16th Deep Learning Meetup in Vienna

Vienna, February 2018

Demystifying Neural Word Embedding

Applications in Financial Sentiment Analysis, and Gender Bias Detection

Navid Rekabsaz

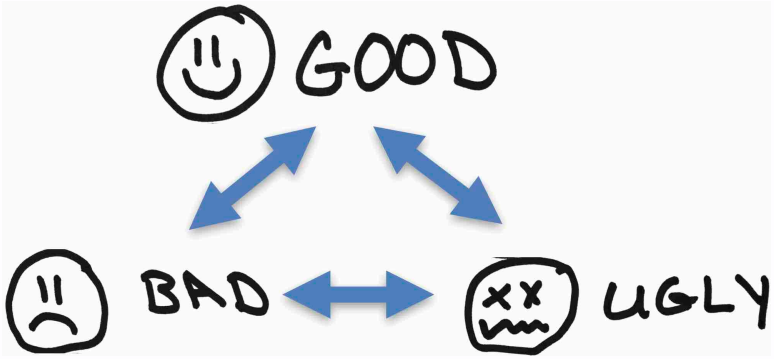
TU Wien (current)

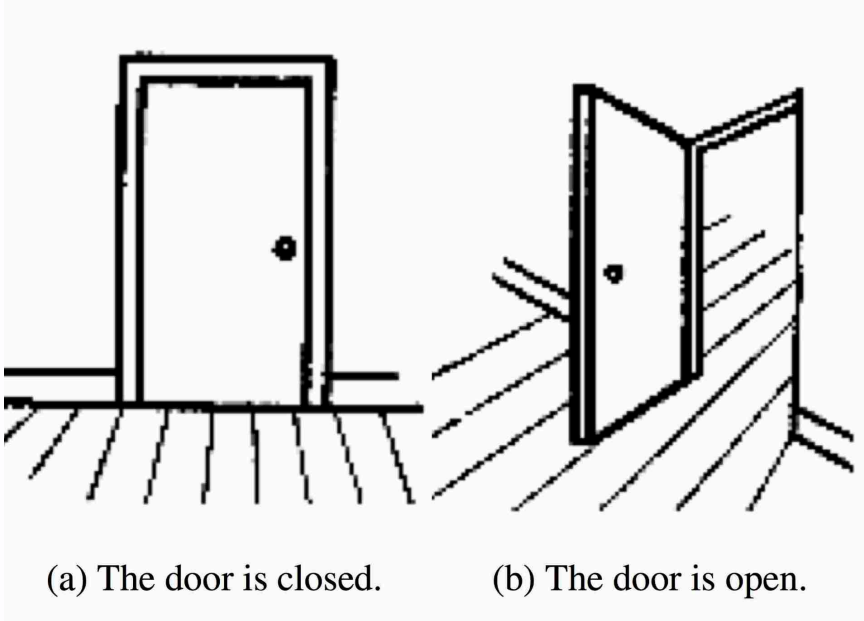
Idiap Research Institute – EPFL (next)


@navidrekabsaz


rekabsaz@ifs.tuwien.ac.at

Semantics

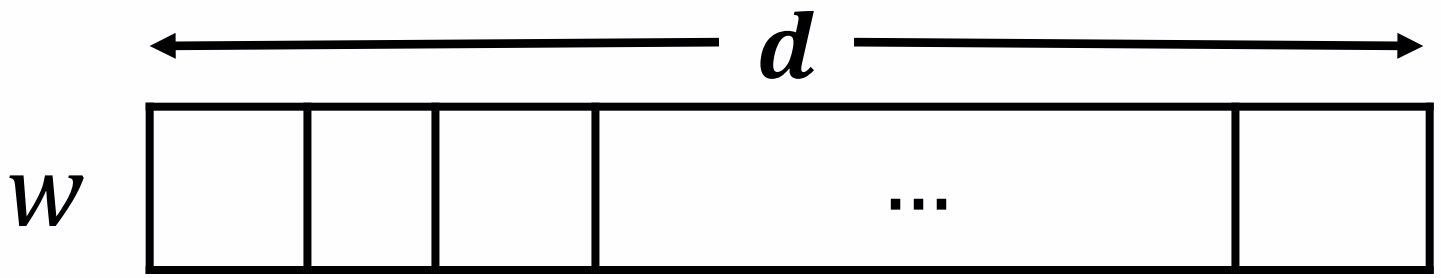




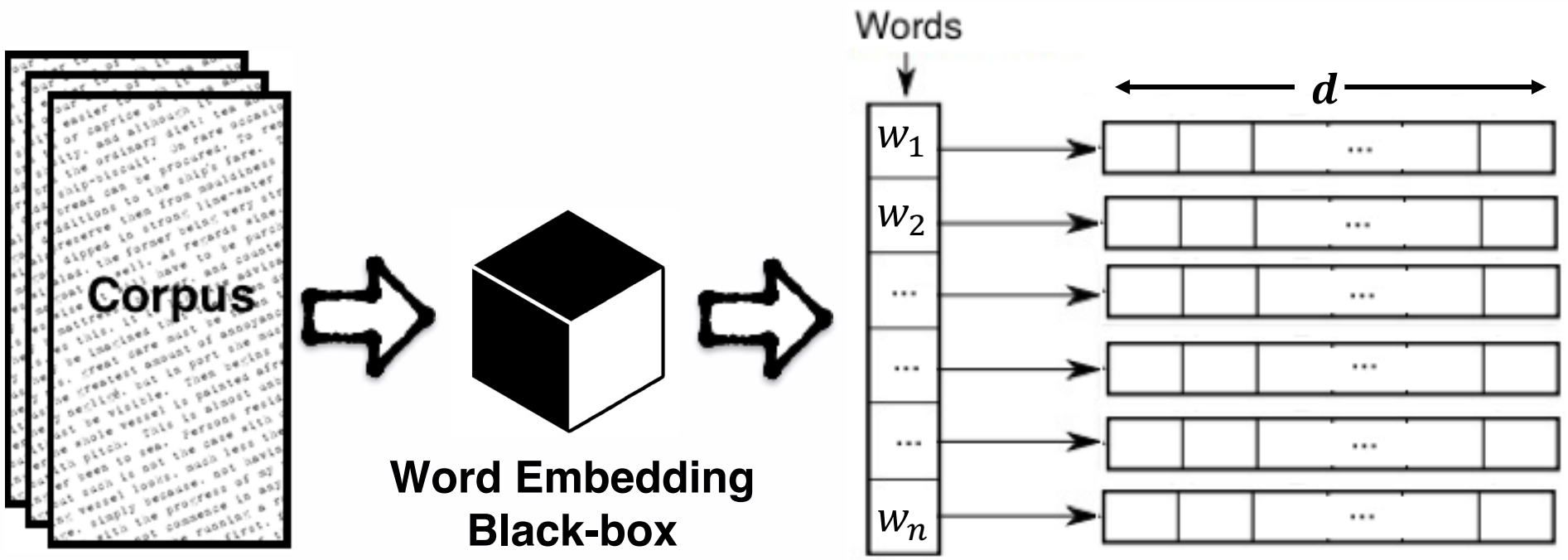
(a) The door is closed. (b) The door is open.

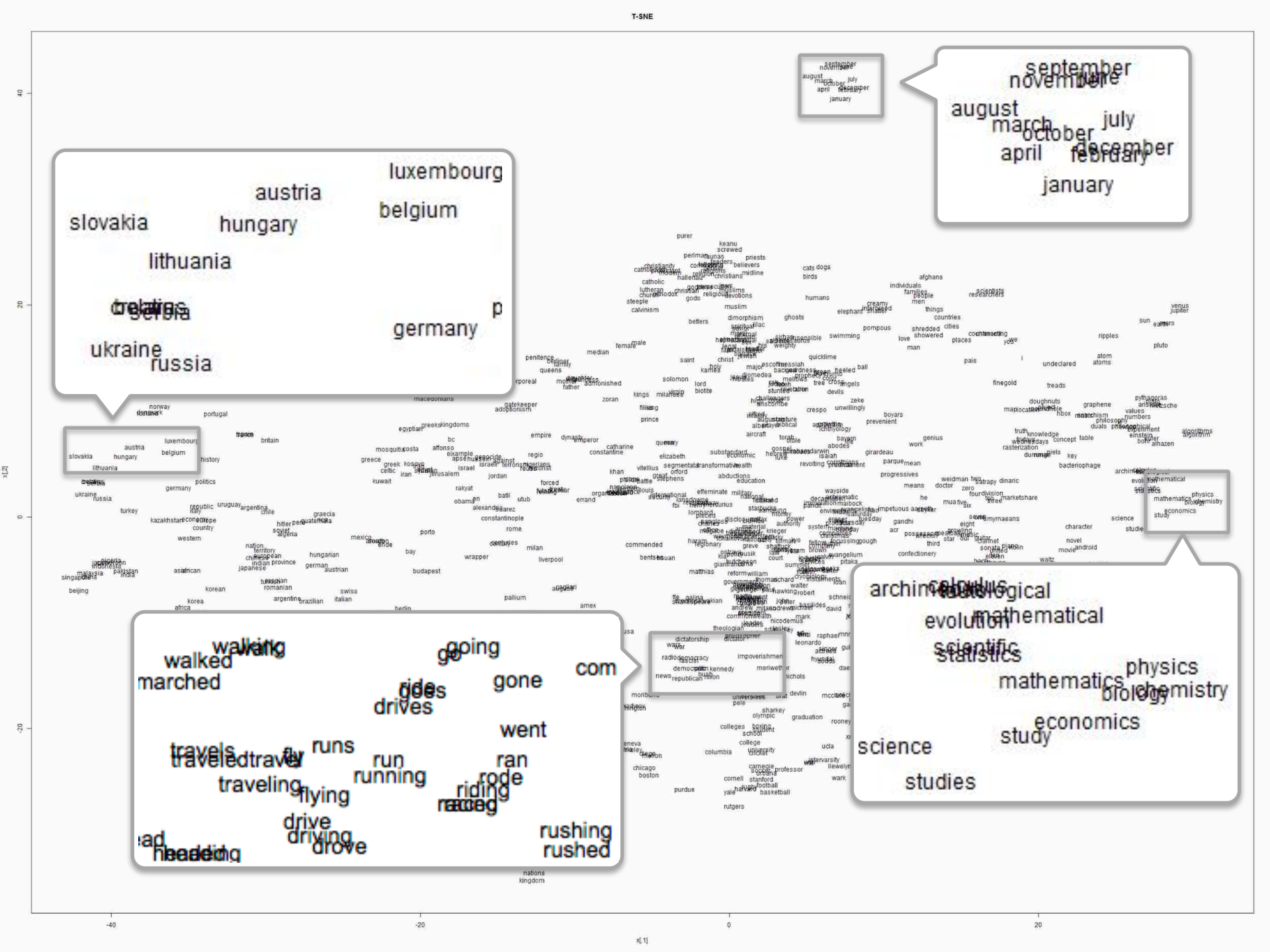
Semantic Vectors

- A vector with *d* dimensions represents each word
  - Also called **Word Embedding**
- Dimensions reflect the *concepts* in language



