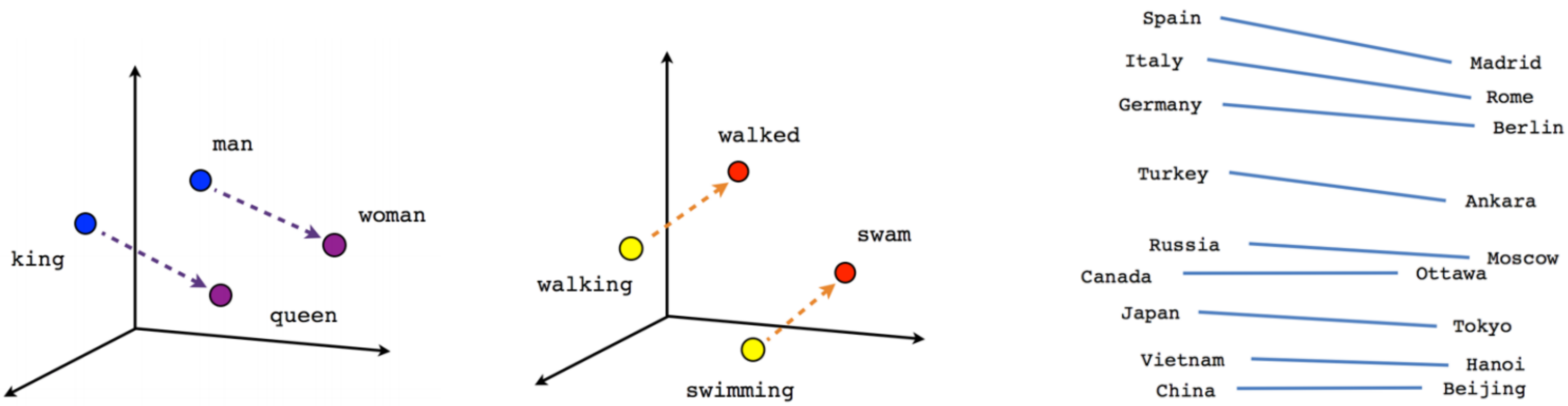
Representation Learning For Computational Imagination

Yong-Yeol (YY) Ahn
Indiana University

yyahn@iu.edu
@yy

Word2vec



Male-Female

Verb tense

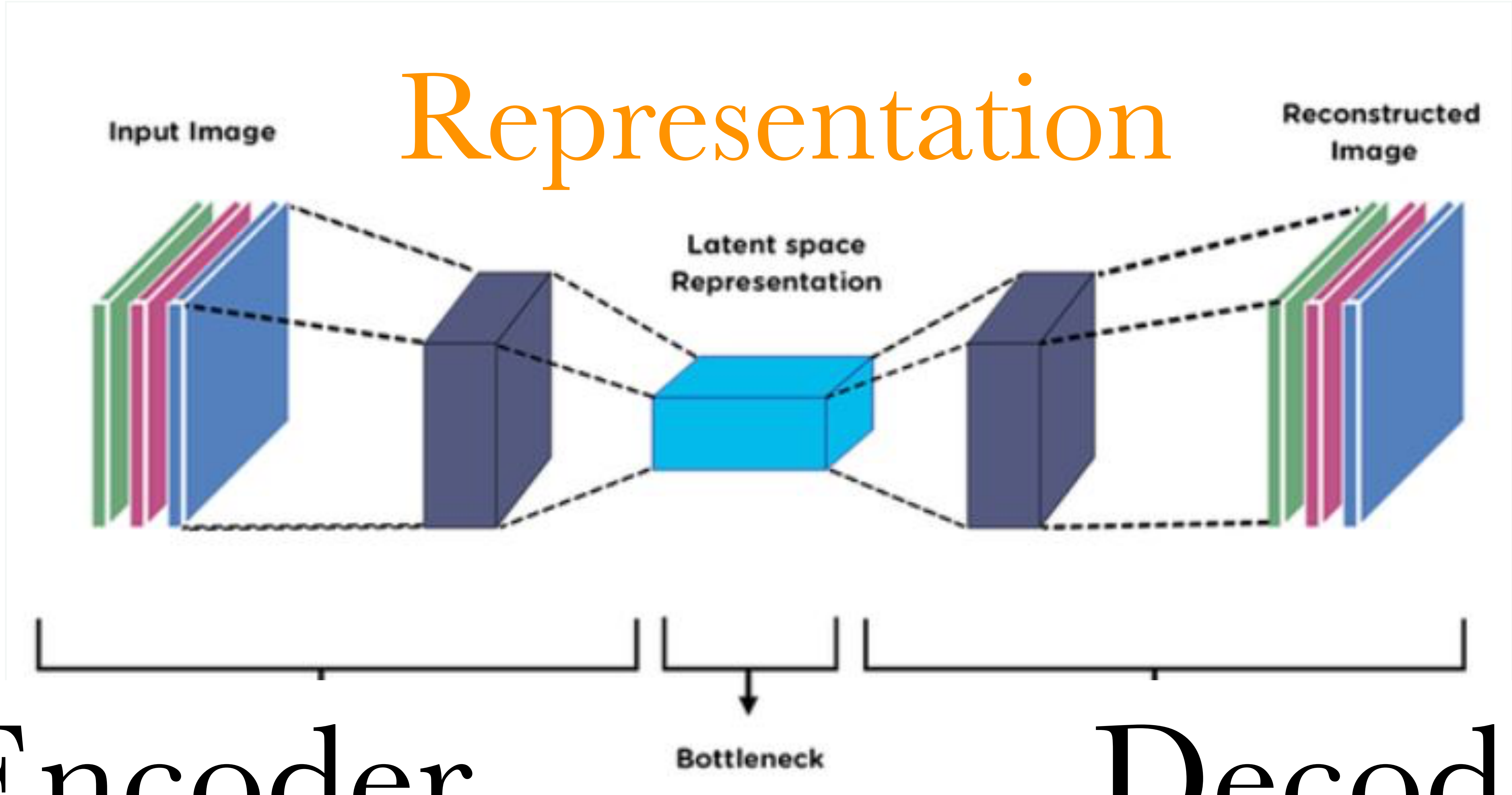
Country-Capital

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. "Distributed representations of words and phrases and their compositionality." *arXiv preprint arXiv:1310.4546* (2013).

Machine Learning

Data → **“feature vectors”** → Task

Deep Learning

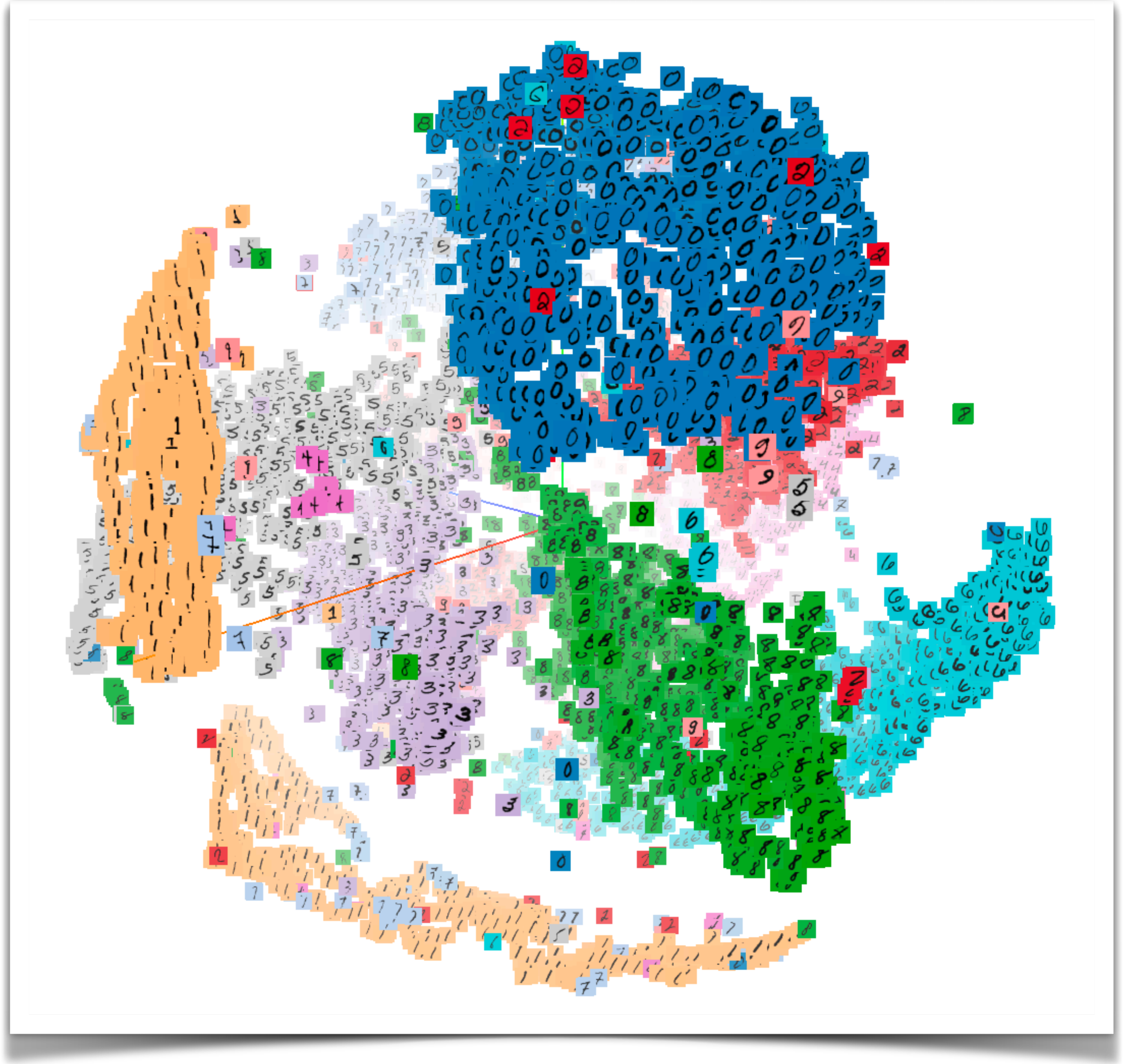


Encoder

Decoder

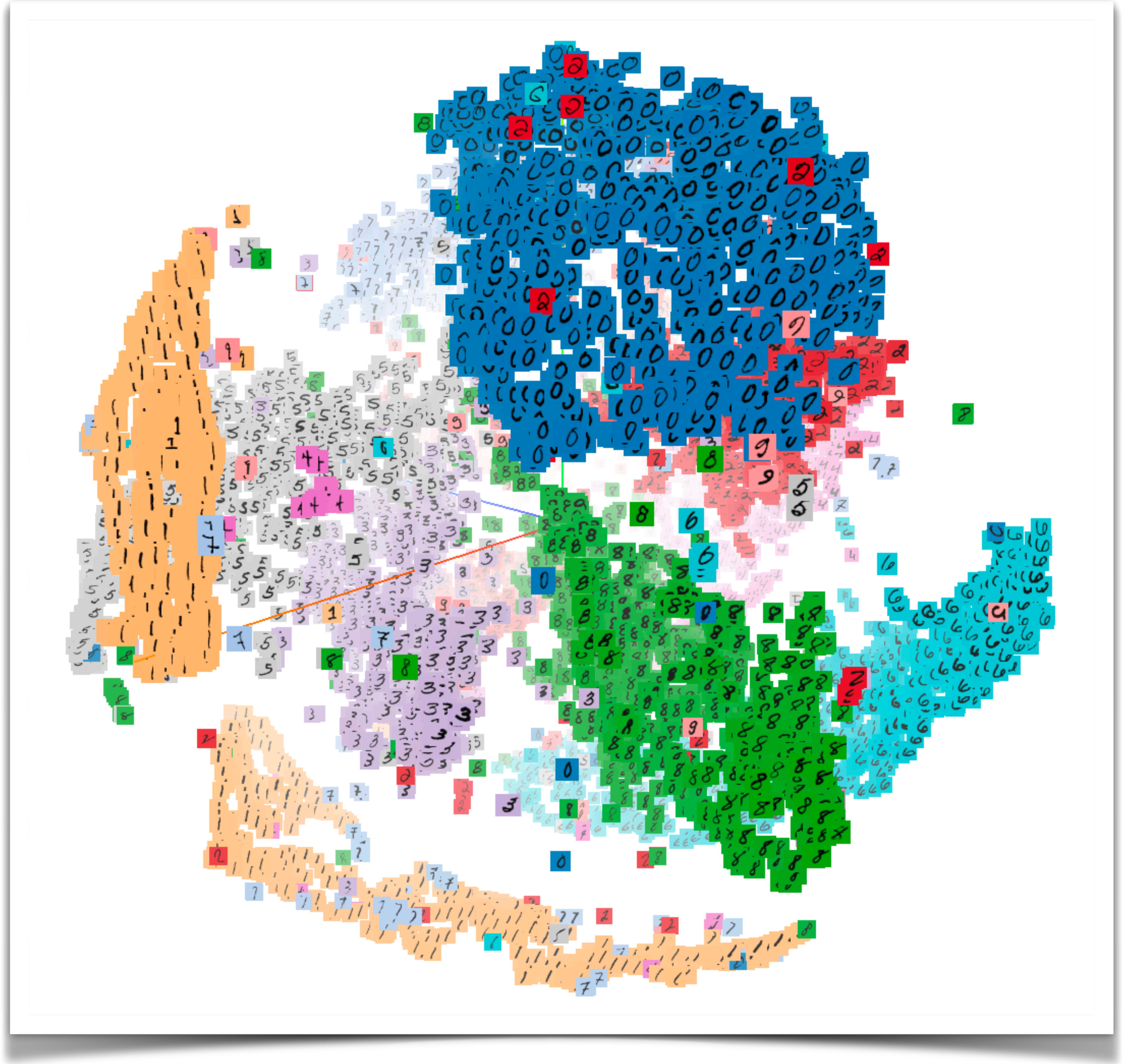
Can we let the machine discover useful features?

Representations live in a vector space.



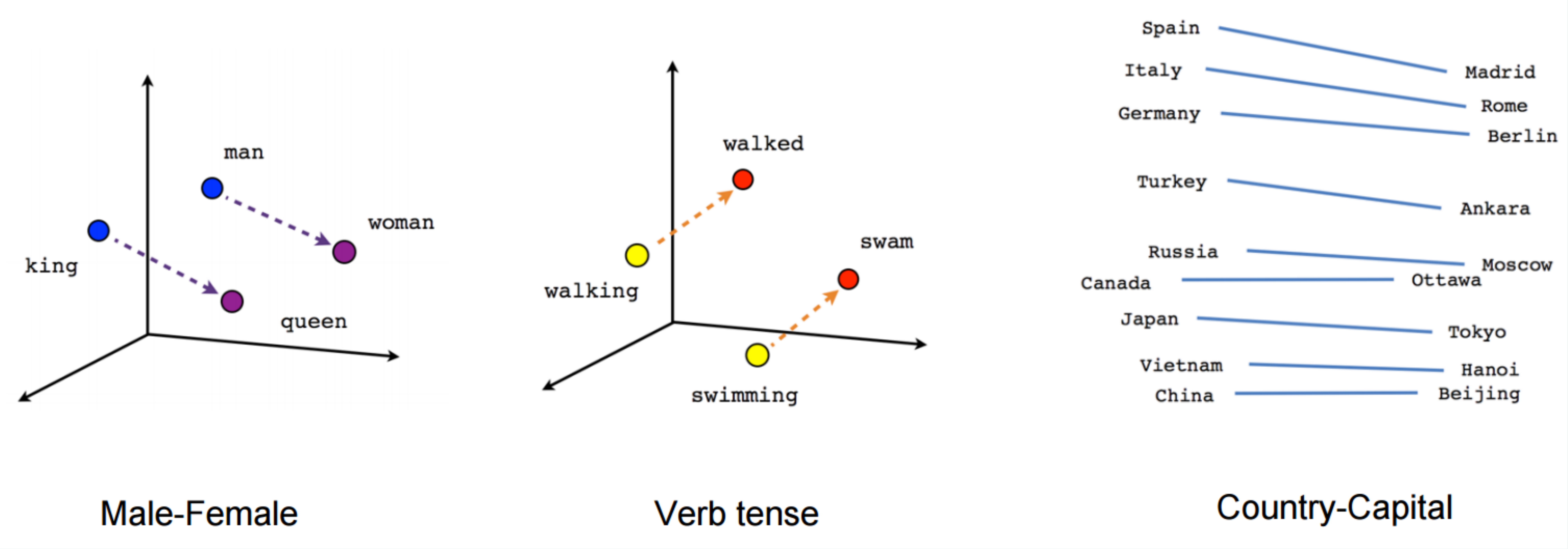
https://projector.tensorflow.org/

Can we interpret this *literally*, as a "**space**"?



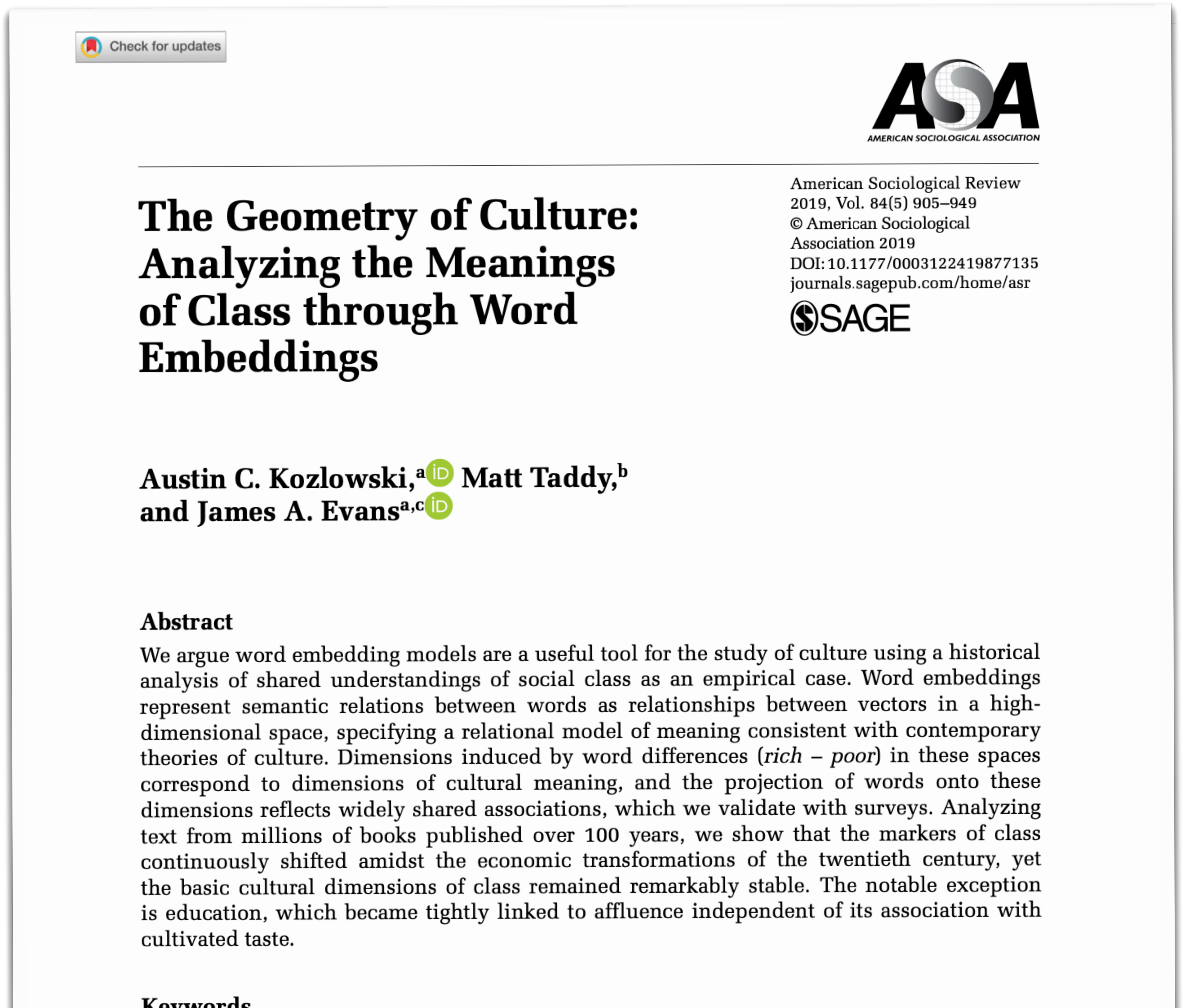
https://projector.tensorflow.org/

We can find meaningful *semantic axes* in the space



Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. "Distributed representations of words and phrases and their compositionality." *arXiv preprint arXiv:1310.4546* (2013).

“Geometry of Culture”



Check for updates

ASA
AMERICAN SOCIOLOGICAL ASSOCIATION

American Sociological Review
2019, Vol. 84(5) 905–949

DOI: 10.1177/0003122419877135
journals.sagepub.com/home/asr
SAGE

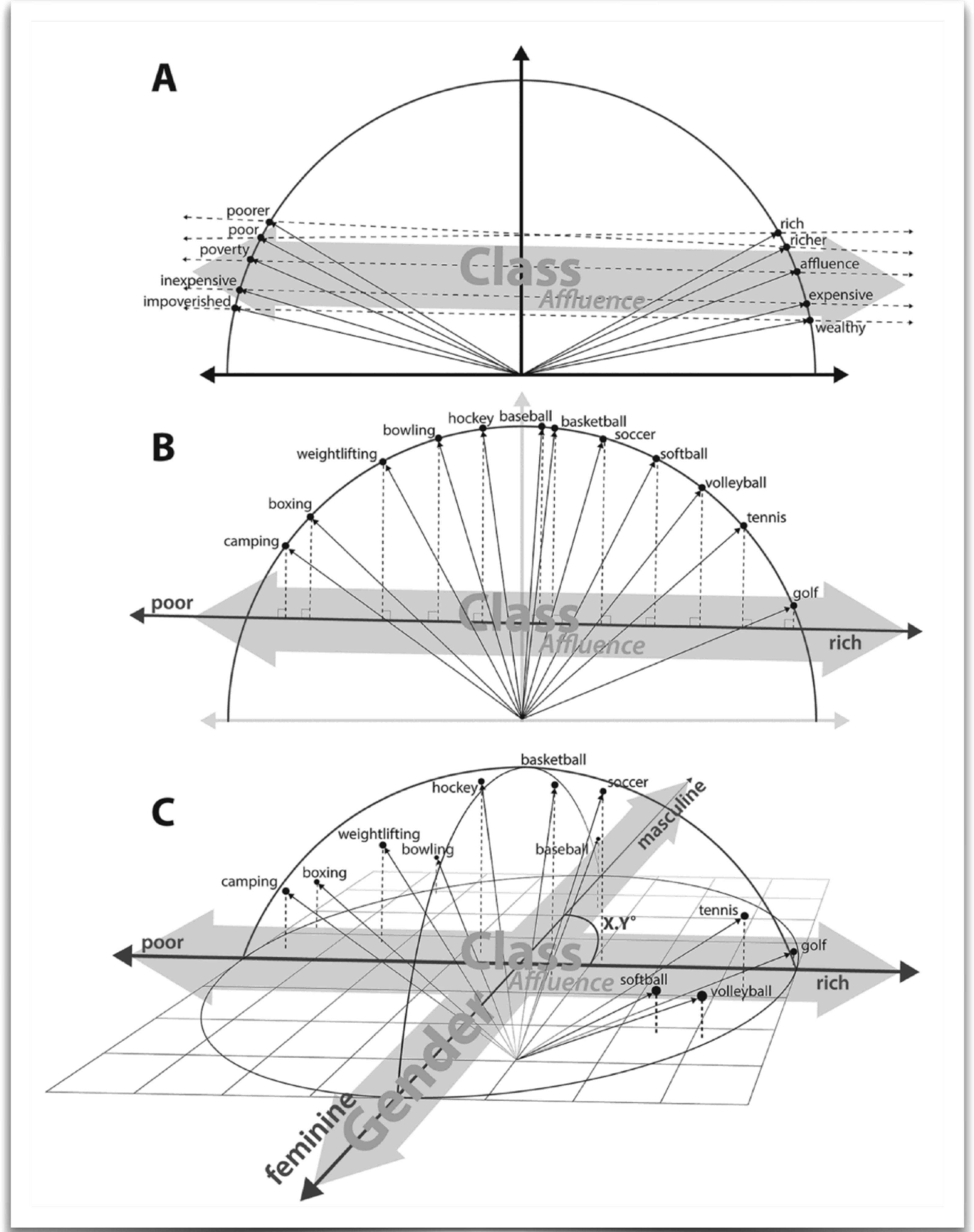
The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings

Austin C. Kozlowski,[a] Matt Taddy,[b] and James A. Evans[a,c]

Abstract

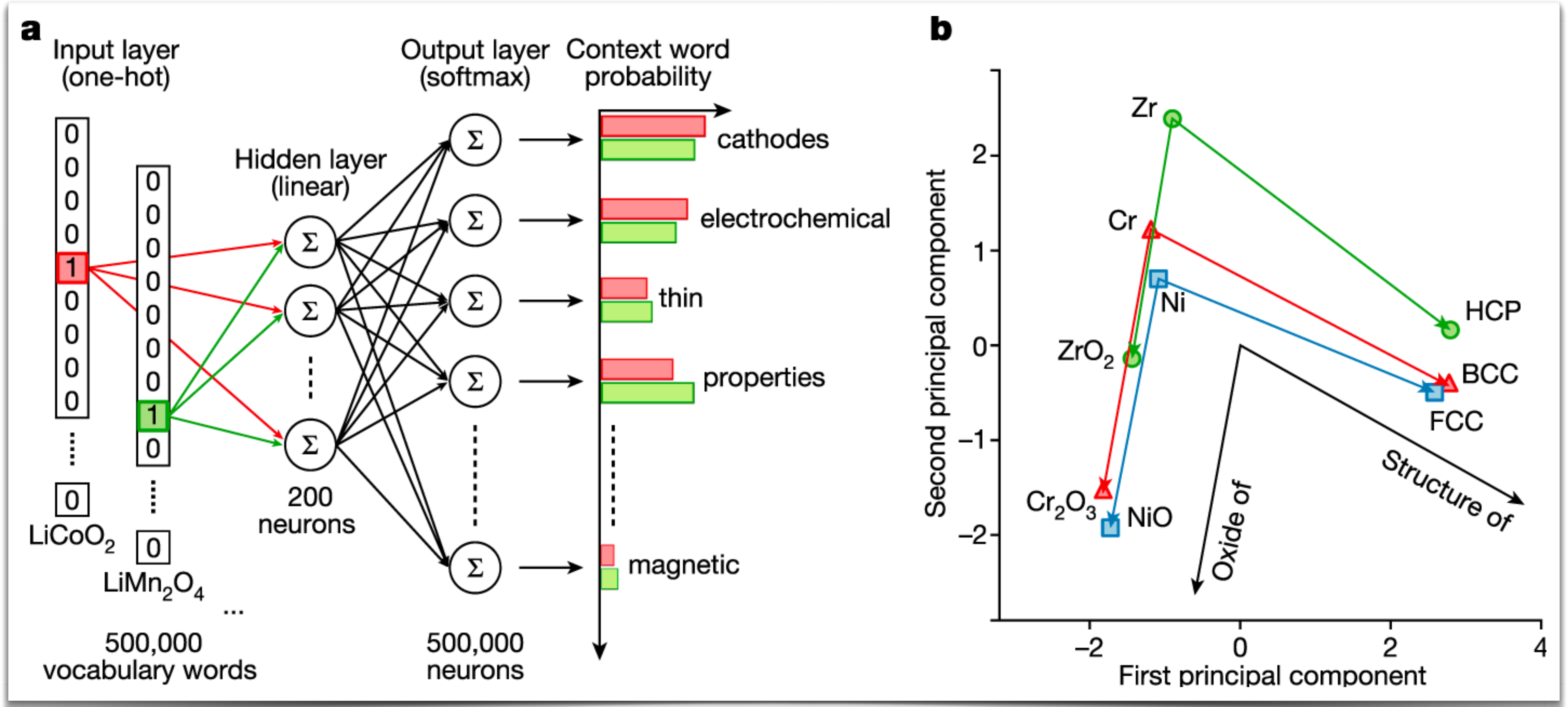
We argue word embedding models are a useful tool for the study of culture using a historical analysis of shared understandings of social class as an empirical case. Word embeddings represent semantic relations between words as relationships between vectors in a high-dimensional space, specifying a relational model of meaning consistent with contemporary theories of culture. Dimensions induced by word differences (*rich* – *poor*) in these spaces correspond to dimensions of cultural meaning, and the projection of words onto these dimensions reflects widely shared associations, which we validate with surveys. Analyzing text from millions of books published over 100 years, we show that the markers of class continuously shifted amidst the economic transformations of the twentieth century, yet the basic cultural dimensions of class remained remarkably stable. The notable exception is education, which became tightly linked to affluence independent of its association with cultivated taste.

