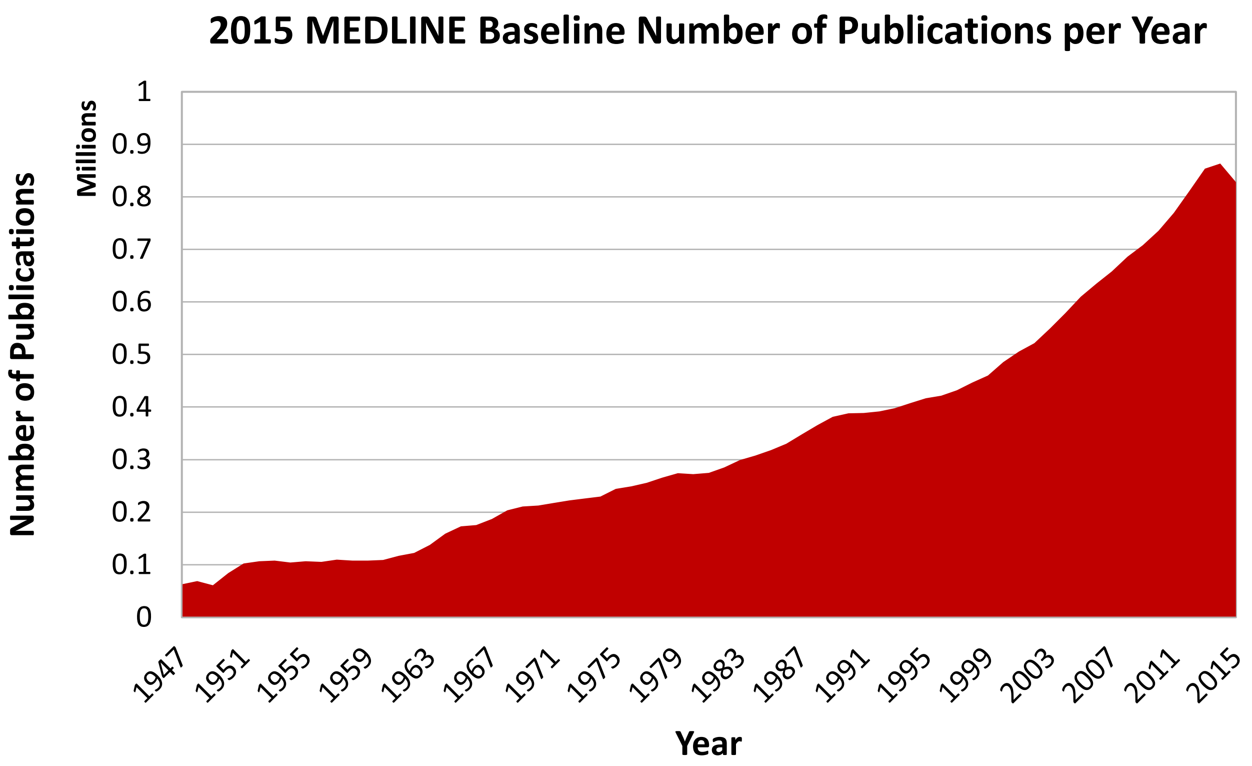
Literature Based Discovery

please hit the record button, Bridget

Increasing Publication Rate



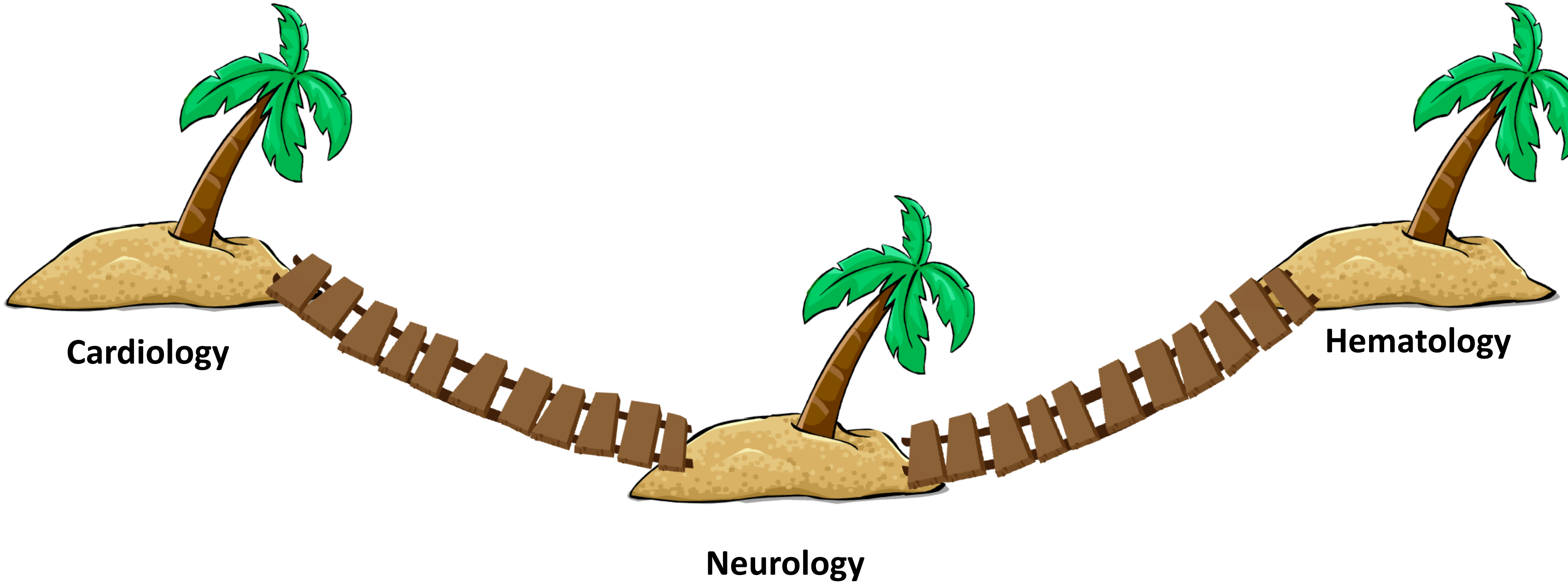
Islands of Knowledge

Cardiology