Word Embedding

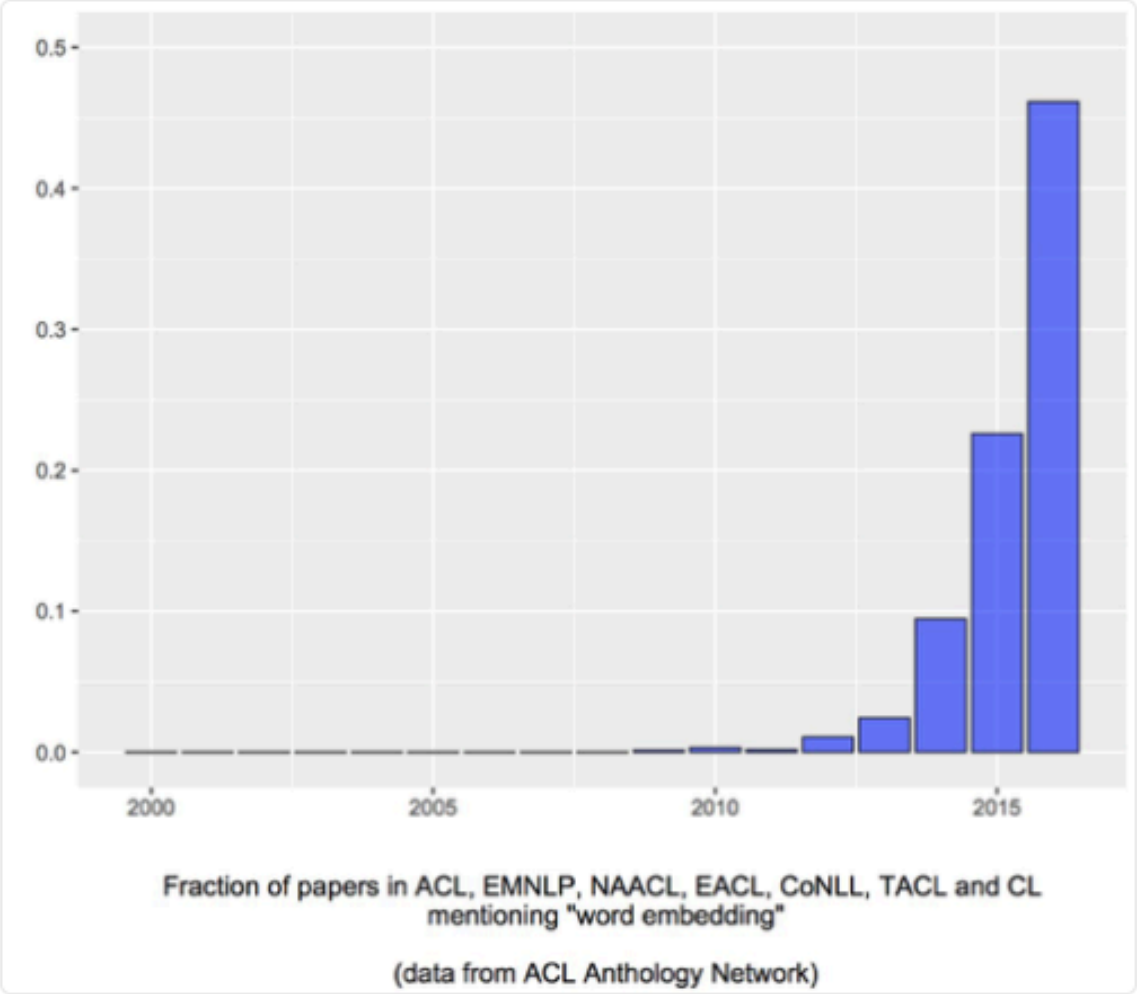



T-SNE
luxembourg
austria
belgium
slovakia
hungary
lithuania
germany
ukraine
russia
september
november
august
march
july
october
december
april
february
january
walked
walking
going
gone
marched
drives
went
travels
runs
ran
running
traveling
flying
riding
drive
driving
drove
rushing
rushed
mathematical
evolution
scientific
statistics
physics
mathematics
chemistry
biology
economics
study
science
studies
x[,1]
x[,2]

David Bamman
@dbamman

Following

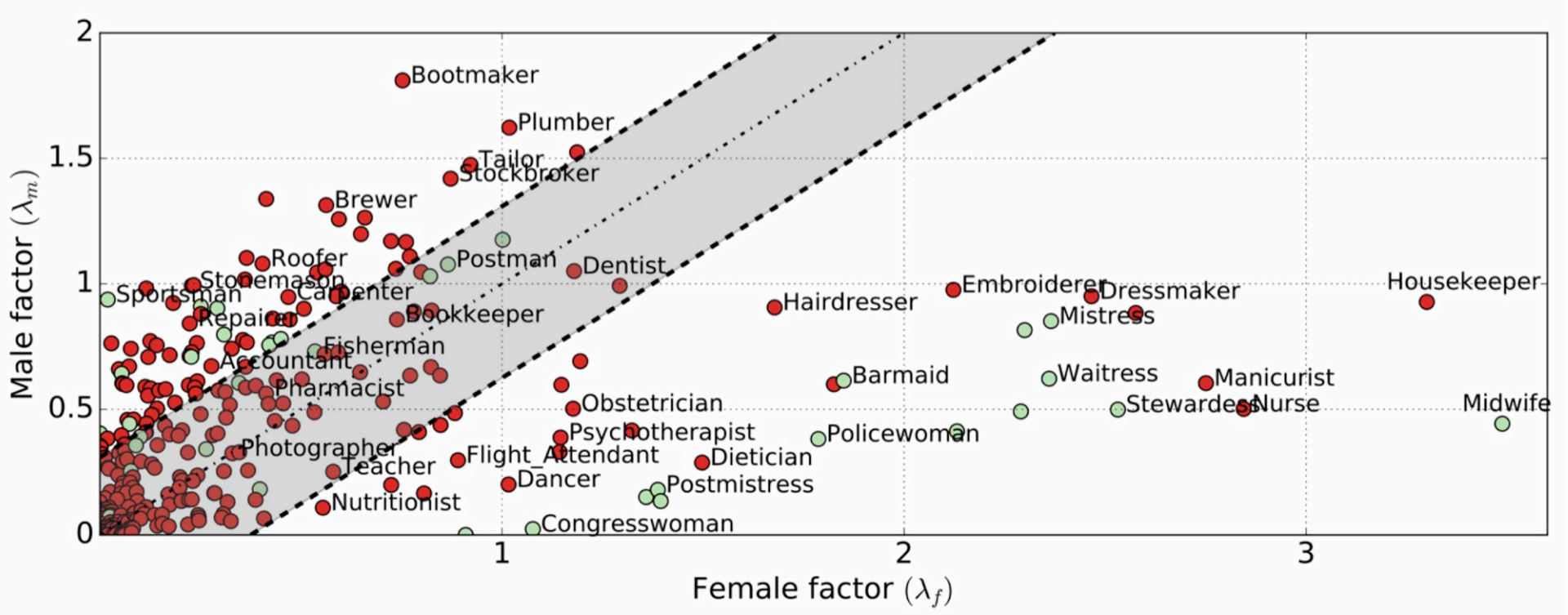
The recent history of NLP: peak embeddings



6:24 PM - 25 Feb 2018

Word Embedding for Gender Bias Study

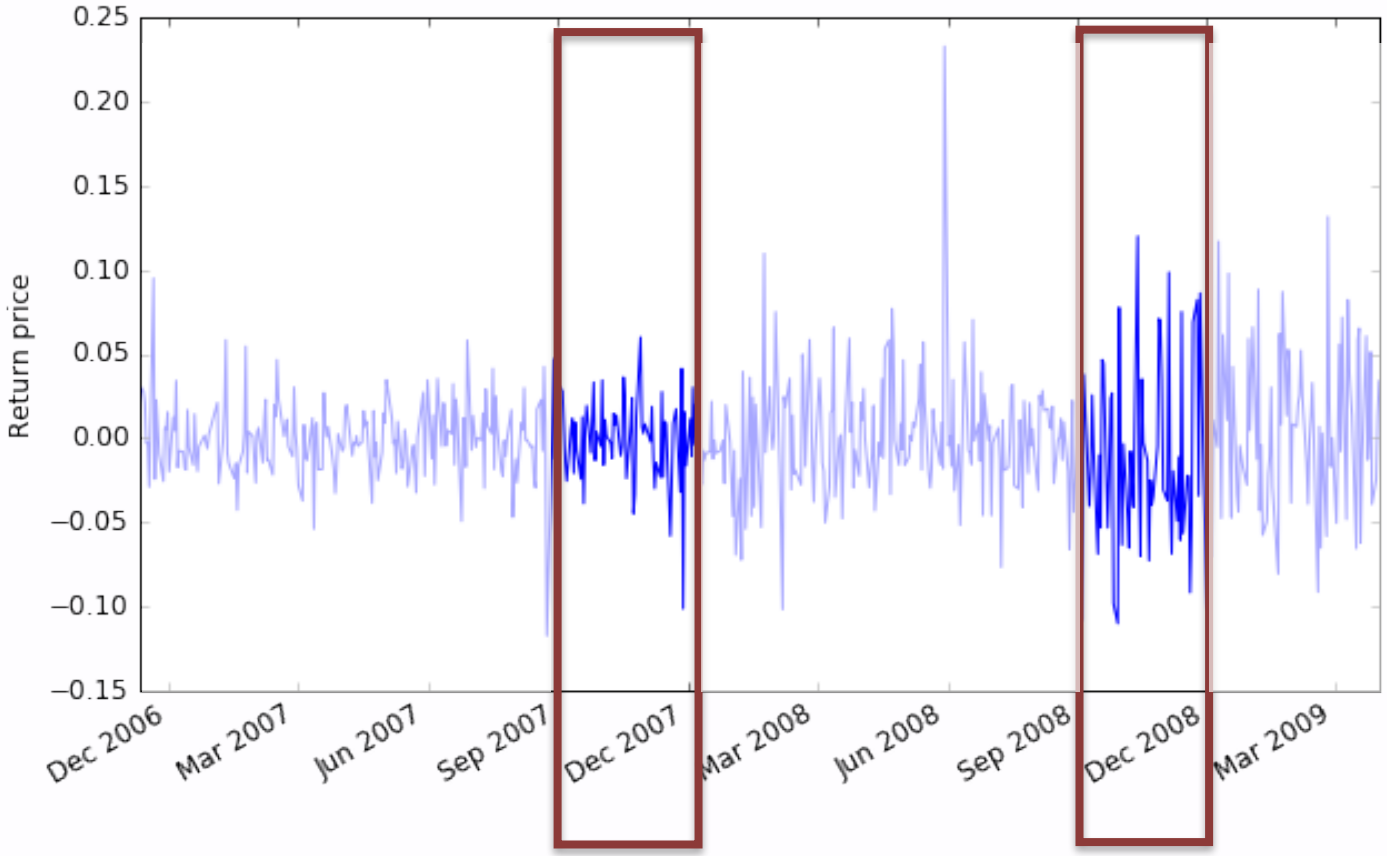
- Processing the text of English Wikipedia
- The inclinations of 350 occupations to female/male factors



Word Embedding for Financial Volatility Prediction

return price=(price$_{(t)}$ / price$_{(t-1)}$)-1

volatility=log(std(return prices)) Kogan et al. [2009]



Volatility Prediction using Financial Disclosures Sentiments with Word Embedding-based IR Models
Navid Rekabsaz, Mihai Lupu, Artem Baklanov, Allan Hanbury, Alexander Duer, Linda Anderson
In Proceedings of the Annual Meeting of the Association for Computational Linguistics (**ACL 2017**)

