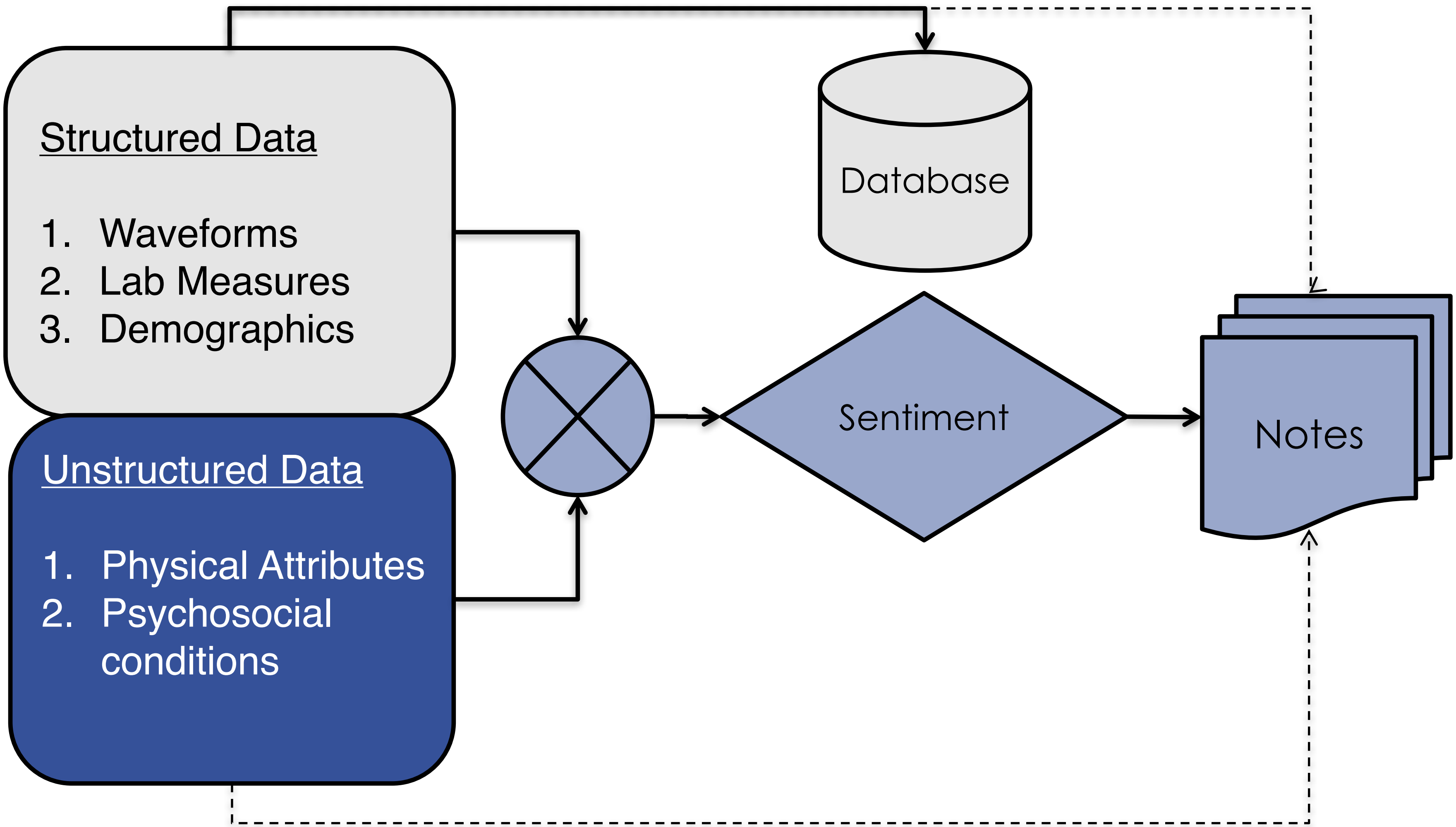
A Visualization of Evolving Clinical Sentiment Using Vector Representations of Clinical Notes

Mohammad M. Ghassemi, MIT
Roger G. Mark, MIT
Shamim Nemati, Emory University

- Motivation
- Hypothesis

INTRODUCTION

Clinical Notes are Shaped by Clinicians' Sentiment



Hypothesis

The judgment of care providers is driven by comprehensive observations of the patient, and this judgment may be reflected in the structural complexity and sentiment of their written patient notes.

We investigate our hypothesis by analyzing the evolution of the **sentiment** and **language** use over time and patient category.

Hypothesis

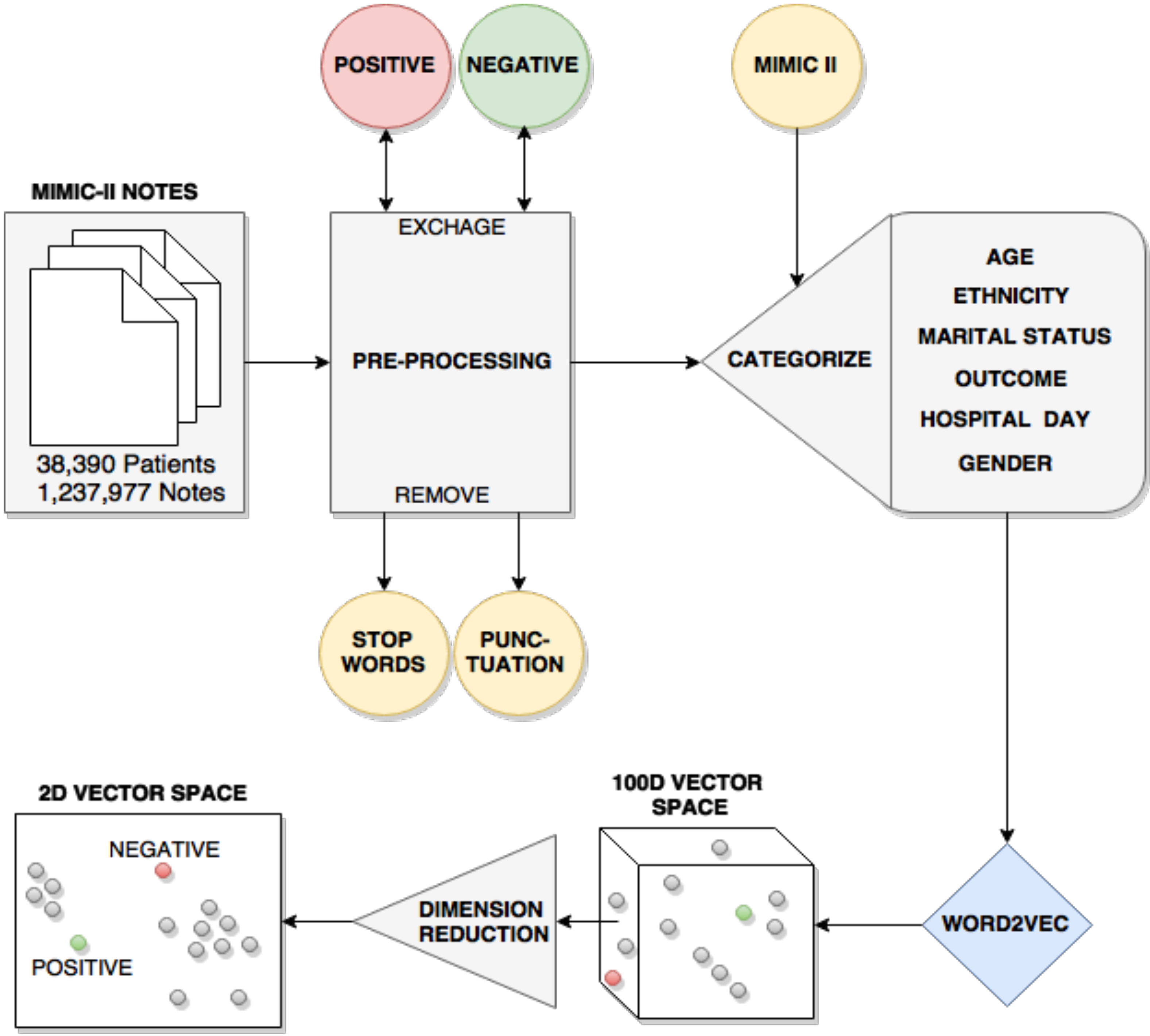
The judgment of care providers is driven by comprehensive observations of the patient, and this judgment may be reflected in the structural com ment of their written patien

- Age group
- Gender
- Marital Status
- Length of hospital stay
- Race/ethnicity
- Patient Outcome

We investiga by analyzing the evolution nd **language** use over time and patient category.

- Methods & Analysis
- Word2Vec Tool
- Illustration

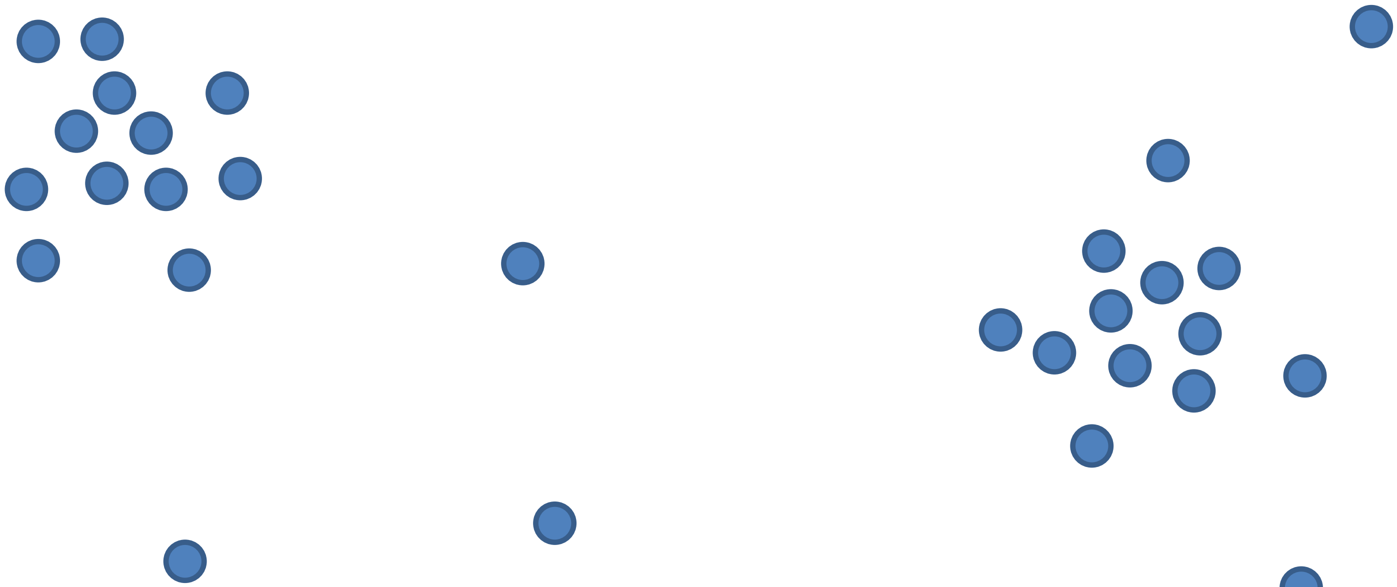
METHODS


POSITIVE
NEGATIVE
MIMIC II
MIMIC-II NOTES
38,390 Patients
1,237,977 Notes
EXCHAGE
PRE-PROCESSING
REMOVE
STOP WORDS
PUNC-TUATION
CATEGORIZE
AGE
ETHNICITY
MARITAL STATUS
OUTCOME
HOSPITAL DAY
GENDER
WORD2VEC
100D VECTOR SPACE
DIMENSION REDUCTION
2D VECTOR SPACE
NEGATIVE
POSITIVE

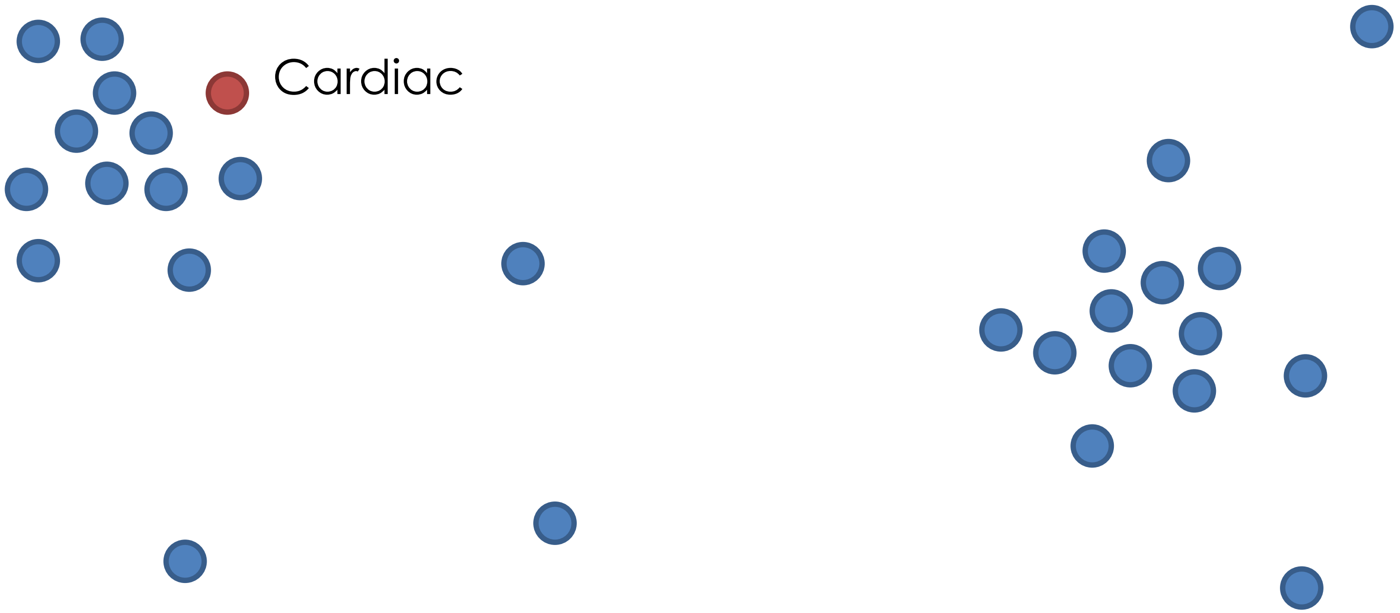
The Word2vec Tool

- Word2Vec describes a class of neural network models that, given an unlabeled training corpus, produce a vector for each word in the corpus that encodes it’s semantic information.
- Semantic similarity is measured by cosine distance.

