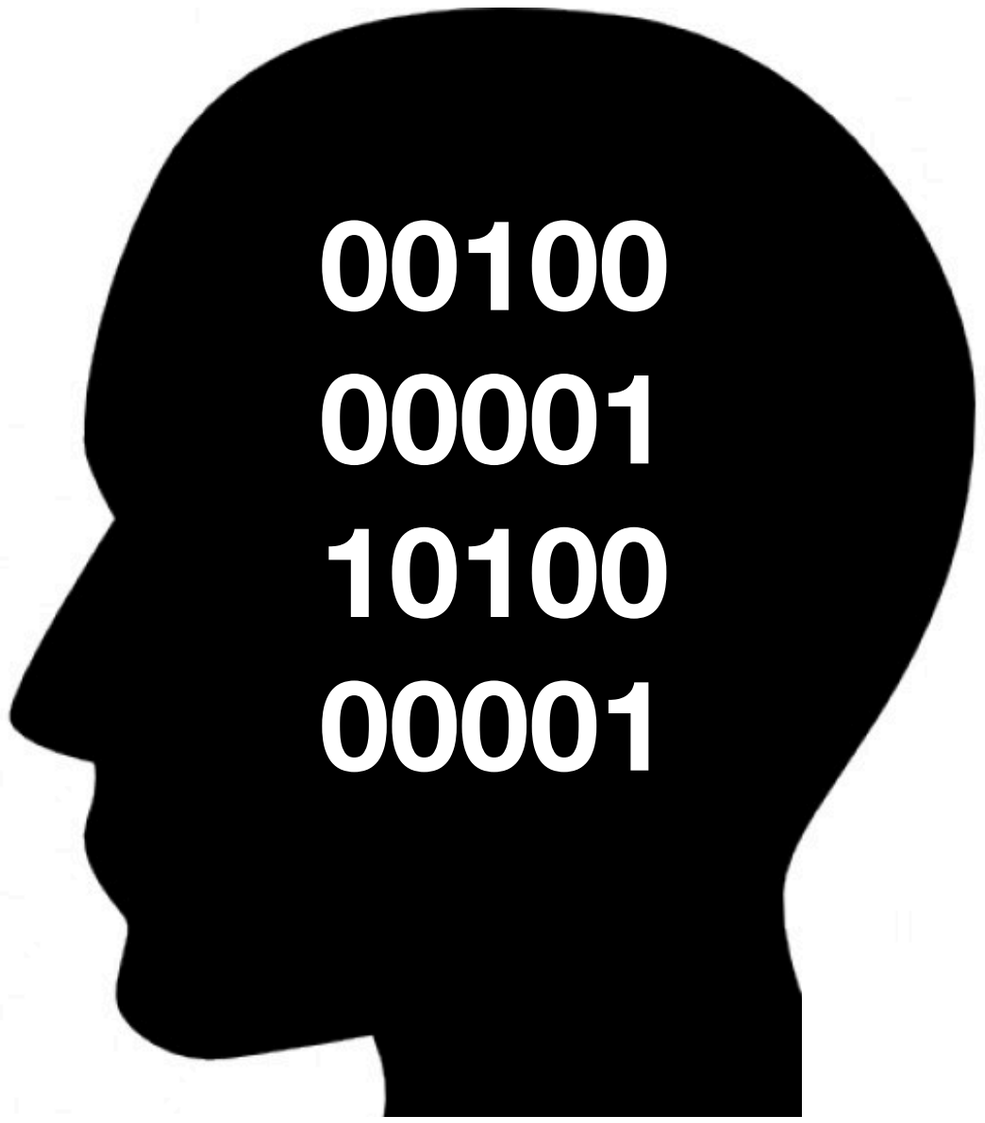


COG260: Data, Computation, and The Mind

Words

Lab 4: Prototypicality

Visualizing high-dimensional data

F1 F2 F3 F4 F5 F6 F7 F8 .....

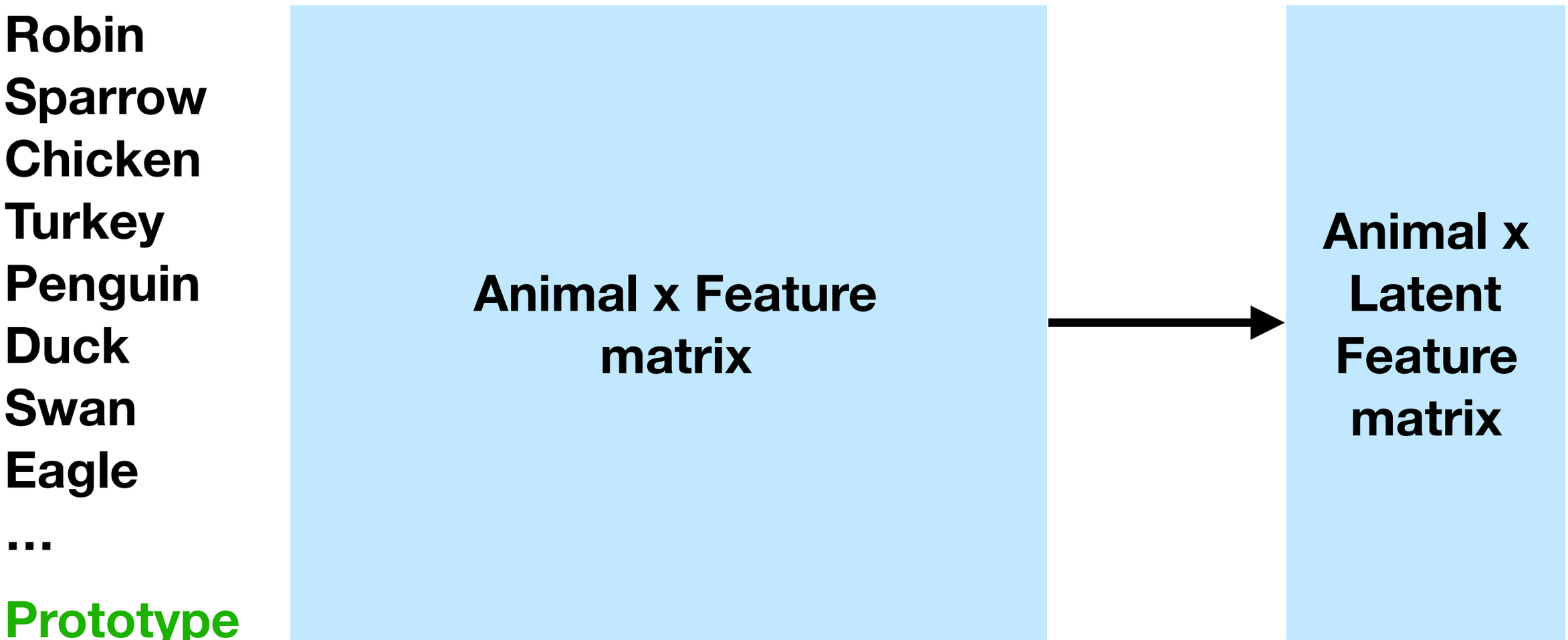
Robin
Sparrow
Chicken
Turkey
Penguin
Duck
Swan
Eagle
...

