РАСШИФРОВЫВАЕМ ЗАГАДОЧНЫЕ КОДЫ: РУКОПИСЬ ВОЙНИЧА, БИОИНФОРМАТИКА И МОЗГОВАЯ АКТИВНОСТЬ

....и как в этом поможет word2wec

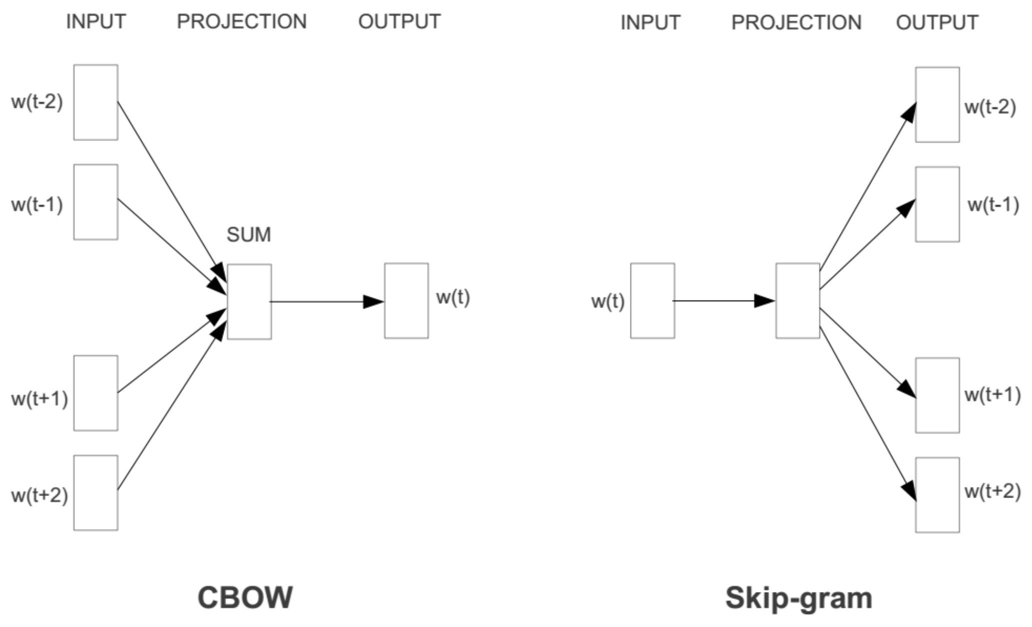
ДМИТРИЙ НОВИЦКИЙ

WORD2VEC : A RECALL

- Represent each word with a low-dimensional vector
- Word similarity = vector similarity
- Key idea: Predict surrounding words of every word
- Faster and can easily incorporate a new sentence/document or add a word to the vocabulary

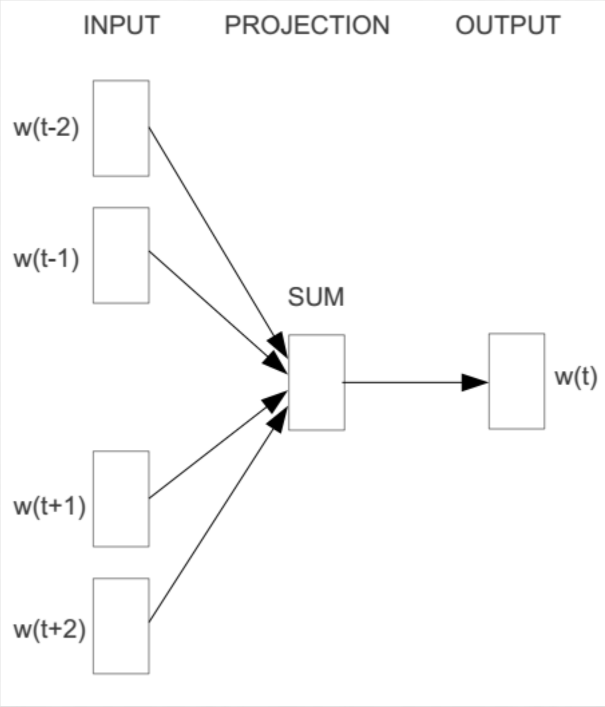
REPRESENT THE MEANING OF WORD – WORD2VEC

- 2 basic neural network models:
  - Continuous Bag of Word (CBOW): use a window of word to predict the middle word
  - Skip-gram (SG): use a word to predict the surrounding ones in window.