Keywords



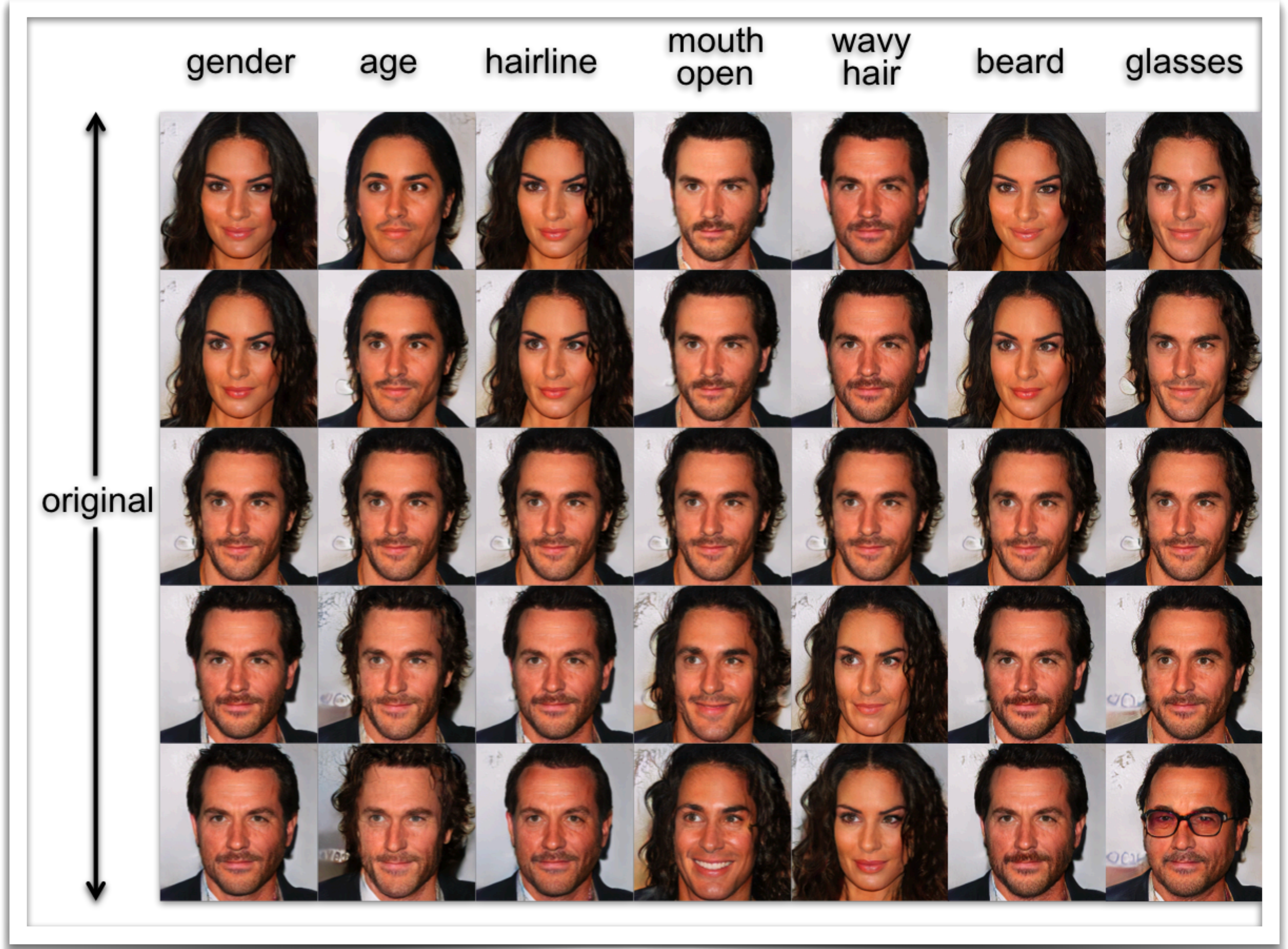
A. C. Kozlowski, M. Taddy, and J. Evans, ASR, 2019
https://arxiv.org/abs/1806.05521

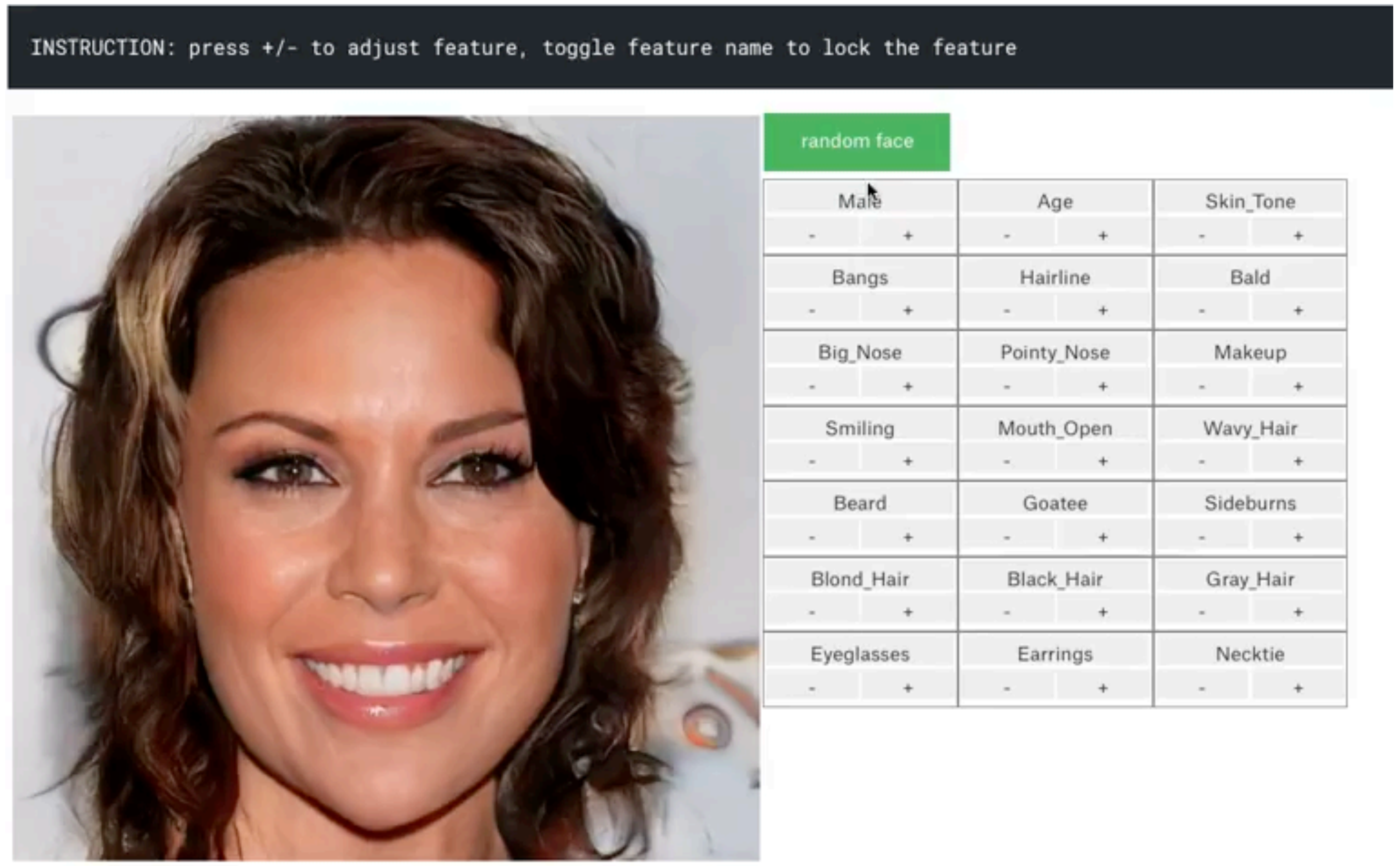
Meaningful axes about material properties



Tshitoyan, Vahe, et al. "Unsupervised word embeddings capture latent knowledge from materials science literature." Nature 571.7763 (2019): 95-98.

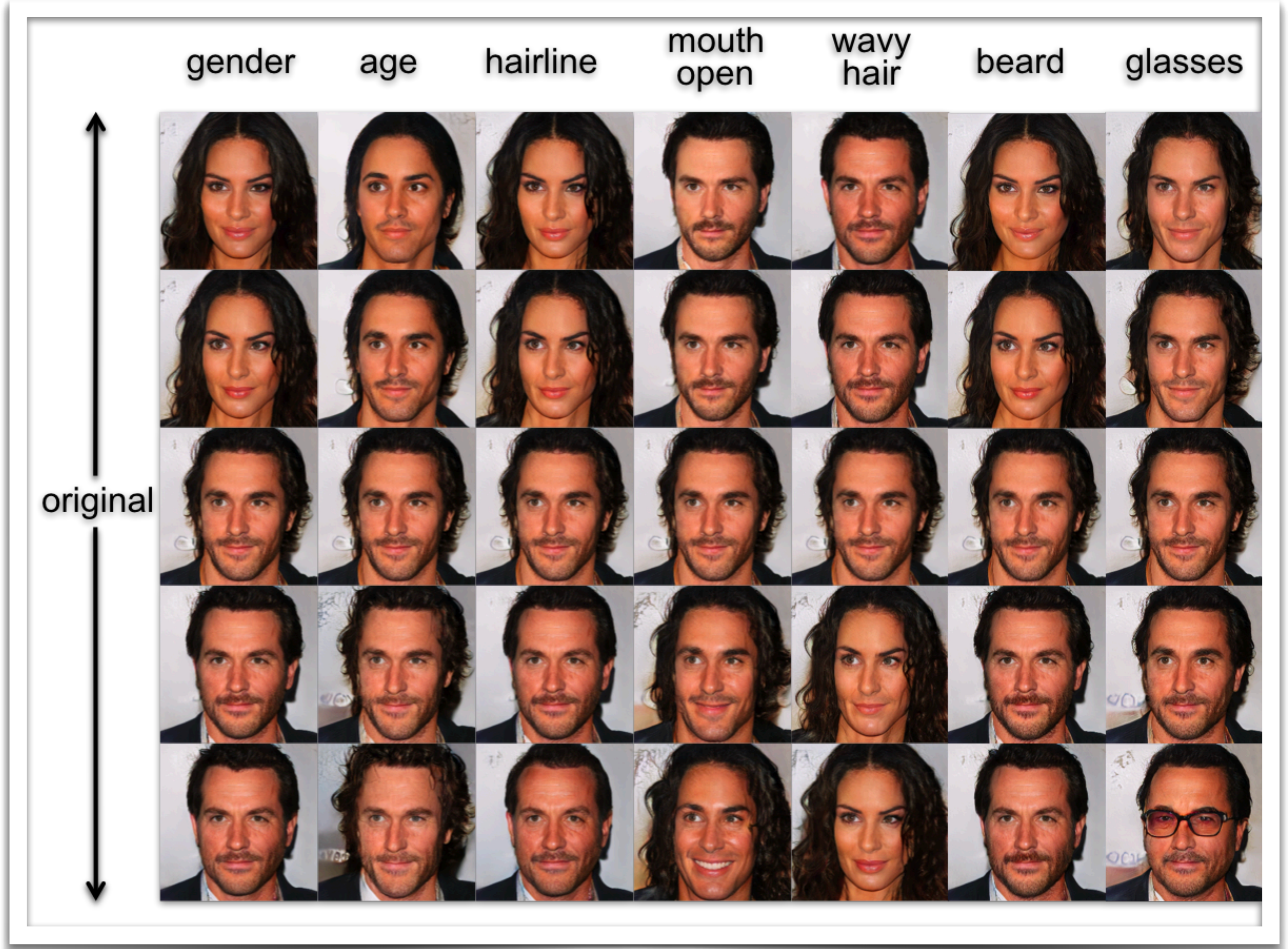
Meaningful axes about facial features

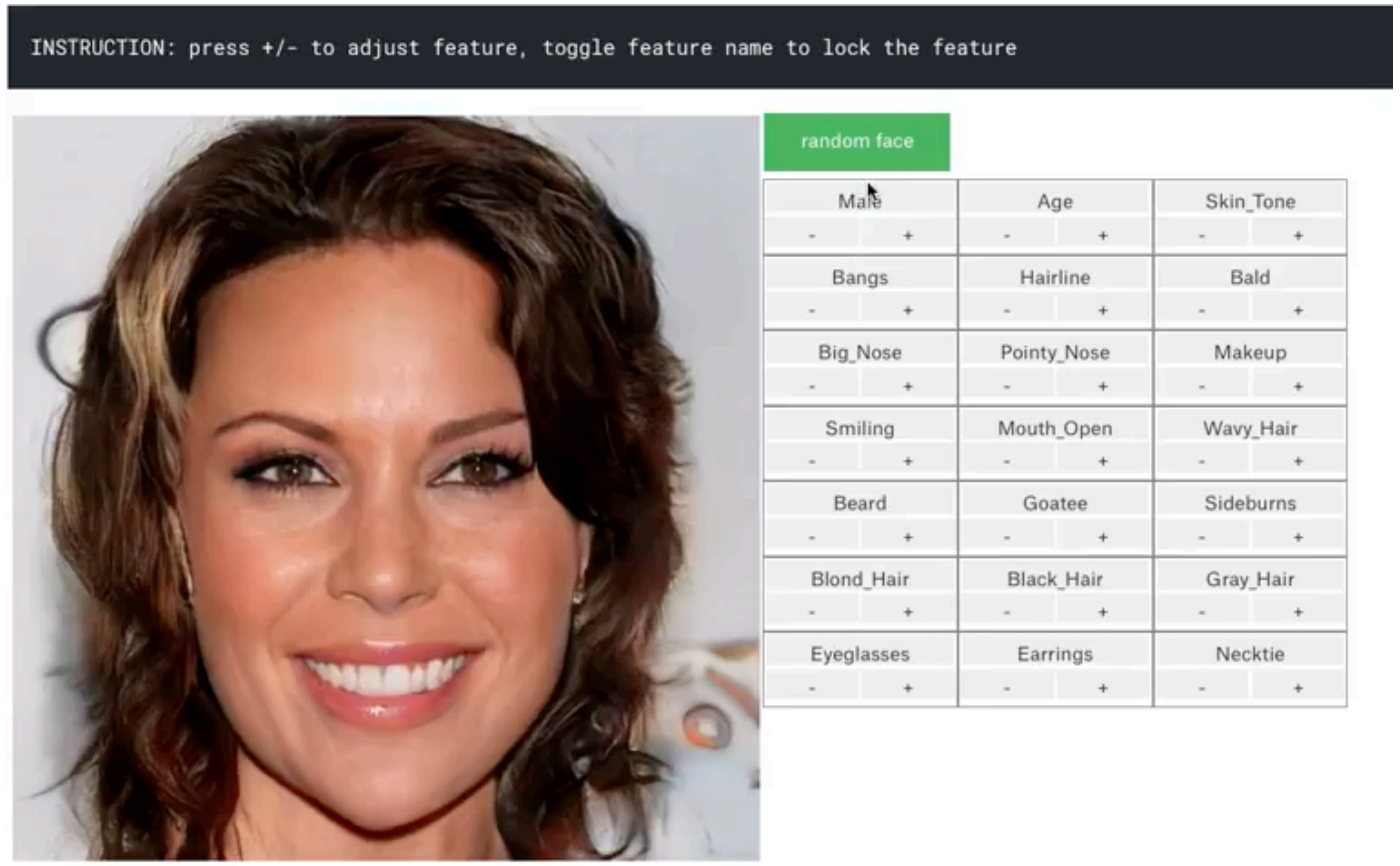




https://github.com/SummitKwan/transparent_latent_gan

Meaningful axes about facial features

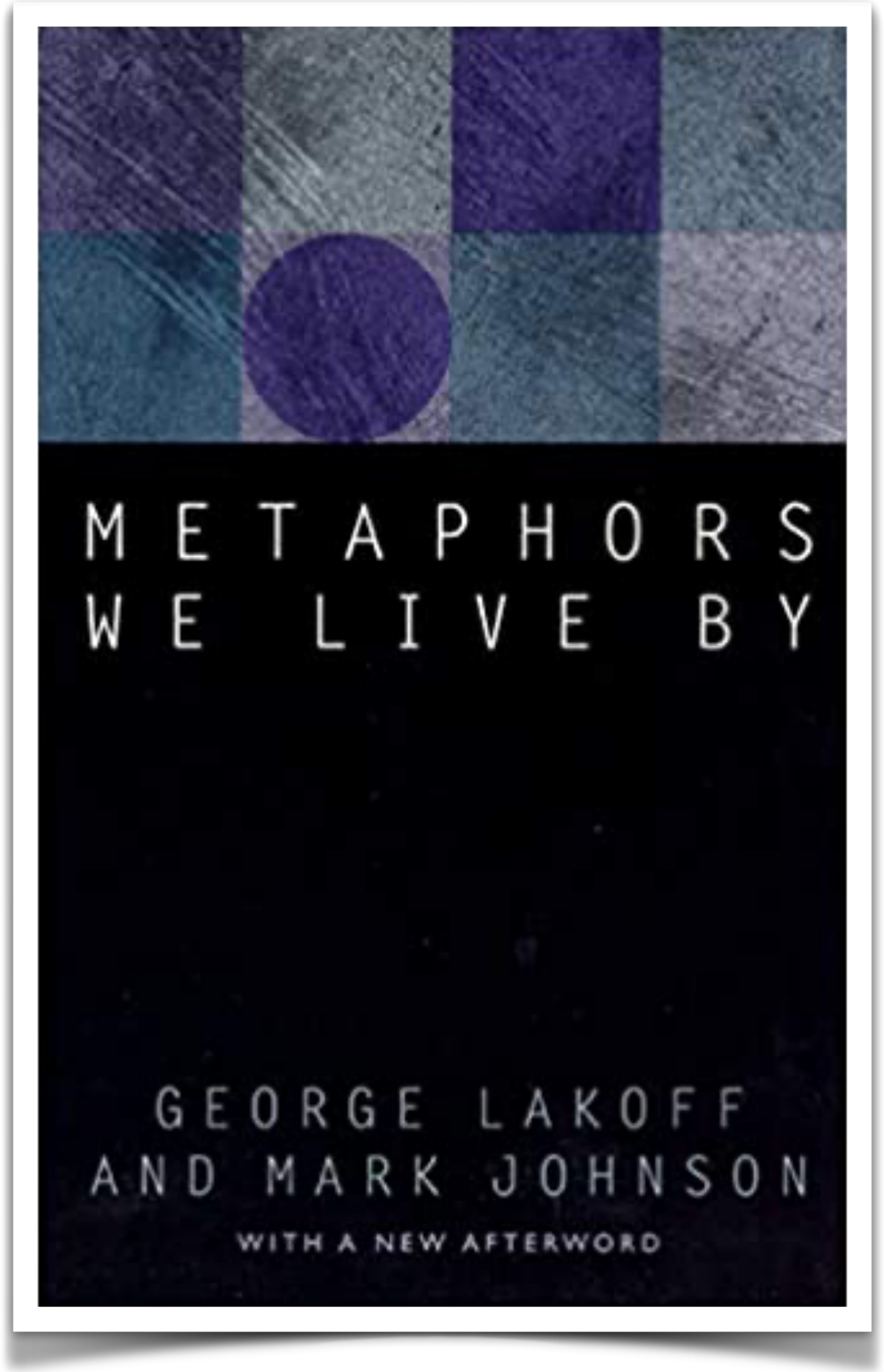




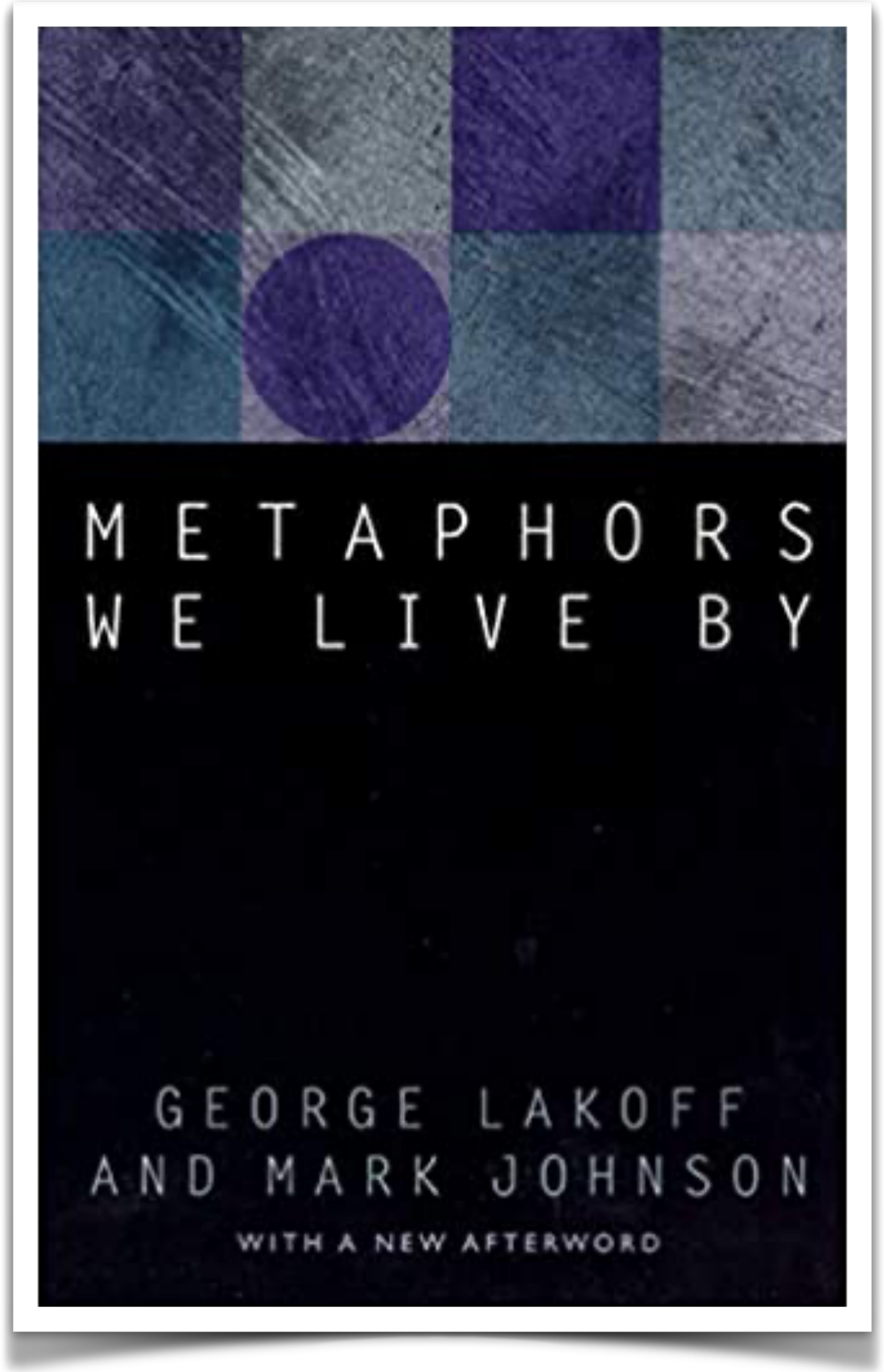
https://github.com/SummitKwan/transparent_latent_gan

The representation space itself is interesting!

We think and imagine *spatially*



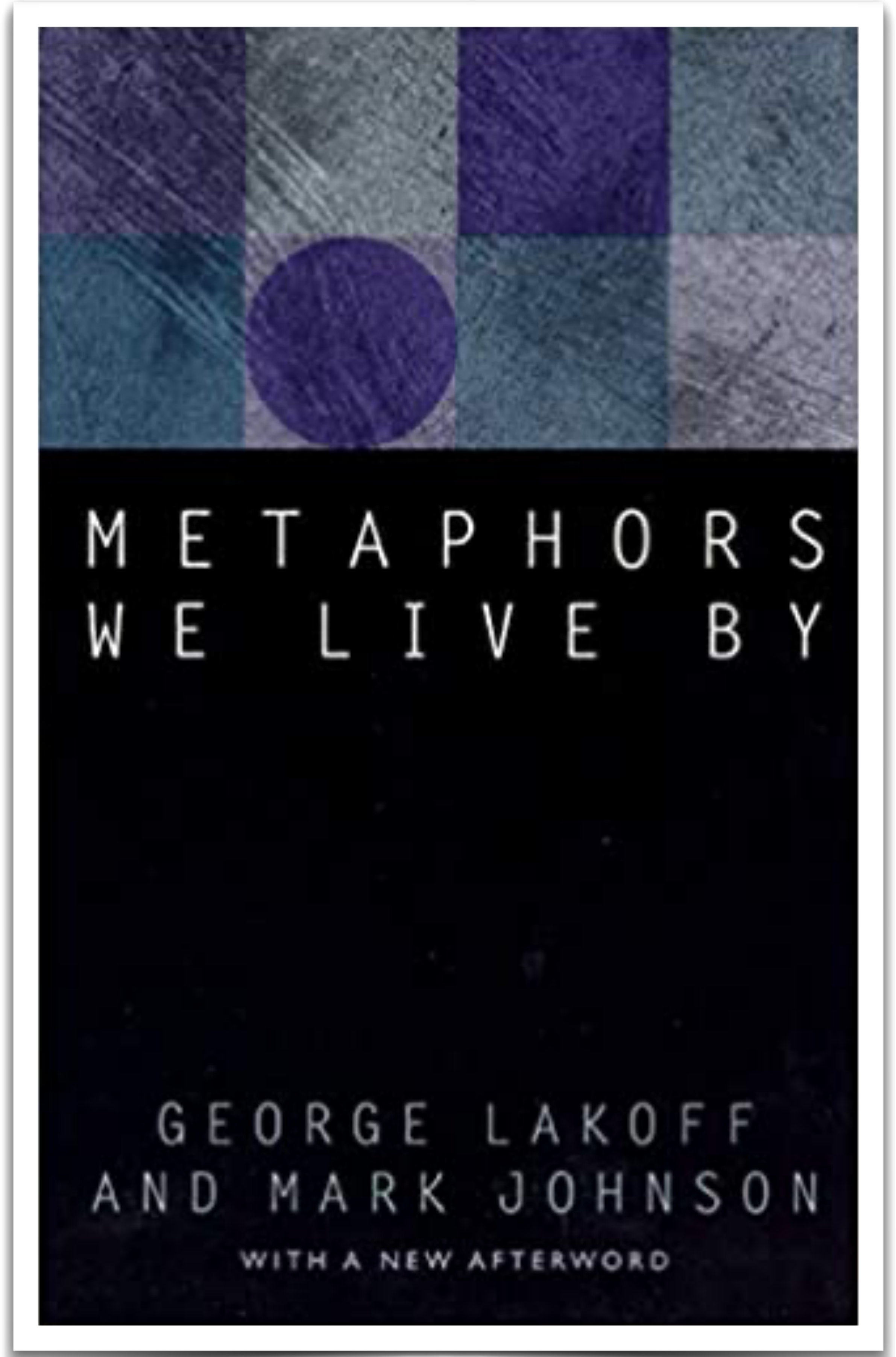
We think and imagine *spatially*



HAPPY IS UP; SAD IS DOWN

I'm feeling *up.* That *boosted* my spirits. My spirits *rose.* You're in *high* spirits. Thinking about her always gives me a *lift.* I'm feeling *down.* I'm *depressed.* He's really *low* these days. I *fell* into a depression. My spirits *sank.*

We think and imagine *spatially*



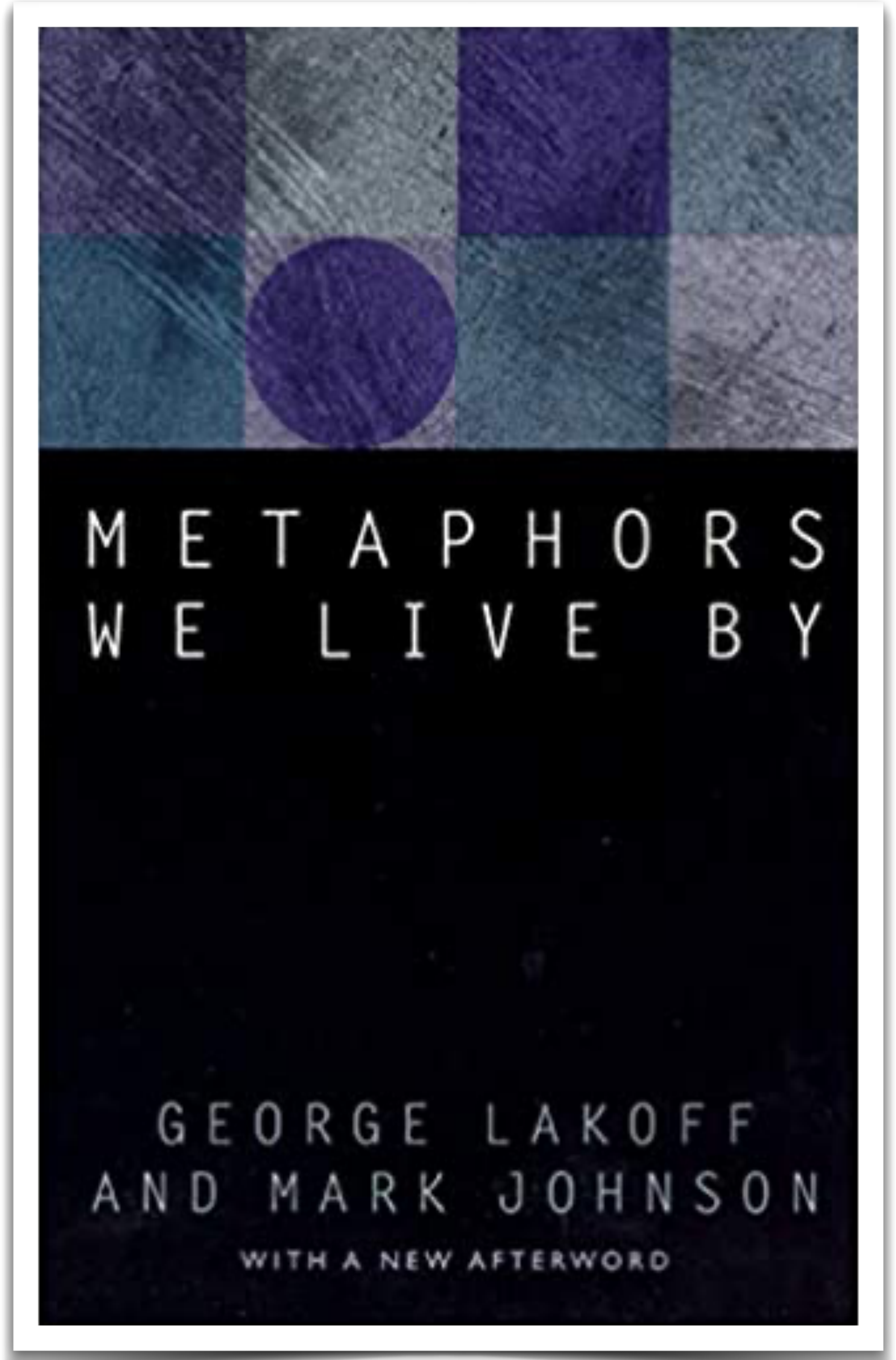
HAPPY IS UP; SAD IS DOWN

I'm feeling *up.* That *boosted* my spirits. My spirits *rose.* You're in *high* spirits. Thinking about her always gives me a *lift.* I'm feeling *down.* I'm *depressed.* He's really *low* these days. I *fell* into a depression. My spirits *sank.*

CONSCIOUS IS UP; UNCONSCIOUS IS DOWN

Get *up.* Wake *up.* I'm *up* already. He *rises* early in the morning. He *fell* asleep. He *dropped* off to sleep. He's *under* hypnosis. He *sank* into a coma.

