Graph-based Concept Weighting for Medical Information Retrieval

Bevan Koopman

Guido Zuccon, Peter Bruza, Michael Lawley, Laurianne Sitbon

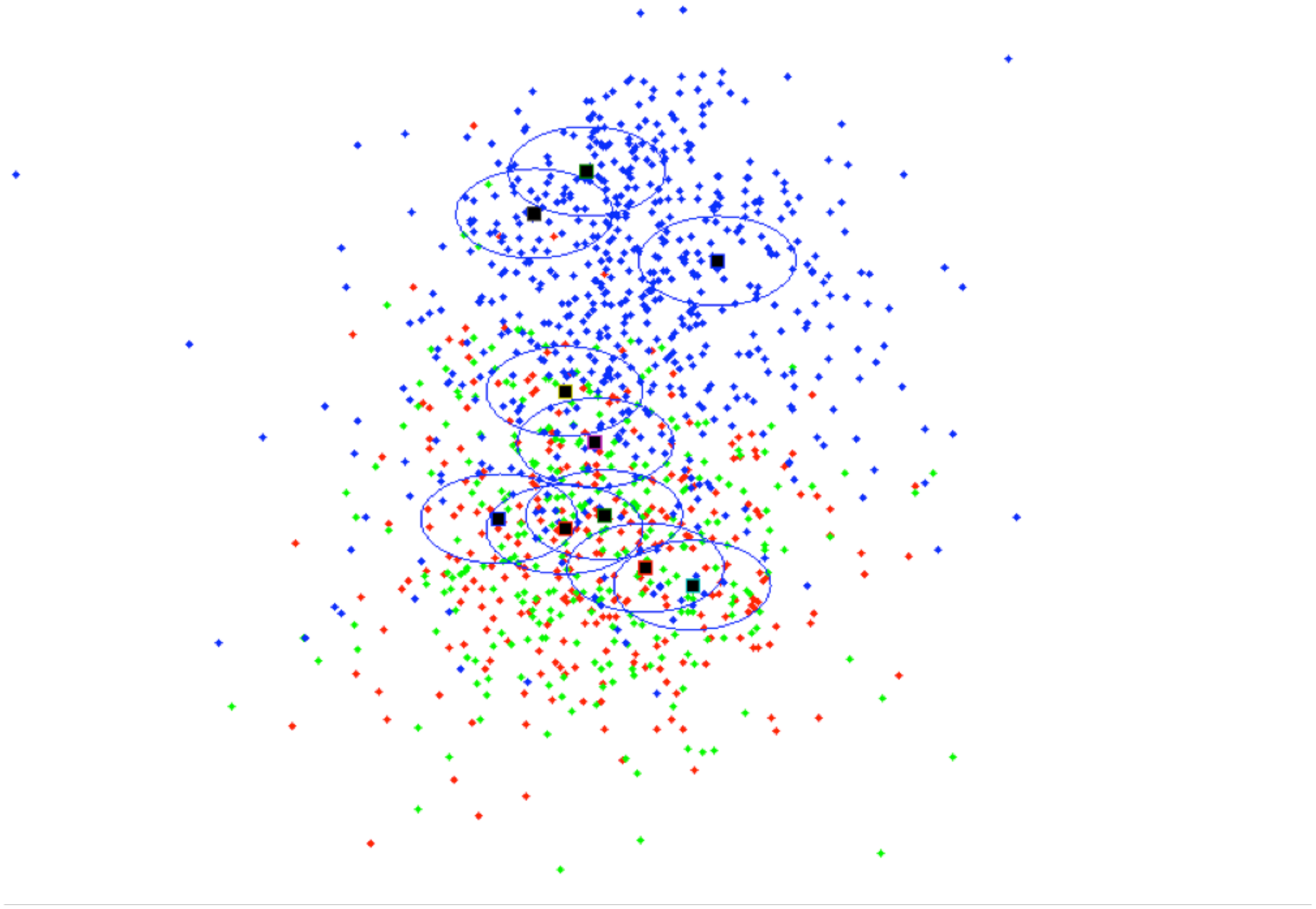
QUT

CSIRO

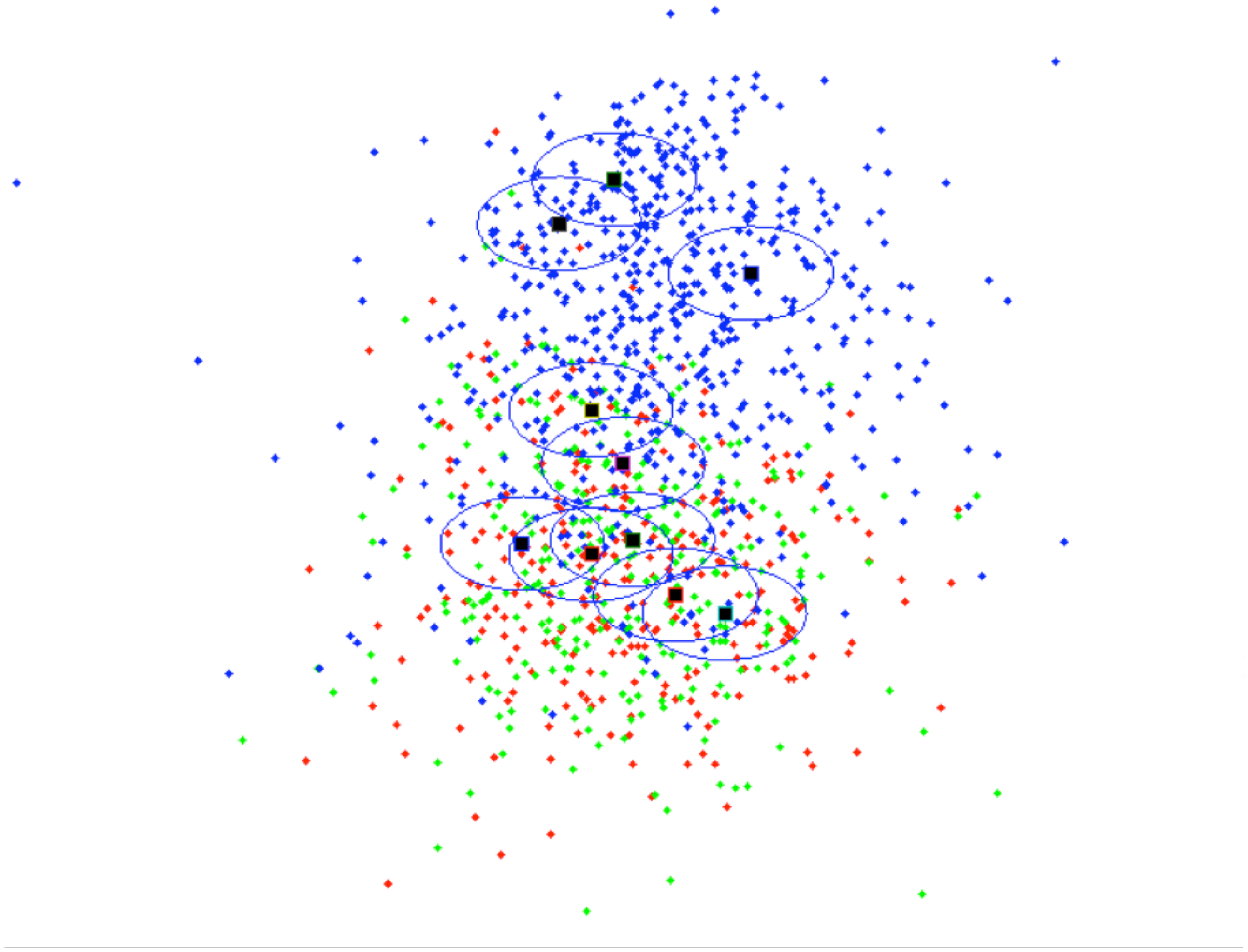
THE AUSTRALIAN E-HEALTH RESEARCH CENTRE

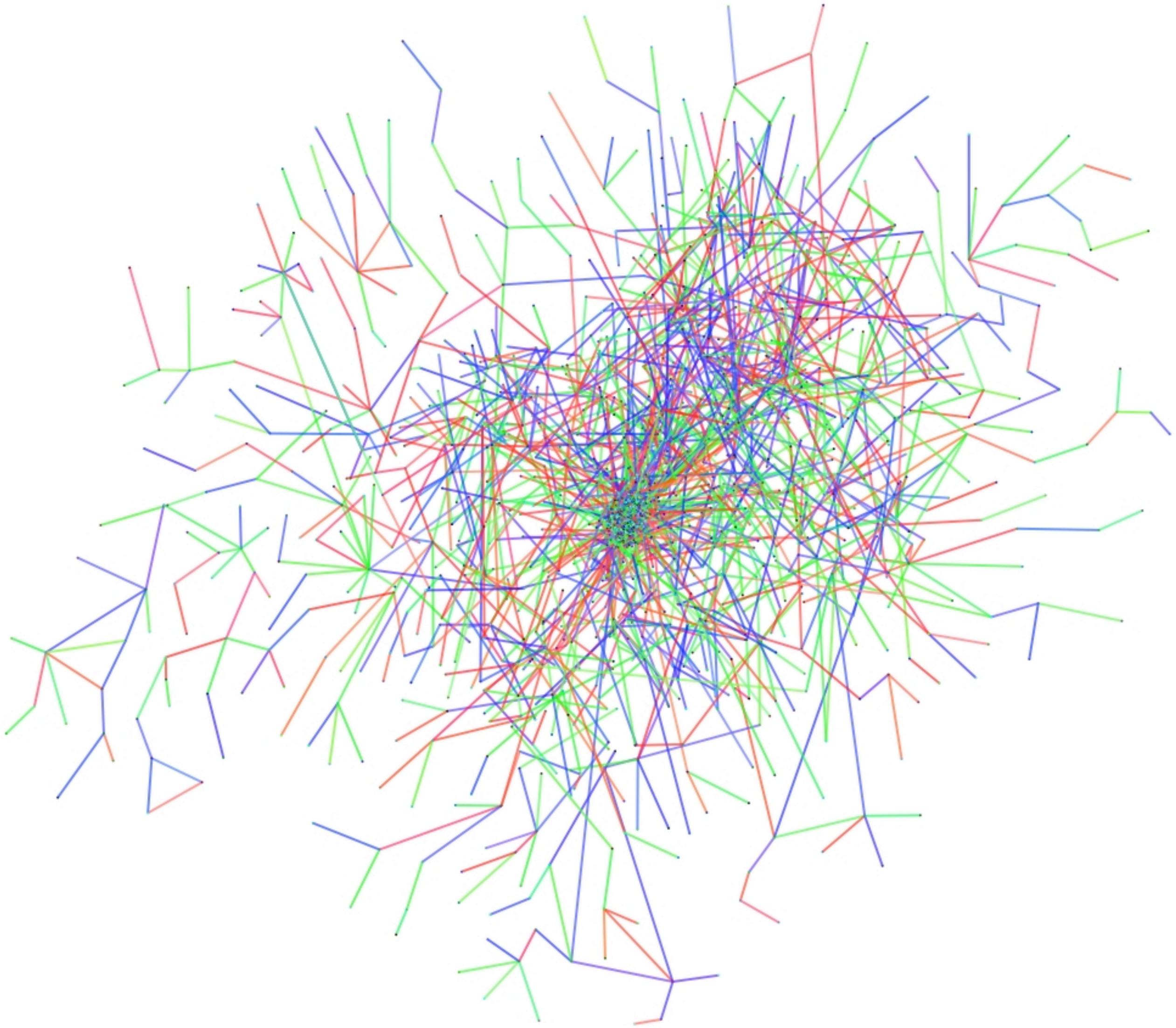
Document Representation for IR