Prototype

Animal x Feature matrix

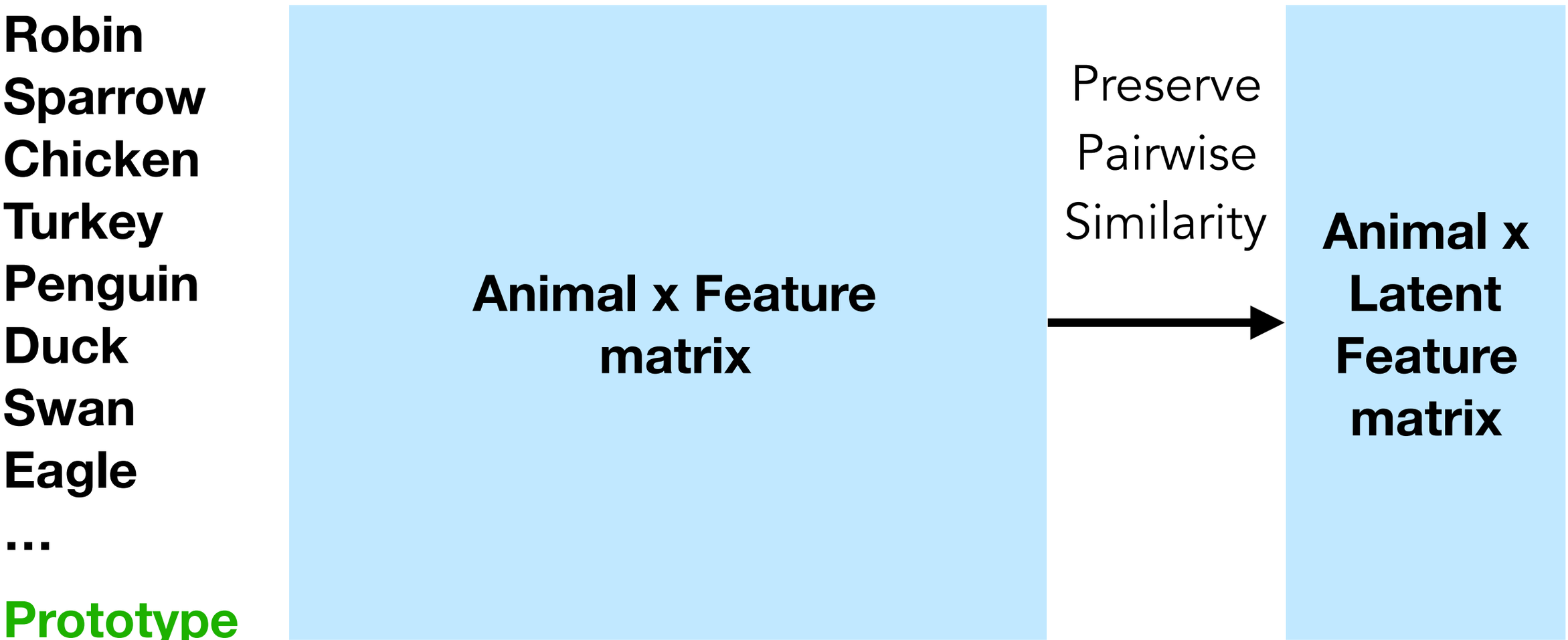
Lab 4: Prototypicality

Visualizing high-dimensional data



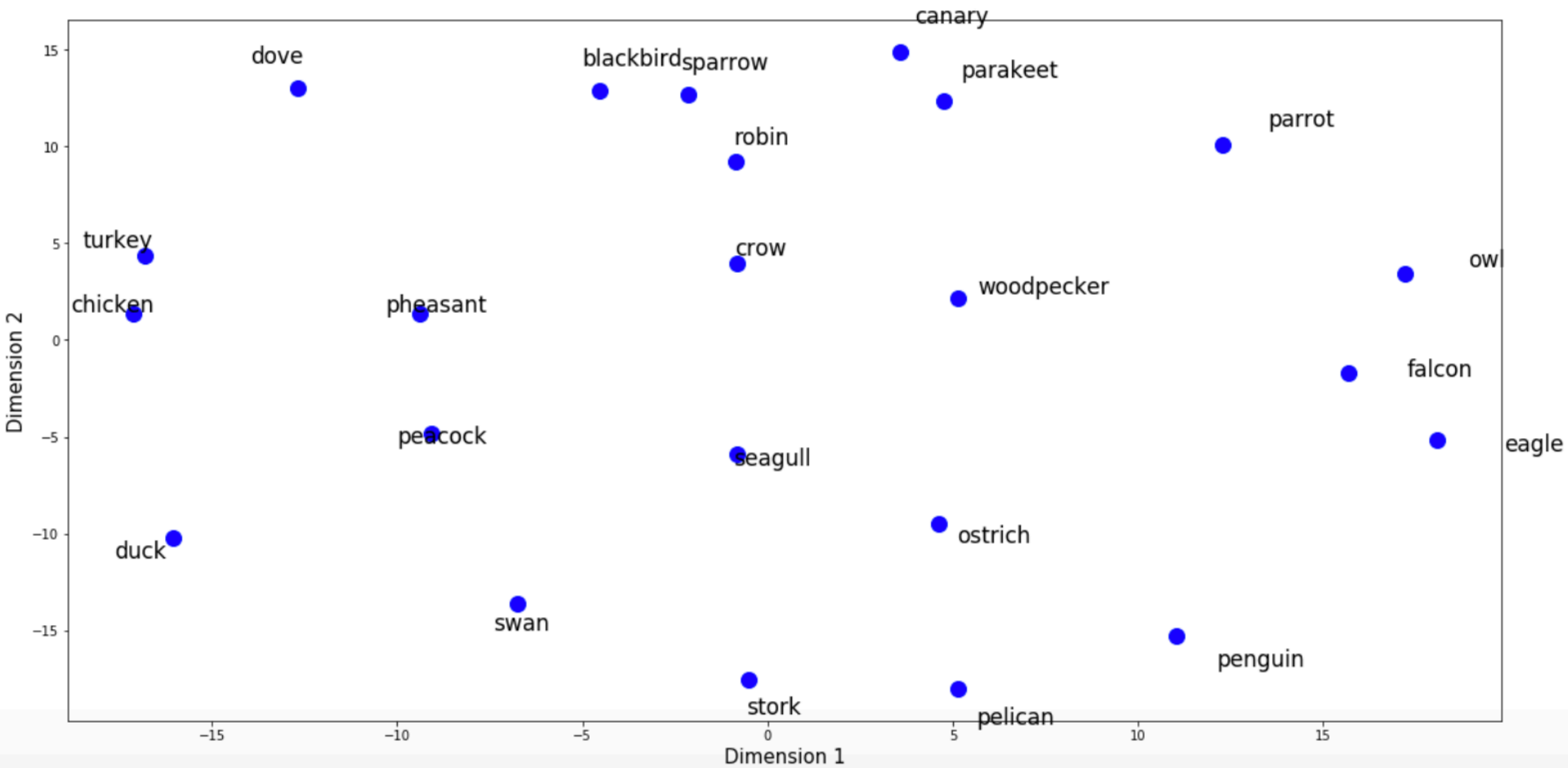
Lab 4: Prototypicality

Visualizing high-dimensional data



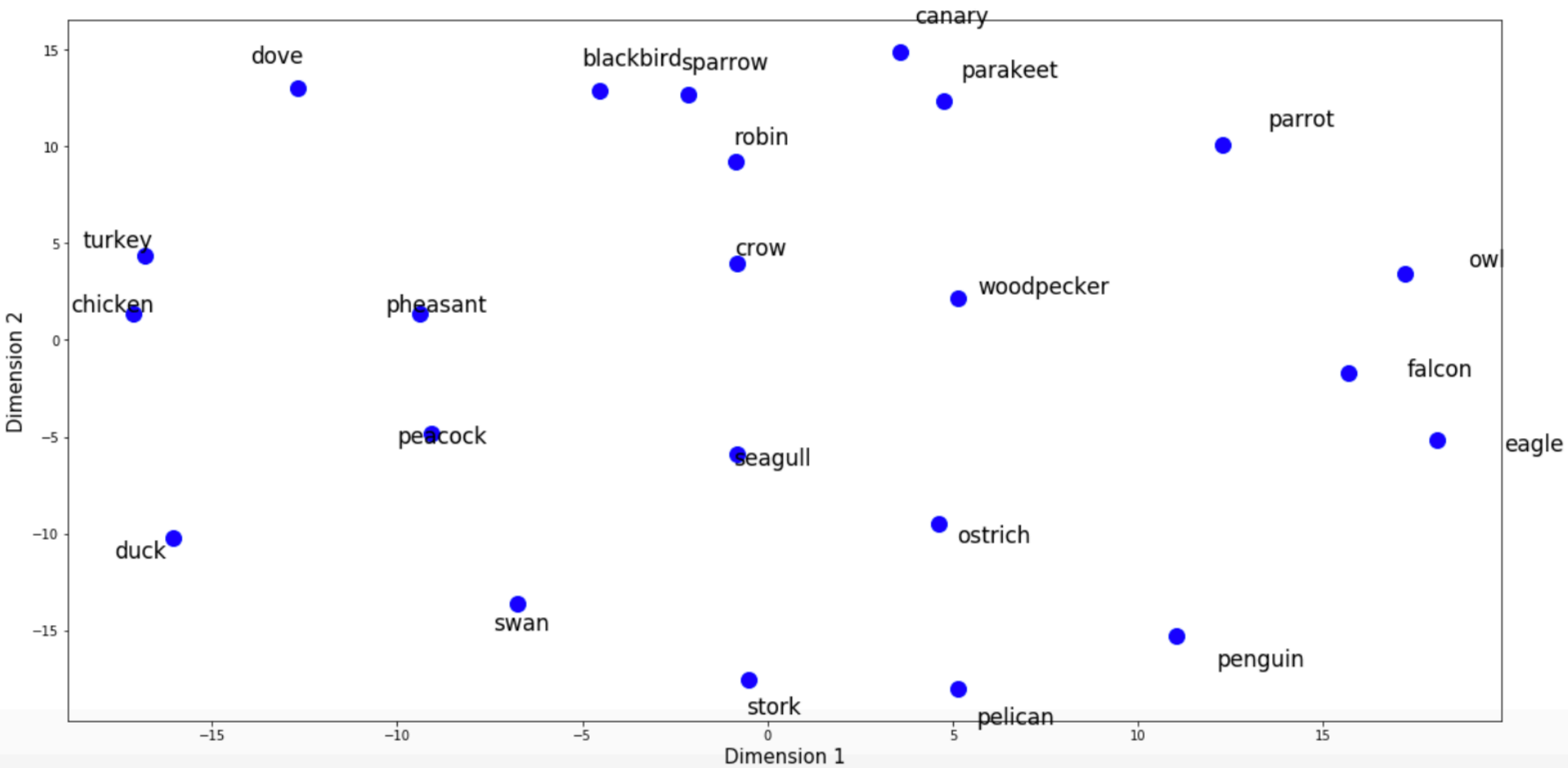
Lab 4: Prototypicality

2D visualization of birds via multidimensional scaling



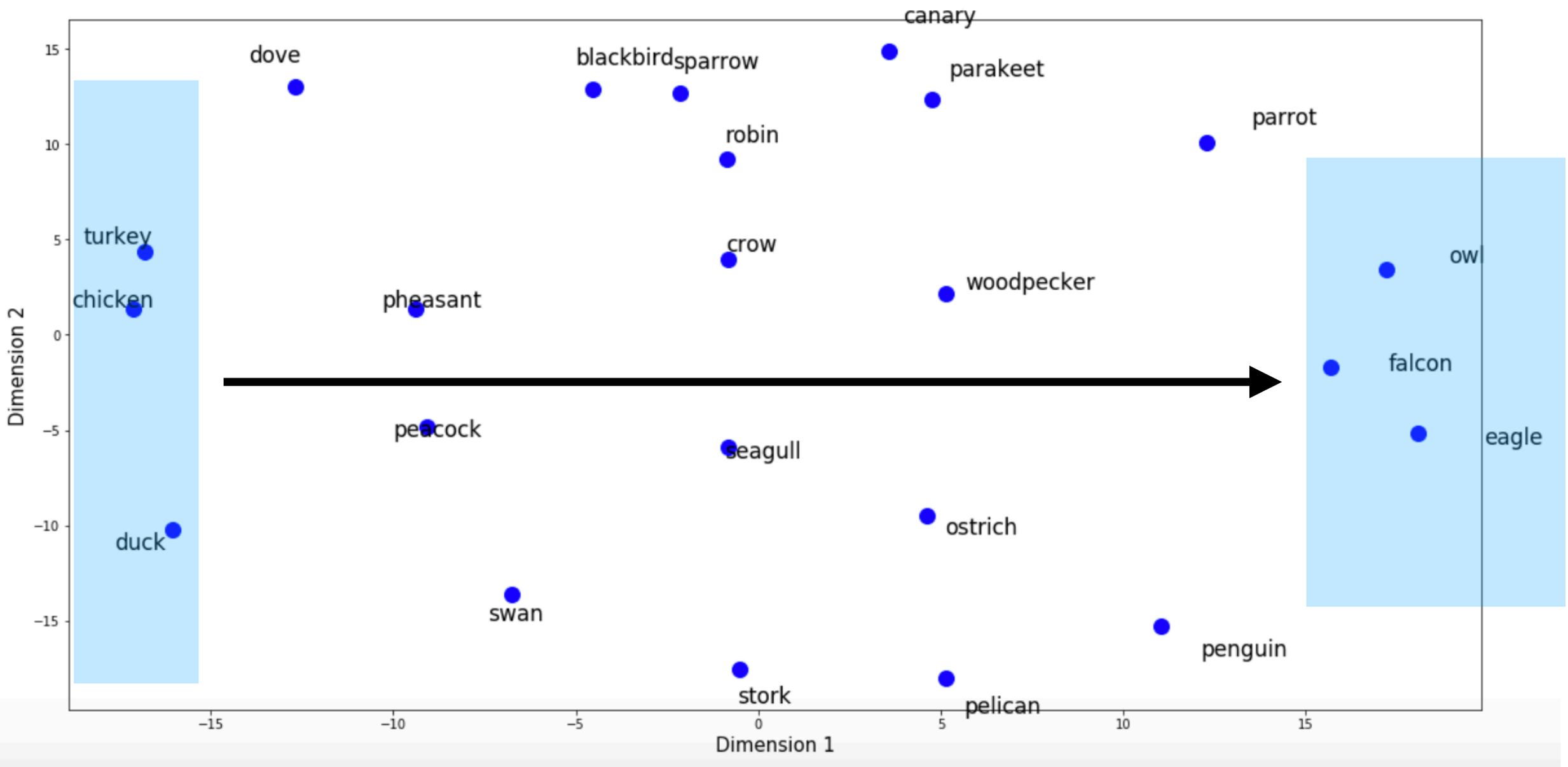
Lab 4: Prototypicality

Discuss: Can you make sense of this 2D space?