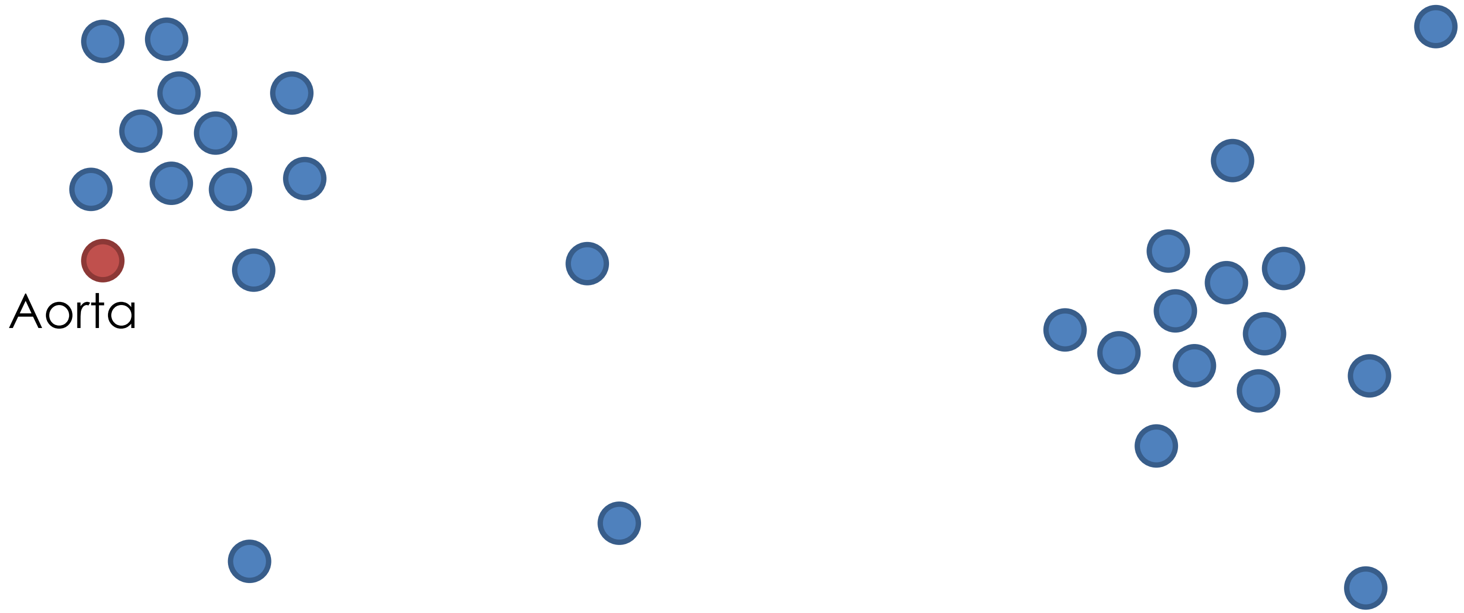
Individual Words in the Vector Space



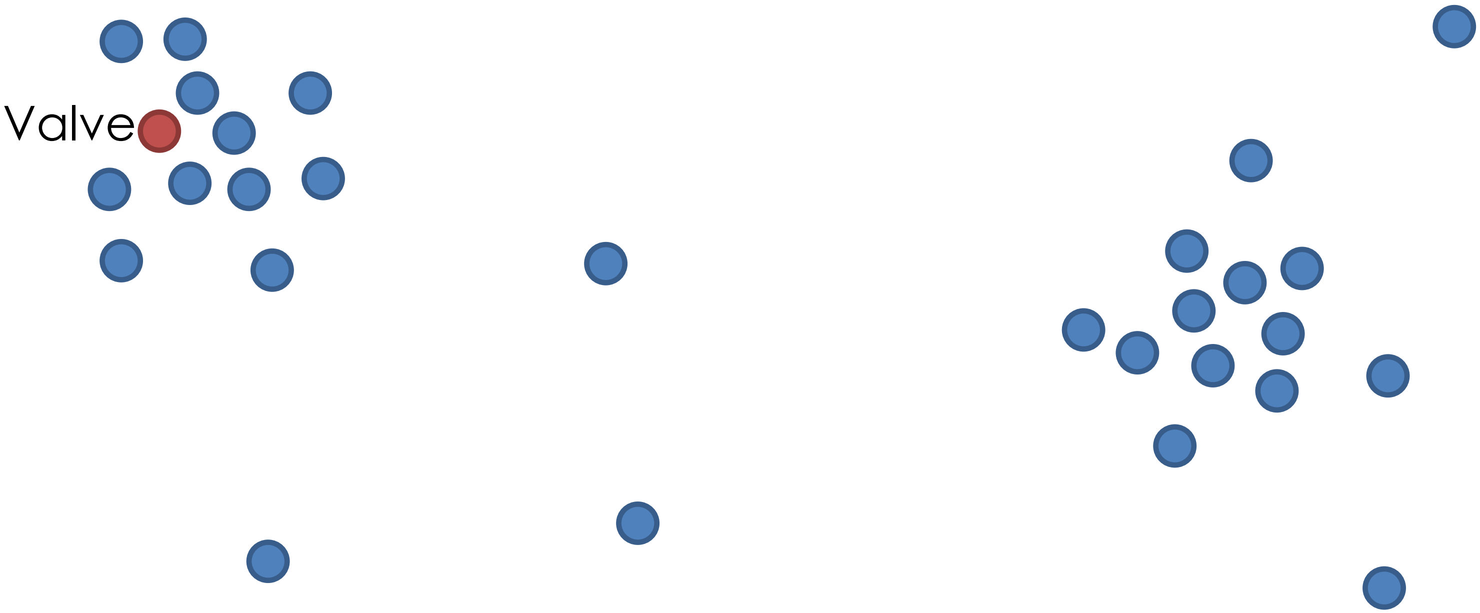
Similar Words Cluster Together



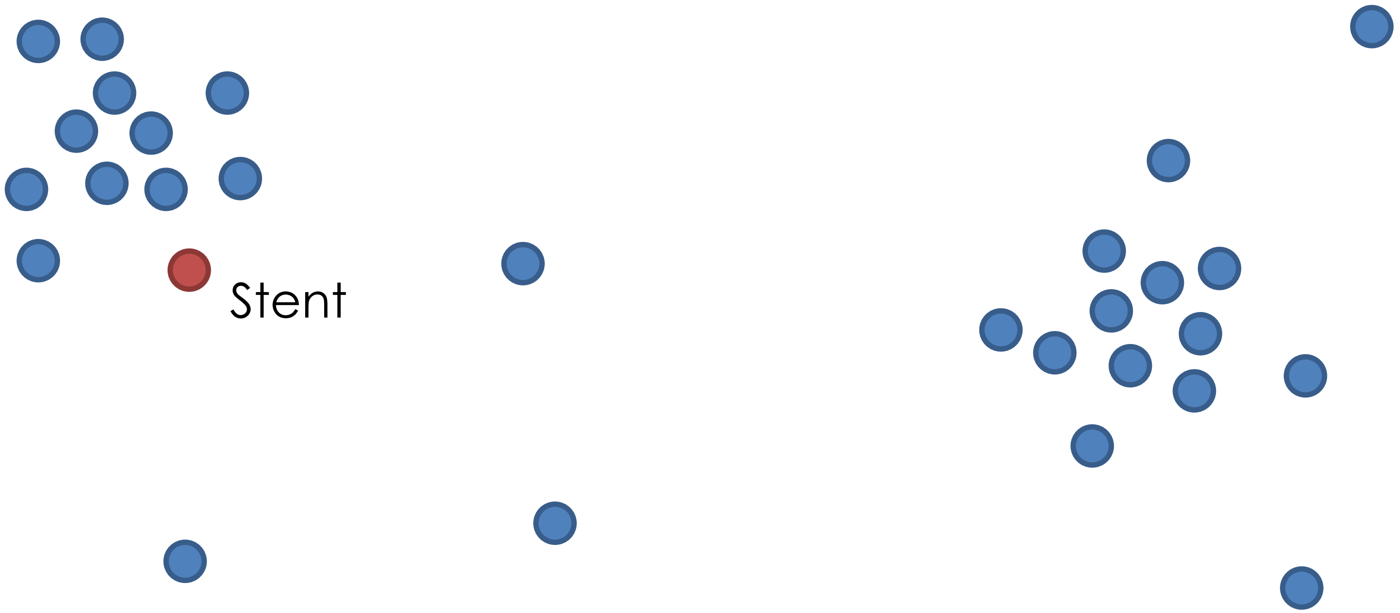
Similar Words Cluster Together



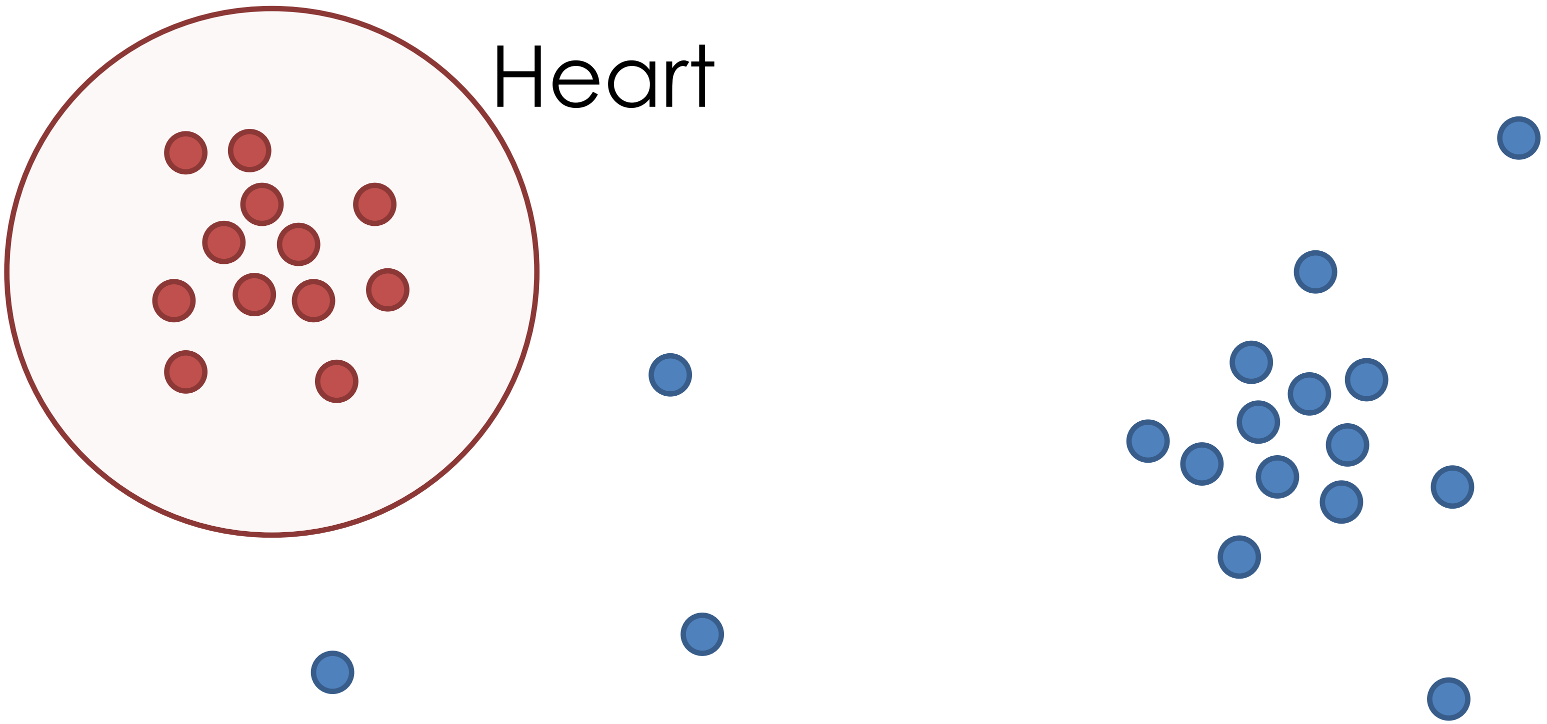
Similar Words Cluster Together



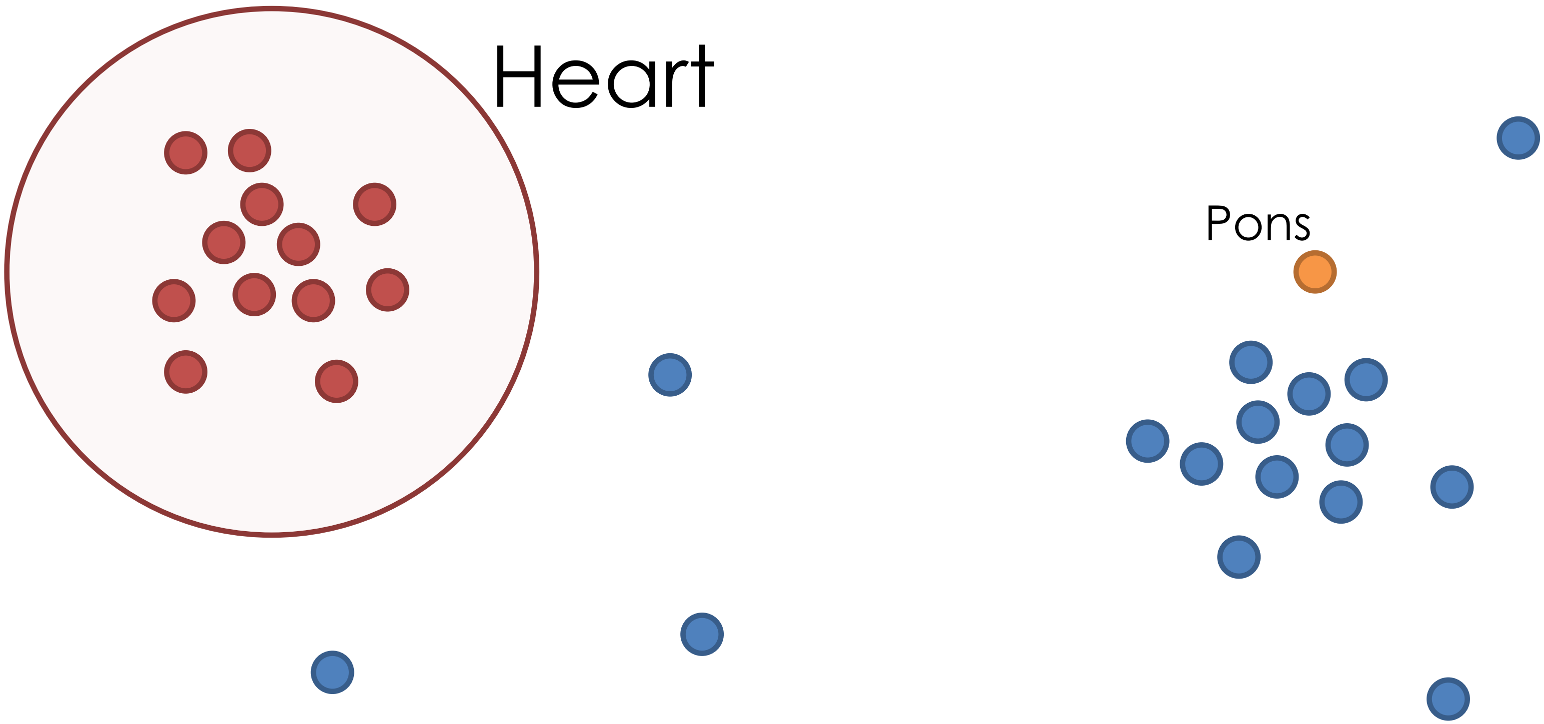