Document Representation for IR



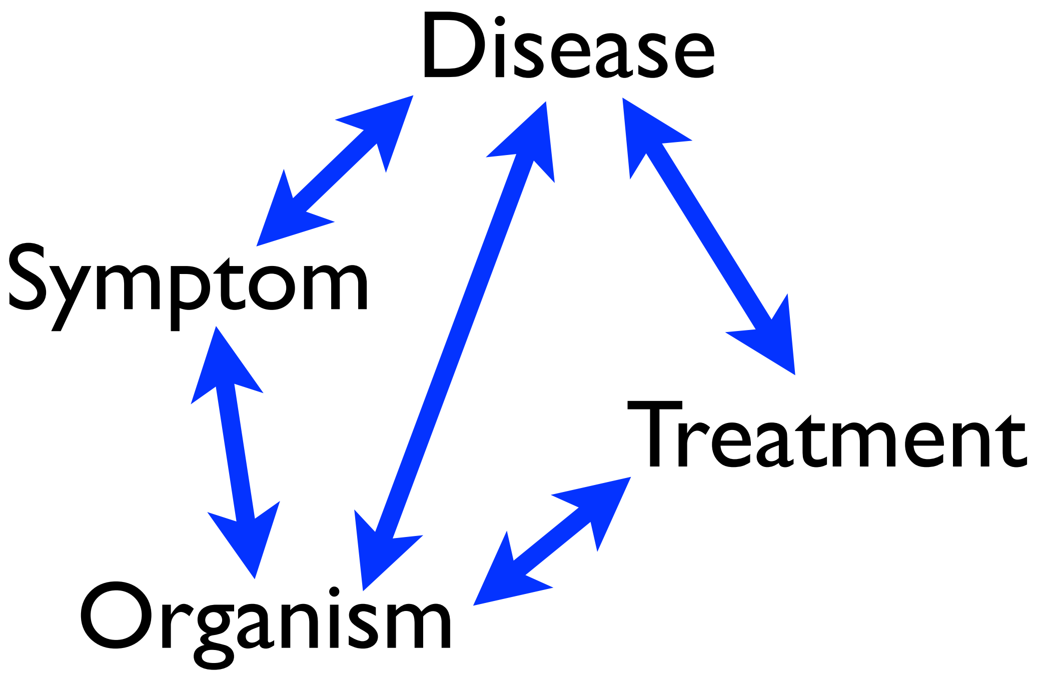
Document Representation for IR





Why Medical IR?

- Vocabulary mismatch
  - hypertension ≈ high blood pressure
- Interdependence between terms



Disease

Symptom

Treatment

Organism

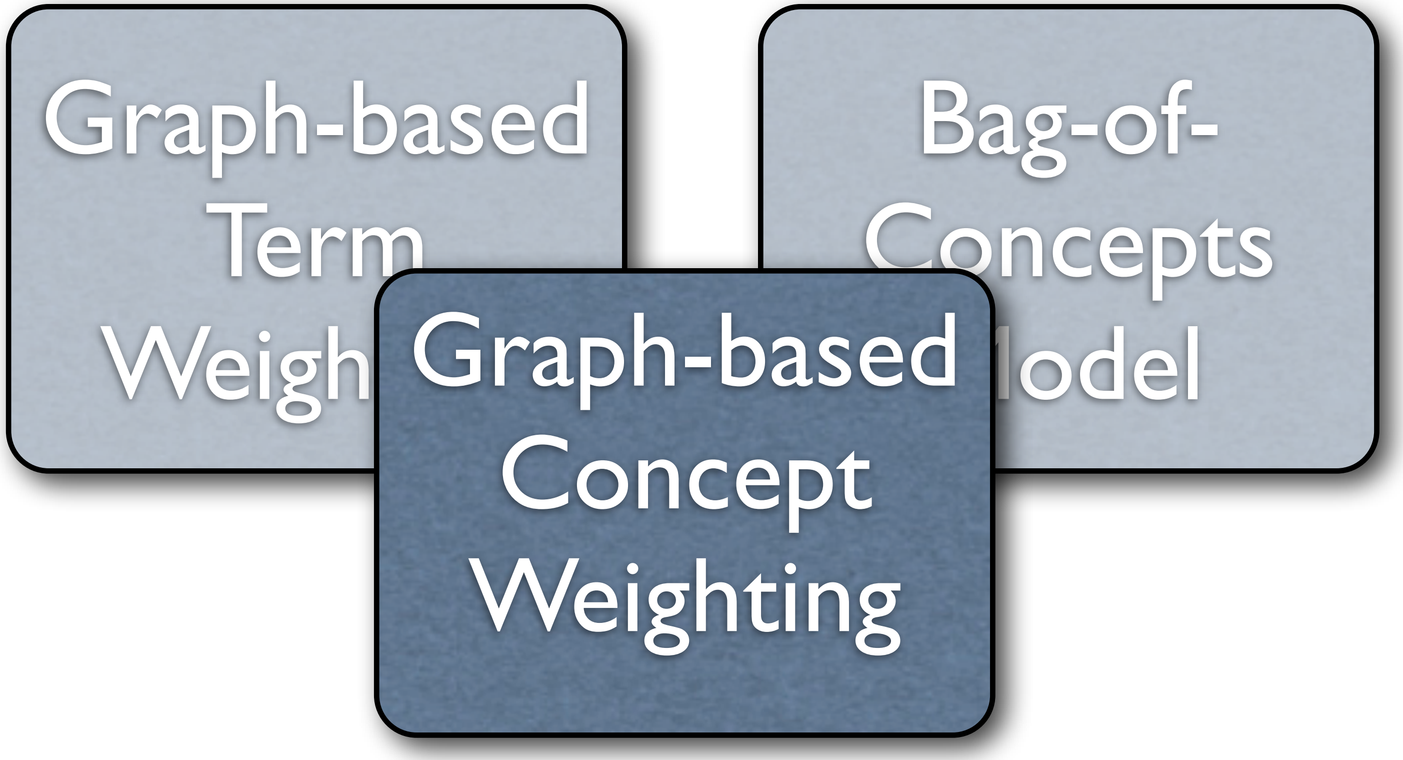
(Patel et al, 2007)