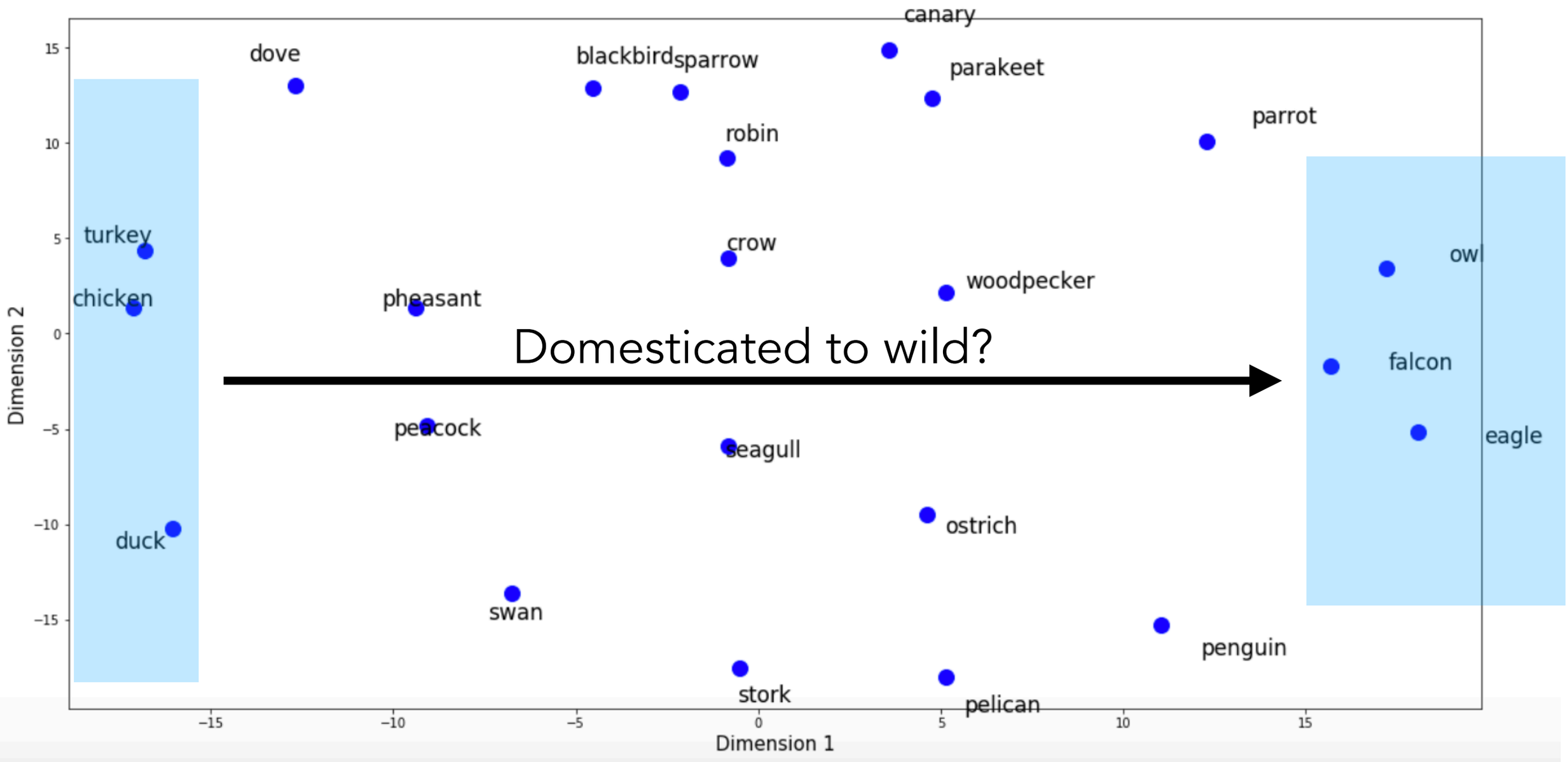
Lab 4: Prototypicality

2D visualization of birds via multidimensional scaling



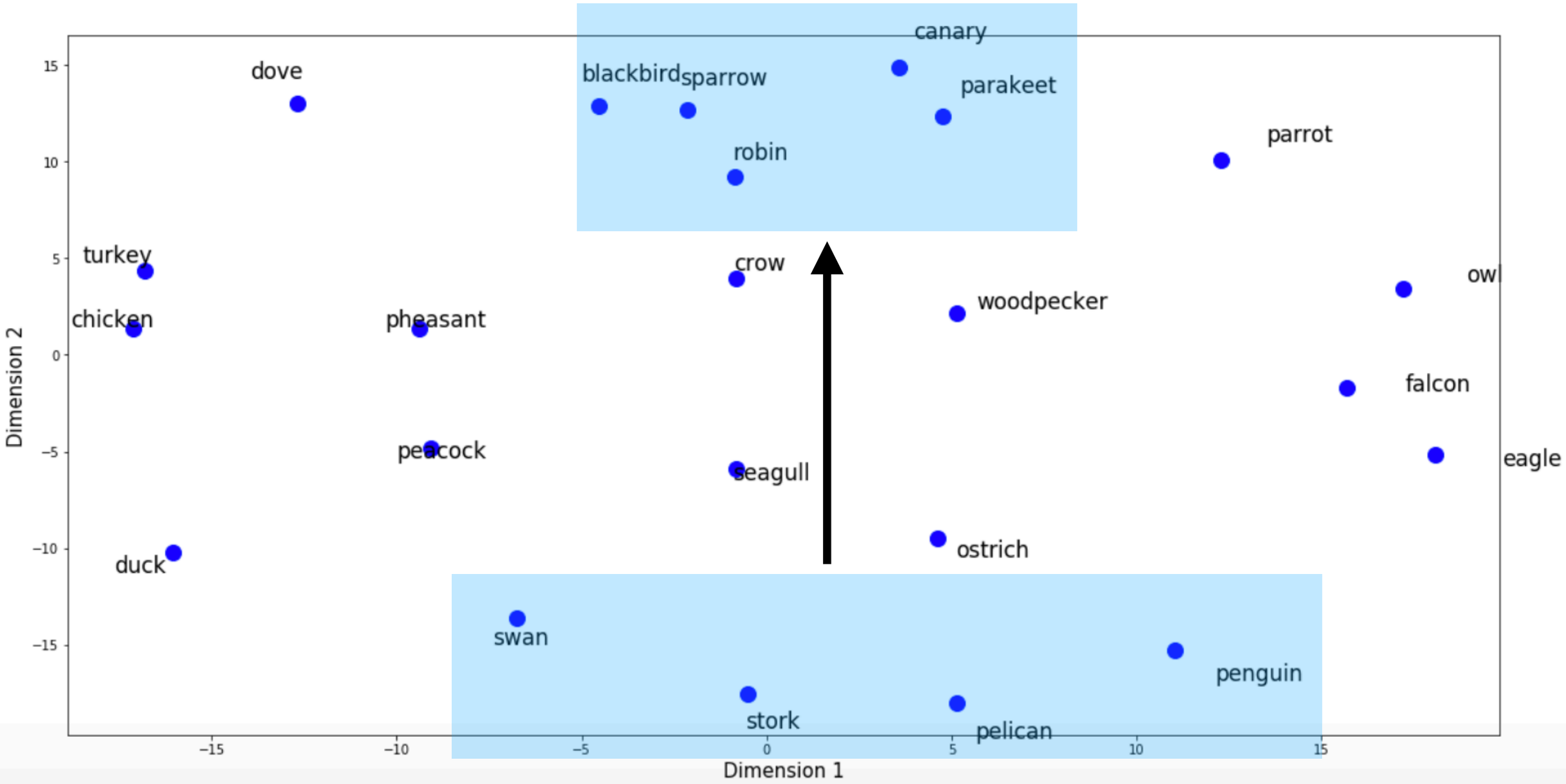
Lab 4: Prototypicality

2D visualization of birds via multidimensional scaling



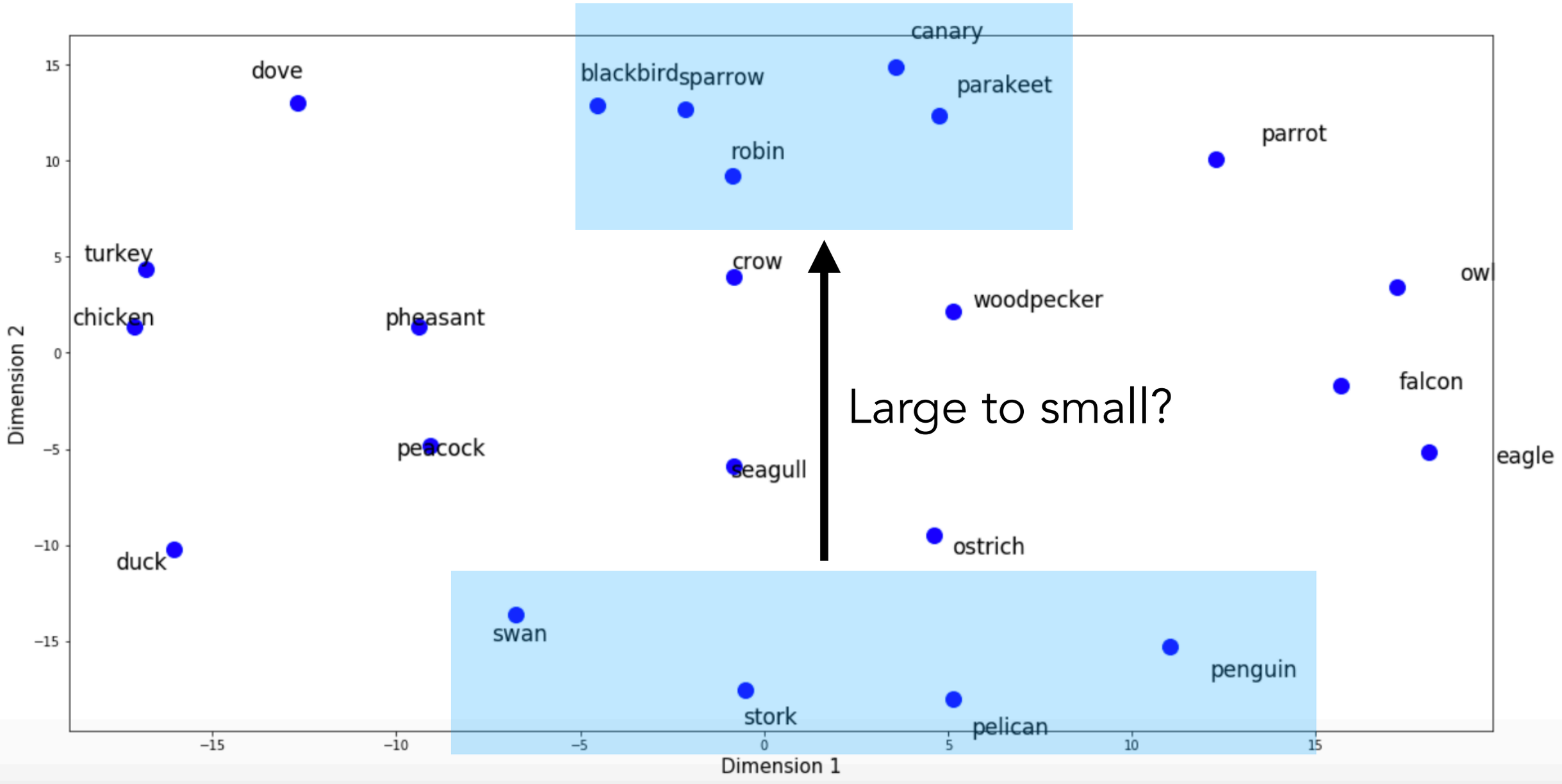
Lab 4: Prototypicality

2D visualization of birds via multidimensional scaling



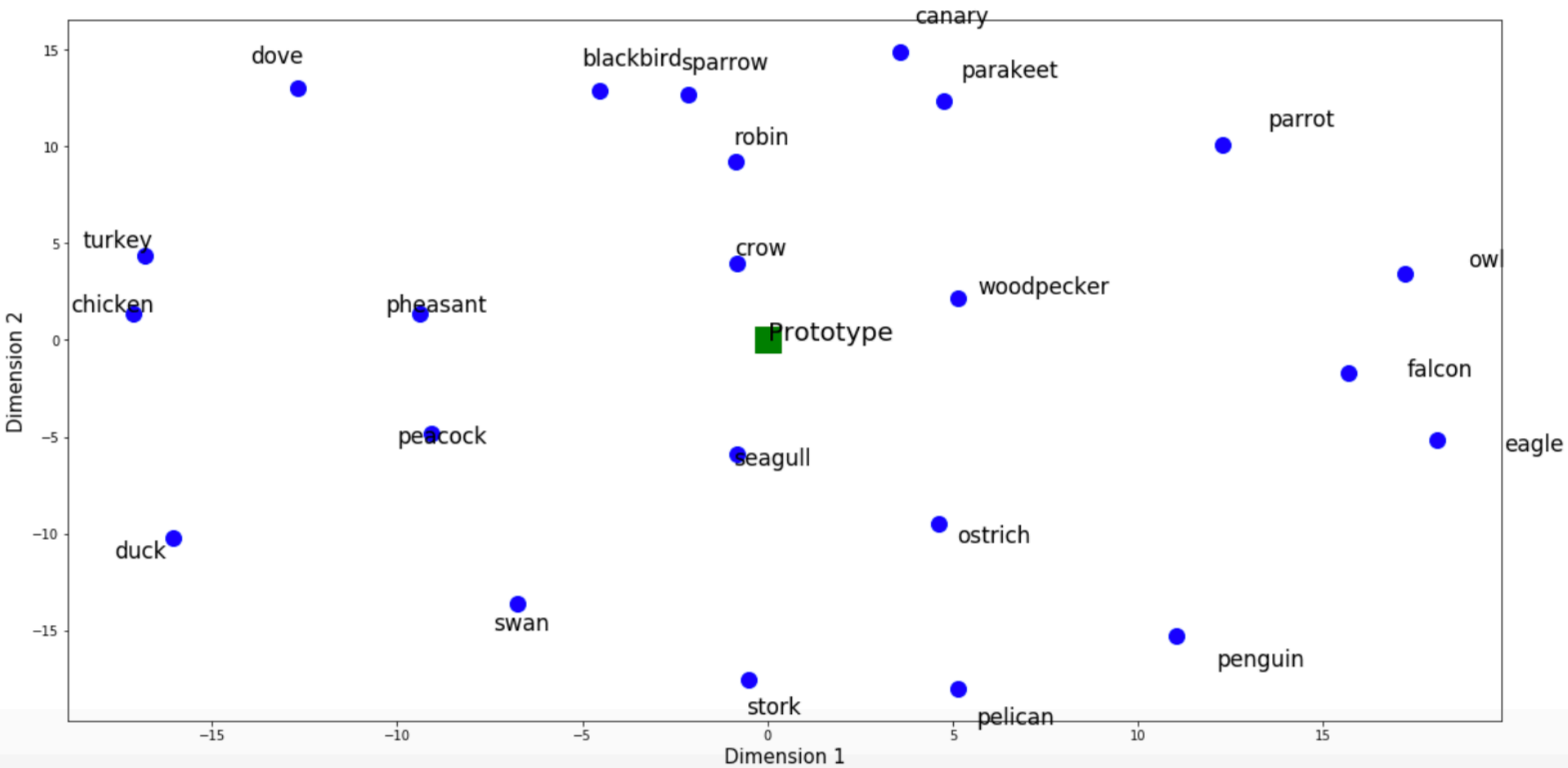
Lab 4: Prototypicality

2D visualization of birds via multidimensional scaling



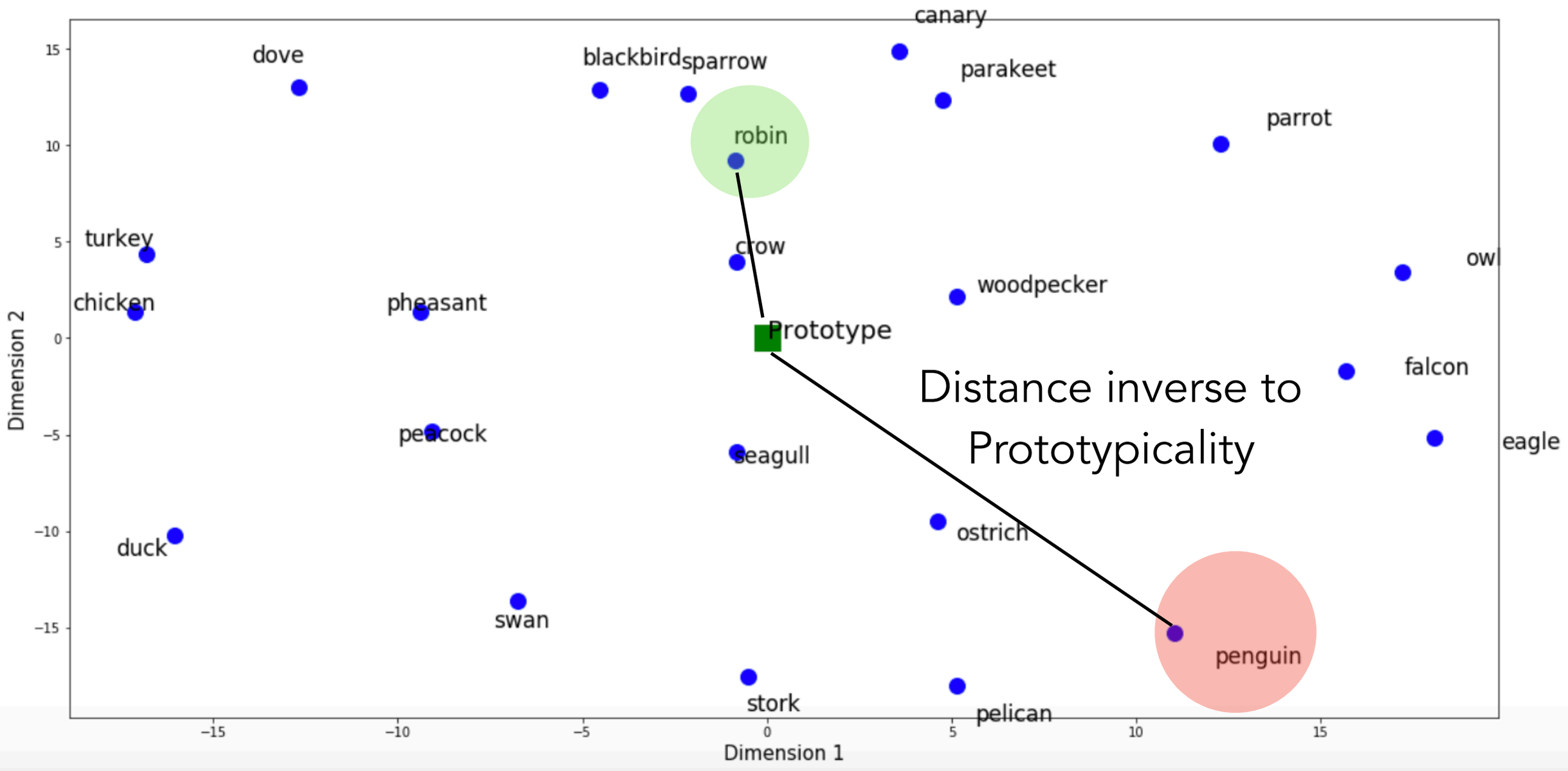
Lab 4: Prototypicality

2D visualization of birds via multidimensional scaling



Lab 4: Prototypicality

2D visualization of birds via multidimensional scaling



Lab 4: Prototypicality

- Discuss:
    - How would you improve the prototype model that you have implemented?

Lab 4: Prototypicality

Some reasonable solutions:

> The prototype model can be improved by measuring prototype as the median of the birds (as opposed to the mean) which is more robust to outliers.

> Adjust the prototype vector to be closer to birds with lower ratings.

Some less ideal solutions:

> Apply a logarithmic transformation on the data.

> Use Spearman correlation as opposed to Pearson correlation.

Lab 4: Prototypicality

- Discuss:
  - How would you improve the prototype model that you have implemented?
    - Feature weighting
    - Frequency or familiarity (of the birds)

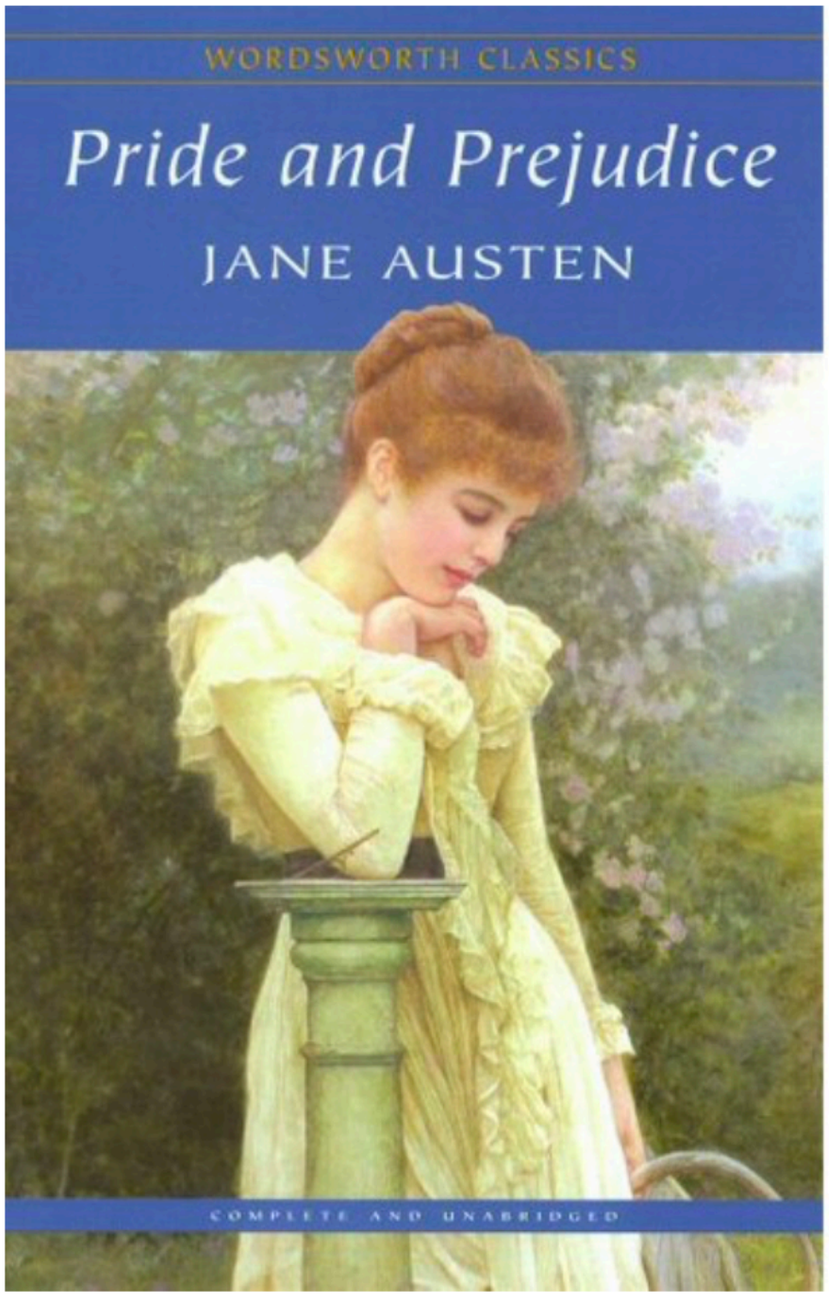
Lab 4: Prototypicality

- Discuss:
  - How would you improve the prototype model that you have implemented?
    - Feature weighting
    - Frequency or familiarity (of the birds)
    - ……
    - Q: How would you know when to stop improving?

Outline

- Word frequency and Zipf's law
- Word sequence and n-gram model
- Lab 6 (final lab!)

Word frequencies vary



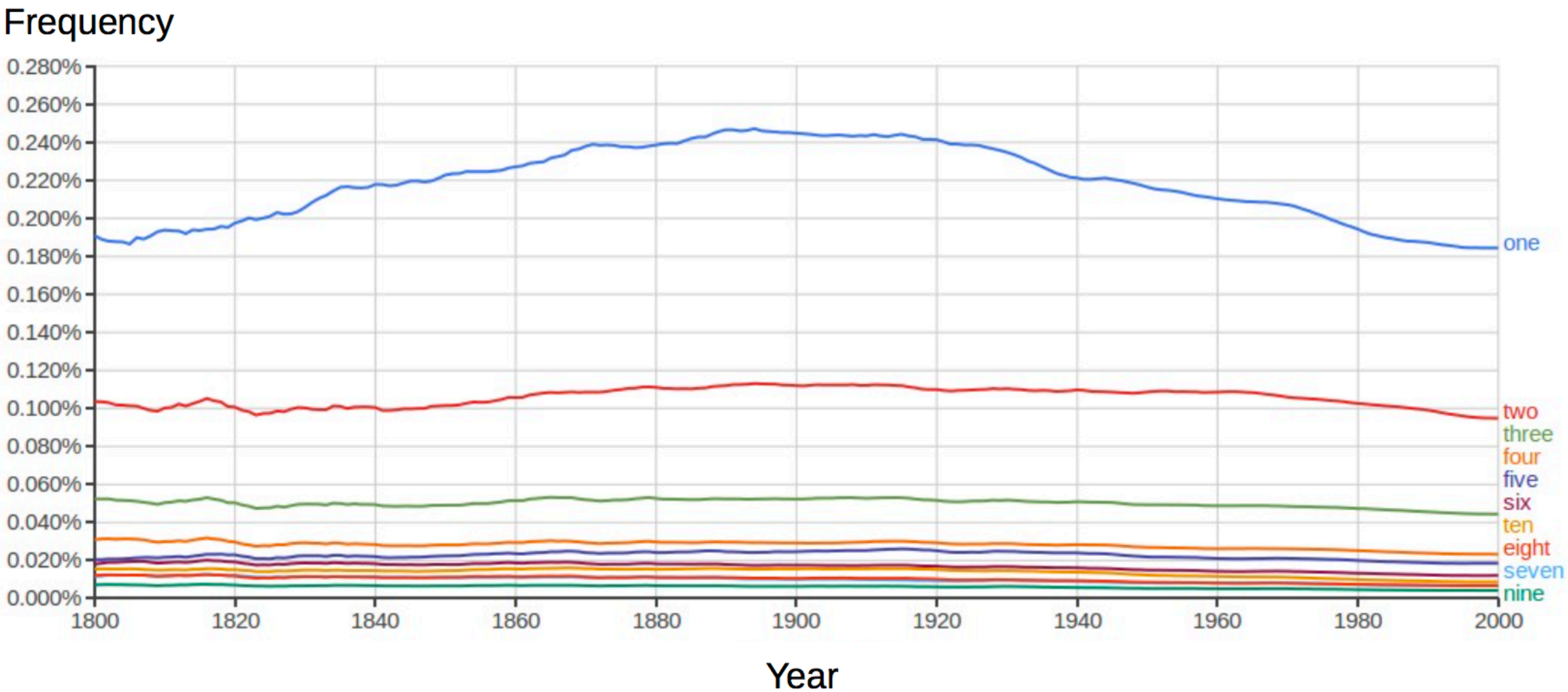
them all, and they were at last obliged to accept the second-hand intelligence of their neighbor, Lady Lucas. Her report was highly favorable. Sir William had been delighted with him. He was quite young, wonderfully handsome, extremely agreeable, and, to crown the whole, he meant to be at the next assembly with a large party. Nothing could be more delightful! To be fond of dancing was a certain step towards falling in love; and very lively hopes of Mr. Bingley's heart were entertained.

"If I can but see one of my daughters happily settled at Netherfield," said Mrs. Bennet to her husband, "and all the others equally well married, I shall have nothing to wish for."

In a few days Mr. Bingley returned Mr. Bennet's visit, and sat about ten minutes with him in his library. He had entertained hopes of being admitted to a sight of the young ladies, of whose beauty he had heard much; but he saw only the father. The ladies were somewhat more fortunate, for they had the advantage of ascertaining from an upper window that he wore a blue coat, and rode a black horse.