We think and imagine *spatially*



HAPPY IS UP; SAD IS DOWN

I'm feeling *up.* That *boosted* my spirits. My spirits *rose.* You're in *high* spirits. Thinking about her always gives me a *lift.* I'm feeling *down.* I'm *depressed.* He's really *low* these days. I *fell* into a depression. My spirits *sank.*

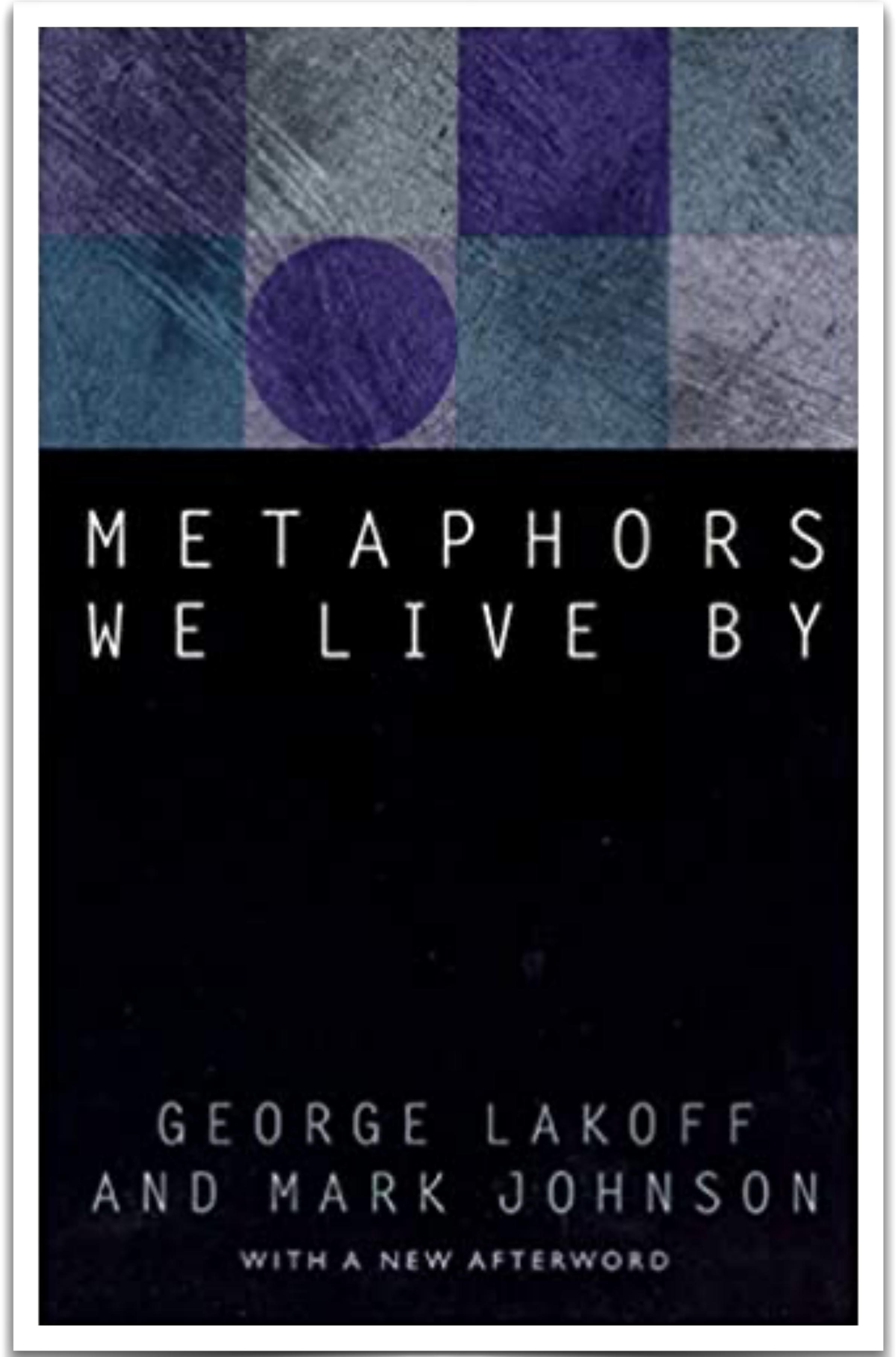
CONSCIOUS IS UP; UNCONSCIOUS IS DOWN

Get *up.* Wake *up.* I'm *up* already. He *rises* early in the morning. He *fell* asleep. He *dropped* off to sleep. He's *under* hypnosis. He *sank* into a coma.

HEALTH AND LIFE ARE UP; SICKNESS AND DEATH ARE DOWN

He's at the *peak* of health. Lazarus *rose* from the dead. He's in *top* shape. As to his health, he's way *up* there. He *fell* ill. He's *sinking* fast. He came *down* with the flu. His health is *declining.* He *dropped* dead.

We think and imagine *spatially*



HAPPY IS UP; SAD IS DOWN

I'm feeling *up.* That *boosted* my spirits. My spirits *rose.* You're in *high* spirits. Thinking about her always gives me a *lift.* I'm feeling *down.* I'm *depressed.* He's really *low* these days. I *fell* into a depression. My spirits *sank.*

CONSCIOUS IS UP; UNCONSCIOUS IS DOWN

Get *up.* Wake *up.* I'm *up* already. He *rises* early in the morning. He *fell* asleep. He *dropped* off to sleep. He's *under* hypnosis. He *sank* into a coma.

HEALTH AND LIFE ARE UP; SICKNESS AND DEATH ARE DOWN

He's at the *peak* of health. Lazarus *rose* from the dead. He's in *top* shape. As to his health, he's way *up* there. He *fell* ill. He's *sinking* fast. He came *down* with the flu. His health is *declining.* He *dropped* dead.

GOOD IS UP; BAD IS DOWN

Things are looking *up.* We hit a *peak* last year, but it's been *downhill* ever since. Things are at an all-time *low.* He does *high*-quality work.

Representation learning ~ matrix factorization

Neural Word Embedding as Implicit Matrix Factorization

Omer Levy
Department of Computer Science
Bar-Ilan University
omerlevy@gmail.com

Yoav Goldberg
Department of Computer Science
Bar-Ilan University
yoav.goldberg@gmail.com

Abstract

Improving Distributional Similarity with Lessons Learned from Word Embeddings

Omer Levy **Yoav Goldberg** **Ido Dagan**
Computer Science Department
Bar-Ilan University
Ramat-Gan, Israel
{omerlevy,yogo,dagan}@cs.biu.ac.il

Abstract

trends suggest that neural-
k-inspired word embedding models
form traditional count-based distri-
l models on word similarity and
y detection tasks. We reveal that

A recent study by Baroni et al. (2014) conducts a set of systematic experiments comparing `word2vec` embeddings to the more traditional distributional methods, such as pointwise mutual information (PMI) matrices (see Turney and Pantel (2010) and Baroni and Lenci (2010) for comprehensive surveys). These results suggest

How should we ***represent*** them?

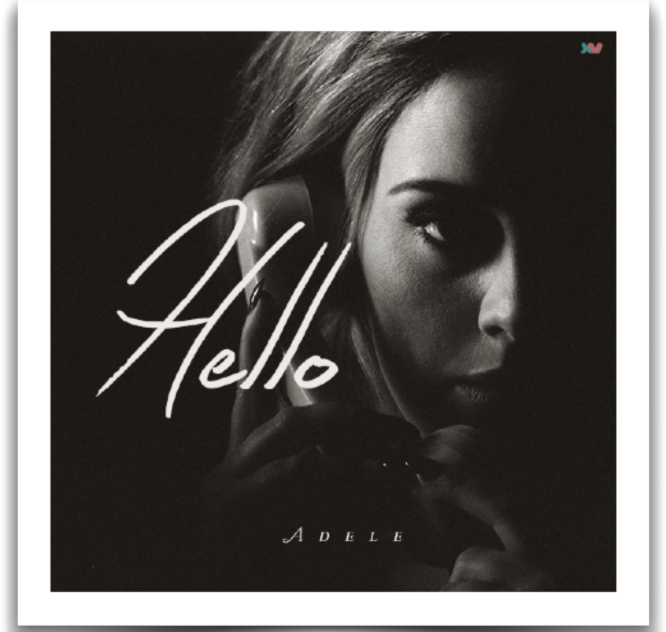
How to encode the "meaning"?

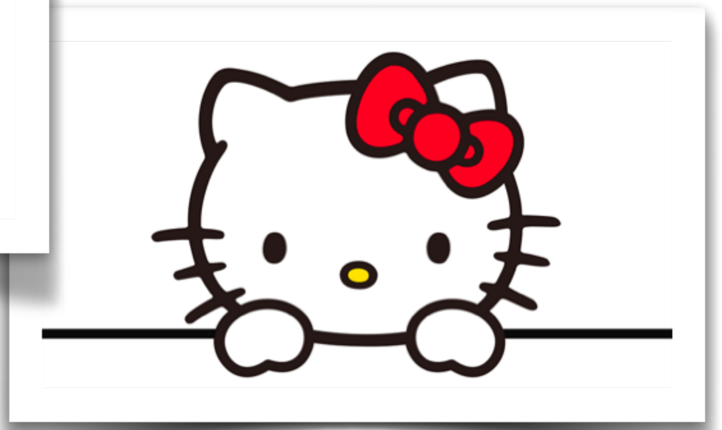
```
hello world!
```

↓

```
01101000 01100101 01101100 01101100
01101111 00100000 01110111 01101111
01110010 01101100 01100100 00100001
```

How should we ***represent*** them?
How to encode the "meaning"?

```
hello world!
```

```
01101000 01100101 01101100 01101100
01101111 00100000 01110111 01101111
01110010 01101100 01100100 00100001
```



Harris, Z. (1954). Distributional structure. Word, 10(23): 146-162.
Firth, J.R. (1957). A synopsis of linguistic theory 1930-1955. In Studies in Linguistic Analysis, pp. 1-32. Oxford: Philological Society. Reprinted in F.R. Palmer (ed.)

Ferdinand de Saussure

"Among all the individuals that are linked together by speech, some sort of average will be set up : all will reproduce — not exactly of course, but approximately — the same signs united with the same concepts."

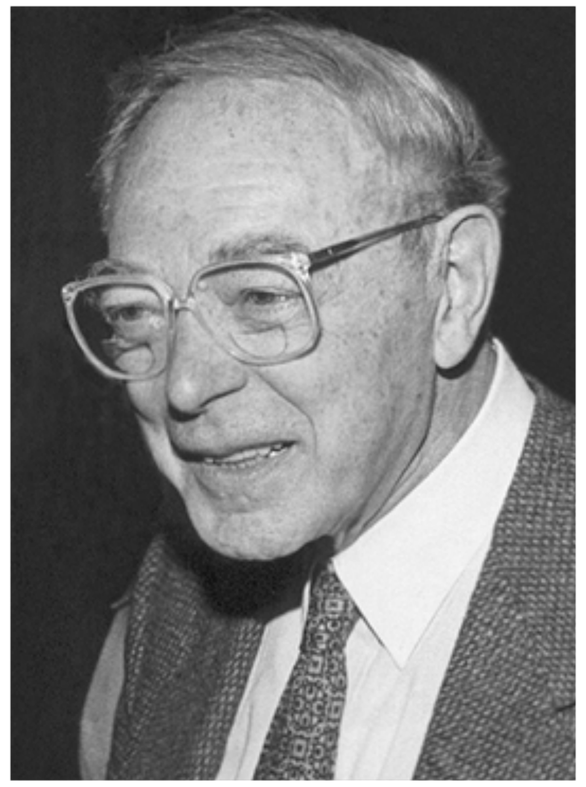
Harris, Z. (1954). Distributional structure. Word, 10(23): 146-162.
Firth, J.R. (1957). A synopsis of linguistic theory 1930-1955. In Studies in Linguistic Analysis, pp. 1-32. Oxford: Philological Society. Reprinted in F.R. Palmer (ed.)

Ferdinand de Saussure

"Among all the individuals that are linked together by speech, some sort of average will be set up : all will reproduce — not exactly of course, but approximately — the same signs united with the same concepts."

Distributional hypothesis*: words that occur in the same contexts tend to have similar meanings.*

We can study language by analyzing how it is used in a corpus.



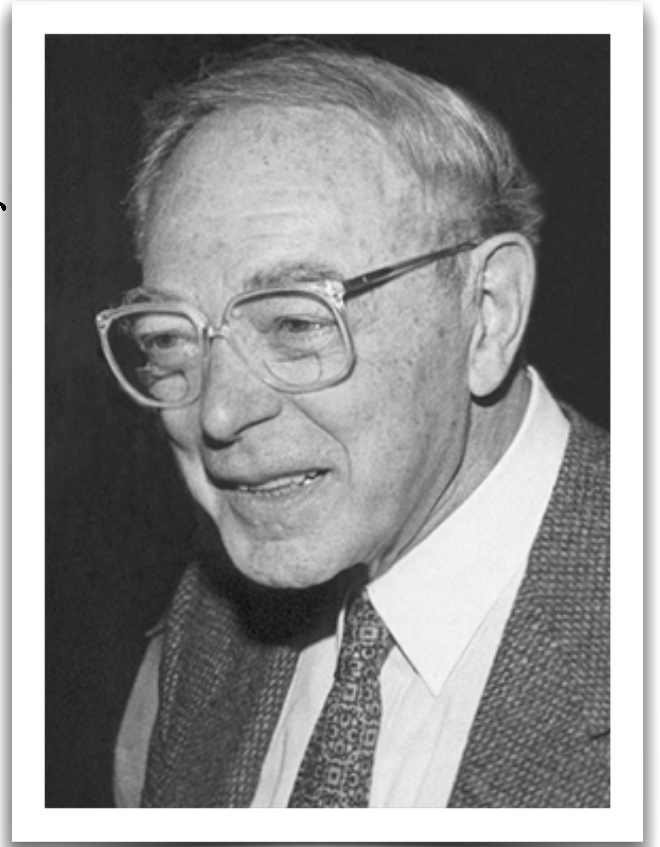
Zellig S. Harris

Harris, Z. (1954). Distributional structure. Word, 10(23): 146-162.
Firth, J.R. (1957). A synopsis of linguistic theory 1930-1955. In Studies in Linguistic Analysis, pp. 1-32. Oxford: Philological Society. Reprinted in F.R. Palmer (ed.)

Ferdinand de Saussure

"Among all the individuals that are linked together by speech, some sort of average will be set up : all will reproduce — not exactly of course, but approximately — the same signs united with the same concepts."

Distributional hypothesis: *words that occur in the same contexts tend to have similar meanings.*



Zellig S. Harris

We can study language by analyzing how it is used in a corpus.

John R. Firth

"You shall know a word by the company it keeps."

Harris, Z. (1954). Distributional structure. Word, 10(23): 146-162.
Firth, J.R. (1957). A synopsis of linguistic theory 1930-1955. In Studies in Linguistic Analysis, pp. 1-32. Oxford: Philological Society. Reprinted in F.R. Palmer (ed.)

Contexts ~ Meaning

The quick brown ____ jumps over the lazy dog.

He is cunning as a ____.

The ____ was already in your chicken house.

…

Contexts ~ Meaning

The quick brown ____ jumps over the lazy dog.

He is cunning as a ____.

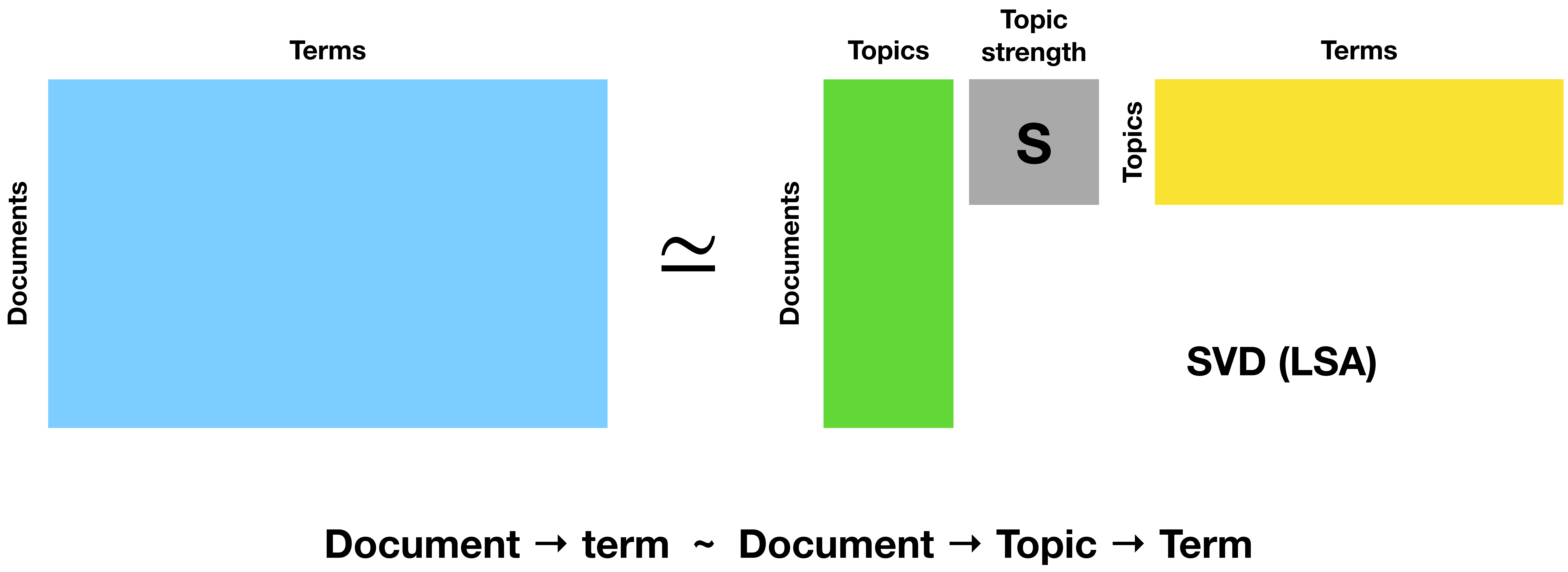
The ____ was already in your chicken house.

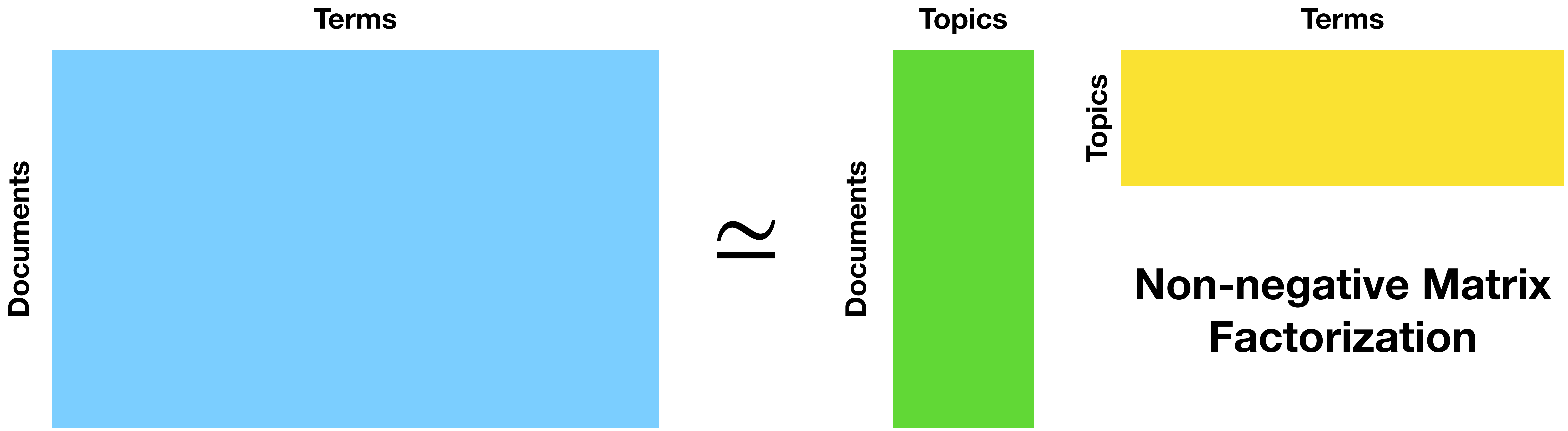
…

What do you mean by “contexts”?

Encoding contexts: term-document matrix

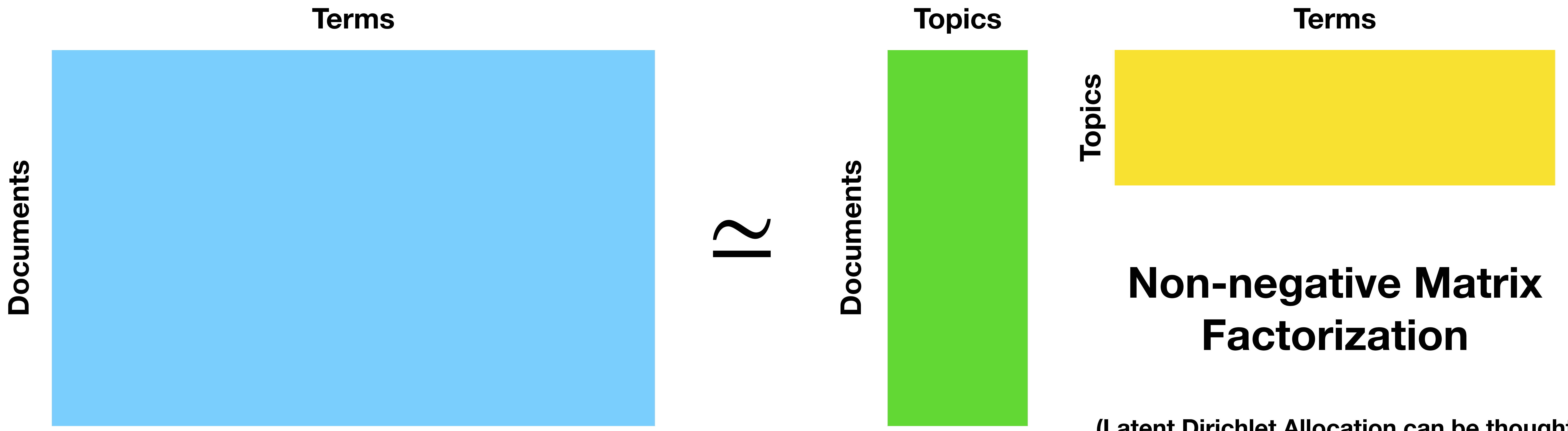
	cat	dog	fox	wolf	coyote	…
D1	15	10	0	0	8	…
D2	2	6	2	2	0	…
D3	0	1	16	15	6	…
…	…	…	…	…	…	…


Terms
Documents
≃
Topics
Topic strength
Terms
Documents
S
Topics
SVD (LSA)
Document → term ~ Document → Topic → Term



Non-negative Matrix Factorization

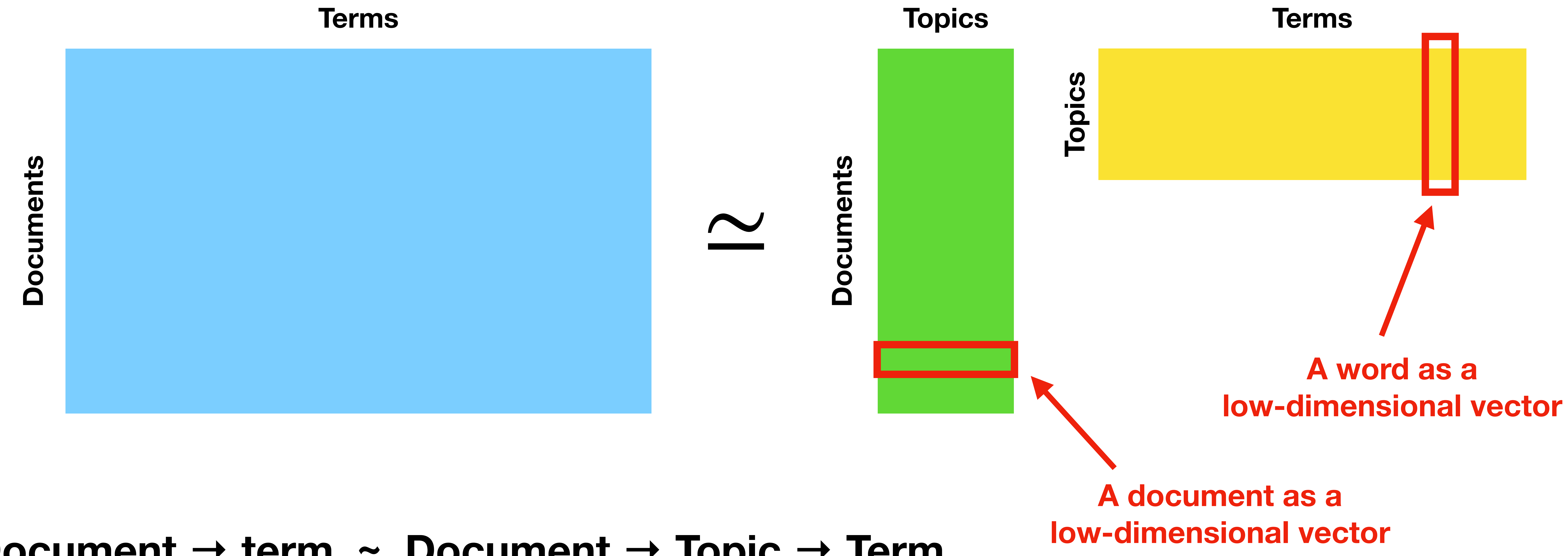
Document → term ~ Document → Topic → Term



Non-negative Matrix Factorization

(Latent Dirichlet Allocation can be thought as a “softer” Bayesian method to do this.)

Document → term ~ Document → Topic → Term

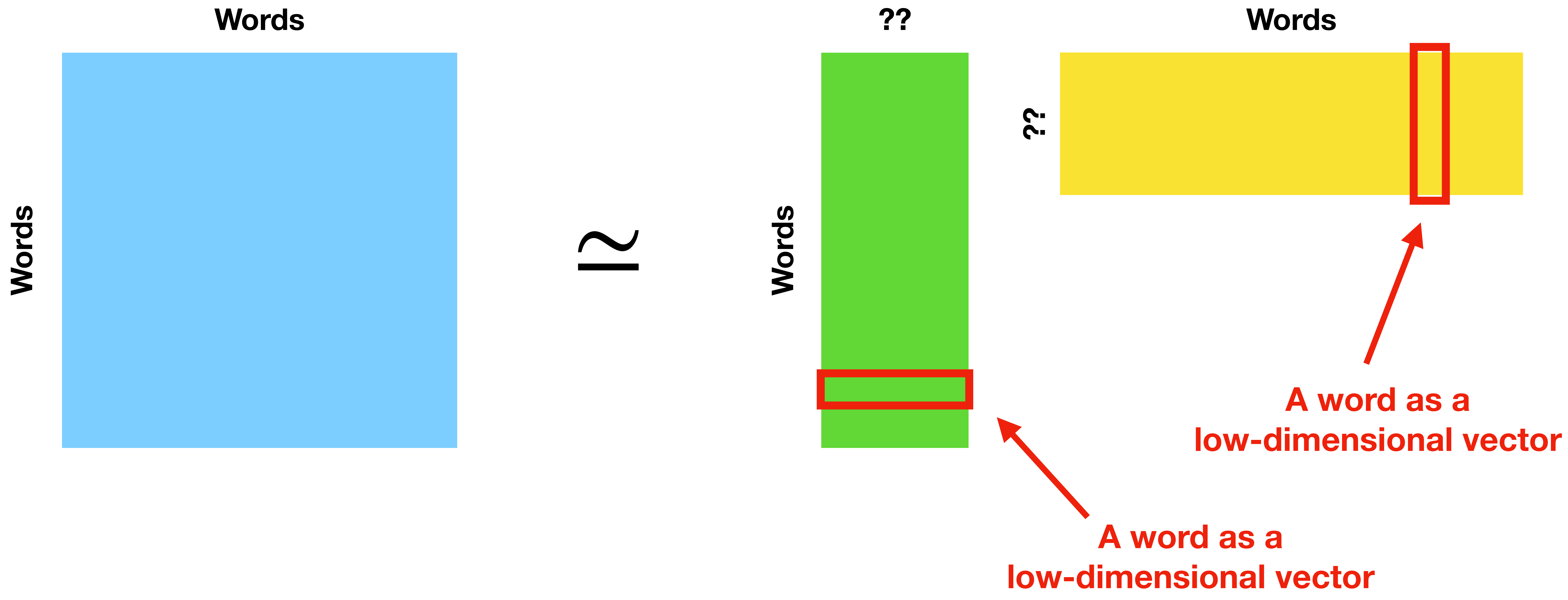


Document → term ~ Document → Topic → Term

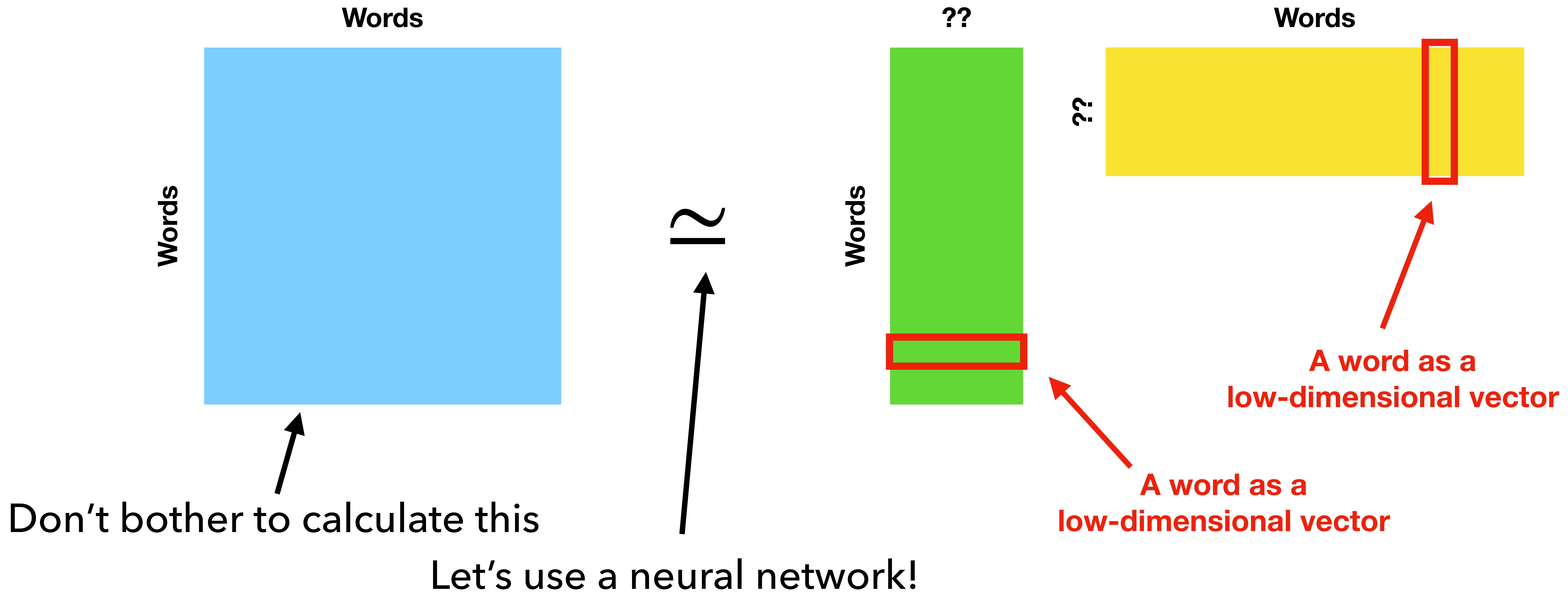
Using *documents* as "contexts" led to nice **models** and **representations**.

Can we think of nearby *words* as *contexts*?

Neural Language Model as a Matrix Factorization



Neural Language Model as a Matrix Factorization



The idea of "language model"

What is the *probability* of this sentence?

A good *language model* should assign high probability for real sentences and low probability for nonsensical sentences.

The idea of "language model"

What is the *probability* of this sentence?

A good *language model* should assign high probability for real sentences and low probability for nonsensical sentences.

$$P(w_1, w_2, \ldots, w_n) = ?$$

The idea of "language model"

What is the *probability* of this sentence?