WORD2VEC – CONTINUOUS BAG OF WORD

- E.g. "The cat sat on floor"
  - Window size = 2



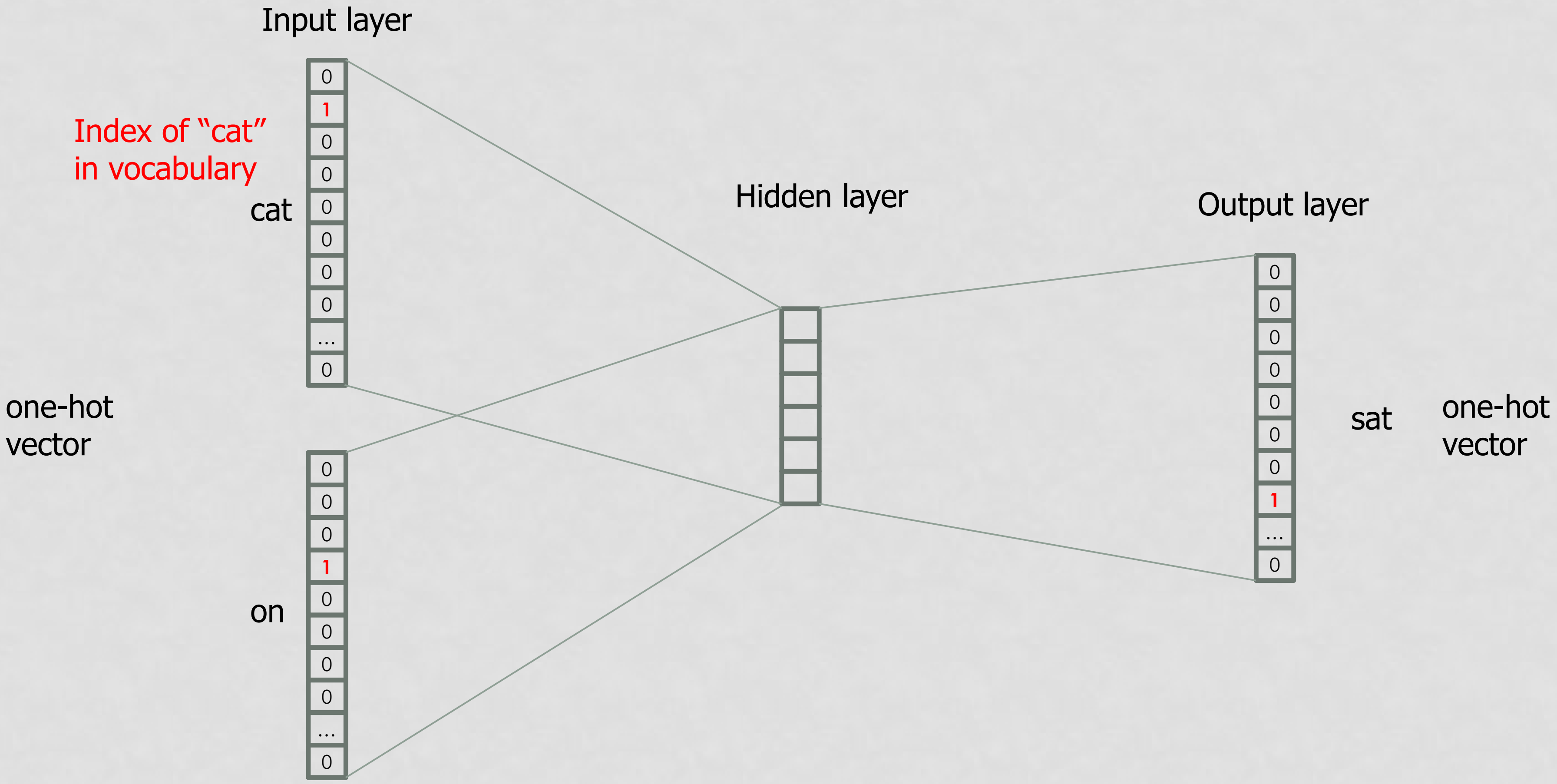
the

cat

on

floor

sat



Input layer

Index of "cat" in vocabulary

cat

on

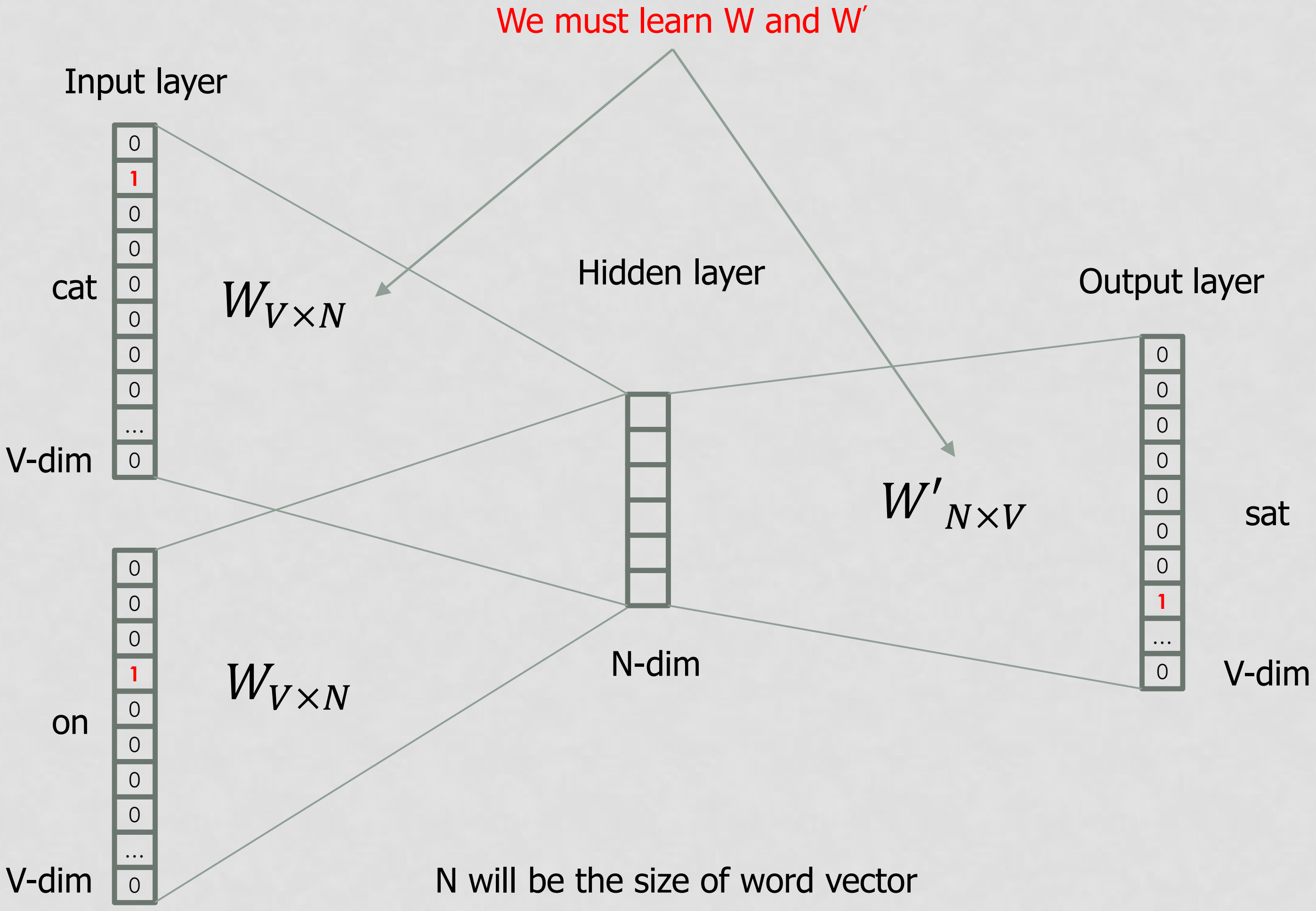
one-hot vector

Hidden layer

Output layer

sat

one-hot vector



We must learn W and W′

Input layer

cat

0 1 0 0 0 0 0 0 … 0

V-dim

on

0 0 0 1 0 0 0 0 … 0

V-dim

$W_{V \times N}$

$W_{V \times N}$

Hidden layer

N-dim

$W'_{N \times V}$

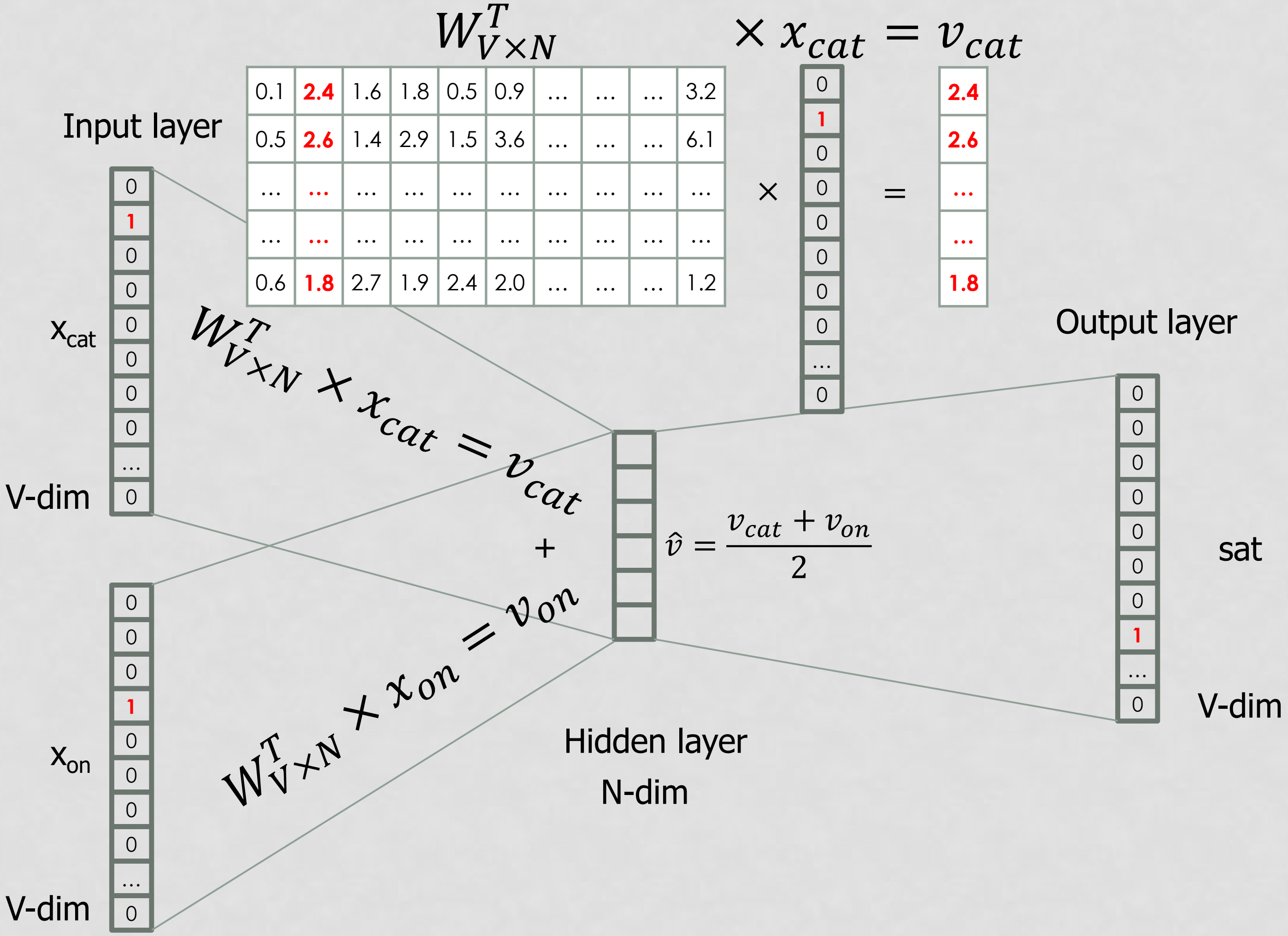
Output layer

0 0 0 0 0 0 0 1 … 0

sat

V-dim

N will be the size of word vector



$$W^T_{V\times N} \times x_{cat} = v_{cat}$$

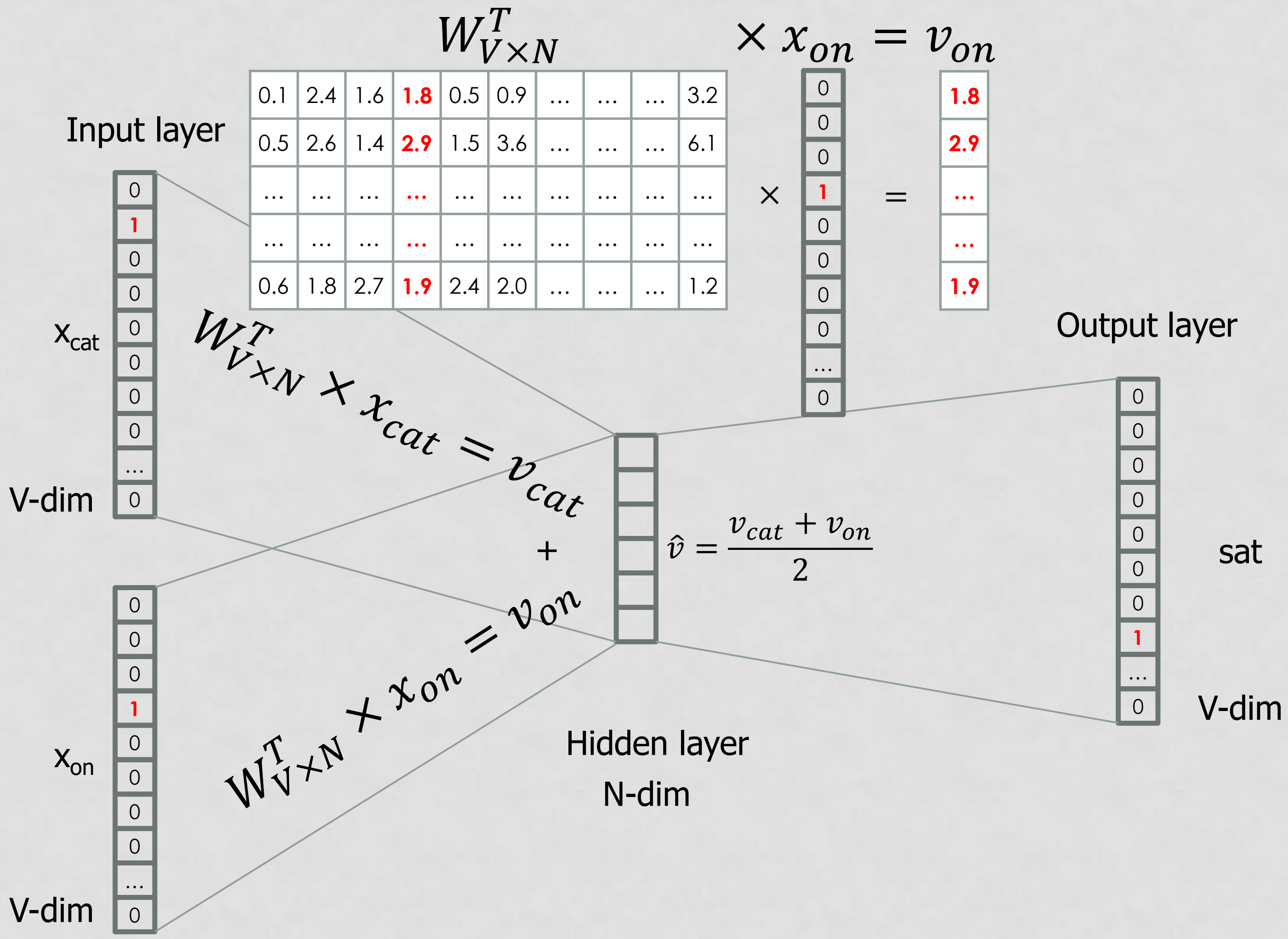
$$\begin{bmatrix} 0.1 & 2.4 & 1.6 & 1.8 & 0.5 & 0.9 & \dots & \dots & \dots & 3.2 \\ 0.5 & 2.6 & 1.4 & 2.9 & 1.5 & 3.6 & \dots & \dots & \dots & 6.1 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0.6 & 1.8 & 2.7 & 1.9 & 2.4 & 2.0 & \dots & \dots & \dots & 1.2 \end{bmatrix} \times \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \dots \\ 0 \end{bmatrix} = \begin{bmatrix} 2.4 \\ 2.6 \\ \dots \\ \dots \\ 1.8 \end{bmatrix}$$

$$W^T_{V\times N} \times x_{cat} = v_{cat}$$

$+$

$$W^T_{V\times N} \times x_{on} = v_{on}$$

$$\hat{v} = \frac{v_{cat} + v_{on}}{2}$$



$$W^T_{V\times N} \times x_{on} = v_{on}$$

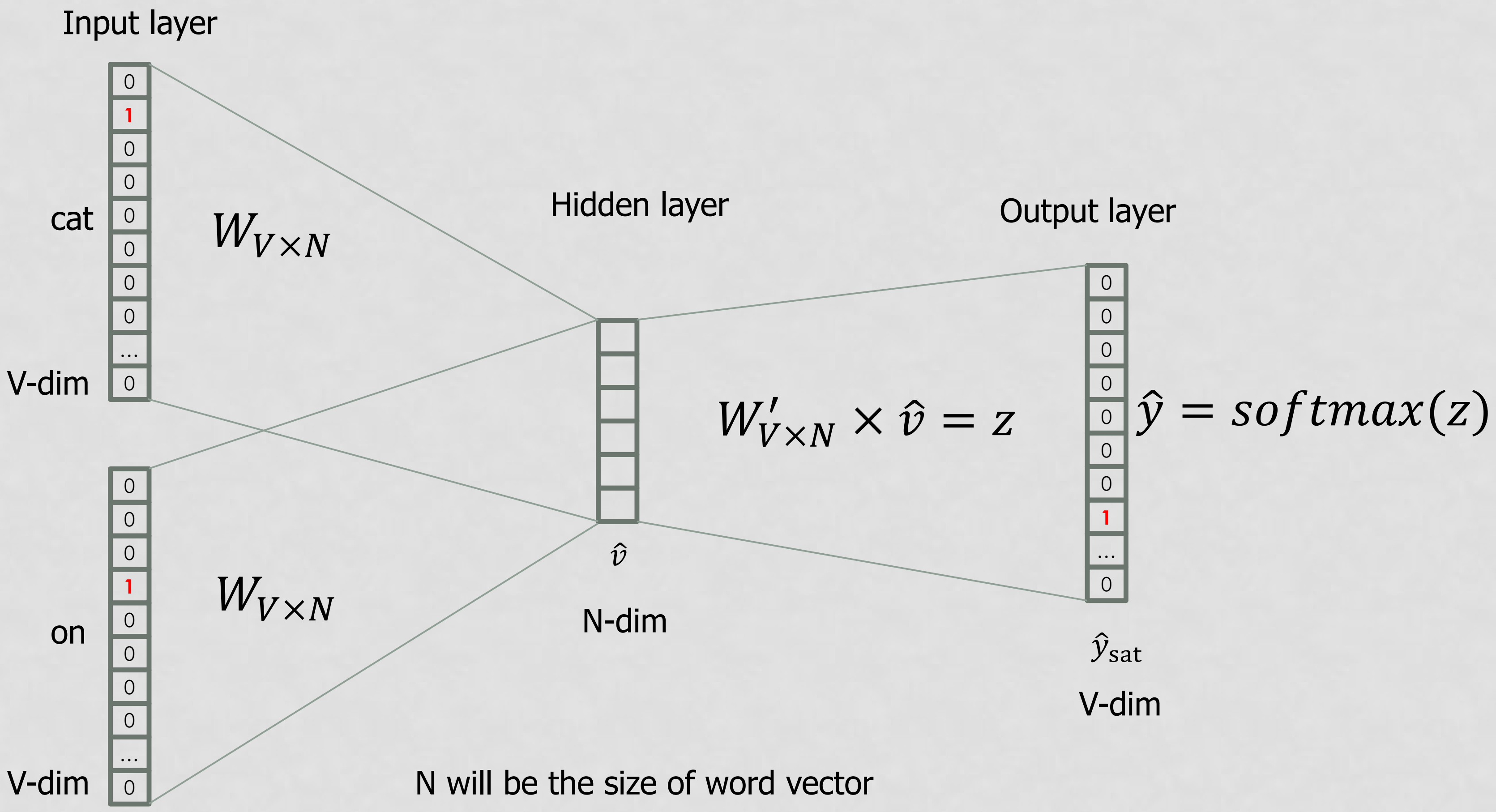
$$\begin{bmatrix} 0.1 & 2.4 & 1.6 & 1.8 & 0.5 & 0.9 & \dots & \dots & \dots & 3.2 \\ 0.5 & 2.6 & 1.4 & 2.9 & 1.5 & 3.6 & \dots & \dots & \dots & 6.1 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0.6 & 1.8 & 2.7 & 1.9 & 2.4 & 2.0 & \dots & \dots & \dots & 1.2 \end{bmatrix} \times \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ \dots \\ 0 \end{bmatrix} = \begin{bmatrix} 1.8 \\ 2.9 \\ \dots \\ \dots \\ 1.9 \end{bmatrix}$$