Companies Annual Reports

UNITED STATES
SECURITIES AND EXCHANGE COMMISSION
Washington, D.C. 20549

FORM 10-K

☒ **ANNUAL REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934**
For the fiscal year ended May 31, 2011
OR

☐ **TRANSITION REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934**
For the transition period from ____ to ____
Commission file number: 000-51788

Oracle Corporation

(Exact name of registrant as specified in its charter)

Delaware	**54-2185193**
(State or other jurisdiction of incorporation or organization)	**(I.R.S. Employer Identification No.)**
500 Oracle Parkway **Redwood City, California** **(Address of principal executive offices)**	**94065** **(Zip Code)**

(650) 506-7000
(Registrant's telephone number, including area code)

Securities registered pursuant to Section 12(b) of the Act:

Title of each class	**Name of each exchange on which registered**
Common Stock, par value $0.01 per share	The NASDAQ Stock Market LLC

Securities registered pursuant to Section 12(g) of the Act:
None

Indicate by check mark if the registrant is a well-known seasoned issuer, as defined in Rule 405 of the Securities Act. YES ☒ NO ☐

Indicate by check mark if the registrant is not required to file reports pursuant to Section 13 or Section 15(d) of the Act. YES ☐ NO ☒

Indicate by check mark whether the registrant (1) has filed all reports required to be filed by Section 13 or 15(d) of the Securities Exchange Act of 1934 during the preceding 12 months (or for such shorter period that the registrant was required to file such reports), and (2) has been subject to such filing requirements for the past 90 days. YES ☒ NO ☐

Indicate by check mark whether the registrant has submitted electronically and posted on its corporate Website, if any, every Interactive Data File required to be submitted and posted pursuant to Rule 405 of Regulation S-T (§232.405 of this chapter) during the preceding 12 months (or for such shorter period that the registrant was required to submit and post such files). YES ☒ NO ☐

Indicate by check mark if disclosure of delinquent filers pursuant to Item 405 of Regulation S-K (§229.405 of this chapter) is not contained herein, and will not be contained, to the best of registrant's knowledge, in definitive proxy or information statements incorporated by reference in Part III of this Form 10-K or any amendment to this Form 10-K. ☐

Indicate by check mark whether the registrant is a large accelerated filer, an accelerated filer, a non-accelerated filer, or a smaller reporting company. See the definitions of "large accelerated filer," "accelerated filer" and "smaller reporting company" in Rule 12b-2 of the Exchange Act.

Large accelerated filer ☒	Accelerated filer ☐
Non-accelerated filer ☐ (Do not check if a smaller reporting company)	Smaller reporting company ☐

Indicate by check mark whether the registrant is a shell company (as defined in Rule 12b-2 of the Exchange Act). YES ☐ NO ☒

The aggregate market value of the voting stock held by non-affiliates of the registrant was $107,183,061,000 based on the number of shares held by non-affiliates of the registrant as of May 31, 2011, and based on the closing sale price of common stock as reported by the NASDAQ Global Select Market on November 30, 2010, which is the last business day of the registrant's most recently completed second fiscal quarter. This calculation does not reflect a determination that persons are affiliates for any other purposes.

Number of shares of common stock outstanding as of June 20, 2011: 5,065,515,000.

Documents Incorporated by Reference:

Portions of the registrant's definitive proxy statement relating to its 2011 annual stockholders' meeting are incorporated by reference into Part III of this Annual Report on Form 10-K where indicated.

manufacturing, professional services, public sector, retail, travel, transportation and utilities. For example, we offer the banking and financial services sector a suite of applications addressing cash management, trade, treasury, payments, lending, private wealth management, asset management, compliance, enterprise risk and business analytics, among others. We offer the retail sector software solutions designed to provide unified and actionable data among store, merchandising and financial operations. Our applications for consumer goods manufacturers are designed to provide them with the ability to build their brand against retail private label programs by engaging directly with the consumer. Our ability to offer applications to address industry-specific complex processes provides us an opportunity to expand our customers' knowledge of our broader product offerings and address customer specific technology challenges.

Software License Updates and Product Support

We seek to protect and enhance our customers' current investments in Oracle software by offering proactive and personalized support
upgrades. Software li
maintenance releases
internet and telephon
internet access to tecl
contracts are general
customers purchase
licenses and renew th
updates and product
and 2009, respectivel

Hardware Systems

As a result of our a
hardware systems bus
support.

Hardware Systems P

Our customers dema
computational requir
space, and operationa
offerings, including
hardware-related soft
customer environmen
also engineered our
customers when our
Exadata and Oracle
our open, integrated
security, ease of man
12% and 6% of our to

Servers

We offer a wide r
differentiated by thei
(general purpose or sp
and lower total cost
having more computa
are designed for the
standing relationship
SPARC server compo

manufacturing, professional services, public sector, retail, travel, transportation and utilities. For example, we offer the banking and financial services sector a suite of applications addressing cash management, trade, treasury, payments, lending, private wealth management, asset management, compliance, enterprise risk and business analytics, among others. We offer the retail sector software solutions designed to provide unified and actionable data among store, merchandising and financial operations. Our applications for consumer goods manufacturers are designed to provide them with the ability to build their brand against retail private label programs by engaging directly with the consumer. Our ability to offer applications to address industry-specific complex processes provides us an opportunity to expand our customers' knowledge of our broader product offerings and address customer specific technology challenges.

Software License Updates and Product Support

We seek to protect and enhance our customers' current investments in Oracle software by offering proactive and personalized support services, including our Lifetime Support policy, and unspecified product enhancements and upgrades. Software license updates provide customers with rights to unspecified software product upgrades and maintenance releases and patches released during the term of the support period. Product support includes internet and telephone access to technical support personnel located in our global support centers, as well as internet access to technical content through "My Oracle Support." Software license updates and product support contracts are generally priced as a percentage of the net new software license fees. Substantially all of our customers purchase software license updates and product support contracts when they acquire new software licenses and renew their software license updates and product support contracts annually. Our software license updates and product support revenues represented 42%, 49% and 50% of our total revenues in fiscal 2011, 2010 and 2009, respectively.

Hardware Systems Business

As a result of our acquisition of Sun in January 2010, we entered into the hardware systems business. Our hardware systems business consists of two operating segments: hardware systems products and hardware systems support.

Hardware Systems Products

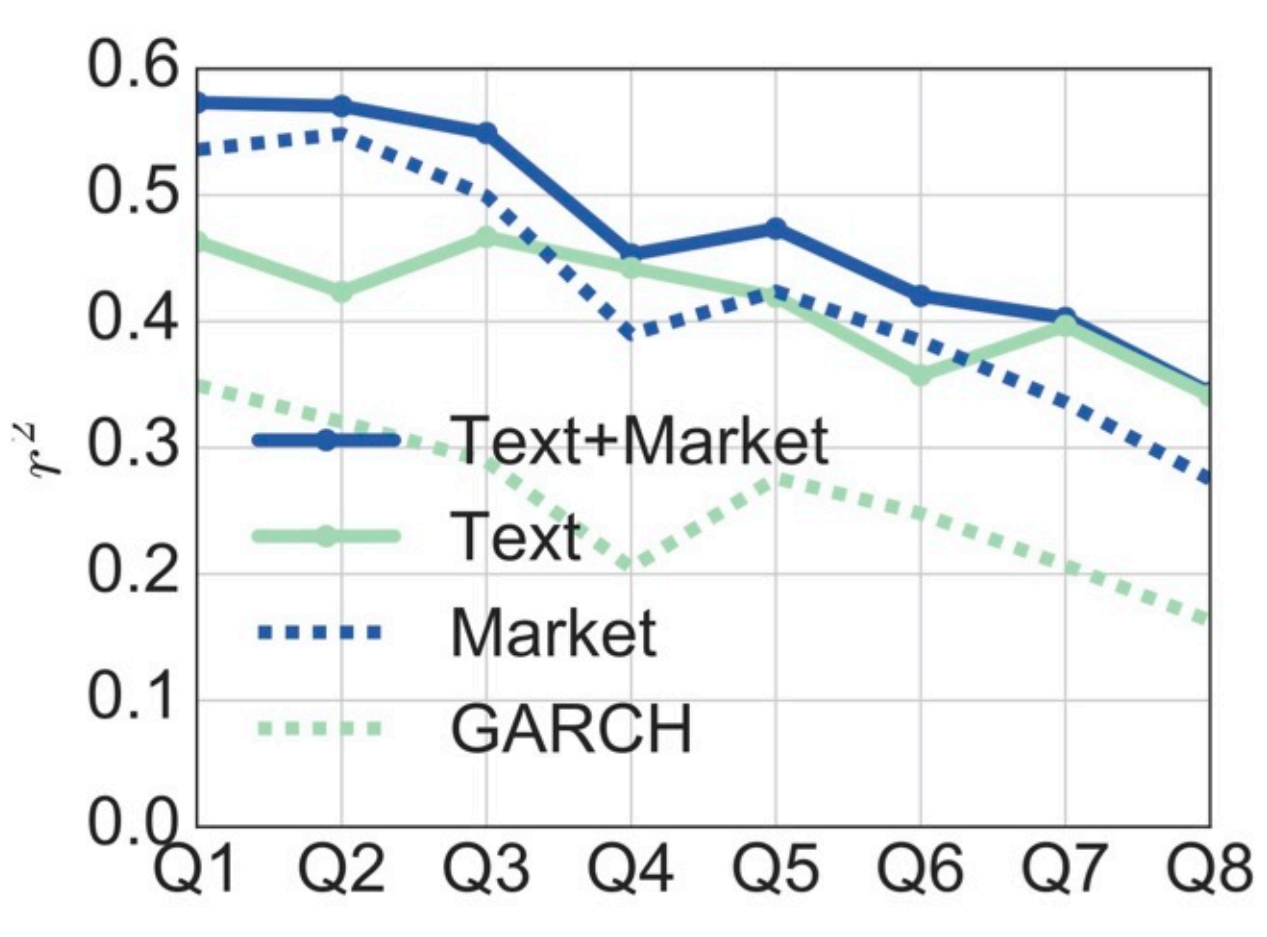
Our customers demand a broad set of hardware systems solutions to manage growing amounts of data and computational requirements, to meet increasing compliance and regulatory demands, and to reduce energy, space, and operational costs. To meet these demands, we have a wide variety of innovative hardware systems offerings, including servers and storage products, networking components, operating systems and other hardware-related software. Our hardware systems component products are designed to be "open," or to work in customer environments that may include other Oracle or non-Oracle hardware or software components. We have also engineered our hardware systems products to create performance and operational cost advantages for customers when our hardware and software products are combined as engineered systems, as with Oracle Exadata and Oracle Exalogic Elastic Cloud. By combining our server and storage hardware with our software, our open, integrated products better address customer requirements for performance, scalability, reliability, security, ease of management, and lower total cost of ownership. Our hardware systems products represented 12% and 6% of our total revenues in fiscal 2011 and 2010, respectively.

Servers

We offer a wide range of server systems using our SPARC microprocessor. Our SPARC servers are differentiated by their reliability, security and scalability; and by the customer environments that they target (general purpose or specialized systems). Our midsize and large servers are designed to offer greater performance and lower total cost of ownership than mainframe systems for business critical applications and for customers having more computationally intensive needs. Our SPARC servers run the Oracle Solaris operating system and are designed for the most demanding mission critical enterprise environments at any scale. We have a long-standing relationship with Fujitsu Limited for the development, manufacturing and marketing of certain of our SPARC server components and products.