Overview

Graph-based Term Weighting

Bag-of-Concepts Model

Overview

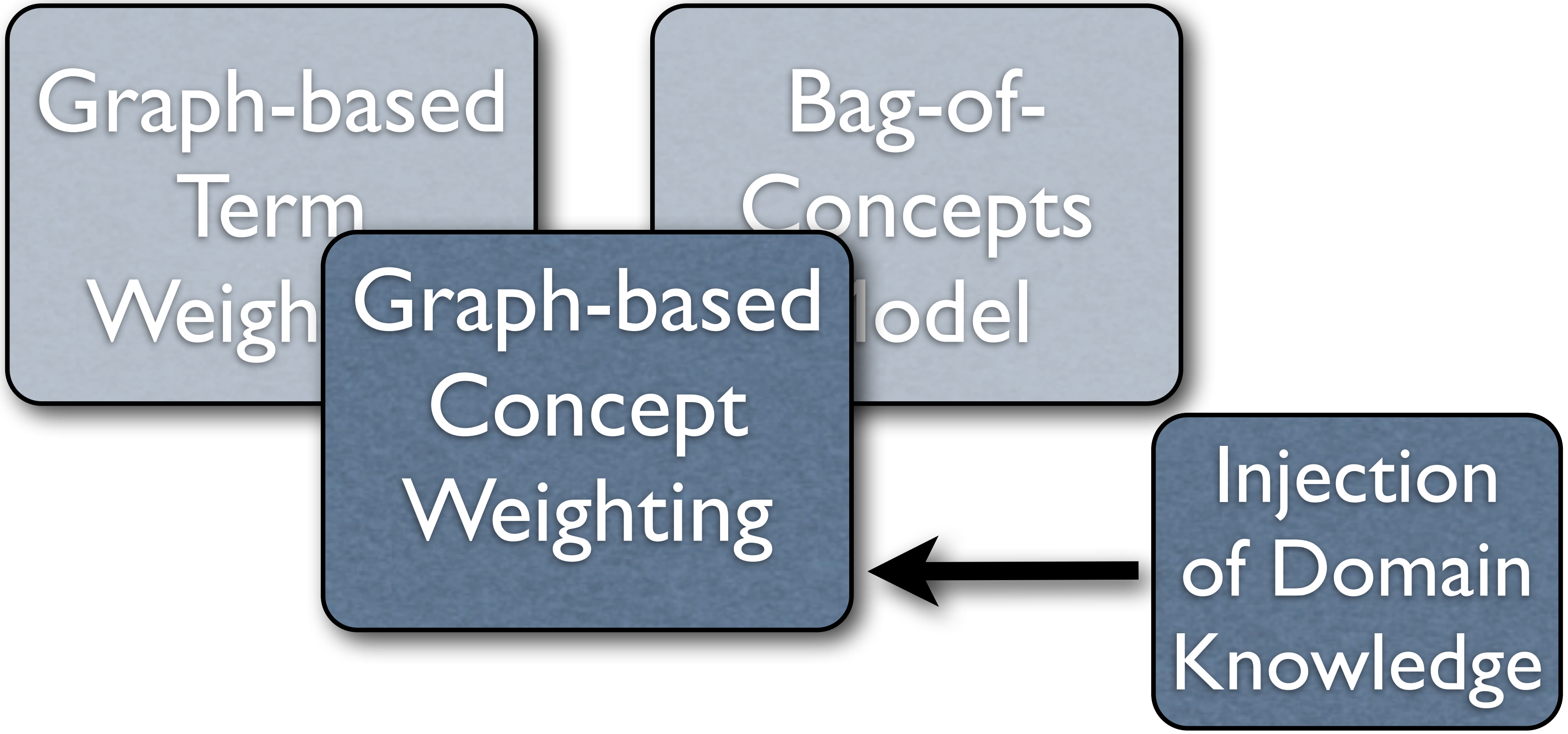


Graph-based Term Weigh

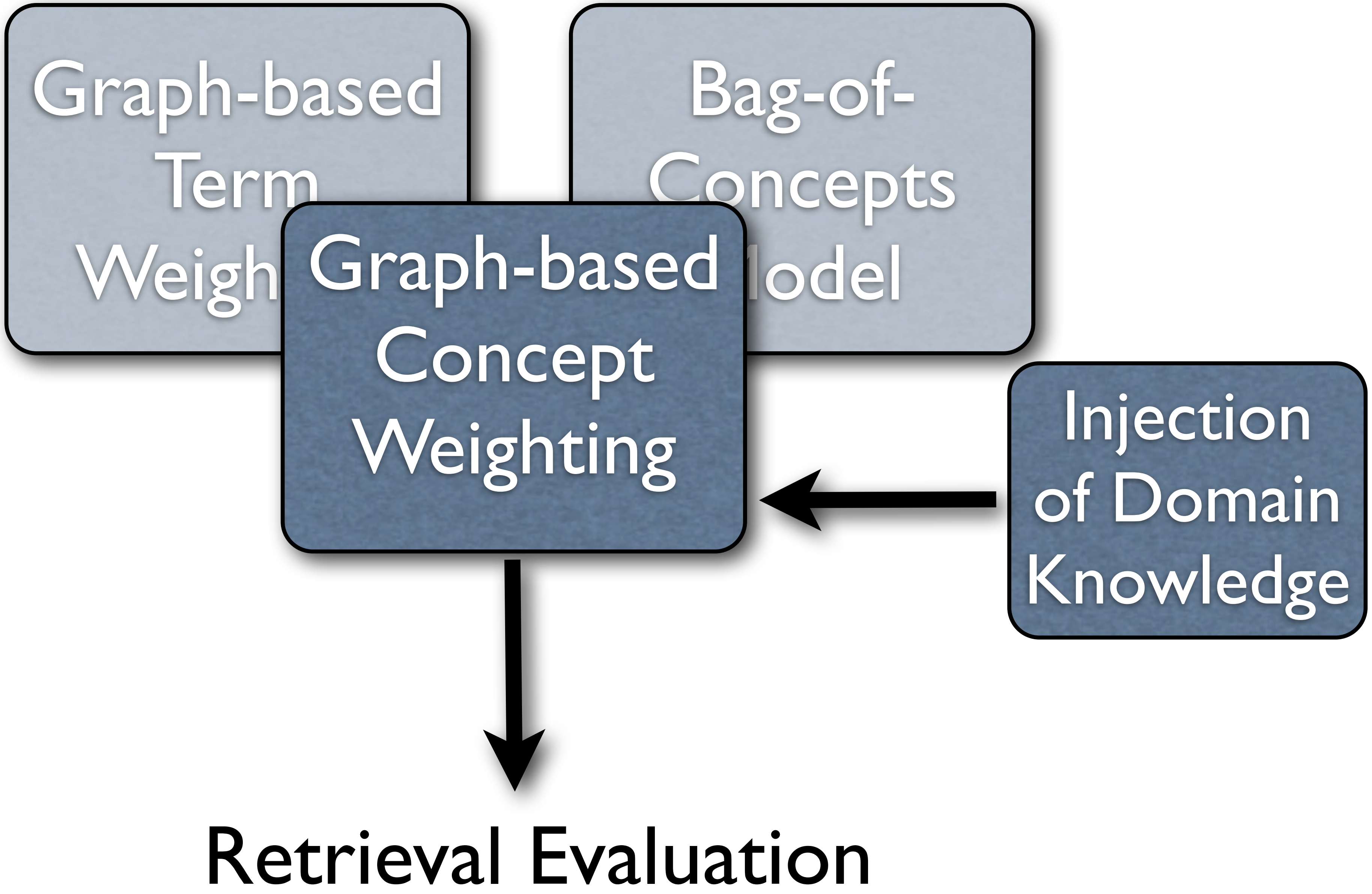
Bag-of-Concepts odel

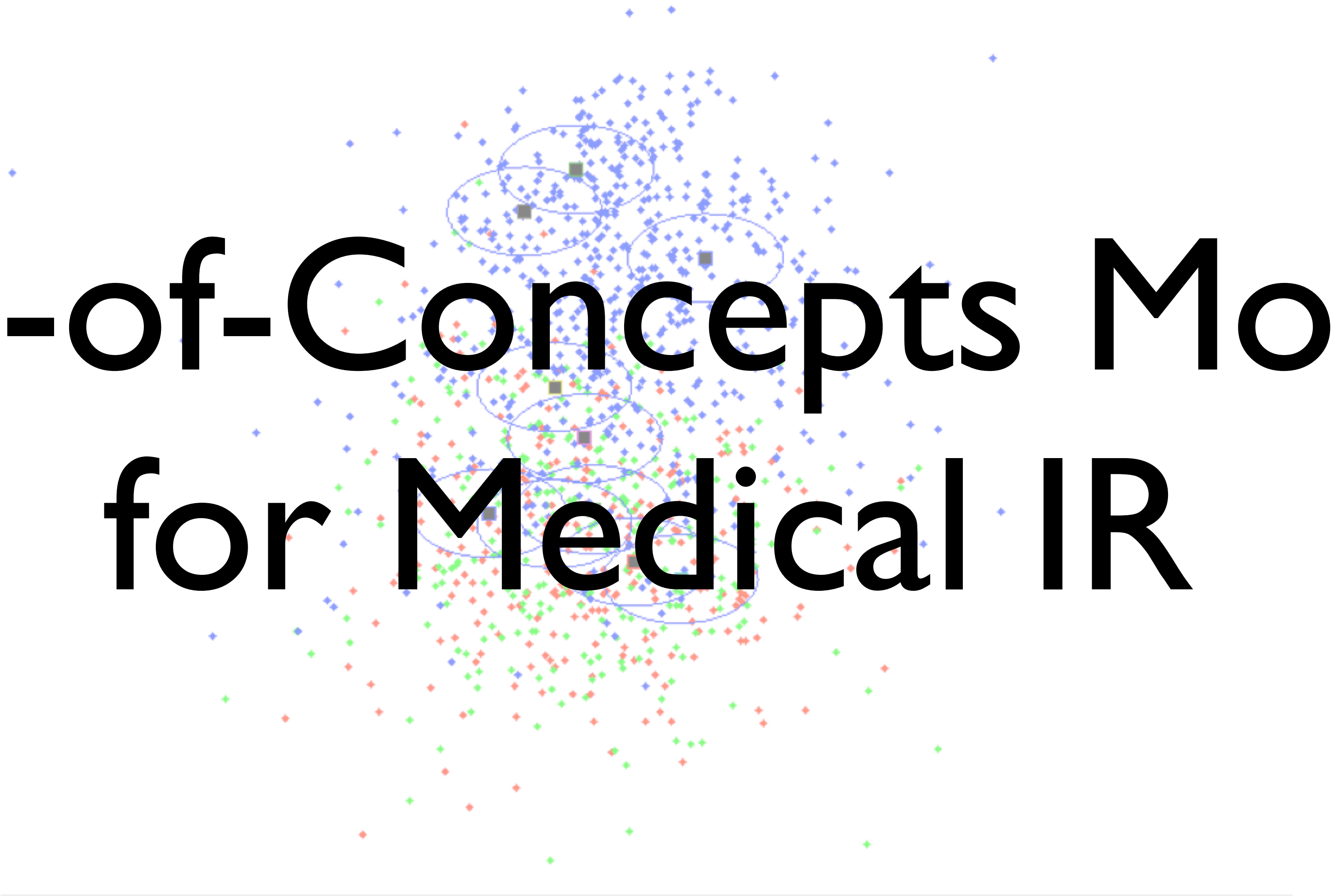
Graph-based Concept Weighting

Overview


Graph-based Term Weigh
Bag-of-Concepts odel
Graph-based Concept Weighting
Injection of Domain Knowledge

Overview


Graph-based Term Weigh
Bag-of-Concepts lodel
Graph-based Concept Weighting
Injection of Domain Knowledge
Retrieval Evaluation

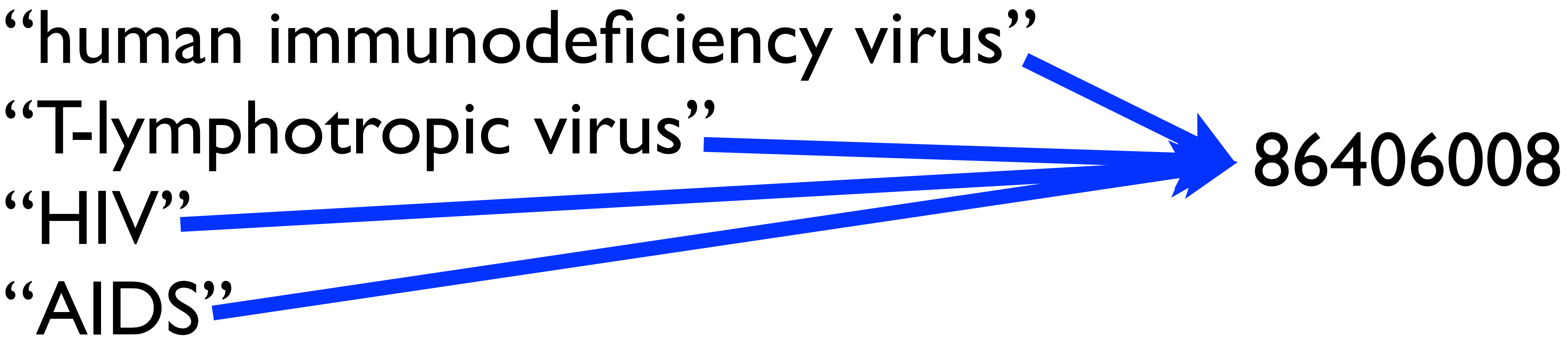


Bag-of-Concepts Model for Medical IR

Convert Terms to Concepts

“human immunodeficiency virus”
“T-lymphotropic virus”
“HIV”
“AIDS”

Convert Terms to Concepts



“human immunodeficiency virus”

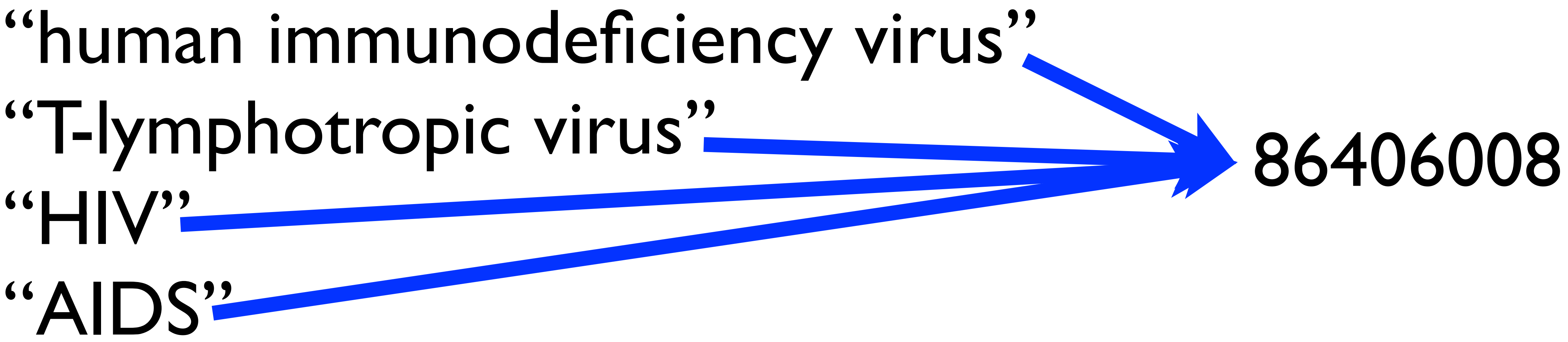
“T-lymphotropic virus”

“HIV”

“AIDS”

86406008

Convert Terms to Concepts



"human immunodeficiency virus"

"T-lymphotropic virus"

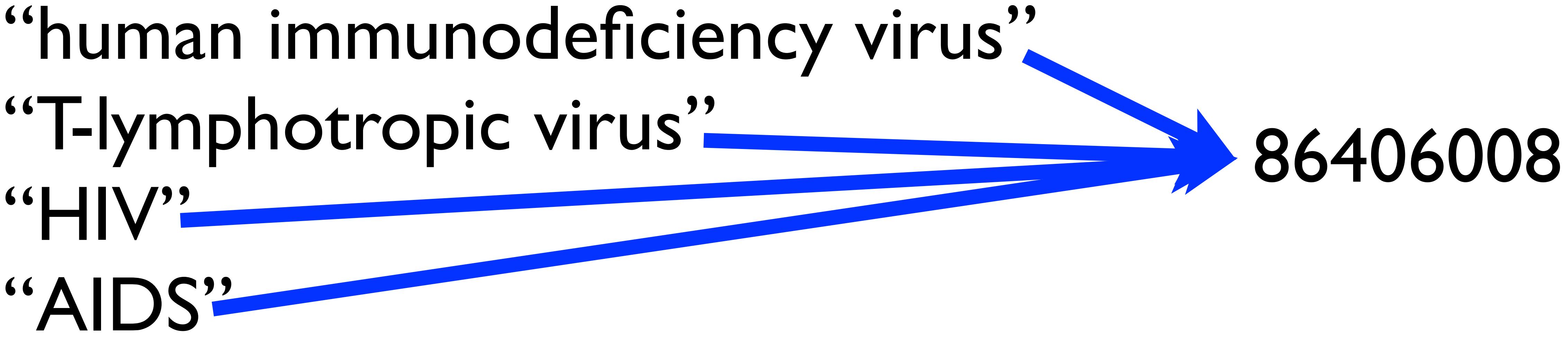
"HIV"