$$W^T_{V\times N} \times x_{cat} = v_{cat}$$

$$+$$

$$W^T_{V\times N} \times x_{on} = v_{on}$$

$$\hat{v} = \frac{v_{cat} + v_{on}}{2}$$



Input layer

cat

0 1 0 0 0 0 0 0 … 0

$W_{V \times N}$

V-dim

on

0 0 0 1 0 0 0 0 … 0

$W_{V \times N}$

V-dim

Hidden layer

$\hat{v}$

N-dim

$$W'_{V \times N} \times \hat{v} = z$$

Output layer

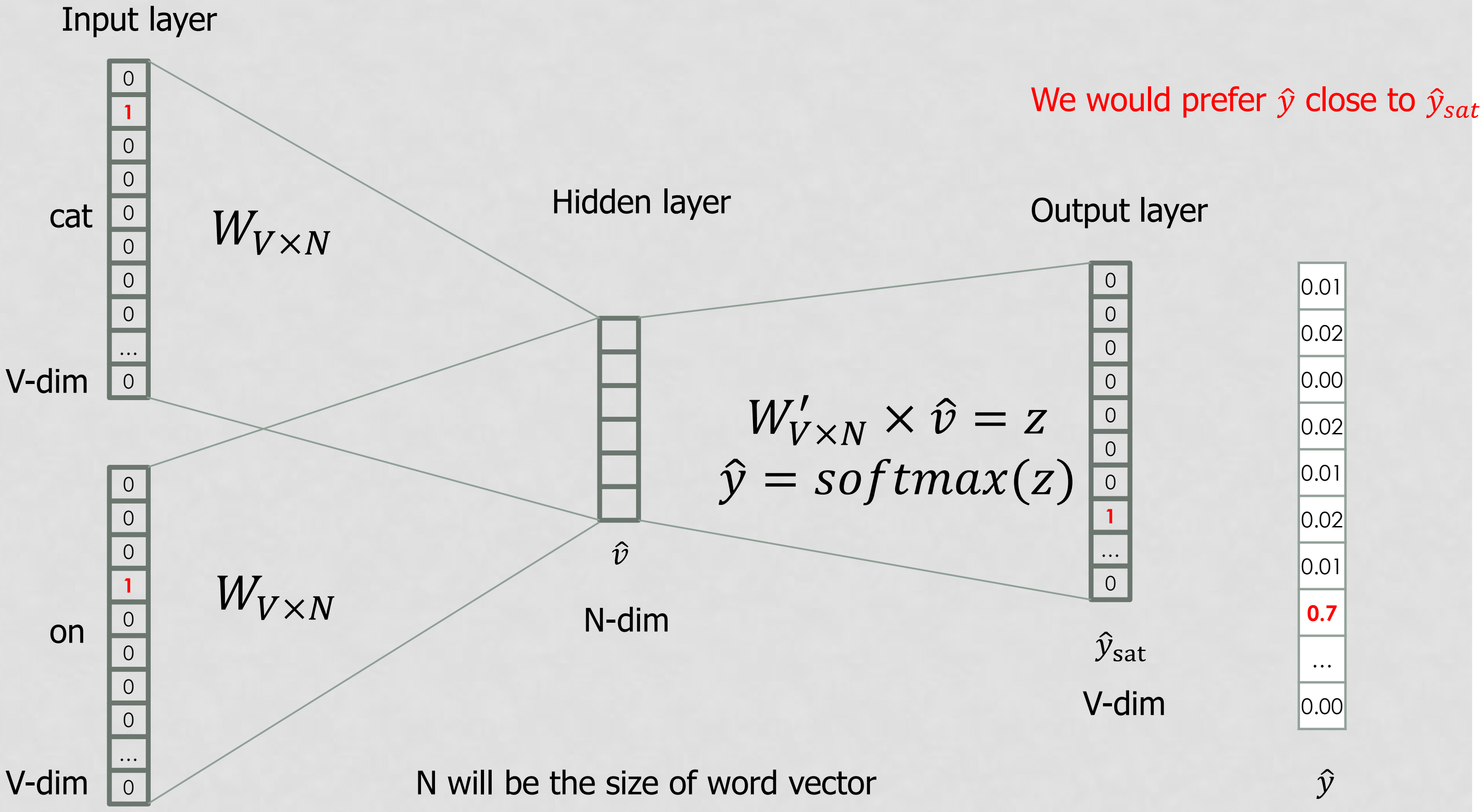
0 0 0 0 0 0 0 1 … 0

$$\hat{y} = softmax(z)$$

$\hat{y}_{\text{sat}}$

V-dim

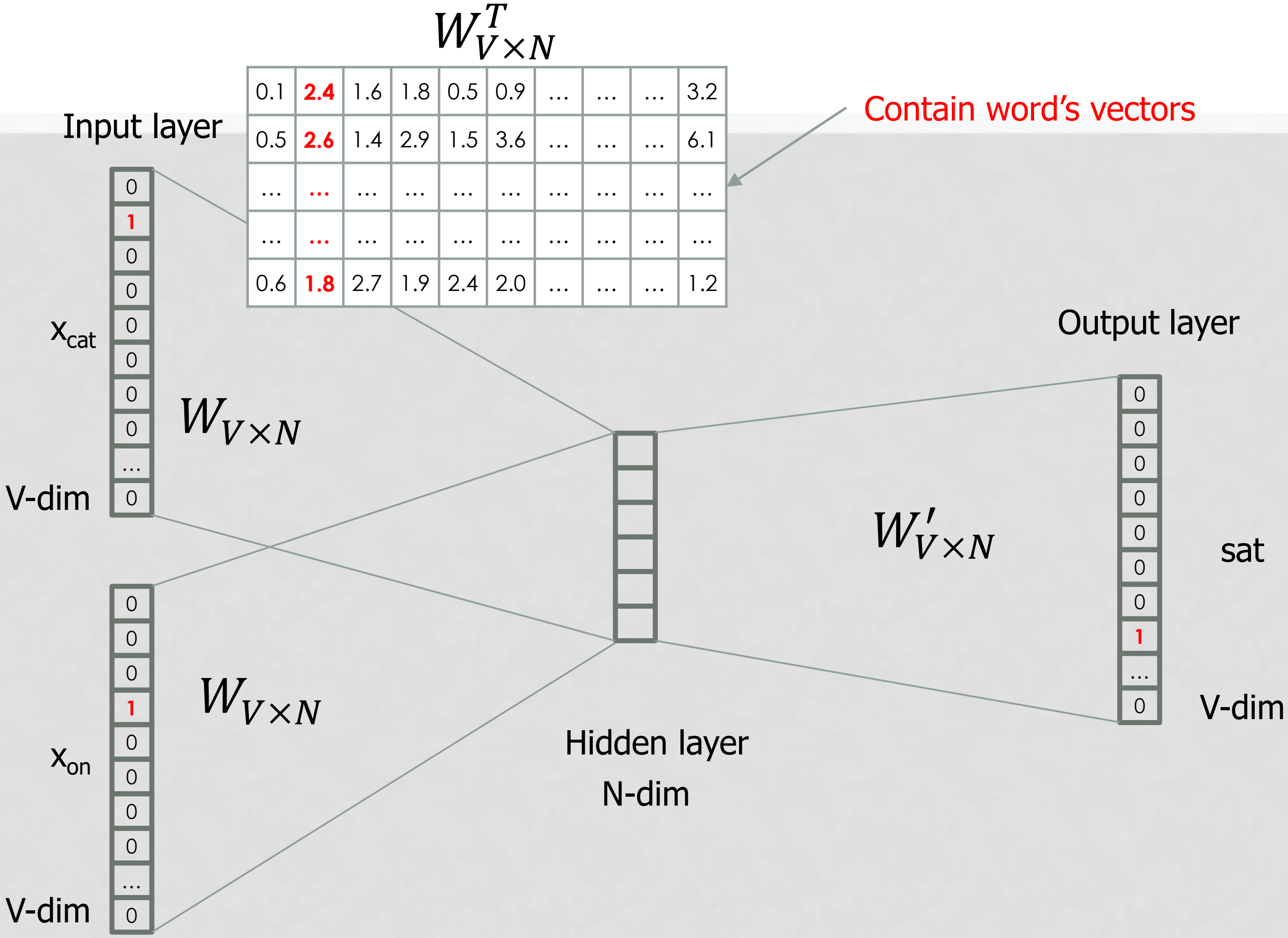
N will be the size of word vector



We would prefer $\hat{y}$ close to $\hat{y}_{sat}$

$$W'_{V \times N} \times \hat{v} = z$$
$$\hat{y} = softmax(z)$$

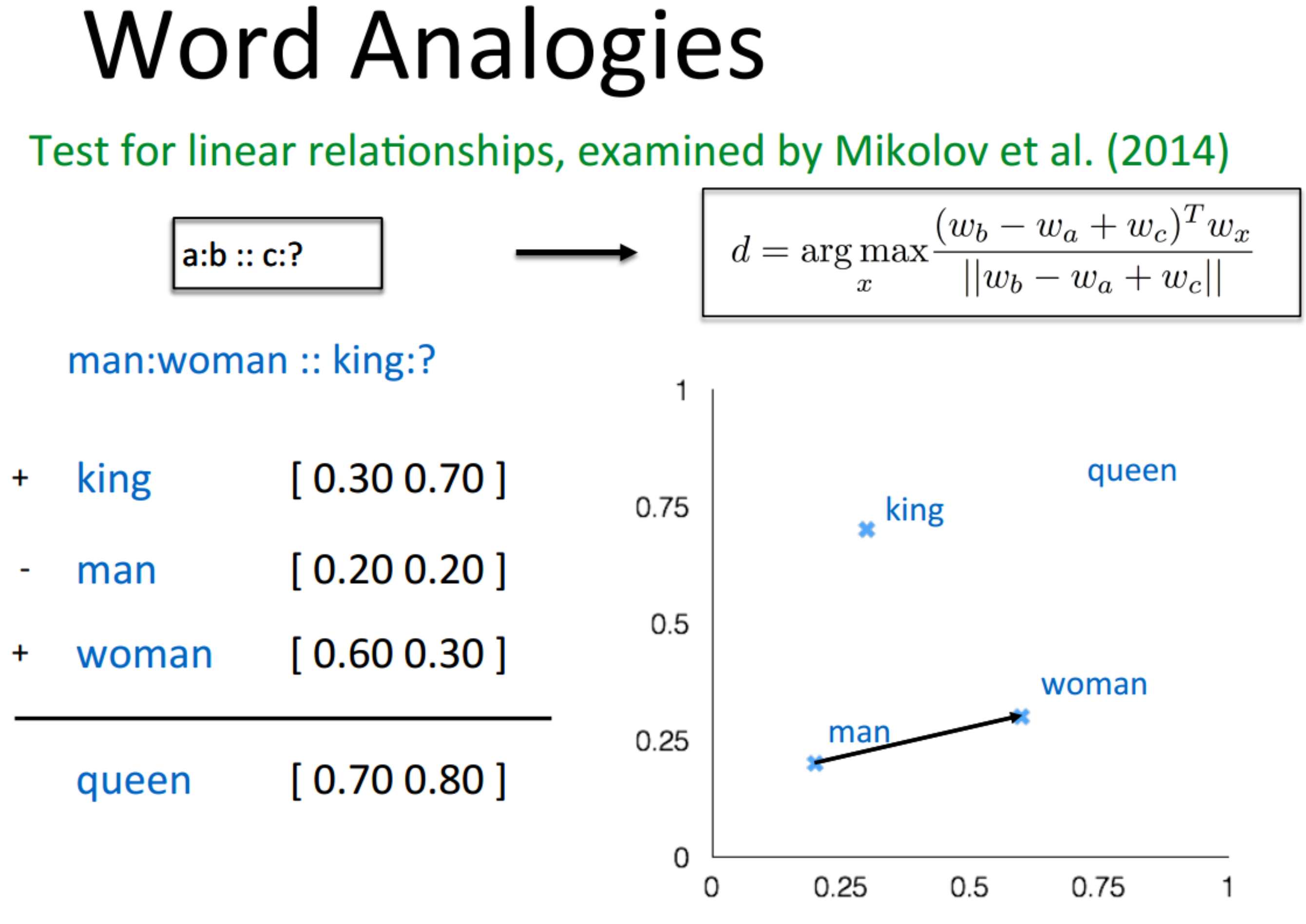
N will be the size of word vector



0.1	2.4	1.6	1.8	0.5	0.9	…	…	…	3.2
0.5	2.6	1.4	2.9	1.5	3.6	…	…	…	6.1
…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…
0.6	1.8	2.7	1.9	2.4	2.0	…	…	…	1.2

We can consider either W or W’ as the word’s representation.
Or even take the average.

SOME INTERESTING RESULTS



Word Analogies

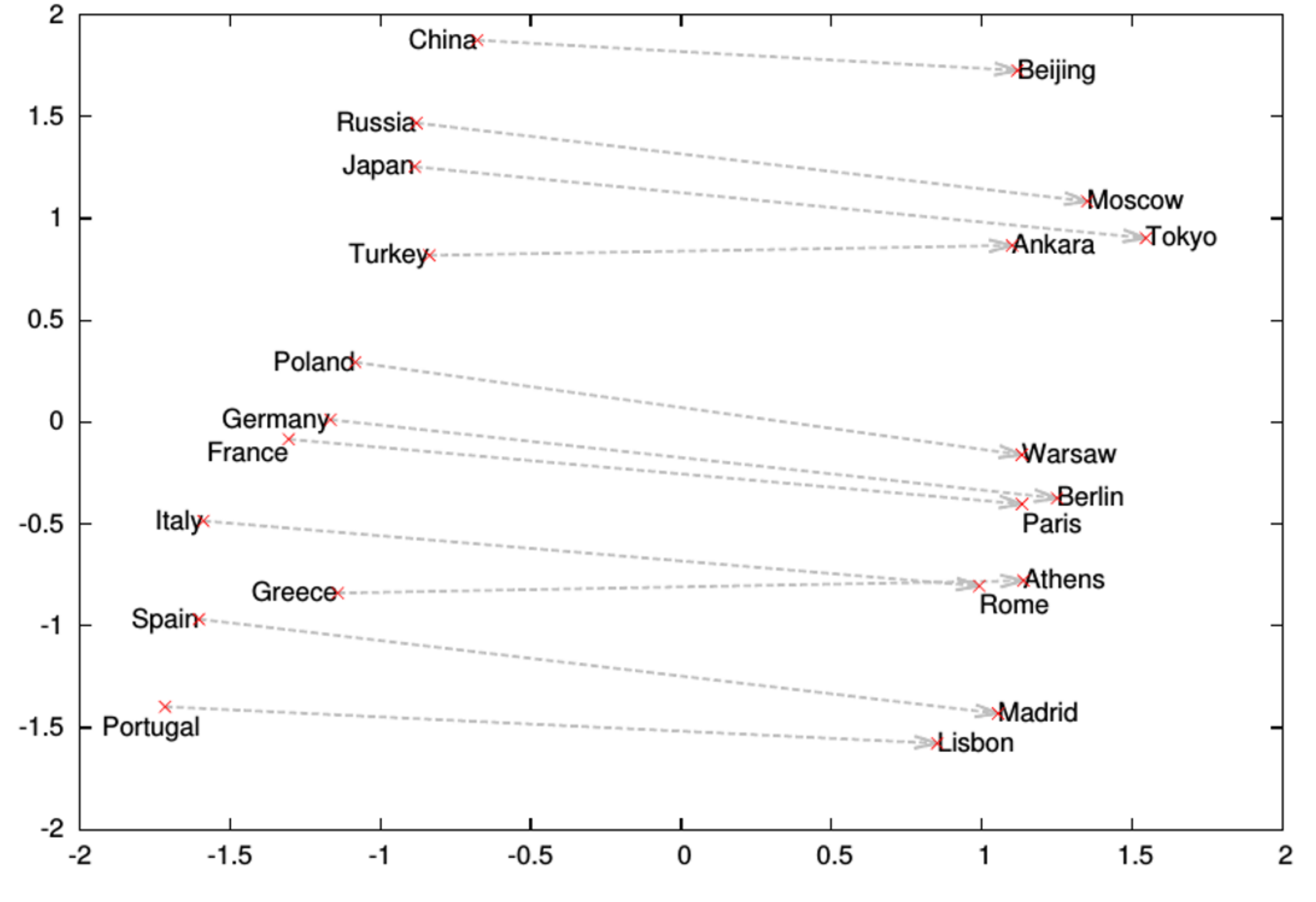
Test for linear relationships, examined by Mikolov et al. (2014)

$$\boxed{a:b :: c:?} \longrightarrow \boxed{d = \arg\max_x \frac{(w_b - w_a + w_c)^T w_x}{||w_b - w_a + w_c||}}$$

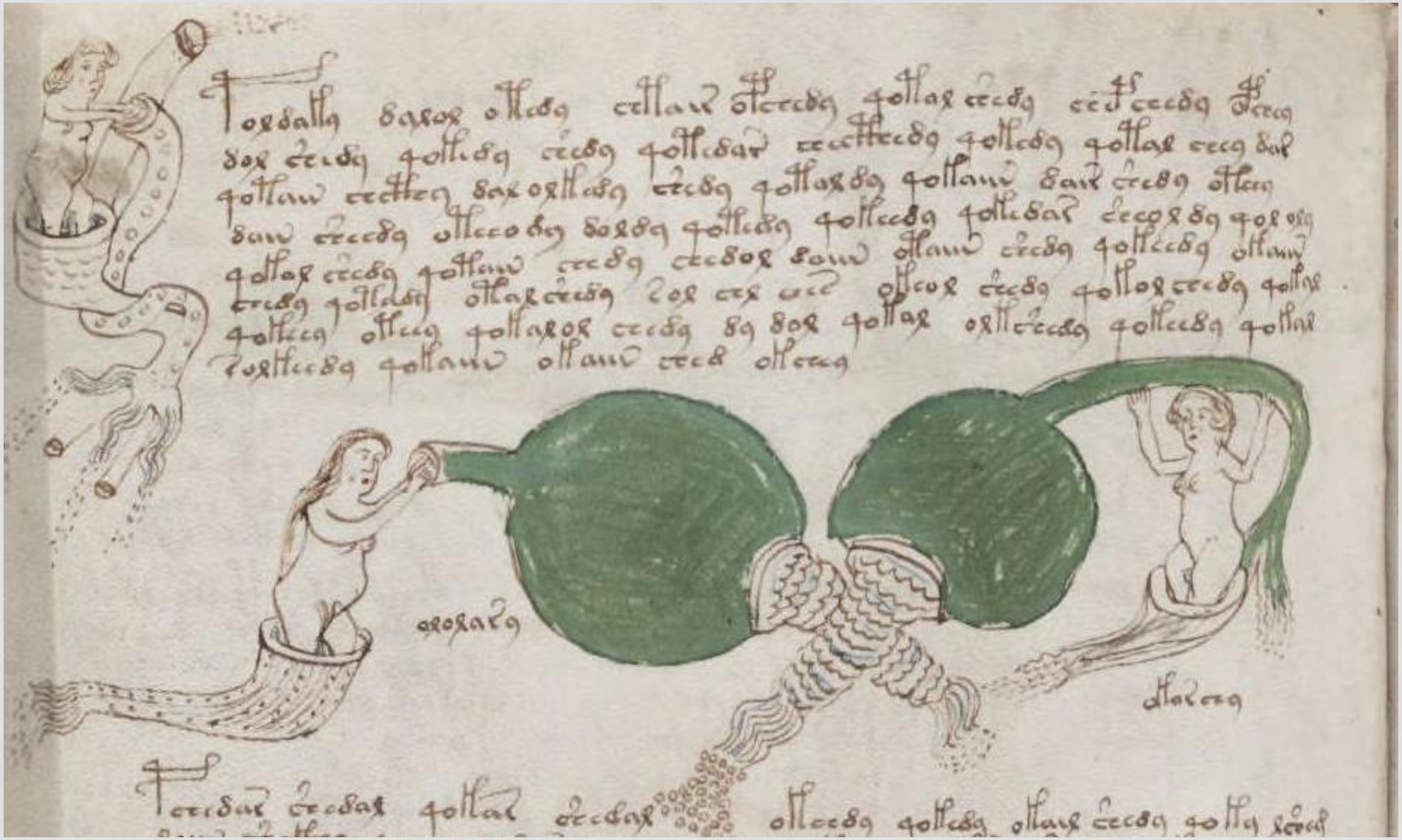
man:woman :: king:?