A good *language model* should assign high probability for real sentences and low probability for nonsensical sentences.

$$P(w_1, w_2, \ldots, w_n) = ?$$

$$P(w_1, w_2, \ldots, w_n) = P(w_n \,|\, w_1, \ldots, w_{n-1}) P(w_{n-1} \,|\, w_1, \ldots, w_{n-2}) \times P(w_{n-2} \,|\, w_1, \ldots, w_{n-3}) \times \cdots \times P(w_1)$$

The idea of "language model"

$$
\begin{aligned}
P(w_1, w_2, \ldots, w_n) = P(w_n \mid w_1, \ldots, w_{n-1}) P(w_{n-1} \mid w_1, \ldots, w_{n-2}) \\
\times P(w_{n-2} \mid w_1, \ldots, w_{n-3}) \times \cdots \times P(w_1)
\end{aligned}
$$

The idea of "language model"

$$P(w_1, w_2, \ldots, w_n) = P(w_n \mid w_1, \ldots, w_{n-1}) P(w_{n-1} \mid w_1, \ldots, w_{n-2}) \times P(w_{n-2} \mid w_1, \ldots, w_{n-3}) \times \cdots \times P(w_1)$$

What is the probability of the *next word*?

The idea of "language model"

$$P(w_1, w_2, \ldots, w_n) = P(w_n \mid w_1, \ldots, w_{n-1}) P(w_{n-1} \mid w_1, \ldots, w_{n-2}) \\ \times P(w_{n-2} \mid w_1, \ldots, w_{n-3}) \times \cdots \times P(w_1)$$

What is the probability of the *next word*?

$$P(w_t \mid w_1, \ldots, w_{t-1}) = \ ?$$

The idea of "language model"

$$P(w_1, w_2, \ldots, w_n) = P(w_n \mid w_1, \ldots, w_{n-1}) P(w_{n-1} \mid w_1, \ldots, w_{n-2}) \times P(w_{n-2} \mid w_1, \ldots, w_{n-3}) \times \cdots \times P(w_1)$$

What is the probability of the *next word*?

$$P(\underbrace{w_t}_{\text{Target}} \mid w_1, \ldots, w_{t-1}) = ?$$

The idea of "language model"

$$P(w_1, w_2, \ldots, w_n) = P(w_n \mid w_1, \ldots, w_{n-1}) P(w_{n-1} \mid w_1, \ldots, w_{n-2}) \\ \times P(w_{n-2} \mid w_1, \ldots, w_{n-3}) \times \cdots \times P(w_1)$$

What is the probability of the *next word*?

$$P(w_t \mid w_1, \ldots, w_{t-1}) = ?$$

Target Context

A slightly different formulation

A slightly different formulation

What is the probability of the *target word given the contexts around it*?

A slightly different formulation

What is the probability of the *target word given the contexts around it*?

The quick brown ____ jumps over the lazy dog.

A slightly different formulation

What is the probability of the *target word given the contexts around it*?

The quick brown ____ *jumps over the lazy dog.*

A slightly different formulation

What is the probability of the *target word given the contexts around it*?

The quick brown ____ *jumps over the lazy dog.*

A slightly different formulation

What is the probability of the *target word given the contexts around it?*

The quick brown _____ *jumps over the lazy dog.*

Context

A slightly different formulation

What is the probability of the *target word given the contexts around it*?

The quick brown _____ *jumps over the lazy dog.*

Context Context

A slightly different formulation

What is the probability of the *target word given the contexts around it?*

The quick brown ____ *jumps over the lazy dog.*

Context | Target | Context

Even the most sophisticated methods are still rooted in this simple core idea.

BERT:

predict the masked word

The quick brown ____ *jumps over the lazy dog.*

Context — Target — Context

GPT:

predict the next word

The quick brown fox jumps ____

Context — Target

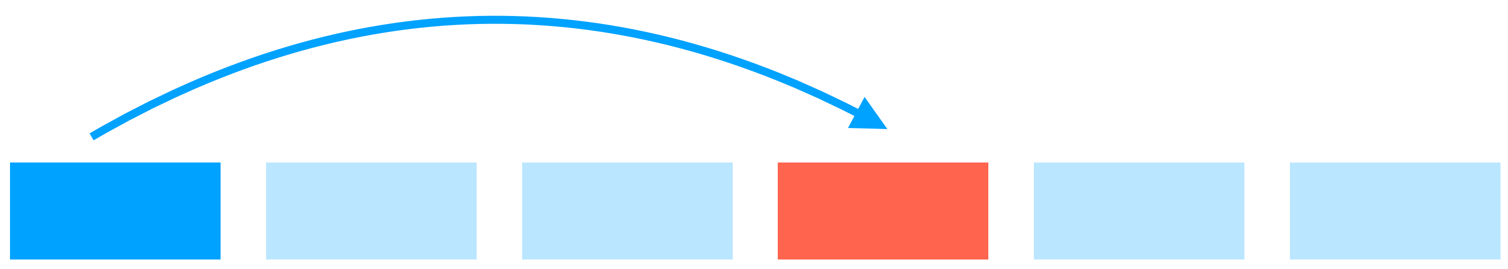
word2vec: Skip-gram model (single-word context)

$$P(w_t \mid w_{t-n}, \ldots, w_{t-1}) = \ ?$$

Can we just think about one word at a time ("**skipping**" the others)?

word2vec: Skip-gram model (single-word context)

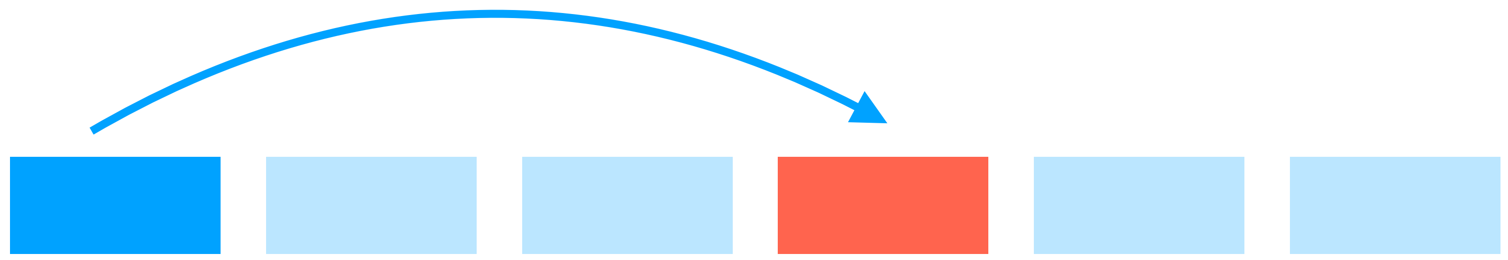
$$P(w_t \mid w_{t-n}, \ldots, w_{t-1}) = \ ?$$



Can we just think about one word at a time ("**skipping**" the others)?

word2vec: Skip-gram model (single-word context)

$$P(w_t \mid w_{t-n}, \ldots, w_{t-1}) = ?$$



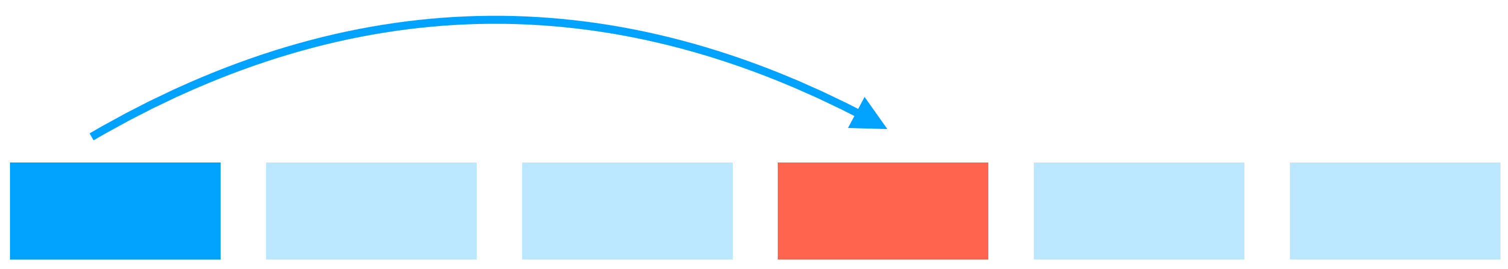
Can we just think about one word at a time ("**skipping**" the others)?

$$P(w_t \mid w_c) = ?$$

A single context word from the n-gram window

word2vec: Skip-gram model (single-word context)

$$P(w_t \mid w_{t-n}, \ldots, w_{t-1}) = ?$$



Can we just think about one word at a time ("**skipping**" the others)?

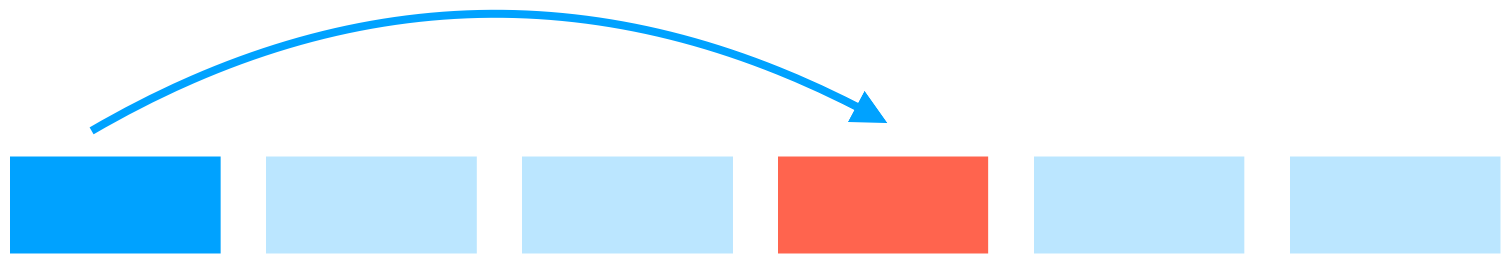
$$P(w_t \mid w_c) = ?$$

A single context word from the n-gram window

$$P(w_t \mid \textbf{context}) \approx \prod_{c \in C} P(w_t \mid w_c)$$

word2vec: Skip-gram model (single-word context)

$$P(w_t \mid w_{t-n}, \ldots, w_{t-1}) = ?$$



Can we just think about one word at a time ("**skipping**" the others)?

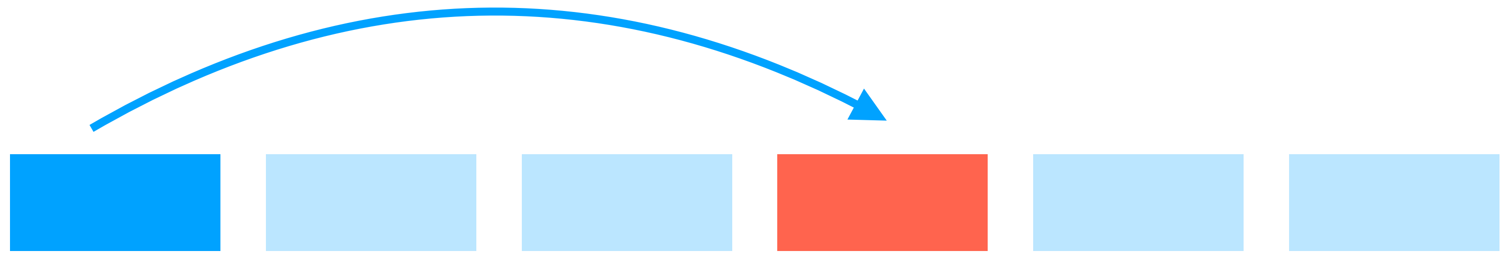
$$P(w_t \mid w_c) = ?$$

A single context word from the n-gram window

$$P(w_t \mid \textbf{context}) \approx \prod_{c \in C} P(w_t \mid w_c)$$

$$C = \{t - w, \ldots, t - 1, t + 1, \ldots, t + w\}$$

word2vec: Skip-gram model (single-word context)



$$P(w_t \mid \textbf{context}) \approx \prod_{c \in C} P(w_t \mid w_c)$$

$C = \{t-w, \ldots, t-1,\ t+1, \ldots, t+w\}$

$$P(w_1, \ldots, w_n) \approx \prod_t \prod_{c \in C} P(w_t \mid w_c)$$

$$\text{Maximize:} \quad \boxed{\frac{1}{T} \sum_t^T \sum_{c \in C} \log P(w_t \mid w_c)}$$

word2vec: using two vectors to evaluate the language model

$$\frac{1}{T}\sum_{t}^{T}\sum_{c \in C} \log P(w_t | w_c)$$

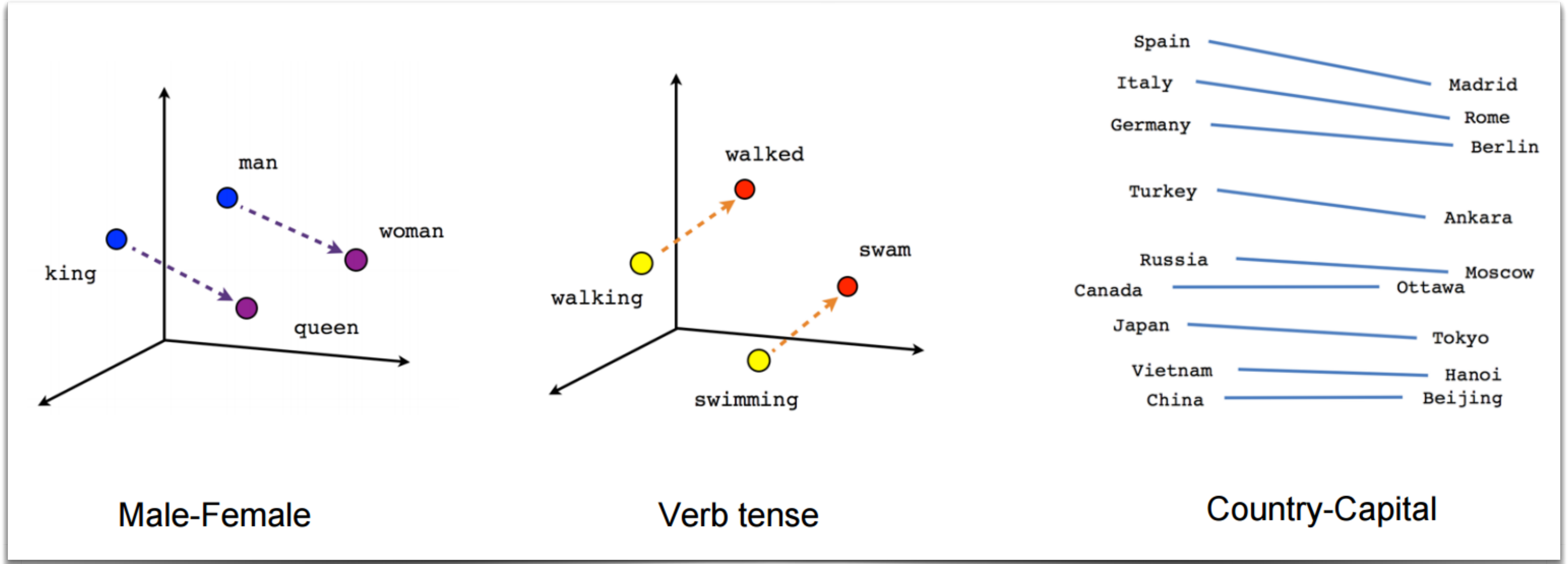
$$P(w_t | w_c) = ?$$

Let's assume that we have really good (two) vector representations for each word.

$(\mathbf{q}_i, \mathbf{k}_i)$ Each word has a '**query**' and a '**key**' vector that approximates the conditional probability. $P(w_t | w_c) \approx f(\mathbf{k}_t, \mathbf{q}_c)$

A Simple Choice: $$\boxed{P(w_t | w_c) \approx \frac{\exp(\mathbf{k}_t \cdot \mathbf{q}_c)}{\sum_i \exp(\mathbf{k}_i \cdot \mathbf{q}_c)}}$$

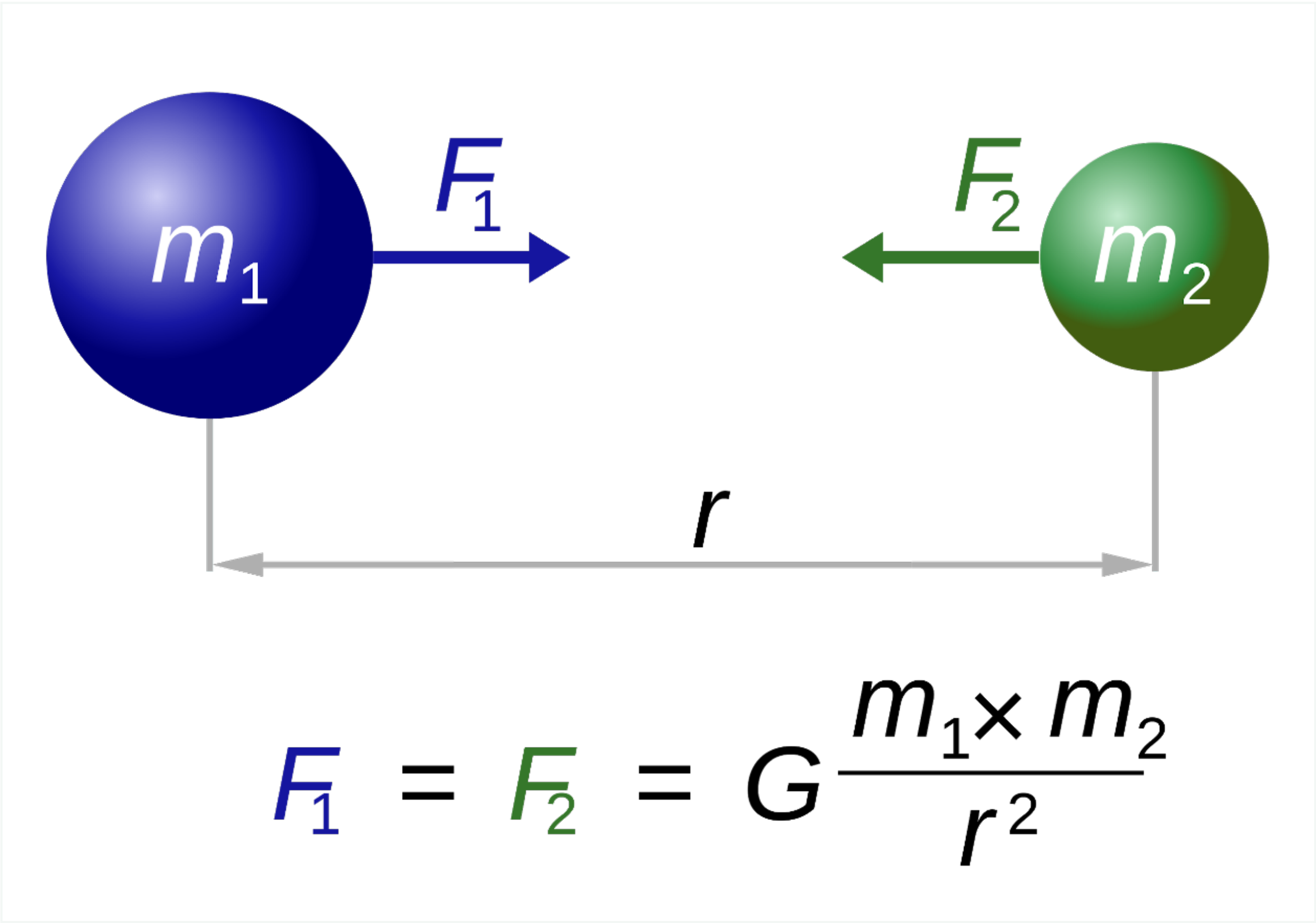
word2vec



Representations that produce a good language model also capture "meaning"!

Correspondence between word2vec model and the gravity law of mobility

Gravity law of mobility



Wikipedia user [Dennis Nilsson]

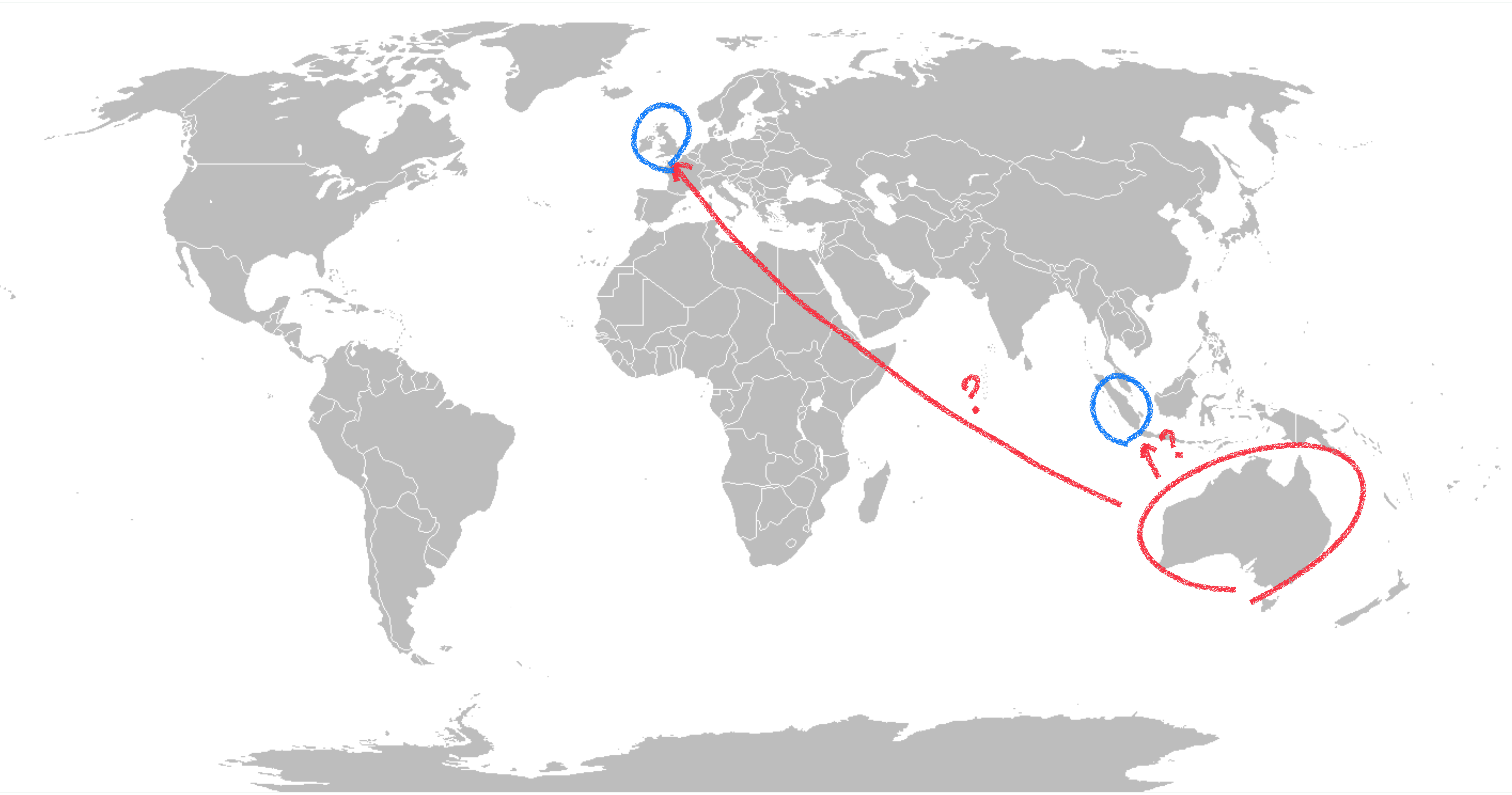
"You are less likely to go somewhere farther away than somewhere close."

$$\hat{T}_{ij} = C m_i m_j f(r_{ij})$$

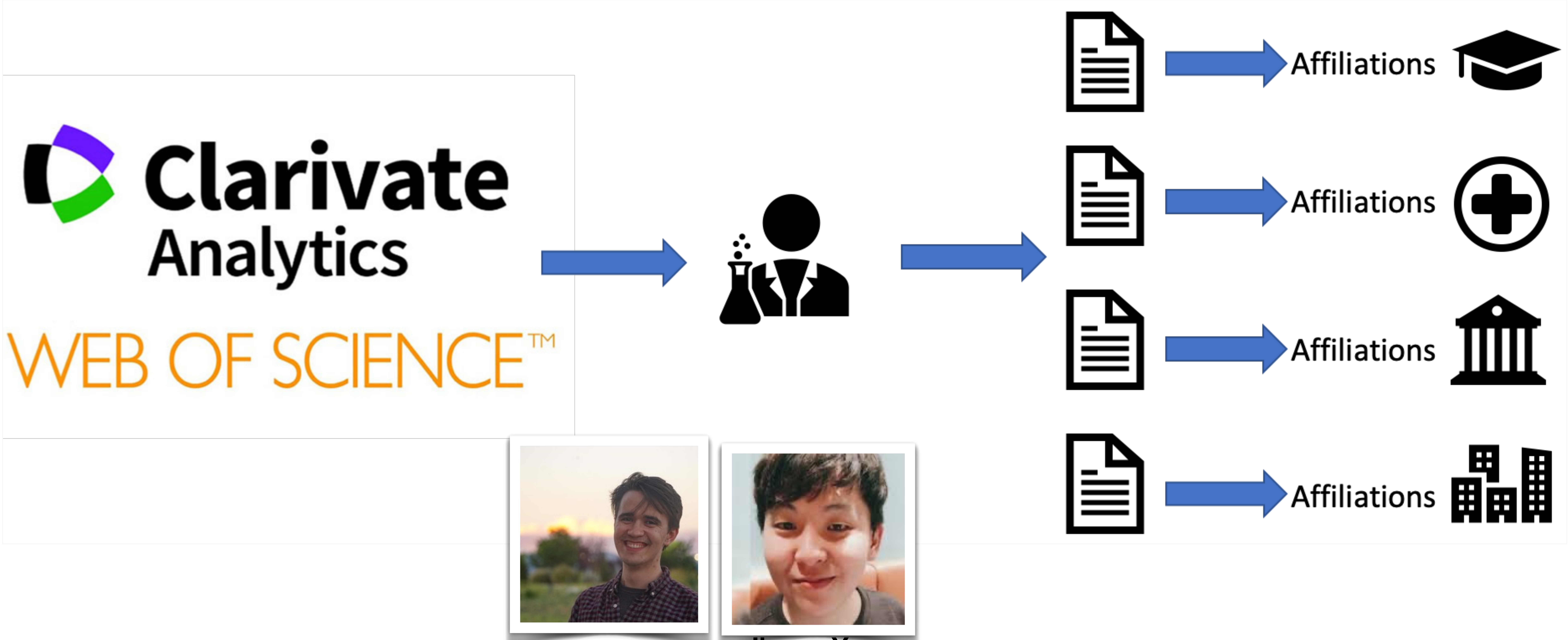
Flux

Population

a decaying function

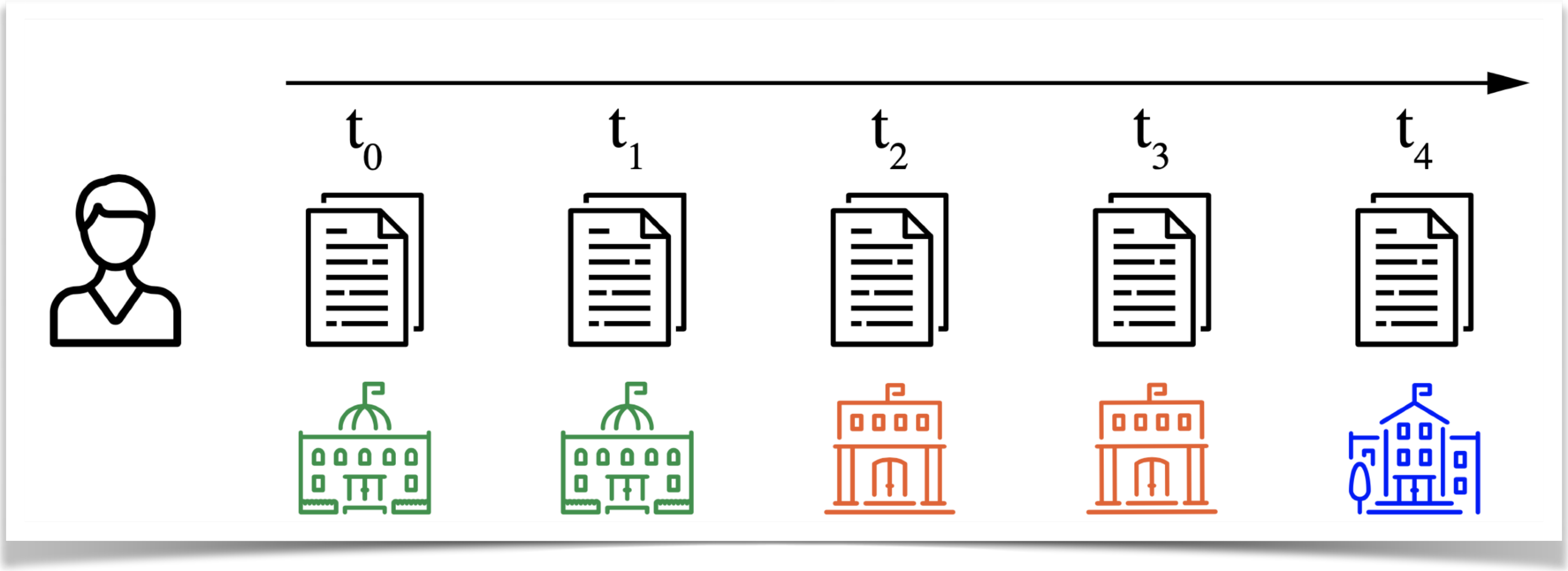


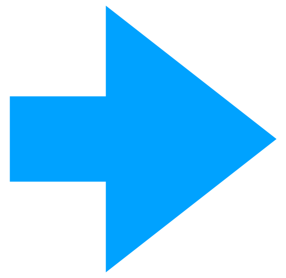
Data: Scientific mobility (2008 - 2019), and several others