Similar Words Cluster Together



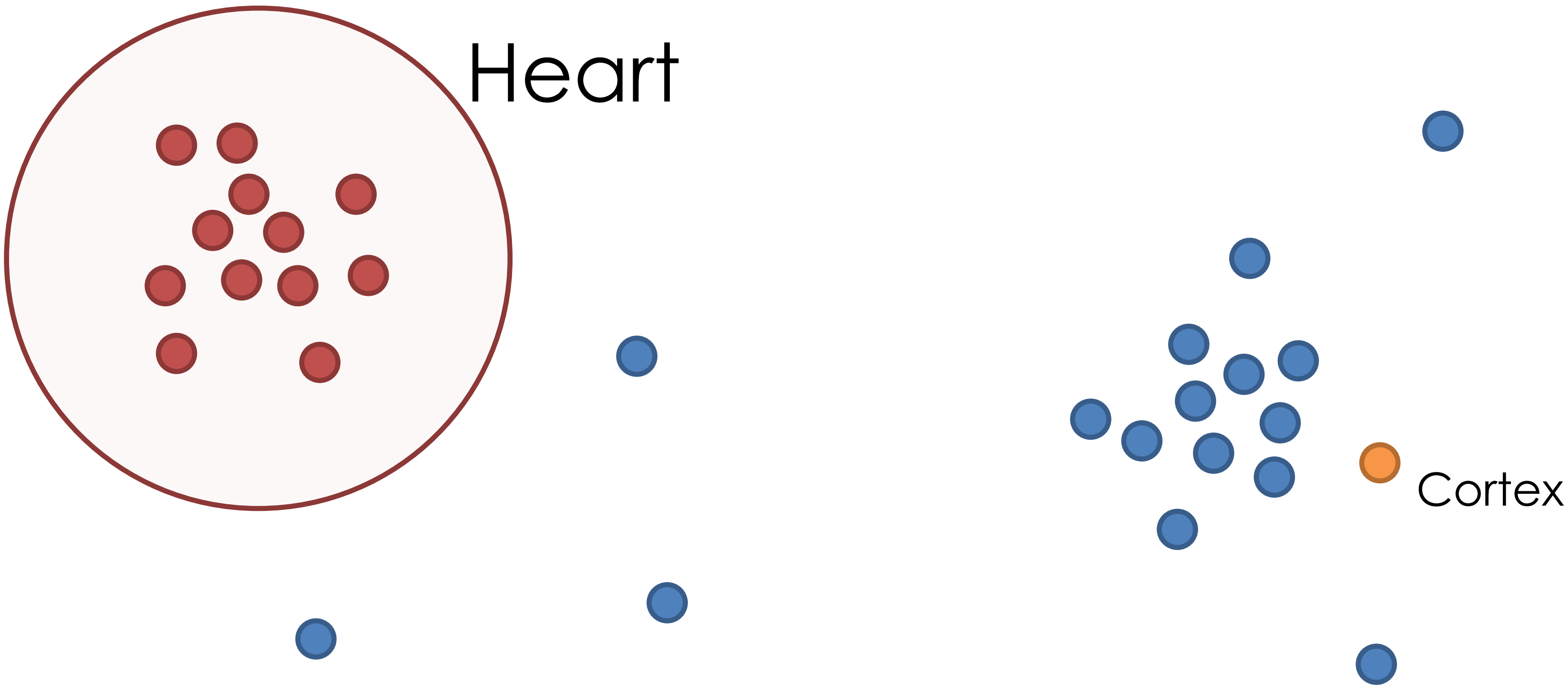
Clusters Form Meaningful Groups



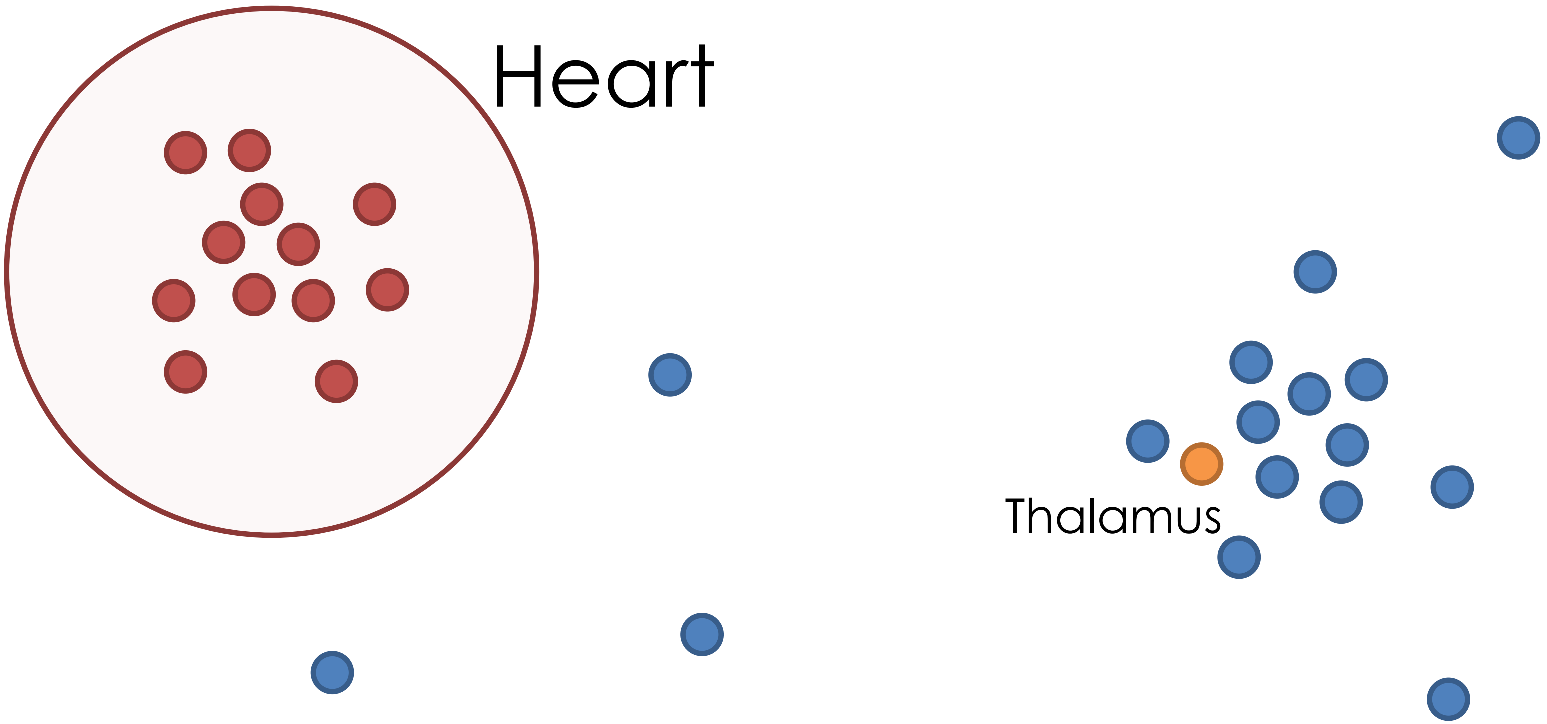
Clusters Form Meaningful Groups



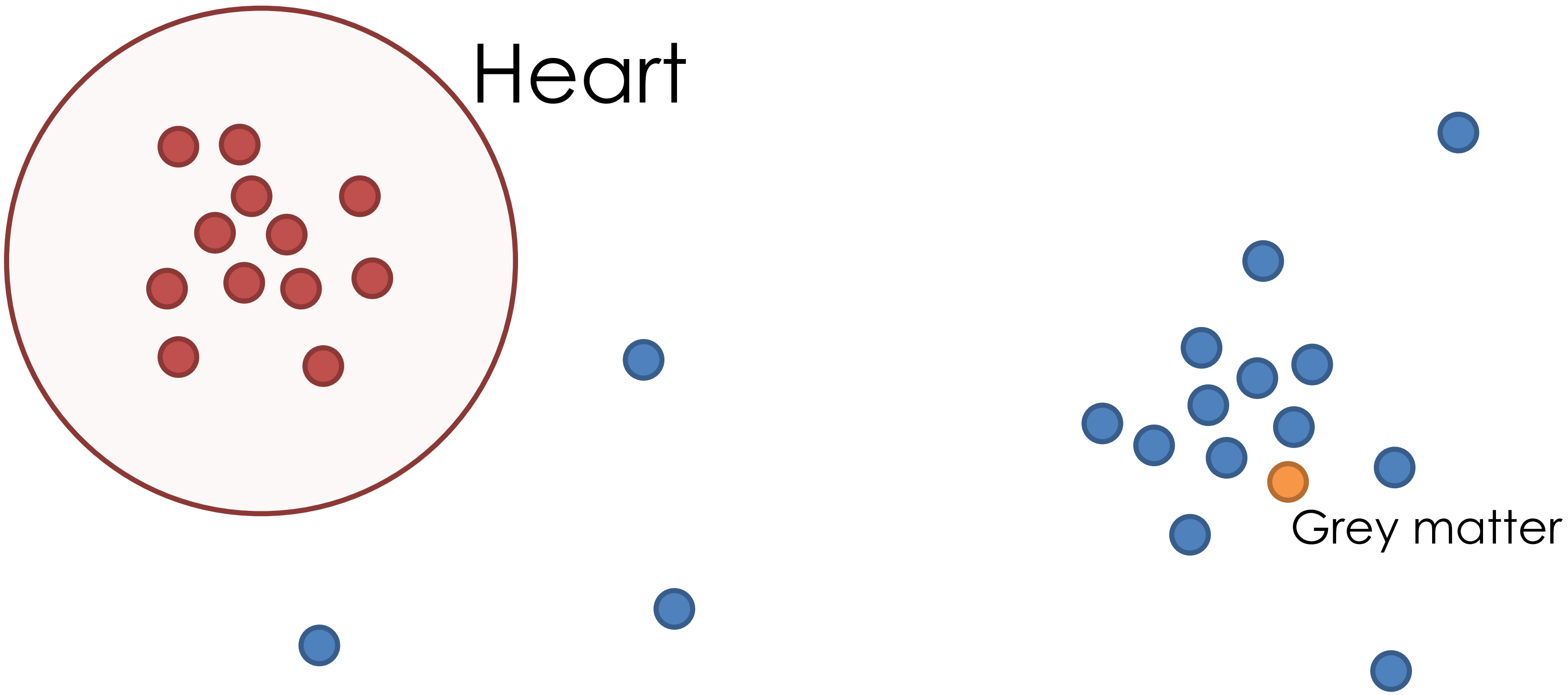
Clusters Form Meaningful Groups



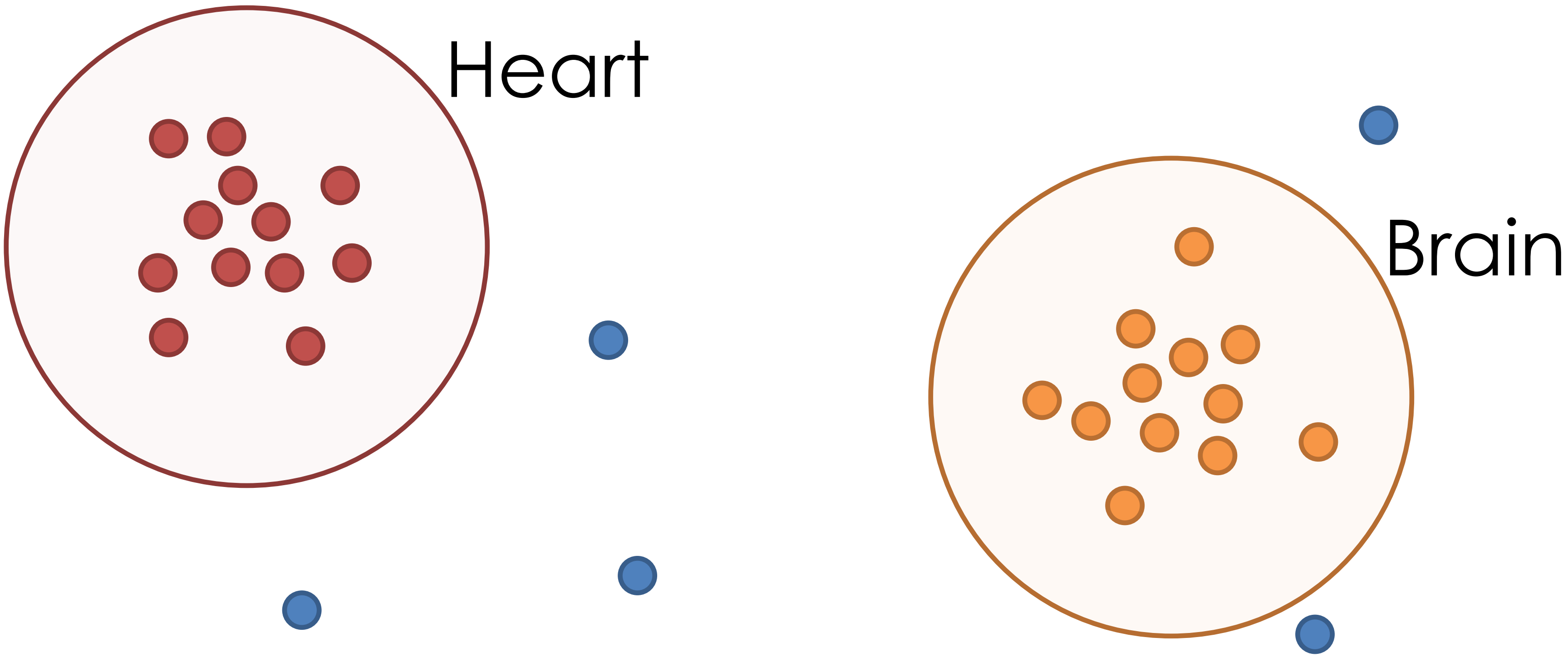
Clusters Form Meaningful Groups



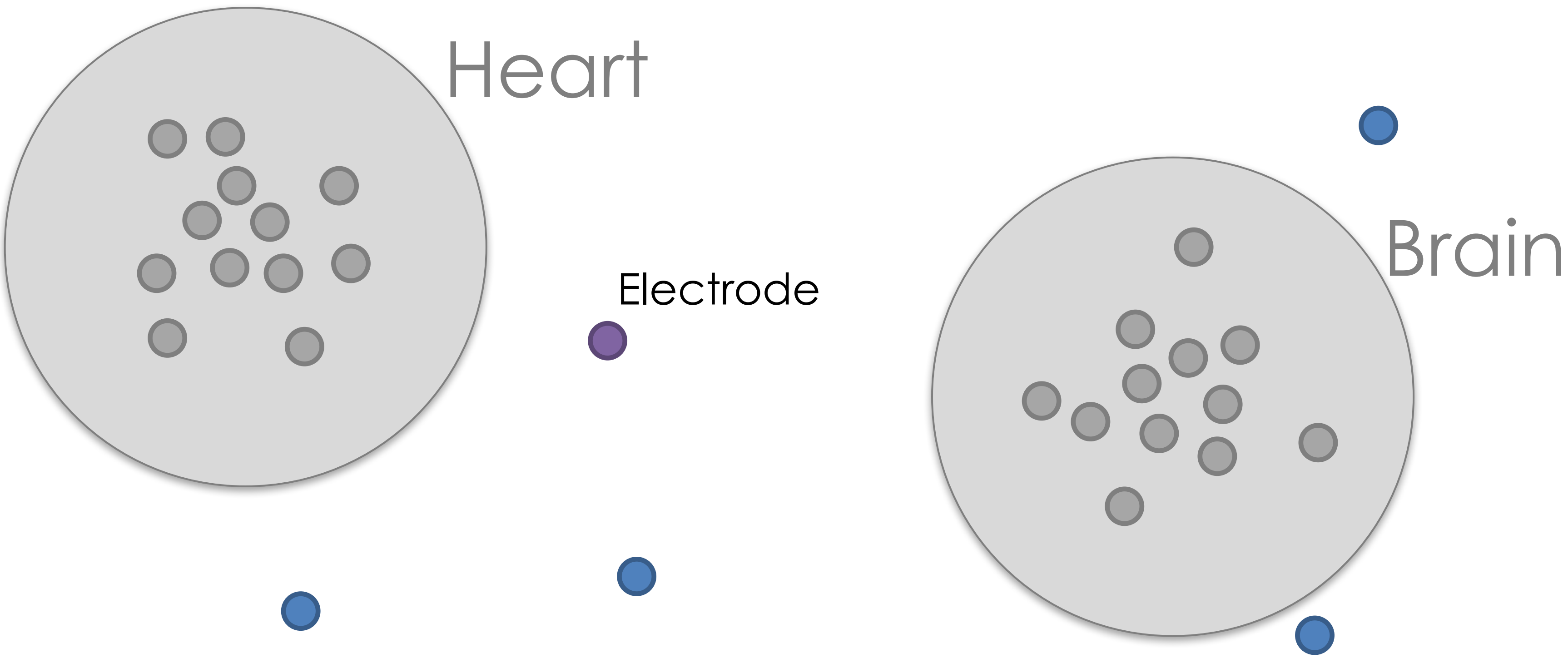