An invitation to dinner was soon afterwards dispatched; and already had Mrs. Bennet planned the courses that were to do credit to her housekeeping, when

Frequency variation in words from the same domain



Adapted from Google Books Ngram Viewer

Word frequency variation in different languages

English (COCA)

Rank	Word	Frequency
1	the	22038615
2	be	12545825
3	and	10741073
4	of	10343885
5	a	10144200
6	in	6996437
7	to	6332195
8	have	4303955
9	to	3856916
10	it	3872477
11	I	3978265
12	that	3430996
13	for	3281454
14	you	3081151
15	he	2909254
16	with	2683014
17	on	2485306
18	do	2573587
19	say	1915138
20	this	1885366
21	they	1865580
22	at	1767638
23	but	1776767

Word frequency variation in different languages

English (COCA)

Rank	Word	Frequency
1	the	22038615
2	be	12545825
3	and	10741073
4	of	10343885
5	a	10144200
6	in	6996437
7	to	6332195
8	have	4303955
9	to	3856916
10	it	3872477
11	I	3978265
12	that	3430996
13	for	3281454
14	you	3081151
15	he	2909254
16	with	2683014
17	on	2485306
18	do	2573587
19	say	1915138
20	this	1885366
21	they	1865580
22	at	1767638
23	but	1776767

Other example languages (Wiktionary)

Bulgarian (Български)

1. да 5906110
2. не 3720562
3. се 2988327
4. на 2397563
5. си 2099118
6. ще 1915058
7. за 1873135
8. ли 1667921
9. това 1624104
10. че 1596504
11. ти 1476416
12. от 1300802
13. ми 1244928
14. какво 1133986
15. го 1102525
16. съм 825557
17. по 774112
18. но 753428
19. аз 712170
20. те 707575
21. добре 678220
22. трябва 634551
23. ме 619025

Danish [edit]

1. er 1478564
2. jeg 1428307
3. det 1424533
4. du 1074169
5. ikke 796635
6. at 652275
7. en 616796
8. og 555335
9. har 535111
10. vi 462267
11. til 433992
12. på 403239
13. hvad 378442
14. mig 365426
15. med 361775
16. de 361028
17. den 358228
18. for 356369
19. der 351822
20. så 345575
21. dig 342354
22. han 318262
23. kan 306986

Dutch [edit]

1. ik 7840843
2. je 7242582
3. het 5233420
4. de 5171977
5. dat 4505789
6. is 4426258
7. een 3715858
8. niet 3678884
9. en 2800789
10. wat 2228941
11. van 2209684
12. we 2076863
13. in 1962675
14. ze 1789915
15. hij 1719998
16. op 1696645
17. te 1691827
18. zijn 1681160
19. er 1615173
20. maar 1598672
21. die 1434016
22. heb 1429701
23. me 1411318

Estonian (Eesti)

1. on 1043820
2. ma 953774
3. ei 756943
4. sa 629312
5. see 598939
6. et 512423
7. ja 467520
8. ta 421577
9. kui 315758
10. seda 266964
11. kas 239828
12. me 237133
13. mis 223191
14. mida 205148
15. pole 184063
16. jah 166757
17. oma 160264
18. aga 154097
19. siis 151521
20. olen 149541
21. nii 147466
22. oled 146760
23. ära 143521

Word frequency variation in different languages

English (COCA)

Rank	Word	Frequency
1	the	22038615
2	be	12545825
3	and	10741073
4	of	10343885
5	a	10144200
6	in	6996437
7	to	6332195
8	have	4303955
9	to	3856916
10	it	3872477
11	I	3978265
12	that	3430996
13	for	3281454
14	you	3081151
15	he	2909254
16	with	2683014
17	on	2485306
18	do	2573587
19	say	1915138
20	this	1885366
21	they	1865580
22	at	1767638
23	but	1776767

Other example languages (Wiktionary)

Bulgarian (Български)

1. да 5906110
2. не 3720562
3. се 2988327
4. на 2397563
5. си 2099118
6. ще 1915058
7. за 1873135
8. ли 1667921
9. това 1624104
10. че 1596504
11. ти 1476416
12. от 1300802
13. ми 1244928
14. какво 1133986
15. го 1102525
16. съм 825557
17. по 774112
18. но 753428
19. аз 712170
20. те 707575
21. добре 678220
22. трябва 634551
23. ме 619025

Danish [edit]

1. er 1478564
2. jeg 1428307
3. det 1424533
4. du 1074169
5. ikke 796635
6. at 652275
7. en 616796
8. og 555335
9. har 535111
10. vi 462267
11. til 433992
12. på 403239
13. hvad 378442
14. mig 365426
15. med 361775
16. de 361028
17. den 358228
18. for 356369
19. der 351822
20. så 345575
21. dig 342354
22. han 318262
23. kan 306986

Dutch [edit]

1. ik 7840843
2. je 7242582
3. het 5233420
4. de 5171977
5. dat 4505789
6. is 4426258
7. een 3715858
8. niet 3678884
9. en 2800789
10. wat 2228941
11. van 2209684
12. we 2076863
13. in 1962675
14. ze 1789915
15. hij 1719998
16. op 1696645
17. te 1691827
18. zijn 1681160
19. er 1615173
20. maar 1598672
21. die 1434016
22. heb 1429701
23. me 1411318

Estonian (Eesti)

1. on 1043820
2. ma 953774
3. ei 756943
4. sa 629312
5. see 598939
6. et 512423
7. ja 467520
8. ta 421577
9. kui 315758
10. seda 266964
11. kas 239828
12. me 237133
13. mis 223191
14. mida 205148
15. pole 184063
16. jah 166757
17. oma 160264
18. aga 154097
19. siis 151521
20. olen 149541
21. nii 147466
22. oled 146760
23. ära 143521

Discuss: What to make of this phenomenon?

Two thought experiments

- Experiment 1:
  - What happens if all words were equally frequent?
- Experiment 2:
  - What happens if all words were of equal length?

Zipf’s law

$$R \times F \cong C$$

R: Rank (or frequency rank) of a word
F: Frequency of a word
C: Constant

Zipf (1949)

Illustration of Zipf's law

English word frequencies revisited

Rank	Word	Frequency
1	the	22038615
2	be	12545825
3	and	10741073
4	of	10343885
5	a	10144200
6	in	6996437
7	to	6332195
8	have	4303955
9	to	3856916
10	it	3872477
11	I	3978265
12	that	3430996
13	for	3281454
14	you	3081151
15	he	2909254
16	with	2683014
17	on	2485306
18	do	2573587
19	say	1915138
20	this	1885366
21	they	1865580
22	at	1767638
23	but	1776767

Illustration of Zipf’s law

English word frequencies revisited

Rank	Word	Frequency
1	the	22038615
2	be	12545825
3	and	10741073
4	of	10343885
5	a	10144200
6	in	6996437
7	to	6332195
8	have	4303955
9	to	3856916
10	it	3872477
11	I	3978265
12	that	3430996
13	for	3281454
14	you	3081151
15	he	2909254
16	with	2683014
17	on	2485306
18	do	2573587
19	say	1915138
20	this	1885366
21	they	1865580
22	at	1767638
23	but	1776767

Zipf's law predicts: $R \times F \cong C$

Illustration of Zipf's law

English word frequencies revisited

Rank	Word	Frequency
1	the	22038615
2	be	12545825
3	and	10741073
4	of	10343885
5	a	10144200
6	in	6996437
7	to	6332195
8	have	4303955
9	to	3856916
10	it	3872477
11	I	3978265
12	that	3430996
13	for	3281454
14	you	3081151
15	he	2909254
16	with	2683014
17	on	2485306
18	do	2573587
19	say	1915138
20	this	1885366
21	they	1865580
22	at	1767638
23	but	1776767

Zipf's law predicts: $R \times F \cong C$

1 x 22038615 ~ 22000000
2 x 12545825 ~ 25000000
… …

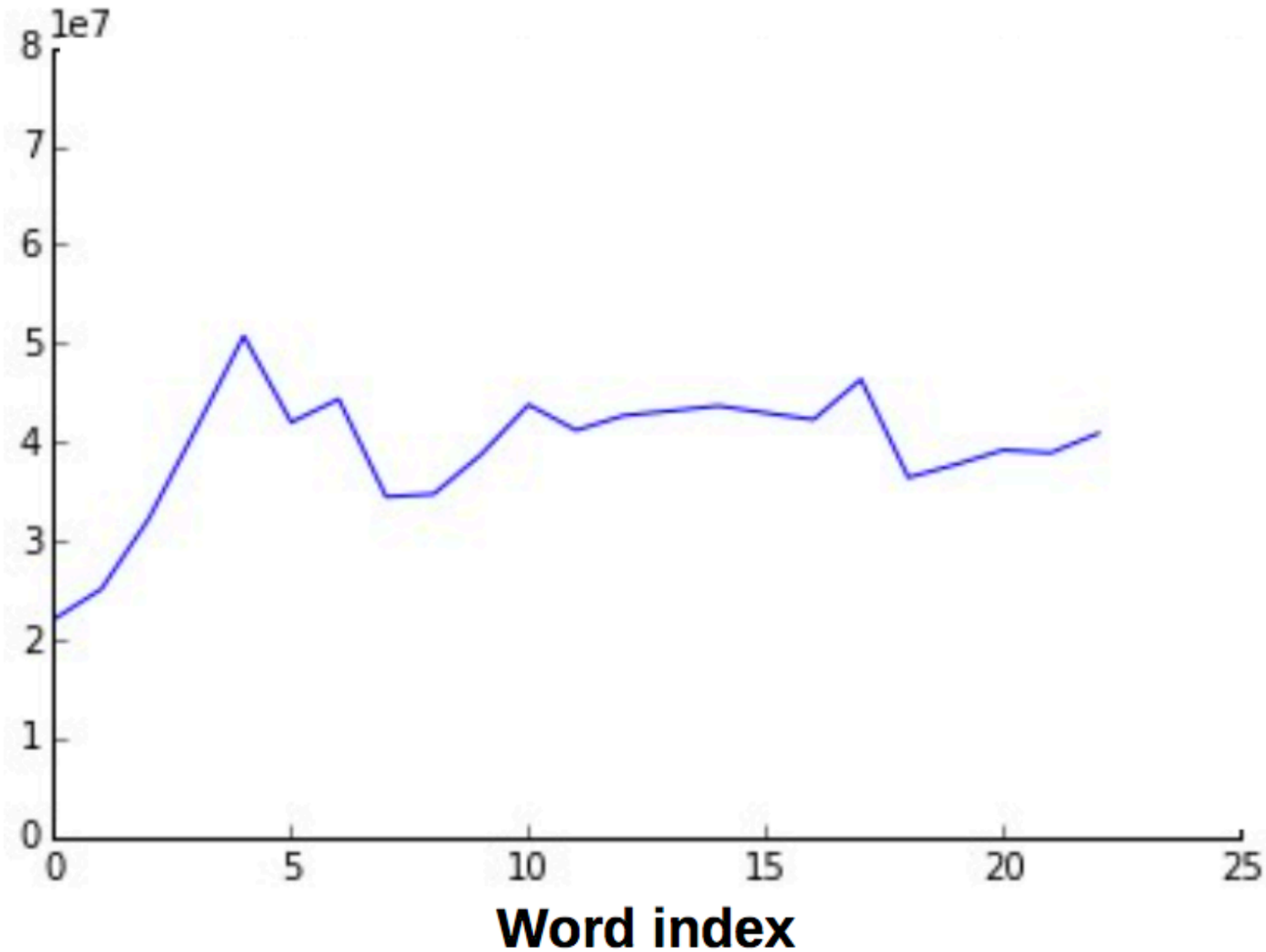


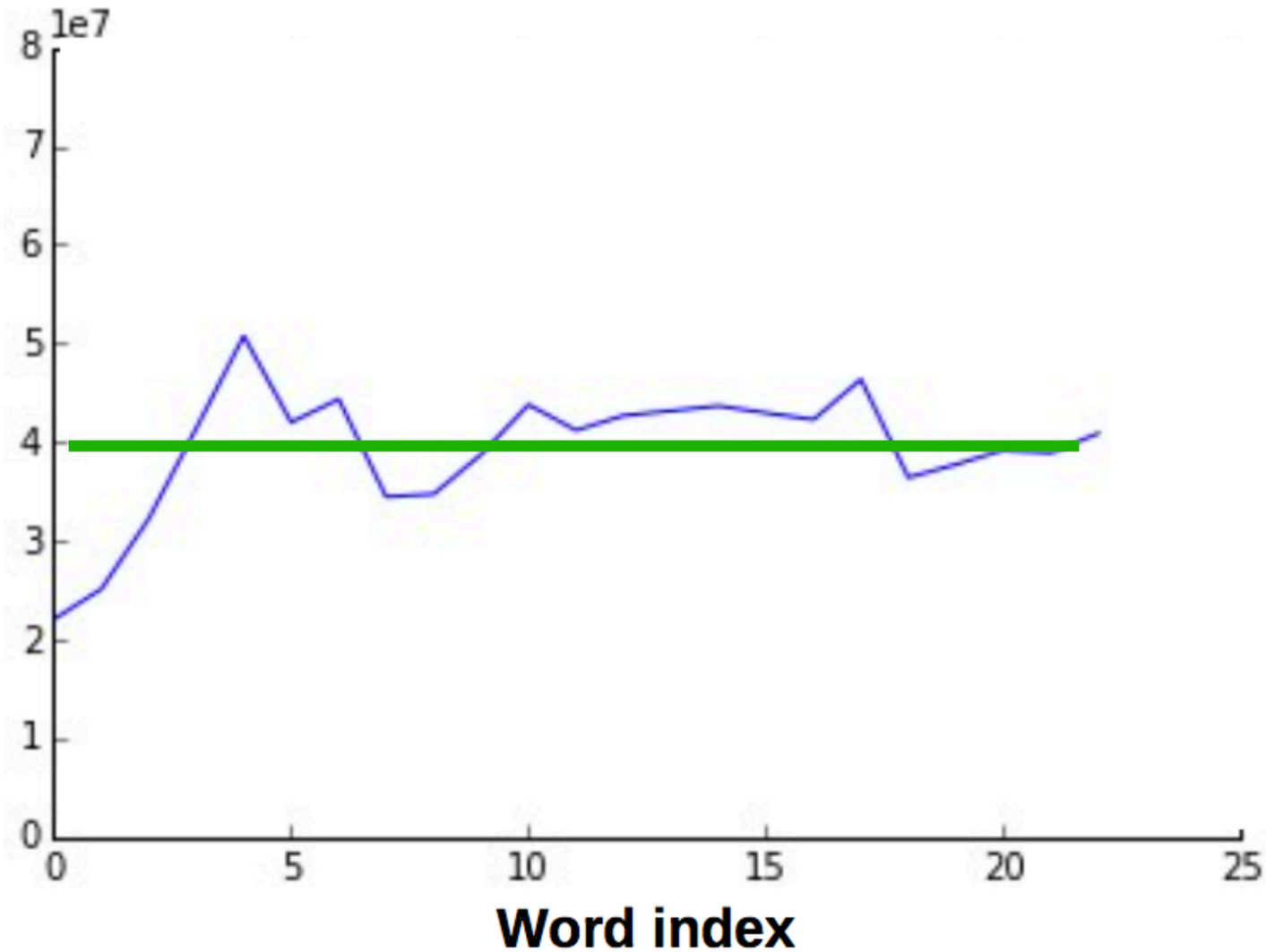
Illustration of Zipf's law

English word frequencies revisited

Rank	Word	Frequency
1	the	22038615
2	be	12545825
3	and	10741073
4	of	10343885
5	a	10144200
6	in	6996437
7	to	6332195
8	have	4303955
9	to	3856916
10	it	3872477
11	I	3978265
12	that	3430996
13	for	3281454
14	you	3081151
15	he	2909254
16	with	2683014
17	on	2485306
18	do	2573587
19	say	1915138
20	this	1885366
21	they	1865580
22	at	1767638
23	but	1776767

Zipf's law predicts: $R \times F \cong C$

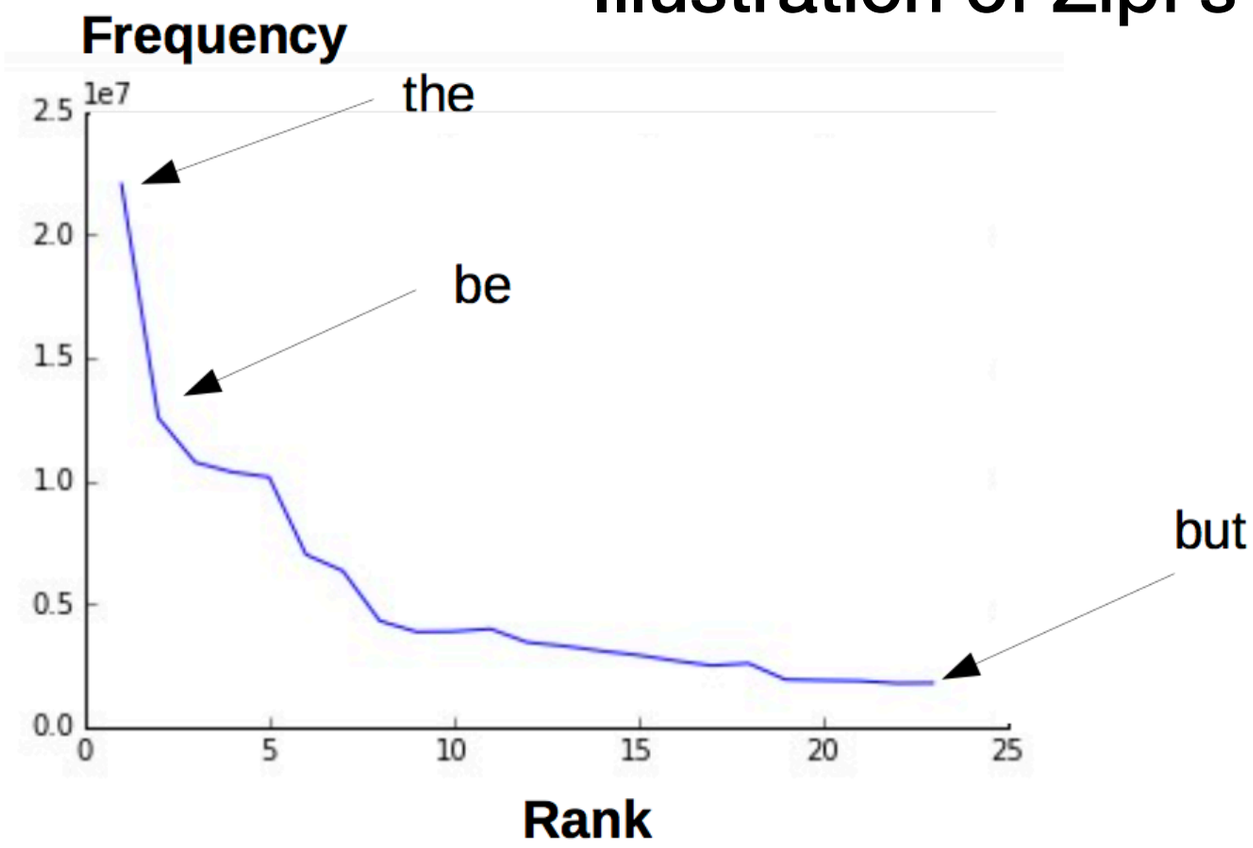
1 x 22038615 ~ 22000000
2 x 12545825 ~ 25000000
… …