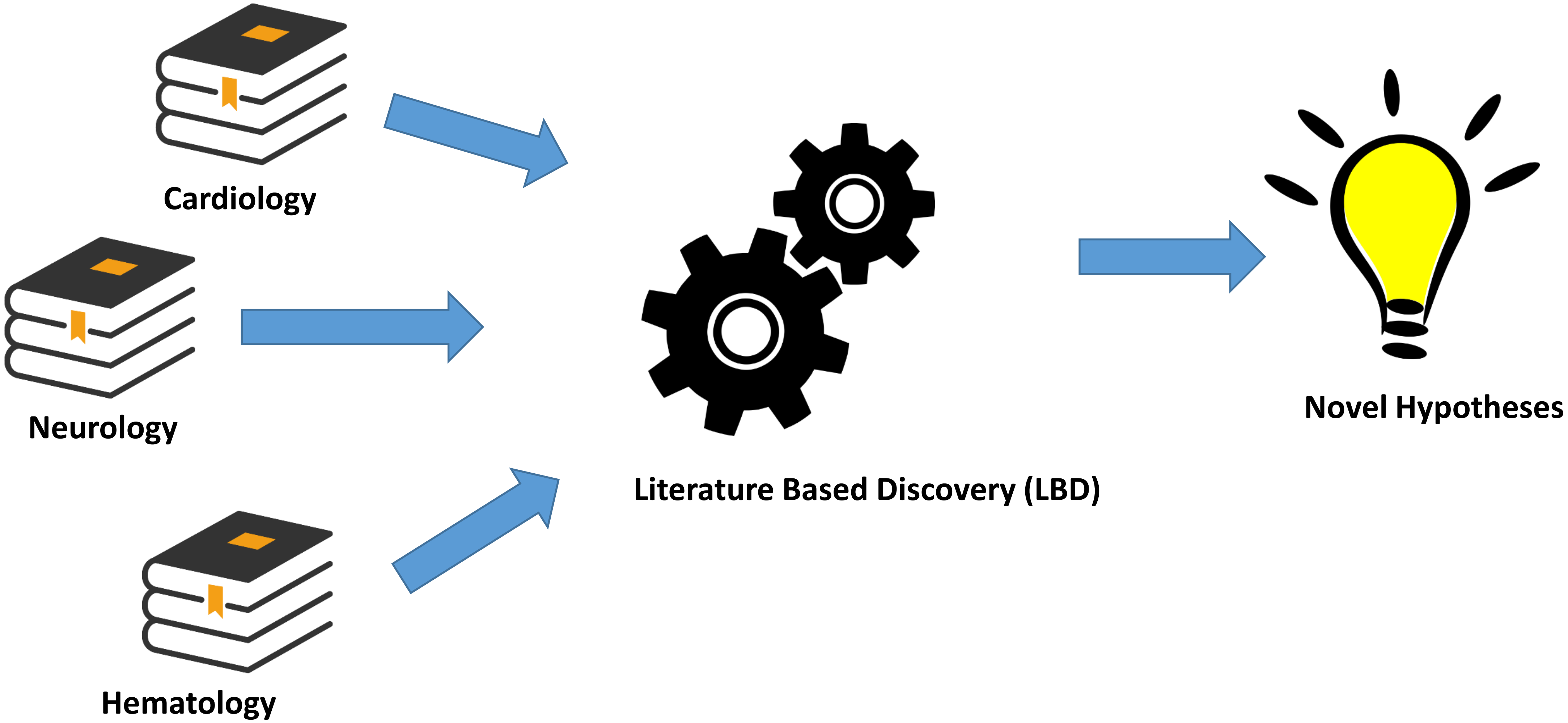
Neurology

Hematology

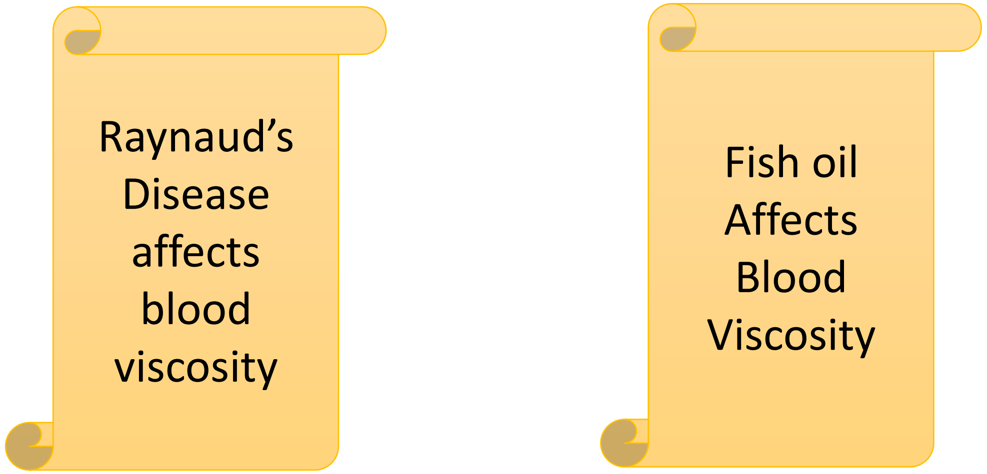
Literature Based Discovery



Literature Based Discovery

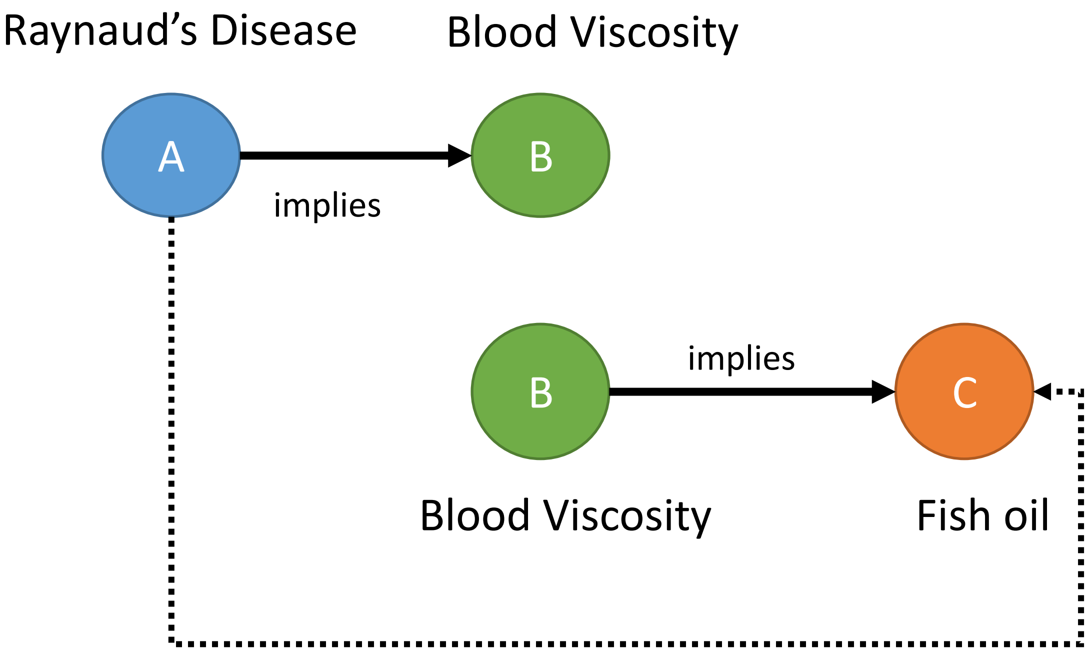