"AIDS"

86406008

"esophageal reflux"

Convert Terms to Concepts



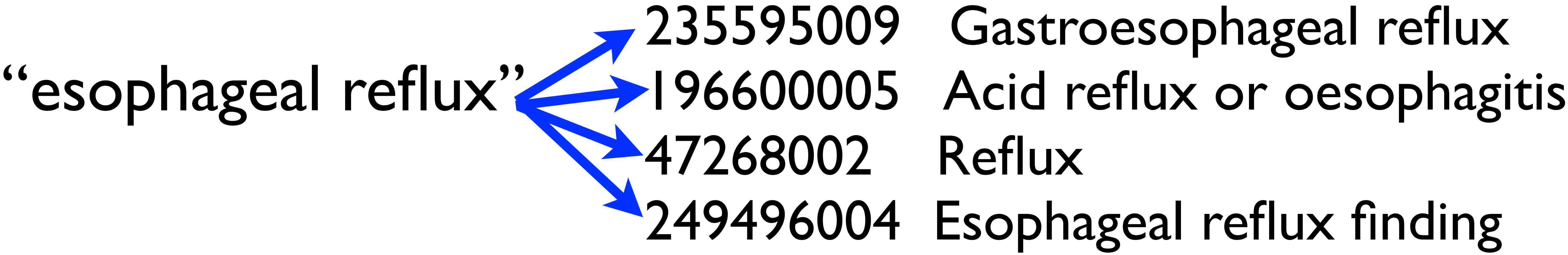
“human immunodeficiency virus”

“T-lymphotropic virus”

“HIV”

“AIDS”

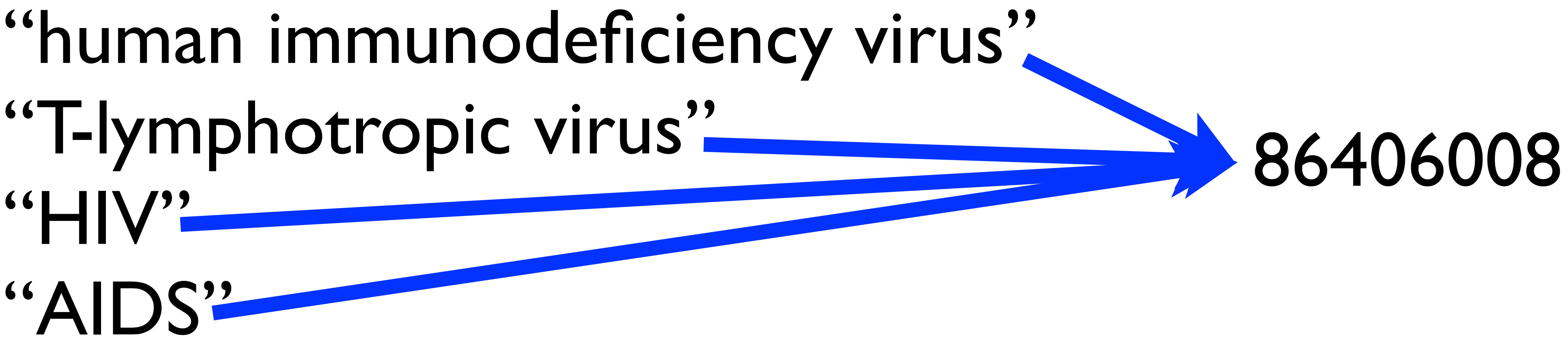
86406008



“esophageal reflux”

235595009	Gastroesophageal reflux
196600005	Acid reflux or oesophagitis
47268002	Reflux
249496004	Esophageal reflux finding

Convert Terms to Concepts



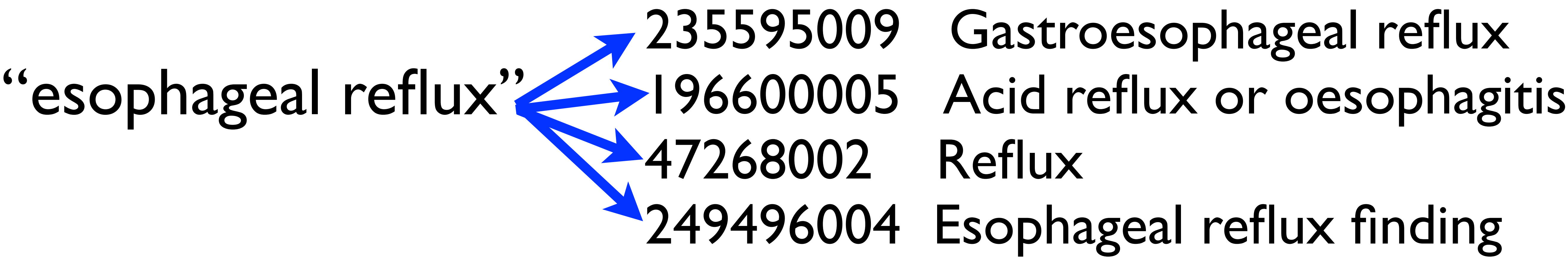
"human immunodeficiency virus"

"T-lymphotropic virus"

"HIV"

"AIDS"

86406008



"esophageal reflux"

235595009 Gastroesophageal reflux

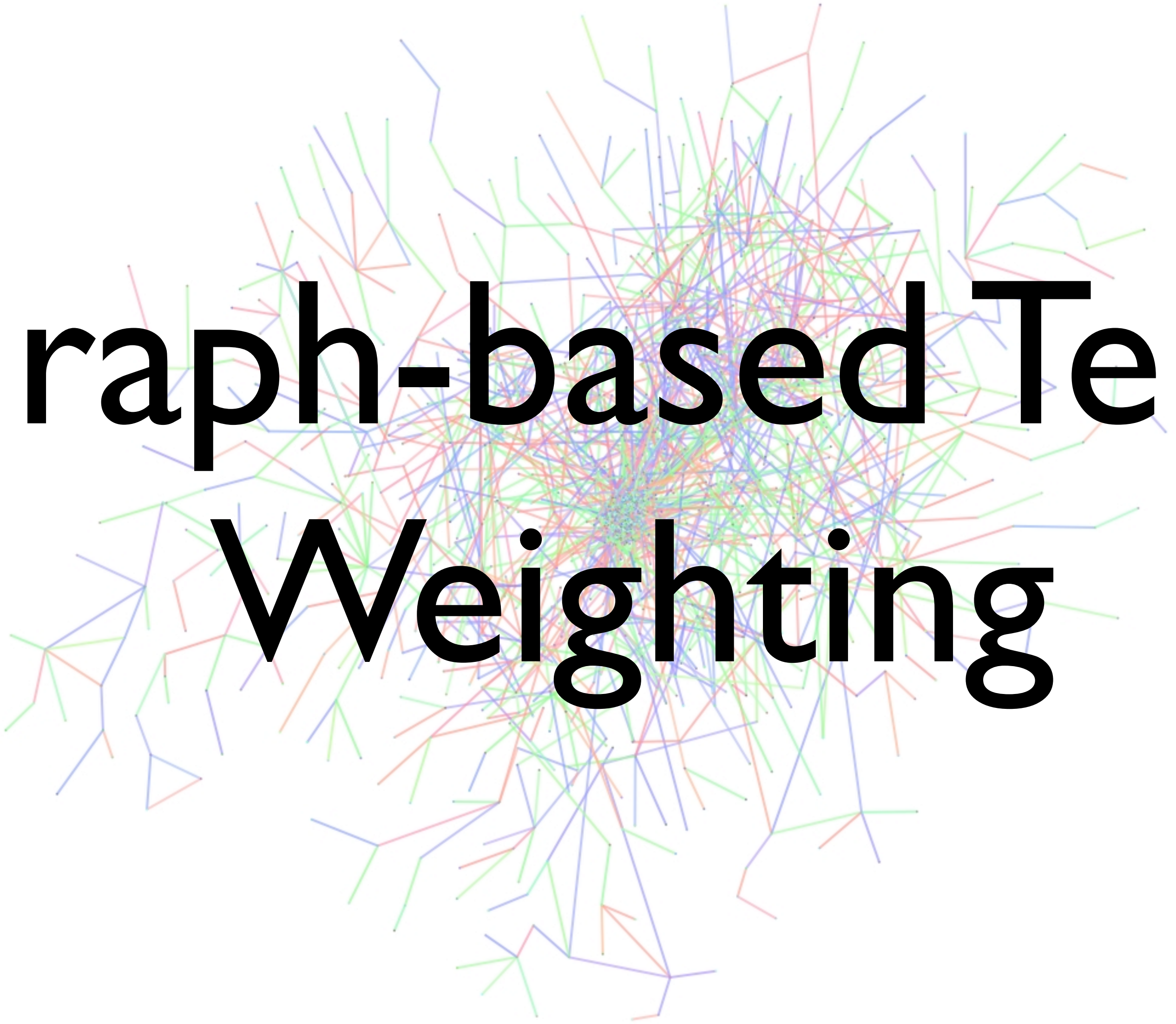
196600005 Acid reflux or oesophagitis

47268002 Reflux

249496004 Esophageal reflux finding

Index & retrieval on "bag-of-concepts" (Koopman et al, 2012)