English word frequencies revisited

Rank	Word	Frequency
1	the	22038615
2	be	12545825
3	and	10741073
4	of	10343885
5	a	10144200
6	in	6996437
7	to	6332195
8	have	4303955
9	to	3856916
10	it	3872477
11	I	3978265
12	that	3430996
13	for	3281454
14	you	3081151
15	he	2909254
16	with	2683014
17	on	2485306
18	do	2573587
19	say	1915138
20	this	1885366
21	they	1865580
22	at	1767638
23	but	1776767

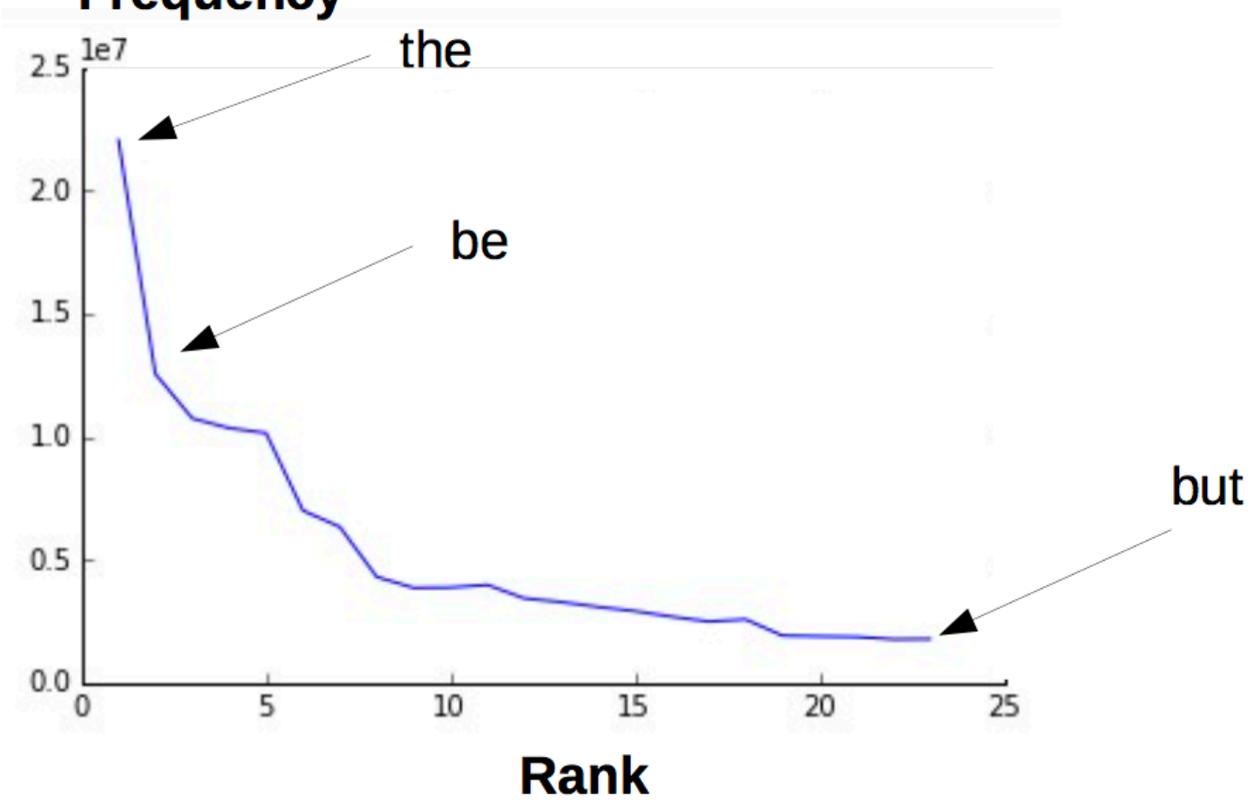
Illustration of Zipf's law

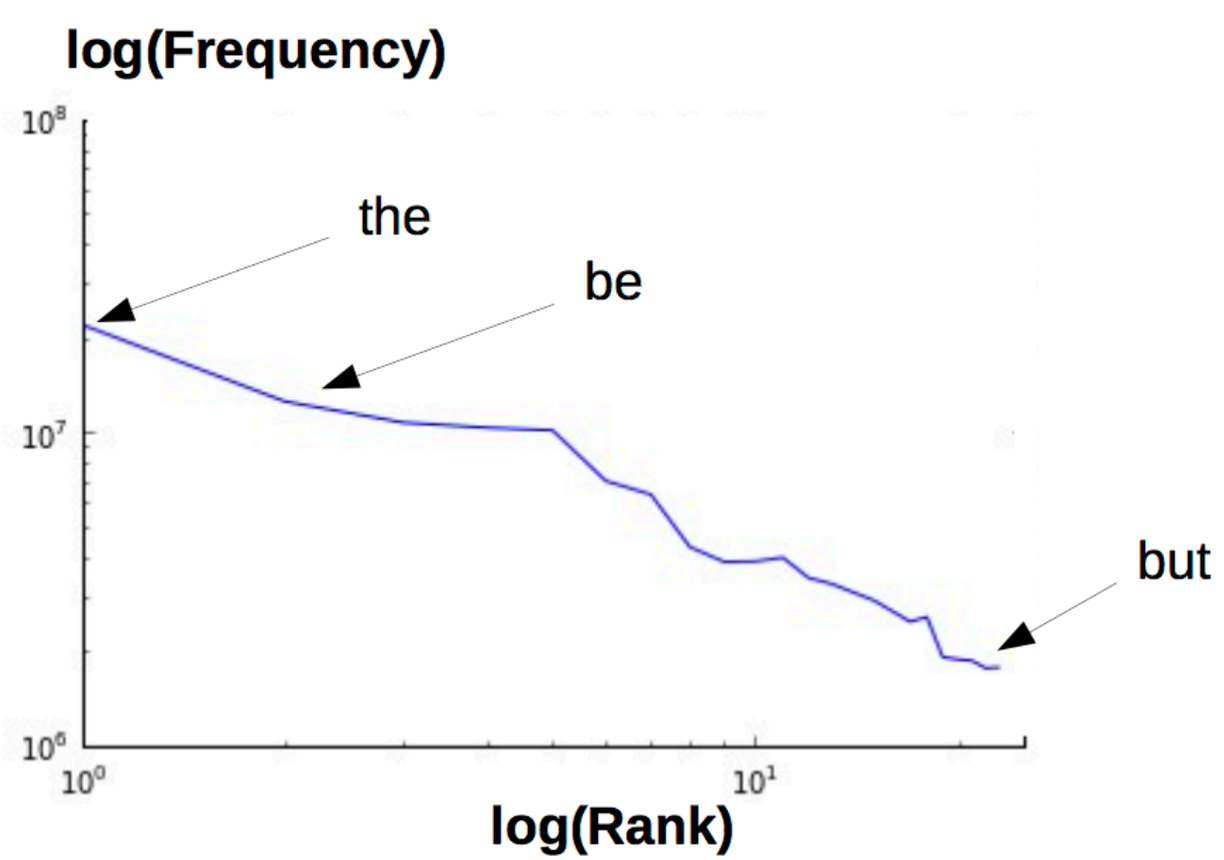


English word frequencies revisited

Rank	Word	Frequency
1	the	22038615
2	be	12545825
3	and	10741073
4	of	10343885
5	a	10144200
6	in	6996437
7	to	6332195
8	have	4303955
9	to	3856916
10	it	3872477
11	I	3978265
12	that	3430996
13	for	3281454
14	you	3081151
15	he	2909254
16	with	2683014
17	on	2485306
18	do	2573587
19	say	1915138
20	this	1885366
21	they	1865580
22	at	1767638
23	but	1776767

Illustration of Zipf's law

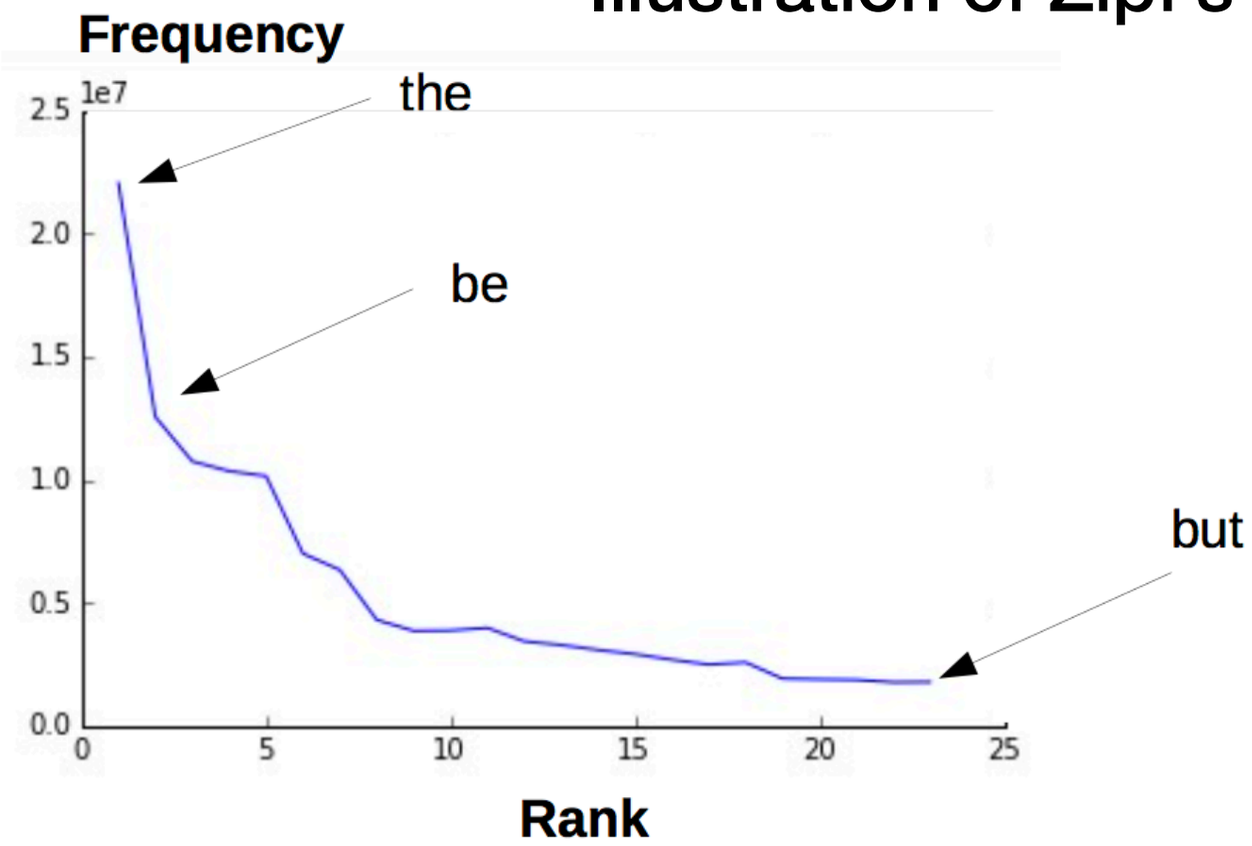


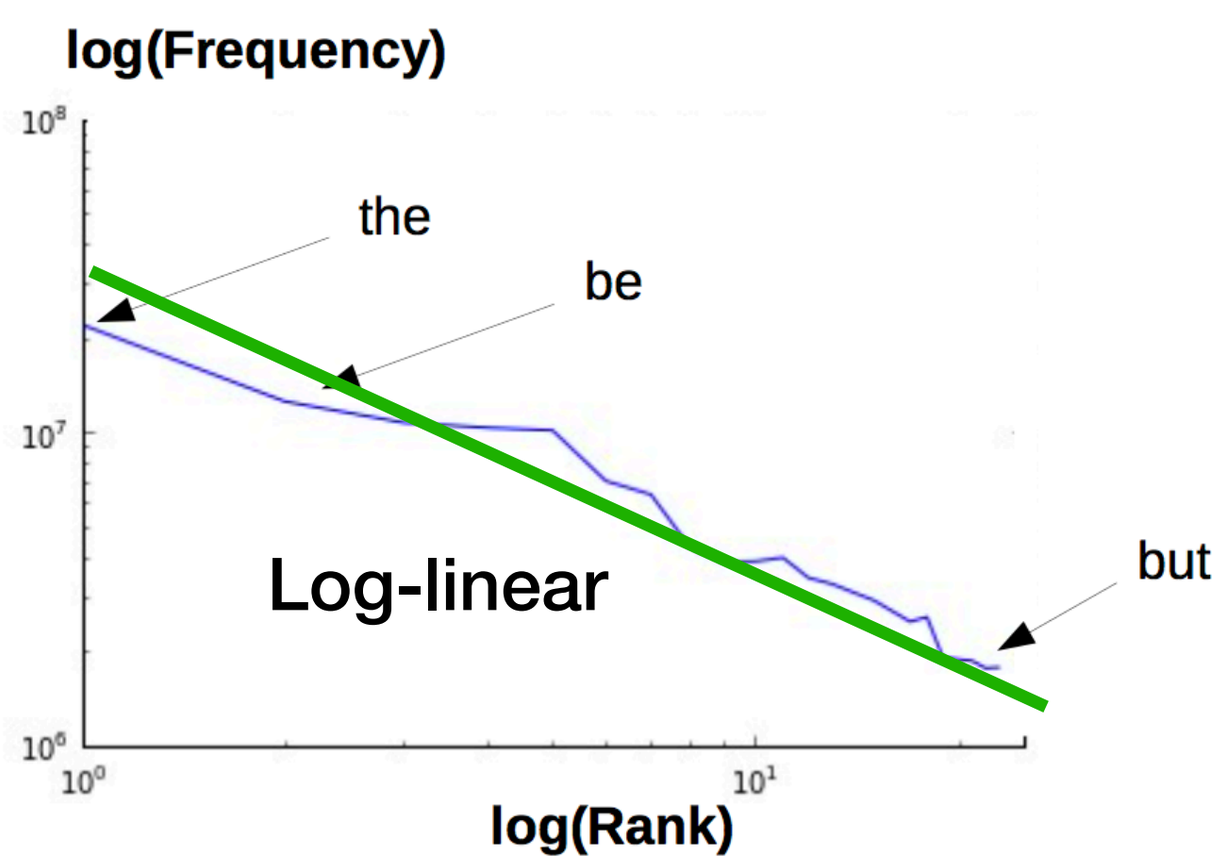


English word frequencies revisited

Rank	Word	Frequency
1	the	22038615
2	be	12545825
3	and	10741073
4	of	10343885
5	a	10144200
6	in	6996437
7	to	6332195
8	have	4303955
9	to	3856916
10	it	3872477
11	I	3978265
12	that	3430996
13	for	3281454
14	you	3081151
15	he	2909254
16	with	2683014
17	on	2485306
18	do	2573587
19	say	1915138
20	this	1885366
21	they	1865580
22	at	1767638
23	but	1776767

Illustration of Zipf's law





Derivation of Zip’f law

Equation of a line: y = ax + b

Derivation of Zip’f law

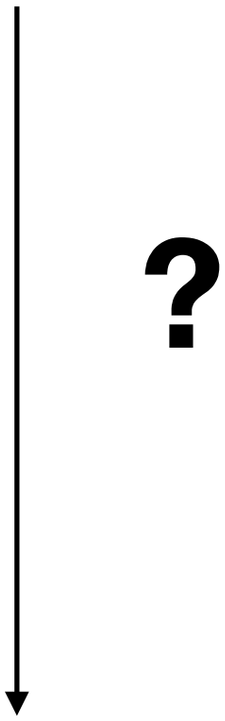
Equation of a line: y = ax + b

logF(requency) = a logR(ank) + b

Derivation of Zip’f law

Equation of a line: y = ax + b

logF(requency) = a logR(ank) + b



?