Literature Based Discovery



Document 1

Document 2

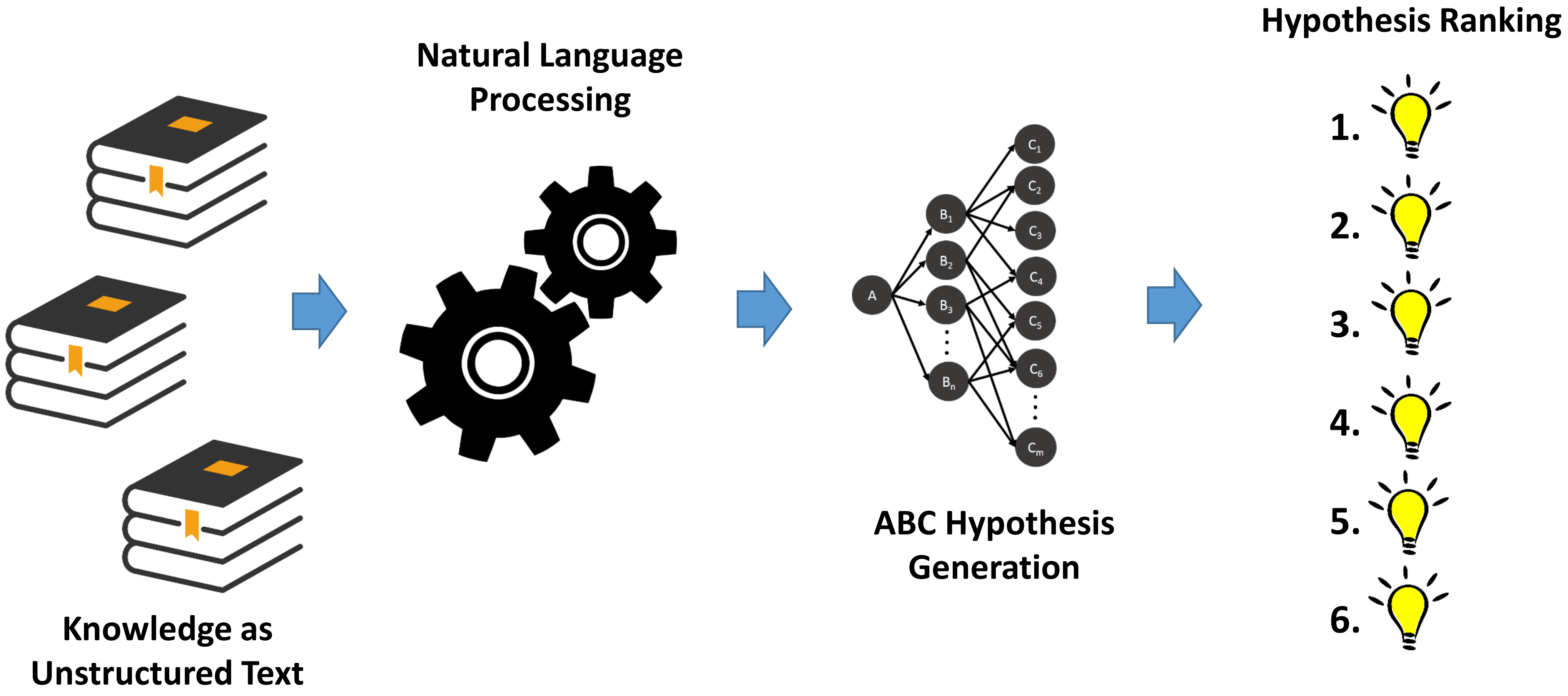


Therefore A implies C

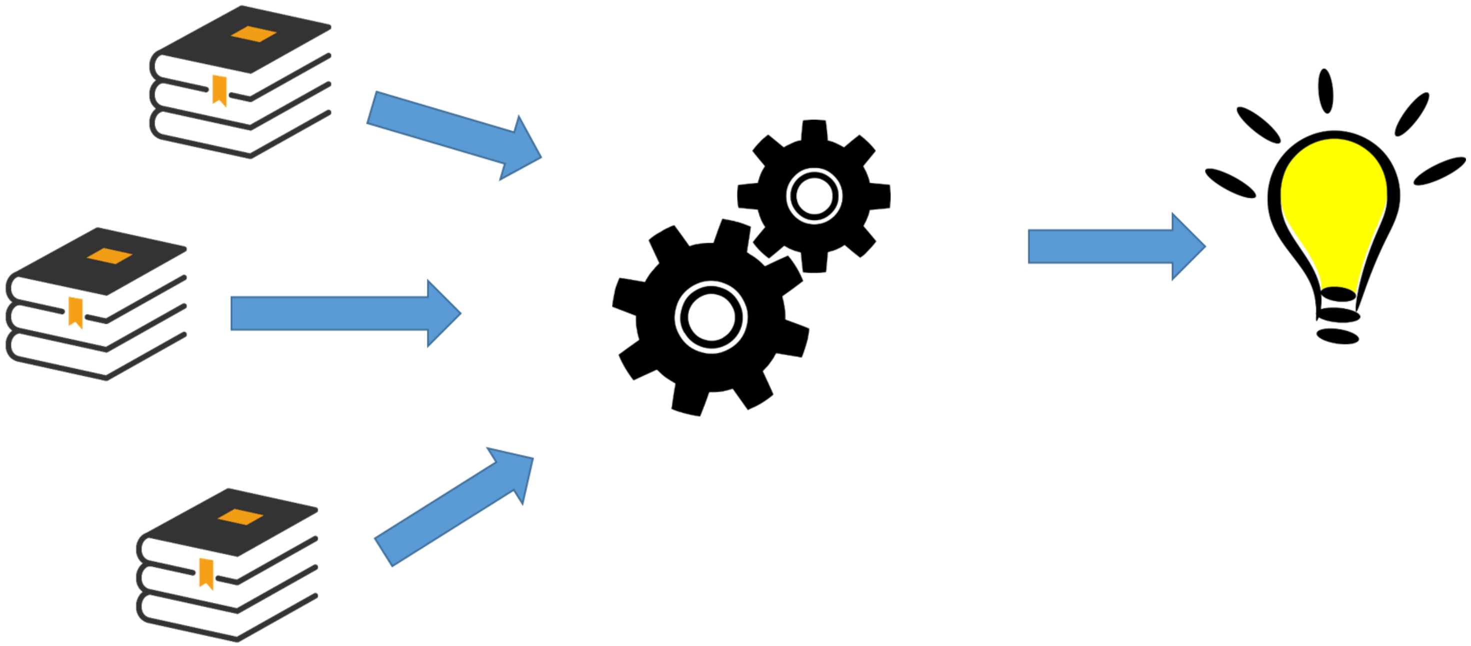
1986 – Fish Oil as a treatment for Raynaud's Disease[1]