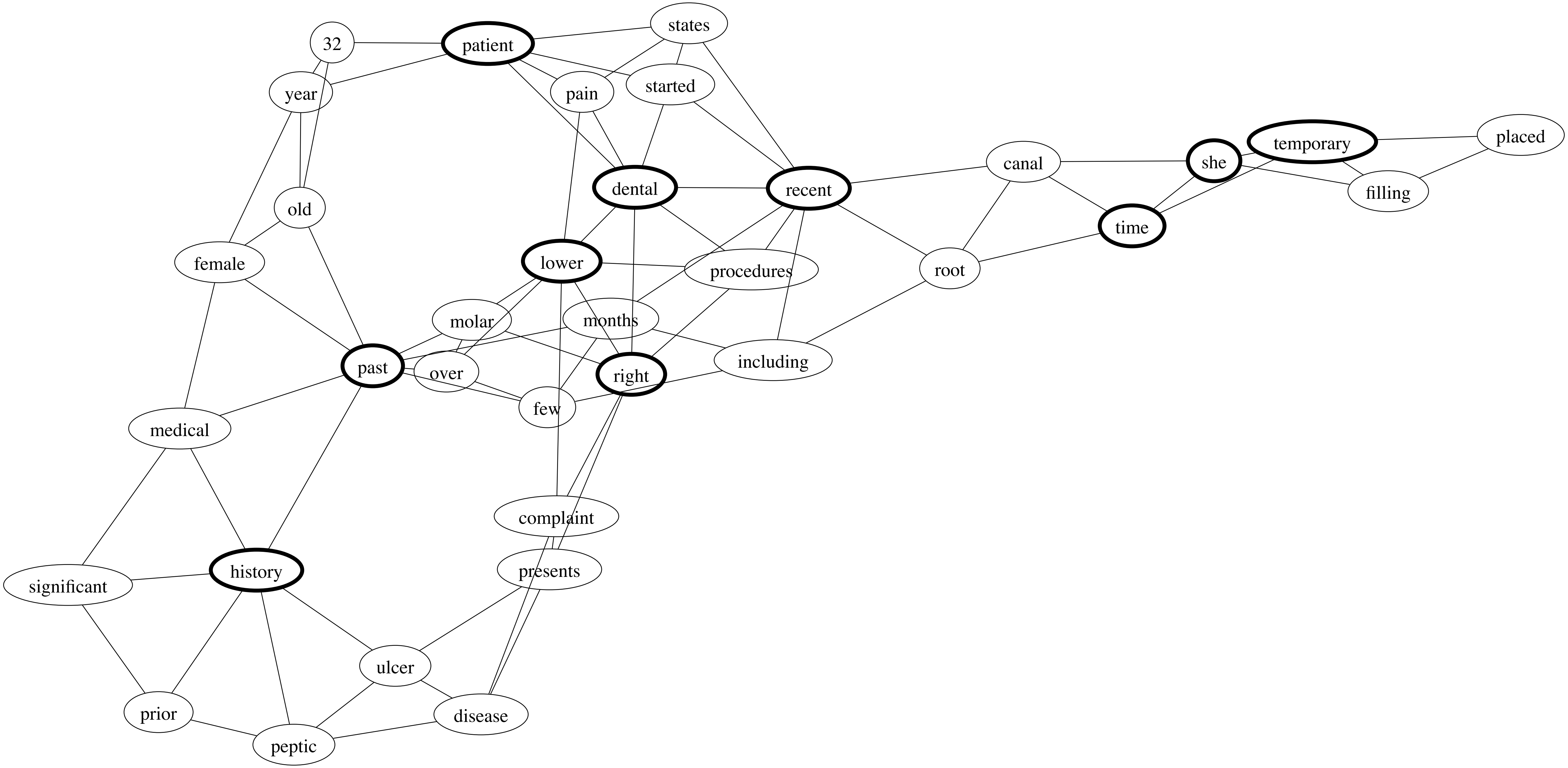
Graph-based Term Weighting



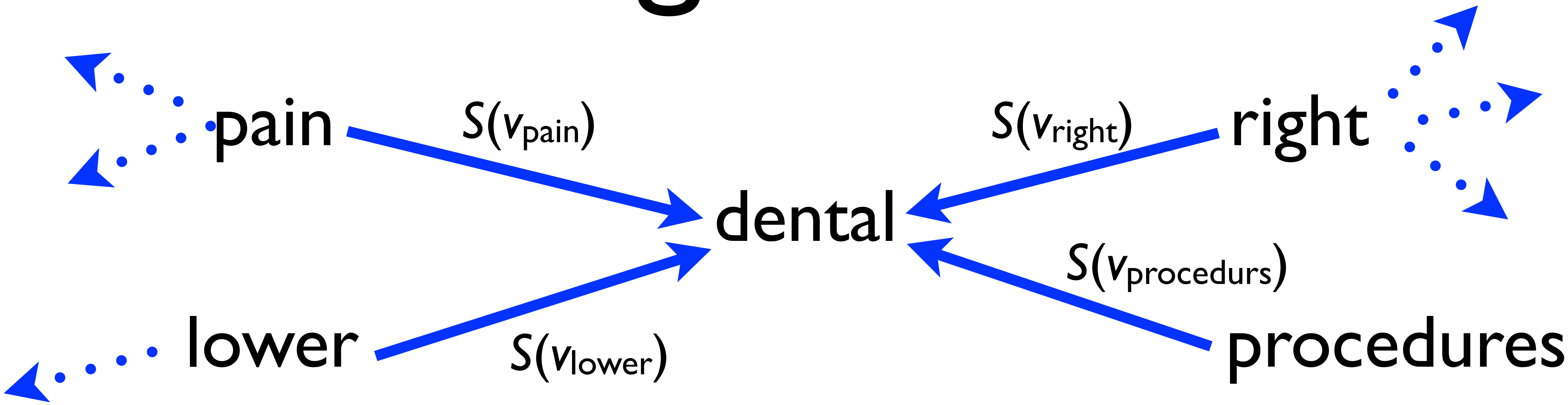
Example Medical Doc

"The patient is a 32-year-old female with a past medical history significant for a prior history of peptic ulcer disease who presents with a complaint of right lower dental pain. The patient states that she was started on recent dental procedures, on a right lower molar, over the past few months, including a recent root canal, at which time she had a temporary filling placed."

Document Term Graph

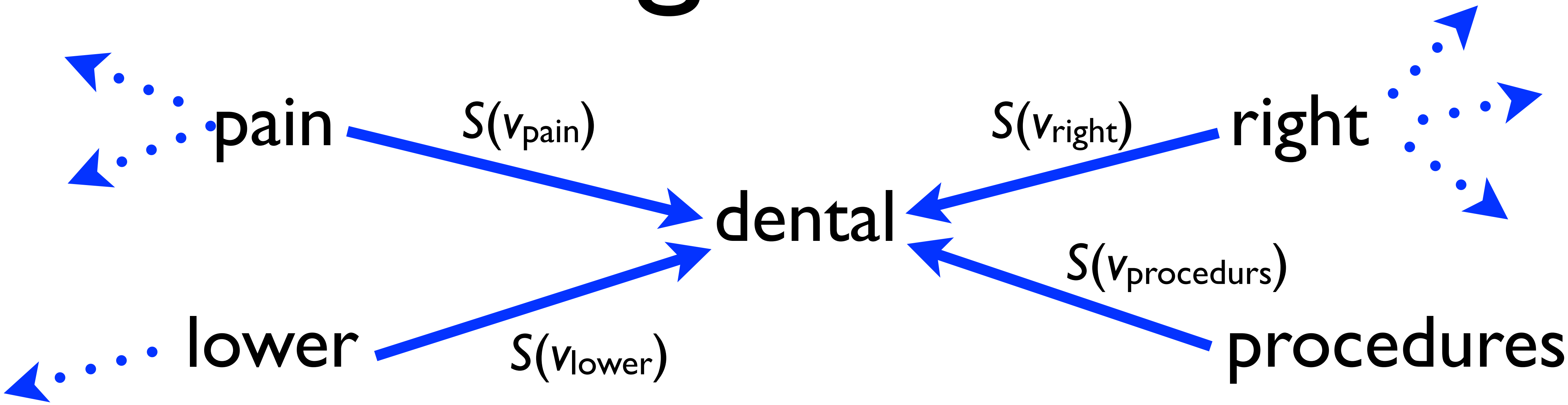


Term Weighting using PageRank



(Blanco & Lioma, 2012)

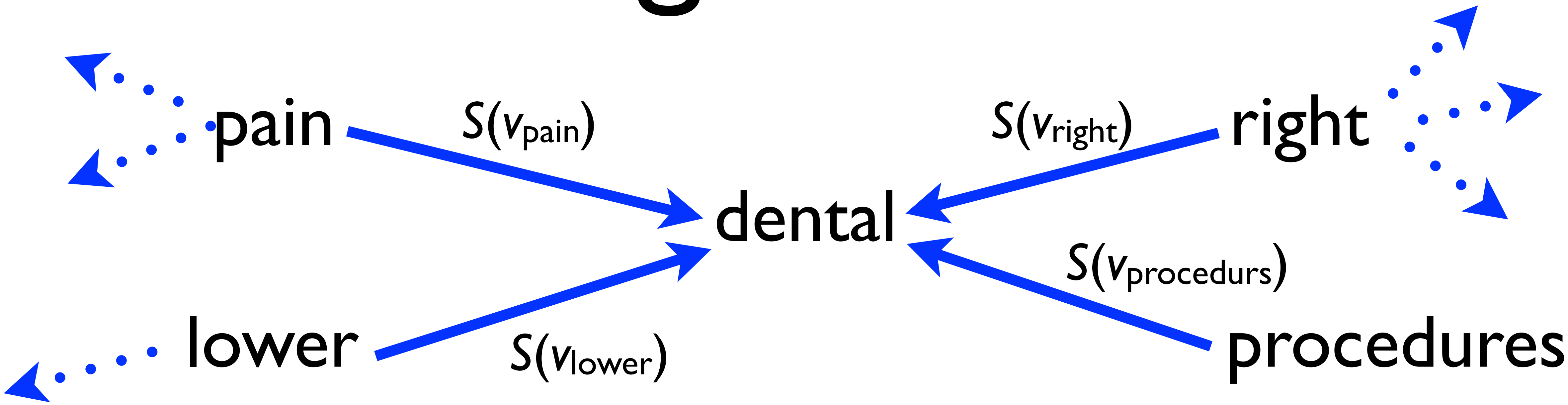
Term Weighting using PageRank



$$S(v_t) = \sum_{v_j \in \mathcal{V}(v_t)} \frac{S(v_j)}{|\mathcal{V}(v_j)|}$$

(Blanco & Lioma, 2012)

Term Weighting using PageRank



$$S(v_t) = (1 - \phi) + \phi * \sum_{v_j \in \mathcal{V}(v_t)} \frac{S(v_j)}{|\mathcal{V}(v_j)|}$$

(Blanco & Lioma, 2012)

Retrieval Function

$$S(v_t) = (1 - \phi) + \phi * \sum_{v_j \in \mathcal{V}(v_t)} \frac{S(v_j)}{|\mathcal{V}(v_j)|}$$

(Blanco & Lioma, 2012)

Retrieval Function

$$S(v_t) = (1 - \phi) + \phi * \sum_{v_j \in \mathcal{V}(v_t)} \frac{S(v_j)}{|\mathcal{V}(v_j)|}$$

$$w(t, d) = idf(t) * S(v_t)$$

(Blanco & Lioma, 2012)