Dakota Murray
Jisung Yoon

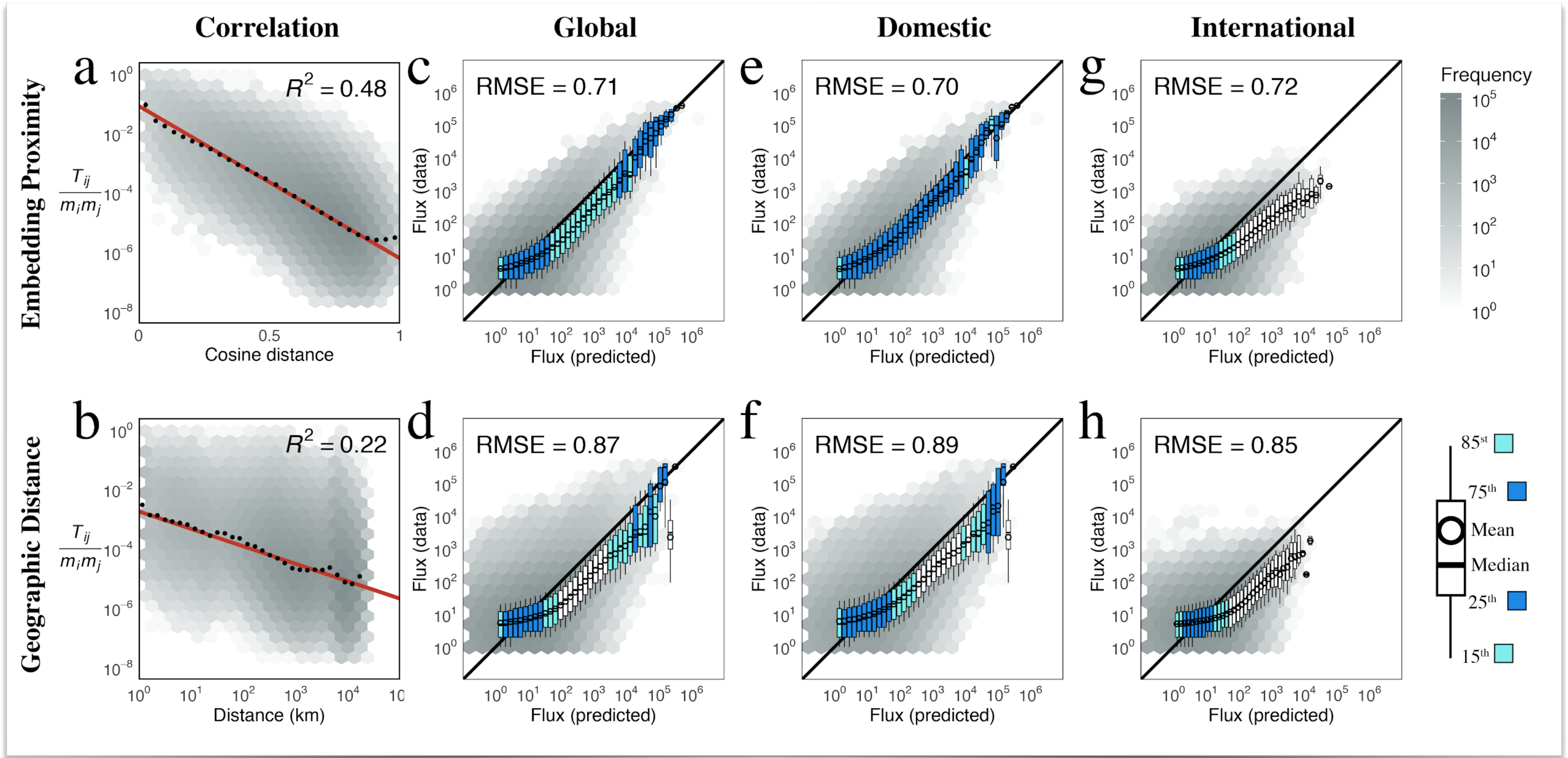
Derive flux between organizations from scientists’ trajectories

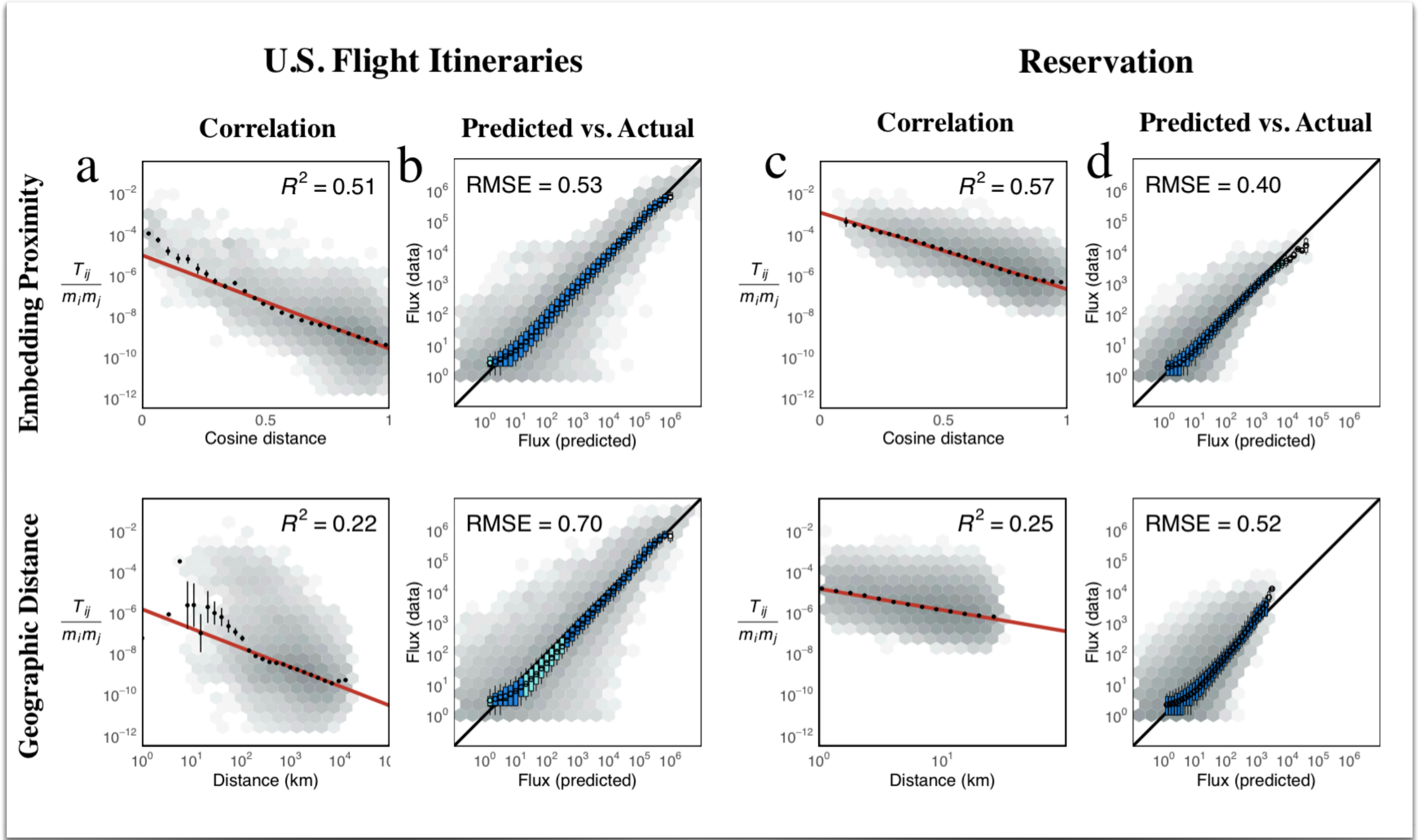




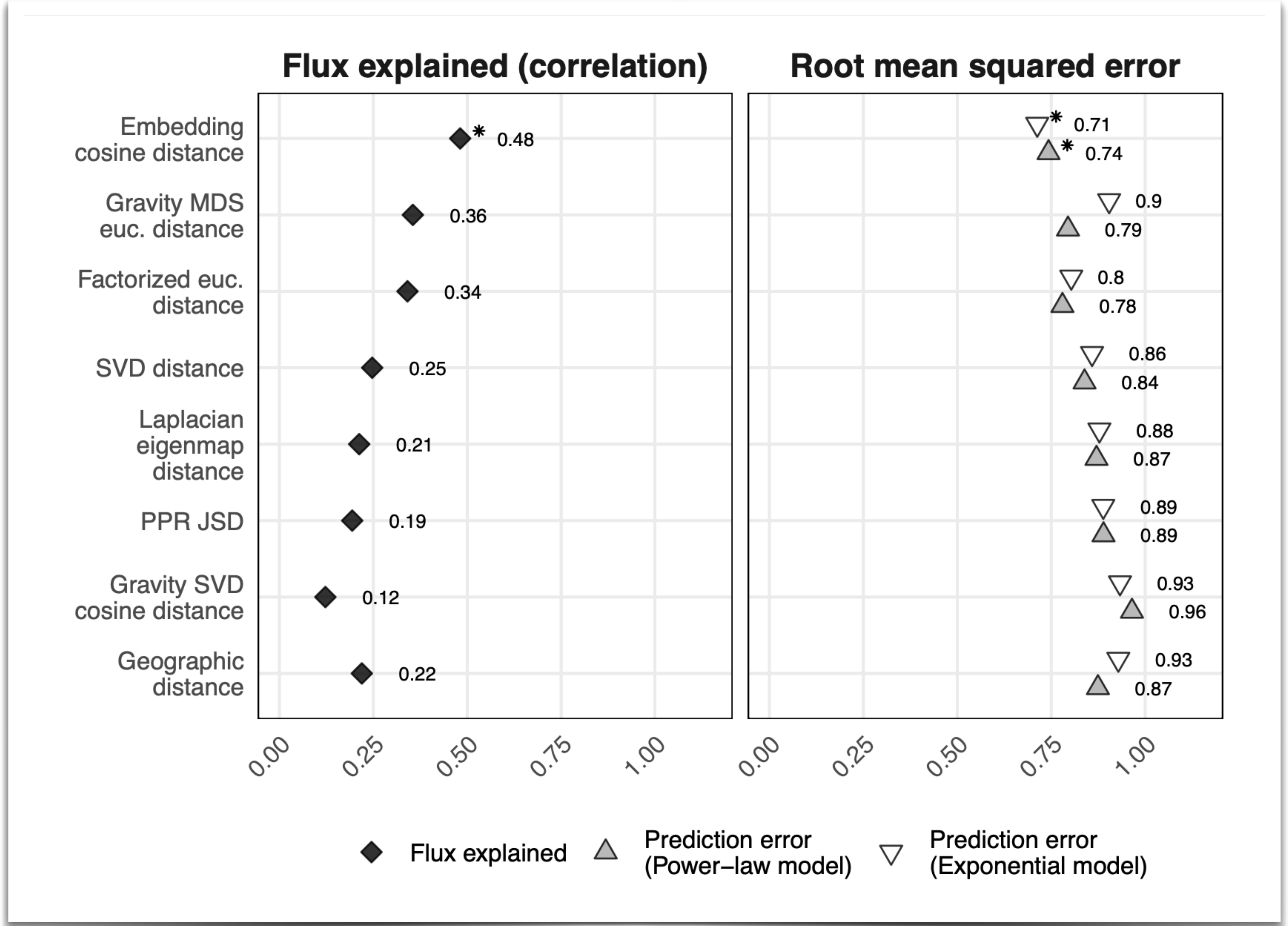
word2vec

Does this embedding better explains flows than geographic distance? **Yes!**




U.S. Flight Itineraries
Reservation
Correlation
Predicted vs. Actual
Correlation
Predicted vs. Actual
Embedding Proximity
Geographic Distance
a
b
c
d
$R^2 = 0.51$
RMSE = 0.53
$R^2 = 0.57$
RMSE = 0.40
$R^2 = 0.22$
RMSE = 0.70
$R^2 = 0.25$
RMSE = 0.52
$\frac{T_{ij}}{m_i m_j}$
Cosine distance
Distance (km)
Flux (data)
Flux (predicted)

Embedding explains the flux best



Murray, Dakota, Jisung Yoon, Sadamori Kojaku, Rodrigo Costas, Woo-Sung Jung, Staša Milojević, and Yong-Yeol Ahn. "Unsupervised embedding of trajectories captures the latent structure of mobility." *arXiv preprint arXiv:2012.02785* (2020).

Why?

Let's go back to the word2vec model

$(\mathbf{q}_i, \mathbf{k}_i)$ Each Word Has a 'Query' and a 'Key' Vector
That Approximates the Conditional Probability. $P(w_t \mid w_c) \approx f(\mathbf{k}_t, \mathbf{q}_c)$

A Simple Choice: $$P(w_t \mid w_c) \approx \frac{\exp(\mathbf{k}_t \cdot \mathbf{q}_c)}{\sum_i \exp(\mathbf{k}_i \cdot \mathbf{q}_c)}$$

$$P(w_t \mid w_c) \approx \frac{\exp(\mathbf{k}_t \cdot \mathbf{q}_c)}{\sum_i \exp(\mathbf{k}_i \cdot \mathbf{q}_c)}$$

nasty!

$$P(w_t \mid w_c) \approx \frac{\exp(\mathbf{k}_t \cdot \mathbf{q}_c)}{\sum_i \exp(\mathbf{k}_i \cdot \mathbf{q}_c)}$$

nasty!

Negative sampling! Let's formulate a classification task.

$$P(w_t \mid w_c) \approx \frac{\exp(\mathbf{k}_t \cdot \mathbf{q}_c)}{\sum_i \exp(\mathbf{k}_i \cdot \mathbf{q}_c)}$$

nasty!

Negative sampling! Let's formulate a classification task.

$$P^{\mathrm{NS}}(Y_j = 1; \boldsymbol{v}_i, \boldsymbol{u}_j) = \frac{1}{1 + \exp(-\boldsymbol{u}_j \cdot \boldsymbol{v}_i)},$$

$$P(w_t | w_c) \approx \frac{\exp(\mathbf{k}_t \cdot \mathbf{q}_c)}{\sum_i \exp(\mathbf{k}_i \cdot \mathbf{q}_c)}$$

nasty!

Negative sampling! Let's formulate a classification task.

$$P^{\mathrm{NS}}(Y_j = 1; \boldsymbol{v}_i, \boldsymbol{u}_j) = \frac{1}{1 + \exp(-\boldsymbol{u}_j \cdot \boldsymbol{v}_i)},$$

$$\mathcal{J}^{\mathrm{NS}} = \sum_{i \in \mathcal{A}} \sum_{j \in \mathcal{D}} \left[Y_j \log P^{\mathrm{NS}}(Y_j = 1; \boldsymbol{v}_i, \boldsymbol{u}_j) + (1 - Y_j) \log P^{\mathrm{NS}}(Y_j = 0; \boldsymbol{v}_i, \boldsymbol{u}_j) \right],$$

Noise contrastive estimation and negative sampling

“Noise Contrastive Estimation” [Gutmann & Hyvärinen, 2010], an unbiased estimator, is subtly different.

$$P^{\mathrm{NS}}(Y_j = 1; \boldsymbol{v}_i, \boldsymbol{u}_j) = \frac{1}{1 + \exp(-\boldsymbol{u}_j \cdot \boldsymbol{v}_i)},$$

$$\mathcal{J}^{\mathrm{NS}} = \sum_{i \in \mathcal{A}} \sum_{j \in \mathcal{D}} \left[Y_j \log P^{\mathrm{NS}}(Y_j = 1; \boldsymbol{v}_i, \boldsymbol{u}_j) + (1 - Y_j) \log P^{\mathrm{NS}}(Y_j = 0; \boldsymbol{v}_i, \boldsymbol{u}_j) \right],$$

Noise contrastive estimation and negative sampling

"Noise Contrastive Estimation" [Gutmann & Hyvärinen, 2010], an unbiased estimator, is subtly different.

$$P^{\mathrm{NCE}}\left(Y_j = 1|j\right) = \frac{1}{1 + \exp\left[-\ln f(\boldsymbol{u}_j \cdot \boldsymbol{v}_i) + \ln p_0(j) + c\right]},$$

$$P^{\mathrm{NS}}(Y_j = 1; \boldsymbol{v}_i, \boldsymbol{u}_j) = \frac{1}{1 + \exp(-\boldsymbol{u}_j \cdot \boldsymbol{v}_i)},$$

$$\mathcal{J}^{\mathrm{NS}} = \sum_{i \in \mathcal{A}} \sum_{j \in \mathcal{D}} \left[Y_j \log P^{\mathrm{NS}}(Y_j = 1; \boldsymbol{v}_i, \boldsymbol{u}_j) + (1 - Y_j) \log P^{\mathrm{NS}}(Y_j = 0; \boldsymbol{v}_i, \boldsymbol{u}_j)\right],$$

Noise contrastive estimation and negative sampling

“Noise Contrastive Estimation” [Gutmann & Hyvärinen, 2010], an unbiased estimator, is subtly different.

$$P^{\mathrm{NCE}}\left(Y_j = 1|j\right) = \frac{1}{1 + \exp\left[-\ln f(\boldsymbol{u}_j \cdot \boldsymbol{v}_i) + \ln p_0(j) + c\right]},$$

$$\mathcal{J}^{\mathrm{NCE}} = \sum_{i \in \mathcal{A}} \sum_{j \in \mathcal{D}} \left[Y_j \log P^{\mathrm{NCE}}(Y_j = 1|j) + (1 - Y_j) \log P^{\mathrm{NCE}}(Y_j = 0|j)\right].$$

$$P^{\mathrm{NS}}(Y_j = 1; \boldsymbol{v}_i, \boldsymbol{u}_j) = \frac{1}{1 + \exp(-\boldsymbol{u}_j \cdot \boldsymbol{v}_i)},$$

$$\mathcal{J}^{\mathrm{NS}} = \sum_{i \in \mathcal{A}} \sum_{j \in \mathcal{D}} \left[Y_j \log P^{\mathrm{NS}}(Y_j = 1; \boldsymbol{v}_i, \boldsymbol{u}_j) + (1 - Y_j) \log P^{\mathrm{NS}}(Y_j = 0; \boldsymbol{v}_i, \boldsymbol{u}_j)\right],$$

Noise contrastive estimation and negative sampling

"Noise Contrastive Estimation" [Gutmann & Hyvärinen, 2010], an unbiased estimator, is subtly different.

$$P^{\mathrm{NCE}}\left(Y_j = 1 | j\right) = \frac{1}{1 + \exp\left[-\ln f(\boldsymbol{u}_j \cdot \boldsymbol{v}_i) + \ln p_0(j) + c\right]},$$

$$\mathcal{J}^{\mathrm{NCE}} = \sum_{i \in \mathcal{A}} \sum_{j \in \mathcal{D}} \left[Y_j \log P^{\mathrm{NCE}}(Y_j = 1|j) + (1 - Y_j) \log P^{\mathrm{NCE}}(Y_j = 0|j)\right].$$

$$P^{\mathrm{NS}}(Y_j = 1; \boldsymbol{v}_i, \boldsymbol{u}_j) = \frac{1}{1 + \exp(-\boldsymbol{u}_j \cdot \boldsymbol{v}_i)},$$

$$\mathcal{J}^{\mathrm{NS}} = \sum_{i \in \mathcal{A}} \sum_{j \in \mathcal{D}} \left[Y_j \log P^{\mathrm{NS}}(Y_j = 1; \boldsymbol{v}_i, \boldsymbol{u}_j) + (1 - Y_j) \log P^{\mathrm{NS}}(Y_j = 0; \boldsymbol{v}_i, \boldsymbol{u}_j)\right],$$

SGNS word2vec actually optimzes...

Sadamori Kojaku

$$P^{NS}(j \mid i) = P_m^{NS}(\boldsymbol{u}_j \cdot \boldsymbol{v}_i) = \frac{P^{\gamma}(j)\exp(\boldsymbol{u}_j \cdot \boldsymbol{v}_i)}{Z'_i},$$

SGNS word2vec actually optimzes…

$$P(w_t \mid w_c) \approx \frac{\exp(\mathbf{k}_t \cdot \mathbf{q}_c)}{\sum_i \exp(\mathbf{k}_i \cdot \mathbf{q}_c)}$$

Sadamori Kojaku

$$P^{NS}(j \mid i) = P_m^{NS}(\boldsymbol{u}_j \cdot \boldsymbol{v}_i) = \frac{P^{\gamma}(j)\exp(\boldsymbol{u}_j \cdot \boldsymbol{v}_i)}{Z_i'},$$

SGNS word2vec actually optimzes...

$$\cancel{P(w_t | w_c) \propto \frac{\exp(\mathbf{k}_t \cdot \mathbf{q}_c)}{\sum_i \exp(\mathbf{k}_i \cdot \mathbf{q}_c)}}$$

Sadamori Kojaku

$$P^{NS}(j \mid i) = P_m^{NS}(\boldsymbol{u}_j \cdot \boldsymbol{v}_i) = \frac{P^{\gamma}(j)\exp(\boldsymbol{u}_j \cdot \boldsymbol{v}_i)}{Z_i'},$$

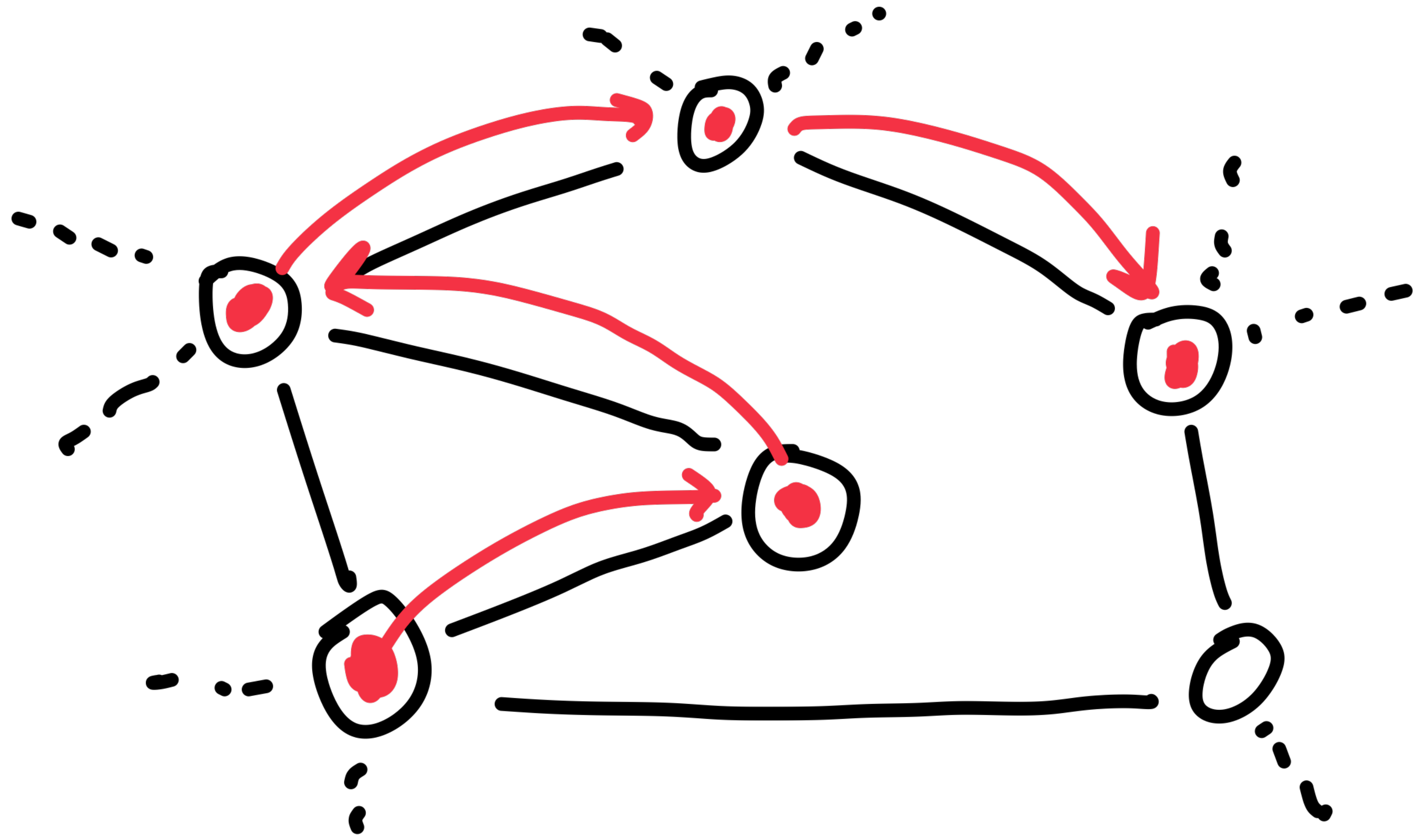
SGNS word2vec actually optimzes...

$$\cancel{P(w_t | w_c) \propto \frac{\exp(\mathbf{k}_t \cdot \mathbf{q}_c)}{\sum_i \exp(\mathbf{k}_i \cdot \mathbf{q}_c)}}$$

Sadamori Kojaku

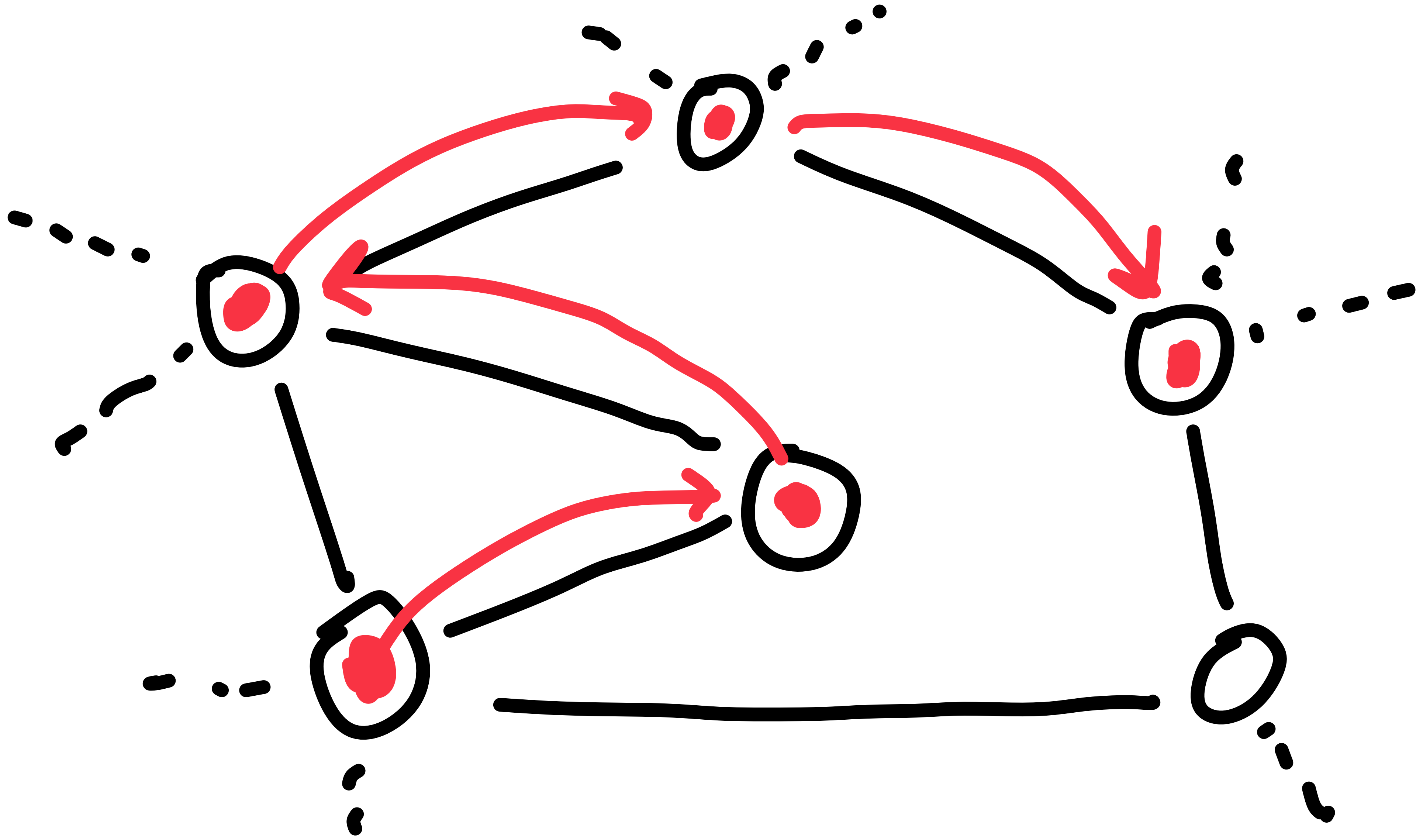
$$P^{NS}(j \mid i) = P_m^{NS}(\boldsymbol{u}_j \cdot \boldsymbol{v}_i) = \frac{P^{\gamma}(j) \exp(\boldsymbol{u}_j \cdot \boldsymbol{v}_i)}{Z_i'},$$

What happens if we apply word2vec to mobility trajectories?



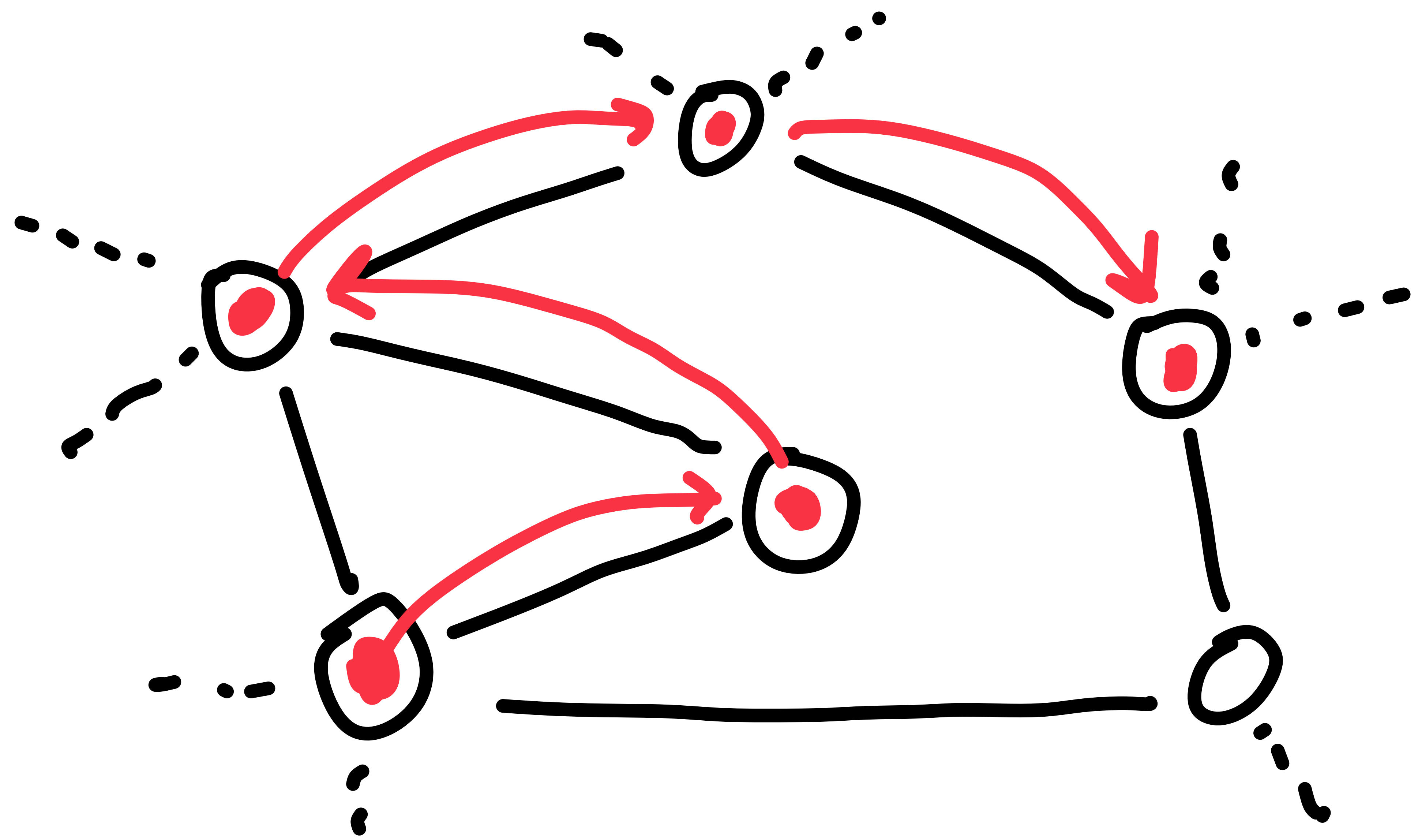
What happens if we apply word2vec to mobility trajectories?

$$P(w_t \mid w_c) \approx \frac{p_n(t)\exp(\mathbf{k}_t \cdot \mathbf{q}_c)}{\sum_i p_n(i)\exp(\mathbf{k}_i \cdot \mathbf{q}_c)}$$



What happens if we apply word2vec to mobility trajectories?