Clusters Form Meaningful Groups



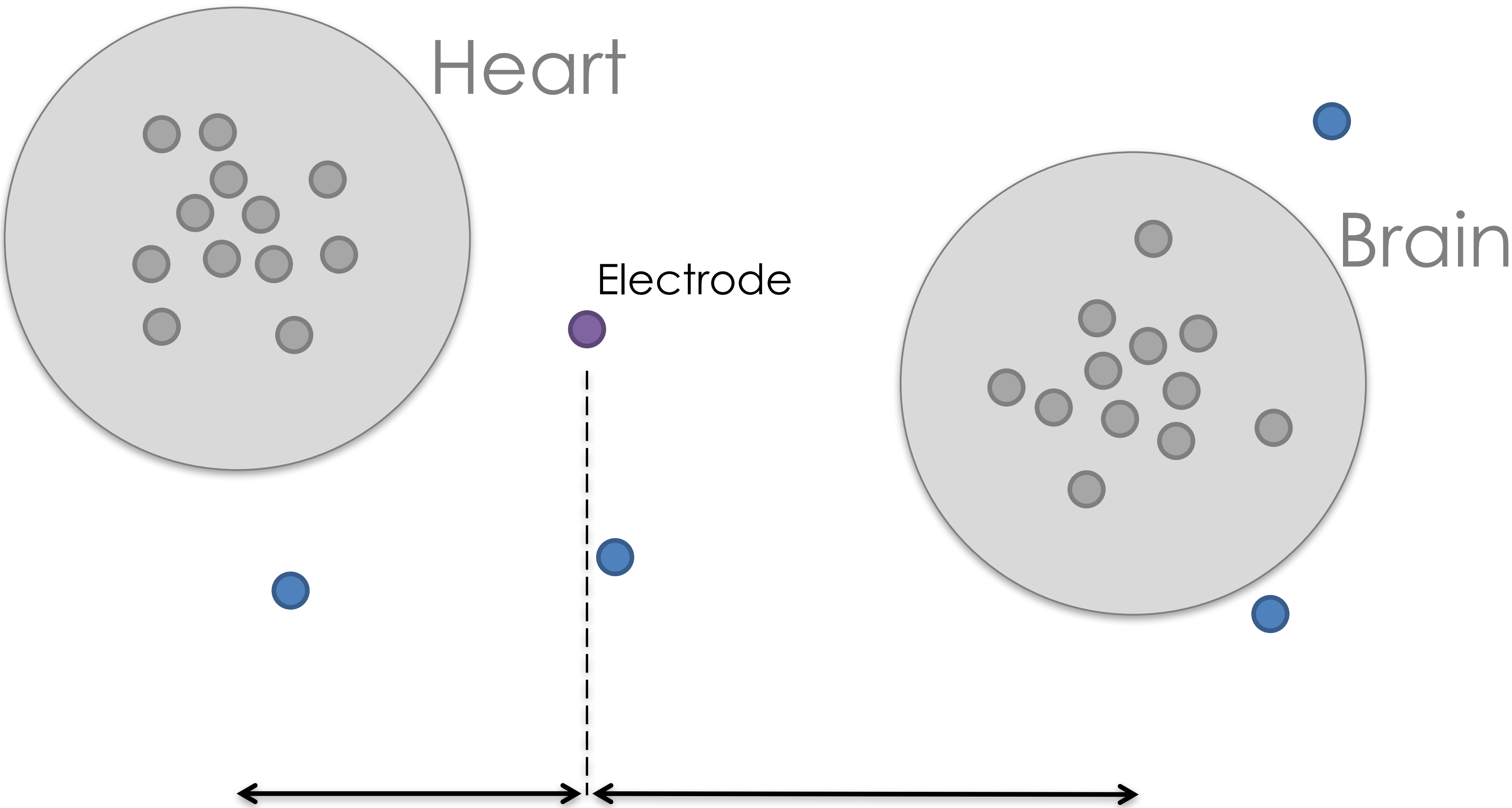
Clusters Form Meaningful Groups



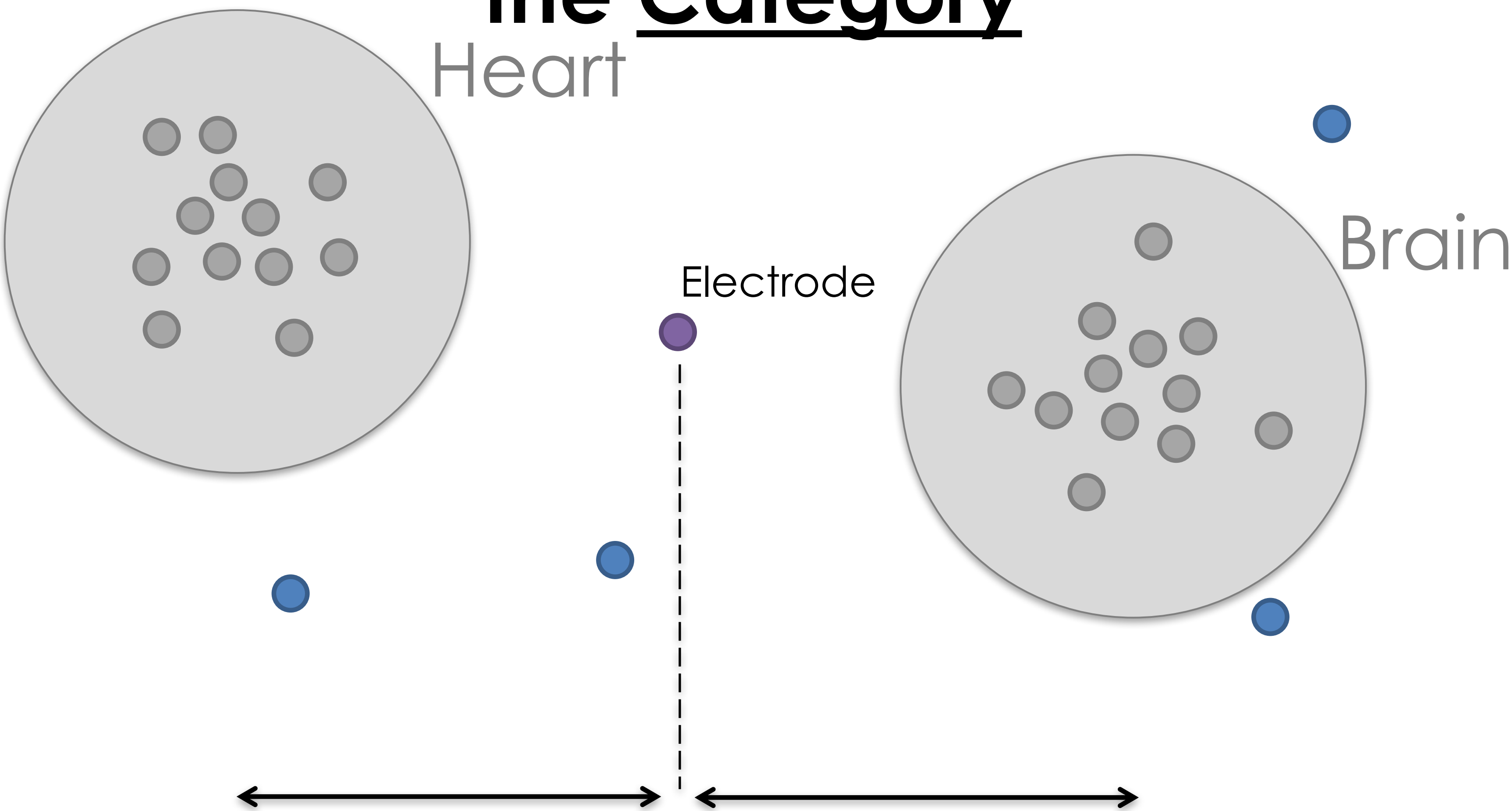
Ungrouped Words



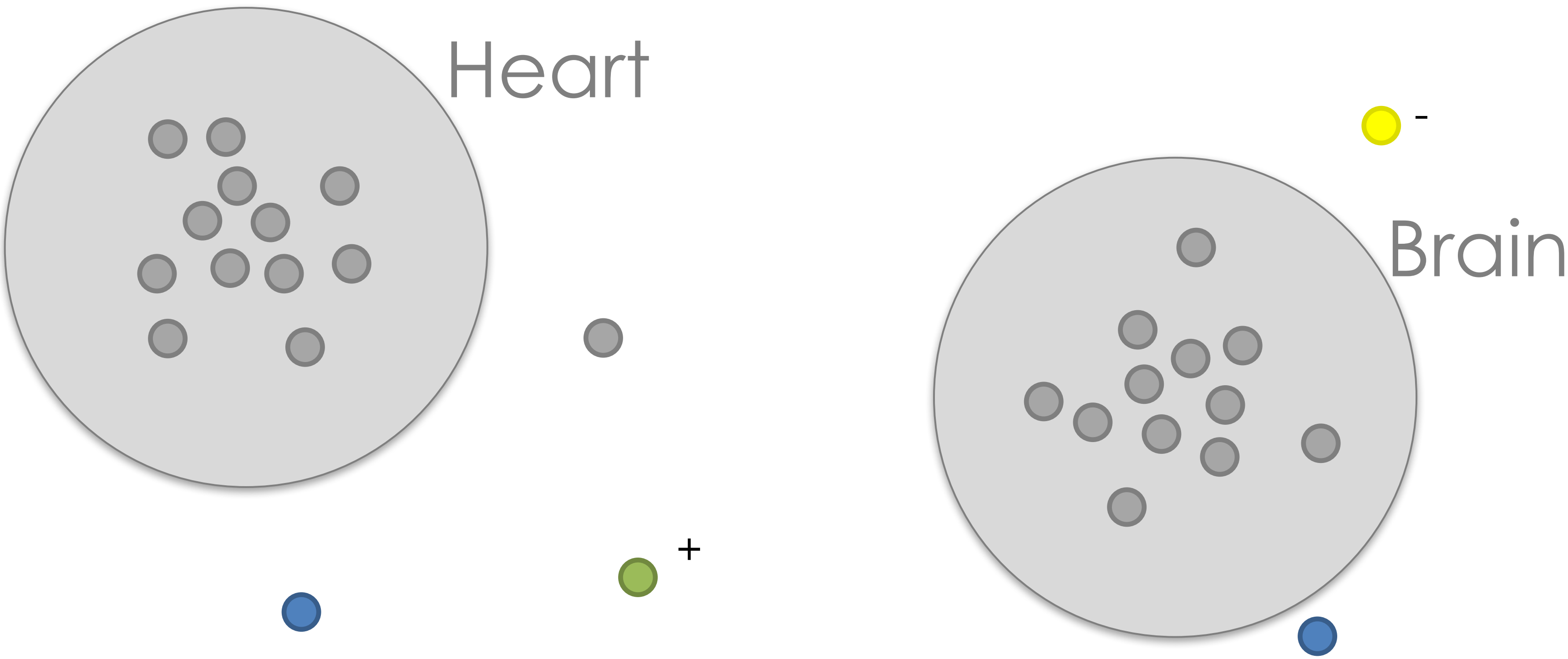
‘Electrode’ is Related to both Groups



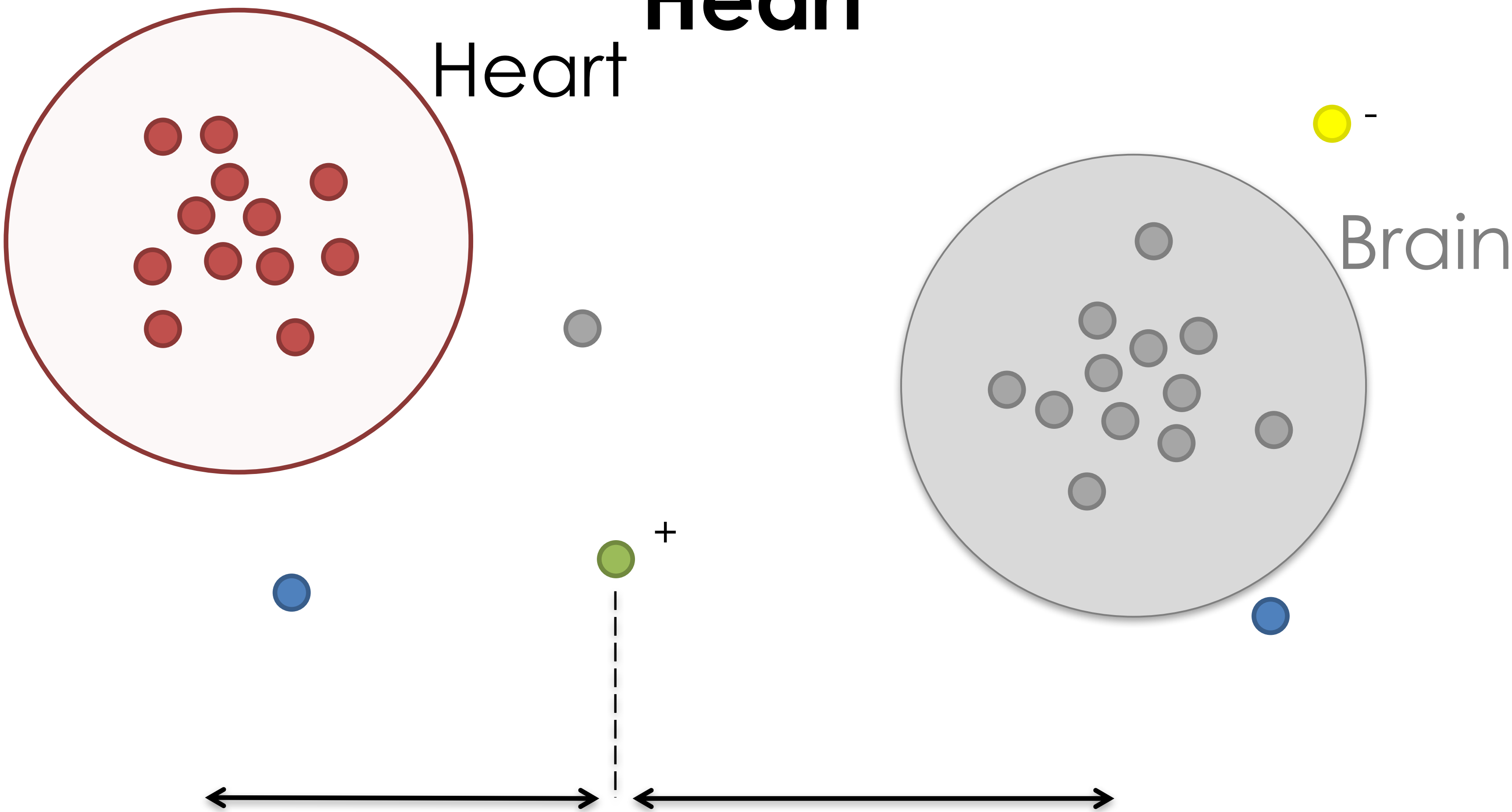
Centrality Indicates Importance for the <u>Category</u>