+	king	[0.30 0.70]
-	man	[0.20 0.20]
+	woman	[0.60 0.30]
	queen	[0.70 0.80]

WORD ANALOGIES



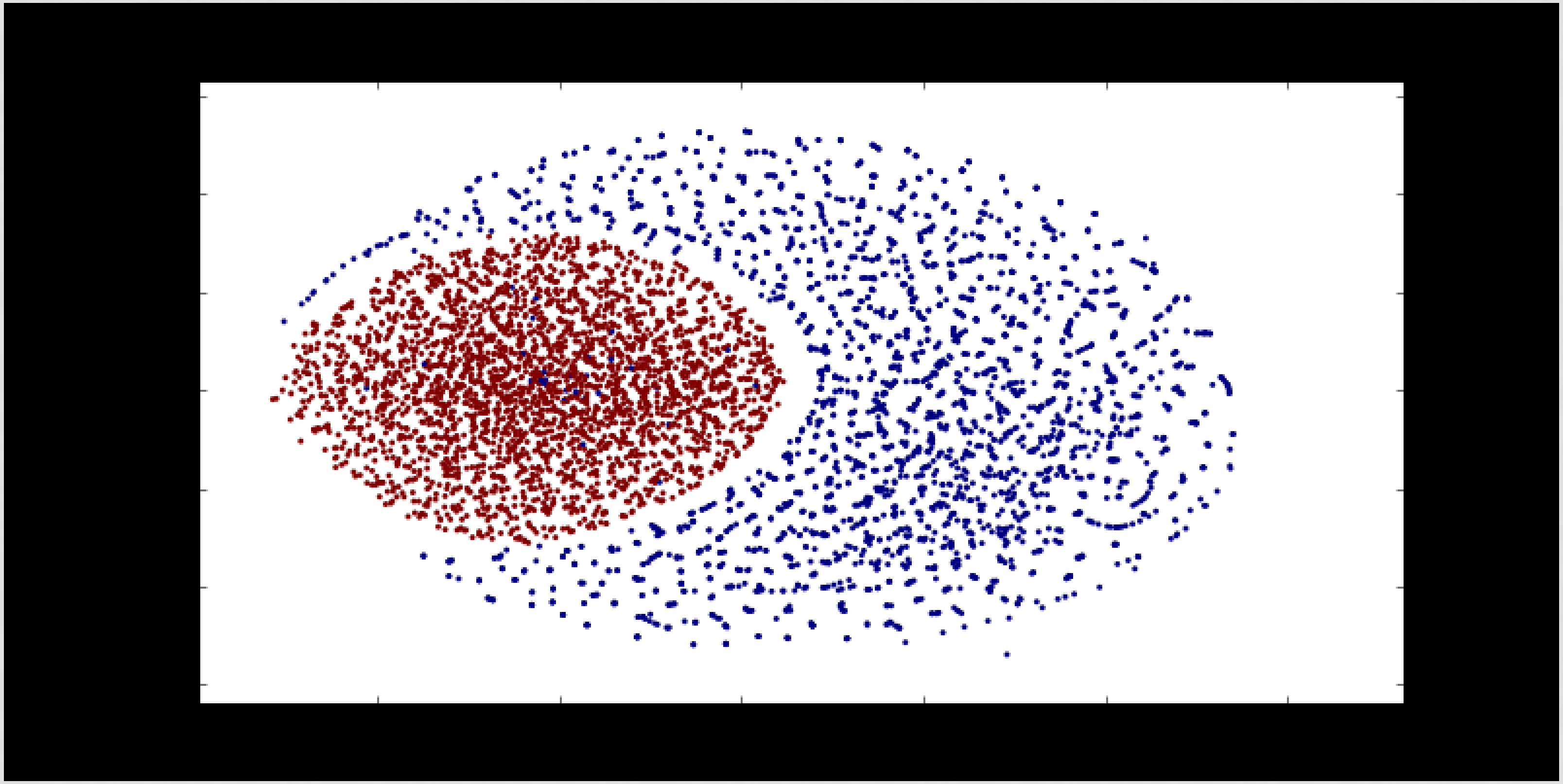
МАНУСКРИПТ ВОЙНИЧА



EVA TRANSCRIPTION

- To train this model, I had to parse and extract the transcription from the EVA (European Voynich Alphabet) to be able to feed the Voynich sentences into the word2vec model. This EVA transcription has the following format:
- <f1r.P1.1;U> fya!ys.ykal.ar.ytaiin.shol.shory.***!r*s.y.kor.sholdo*- #
- <f1r.P1.2;H> sory.ckhar.o!r.y.kair.chtaiin.shar.are.cthar.cthar.dan!-

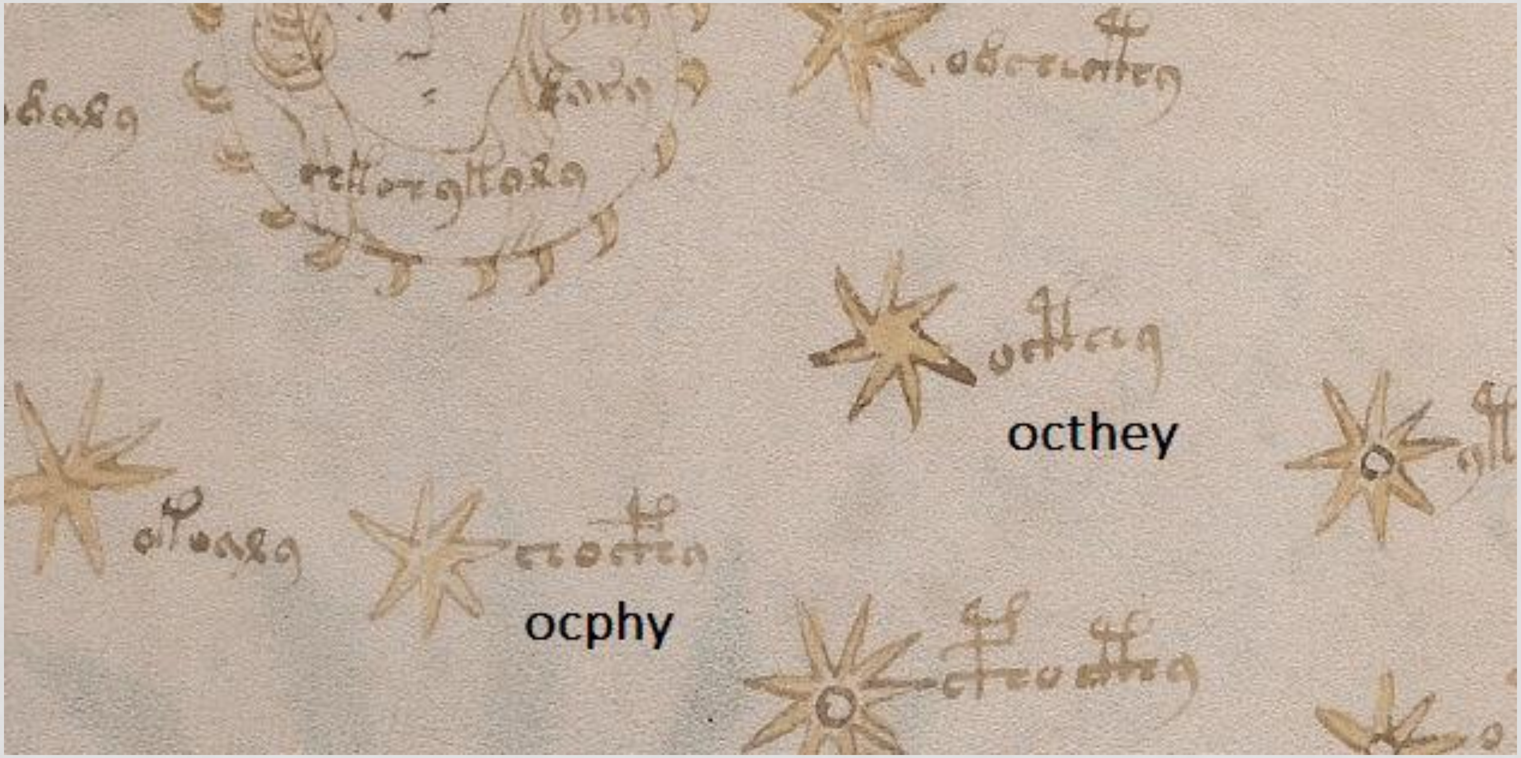
T-SNE VISUALIZATION



RESULTS

- >>> w2v_model.most_similar("octhey")
- [('qoekaiin', 0.6402825713157654),
- ('otcheody', 0.6389687061309814),
- ('ytchos', 0.566596269607544),
- ('ocphy', 0.5415685176849365),
- ('dolchedy', 0.5343093872070312),
- ('aiicthy', 0.5323750376701355),
- ('odchecthy', 0.5235849022865295),
- ('okeeos', 0.5187858939170837),
- ('cphocthy', 0.5159749388694763),
- ('oteor', 0.5050544738769531)]

„ASTRONOMICAL WORDS”



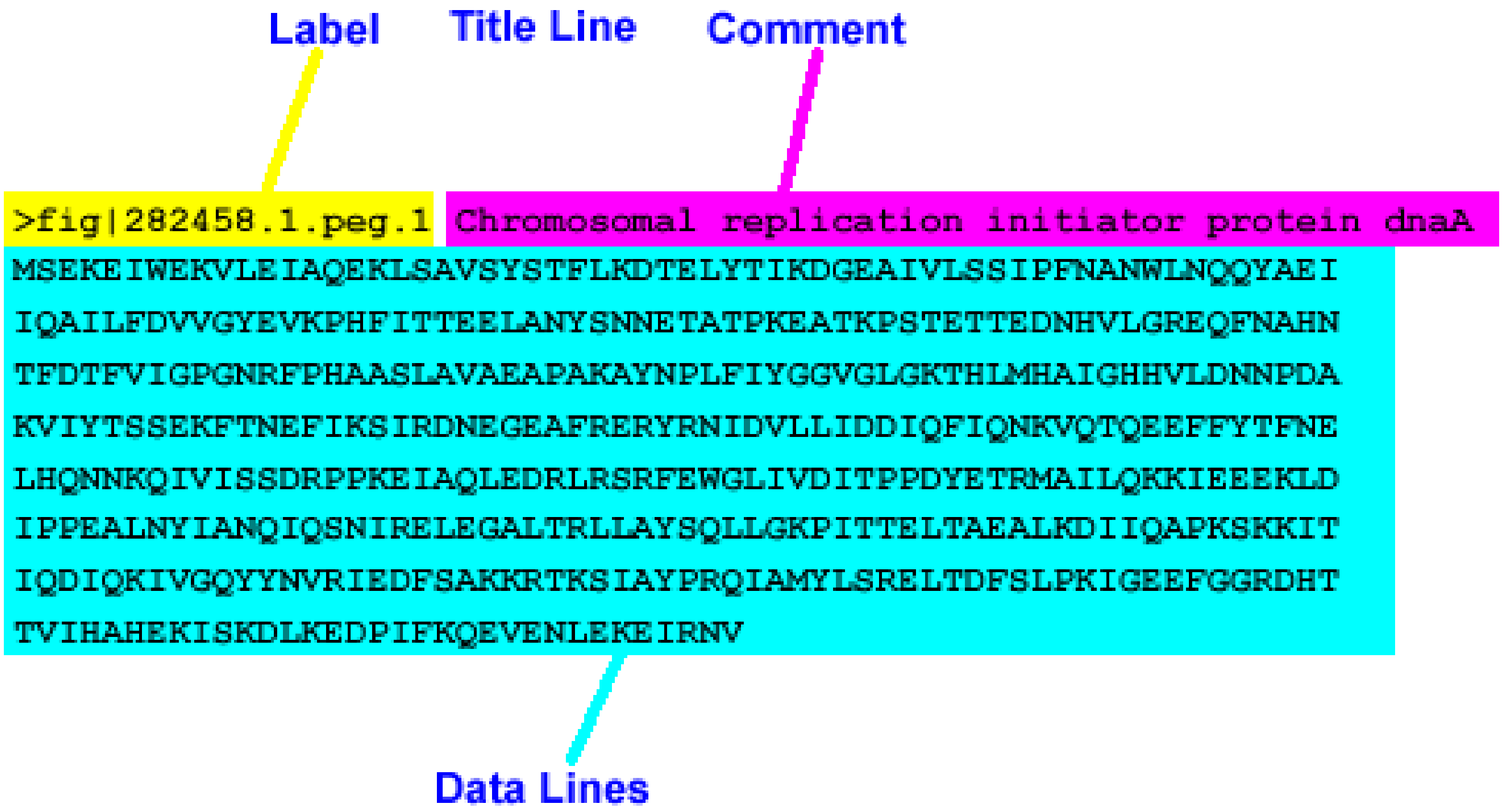
ВВЕДЕНИЕ: ЧТО ТАКОЕ БИОИНФОРМАТИКА

- математические методы компьютерного анализа генома, транскриптома, протеома (омикс- биоинформатика).
- разработка алгоритмов и программ для предсказания пространственной структуры биополимеров– РНК и белок - структурная биоинформатика ~ ФОЛДНИНГ
- [1]моделирование белковых каскадов,предсказание функции белка, регуляторных контуров и т.