Retrieval Function

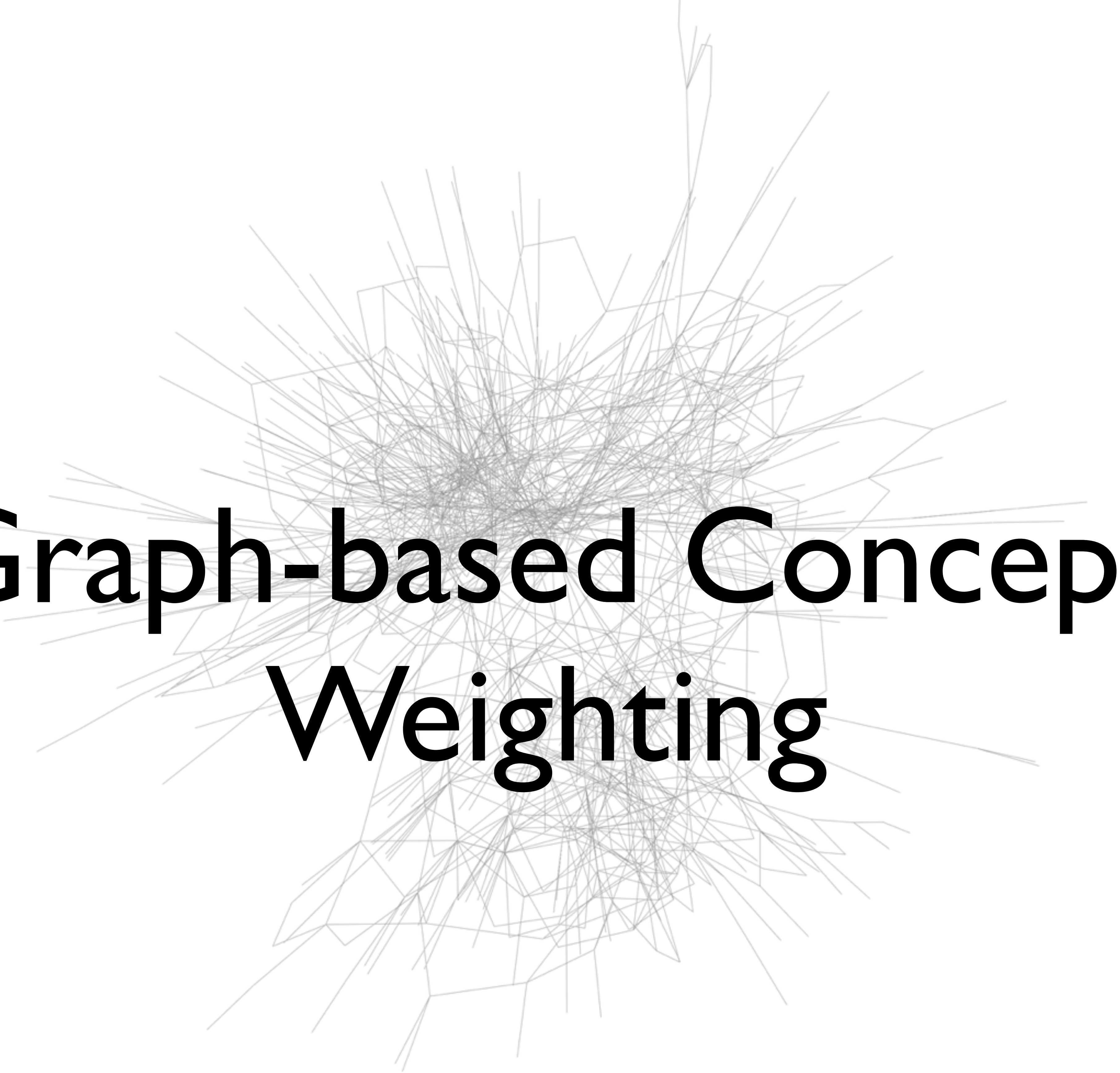
$$S(v_t) = (1 - \phi) + \phi * \sum_{v_j \in \mathcal{V}(v_t)} \frac{S(v_j)}{|\mathcal{V}(v_j)|}$$

$$w(t, d) = idf(t) * S(v_t)$$

$$R(d, q) = \sum_{t \in q} w(t, d)$$

(Blanco & Lioma, 2012)

Graph-based Concept Weighting



Concept-based Retrieval Function

$$w(c, d_c) = idf(c) * S(v_c)$$

$$R(d_c, q_c) = \sum_{c \in q_c} w(c, d_c)$$

Concept-based Retrieval Function

Concept *c*

$$w(c, d_c) = idf(c) * S(v_c)$$

$$R(d_c, q_c) = \sum_{c \in q_c} w(c, d_c)$$

Concept-based Retrieval Function

Concept *c*

Document (concepts) *d_c*

$$w(c, d_c) = idf(c) * S(v_c)$$

$$R(d_c, q_c) = \sum_{c \in q_c} w(c, d_c)$$

Concept-based Retrieval Function

Concept c

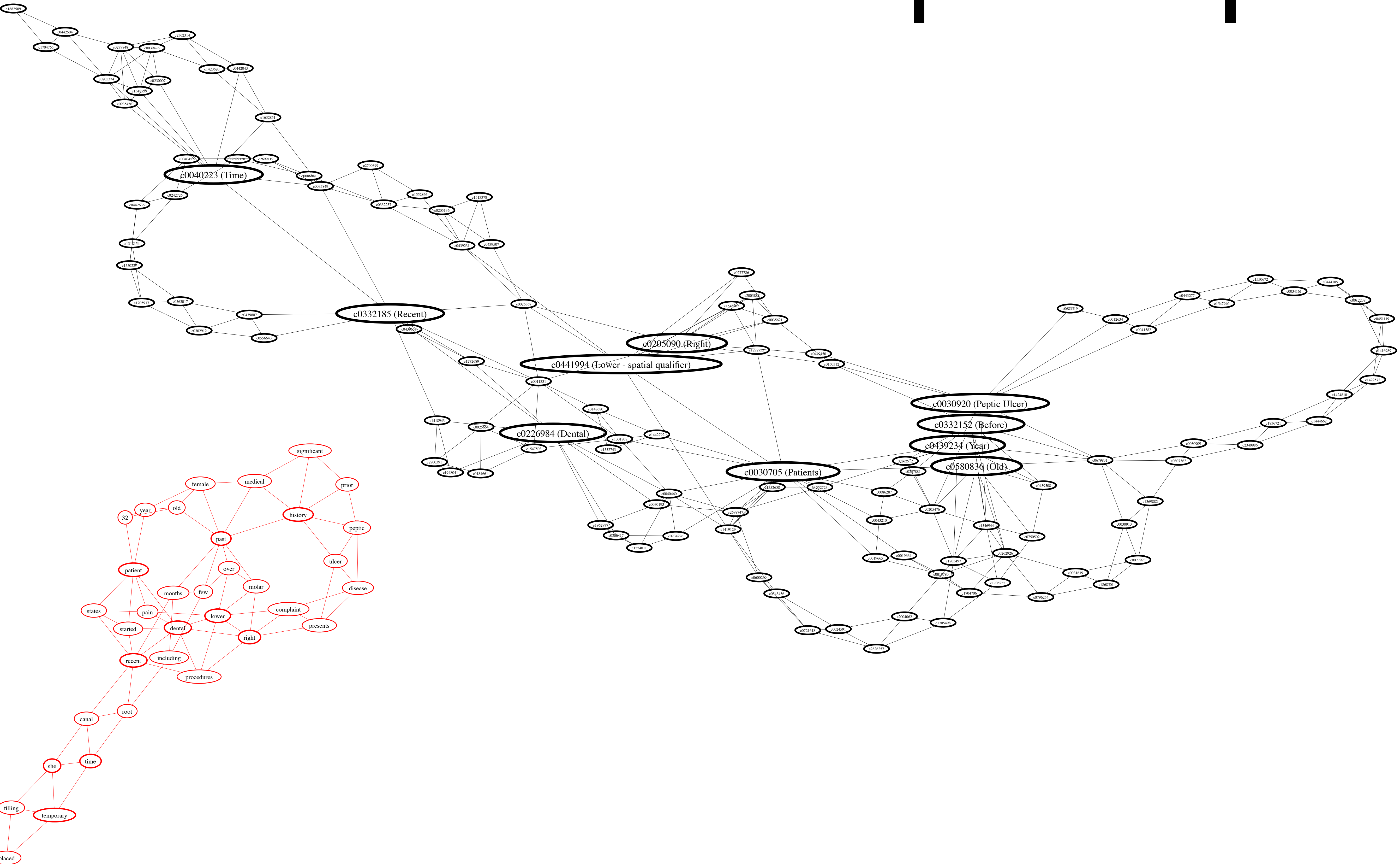
Document (concepts) d_c

$$w(c, d_c) = idf(c) * S(v_c)$$

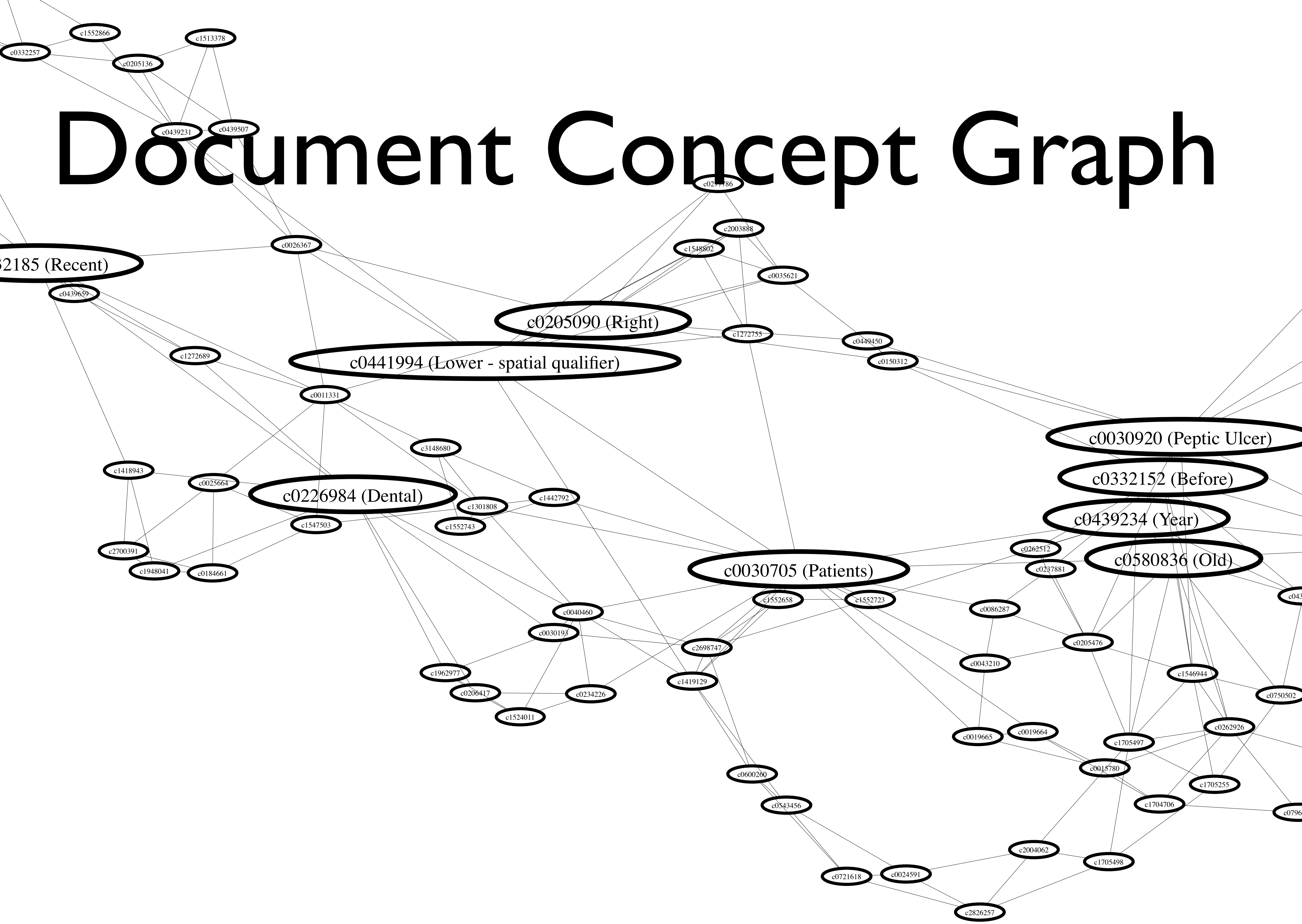
$$R(d_c, q_c) = \sum_{c \in q_c} w(c, d_c)$$