[1]D. R. Swanson, Fish oil, raynaud's syndrome, and undiscovered public knowledge, Perspectives in biology and medicine 30 (1) (1986) 7–18

Baseline LBD System

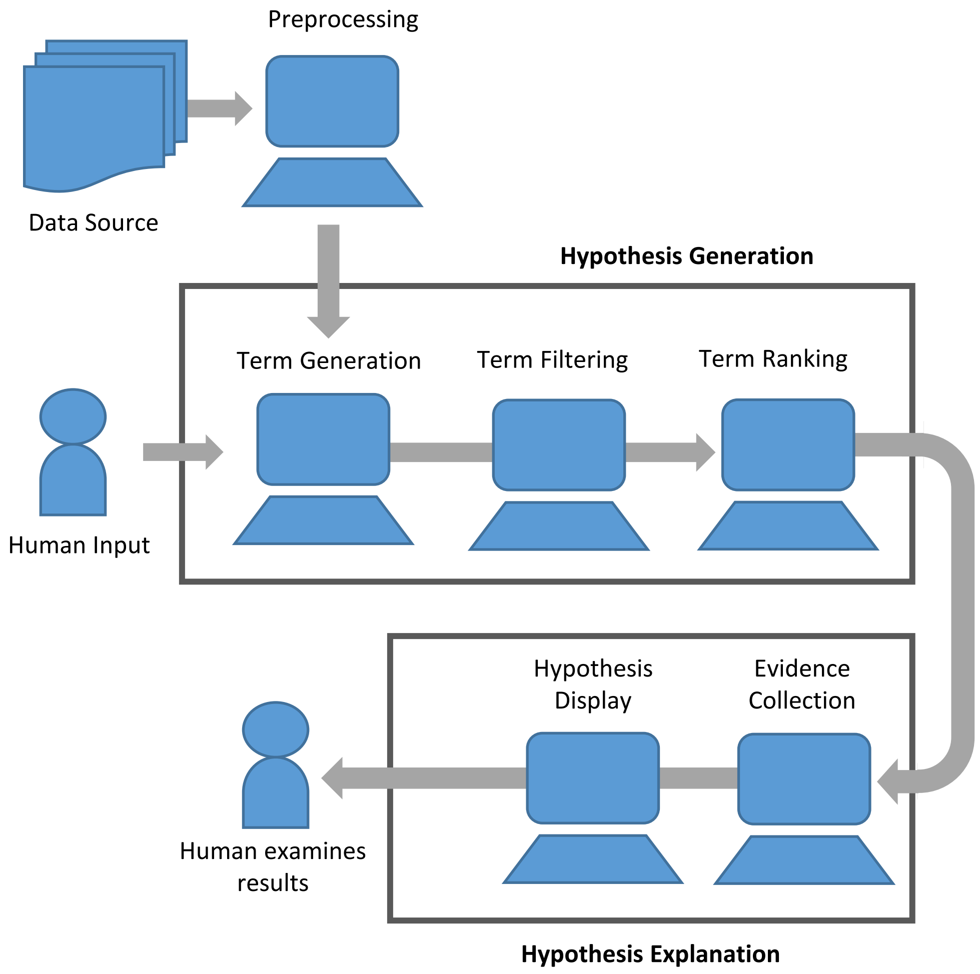


LBD Models and Components



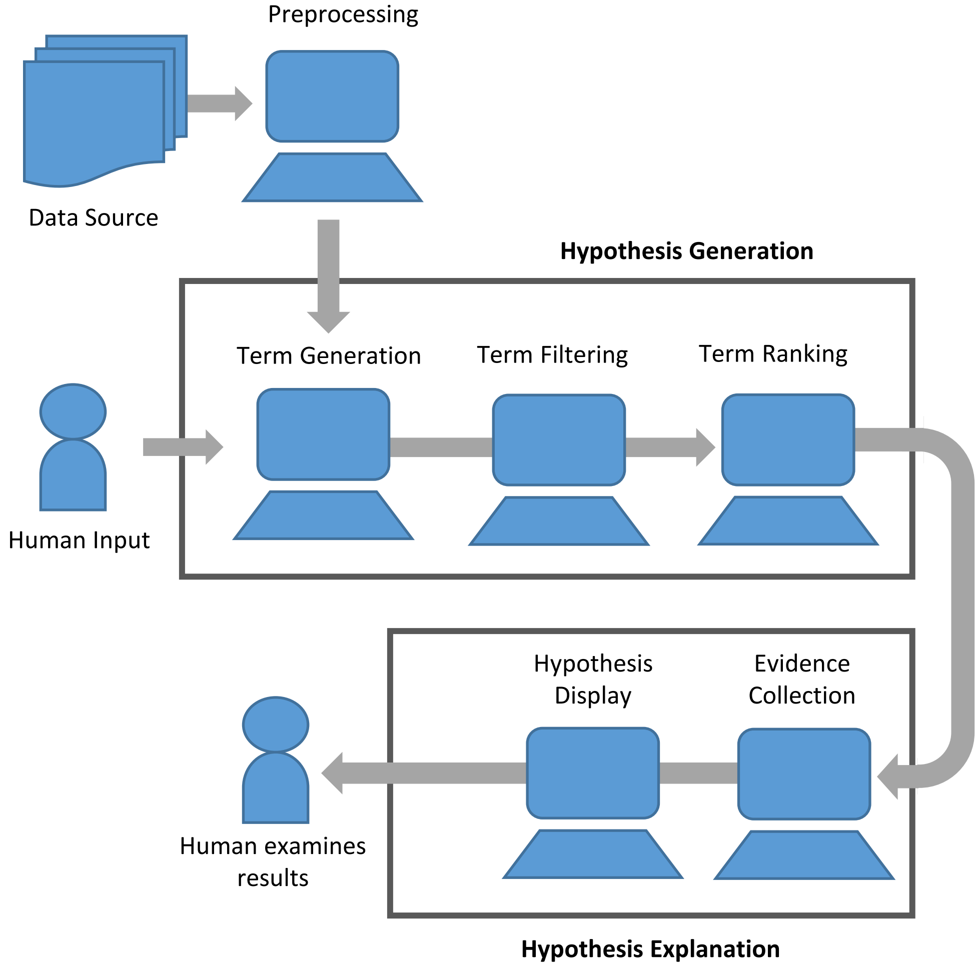
System Components

1. Data Source
2. Preprocessing
3. Hypothesis Generation
4. Hypothesis Explanation



System Components

1. Data Source
2. Preprocessing
3. Hypothesis Generation
4. Hypothesis Explanation



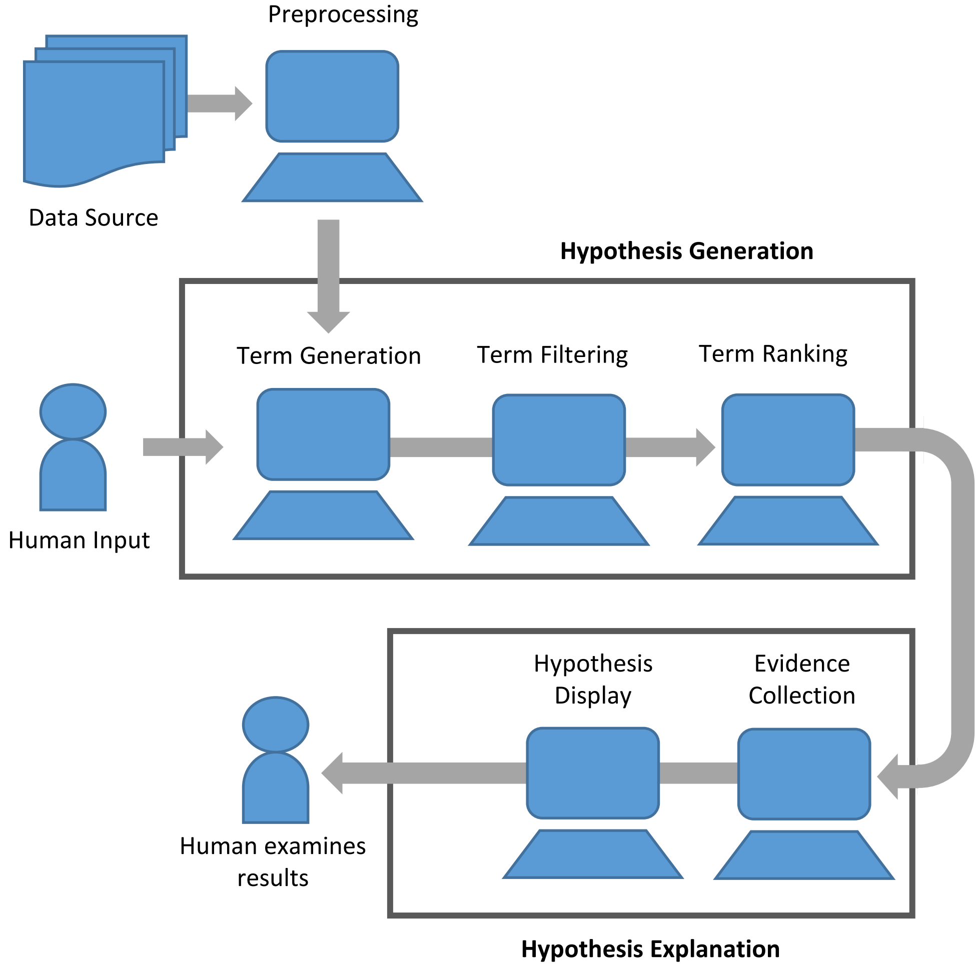
Data Source (Corpus)

- MEDLINE
  - A repository of biomedical publications
  - ~5,600 journals
  - > 22,775,609 citations from 1809 to present
    - 13,835,206 contain an abstract
  - We use 1975 onward
    - 2% of citations contained and abstract prior to that

...
</**JournalIssue**>
<Title>Federal register</Title>
<ISOAbbreviation>Fed Regist</ISOAbbreviation>
</Journal>
<**ArticleTitle**>Elimination of sanctions for refusal of vocational rehabilitation services without good cause. Final rule.</ArticleTitle>
<Pagination>
<MedlinePgn>40119-25</MedlinePgn>
</Pagination>
<Abstract>
<**AbstractText**>We are amending our regulations to remove provisions relating to the imposition of benefit sanctions on account of a beneficiary's refusal of rehabilitation services. We are making these changes to reflect the repeal of sections 222(b) and 1615(c) of the Social Security Act (the Act). Prior to their repeal, these sections of the Act authorized the