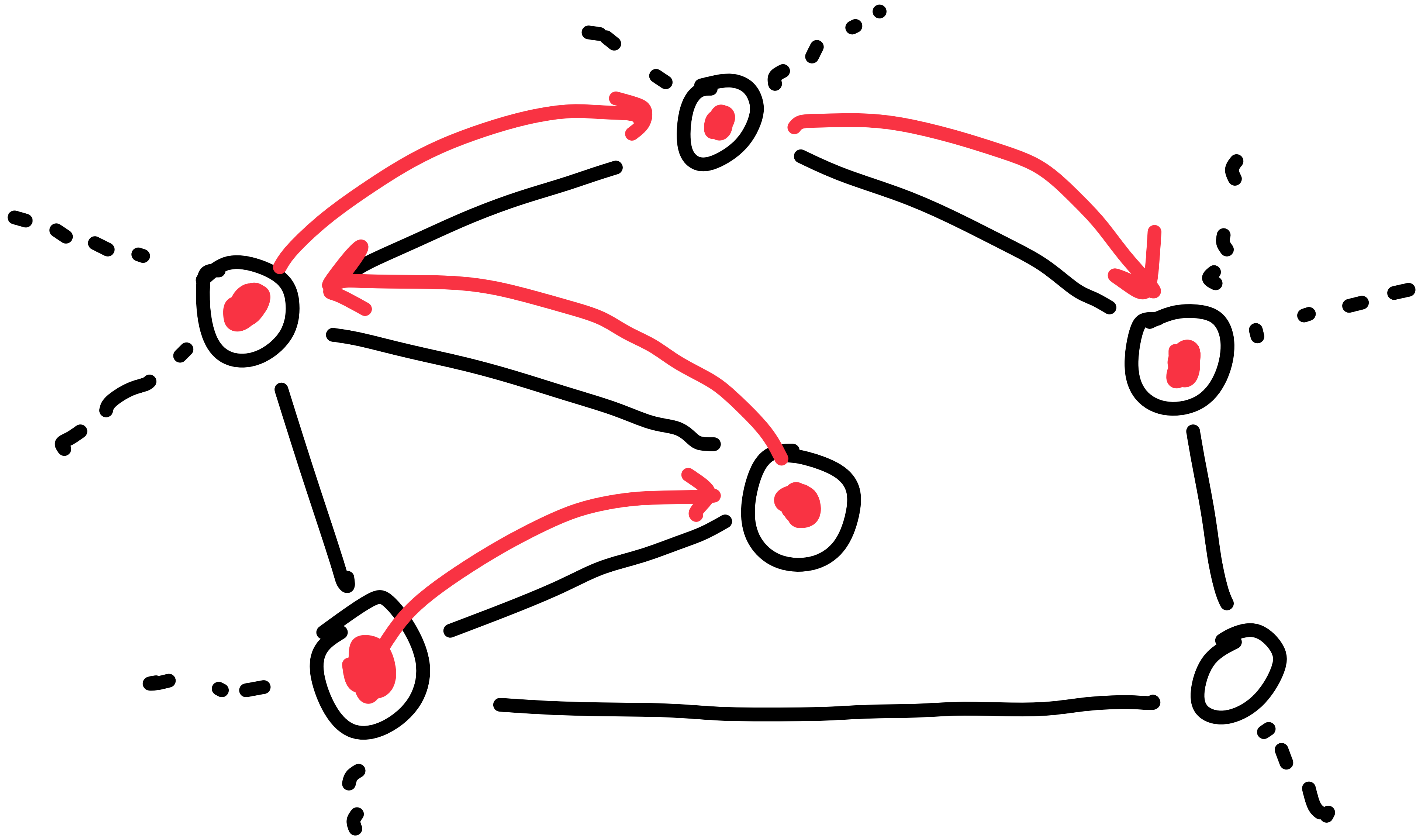
$$P(w_t \mid w_c) \approx \frac{p_n(t)\exp(\mathbf{k}_t \cdot \mathbf{q}_c)}{\sum_i p_n(i)\exp(\mathbf{k}_i \cdot \mathbf{q}_c)}$$



$$\hat{T}_{ij} \propto P(j \mid i)P(i) \propto \frac{P(i)P(j)\exp(\mathbf{k}_j \cdot \mathbf{q}_i)}{\sum_{j'} P(j')\exp(\mathbf{k}_{j'} \cdot \mathbf{q}_i)}$$

What happens if we apply word2vec to mobility trajectories?

$$P(w_t \mid w_c) \approx \frac{p_n(t)\exp(\mathbf{k}_t \cdot \mathbf{q}_c)}{\sum_i p_n(i)\exp(\mathbf{k}_i \cdot \mathbf{q}_c)}$$

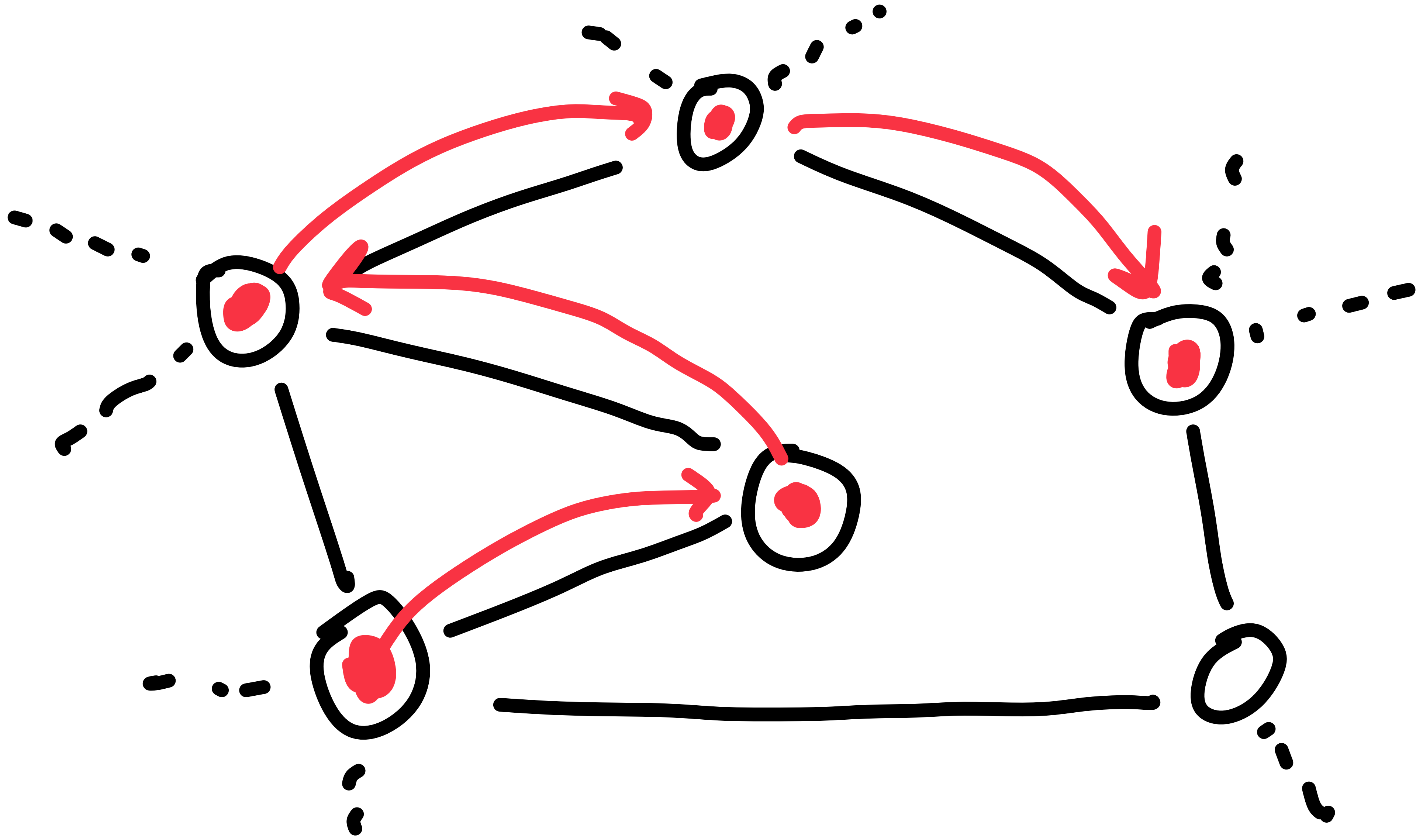


flux

$$\hat{T}_{ij} \propto P(j \mid i)P(i) \propto \frac{P(i)P(j)\exp(\mathbf{k}_j \cdot \mathbf{q}_i)}{\sum_{j'} P(j')\exp(\mathbf{k}_{j'} \cdot \mathbf{q}_i)}$$

What happens if we apply word2vec to mobility trajectories?

$$P(w_t | w_c) \approx \frac{p_n(t)\exp(\mathbf{k}_t \cdot \mathbf{q}_c)}{\sum_i p_n(i)\exp(\mathbf{k}_i \cdot \mathbf{q}_c)}$$



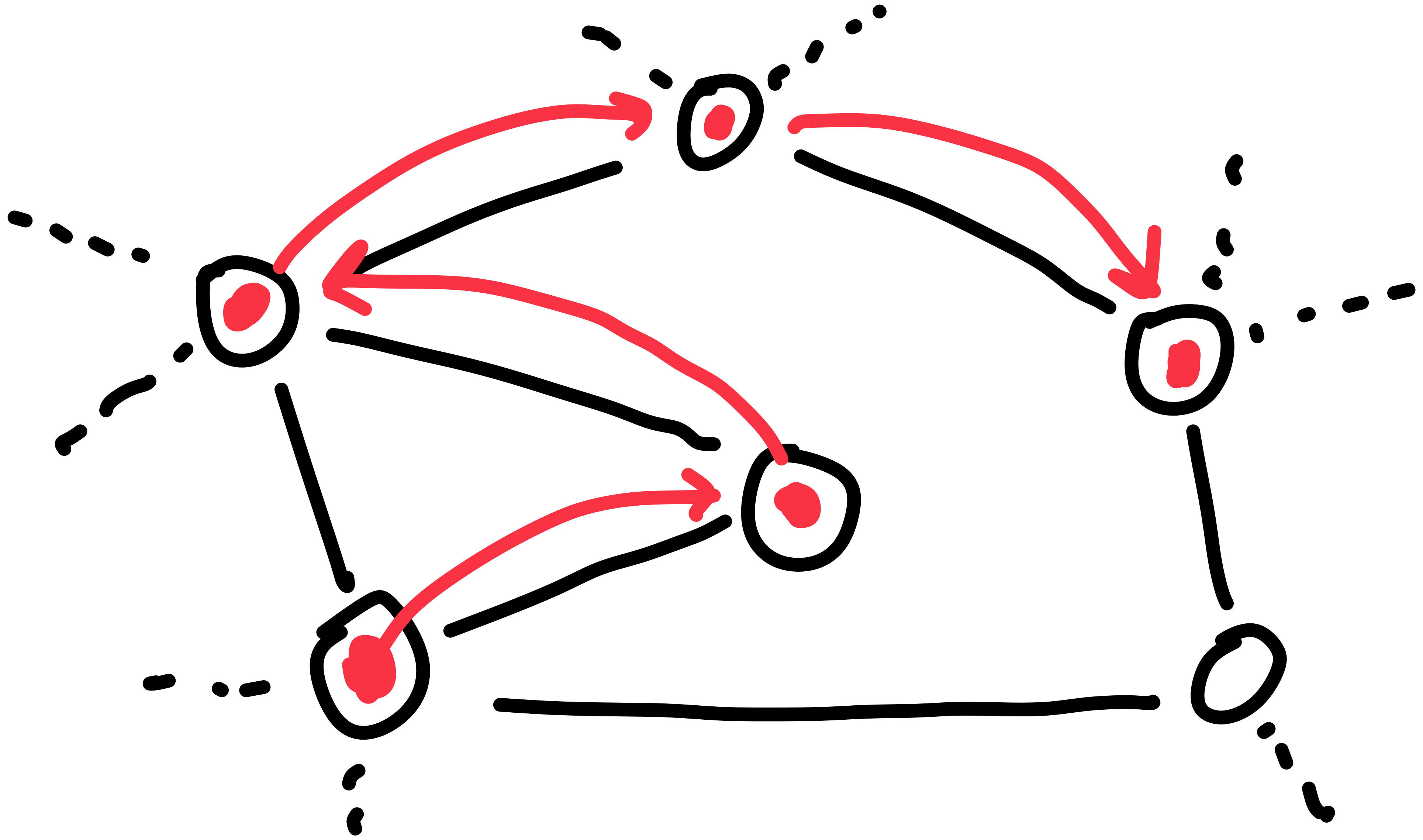
flux

$$\hat{T}_{ij} \propto P(j | i)P(i) \propto \frac{P(i)P(j)\exp(\mathbf{k}_j \cdot \mathbf{q}_i)}{\sum_{j'} P(j')\exp(\mathbf{k}_{j'} \cdot \mathbf{q}_i)}$$

when embedding dimension is sufficiently lar

What happens if we apply word2vec to mobility trajectories?

$$P(w_t \mid w_c) \approx \frac{p_n(t)\exp(\mathbf{k}_t \cdot \mathbf{q}_c)}{\sum_i p_n(i)\exp(\mathbf{k}_i \cdot \mathbf{q}_c)}$$



flux

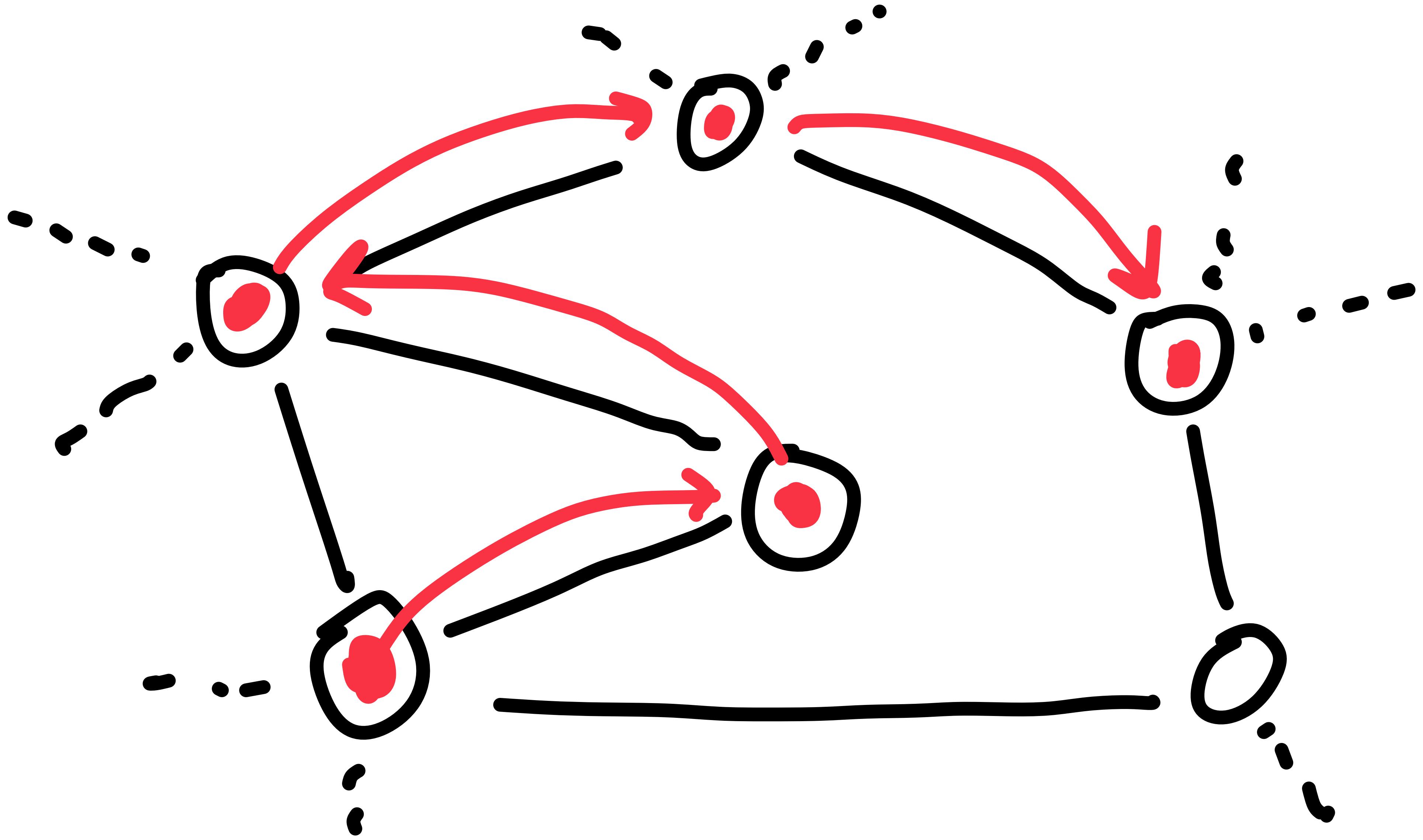
$$\hat{T}_{ij} \propto P(j \mid i)P(i) \propto \frac{P(i)P(j)\exp(\mathbf{k}_j \cdot \mathbf{q}_i)}{\sum_{j'} P(j')\exp(\mathbf{k}_{j'} \cdot \mathbf{q}_i)}$$

when embedding dimension is sufficiently lar

$$\hat{T}_{ij} = \hat{T}_{ji} \propto P(i)P(j)\exp(\mathbf{k}_i \cdot \mathbf{k}_j)$$

What happens if we apply word2vec to mobility trajectories?

$$P(w_t \mid w_c) \approx \frac{p_n(t)\exp(\mathbf{k}_t \cdot \mathbf{q}_c)}{\sum_i p_n(i)\exp(\mathbf{k}_i \cdot \mathbf{q}_c)}$$



flux

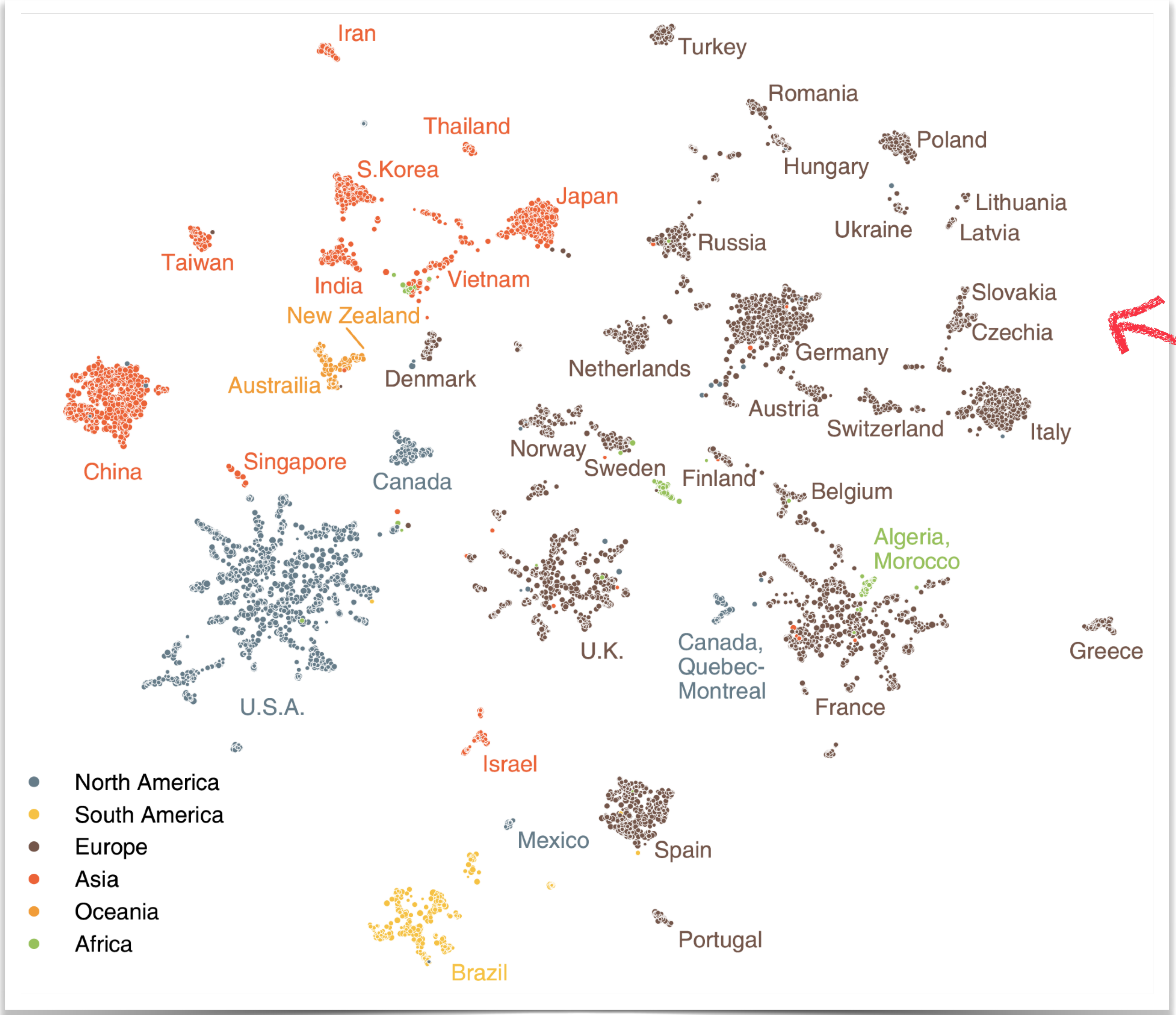
$$\hat{T}_{ij} \propto P(j \mid i)P(i) \propto \frac{P(i)P(j)\exp(\mathbf{k}_j \cdot \mathbf{q}_i)}{\sum_{j'} P(j')\exp(\mathbf{k}_{j'} \cdot \mathbf{q}_i)}$$

when embedding dimension is sufficiently large

$$\hat{T}_{ij} = \hat{T}_{ji} \propto P(i)P(j)\exp(\mathbf{k}_i \cdot \mathbf{k}_j)$$

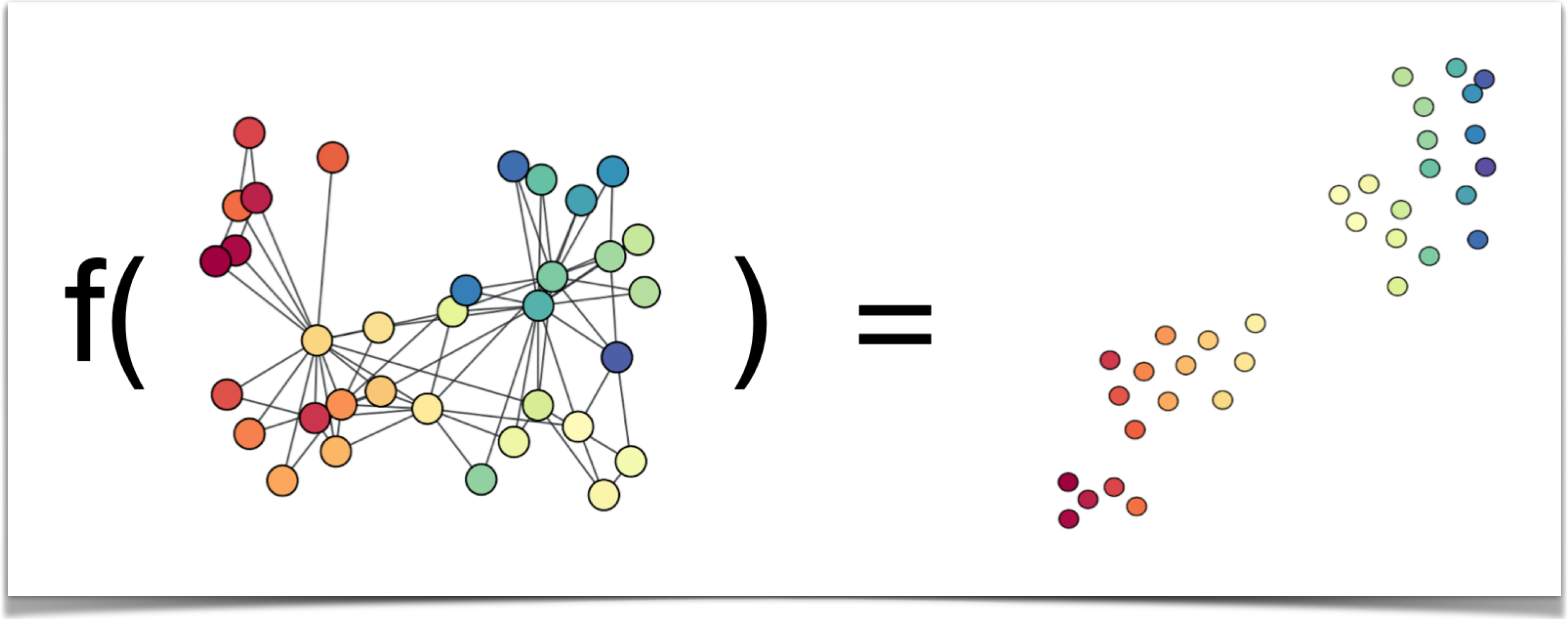
Gravity law!

word2vec model ~ gravity law



The space where the institutions are arranged so that the flux and distance between them satisfies the gravity law of mobility!

Implications in Graph Embedding



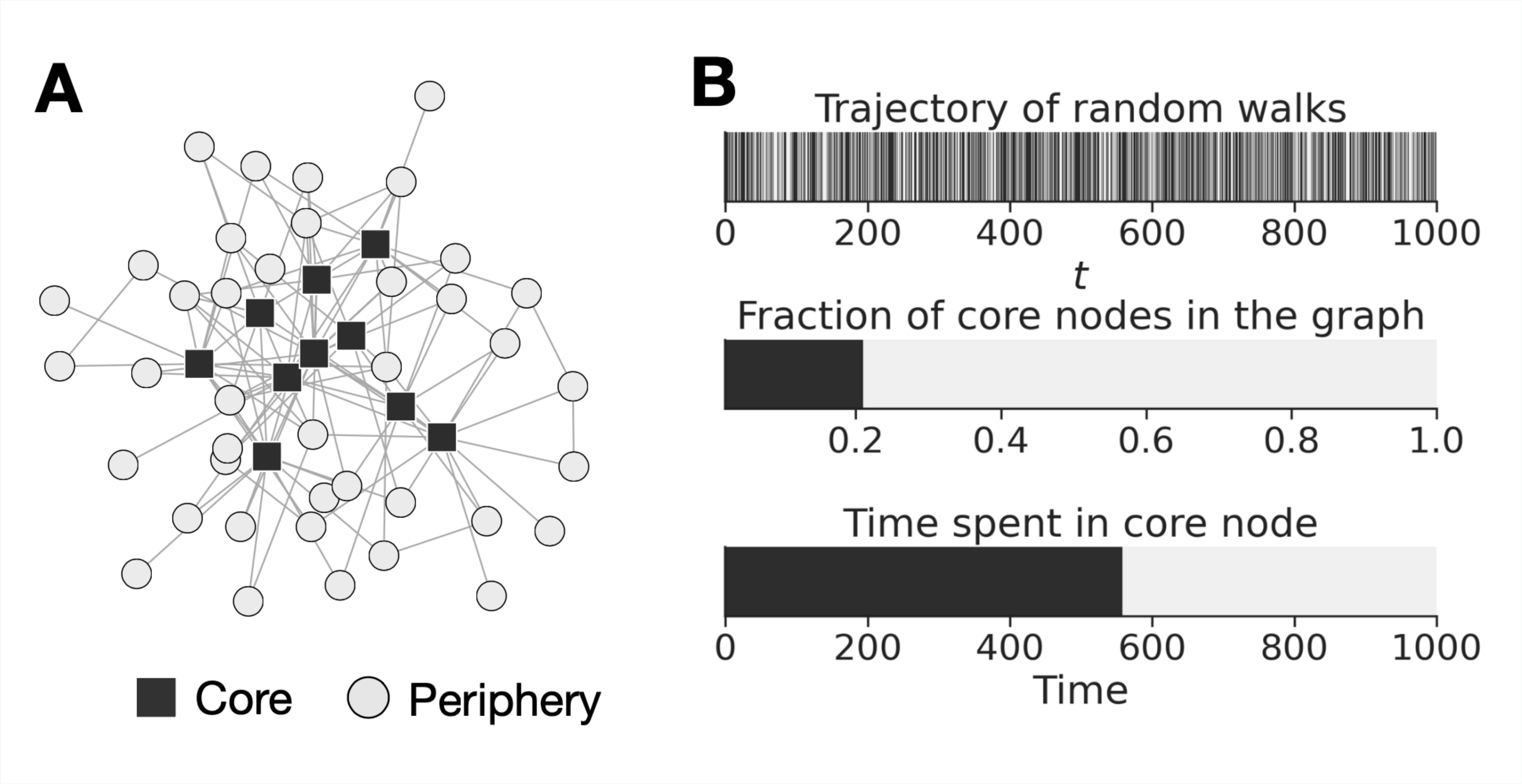
Random walk → “Sentences” (DeepWalk, node2vec, etc.)