R x F = C

Derivation of Zip’f law

Equation of a line: y = ax + b

logF(requency) = a logR(ank) + b

logF - a logR = b

Derivation of Zip’f law

Equation of a line: y = ax + b

logF(requency) = a logR(ank) + b

logF - a logR = b

logF + (-a logR) = b

Derivation of Zip’f law

Equation of a line: y = ax + b

logF(requency) = a logR(ank) + b

logF - a logR = b

logF + (-a logR) = b

logF + (logR^{-a}) = b

Derivation of Zip’f law

Equation of a line: y = ax + b

logF(requency) = a logR(ank) + b

logF - a logR = b

logF + (-a logR) = b

logF + ($\log R^{-a}$) = b ($A\log B = \log B^{A}$)

Derivation of Zip’f law

Equation of a line: y = ax + b

logF(requency) = a logR(ank) + b

logF - a logR = b

logF + (-a logR) = b

logF + ($\log R^{-a}$) = b ($A\log B = \log B^{A}$)

log (F x R^{-a}) = b (logA + logB = logAB)

Derivation of Zip’f law

Equation of a line: y = ax + b

$$\log F(\text{requency}) = a \log R(\text{ank}) + b$$

$$\log F - a \log R = b$$

$$\log F + (-a \log R) = b$$

$$\log F + (\log R^{-a}) = b \quad (A \log B = \log B^{A})$$

$$\log (F \times R^{-a}) = b \quad (\log A + \log B = \log AB)$$

$$F \times R^{-a} = \exp(b)$$

Derivation of Zip’f law

Equation of a line: y = ax + b

logF(requency) = a logR(ank) + b

logF - a logR = b

logF + (-a logR) = b

logF + ($\log R^{-a}$) = b ($A \log B = \log B^{A}$)

log (F x R^{-a}) = b (logA + logB = logAB)

F x R^{-a} = exp(b)

F x R^{-a} = C (Zipf found that a=-1)

Derivation of Zip’f law

Equation of a line: y = ax + b

logF(requency) = a logR(ank) + b

logF - a logR = b

logF + (-a logR) = b

$\log F + (\log R^{-a}) = b$ ($A \log B = \log B^{A}$)

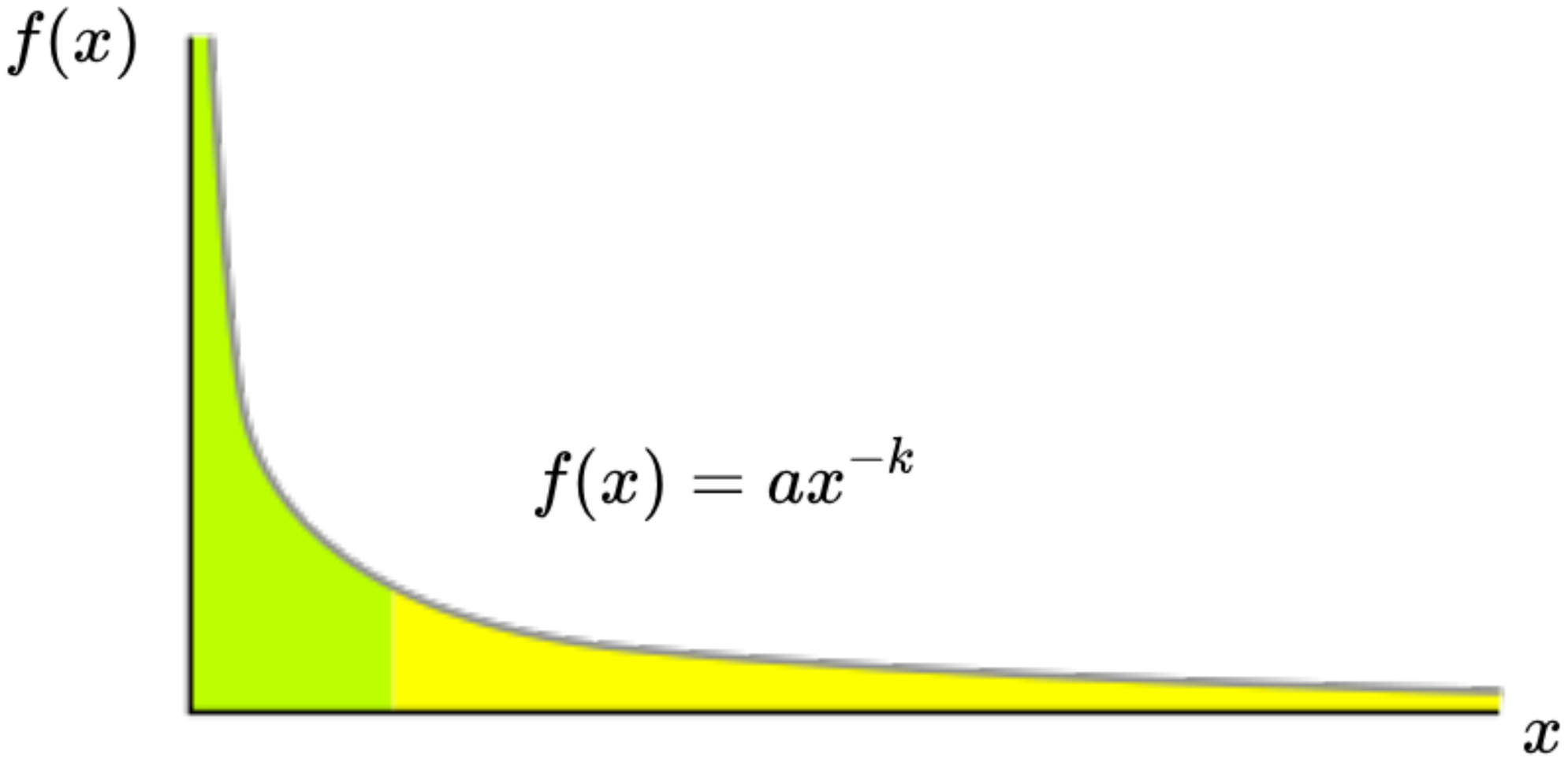
$\log (F \times R^{-a}) = b$ ($\log A + \log B = \log AB$)

$F \times R^{-a} = \exp(b)$

$F \times R^{-a} = C$ (Zipf found that a=-1)

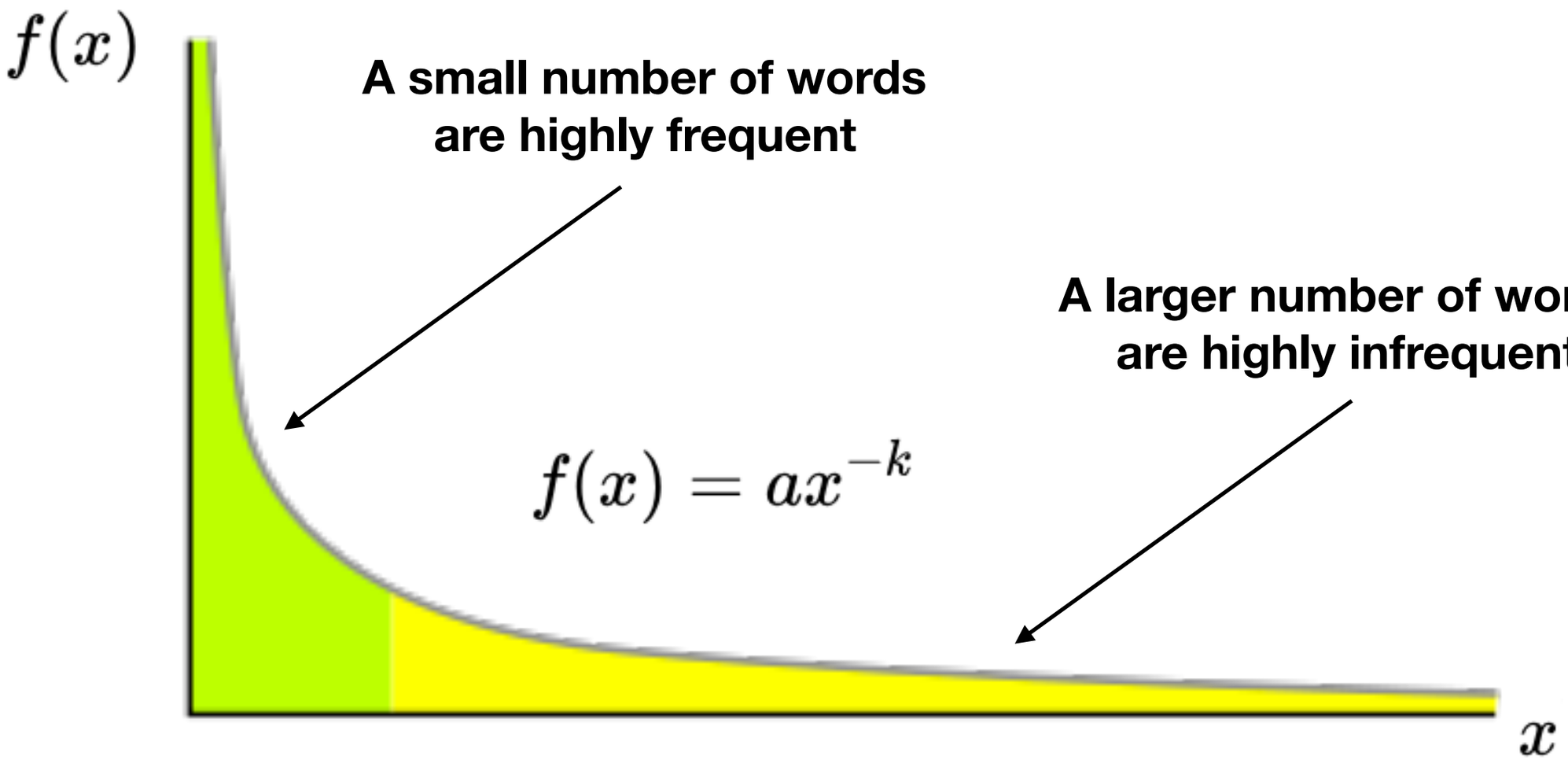
R x F = C

Link to power law



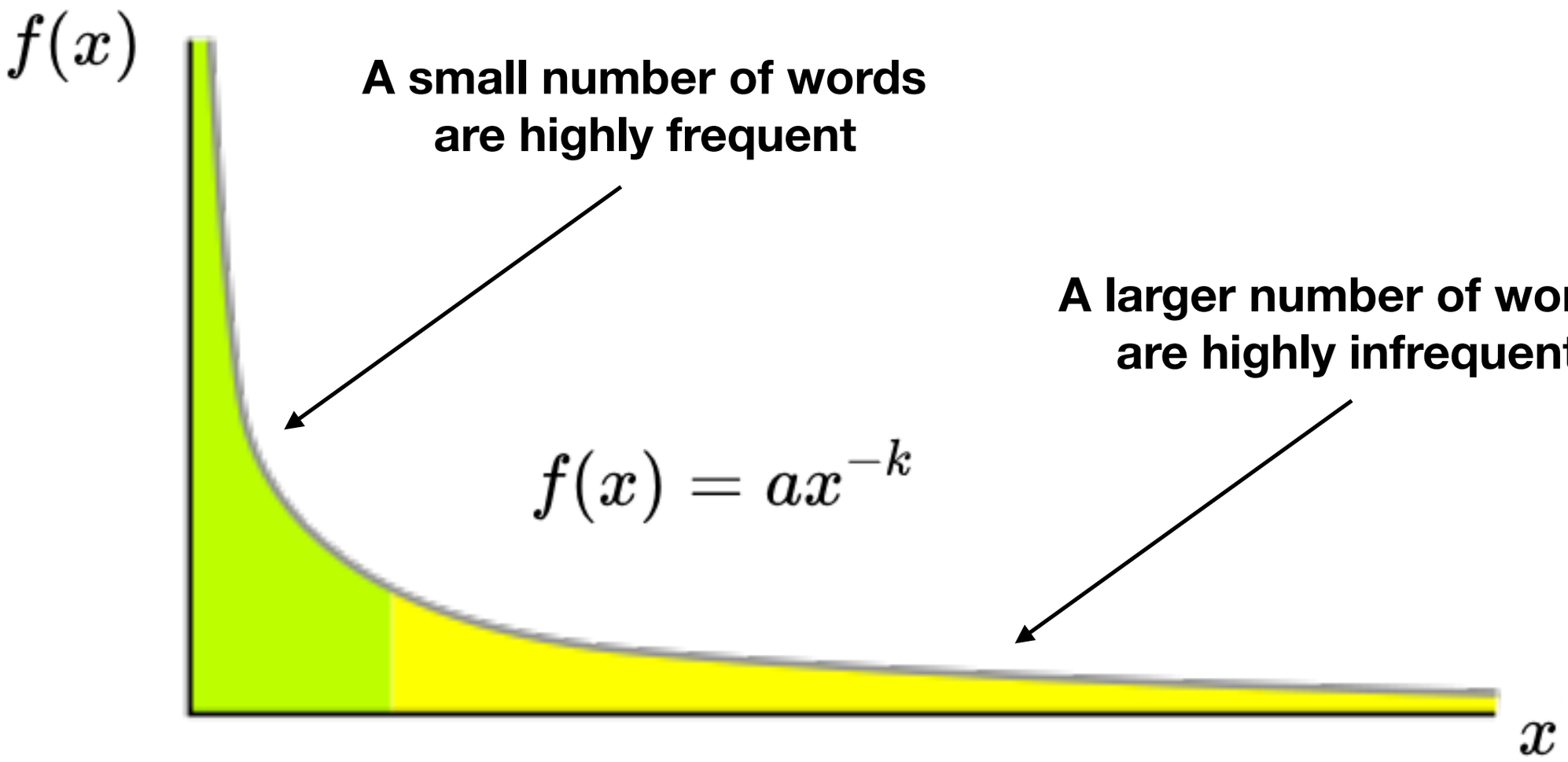
Here F = C x R^{-1}, where *a*=C and *k*=1

Link to power law



Here F = C x R^{-1}, where *a*=C and *k*=1

Discuss: Why?

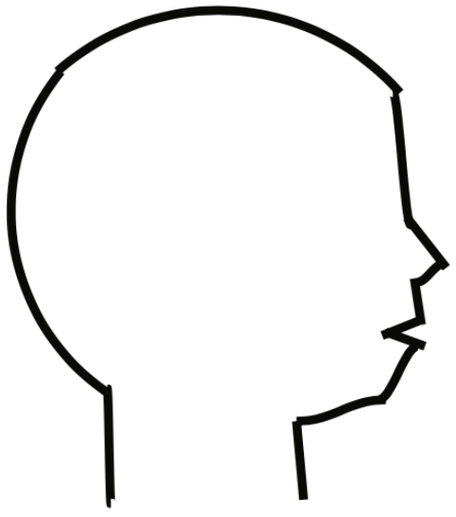


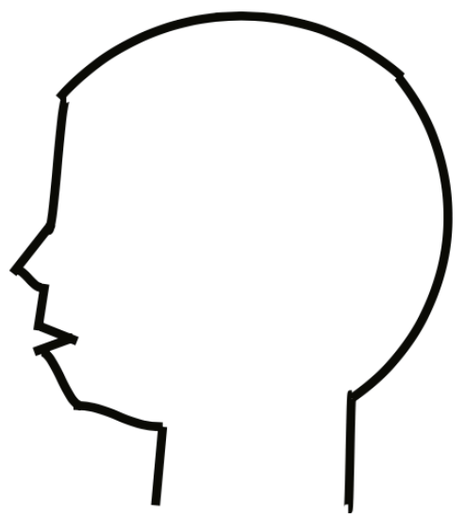
Here F = C x R^{-1}, where *a*=C and *k*=1

The principle of least effort

- Zipf's interpretation of his own law (Zipf, 1949):

Speaker **Listener**

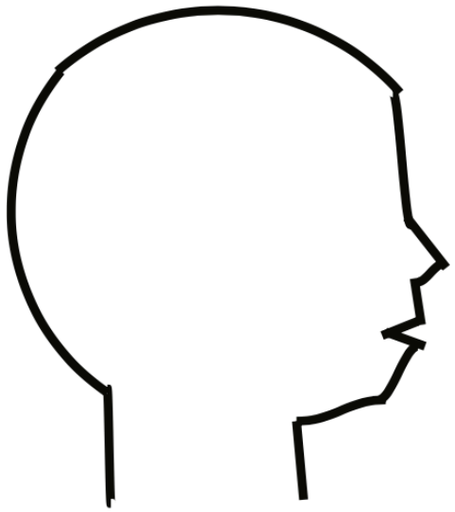




The principle of least effort

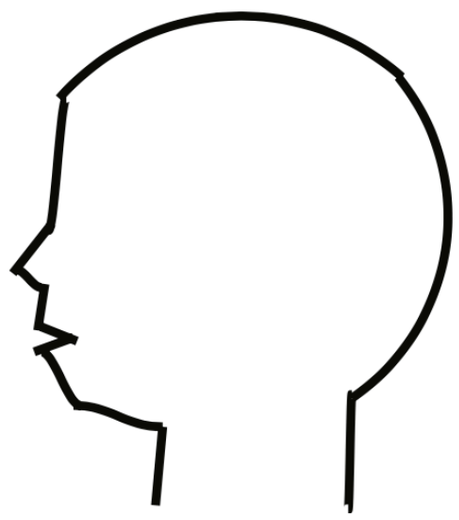
- Zipf's interpretation of his own law (Zipf, 1949):