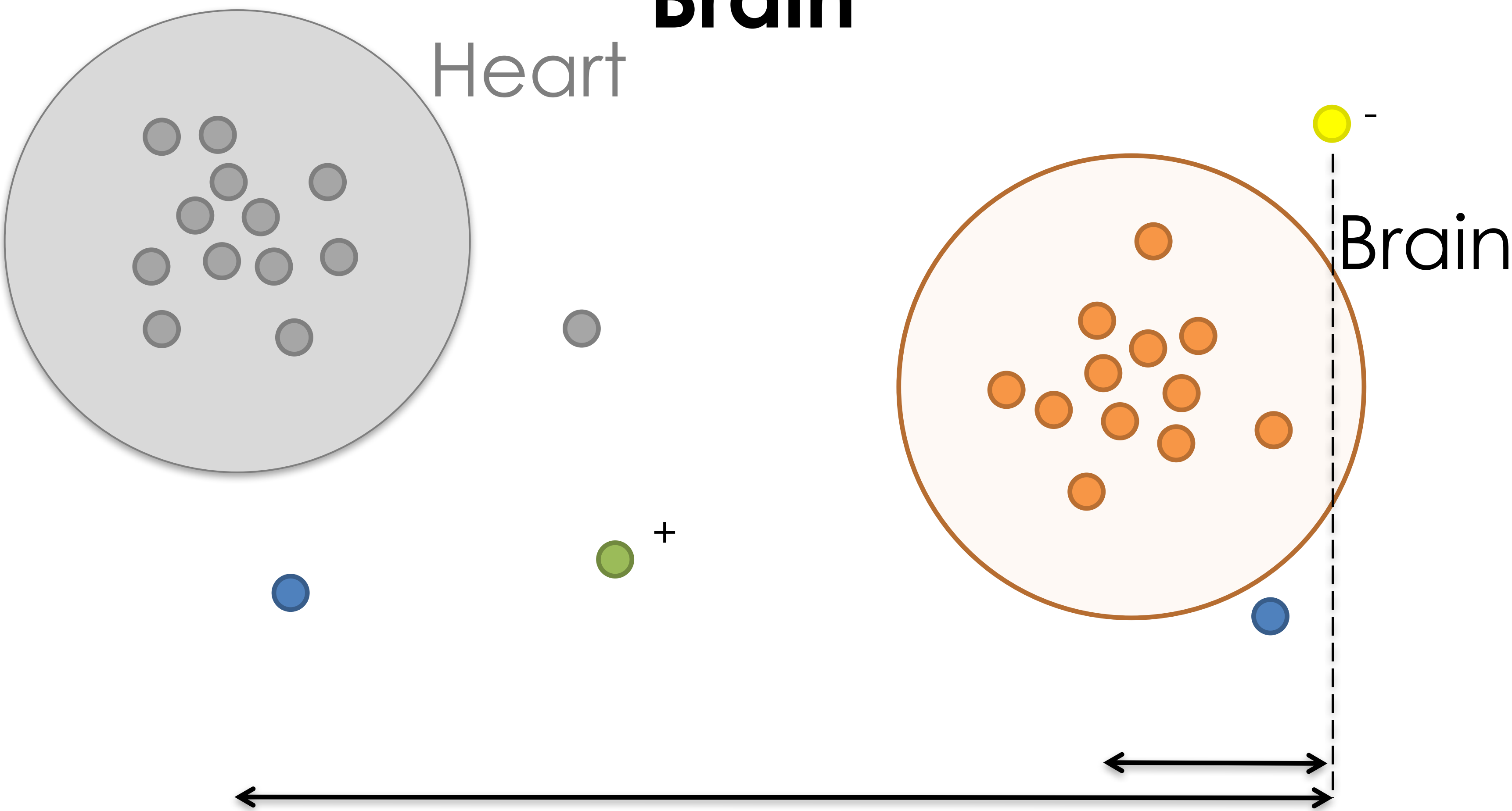
What about the Sentiment Terms?



Positive: Slightly More Related to the Heart



Negative: Strongly Related to the Brain



Conclusions Drawn from Illustrative Example

- There is greater positive sentiment for the <u>patient category</u> than negative sentiment

- There is greater negative sentiment for the brain <u>group</u> than the heart <u>group</u>

Sentiment Score

$$s_s = (\frac{s_p}{s_n} - 1) * 100$$

- Where Sp and Sn are the average cosine similarity between the 'positive' and 'negative' terms, and all other terms in the space

$$\cos(x, y) = \frac{x \cdot y}{\| x \| \| y \|}$$

- We can compare this against another metric that uses word counts

$$100 * (n_{positive}/n_{negative} - 1)$$

Complexity Score

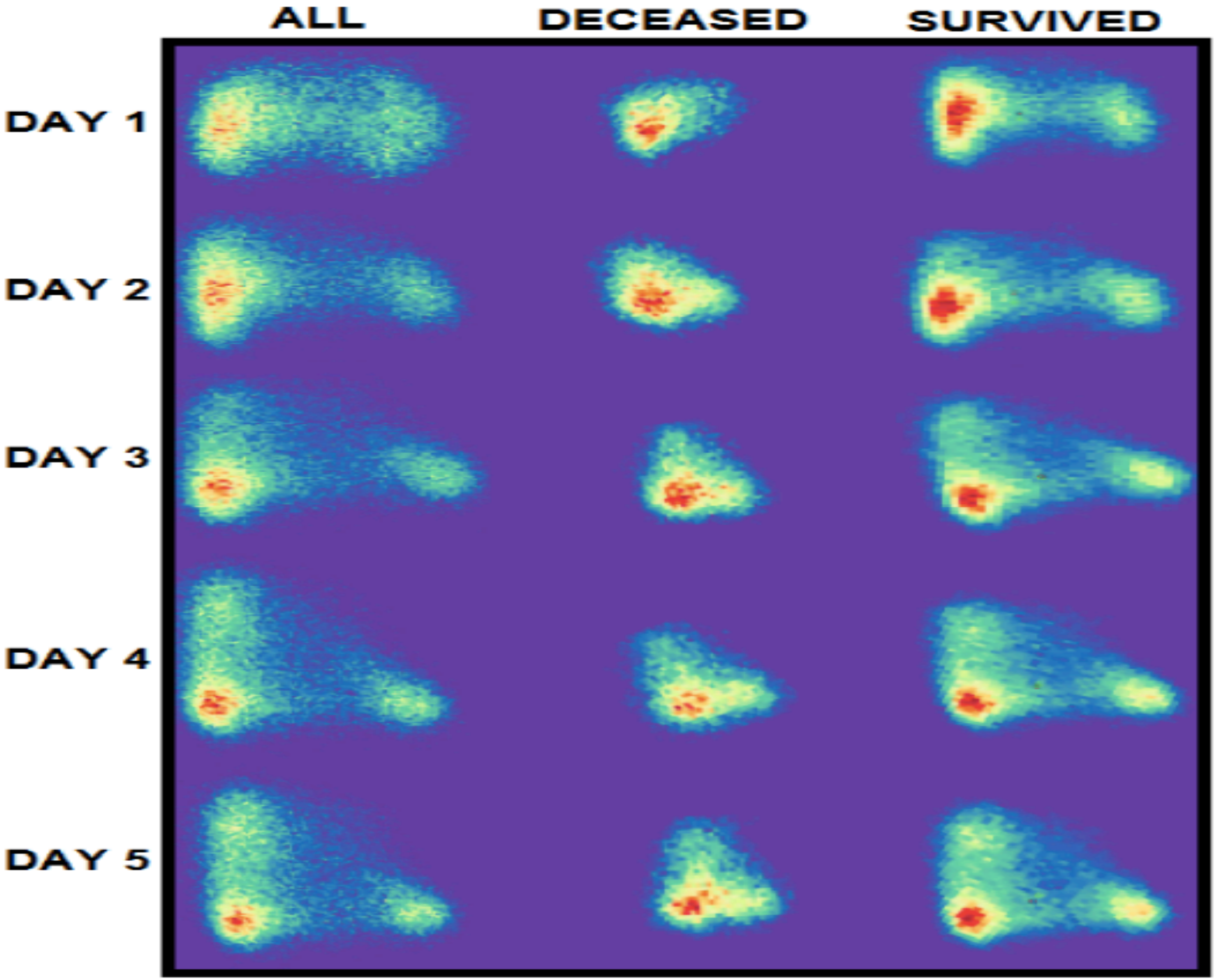
- Track the evolution of an optimal k, in the k-means algorithm
- Where optimality is determined by the Silhouette value

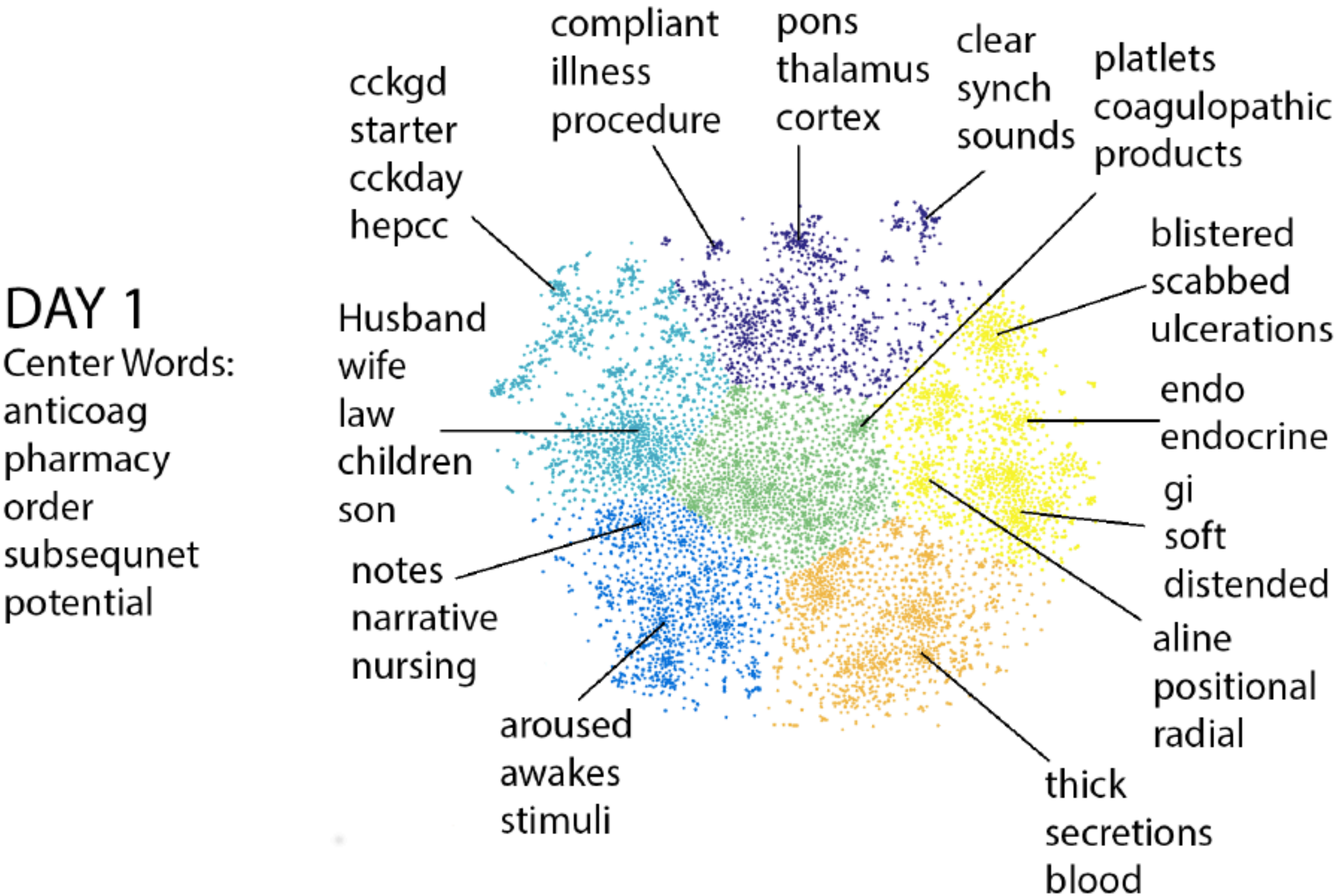
Visualizations

- Distributed Stochastic Neighbor Embedding (tSNE)
  - visualize distinctive word clusters
  - the evolution of language structure
- Principal Component Analysis (PCA)
  - characterize evolution over time