SHOTGUN & NEXT GEN. SEQUENCING

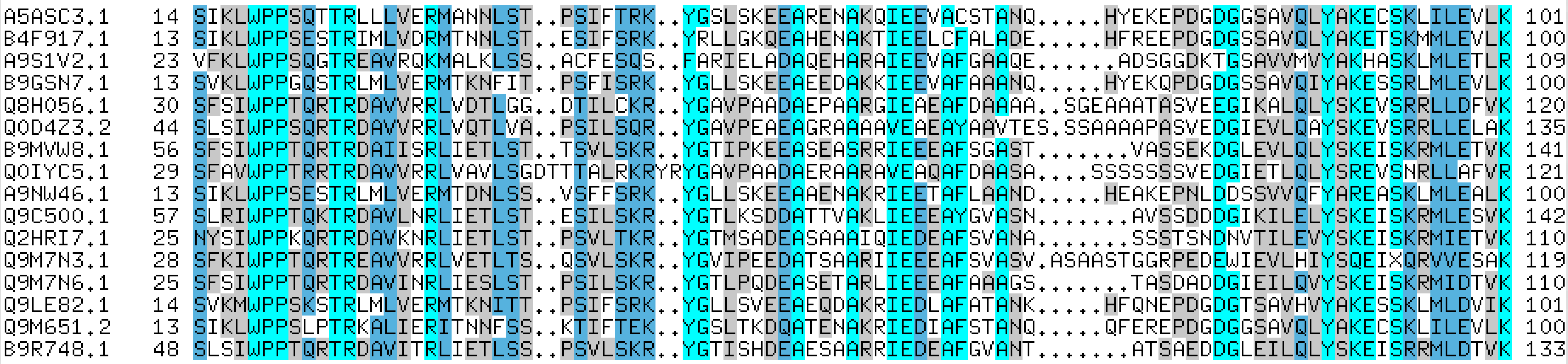
Strand	Sequence
Original	AGCATGCTGCAGTCATGCTTAGG CTA
First shotgun sequence	AGCATGCTGCAGTCATGCT------- -------------------TAGGCTA
Second shotgun sequence	AGCATG-------------------- ------CTGCAGTCATGCTTAGGCTA
Reconstruction	AGCATGCTGCAGTCATGCTTAGG CTA

ПРИМЕР БЕЛКОВОЙ ПОСЛЕДОВАТЕЛЬНОСТИ



```
>fig|282458.1.peg.1 Chromosomal replication initiator protein dnaA
MSEKEIWEKVLEIAQEKLSAVSYSTFLKDTELYTIKDGEAIVLSSIPFNANWLNQQYAEI
IQAILFDVVGYEVKPHFITTEELANYSNNETATPKEATKPSTETTEDNHVLGREQFNAHN
TFDTFVIGPGNRFPHAASLAVAEAPAKAYNPLFIYGGVGLGKTHLMHAIGHHVLDNNPDA
KVIYTSSEKFTNEFIKSIRDNEGEAFRERYRNIDVLLIDDIQFIQNKVQTQEEFFYTFNE
LHQNNKQIVISSDRPPKEIAQLEDRLRSRFEWGLIVDITPPDYETRMAILQKKIEEEKLD
IPPEALNYIANQIQSNIRELEGALTRLLAYSQLLGKPITTELTAEALKDIIQAPKSKKIT
IQDIQKIVGQYYNVRIEDFSAKKRTKSIAYPRQIAMYLSRELTDFSLPKIGEEFGGRDHT
TVIHAHEKISKDLKEDPIFKQEVENLEKEIRNV
```

ВЫРАВНИВАНИЕ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

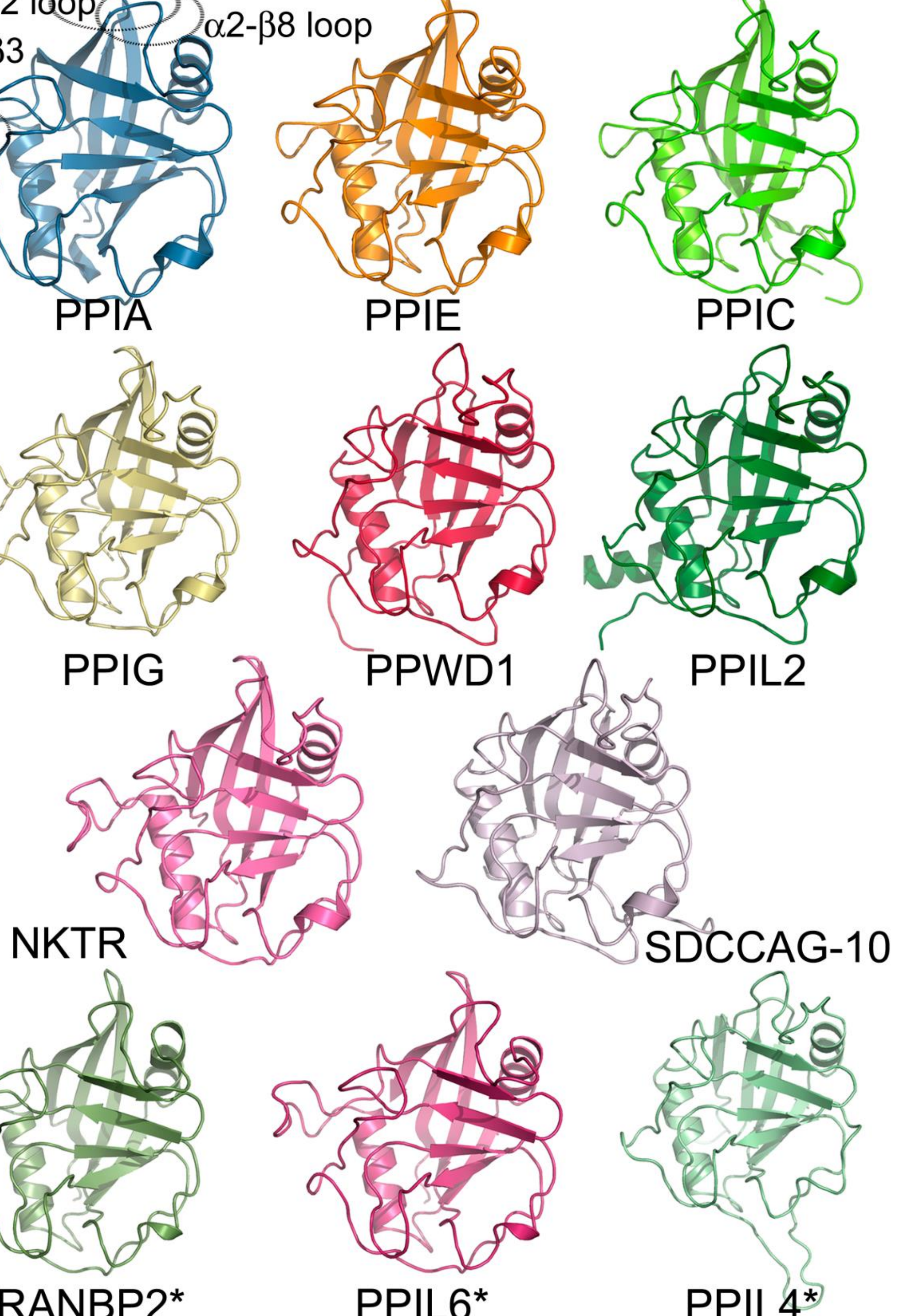


```
A5ASC3.1   14 SIKLWPPSQTTRLLLVERMANNLST..PSIFTRK..YGSLSKEEARENAKQIEEVACSTANQ.....HYEKEPDGDGGSAVQLYAKECSKLILEVLK 101
B4F917.1   13 SIKLWPPSESTRIMLVDRMTNNLST..ESIFSRK..YRLLGKQEAHENAKTIEELCFALADE.....HFREEPDGDGSSAVQLYAKETSKMMLEVLK 100
A9S1V2.1   23 VFKLWPPSQGTREAVRQKMALKLSS..ACFESQS..FARIELADAQEHARAIEEVAFGAAQE......ADSGGDKTGSAVVMVYAKHASKLMLETLR 109
B9GSN7.1   13 SVKLWPPGQSTRLMLVERMTKNFIT..PSFISRK..YGLLSKEEAEEDAKKIEEVAFAAANQ.....HYEKQPDGDGSSAVQIYAKESSRLMLEVLK 100
Q8H056.1   30 SFSIWPPTQRTRDAVVRRLVDTLGG..DTILCKR..YGAVPAADAEPAARGIEAEAFDAAAA..SGEAAATASVEEGIKALQLYSKEVSRRLLDFVK 120
Q0D4Z3.2   44 SLSIWPPSQRTRDAVVRRLVQTLVA..PSILSQR..YGAVPEAEAGRAAAAVEAEAYAAVTES.SSAAAAPASVEDGIEVLQAYSKEVSRRLLELAK 135
B9MVW8.1   56 SFSIWPPTQRTRDAIISRLIETLST..TSVLSKR..YGTIPKEEASEASRRIEEEAFSGAST.......VASSEKDGLEVLQLYSKEISKRMLETVK 141
Q0IYC5.1   29 SFAVWPPTRRTRDAVVRRLVAVLSGDTTTALRKRYRYGAVPAADAERAARAVEAQAFDAASA....SSSSSSSVEDGIETLQLYSREVSNRLLAFVR 121
A9NW46.1   13 SIKLWPPSESTRLMLVERMTDNLSS..VSFFSRK..YGLLSKEEAAENAKRIEETAFLAAND.....HEAKEPNLDDSSVVQFYAREASKLMLEALK 100
Q9C500.1   57 SLRIWPPTQKTRDAVLNRLIETLST..ESILSKR..YGTLKSDDATTVAKLIEEEAYGVASN.......AVSSDDDGIKILELYSKEISKRMLESVK 142
Q2HRI7.1   25 NYSIWPPKQRTRDAVKNRLIETLST..PSVLTKR..YGTMSADEASAAAIQIEDEAFSVANA.......SSSTSNDNVTILEVYSKEISKRMIETVK 110
Q9M7N3.1   28 SFKIWPPTQRTREAVVRRLVETLTS..QSVLSKR..YGVIPEEDATSAARIIEEEAFSVASV.ASAASTGGRPEDEWIEVLHIYSQEIXQRVVESAK 119
Q9M7N6.1   25 SFSIWPPTQRTRDAVINRLIESLST..PSILSKR..YGTLPQDEASETARLIEEEAFAAAGS.......TASDADDGIEILQVYSKEISKRMIDTVK 110
Q9LE82.1   14 SVKMWPPSKSTRLMLVERMTKNITT..PSIFSRK..YGLLSVEEAEQDAKRIEDLAFATANK.....HFQNEPDGDGTSAVHVYAKESSKLMLDVIK 101
Q9M651.2   13 SIKLWPPSLPTRKALIERITNNFSS..KTIFTEK..YGSLTKDQATENAKRIEDIAFSTANQ.....QFEREPDGDGGSAVQLYAKECSKLILEVLK 100
B9R748.1   48 SLSIWPPTQRTRDAVITRLIETLSS..PSVLSKR..YGTISHDEAESAARRIEDEAFGVANT.......ATSAEDDGLEILQLYSKEISRRMLDTVK 133
```

- Обнаружение внутривидового и межвидового полиморфизма.
- Таксономия
- Молекулярные часы