Query (concepts) q_c

Document Concept Graph



Document Concept Graph



Injection of Domain Knowledge

Injection of Domain Knowledge

- Document is a graph of SNOMED CT concepts

Injection of Domain Knowledge

- Document is a graph of SNOMED CT concepts

- SNOMED CT ontology is also a graph

Injection of Domain Knowledge

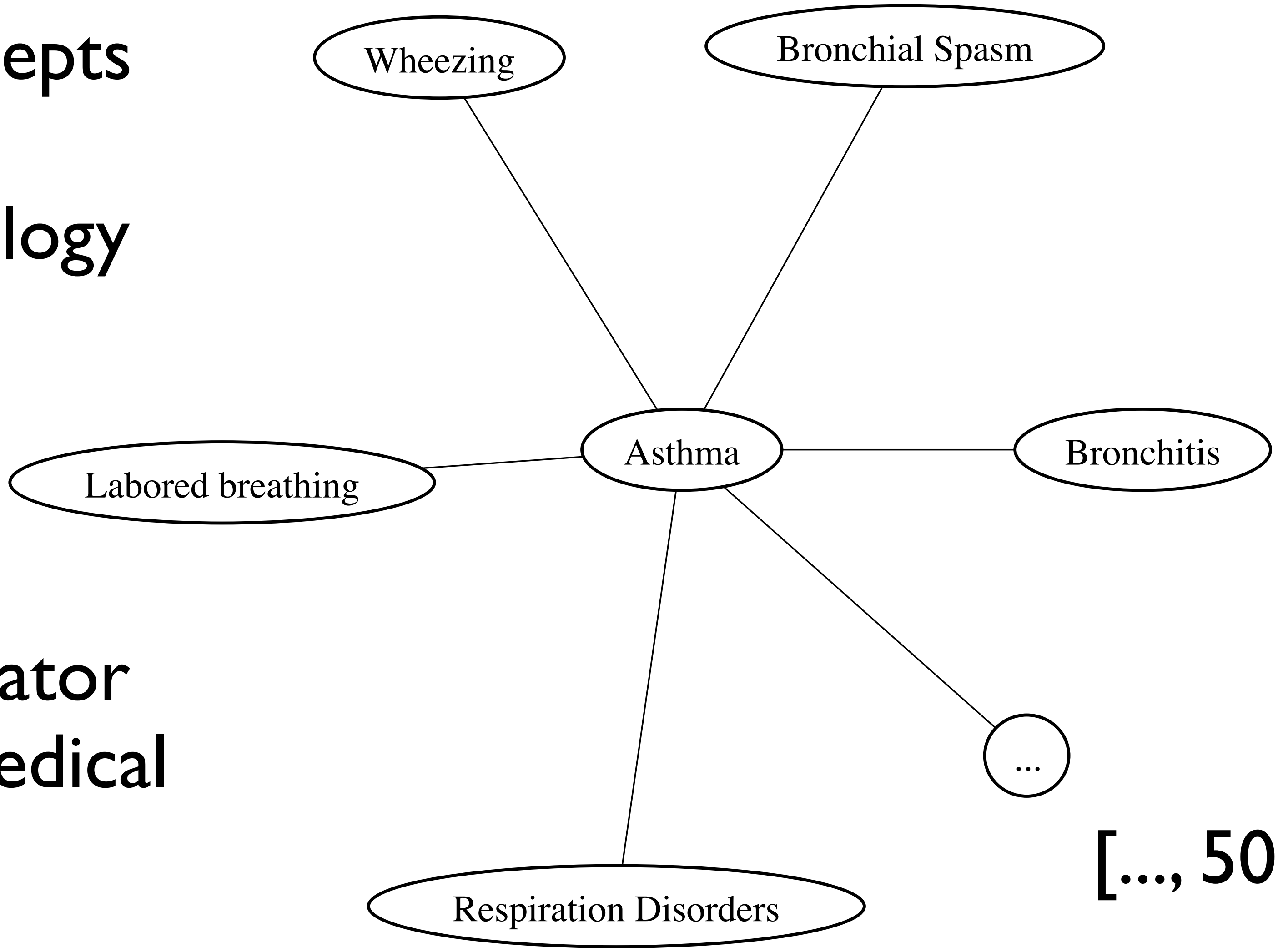
- Document is a graph of SNOMED CT concepts

- SNOMED CT ontology is also a graph

- Concepts “connectedness” in SNOMED CT indicator of importance in medical domain

Injection of Domain Knowledge

- Document is a graph of SNOMED CT concepts
- SNOMED CT ontology is also a graph
- Concepts "connectedness" in SNOMED CT indicator of importance in medical domain



[..., 50]

Domain Importance Concept Weighting

- Adjust original concept weight by the "background" importance of concept in medical domain:

Domain Importance Concept Weighting

- Adjust original concept weight by the "background" importance of concept in medical domain:

$$w(c, d_c) = idf(c) * S(v_c) * \log(|\mathcal{V}_s(c)|)$$

Domain Importance Concept Weighting

- Adjust original concept weight by the "background" importance of concept in medical domain:

$$w(c, d_c) = idf(c) * S(v_c) * \log(|\mathcal{V}_s(c)|)$$

document

Domain Importance Concept Weighting

- Adjust original concept weight by the “background” importance of concept in medical domain:

$$w(c, d_c) = idf(c) * S(v_c) * \log(|\mathcal{V}_s(c)|)$$

corpus

document

Domain Importance Concept Weighting

- Adjust original concept weight by the "background" importance of concept in medical domain:

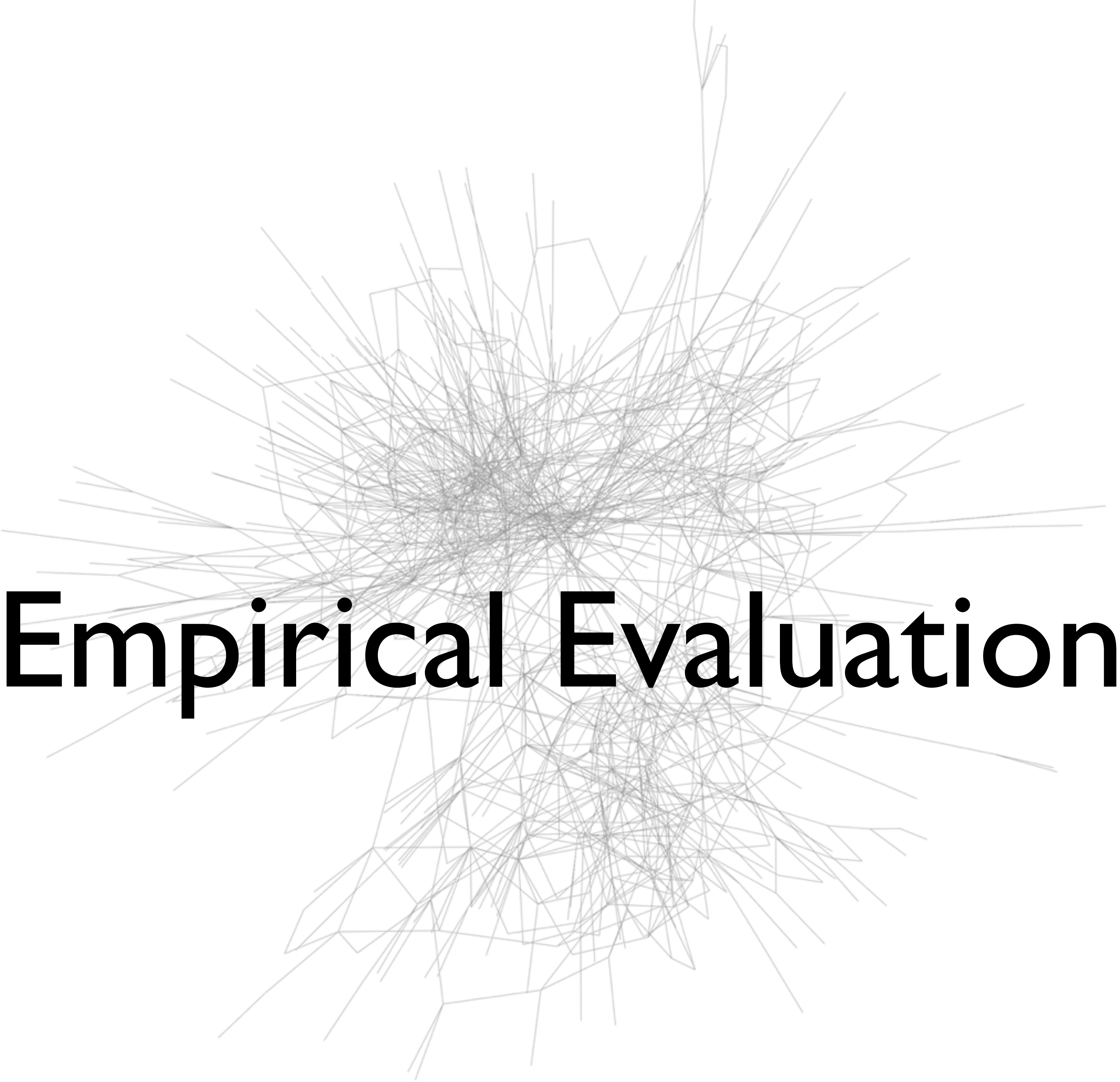
$$w(c, d_c) = idf(c) * S(v_c) * \log(|\mathcal{V}_s(c)|)$$

corpus

document

domain

Empirical Evaluation



Test Collection

- TREC 2011 Medical Records Track
  - 100,866 clinical records
  - 34 clinical queries + qrels
- Entire collection converted to SNOMED-CT concepts using MetaMap