- Findings
- Conclusions

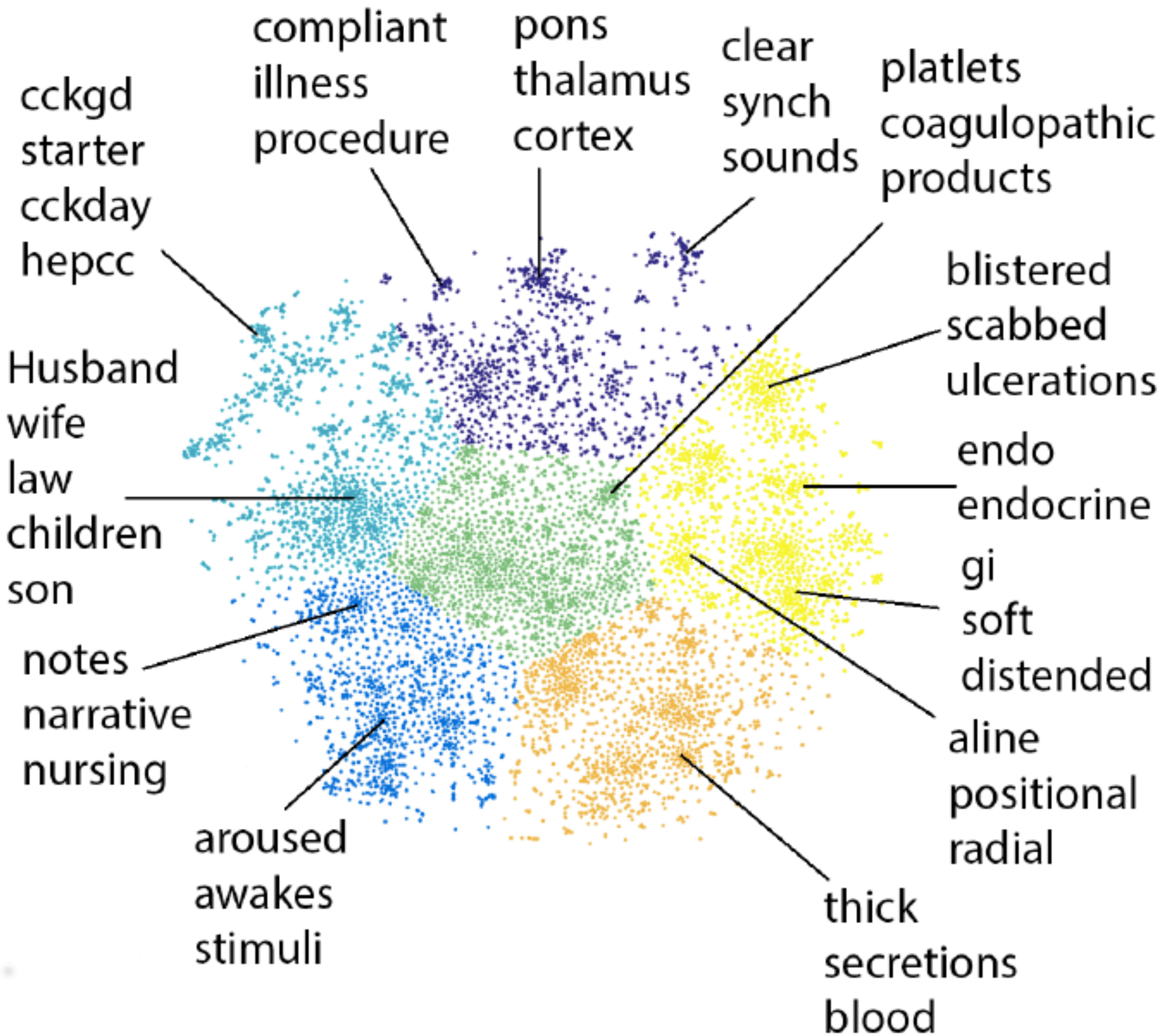
STUDY RESULTS

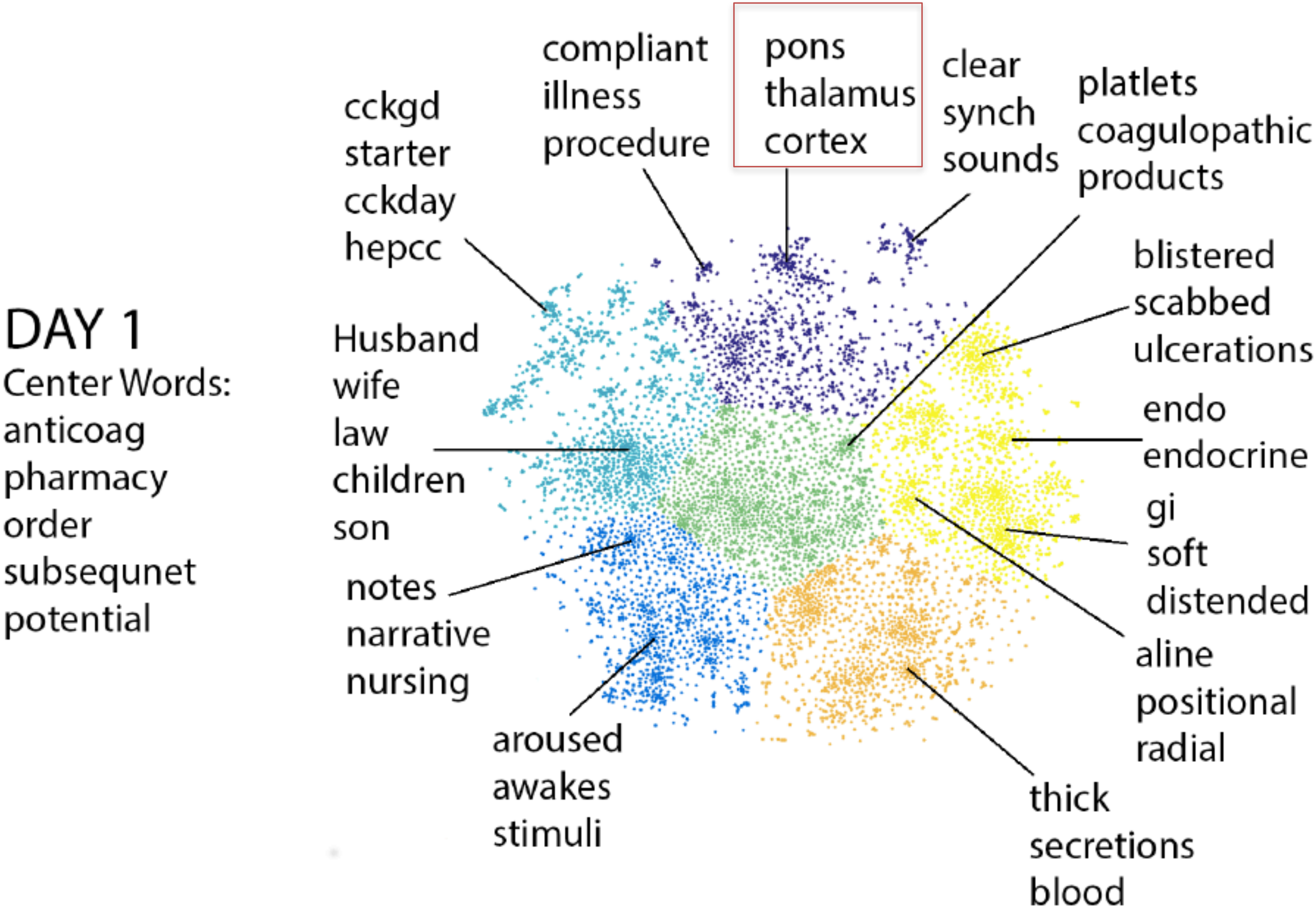

ALL
DECEASED
SURVIVED
DAY 1
DAY 2
DAY 3
DAY 4
DAY 5

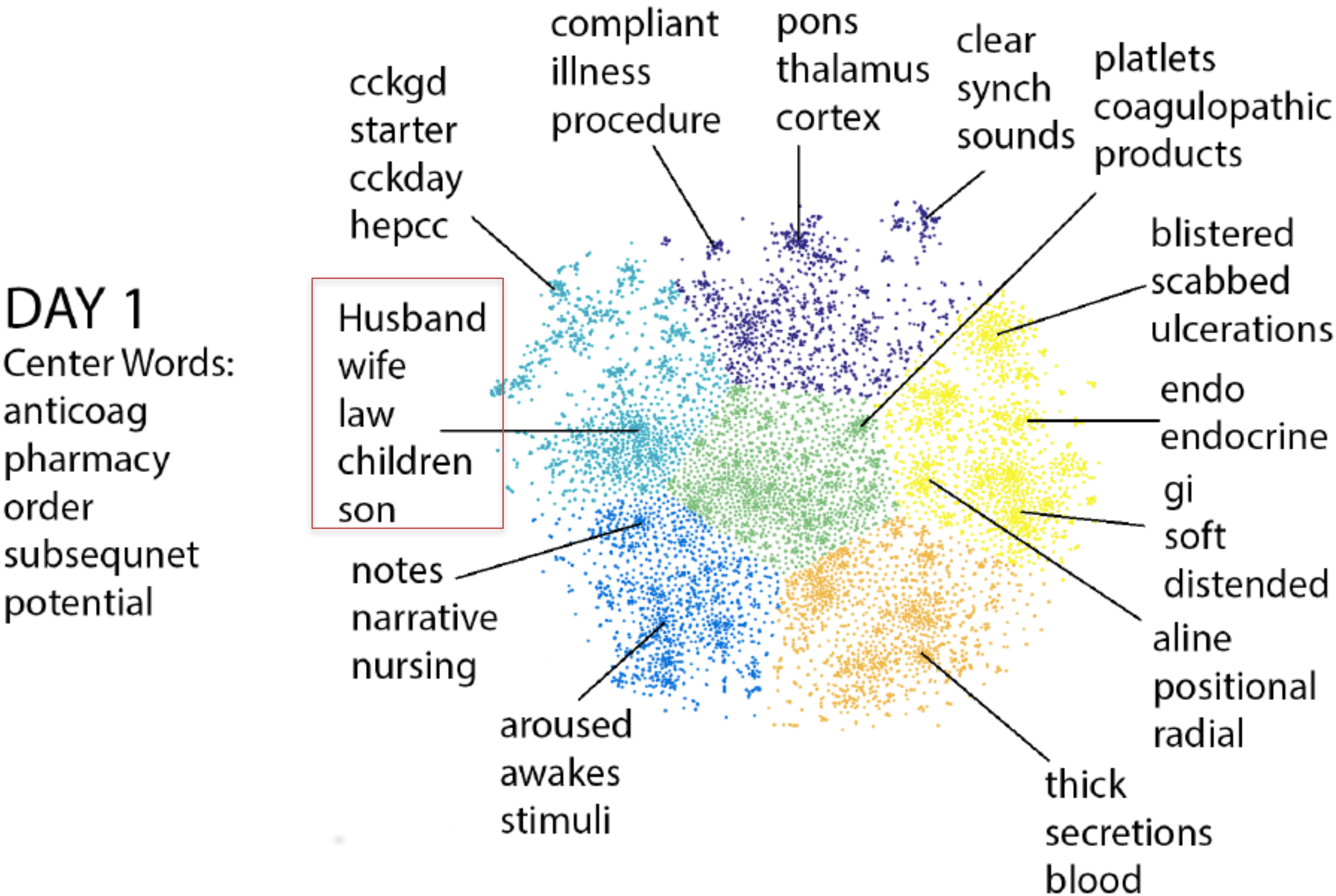

DAY 1
Center Words:
anticoag
pharmacy
order
subsequnet
potential
cckgd
starter
cckday
hepcc
compliant
illness
procedure
pons
thalamus
cortex
clear
synch
sounds
platlets
coagulopathic
products
blistered
scabbed
ulcerations
endo
endocrine
gi
soft
distended
aline
positional
radial
thick
secretions
blood
aroused
awakes
stimuli
notes
narrative
nursing
Husband
wife
law
children
son

DAY 1

Center Words:
anticoag
pharmacy
order
subsequnet
potential




DAY 1
Center Words:
anticoag
pharmacy
order
subsequnet
potential
cckgd
starter
cckday
hepcc
compliant
illness
procedure
pons
thalamus
cortex
clear
synch
sounds
platlets
coagulopathic
products
blistered
scabbed
ulcerations
Husband
wife
law
children
son
endo
endocrine
gi
soft
distended
notes
narrative
nursing
aline
positional
radial
aroused
awakes
stimuli
thick
secretions
blood



DAY 1
Center Words:
anticoag
pharmacy
order
subsequnet
potential
cckgd
starter
cckday
hepcc
compliant
illness
procedure
pons
thalamus
cortex
clear
synch
sounds
platlets
coagulopathic
products
blistered
scabbed
ulcerations
Husband
wife
law
children
son
endo
endocrine
gi
soft
distended
notes
narrative
nursing
aline
positional
radial
aroused
awakes
stimuli
thick
secretions
blood

Day 3

Center Words:

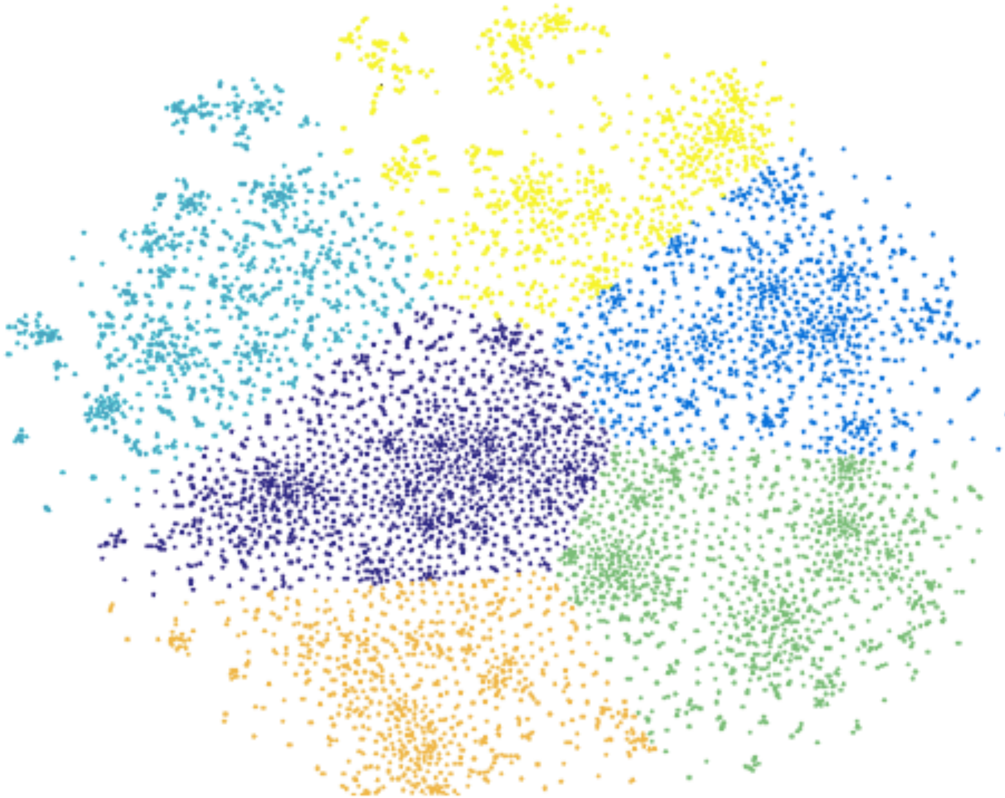
exertion

tolerated

alarms

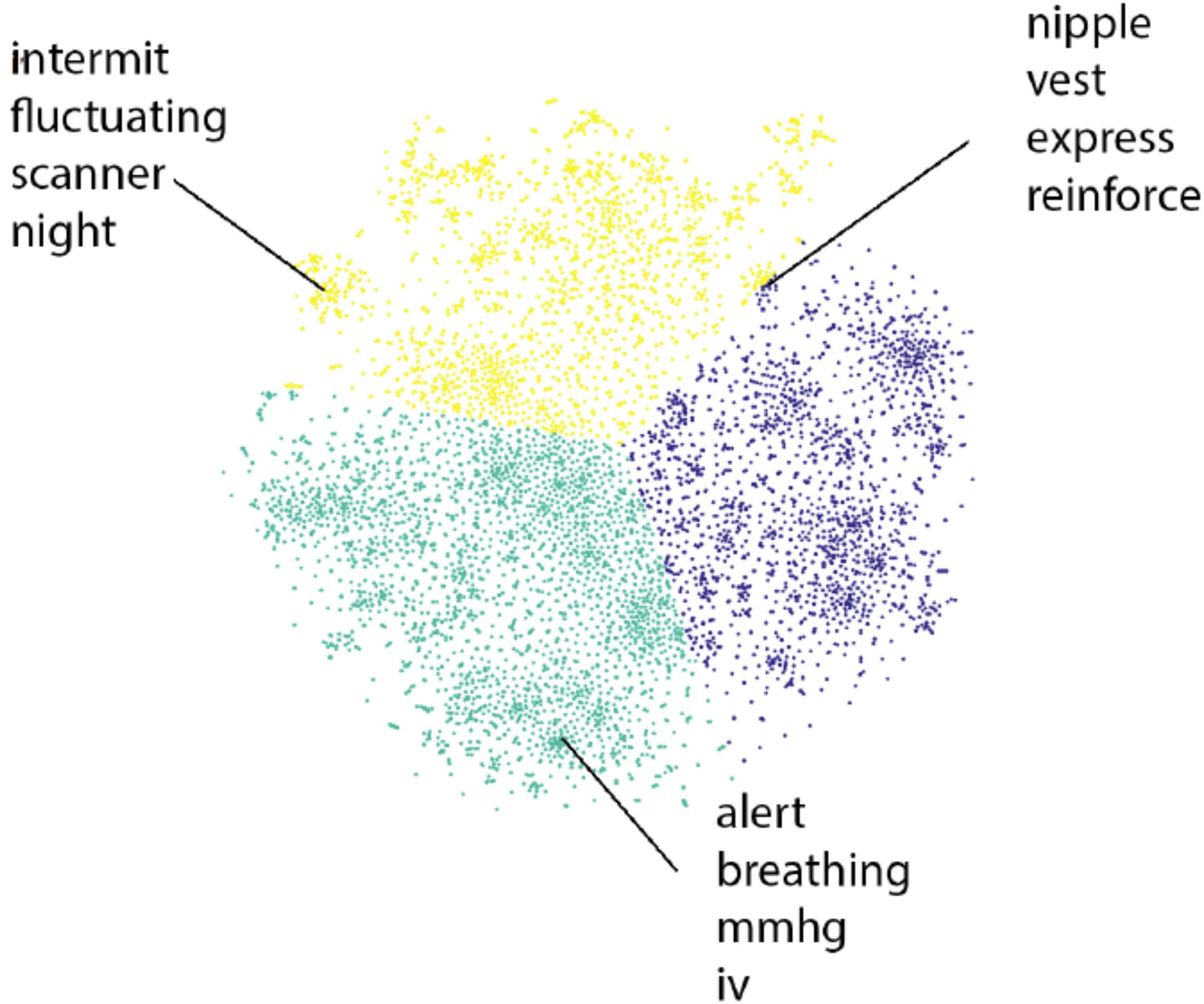
specimines

penicilins

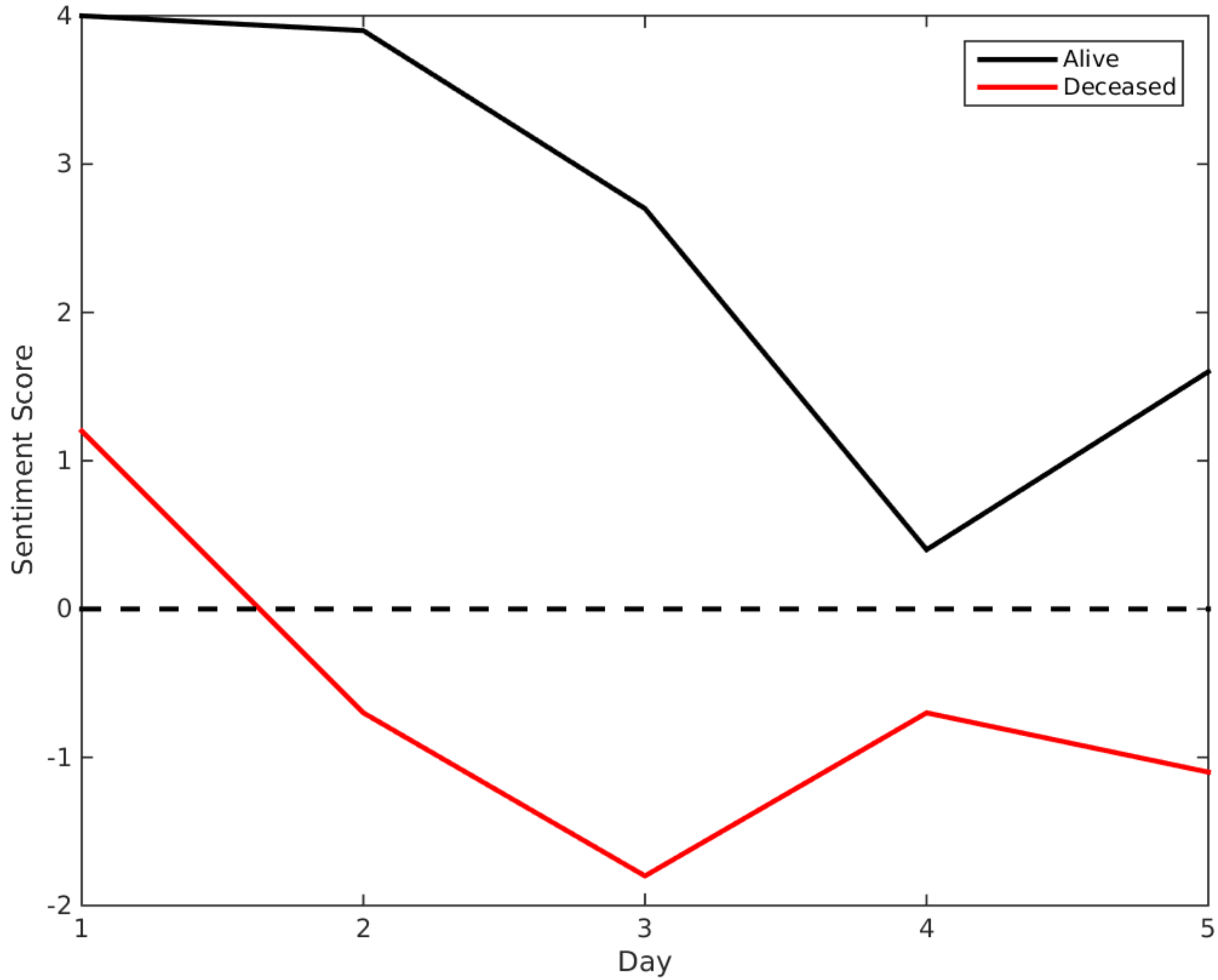


Day 5

Center Words:
prolonged
cardiac
hemorrhage
photo



Sentiment Score also Evolves over Time and Outcome



Sentiment over Categories

- Sentiment differences across categories
- The vector-based score is more aligned with expected results than the score that uses ratio of word counts (RWC)

Sentiment Score	RWC	Note Category
-1.98	-6.84	Deceased
0.57	63.3	Survived
0.63	72.16	Age <25
-0.31	-2.97	Age: 25 - 49
-1.36	-0.42	Age: 50 - 75
-1.82	1.65	Age >75
-0.28	2.16	Married
-0.08	5.25	Single
0.90	54.54	Female
0.39	47.59	Male
-1.07	115.51	Asian
0.14	41.06	White
0.45	62.99	African

Conclusions

- Two main findings:
  - the sentiment of clinical notes evolve over time, patient condition, and patient background
  - The structure / complexity of clinical notes also evolves

- Results arepreliminary, and will require further investigation to reach firm conclusions

Thank you

- Contact me for questions, or to collaborate!

  ghassemi@mit.edu

Example: Similar Patients, Different Treatment Decisions

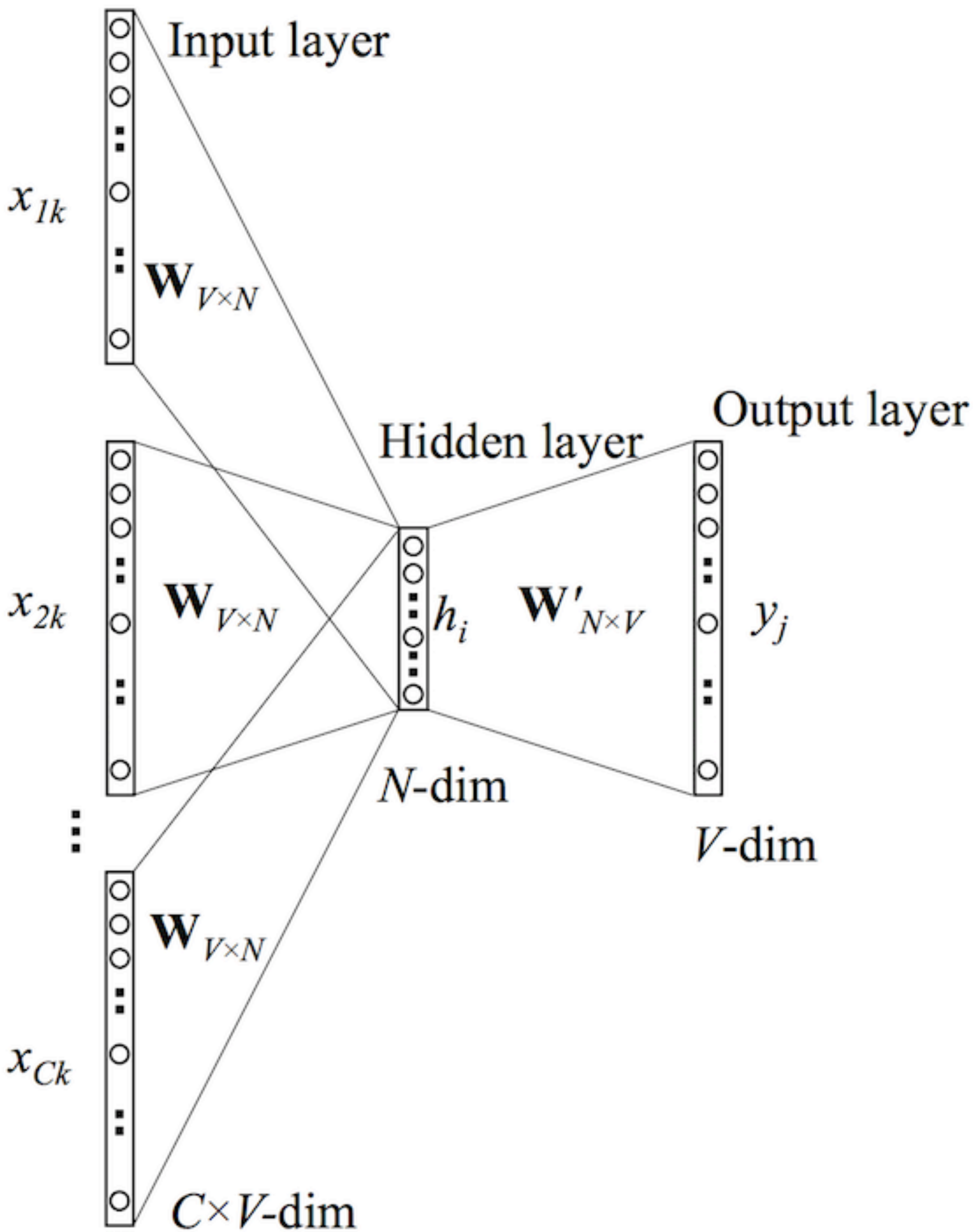
	Patient 1	Patient 2
Structured Data	Age = 35	Age = 35
Treatment Decision	Intubate	Don't Intubate

*Source: Multiparameter Intelligent Monitoring in Critical Care Database

Example: Similar Patients, Different Treatment Decisions

	Patient 1	Patient 2
Structured Data	Age = 35	Age = 35
Treatment Decision	Intubate	Don’t Intubate
Unstructured Data	“blah blah blah”	“blah blah blah?”

*Source: Multiparameter Intelligent Monitoring in Critical Care Database



CBoW

- C – word window size
- V – vocabulary size
- Y – output word
- X – input context words

tf-idf (term frequency * inverse document frequency)

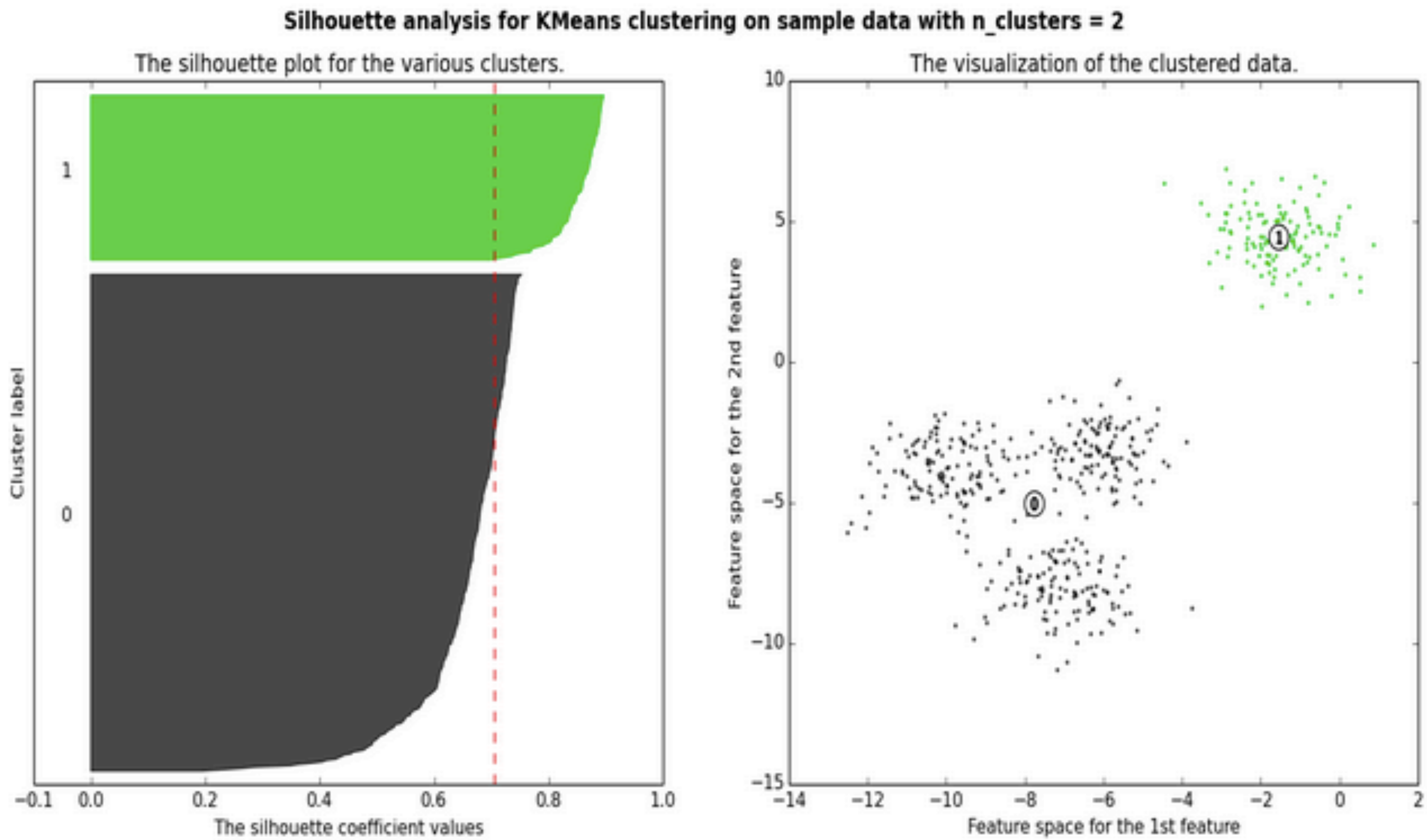
- Find “the red fox”
  - Remove all documents without the words
  - Count the number of times the words show up in each document (term frequency)
  - Because ‘the’ is common, we may over emphasize it and need to eliminate it’s effects
  - We diminish the weight of such terms using the inverse of their frequency in the set of documents (inverse document frequency).