Volatility Prediction with Sentiment Analysis

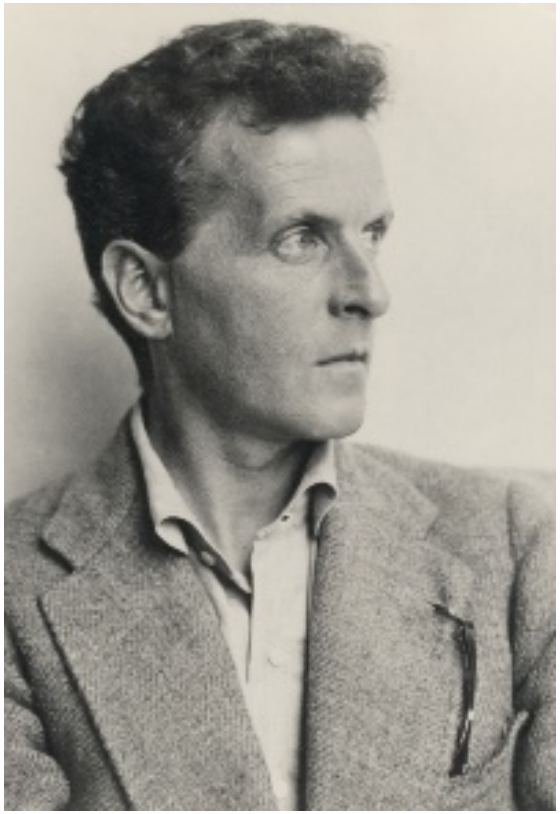
- Sentiment analysis with a word embedding-based method
- Text data significantly improves prediction
- Prediction performance in the upcoming Quartiles



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

J. R. Firth,

A synopsis of linguistic theory (1957)



"In most cases, the meaning of a word is its use."

Ludwig Wittgenstein,
Philosophical Investigations (1953)

drink

drunk

alcohol

on the table

Tesgüino

make

out of corn

fermented

bottle of

Mexico

Nida[1975]


Dutch
drunk
pale
Heineken
brew
red star
bar
drink
green bottle
alcohol

Tesgüino ←→ Heineken

Intuition for algorithms:

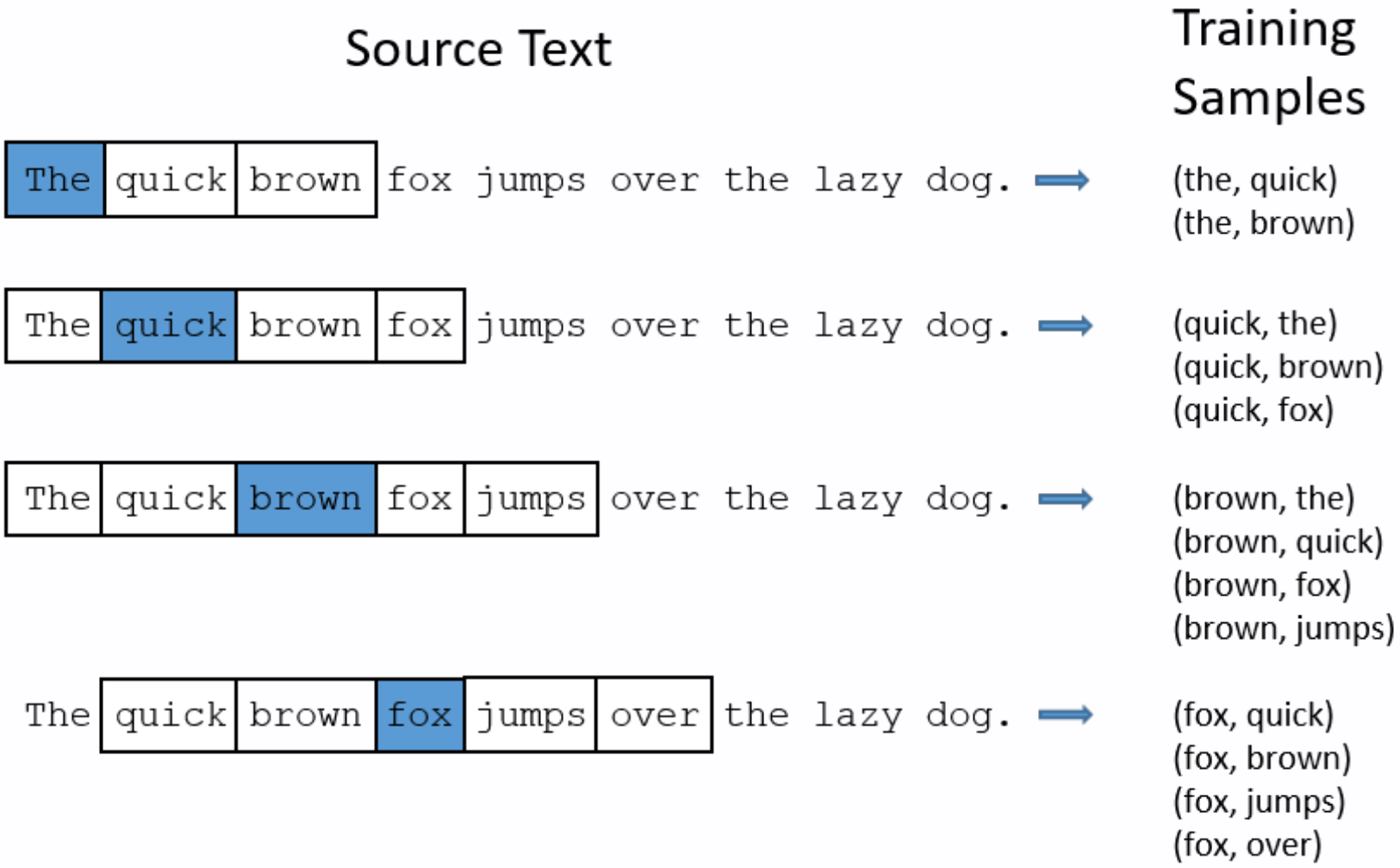
Two words are semantically related when they have similar context words

word2vec

A neural word embedding algorithm

Training Data

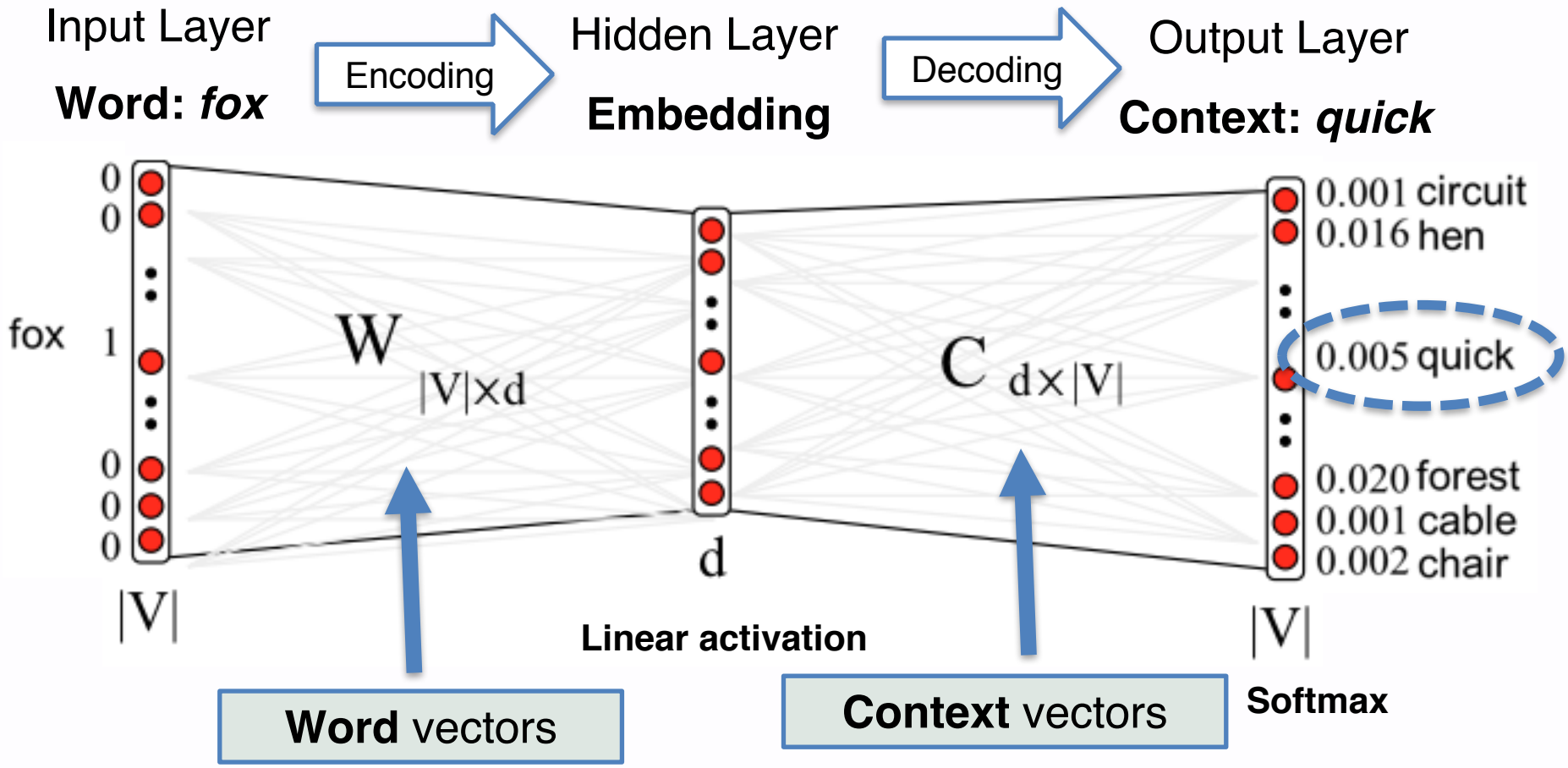
- Window size of 2



http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

Basic Neural Architecture

(*fox*, *quick*)