Speaker



Say as few words as possible:
e.g., 1 word to be used highly frequently

Listener

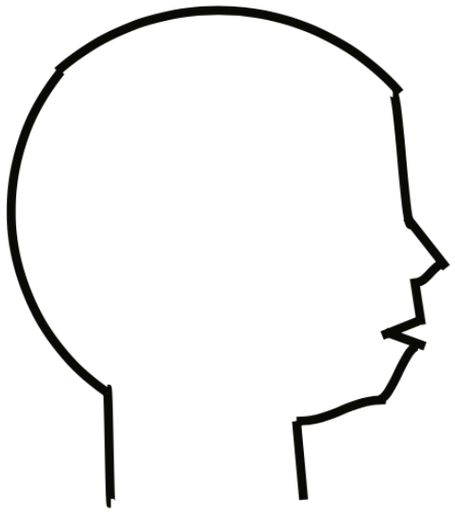


Hear as many words as possible:
e.g., 1 word for each possible meaning

The principle of least effort

- Zipf's interpretation of his own law (Zipf, 1949):

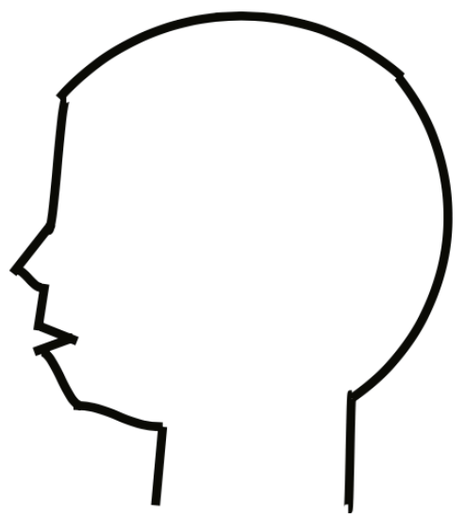
Speaker



Say as few words as possible:
e.g., 1 word to be used highly frequently

Predicts a lexicon with few words at very high frequencies

Listener



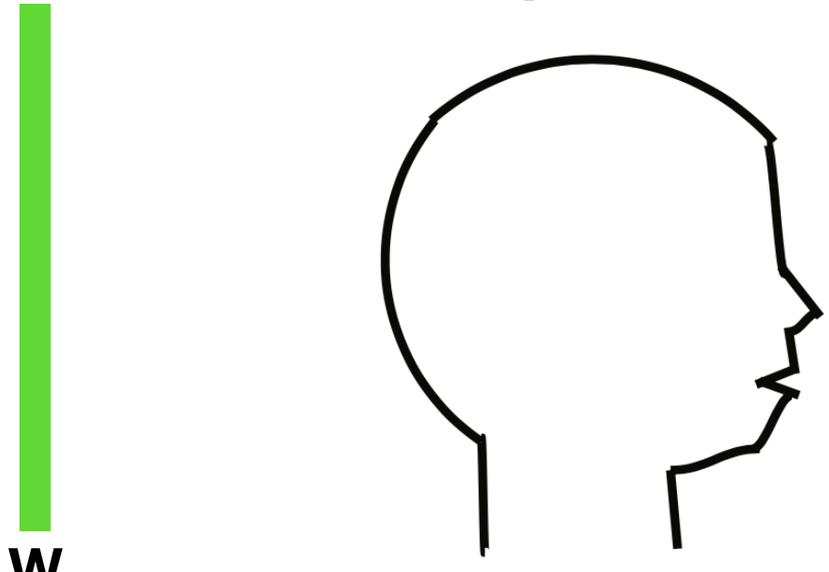
Hear as many words as possible:
e.g., 1 word for each possible meaning

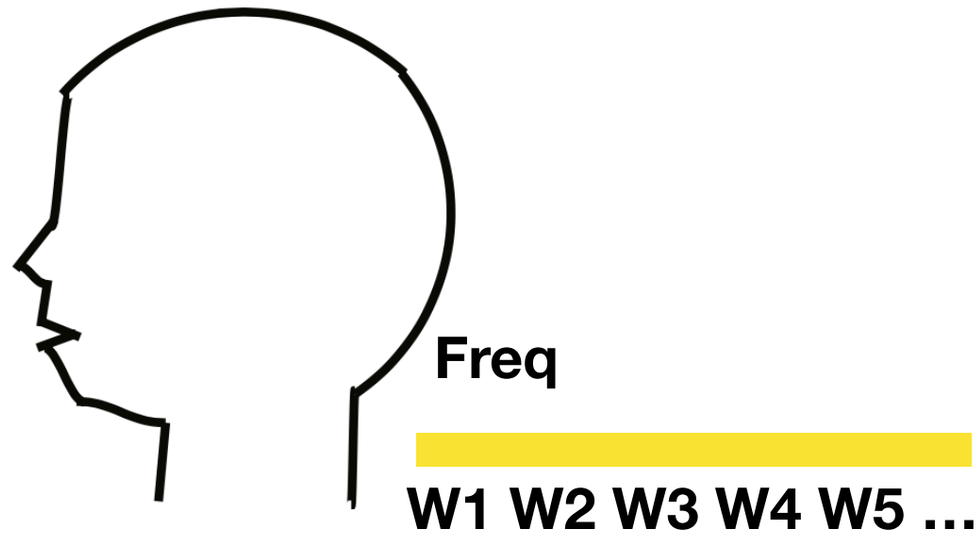
Predicts a lexicon with many words at very low frequencies

The principle of least effort

- Zipf's interpretation of his own law (Zipf, 1949):

Freq **Speaker** **Listener**





Say as few words as possible:
e.g., 1 word to be used highly frequently

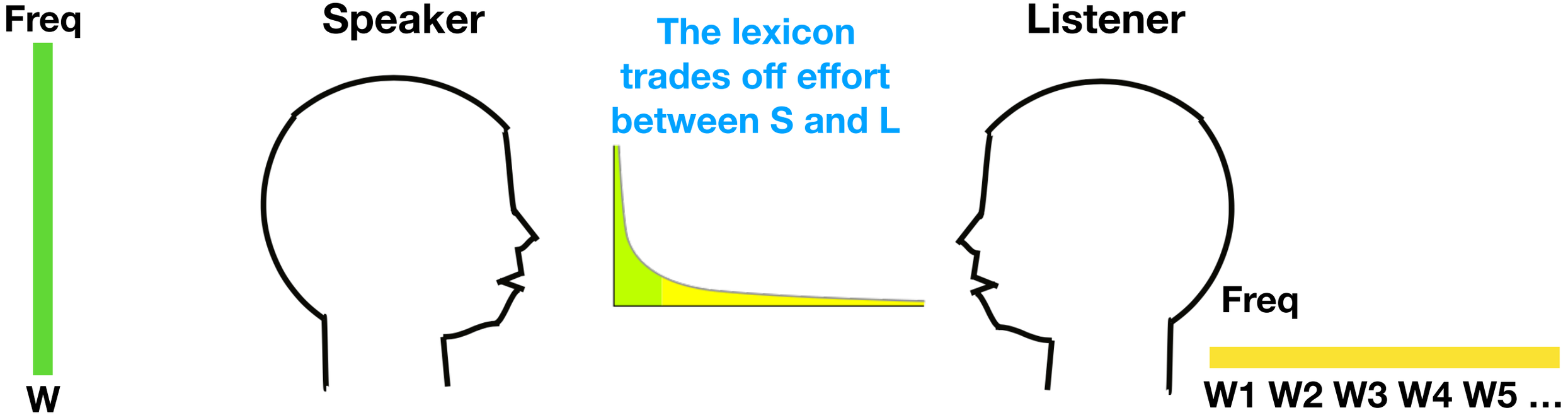
Predicts a lexicon with few words at very high frequencies

Hear as many words as possible:
e.g., 1 word for each possible meaning

Predicts a lexicon with many words at very low frequencies

The principle of least effort

- Zipf's interpretation of his own law (Zipf, 1949):



Two thought experiments

- Experiment 1:
  - What happens if all words were equally frequent?
- Experiment 2:
  - What happens if all words were of equal length?

The economy of words

Shorter words tend to be more frequent,
therefore the *expected word length* is short,
i.e. words are designed to save (cognitive) effort.

Zipf (1949)

The economy of words

Shorter words tend to be more frequent,
therefore the *expected word length* is short,
i.e. words are designed to save (cognitive) effort.

Lab 6

Zipf (1949)

An alternative view

– Piantadosi, S. T., Tily, H., and Gibson, E. (2011). Word lengths are optimized for efficient communication. Proceedings of the National Academy of Sciences, 108(9), 3526–3529.

5-minute break

Guess the sentence

Someone has a secret message for you.
Guess this sentence letter by letter.
You'll get a confirmation if correct.
Then move onto the next letter.

Shannon (1951)

Guess the sentence

Guess the sentence

F

Guess the sentence

FR

Guess the sentence

FRI

Guess the sentence

FRIE

Guess the sentence

FRIEN

Guess the sentence

FRIEND

Guess the sentence

FRIEND O

Guess the sentence

FRIEND OF

Guess the sentence

FRIEND OF M

Guess the sentence

FRIEND OF MI

Guess the sentence

FRIEND OF MIN

Guess the sentence

FRIEND OF MINE

Guess the sentence

FRIEND OF MINE F

Guess the sentence

FRIEND OF MINE FO

Guess the sentence

FRIEND OF MINE FOU

Guess the sentence

FRIEND OF MINE FOUN

Guess the sentence

FRIEND OF MINE FOUND

Guess the sentence

FRIEND OF MINE FOUND I

Guess the sentence

FRIEND OF MINE FOUND IT

Predictability in English

- Theory of communication (a.k.a. information theory)
    - Claude Shannon (1948; 1951)
    - Letter sequence in English words is non-arbitrary
    - e.g., if the current letter is “t”, the next letter is likely “h”

Predictability in English

- Theory of communication (a.k.a. information theory)
  - Claude Shannon (1948; 1951)
    - Letter sequence in English words is non-arbitrary
    - e.g., if the current letter is “t”, the next letter is likely “h”
    - The same observation holds for word sequence
    - e.g., if the current word is “it”, the next word is likely “is”

n-gram model

- A simple (based on co-occurring frequencies) model that captures basic statistical relations between letters and words

- Marginal probability: p(Li=“h”) ~ frequency(“h”)

- Conditional probability: p(Li=“h”|Li-1=“t”) ~ #“h” after “t”

n-gram model

- A simple (based on co-occurring frequencies) model that captures basic statistical relations between letters and words

- Marginal probability: p(Li=“h”) ~ frequency(“h”)

- Conditional probability: p(Li=“h”|Li-1=“t”) ~ #“h” after “t”

- N determines the window size in conditional probability:

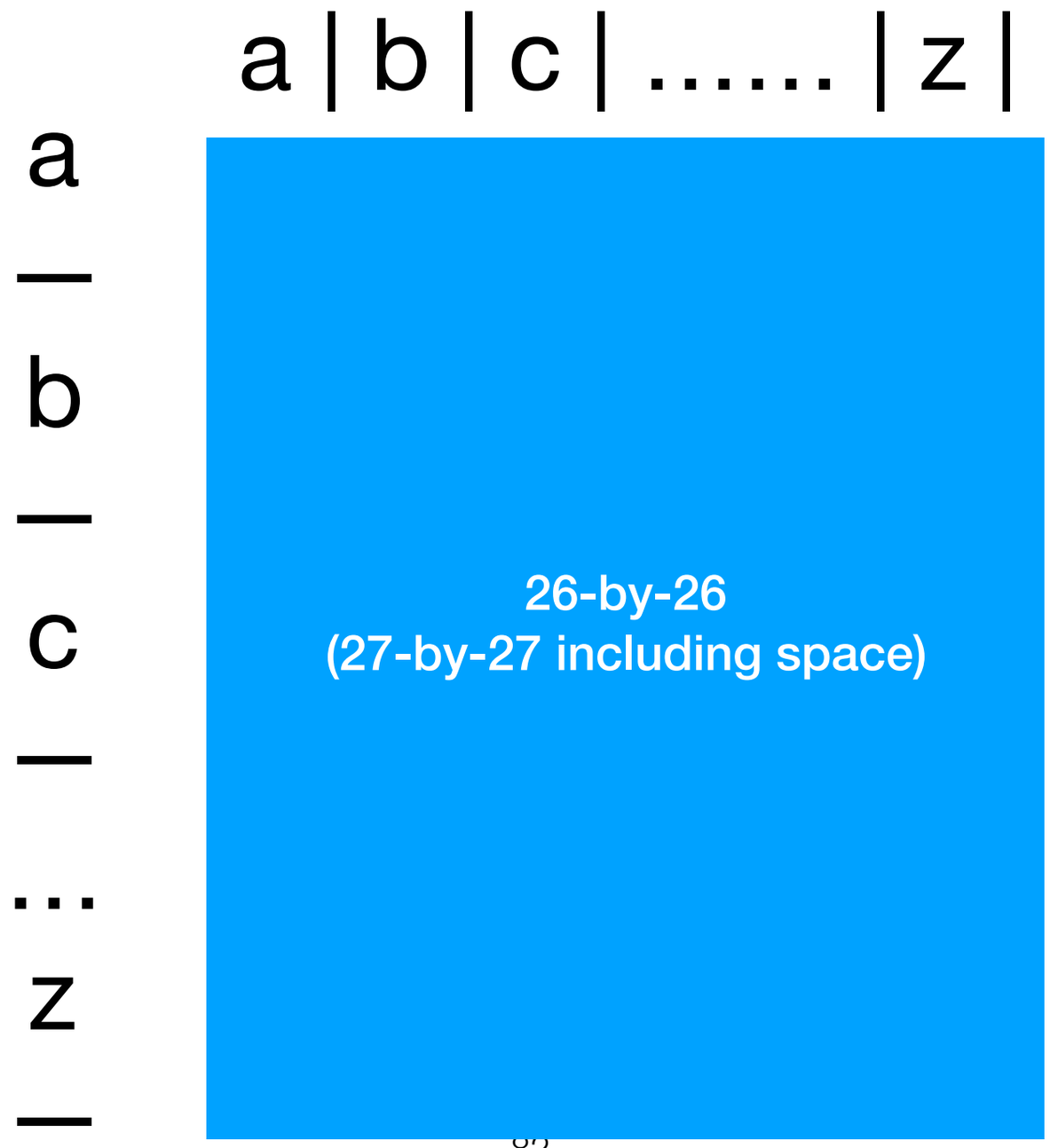
- N = 1: p(Li); N = 2: p(Li|Li-1); … N = n: p(Li|Li-1,Li-2,…,Li-n+1)

Conditional probability table

p(*Li*|*Li-1*)		*Li*		
		a	b	c
Li-1	a	0	$\frac{4}{5}$	$\frac{1}{5}$
	b	$\frac{1}{2}$	$\frac{1}{2}$	0
	c	$\frac{1}{2}$	$\frac{2}{5}$	$\frac{1}{10}$

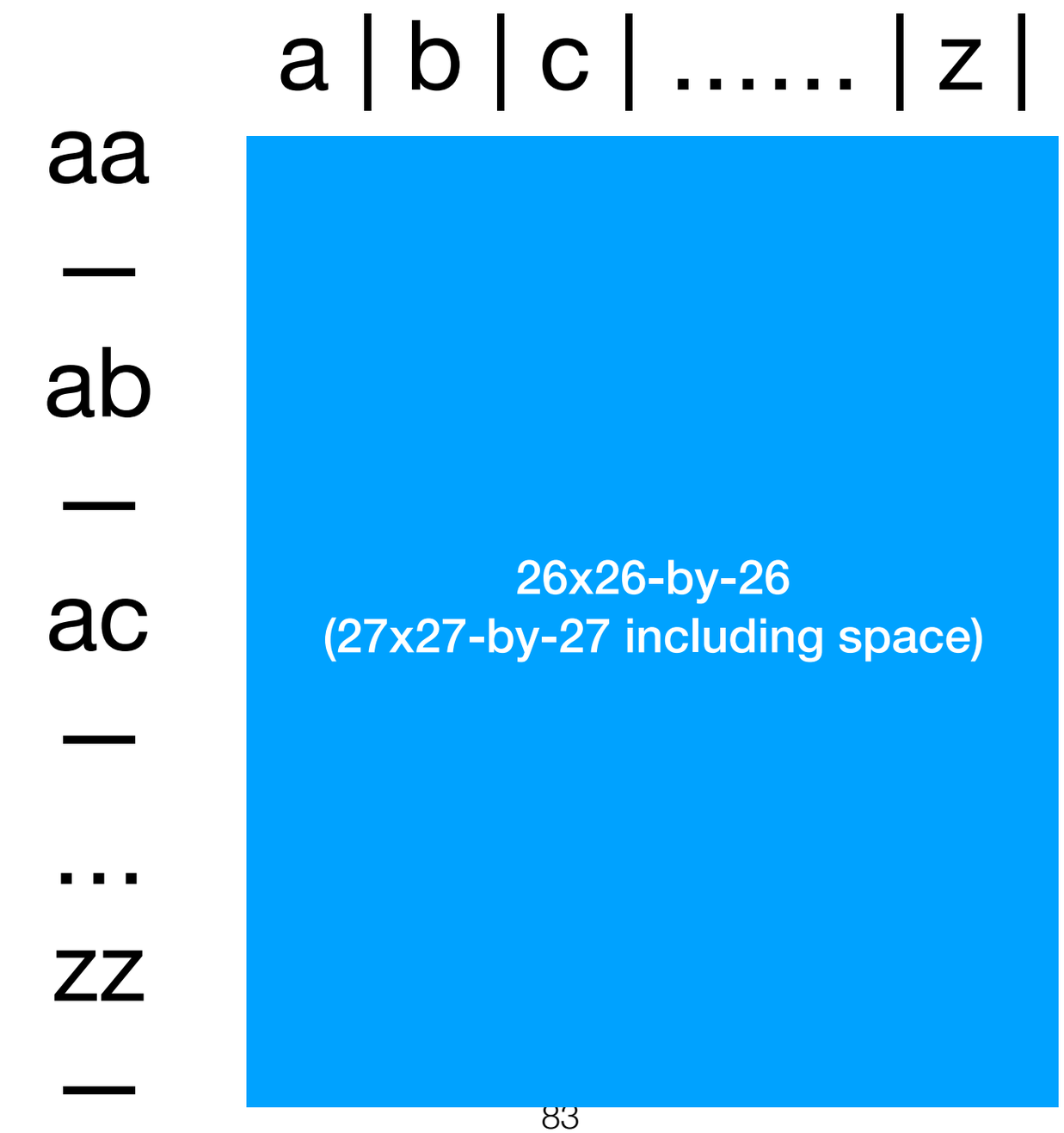
Bi-gram model
Second-order probabilities

p(*Li*|*Li-1*)



Trigram model
Third-order probabilities

p(*Li*|*Li-1*, *Li-2*)



1. Zero-order approximation (symbols independent and equiprobable).

Uniform frequency

XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD QPAAMKBZAACIBZL-HJQD.