Commissioner of Social Security to impose sanctions against the benefits of a disabled or blind beneficiary who refused, without good cause, to accept rehabilitation services made available by a State vocational rehabilitation (VR) agency. The Ticket to Work and Work Incentives Improvement Act of 1999 repealed these sections of the Act, effective January 1, 2001. We are amending our regulations by removing rules and related provisions that are obsolete as a result of the repeal of these sections of the Act to conform our regulations to the changes in the statute.</AbstractText>
</Abstract>
<**AuthorList CompleteYN="Y">
<Author ValidYN="Y">
...

System Components

1. Data Source
2. Preprocessing
3. Hypothesis Generation
4. Hypothesis Explanation



Preprocessing

- Convert data from its raw form to a form accepted by the hypothesis generation step of LBD.
- It is often tightly coupled with the data source,
- Preprocessing steps:
  - Identify the terms:
    - Stop word removal
    - Text normalization
    - Named entity recognition
  - Identify the explicit relationships
    - Collecting co-occurrence information
    - Relation extraction.

Unified Medical Language System (UMLS)

UMLS

- Concept Hierarchy of Biomedical Terms

CUI

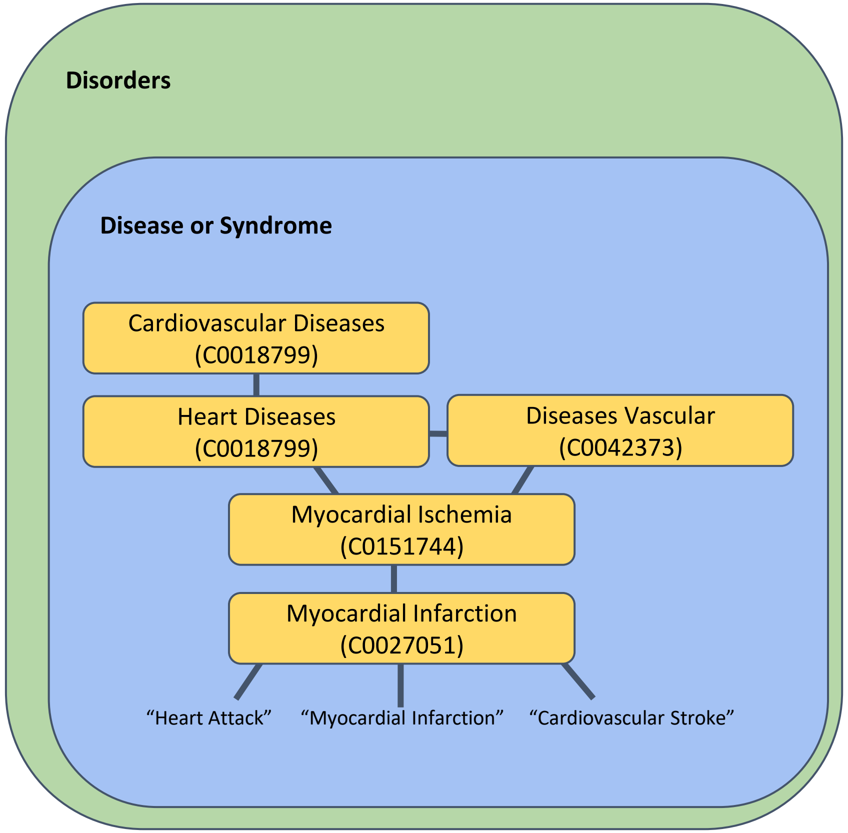
- Concepts unique identifier

Semantic Type

- Broad sets of terms

Semantic Group

- Even broader sets of concepts



Text Processing Tools to Identify Terms

“Raynaud’s Disease changes blood viscosity”