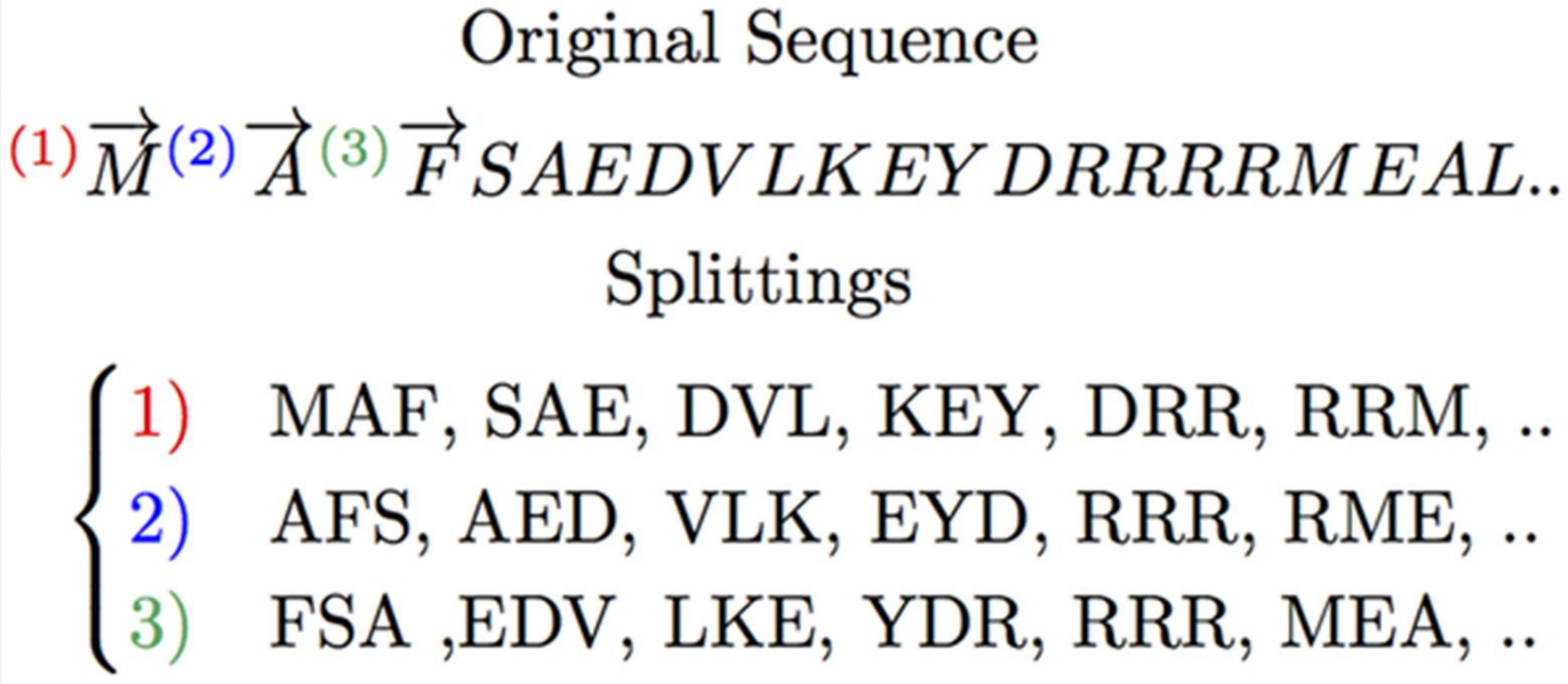
ОСНОВНАЯ СТАТЬЯ

- Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics
- Ehsaneddin Asgari,Mohammad R. K. Mofrad
- PLOS ONE November 10, 2015
- https://doi.org/10.1371/journal.pone.0141287



СЕМЕЙСТВА БЕЛКОВ

РАЗБИВКА БЕЛКОВОЙ ПОСЛЕДОВАТЕЛЬНОСТИ



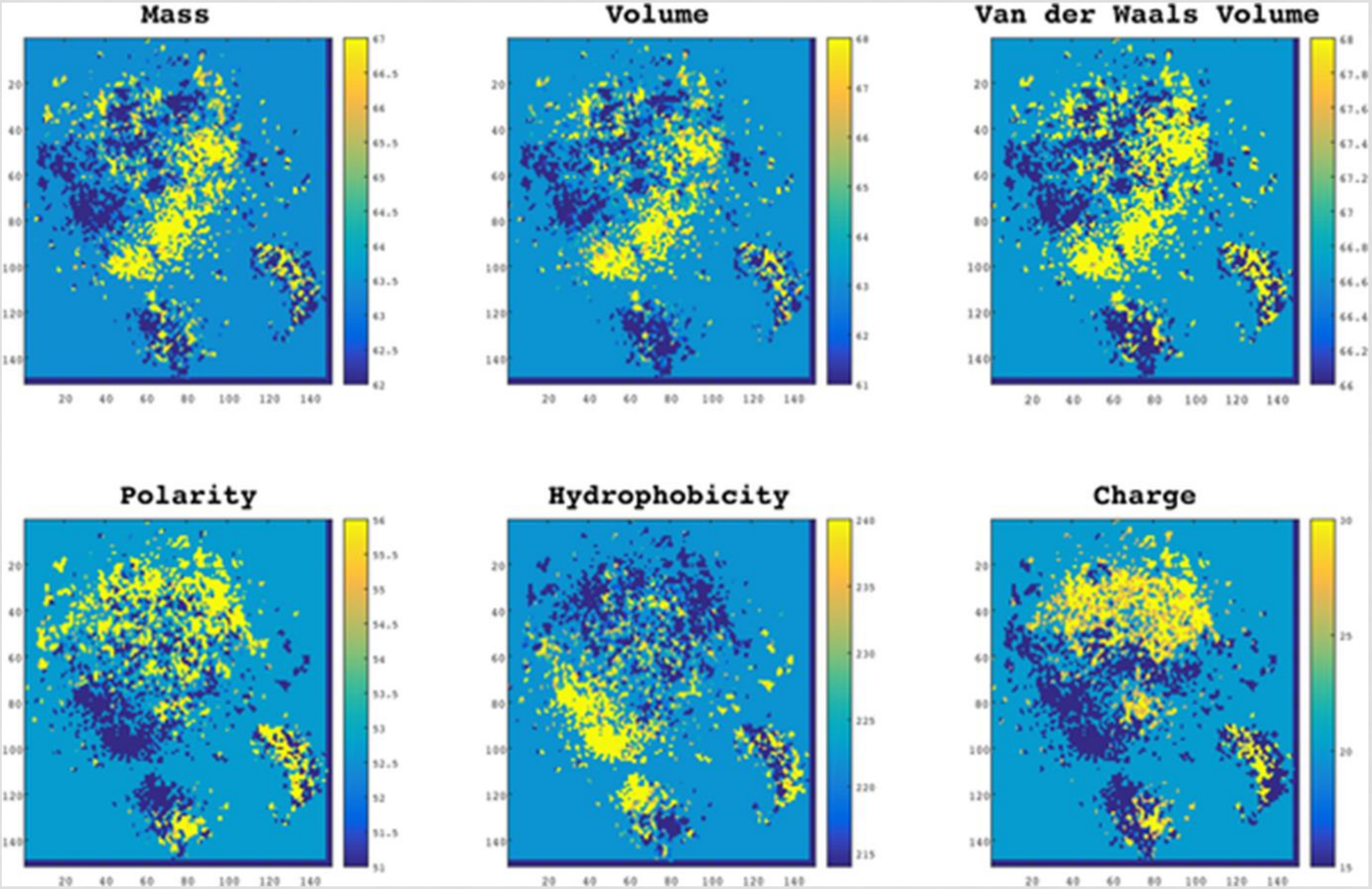
Original Sequence

$$^{(1)}\overrightarrow{M}^{(2)}\overrightarrow{A}^{(3)}\overrightarrow{F}SAEDVLKEYDRRRRMEAL..$$

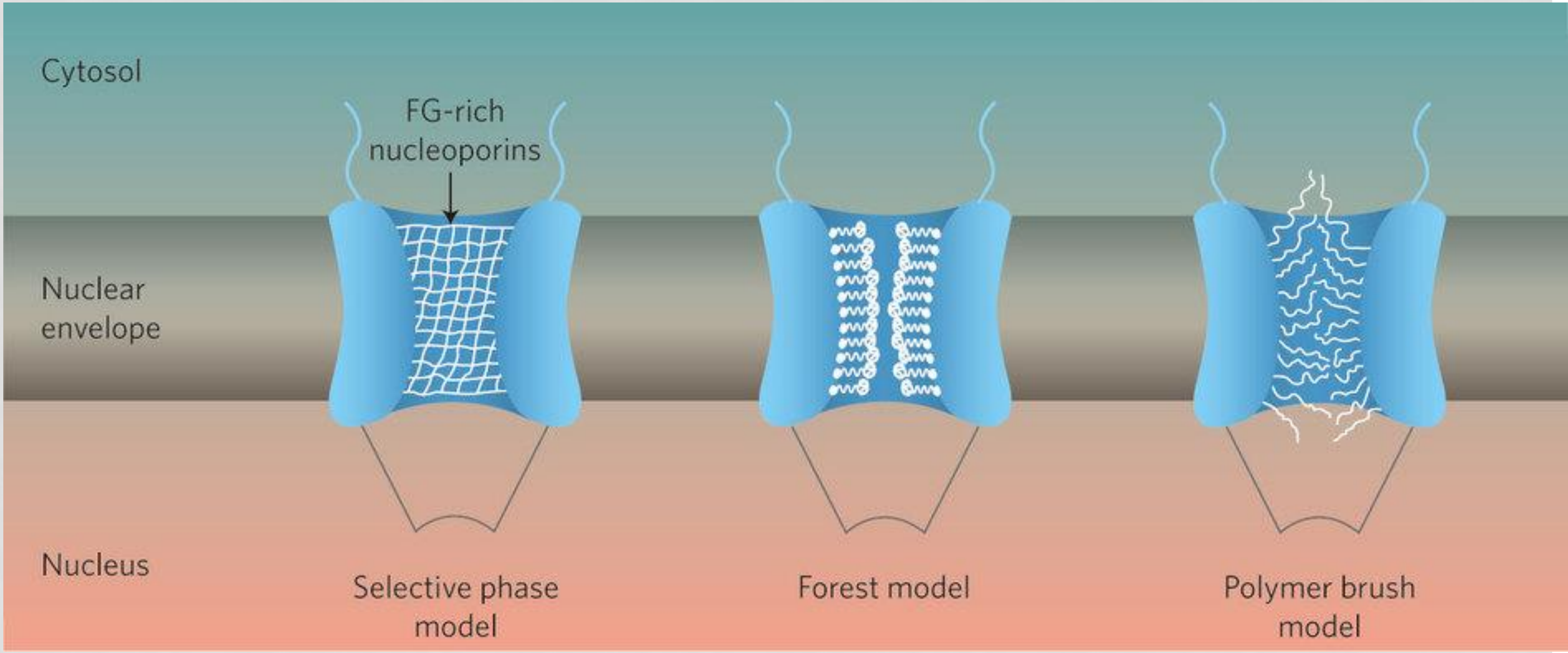
Splittings

$$\left\{\begin{array}{ll} 1) & \text{MAF, SAE, DVL, KEY, DRR, RRM, ..} \\ 2) & \text{AFS, AED, VLK, EYD, RRR, RME, ..} \\ 3) & \text{FSA ,EDV, LKE, YDR, RRR, MEA, ..} \end{array}\right.$$