https://web.stanford.edu/~jurafsky/slp3/

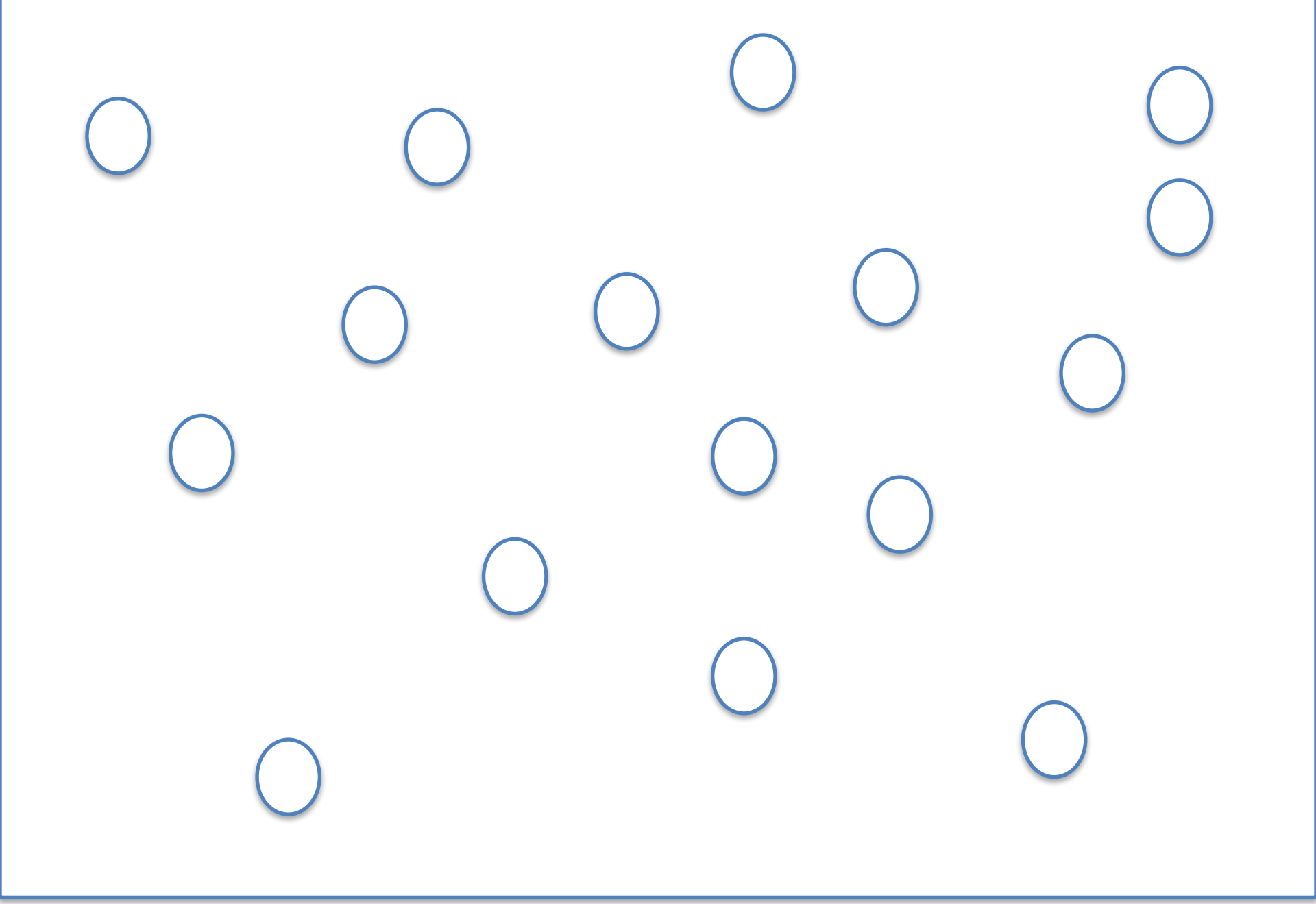
Basic Neural Architecture

- **Normalize** the output layer values over **all context vectors** by ***softmax***

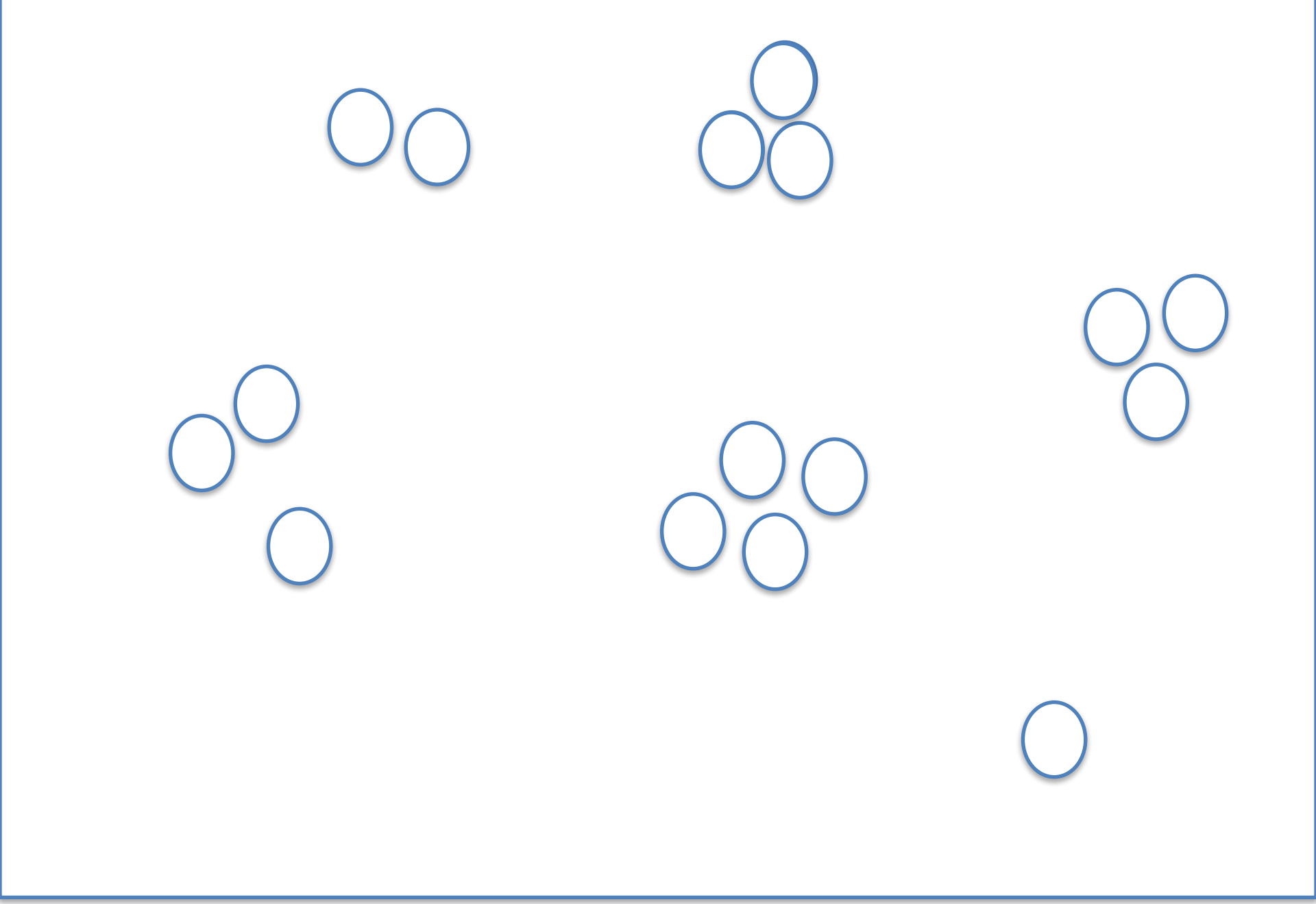
$$p(\text{quick}|\text{fox}) = \frac{e^{W_{\text{fox}} \cdot C_{\text{quick}}}}{\sum_{l \in V} e^{W_{\text{fox}} \cdot C_{\text{l}}}}$$

- Minimize the cost function on all the training

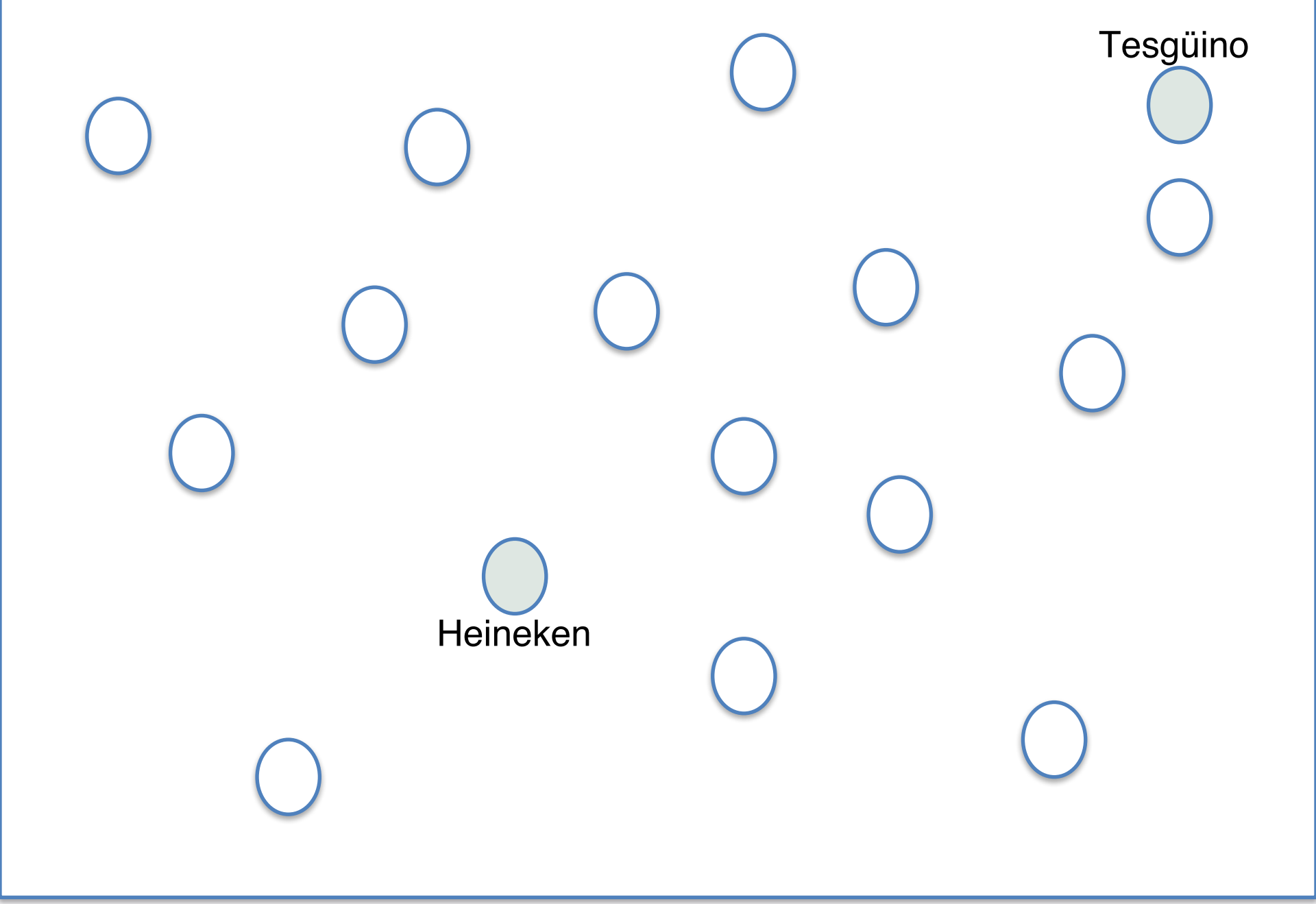
$$J = -\frac{1}{T} \sum_{1}^{T} \log p(c|w)$$