- **Compoundify** – identifies compound words in text

  raynauds_disease changes blood_viscosity

- **MetaMap** – maps text to CUIs

  C0034734 changes C0005848

- **MedaCy –** NER system to identify entities

  [Raynaud’s Disease: Disease] changes [blood viscosity: Symptom]

Text Processing Tools to Identify Relations

“Raynaud’s Disease changes blood viscosity”

- **Text::NSP** – extracts co-occurrence information

  raynauds_disease co-occurs blood_viscosity

- **SemRep** – extracts CUI Relation CUI triplets from text

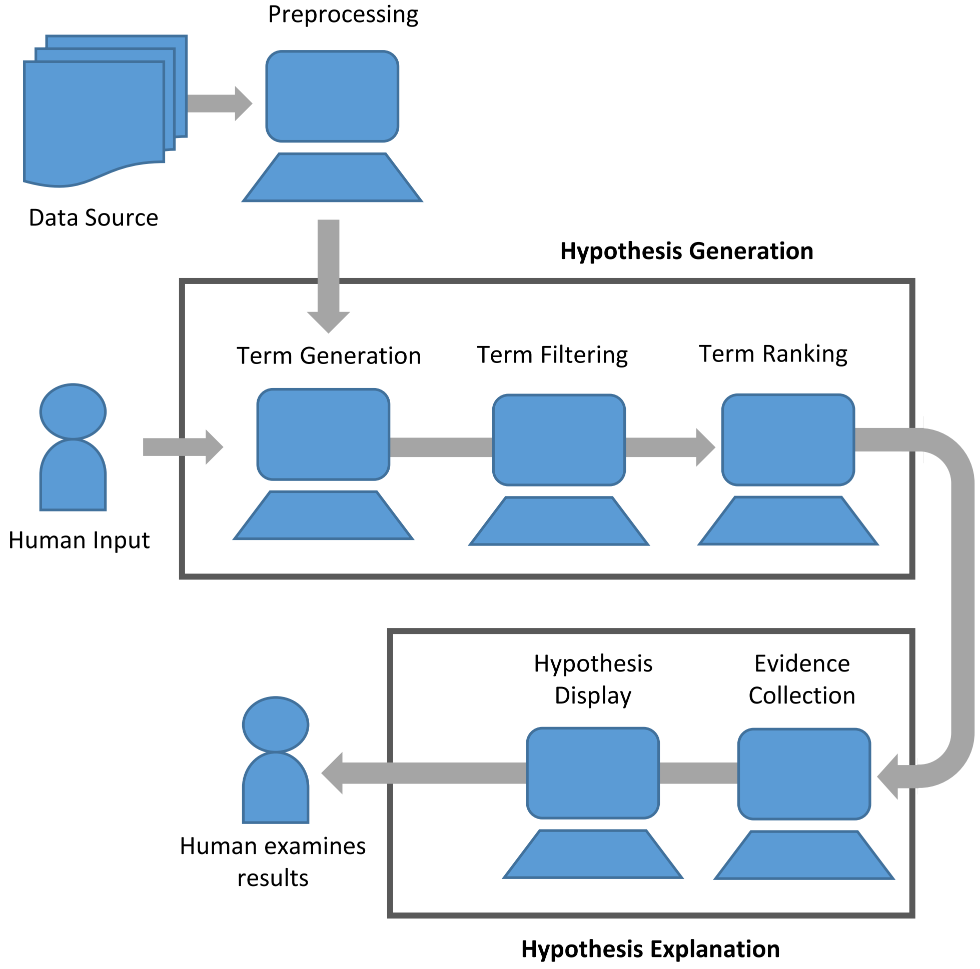
  C0034734 AFFECTS C0005848

- **ReLex** – relation extraction system to identify relations

  [Raynauds Disease: Disease] causes [blood viscosity: Symptom]

System Components

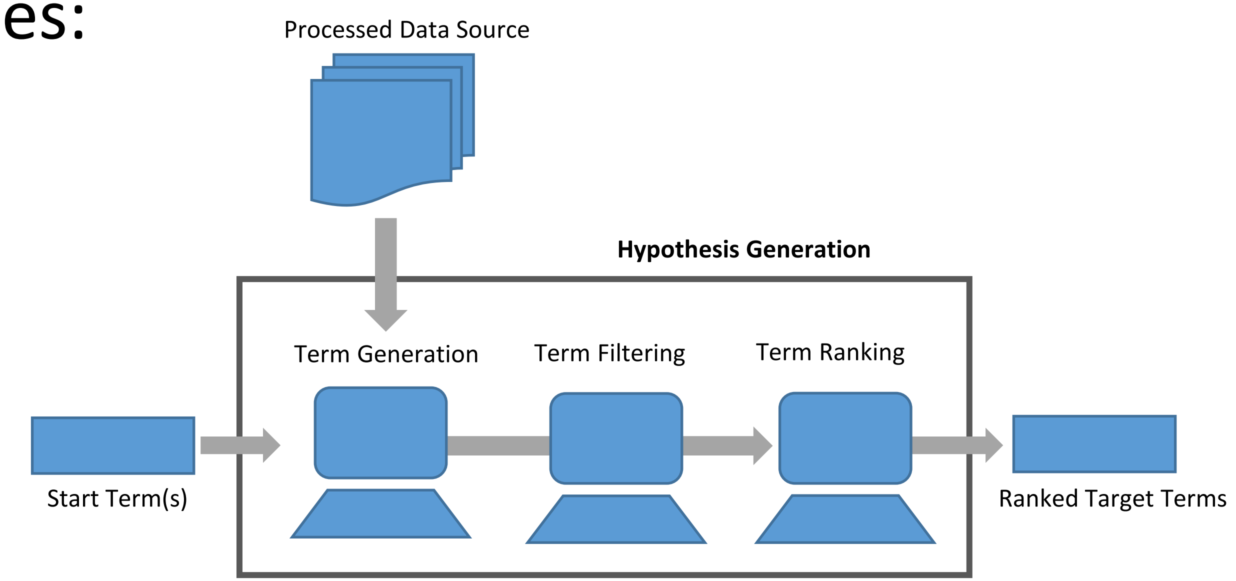
1. Data Source
2. Preprocessing
3. Hypothesis Generation
4. Hypothesis Explanation