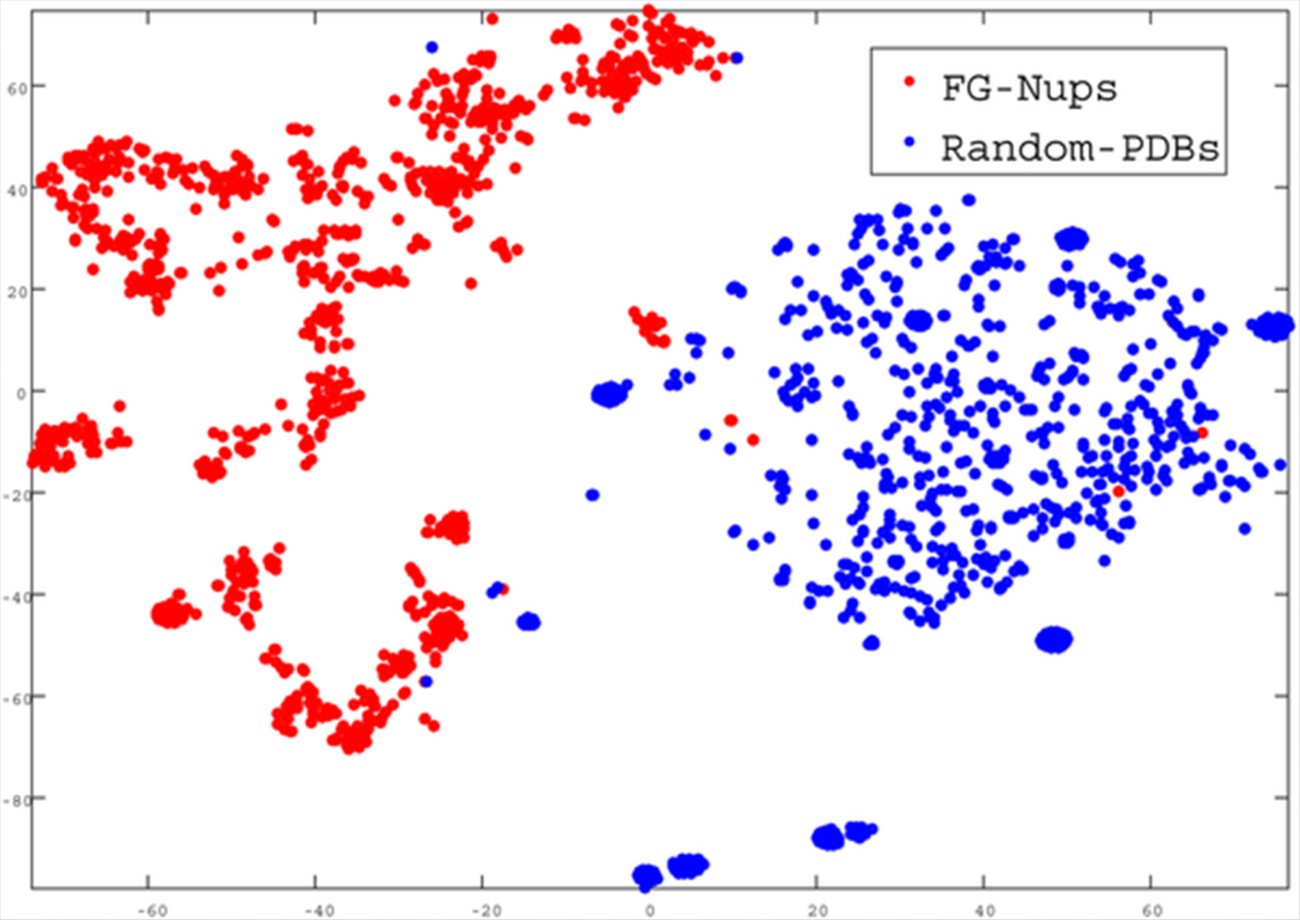
РАСПРЕДЕЛЕНИЕ БЕЛКОВ В ПРОСТРАНСТВЕ 2Х КОМПОНЕНТ
ЦВЕТ ОБОЗНАЧАЕТ ЗНАЧЕНИЕ СООТВ. ПРИЗНАКА



PHENYLALANINE-GLYCINE NUCLEOPORINS (FG-NUPS)



ПОСЛЕДОВАТЕЛЬНОСТИ FG-NUP VS СТРУКТУРИРОВАННЫЕ БЕЛКОВЫЕ ПОСЛЕДОВАТЕЛЬНОСТИ

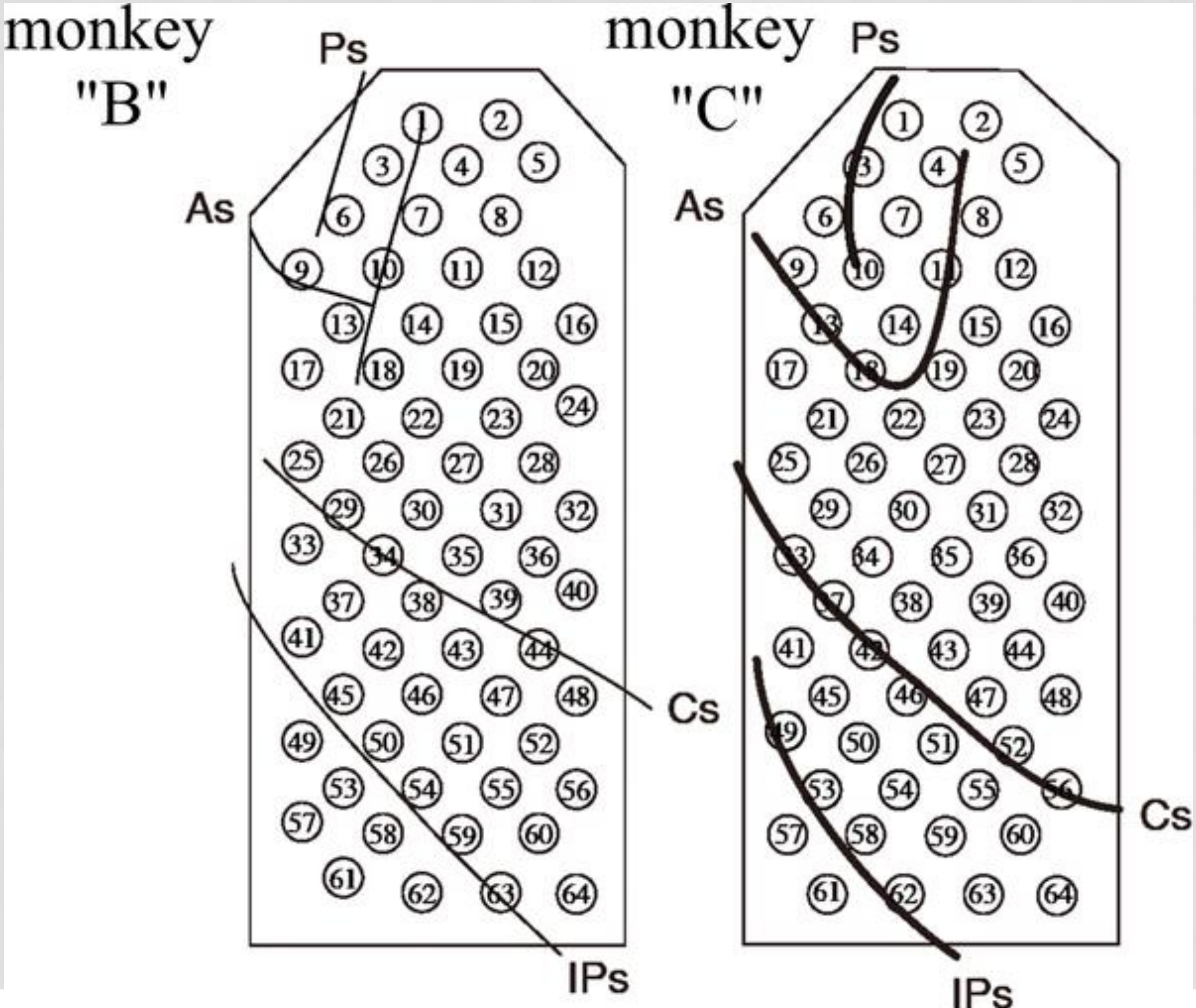


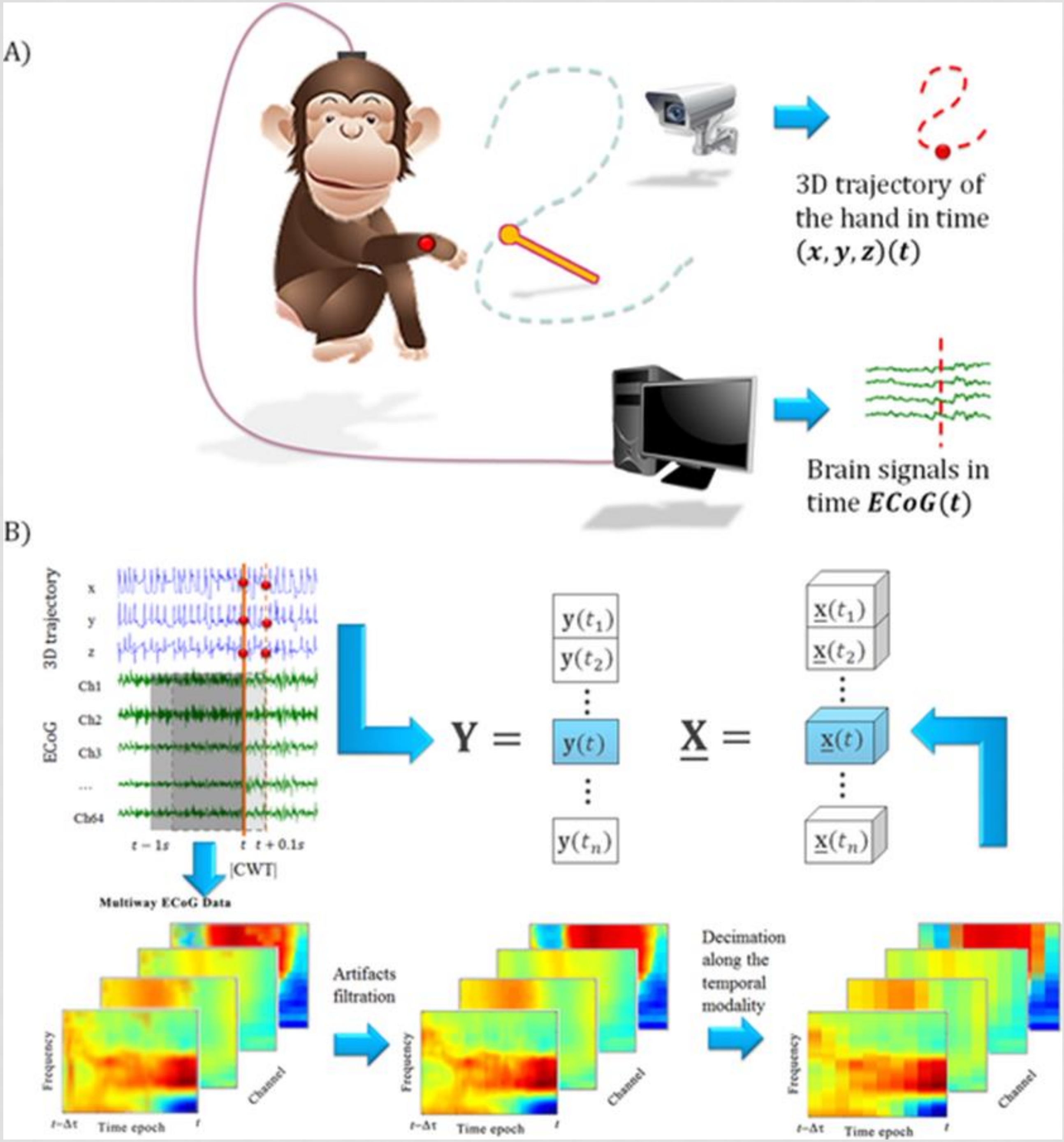
РЕАЛИЗАЦИЯ

- https://github.com/peter-volkov/biovec
- **https://github.com/ehsanasgari/Deep-Proteomics**

Electrocorticography (ECoG)

Electrodes fo epidural recordings.




A)
3D trajectory of the hand in time $(x, y, z)(t)$
Brain signals in time $ECoG(t)$
B)
3D trajectory
x
y
z
ECoG
Ch1
Ch2
Ch3
...
Ch64
$t-1s$
t
$t+0.1s$
|CWT|
Multiway ECoG Data
$\mathbf{Y} =$
$\mathbf{y}(t_1)$
$\mathbf{y}(t_2)$
$\mathbf{y}(t)$
$\mathbf{y}(t_n)$
$\underline{\mathbf{X}} =$
$\underline{\mathbf{x}}(t_1)$
$\underline{\mathbf{x}}(t_2)$
$\underline{\mathbf{x}}(t)$
$\underline{\mathbf{x}}(t_n)$
Frequency
Time epoch
$t-\Delta t$
t
Channel
Artifacts filtration
Decimation along the temporal modality

СПАСИБО ЗА ВНИМАНИЕ!