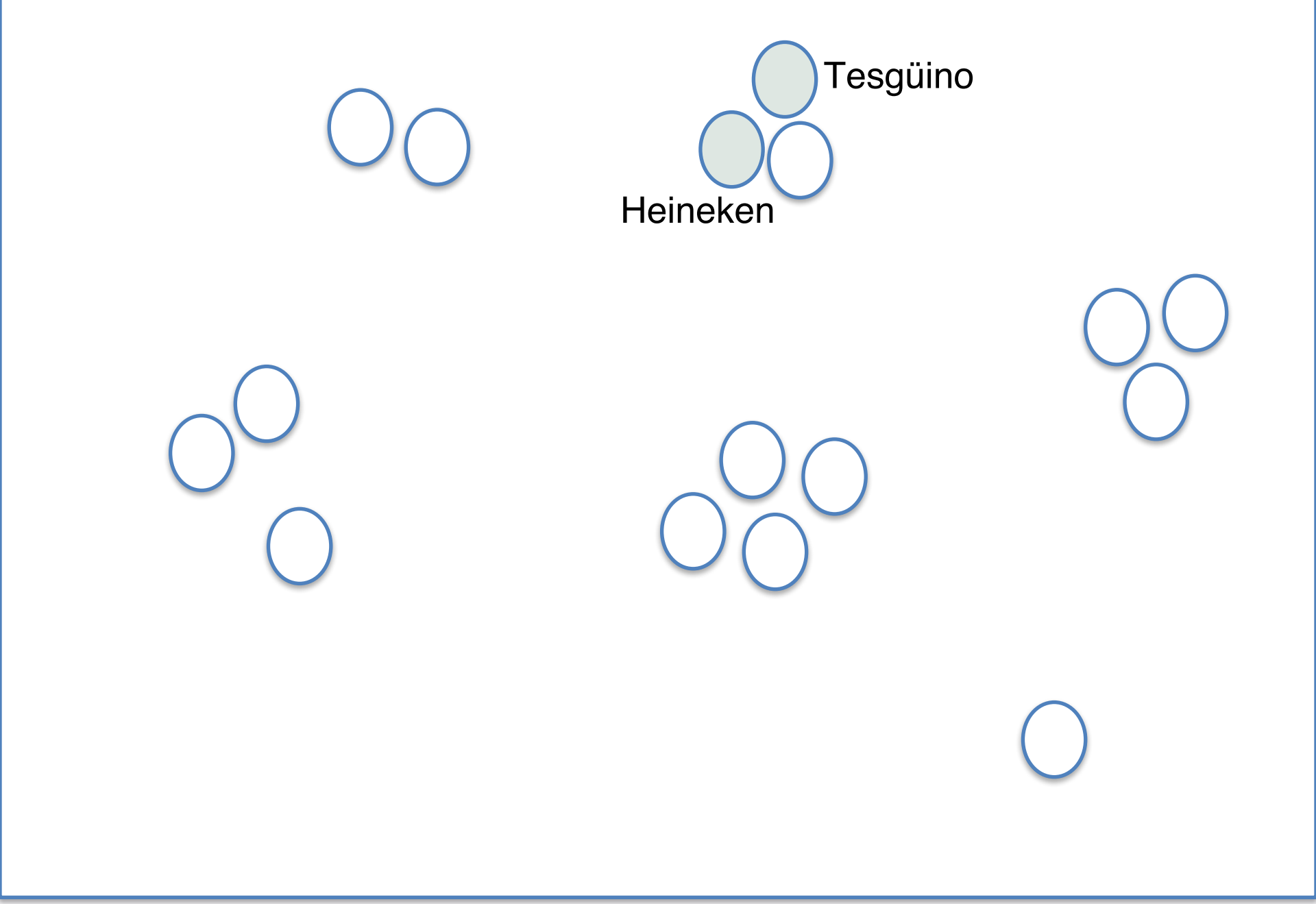
Word vector



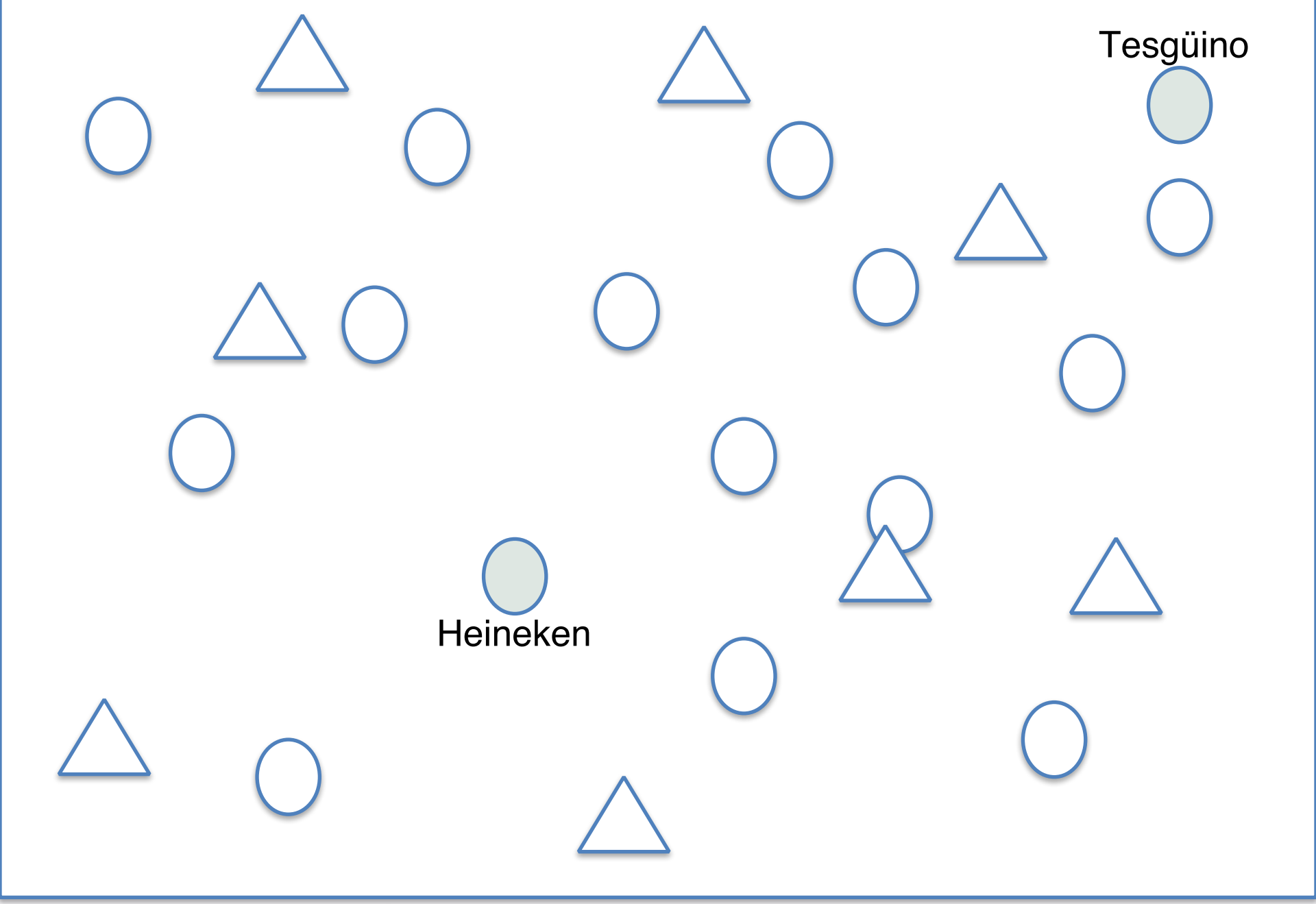
Word vector



Word vector

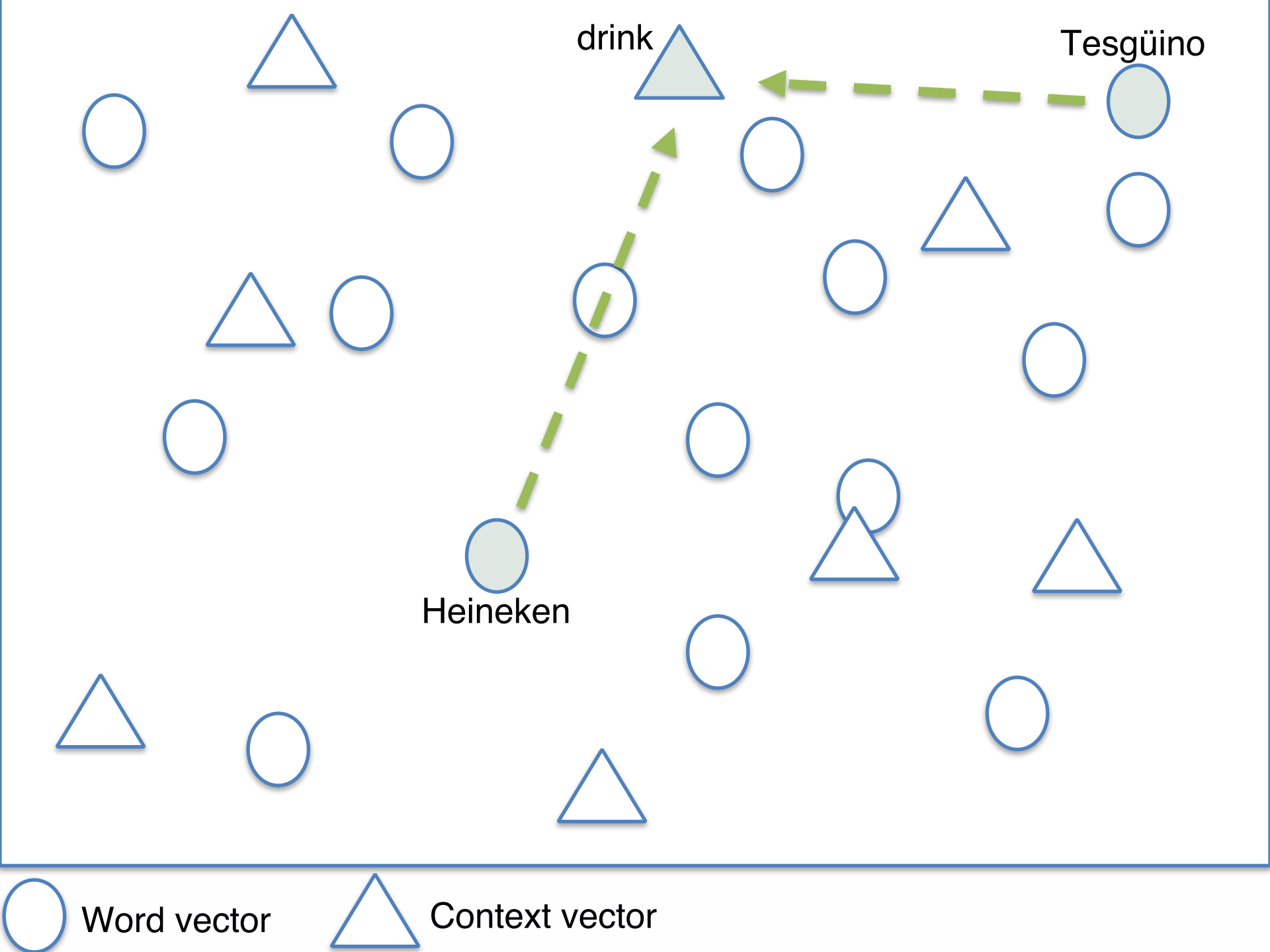


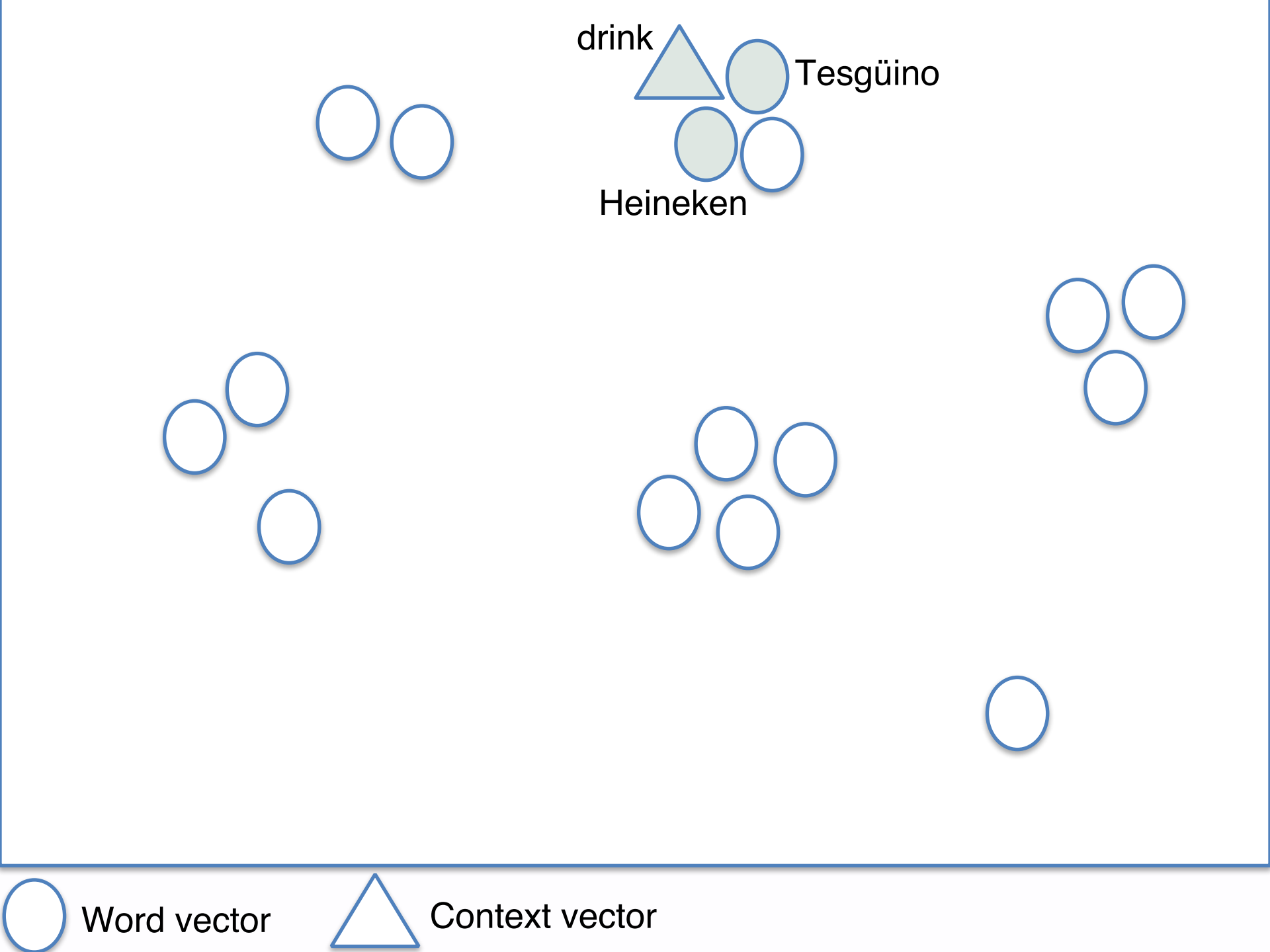
Word vector

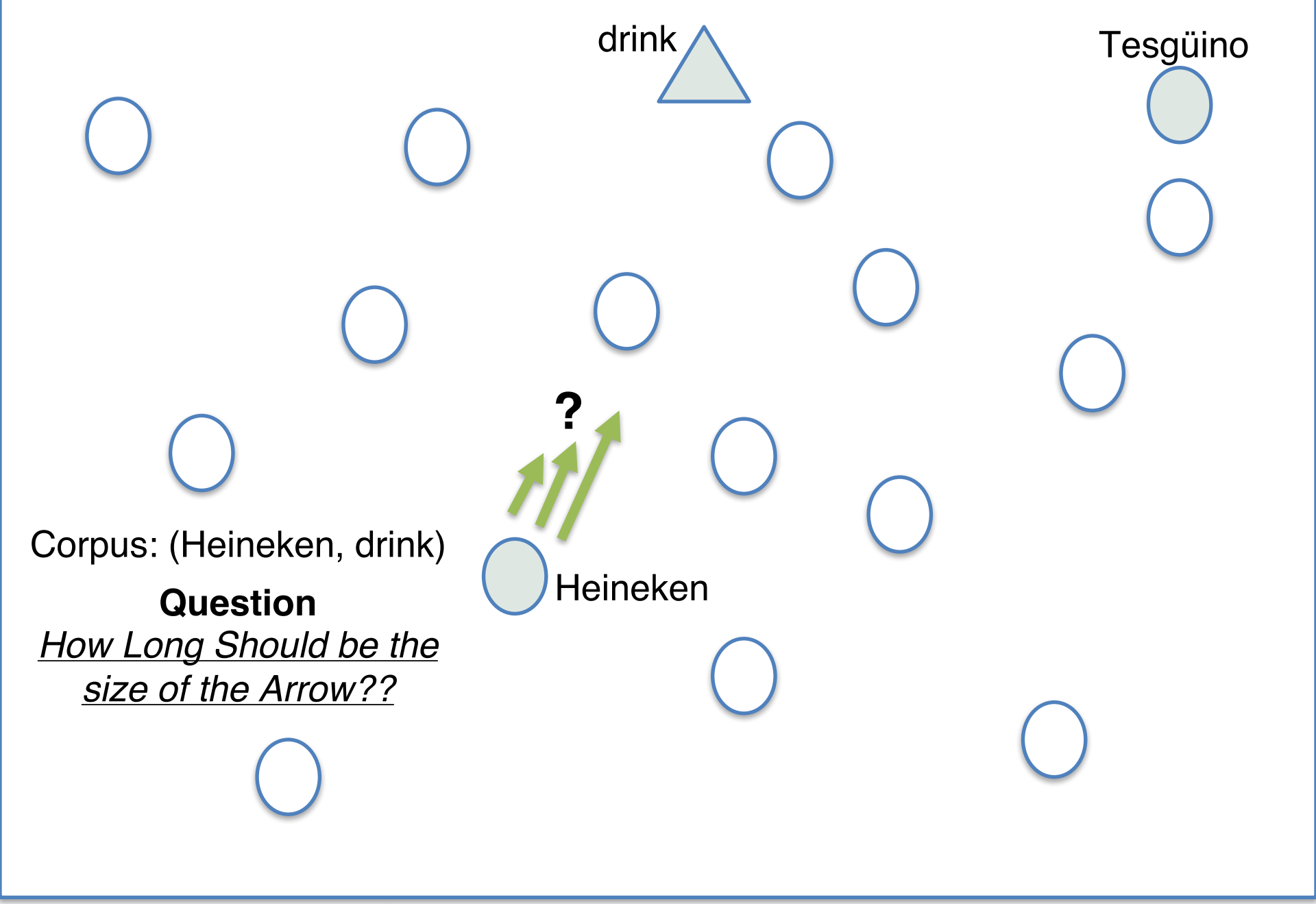


Word vector

Context vector


drink
Tesgüino
Heineken
Word vector
Context vector


drink
Tesgüino
Heineken
Word vector
Context vector



Corpus: (Heineken, drink)