Random walk is biased

Friendship paradox. When we follow an edge, the expected degree is proportional to the degree

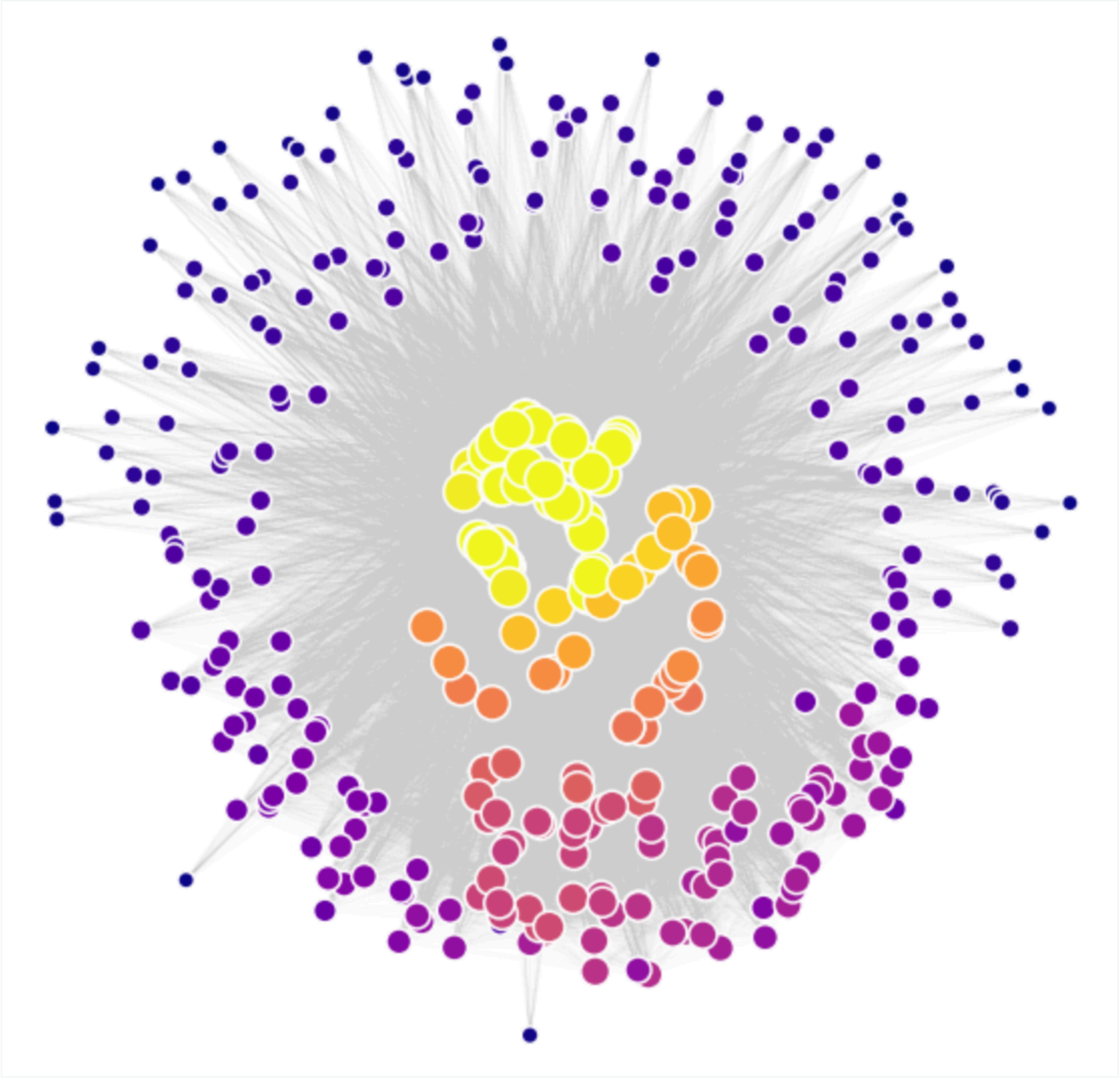
$$\sim p(k) \qquad\qquad \sim kp(k)$$

Random walk is biased



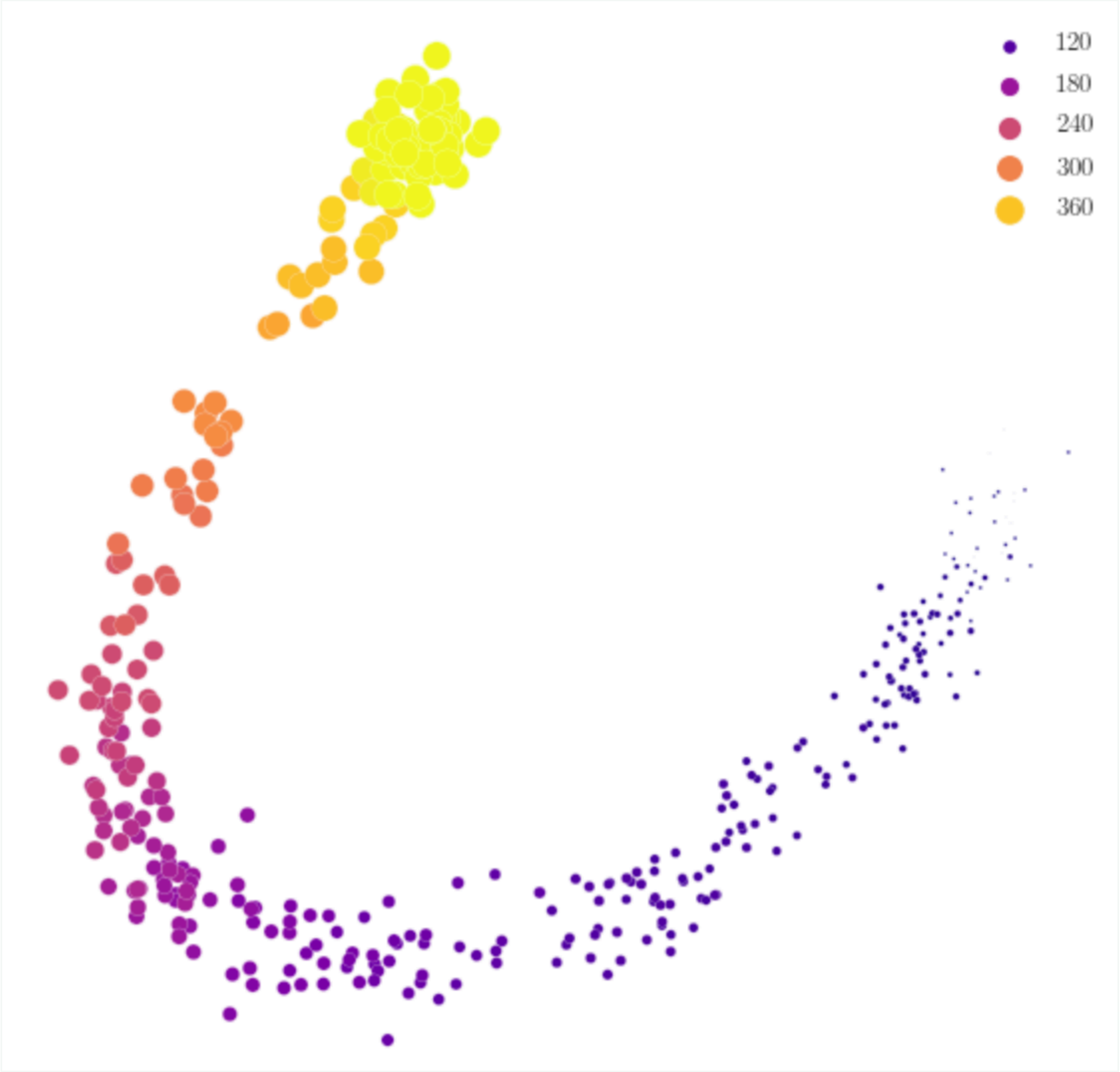
What's the Implication?

Random Walk Bias → Biased Embedding Space



DeepWalk

→



But word2vec's bias negates this random walk bias!

Recall $$P(w_t | w_c) \approx \frac{p_n(t)\exp(\mathbf{k}_t \cdot \mathbf{q}_c)}{\sum_i p_n(i)\exp(\mathbf{k}_i \cdot \mathbf{q}_c)}$$

If negative samples are proportionally sampled based on their degree, **SGNS negates the bias of the random walker**!

But word2vec's bias negates this random walk bias!

Recall $$P(w_t \mid w_c) \approx \frac{p_n(t)\exp(\mathbf{k}_t \cdot \mathbf{q}_c)}{\sum_i p_n(i)\exp(\mathbf{k}_i \cdot \mathbf{q}_c)}$$

If negative samples are proportionally sampled based on their degree, **SGNS negates the bias of the random walker**!

🤔 Can we remove other statistical biases as well?

Residual2vec

We can extract out the *expected conditional probability* based on a null model.

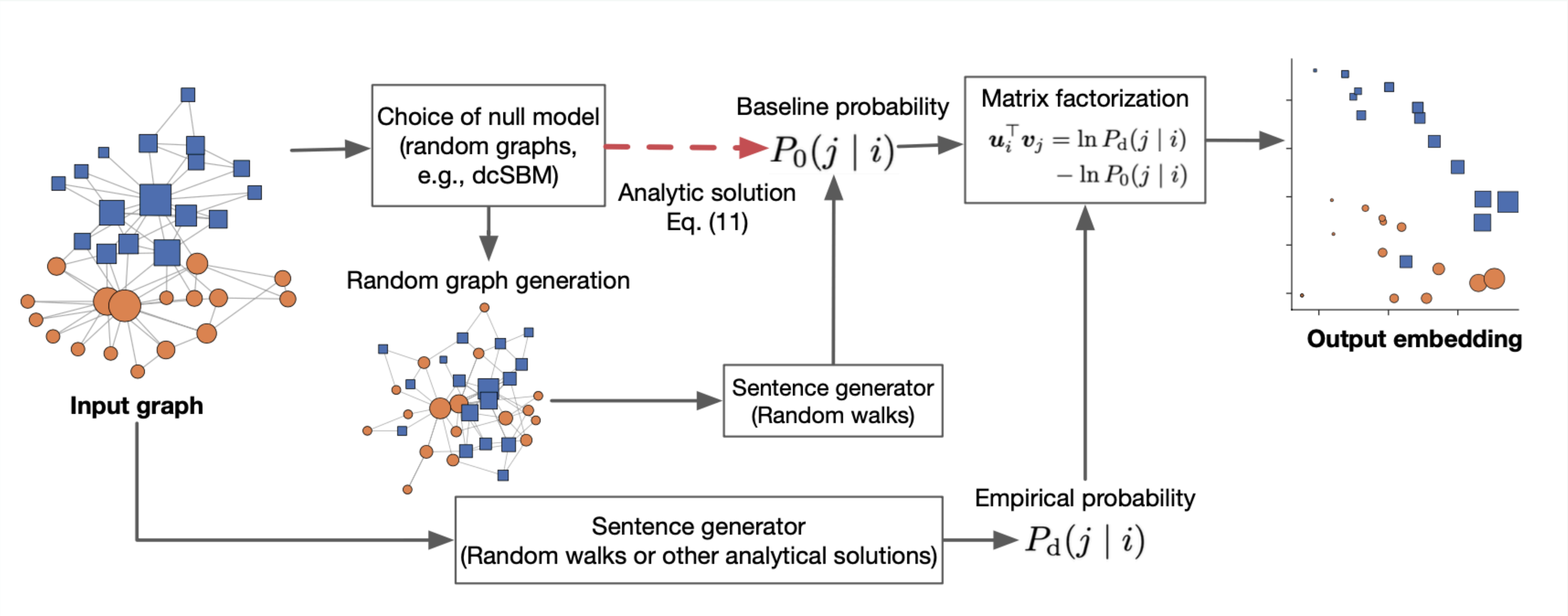
Sadamori Kojaku

$$P_{\mathrm{r2v}}(j \mid i) = \frac{\overbrace{P_0(j \mid i)}^{\text{null model}} \exp(\overbrace{\boldsymbol{u}_i^\top \boldsymbol{v}_j}^{\text{"residual" information not captured by the null model.}})}{Z_i'}$$

Sadamori Kojaku, Jisung Yoon, Isabel Constantino, Yong-Yeol, "Residual2Vec: Debiasing graph embedding with random graphs", NeurIPS'21

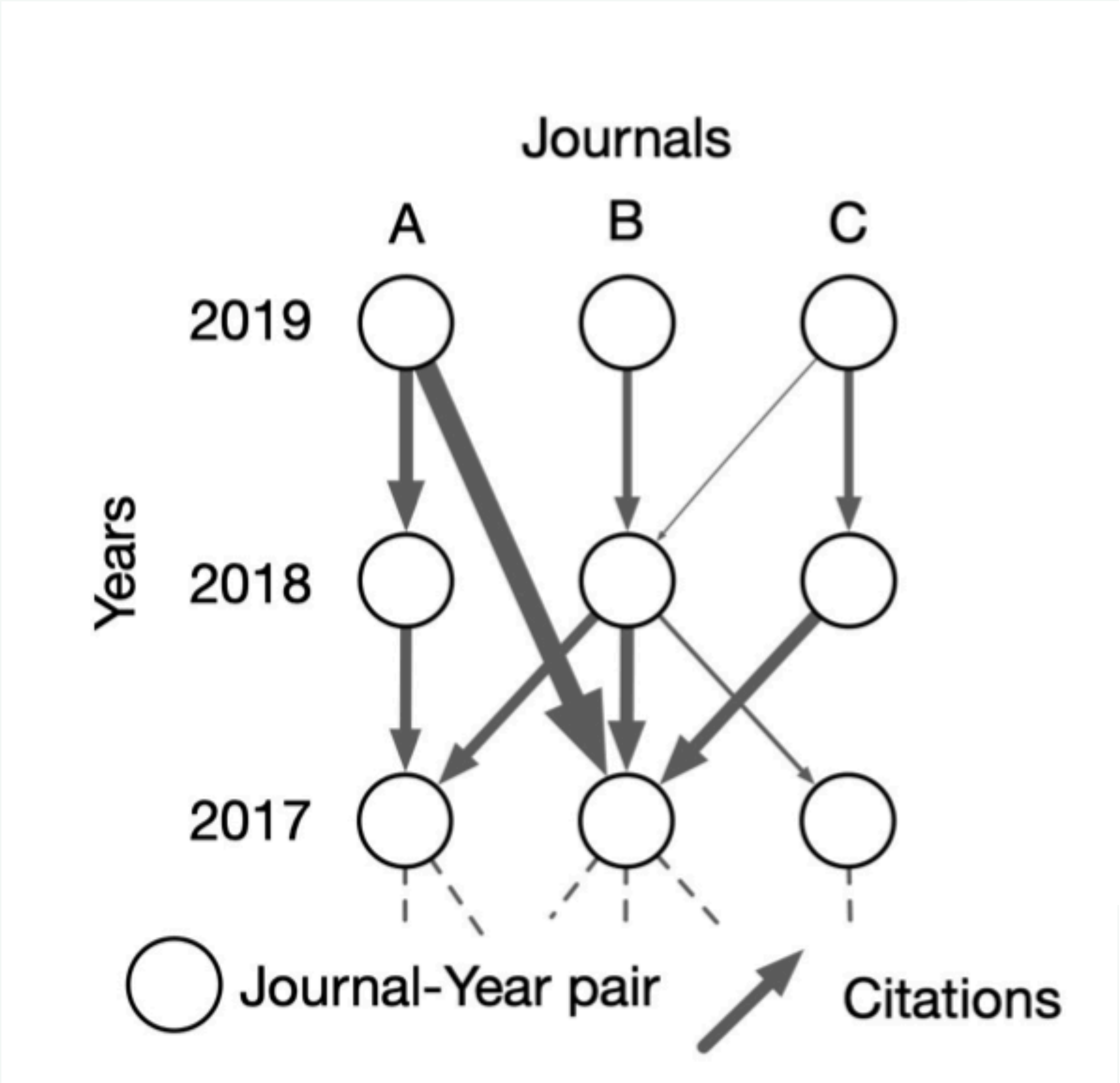
Residual2vec

We can extract out the *expected conditional probability* based on a null model.



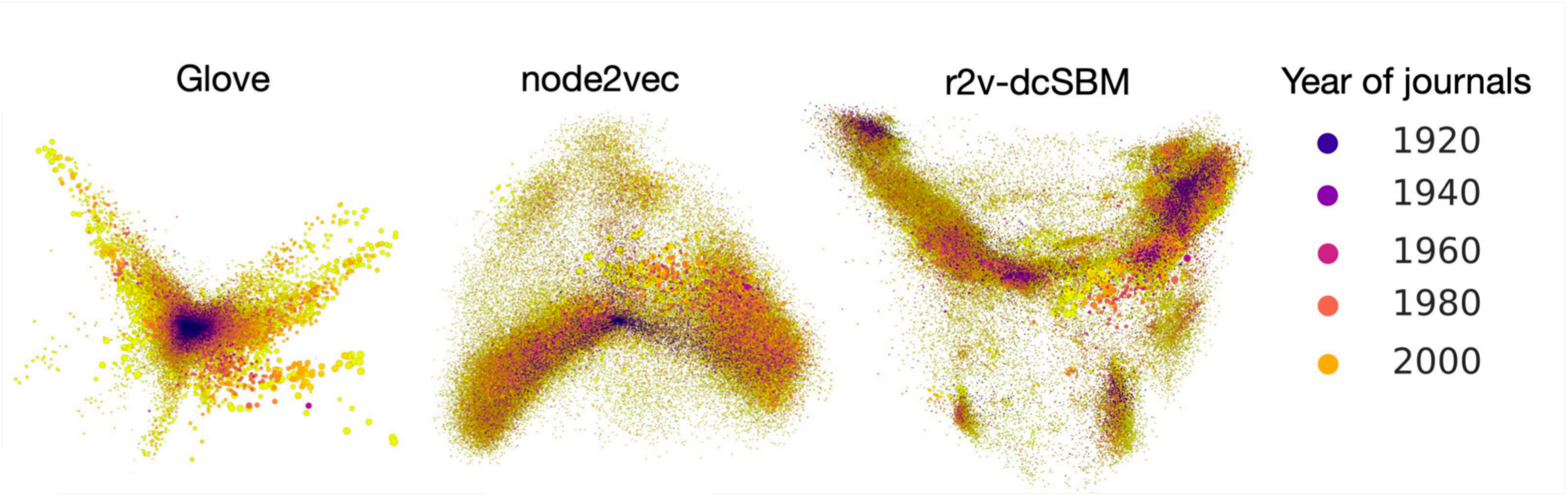
Sadamori Kojaku, Jisung Yoon, Isabel Constantino, Yong-Yeol, "Residual2Vec: Debiasing graph embedding with random graphs", NeurIPS'21

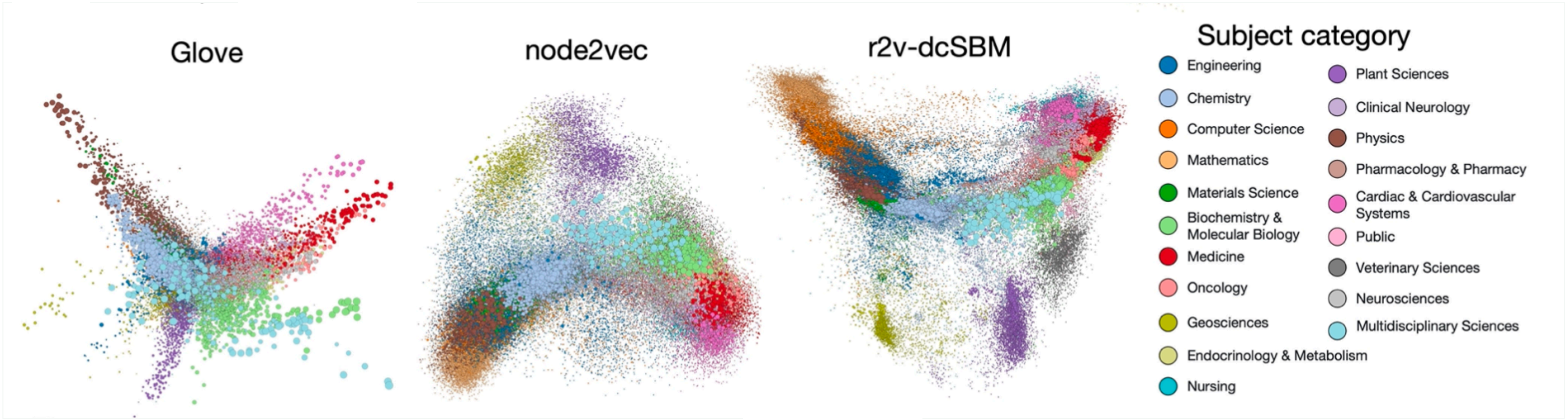
It also allows us to remove specific structural biases



https://observablehq.com/@skojaku/journal-trajector

It also allows us to remove specific structural biases





Sadamori Kojaku, Jisung Yoon, Isabel Constantino, Yong-Yeol, “Residual2Vec: Debiasing graph embedding with random graphs”, NeurIPS’21

Summary

Summary

- It is possible to identify meaningful dimensions and axes in the representation space obtained from neural networks. We can use them to *orient* entities in the space.

Summary

- It is possible to identify meaningful dimensions and axes in the representation space obtained from neural networks. We can use them to *orient* entities in the space.
- Word2vec (SGNS) is *biased*! But, thanks to this bias, the word2vec's objective function corresponds to the gravity law of mobility.

Summary

- It is possible to identify meaningful dimensions and axes in the representation space obtained from neural networks. We can use them to *orient* entities in the space.
- Word2vec (SGNS) is *biased*! But, thanks to this bias, the word2vec's objective function corresponds to the gravity law of mobility.
- This bias also negates the random walk bias in graph embedding!

Summary

- It is possible to identify meaningful dimensions and axes in the representation space obtained from neural networks. We can use them to *orient* entities in the space.
- Word2vec (SGNS) is *biased*! But, thanks to this bias, the word2vec's objective function corresponds to the gravity law of mobility.
- This bias also negates the random walk bias in graph embedding!
- We can further leverage this to remove *specific biases* from a model.

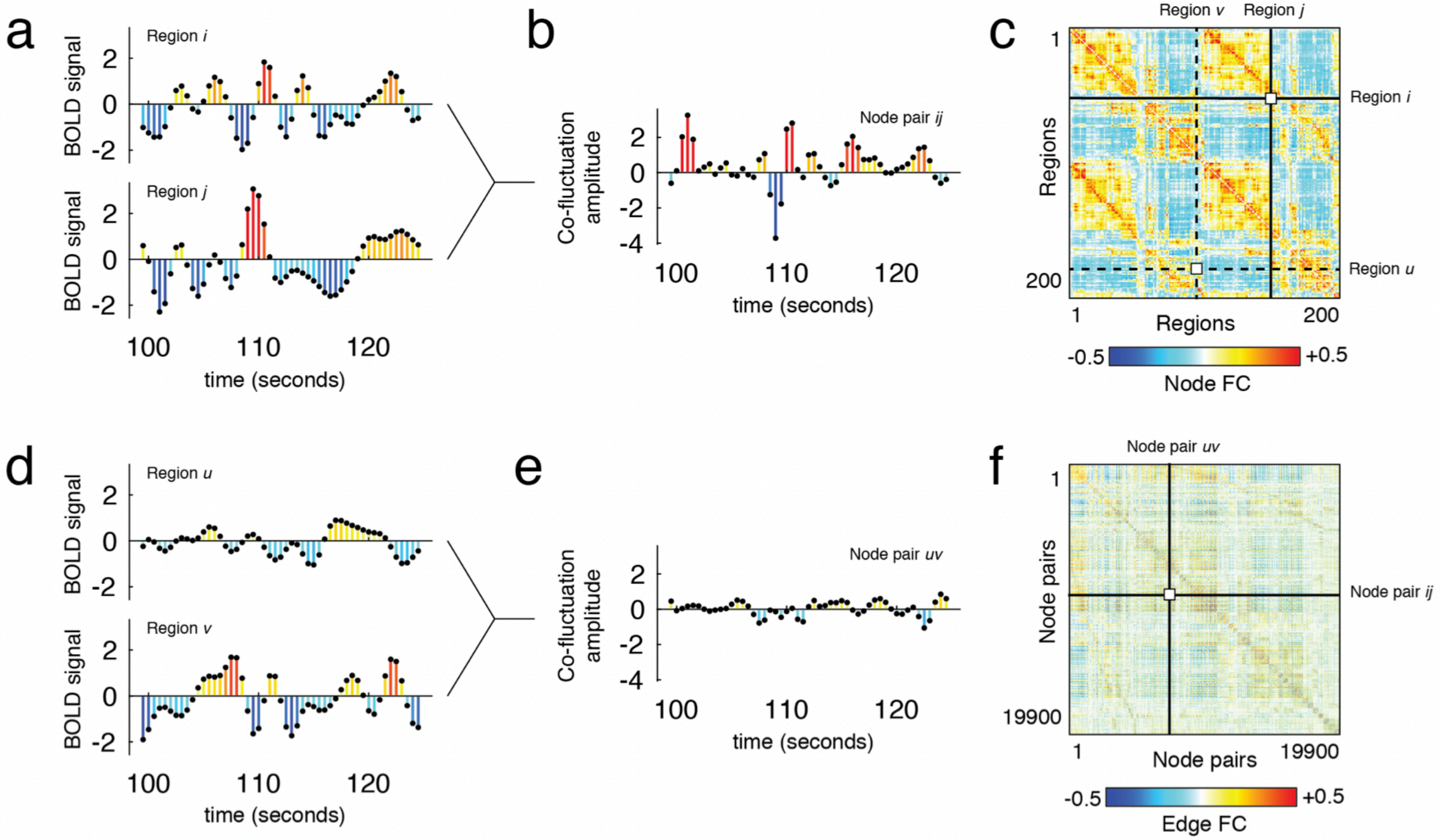
Summary

- It is possible to identify meaningful dimensions and axes in the representation space obtained from neural networks. We can use them to *orient* entities in the space.
- Word2vec (SGNS) is *biased*! But, thanks to this bias, the word2vec's objective function corresponds to the gravity law of mobility.
- This bias also negates the random walk bias in graph embedding!
- We can further leverage this to remove *specific biases* from a model.
- Simple models, when understood well, can take us quite far.

Summary

- It is possible to identify meaningful dimensions and axes in the representation space obtained from neural networks. We can use them to *orient* entities in the space.
- Word2vec (SGNS) is *biased*! But, thanks to this bias, the word2vec's objective function corresponds to the gravity law of mobility.
- This bias also negates the random walk bias in graph embedding!
- We can further leverage this to remove *specific biases* from a model.
- Simple models, when understood well, can take us quite far.
- What could be the ways to obtain **useful, compact representation of dynamic, functional brain networks**?

How about *dense representation* of dynamic neural networks?



Faskowitz, Joshua, et al. "Edge-centric functional network representations of human cerebral cortex reveal overlapping system-level architecture." *Nature neuroscience* 23.12 (2020): 1644-1654.

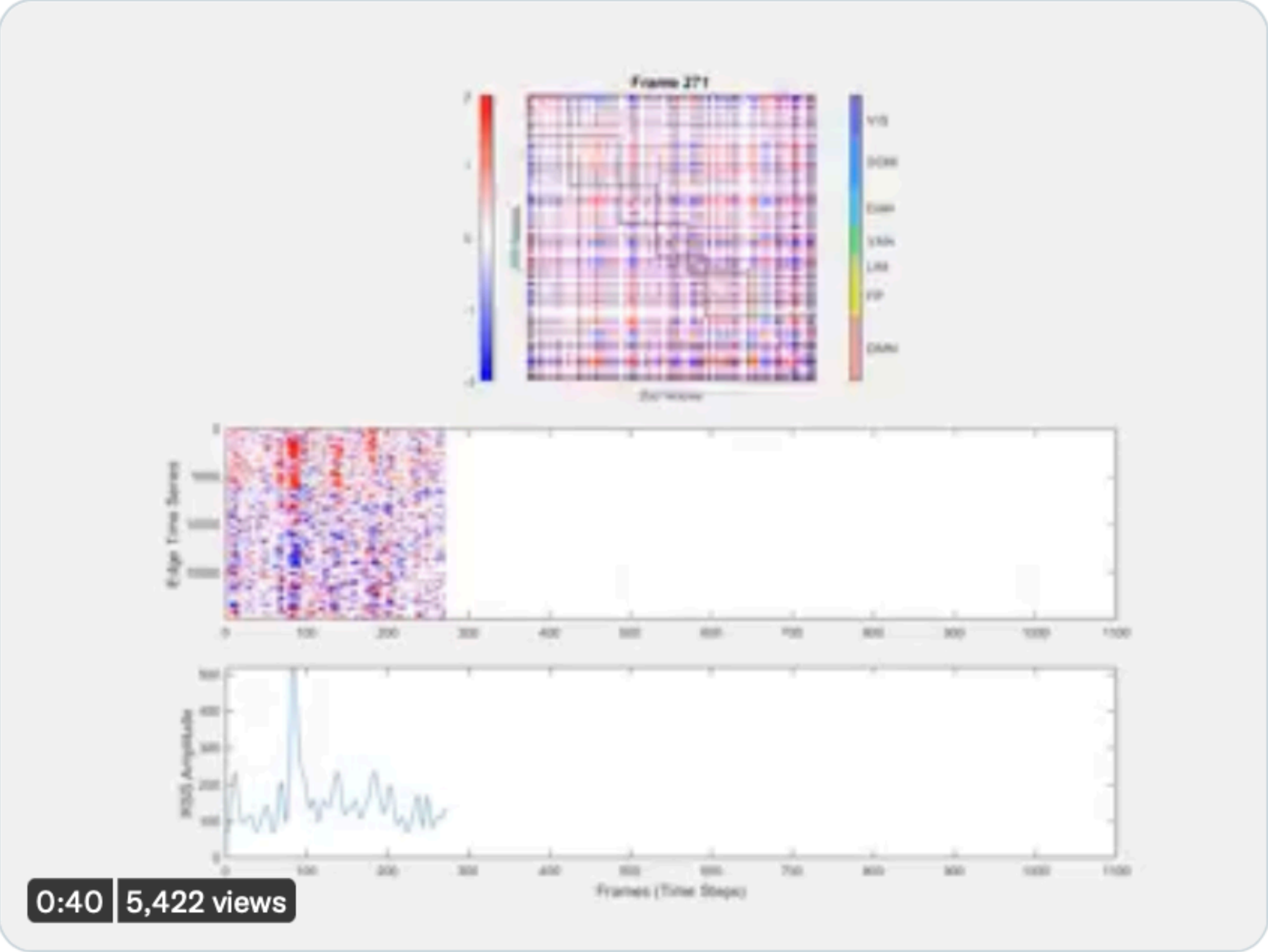
https://twitter.com/spornslab/status/1319390214767378432

Olaf Sporns
@spornslab

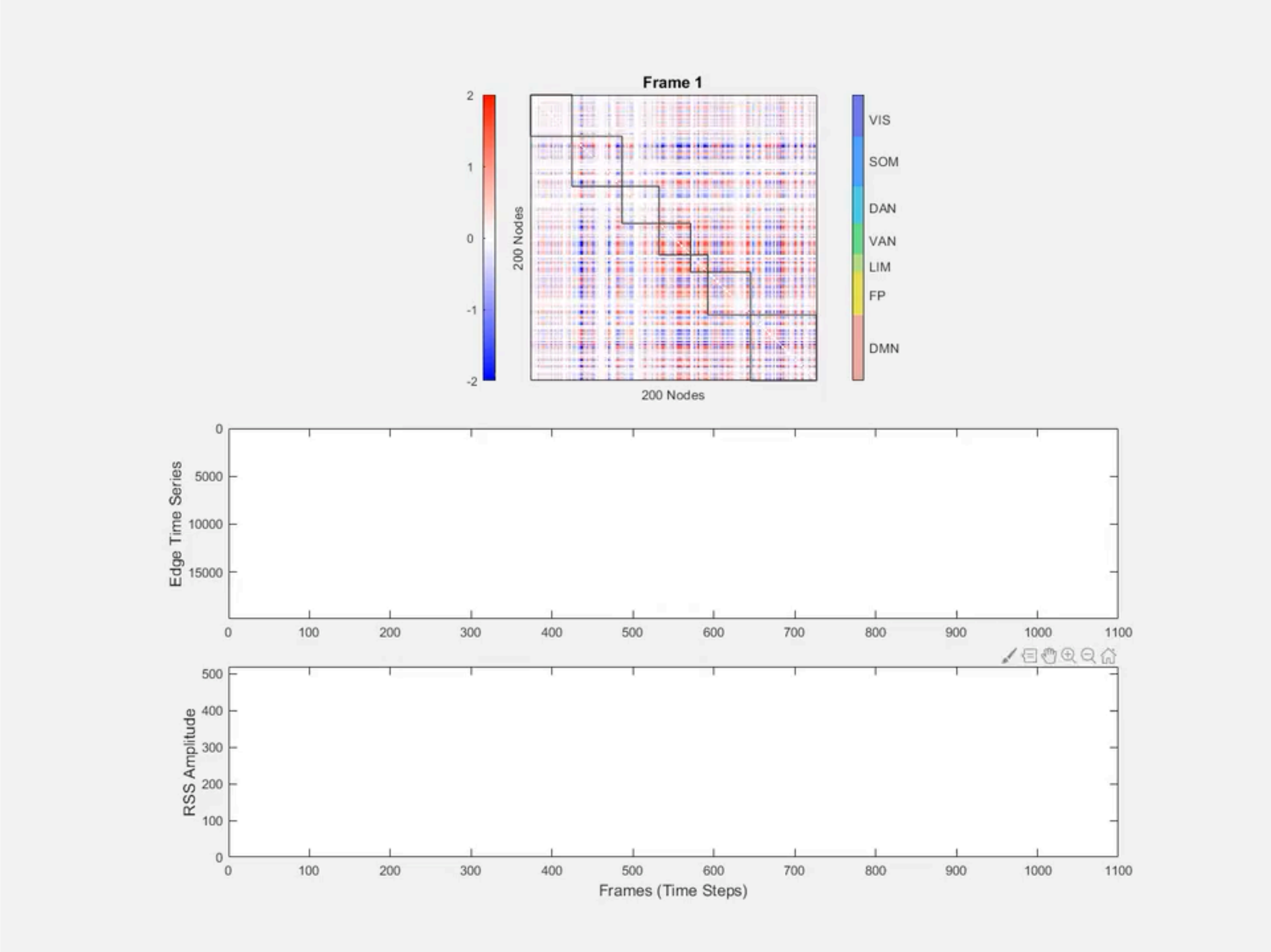
A short thread on this new pub:
pnas.org/content/early/...