Hypothesis Generation

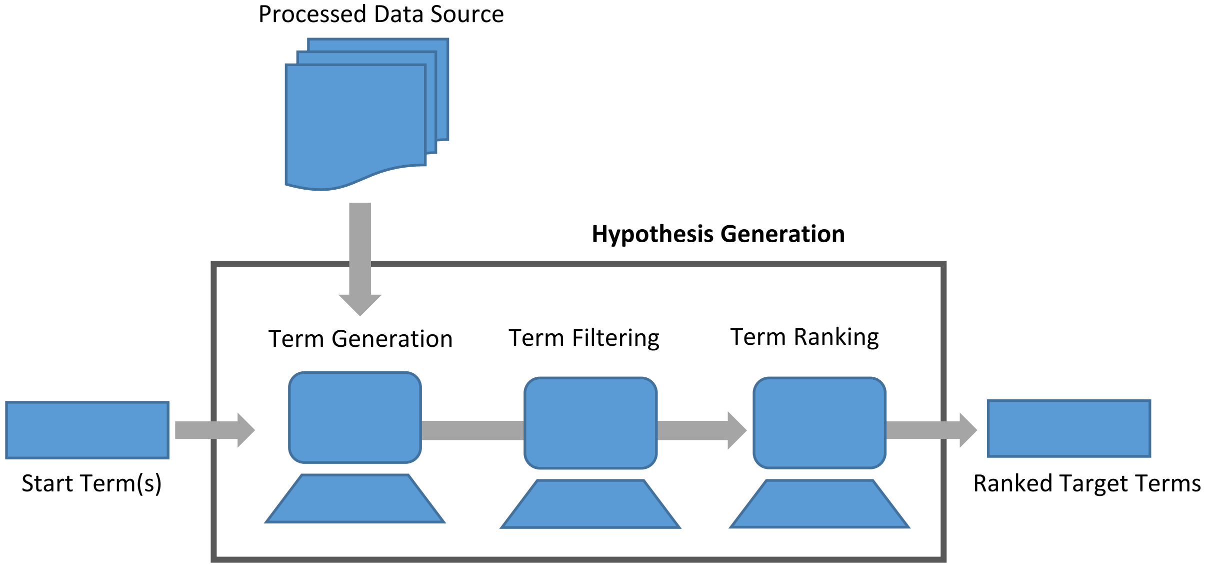
The whole process of generating, filtering, and ranking hypotheses:

1. Term Generation
2. Term Filtering
3. Term Ranking



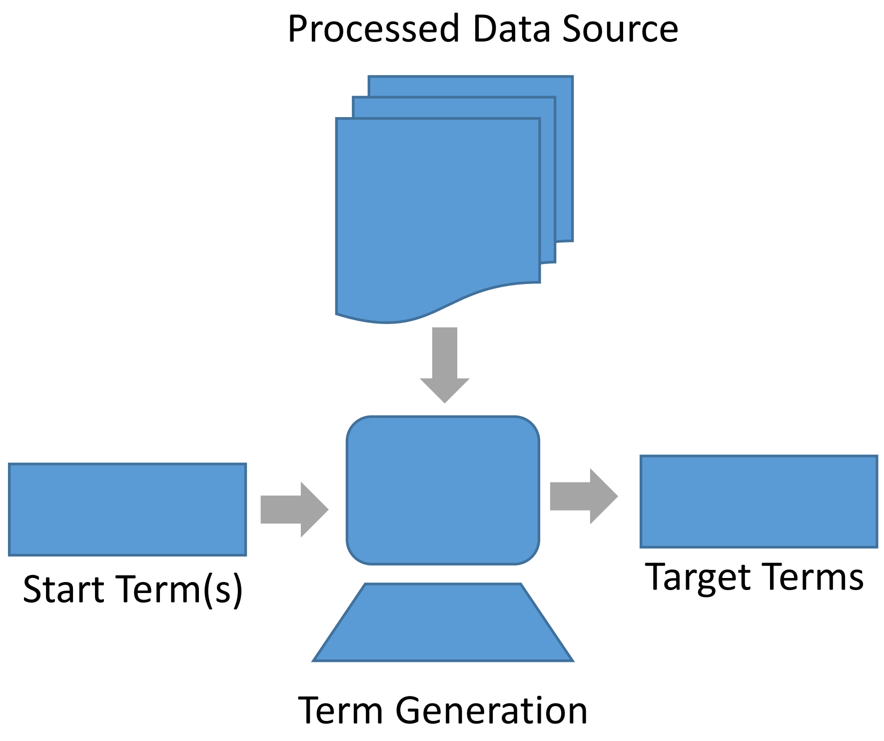
Hypothesis Generation

1. Term Generation
2. Term Filtering
3. Term Ranking



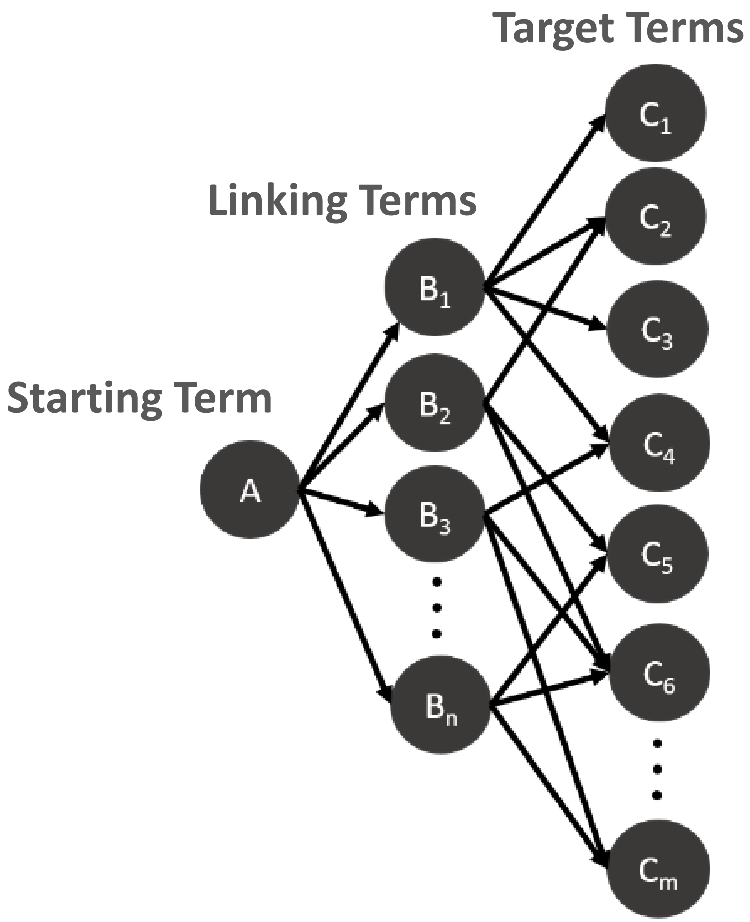
Term Generation

- Creates potential hypotheses
- Typically very noisy
- Examples include:
  - ABC model
  - Discovery patterns
  - Vector-based nearest neighbor searches
  - Discovery by analogy
  - Bibliometric linking
  - User interaction
  - Returning all terms in the vocabulary