Question

How Long Should be the size of the Arrow??

Word vector

Context vector

Answer 1 – Basic NN

- Normalize over **all context vectors** by *softmax*

$$p(\text{drink}|\text{Heineken}) = \frac{e^{W_{\text{Heineken}} \cdot C_{\text{drink}}}}{\sum_{l \subset V} e^{W_{\text{Heineken}} \cdot C_{\text{l}}}}$$

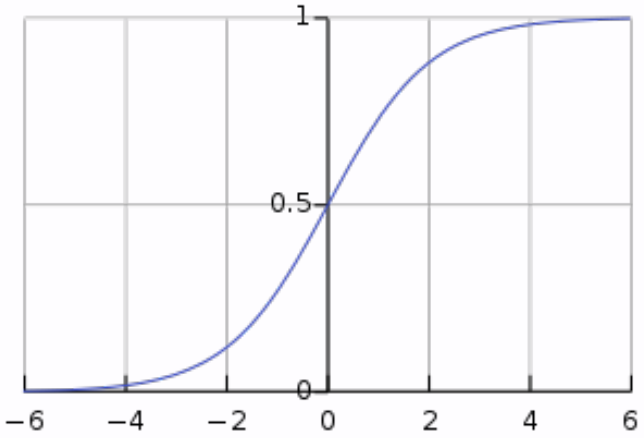
- Problem: denominator is **too expensive** to calculate