Movie below shows functional connectivity unwrapped into ‘edge time series’ (data: single rs-fMRI scan, 200 nodes, 1100 frames, TR=720ms)

Note: the mean of all frames is exactly equal to ‘classic’ FC



3:28 PM · Oct 22, 2020 · Twitter Web App



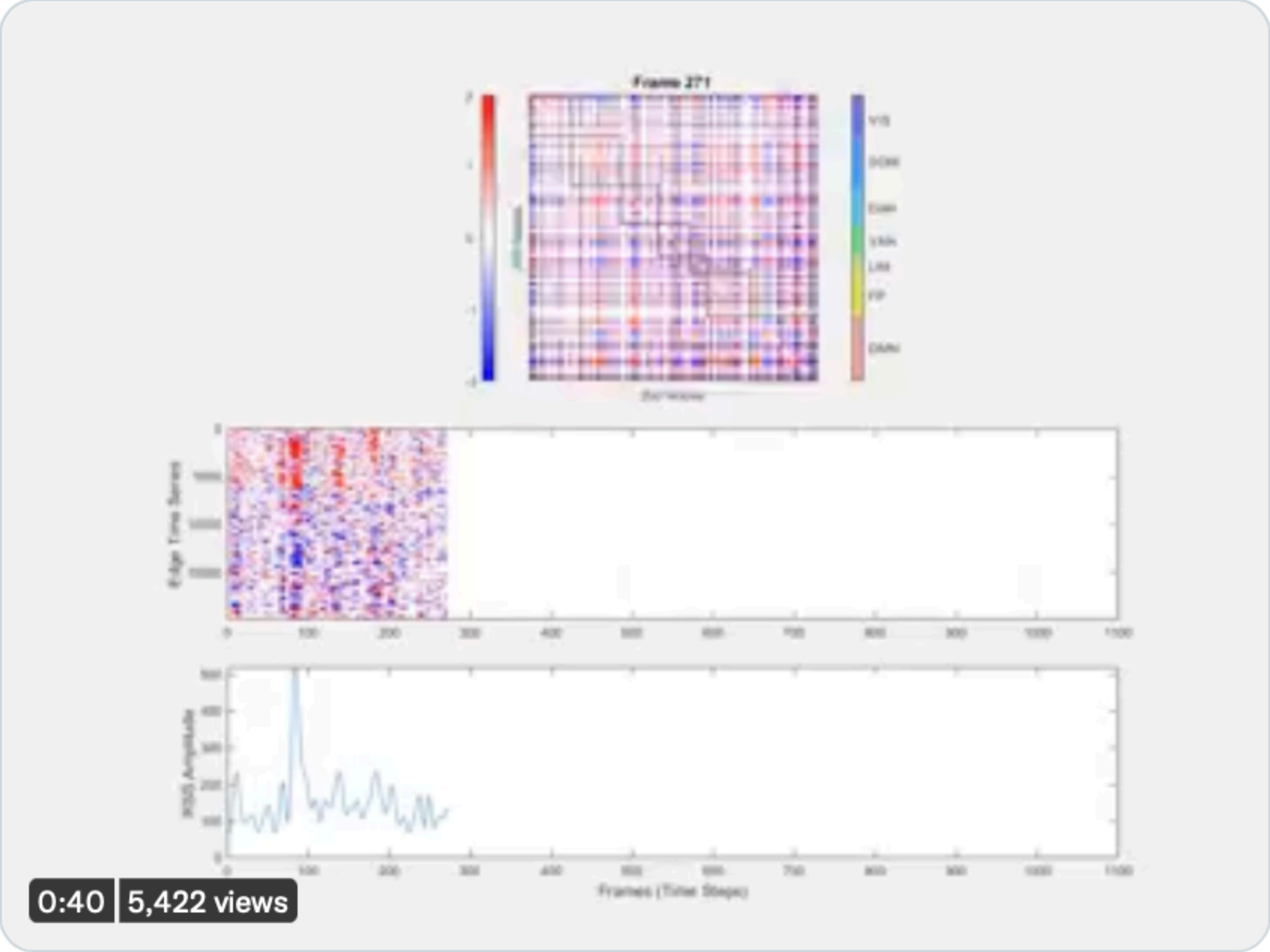
https://twitter.com/spornslab/status/1319390214767378432

Olaf Sporns
@spornslab

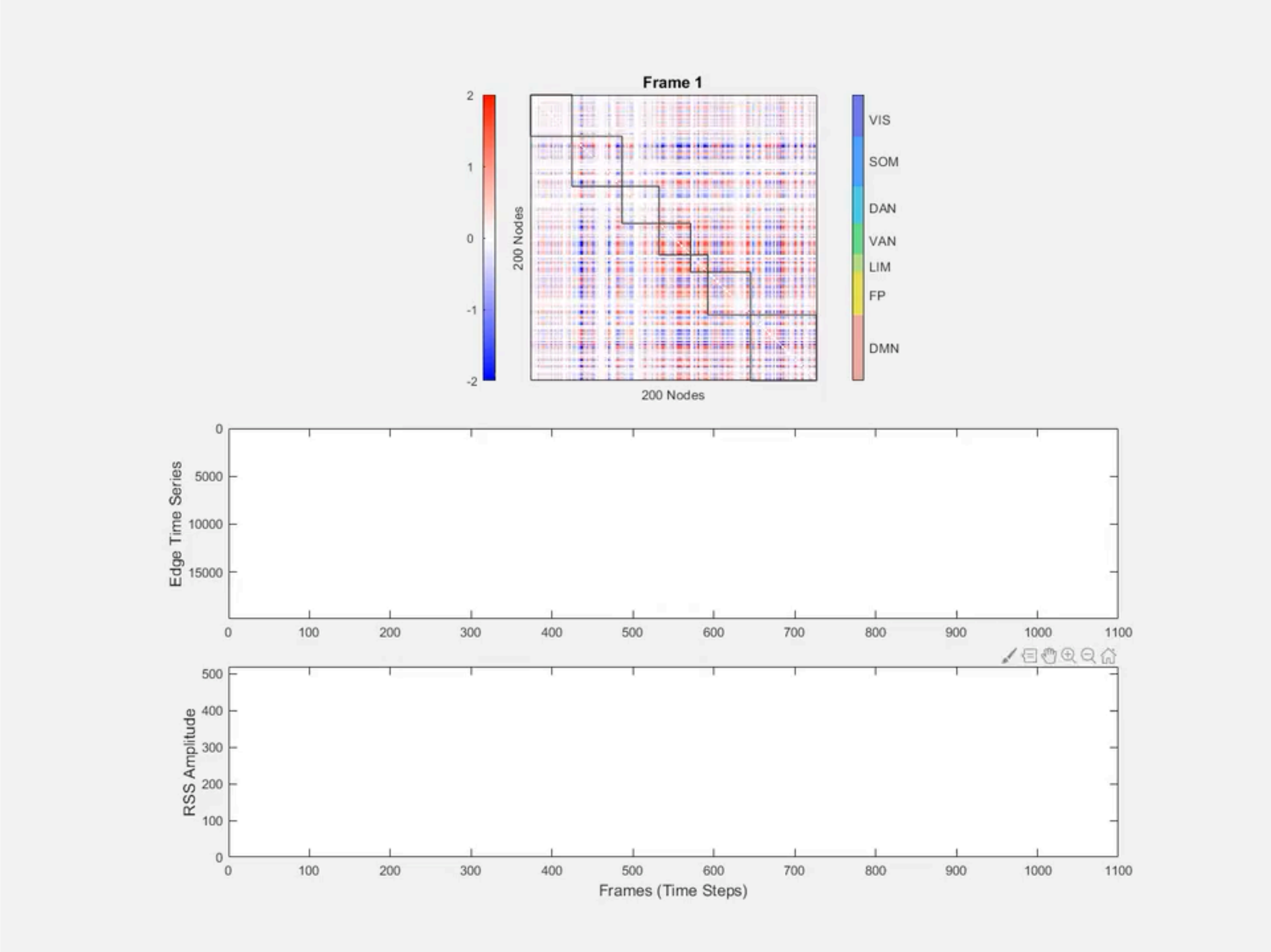
A short thread on this new pub:
pnas.org/content/early/...

Movie below shows functional connectivity unwrapped into ‘edge time series’ (data: single rs-fMRI scan, 200 nodes, 1100 frames, TR=720ms)

Note: the mean of all frames is exactly equal to ‘classic’ FC

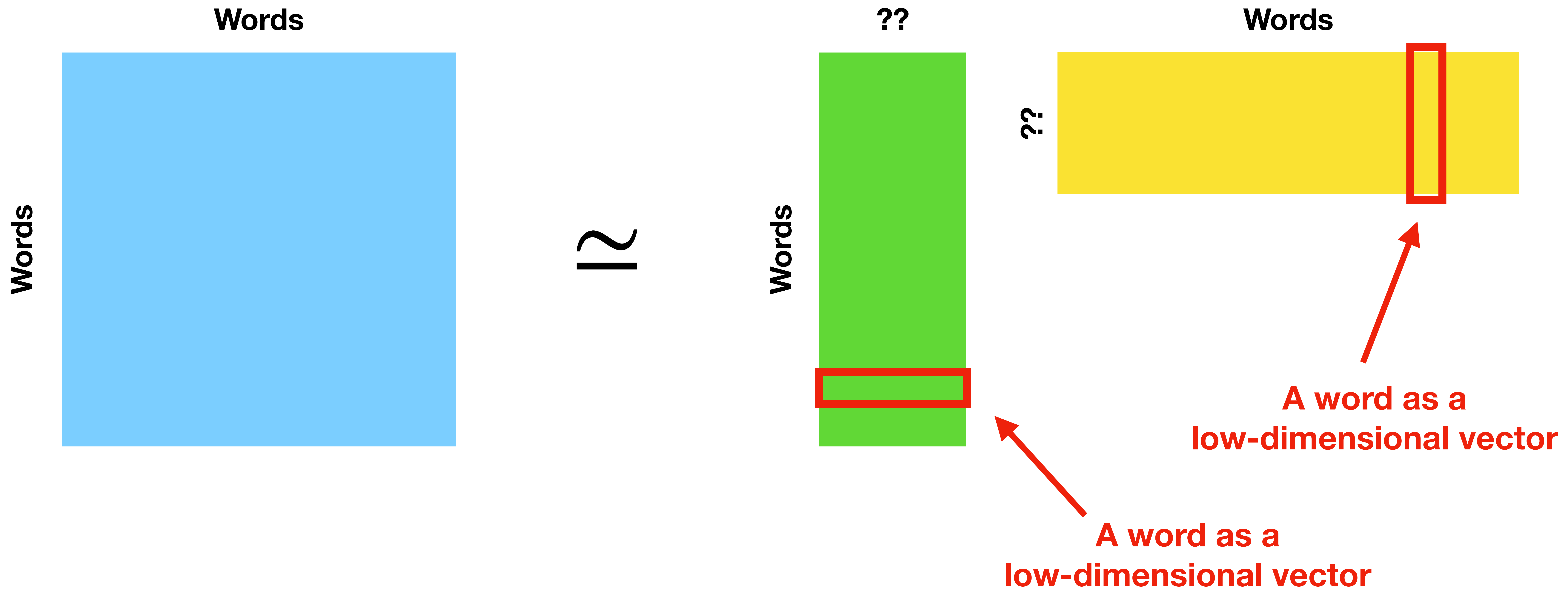


3:28 PM · Oct 22, 2020 · Twitter Web App

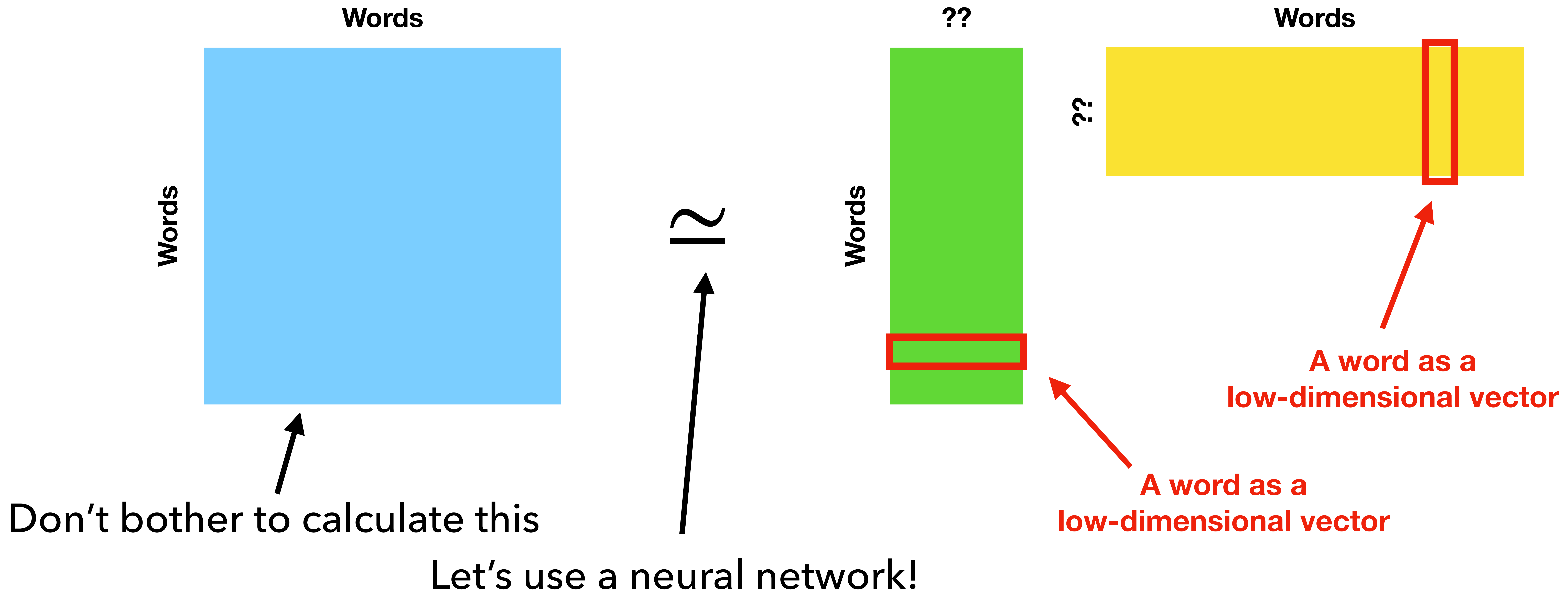


Can we identify universal
& individual
cofluctuation patterns?

Representation learning as Matrix Factorization

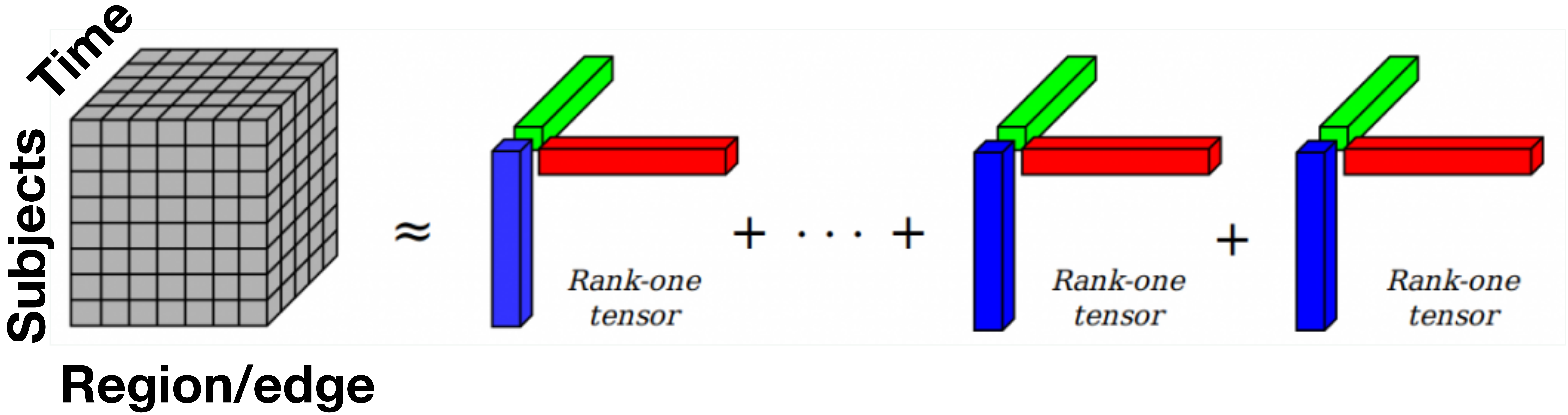


Representation learning as Matrix Factorization

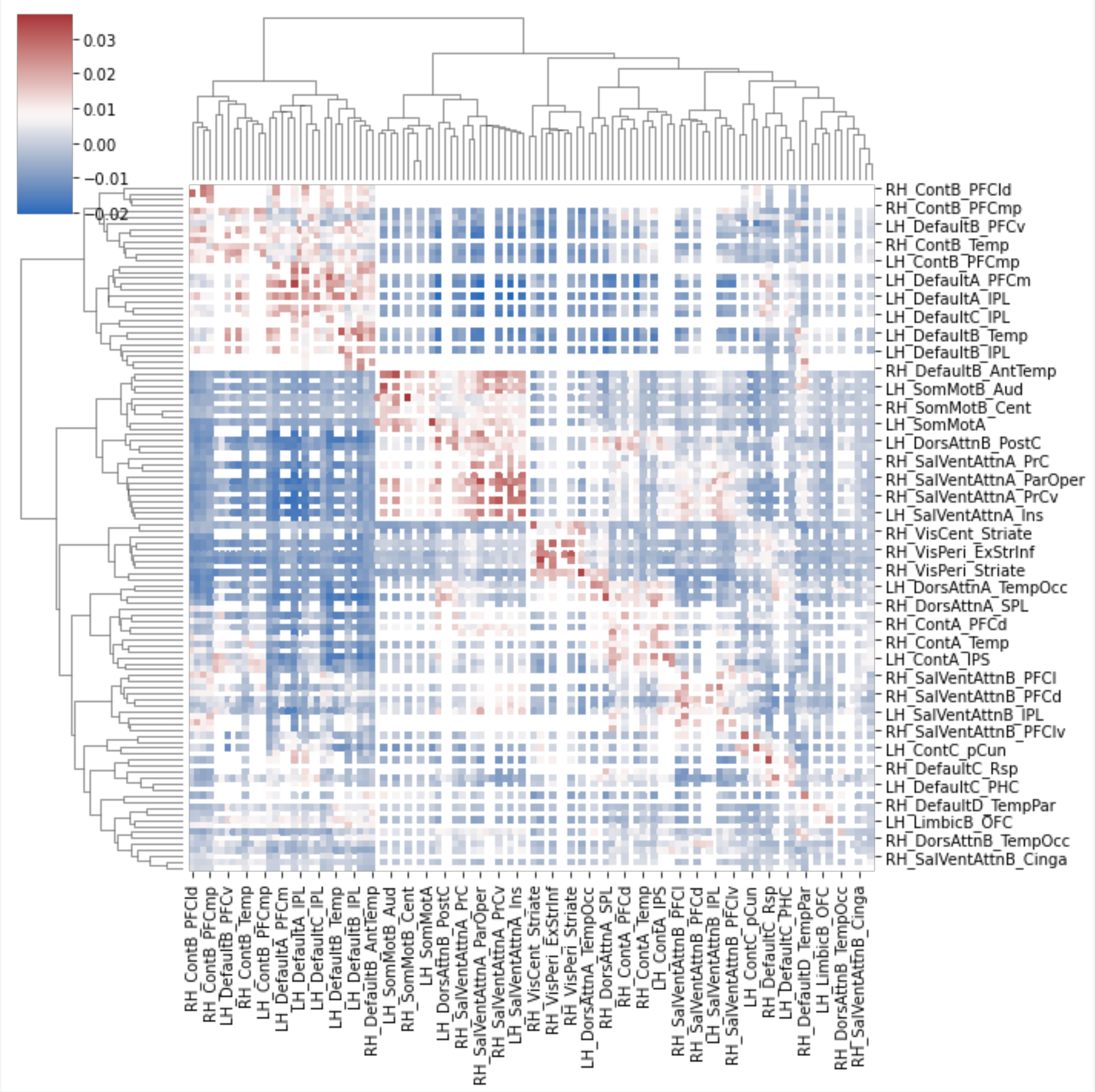


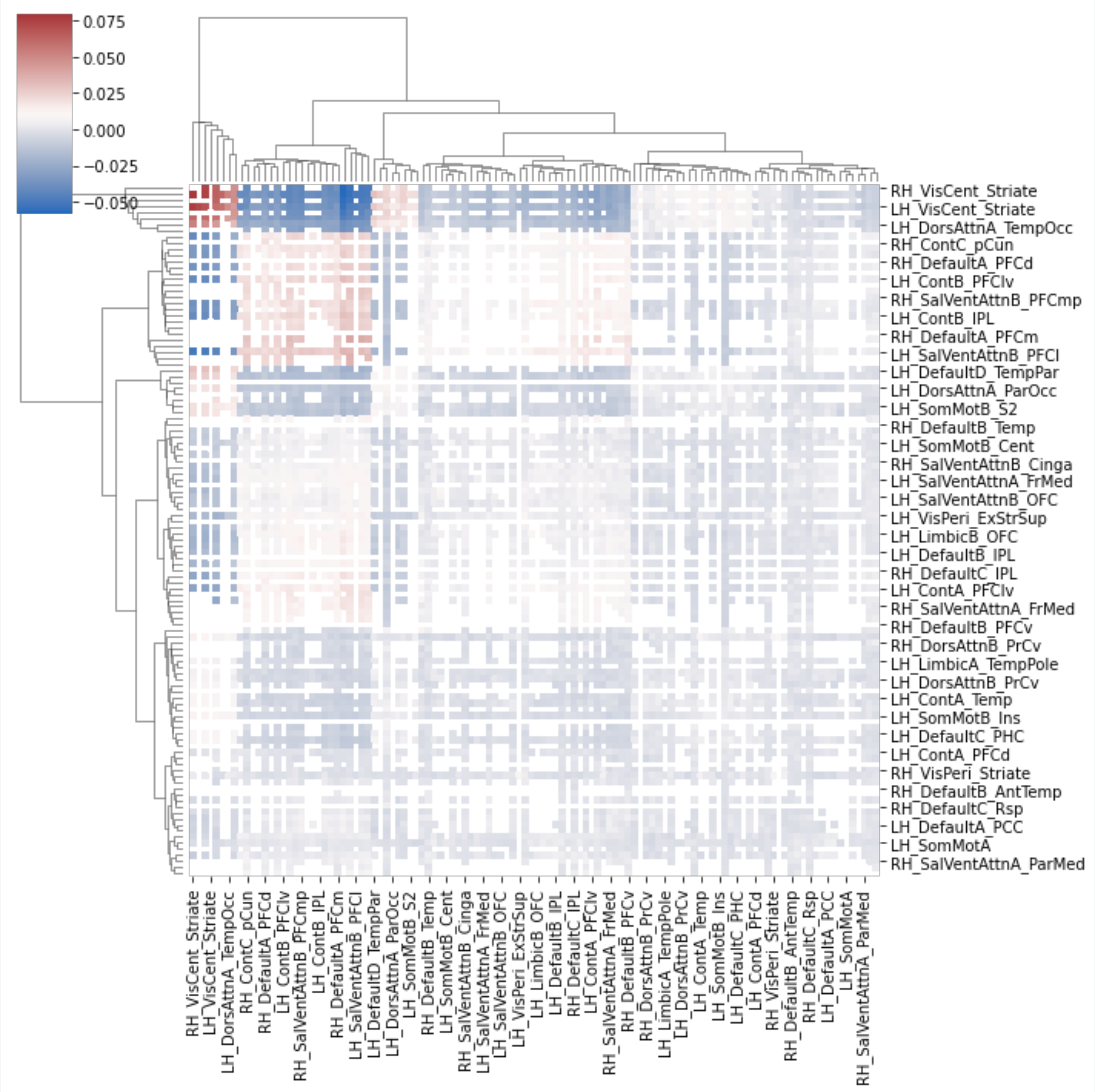
Higher-Rank Tensor?

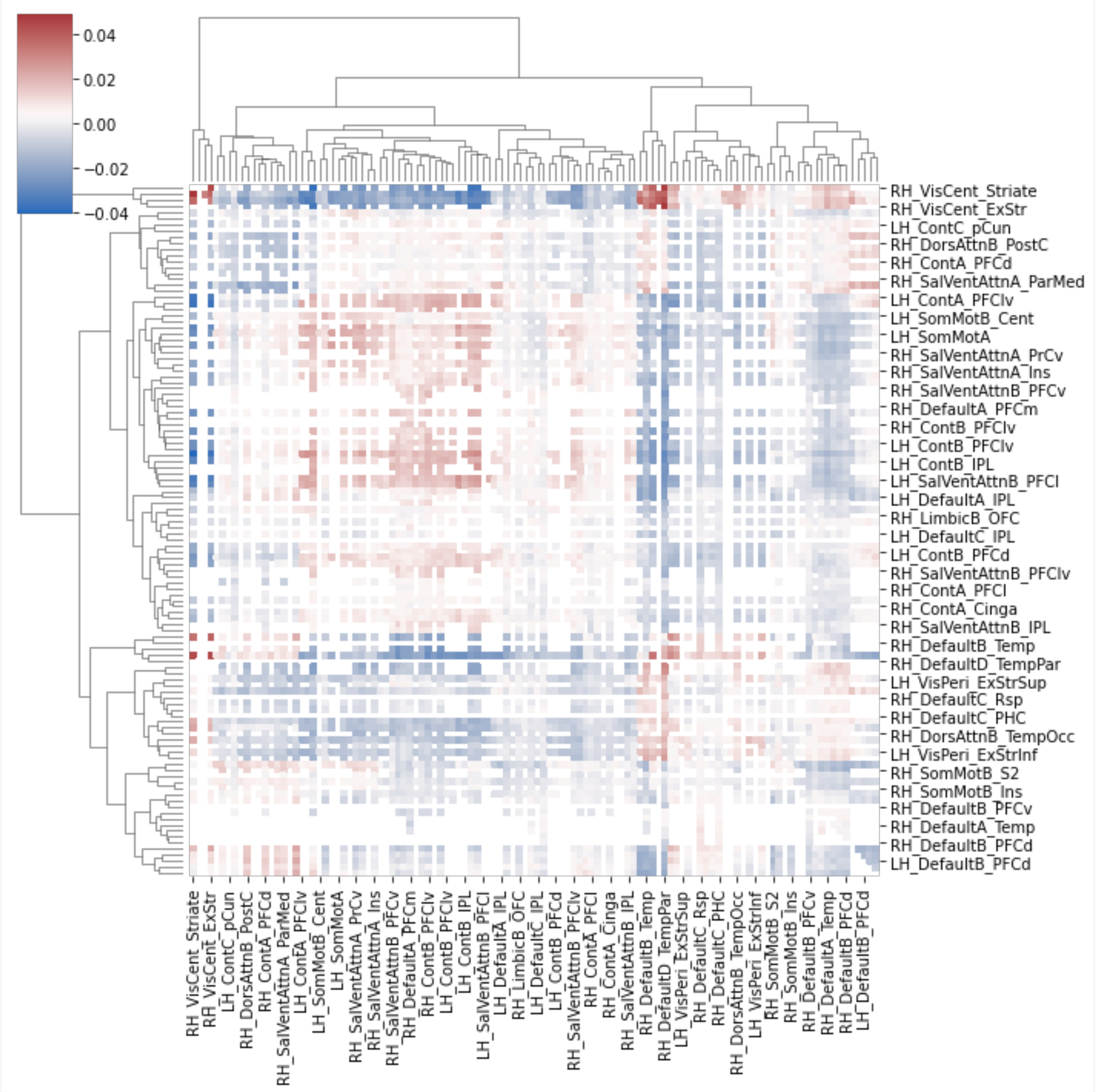
Tensor Decomposition



Edge co-fluctuation components from tensor decomposition







What could be the useful, compact **representations** of the brain's dynamics?

Can we imagine it as a meaningful **space**?

Thanks!


Staša Milojević


Isabel Constantino


Rodrigo Costas


Supun Nakandala


Jisun An


Haewoon Kwak


Sadamori Kojaku


Dakota Murray


Giovanni Luca Ciampaglia


Jisung Yoon


Qing Ke


Jaehyuk Park


Fabio Rojas


Hao Peng


Daniel Romero


Norman Makoto Su


Ceren Budak

Other **Science Genome & CADRE team**: Alessandro Flammini, Filippo Menczer, Sriraam Natarajan, Attila Varga, Xiaoran Yan, Filipi Silva, Clara Boothby, Valentin Pentchev, Matthew Hutchinson, Chathuri Peli Kankanamalage