1. Zero-order approximation (symbols independent and equiprobable).

Uniform frequency

XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD QPAAMKBZAACIBZL-HJQD.

2. First-order approximation (symbols independent but with frequencies of English text).

Eng letter frequency

OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL.

1. Zero-order approximation (symbols independent and equiprobable).

Uniform frequency

XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD QPAAMKBZAACIBZL-HJQD.

2. First-order approximation (symbols independent but with frequencies of English text).

Eng letter frequency

OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL.

3. Second-order approximation (digram structure as in English).

p(Li|Li-1)

ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TU-COOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE.

1. Zero-order approximation (symbols independent and equiprobable).

Uniform frequency

XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD QPAAMKBZAACIBZL-HJQD.

2. First-order approximation (symbols independent but with frequencies of English text).

Eng letter frequency

OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL.

3. Second-order approximation (digram structure as in English).

p(Li|Li-1)

ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TU-COOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE.

4. Third-order approximation (trigram structure as in English).

p(Li|Li-1,Li-2)

IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONS-TURES OF THE REPTAGIN IS REGOACTIONA OF CRE.

1. Zero-order approximation (symbols independent and equiprobable).

Uniform frequency

XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD QPAAMKBZAACIBZL-HJQD.

2. First-order approximation (symbols independent but with frequencies of English text).

Eng letter frequency

OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL.

3. Second-order approximation (digram structure as in English).

p(Li|Li-1)

ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TU-COOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE.

4. Third-order approximation (trigram structure as in English).

p(Li|Li-1,Li-2)

IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONS-TURES OF THE REPTAGIN IS REGOACTIONA OF CRE.

5. First-order word approximation. Rather than continue with tetragram, . . . , *n*-gram structure it is easier and better to jump at this point to word units. Here words are chosen independently but with their appropriate frequencies.

Eng word frequency (Zipf)

REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NAT-URAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE.

1. Zero-order approximation (symbols independent and equiprobable).

Uniform frequency

XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD QPAAMKBZAACIBZL-HJQD.

2. First-order approximation (symbols independent but with frequencies of English text).

Eng letter frequency

OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL.

3. Second-order approximation (digram structure as in English).

p(Li|Li-1)

ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TU-COOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE.

4. Third-order approximation (trigram structure as in English).

p(Li|Li-1,Li-2)

IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONS-TURES OF THE REPTAGIN IS REGOACTIONA OF CRE.

5. First-order word approximation. Rather than continue with tetragram, . . . , *n*-gram structure it is easier and better to jump at this point to word units. Here words are chosen independently but with their appropriate frequencies.

Eng word frequency (Zipf)

REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NAT-URAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE.

6. Second-order word approximation. The word transition probabilities are correct but no further struc-ture is included.

p(Wi|Wi-1)

THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHAR-ACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED.

Shannon (1951)

The applications of n-gram model

- Modeling sequences of natural phenomena, such as words, sentences, DNA sequences
- Sentence identification: Given two sentences, which is more plausible, i.e. higher in n-gram probability?
- Speech recognition, character/word recognition
- Machine translation
- ……

5-minute break

Word meaning

- Syntax can be dissociated with meaning (Chomsky):
  - “Colorless green ideas sleep furiously.”

Word meaning

- Syntax can be dissociated with meaning (Chomsky):
  - “Colorless green ideas sleep furiously.”
- Context/syntax can inform word meaning (Gleitman):
  - “He *daxied* ten dollars in the slot machine.”

Word meaning

- Syntax can be dissociated with meaning (Chomsky):
  - “Colorless green ideas sleep furiously.”
- Context/syntax can inform word meaning (Gleitman):
  - “He *daxied* ten dollars in the slot machine.”
- Word meaning and word vectors (*LSA*, *word2vec*, etc.):
  - Graduate-level (winter): CSC2611 on semantic change

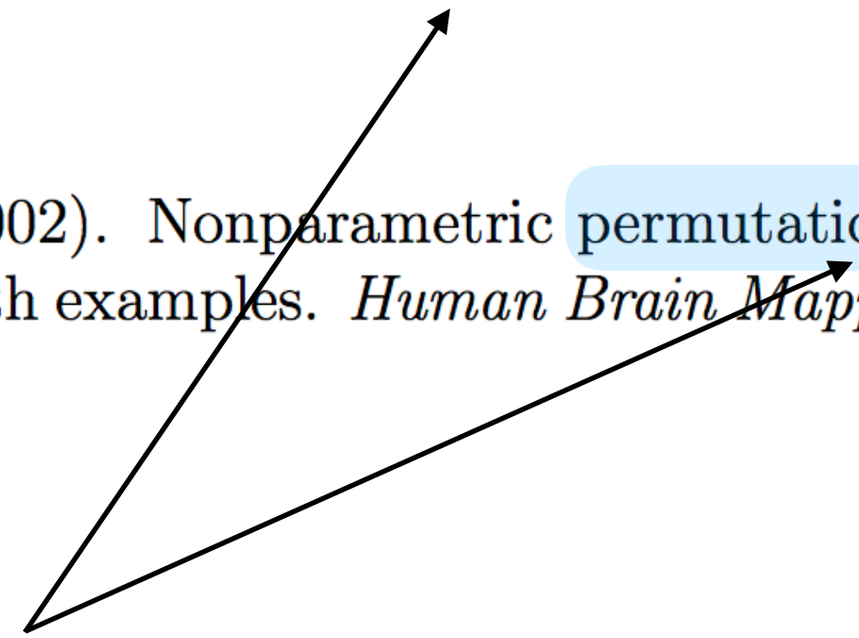
Readings

Required reading:

- Chapter 2 in Zipf, G. K. (1949). *Human behavior and the principle of least effort.* Addison-Wesley Press.

Technical reference:

- Nichols, T. E., and Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, 15(1), 1–25.



Lab 6
(also see Addendum posted)

Optional readings

Optional readings:

- Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27, 379-423 and 623-656.
- Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell system technical journal*, 30, 50–64.
- Piantadosi, S. T., Tily, H., and Gibson, E. (2011). Word lengths are optimized for efficient communication. Proceedings of the National Academy of Sciences, 108(9), 3526–3529.

Recommended book:

- Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use.* Greenwood.

Summary

- **The economy of words**: Zipf's law suggests an inverse relationship between word frequency and rank; further, shorter words tend to be more frequent, so the expected word length is short (you will show this in Lab 6).

- **Predictability of words**: Shannon's work suggests that sentences rely on high-order (conditional) frequencies of words, and models such as n-gram allow English-like sentences to be constructed.

Question of the day

- How did word frequencies become Zipfian distributed?

Lab 6: Word frequency (Zipf's law)

- A self-taught lab:
  - Dictionary: See Section 2.4 of tutorial on syllabus
  - List comprehension: See Sections 2.2.1.7 and 2.2.1.8
- Explain:
  - Expectation $E[X] = \sum x\, p(x) = \sum x\, [\, freq(x) / \sum freq(x)\,]$
  - Permutation: null hypothesis/distribution, test statistic, *p*-value

Permutation test

- A **permutation test** (also called a randomization test, re-randomization test, or an exact test) is a type of statistical significance test in which the distribution of the test statistic under the null hypothesis is obtained by calculating all possible values of the test statistic under rearrangements of the labels on the observed data points. (from Wikipedia)

Permutation test

Observed

A B

Suppose test statistic is: **Dobs = mean(B) - mean(A)**

What is the null hypothesis?

Null hypothesis

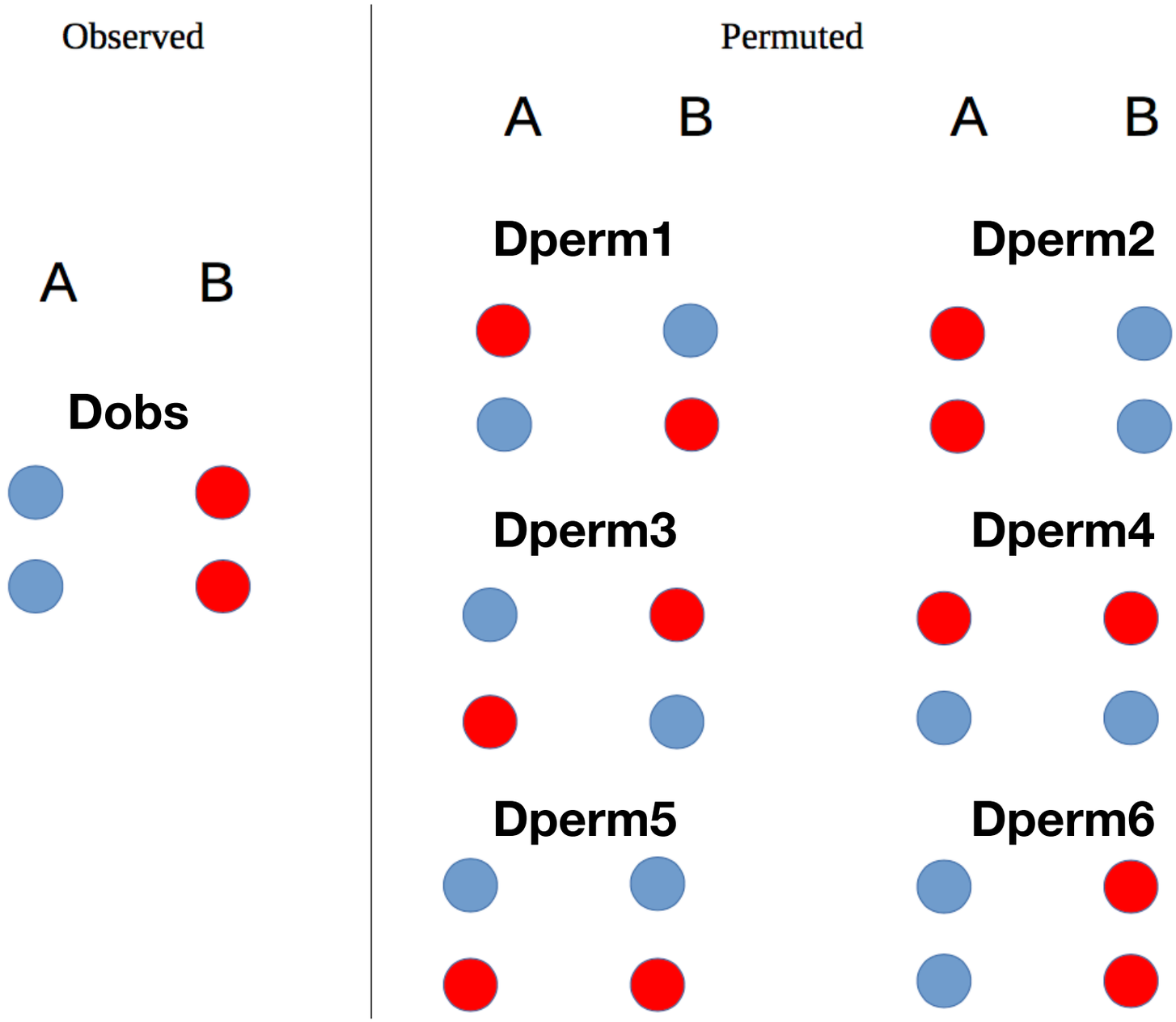
Observed

A B

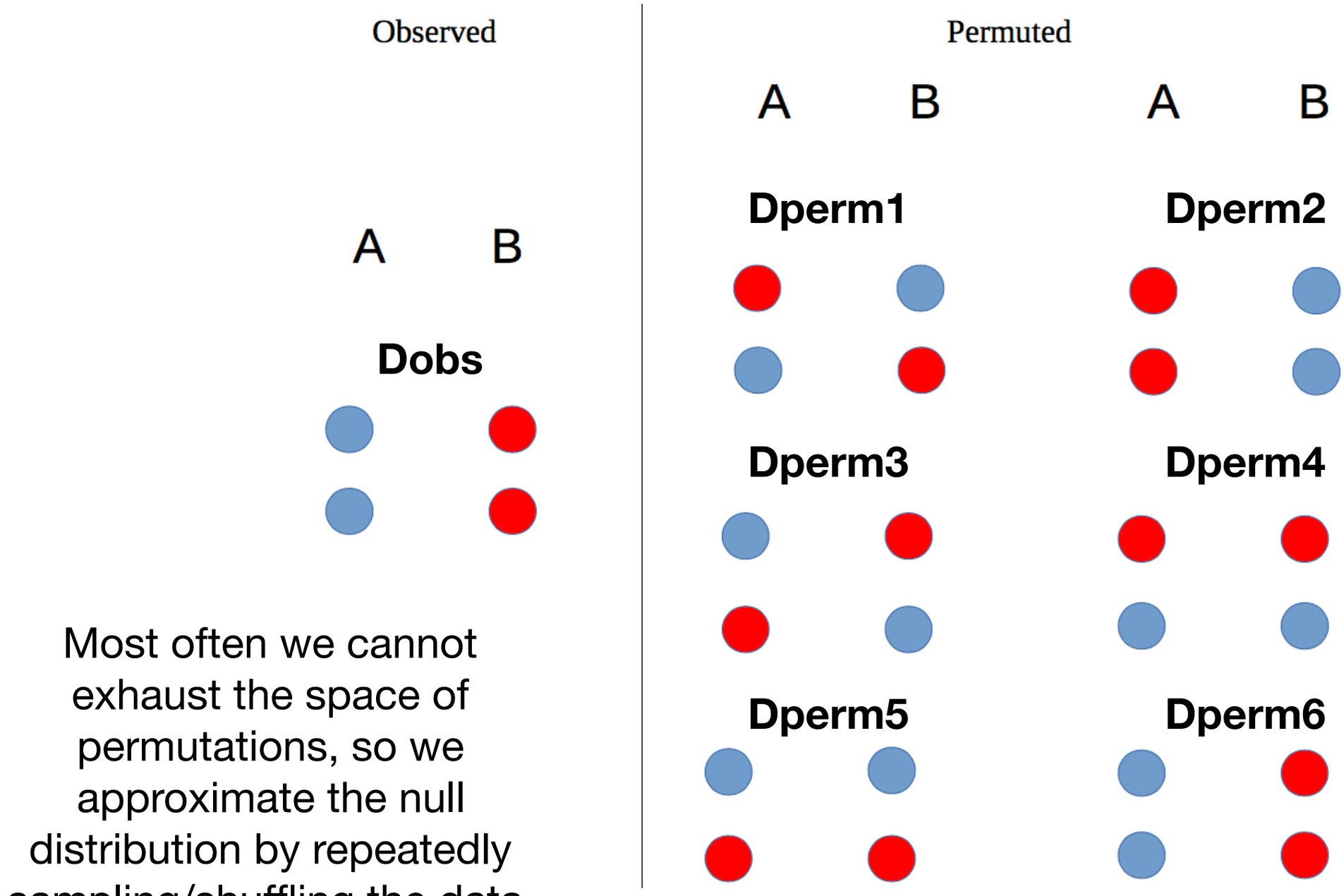
Suppose test statistic is: **Dobs = mean(B) - mean(A)**

What is the null hypothesis? Null: There is no difference between groups, or Dobs ~= 0

Null distribution



Null distribution



Most often we cannot exhaust the space of permutations, so we approximate the null distribution by repeatedly sampling/shuffling the data.

Statistical significance

- *p*-value: The probability of observing the test statistic under the null distribution.

$$p = \frac{\text{sum (number of items in } Dperm \geq Dobs)}{\text{Total number of permutations}}$$

where “p” is typically referred to as the p value which is an indicator of significance. The equation above says that the degree of significance is determined by the fraction of values in *Dperm* that appears to be greater than or equal to the observed value *Dobs*. P-value is always between 0 and 1 (do you see why from the equation above?), and in general, a small p-value (e.g. $p<0.05$) is taken as evidence for establishing statistical significance that the null can be rejected, hence as an alternative way of saying that the observed difference is more extreme than chance.

- Typically, we claim significance if $p < 0.05$