Conclusions

- There is decreasing complexity of the language for patients who do not survive
  - This is not simply an artifact of the number of words

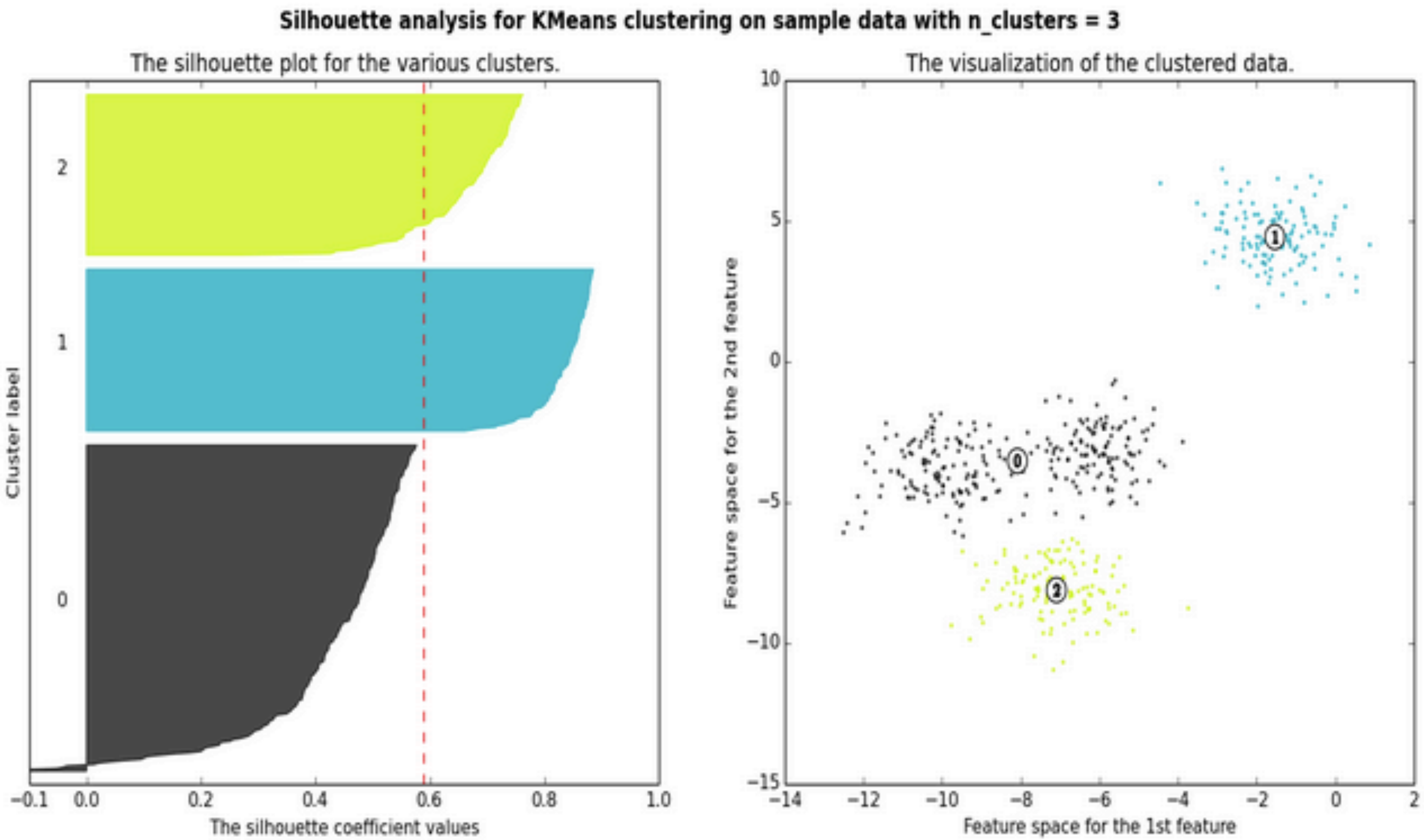
Word Count (Millions)	Note Category
4.44	Age <25
16.66	Age: 25 - 49
44.25	Age: 50 - 75
28.60	Age >75
21.36	Deceased
105.70	Survived
6.65	Day 1
11.32	Day 2
9.26	Day 3
8.69	Day 4
8.18	Day 5
45.08	Married
22.27	Single
32.69	Female
77.50	Male
3.95	Asian
92.87	White
13.14	African

Silhouttee

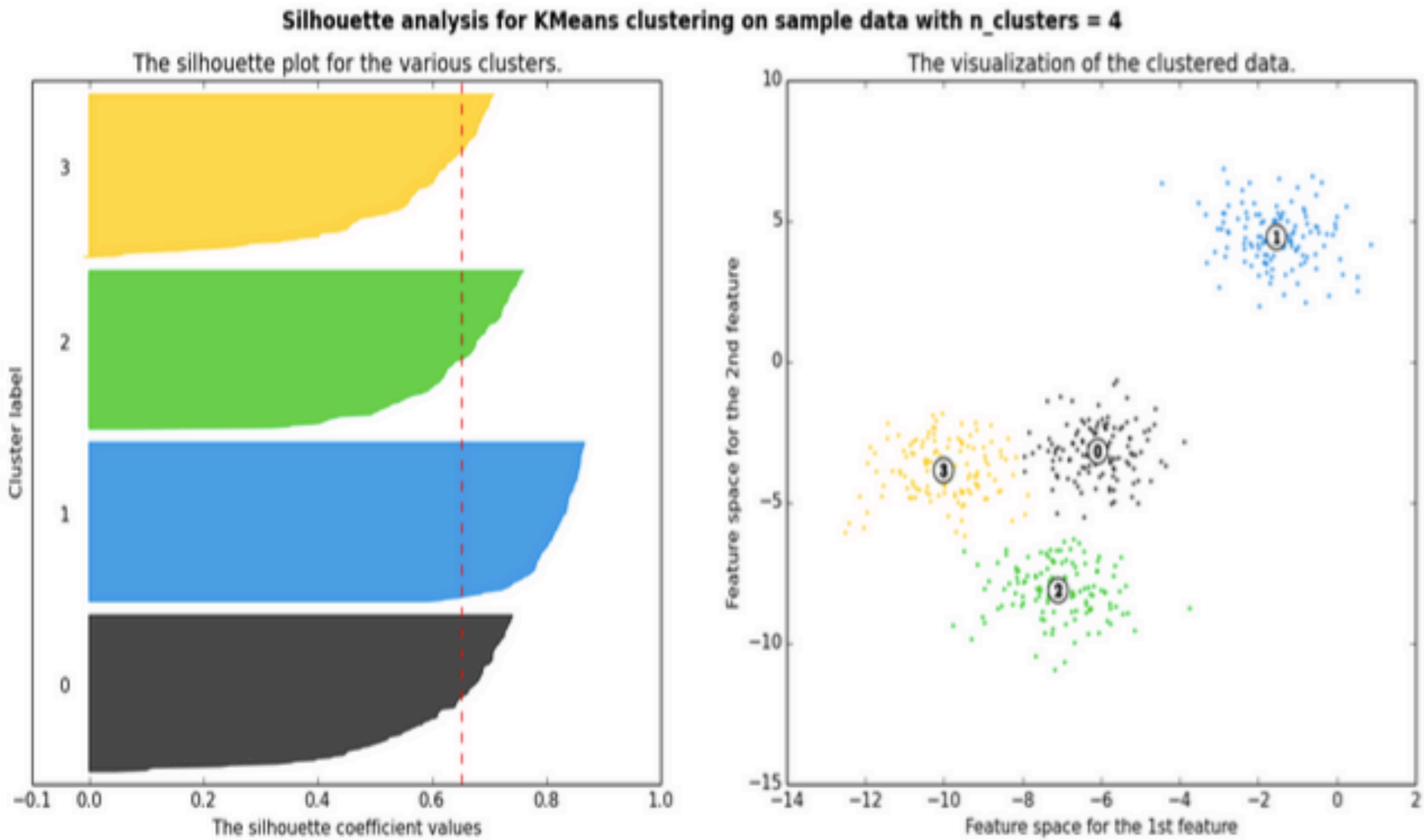


Silhouette analysis for KMeans clustering on sample data with n_clusters = 2
The silhouette plot for the various clusters.
Cluster label
1
0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
10
5
0
−5
−10
−15
−14
−12
−10
−8
−6
−4
−2
0
2
Feature space for the 1st feature

Silhouttee

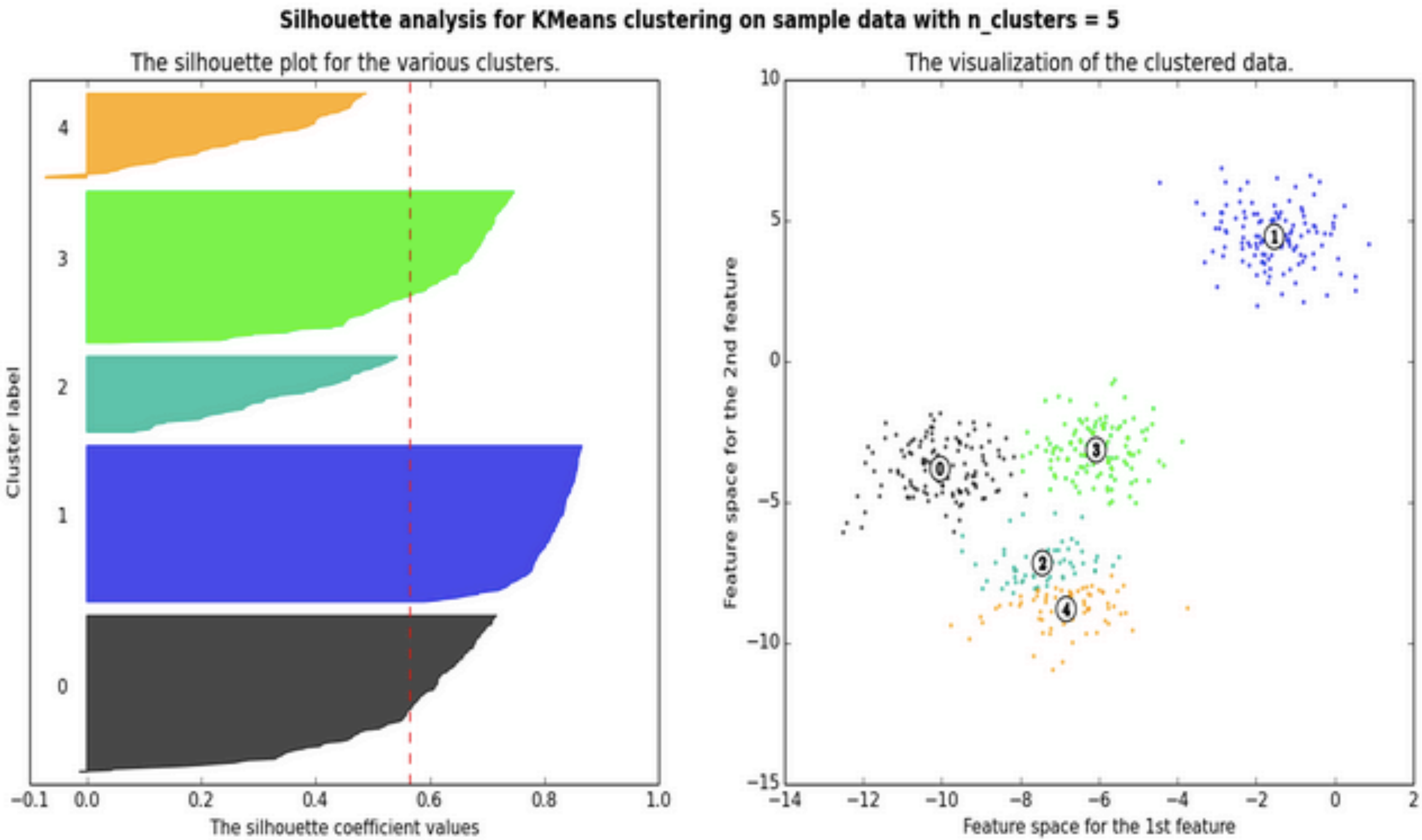

Silhouette analysis for KMeans clustering on sample data with n_clusters = 3
The silhouette plot for the various clusters.
Cluster label
2
1
0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
10
5
0
−5
−10
−15
−14
−12
−10
−8
−6
−4
−2
0
2
Feature space for the 1st feature

Silhouttee



Silhouette analysis for KMeans clustering on sample data with n_clusters = 4
The silhouette plot for the various clusters.
Cluster label
3
2
1
0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
10
5
0
−5
−10
−15
−14
−12
−10
−8
−6
−4
−2
0
2
Feature space for the 1st feature

Silhouttee


Silhouette analysis for KMeans clustering on sample data with n_clusters = 5
The silhouette plot for the various clusters.
Cluster label
The silhouette coefficient values
4
3
2
1
0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The visualization of the clustered data.
Feature space for the 2nd feature
Feature space for the 1st feature
10
5
0
−5
−10
−15
−14
−12
−10
−8
−6
−4
−2
0
2