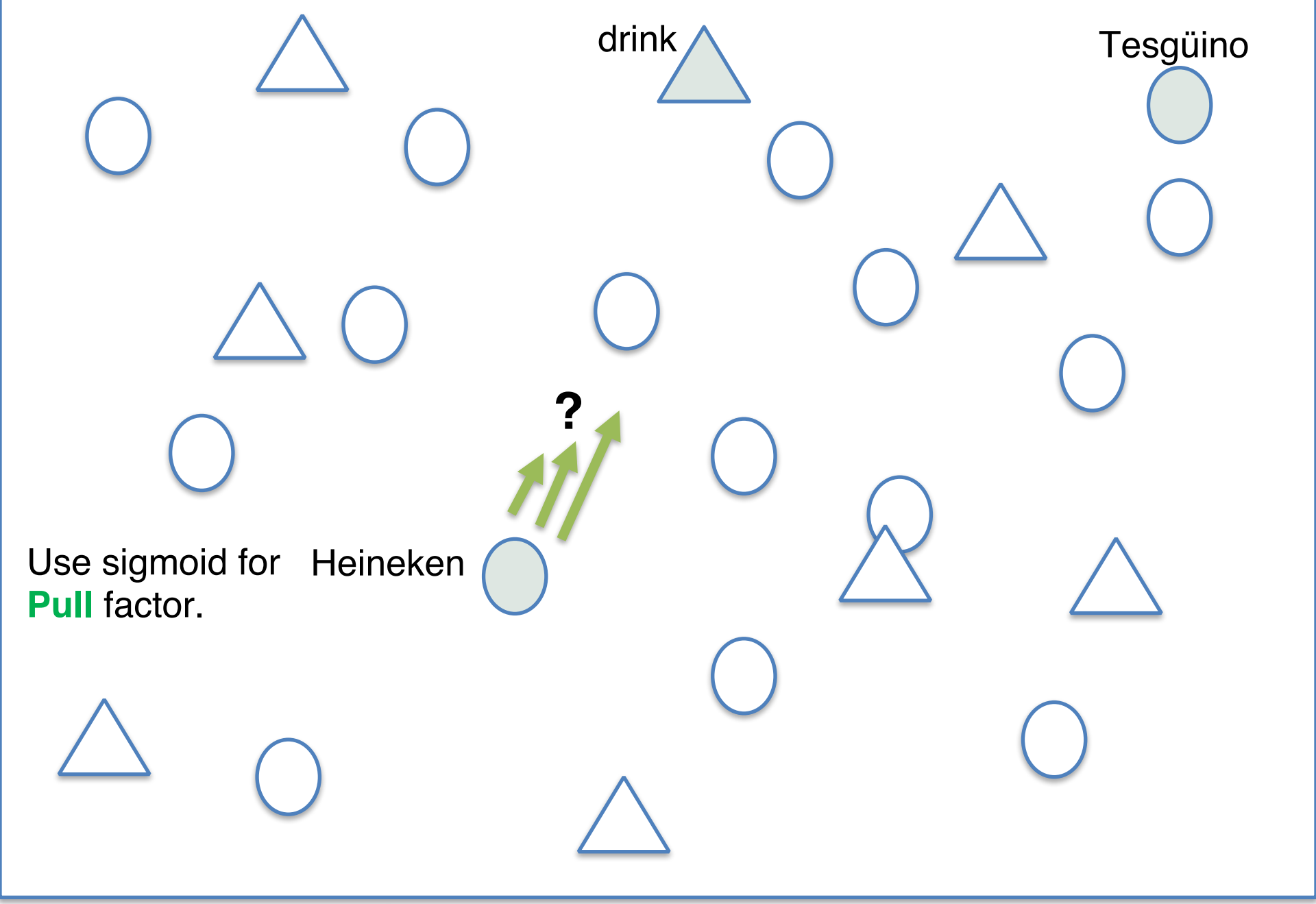
Answer 2 – word2vec Negative Sampling

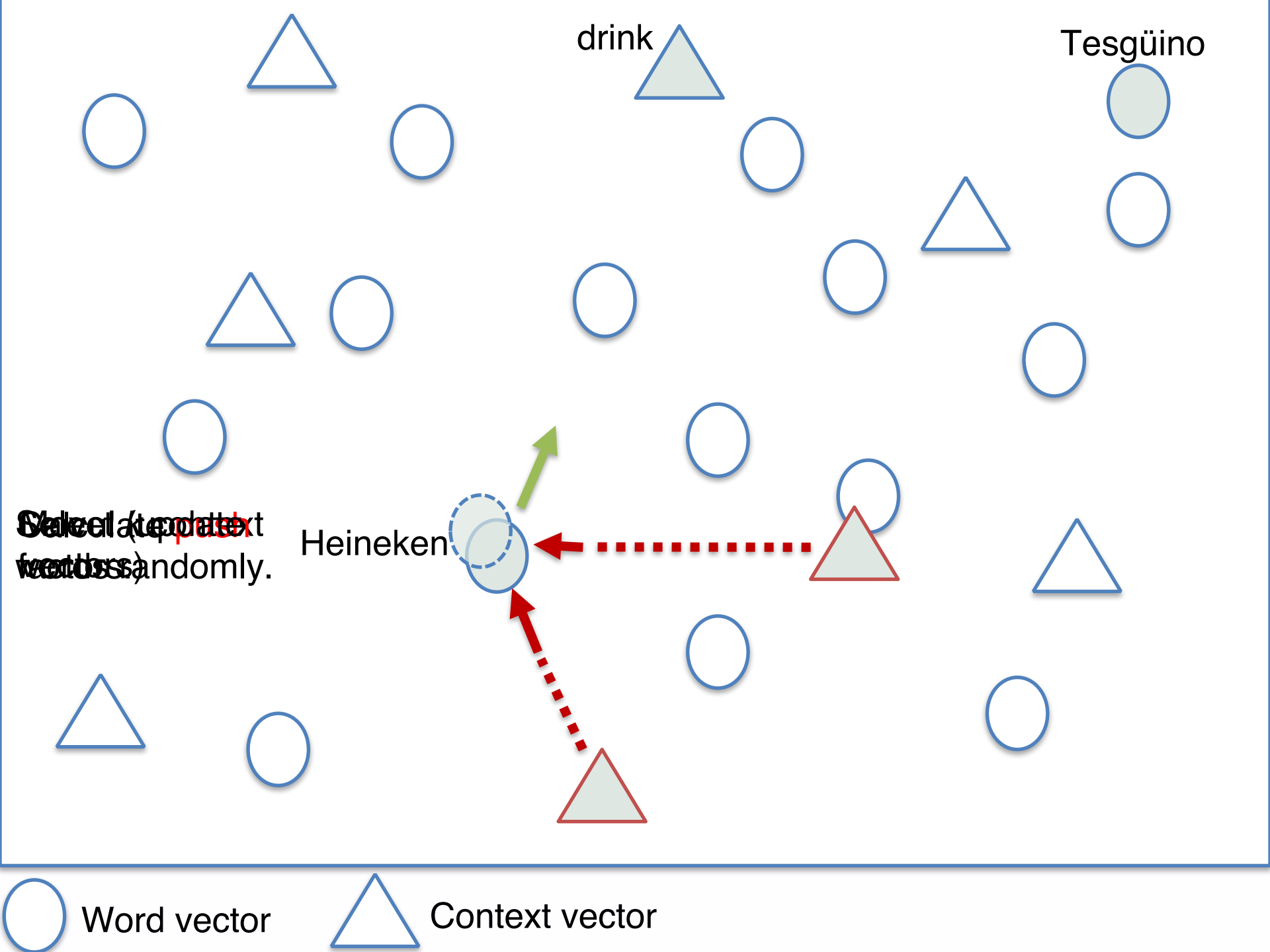
1. Don't normalize! Calculate the probability with *sigmoid*! (**Pull**)

$$p(y = 1|\text{Heineken}, \text{drink}) = \sigma(W_{\text{Heineken}} \cdot C_{\text{drink}})$$

2. **Push** against *k* randomly selected context vectors
   – A random word is an dissimilar word!



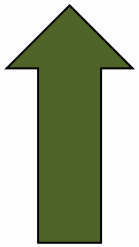

drink
Tesgüino
?
Use sigmoid for
Pull factor.
Heineken
Word vector
Context vector


drink
Tesgüino
Heineken
Word vector
Context vector

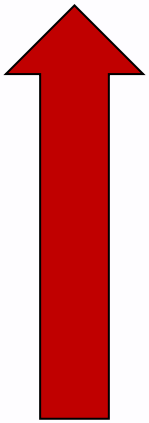
Cost Function of word2vec with Negative Sampling

- $k \sim$ 2-10

$$J = -\frac{1}{T}\sum_{1}^{T}\left[\log p(y=1|w,c) - \sum_{i=1}^{k}\log p(y=1|w,\check{c})\right]$$



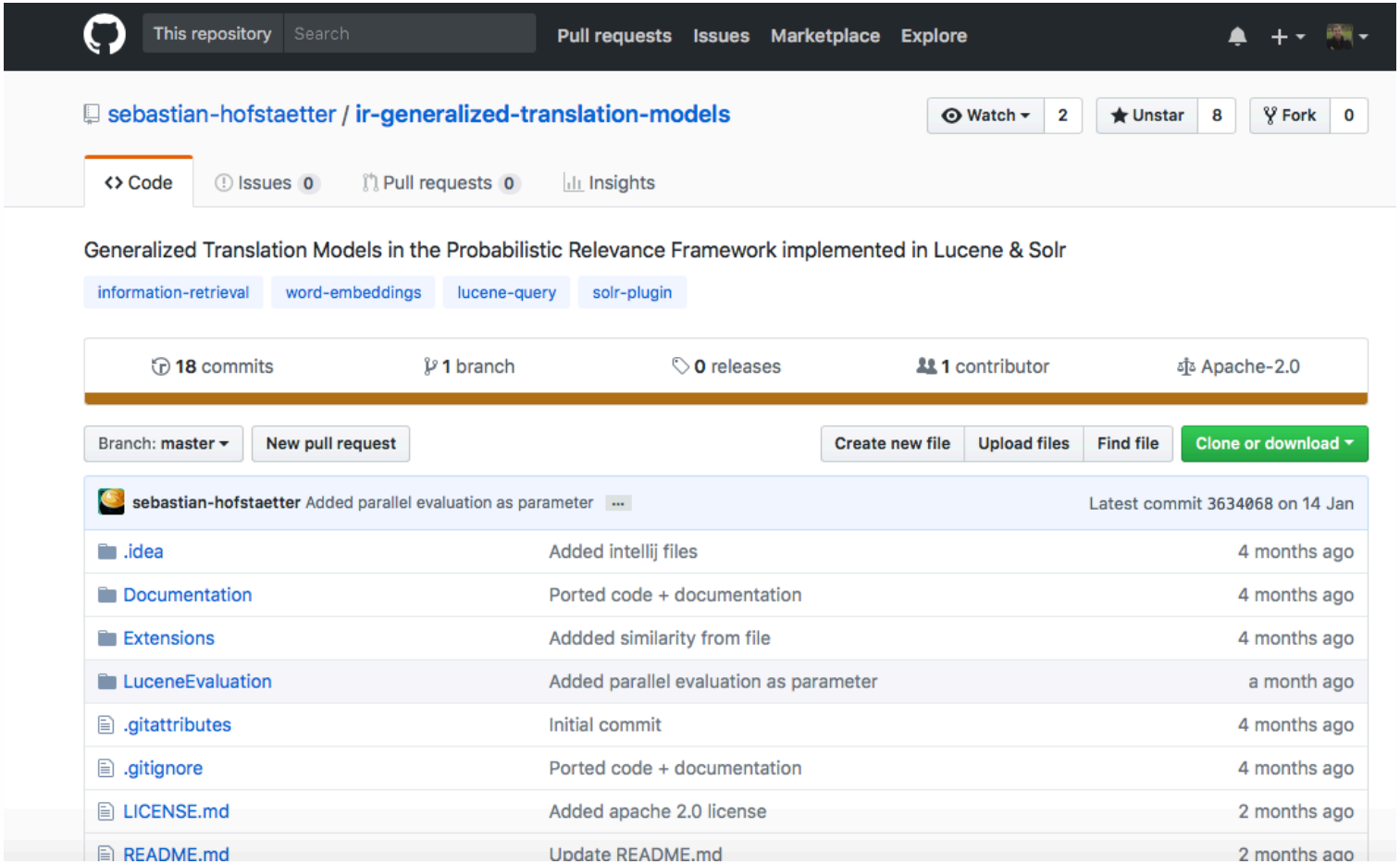
Genuine co-occurrence probability



Randomly sampled probability

Word Embedding for Search Engines

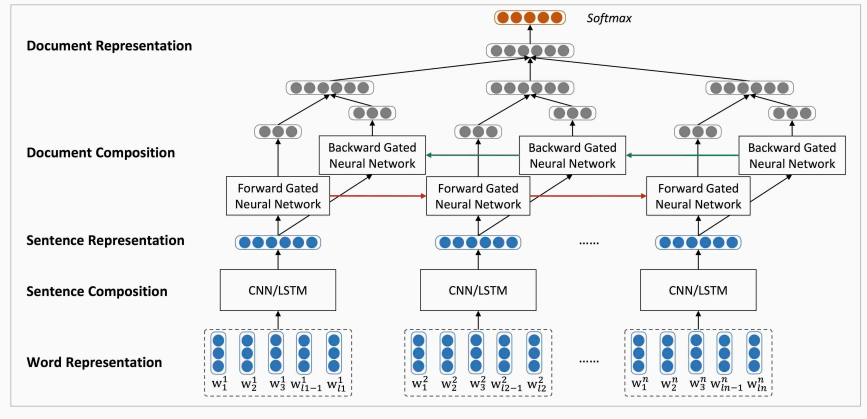
- Open source library for document retrieval
- Effectively combining word embedding with retrieval models
- Implemented for Solr and Lucene



Challenges and Perspective

- Task and domain specific representation learning

- Representation learning bigger language elements



- Interpretability, transparency, and fairness

Thanks!

@navidrekabsaz

rekabsaz@ifs.tuwien.ac.at