Baselines + Models

Baselines + Models

- terms-tfidf

Baselines + Models

- terms-tfidf
- concepts-tfidf

Baselines + Models

- terms-tfidf
- concepts-tfidf
- terms-graph

Baselines + Models

- terms-tfidf
- concepts-tfidf
- terms-graph

- concepts-graph

Baselines + Models

- terms-tfidf
- concepts-tfidf
- terms-graph

- concepts-graph
- concepts-graph-snomed

Baselines + Models

- terms-tfidf
- concepts-tfidf
- terms-graph

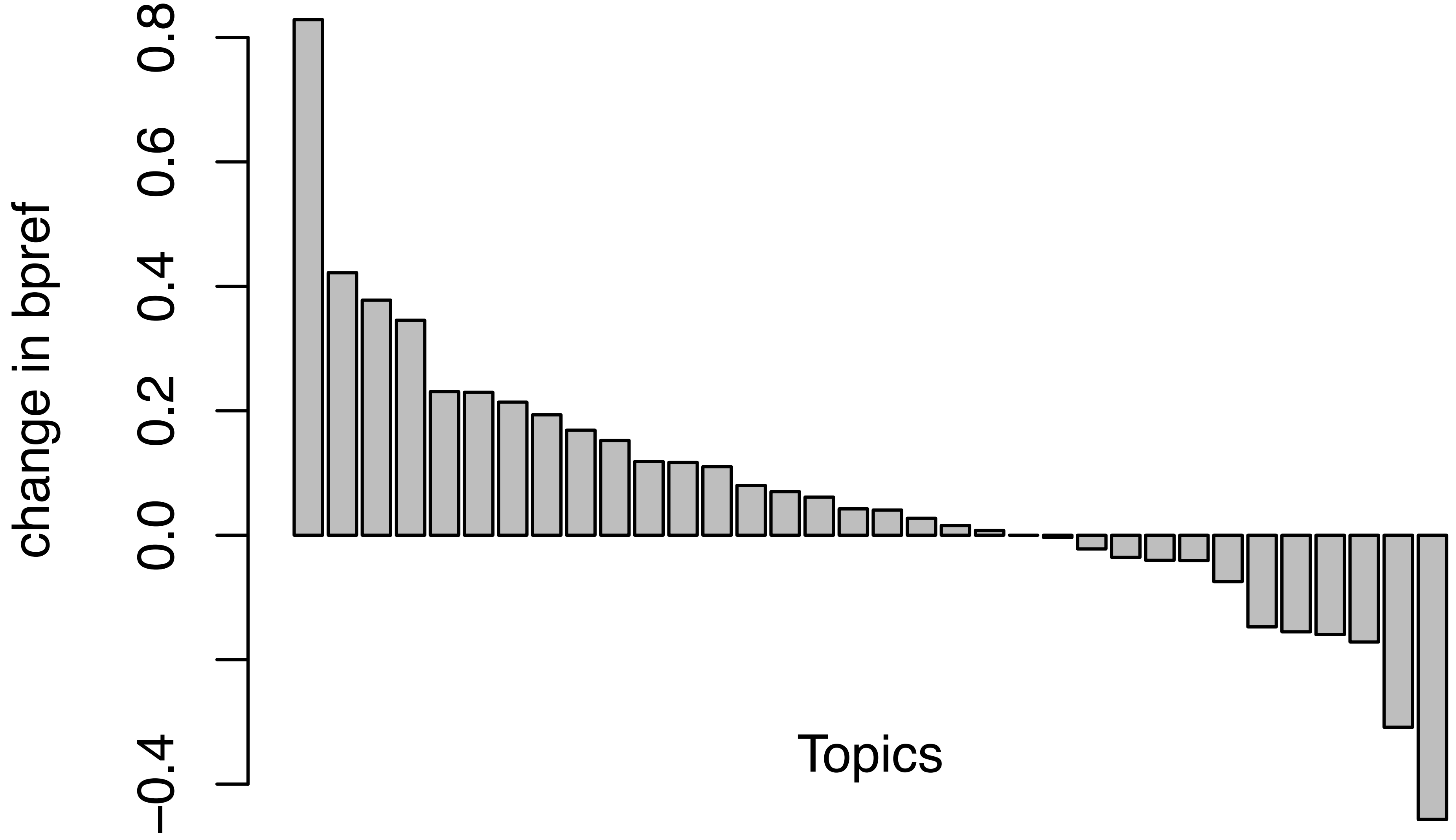
- concepts-graph
- concepts-graph-snomed

Bpref, Precision@10

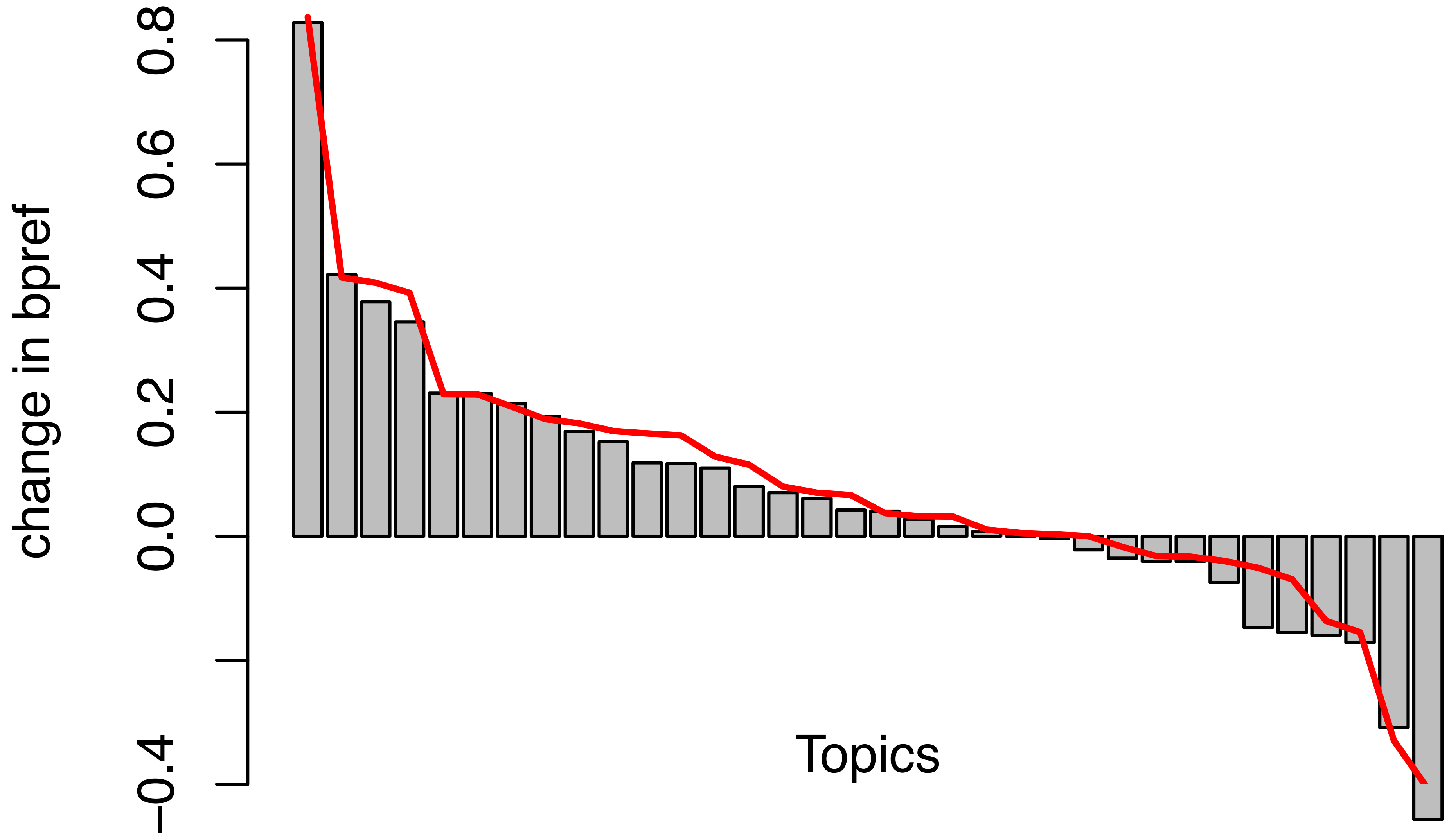
Retrieval Results

Run	**Bpref**	**Prec@10**
terms-tfidf	0.4722	0.4882
concepts-tfidf	0.4993	0.5176
terms-graph	0.4393	0.4882
concepts-graph	0.5050 (+15%)	0.5441 (+11%)
concepts-graph-snomed	**0.5245** (+19%)	**0.5559** (+14%)

Per-query



Per-query



Query reduction..?

$$w(c, d_c) = idf(c) * S(v_c) * \underline{\log(|\mathcal{V}_s(c)|)}$$

Query reduction..?

$$w(c, d_c) = idf(c) * S(v_c) * \underline{\log(|\mathcal{V}_s(c)|)}$$

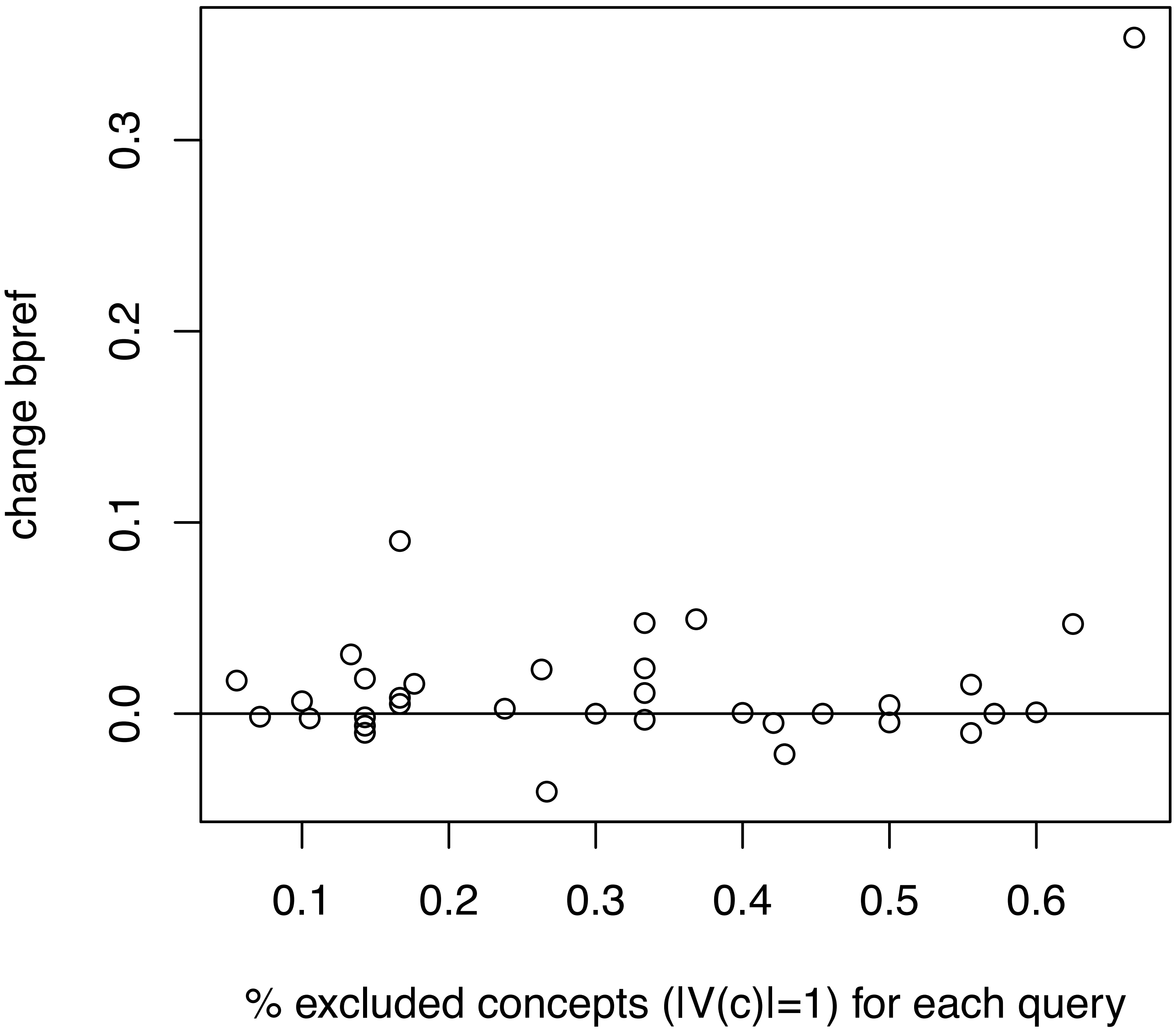
- 34 queries, 448 query concepts

Query reduction..?

$$w(c, d_c) = idf(c) * S(v_c) * \underline{\log(|\mathcal{V}_s(c)|)}$$

- 34 queries, 448 query concepts
- 127 (28%) excluded

Effect of Reduction



Conclusions

Conclusions

- Concept-based representations show improvements over terms representations

Conclusions

- Concept-based representations show improvements over terms representations
- Graph-based concept representation further improves over bag-of-concepts

Conclusions

- Concept-based representations show improvements over terms representations
- Graph-based concept representation further improves over bag-of-concepts
- Injection of domain knowledge provides further improvements & robustness

Conclusions

- Concept-based representations show improvements over terms representations
- Graph-based concept representation further improves over bag-of-concepts
- Injection of domain knowledge provides further improvements & robustness

- Integrating formal background knowledge into data-driven approaches to IR

TREC Medtrack'12

Run	infAP
terms-tfidf	0.1685
concepts-tfidf	0.2027
terms-graph	0.1394
concepts-graph	0.2072
concepts-graph-snomed	0.2123