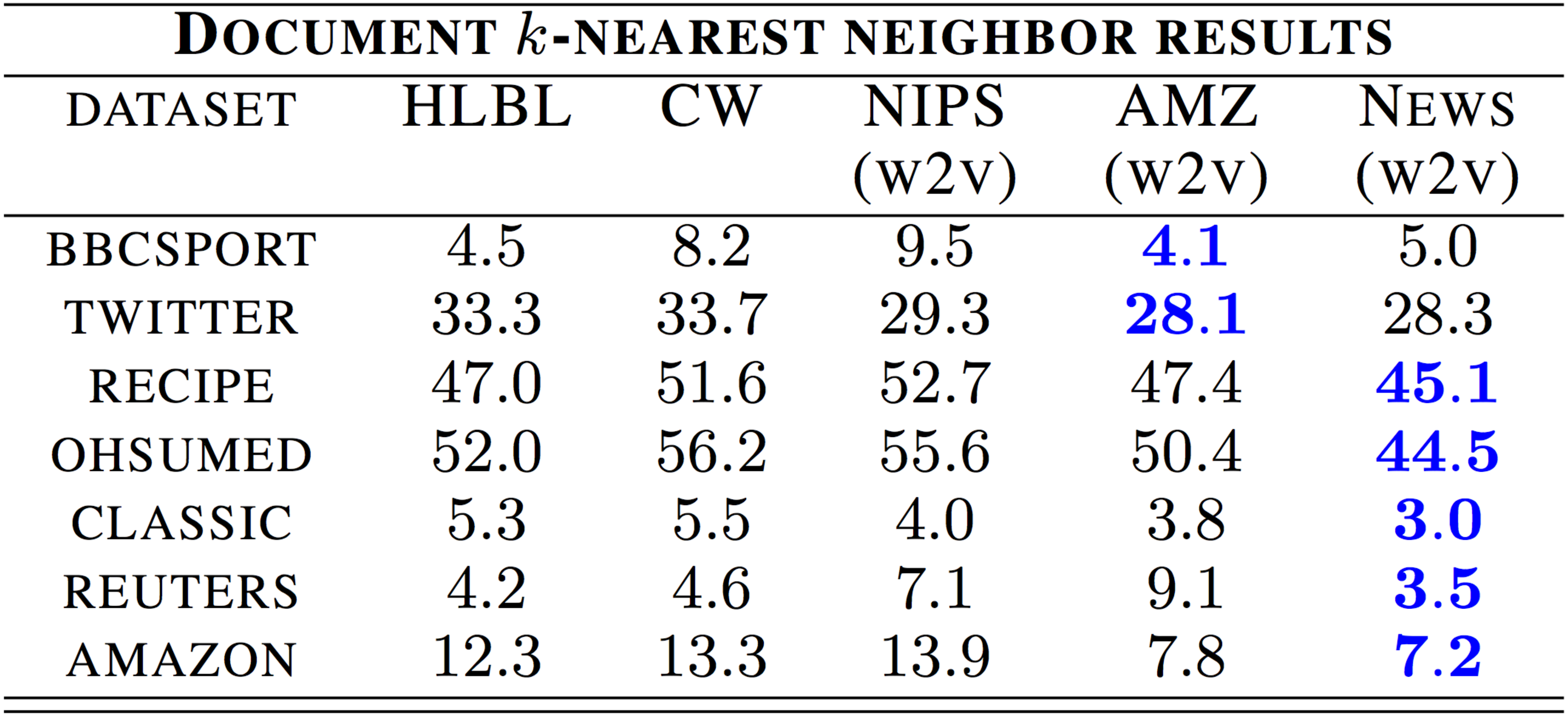
for a random test input...



Faster Approximations



Other Embeddings

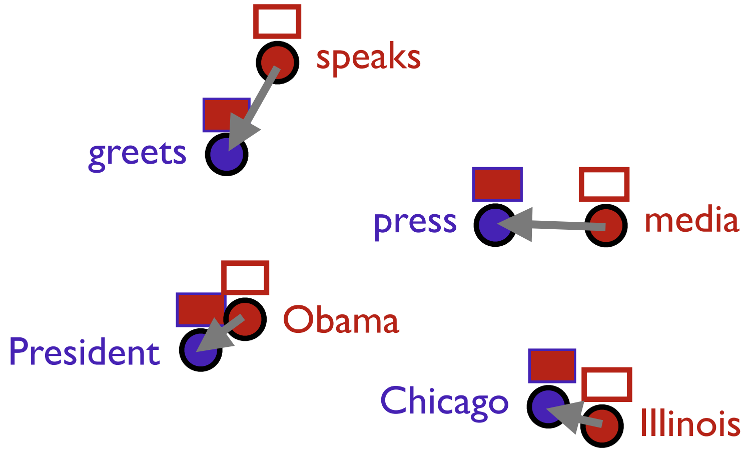


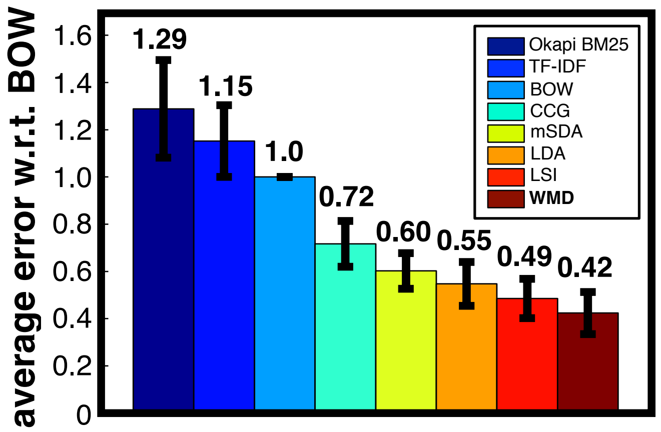
Document k-nearest neighbor results

Dataset	HLBL	CW	NIPS (w2v)	AMZ (w2v)	News (w2v)
BBCSPORT	4.5	8.2	9.5	**4.1**	5.0
TWITTER	33.3	33.7	29.3	**28.1**	28.3
RECIPE	47.0	51.6	52.7	47.4	**45.1**
OHSUMED	52.0	56.2	55.6	50.4	**44.5**
CLASSIC	5.3	5.5	4.0	3.8	**3.0**
REUTERS	4.2	4.6	7.1	9.1	**3.5**
AMAZON	12.3	13.3	13.9	7.8	**7.2**

Conclusion

- Word Mover's Distance:
  - **document distances from word embeddings**
  - Very accurate as it leverages high quality **word2vec** embedding
  - Fast through approximations





WMD	WCD	RWMD
$O(n^3 \log n)$	$O(nd)$	$O(n^2 d)$

Code: `http://matthewkusner.com`

Thank you. Questions?