ABC Hypothesis Generation

- A implies B, B implies C, therefore A implies C
- Relationships as:
  - Co-occurrence
  - Extracted relationships
- Limitations:
  - Co-occurrence isn't necessarily a relationship
  - Relation extraction misses relationships
  - ABCD, ABCDE?
  - Information Explosion

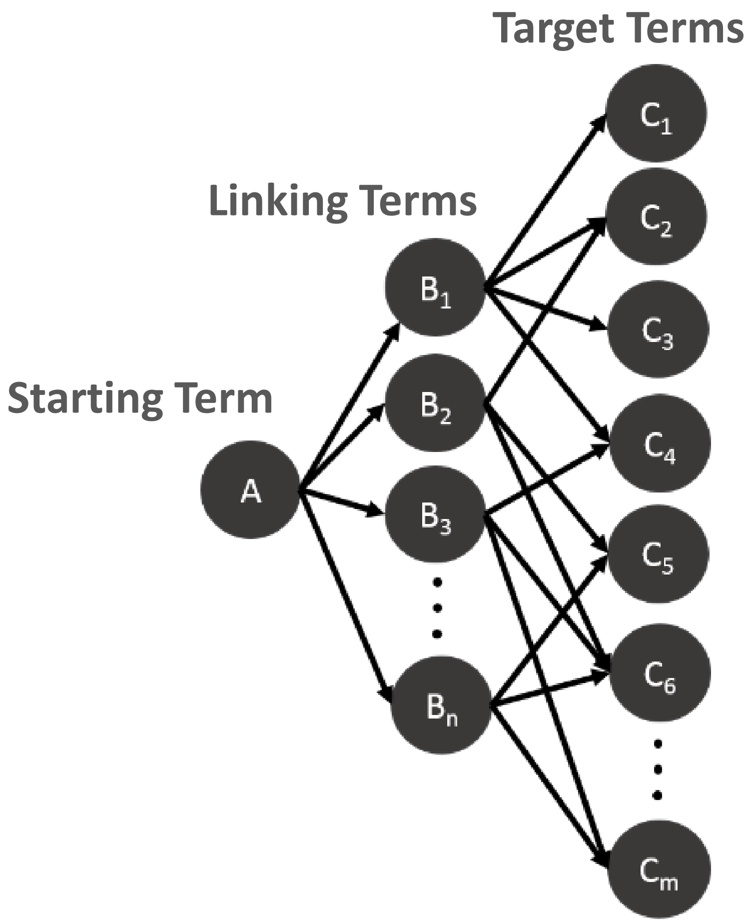


Hypothesis Generation Examples

- ABC Co-occurrence
- Discovery Patterns
- Discovery by Analogy
- Discovery Browsing

ABC Co-occurrence

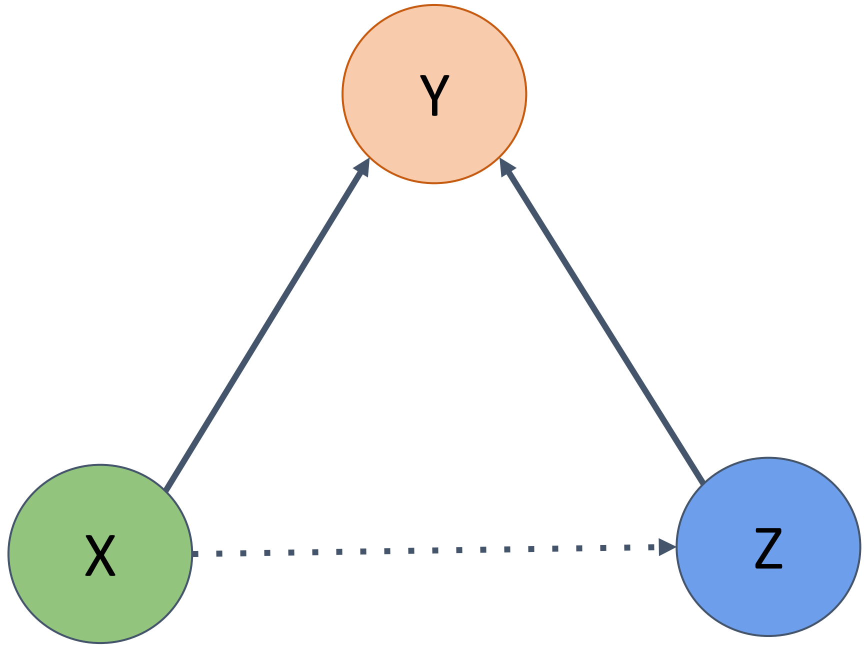
- A co-occurs with B
- B co-occurs with C
- Find all A-B-C terms



Discovery Patterns

Find all Relationships such that:

- Z is a drug
- X is Raynaud's Disease
- X-Y is a stimulates relation
- Y-Z is an disrupts relation
- And X has no relation to Z



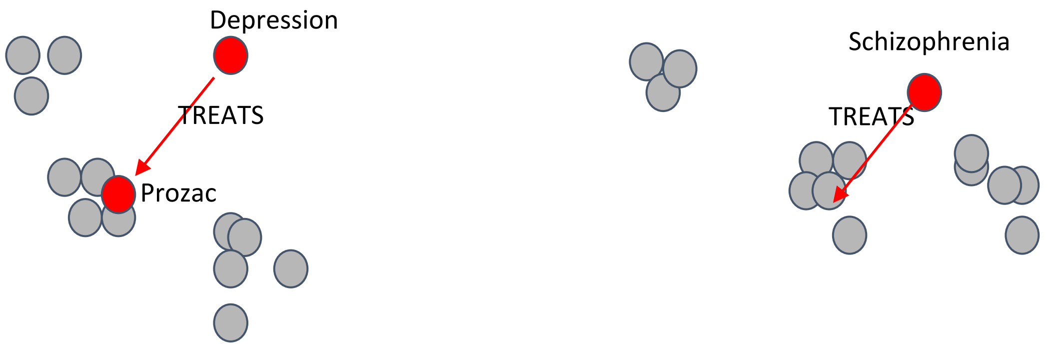
Raynaud's Disease

Discovery By Analogy

Why?

"Prozac is to Depression as **?** is to Schizophrenia"

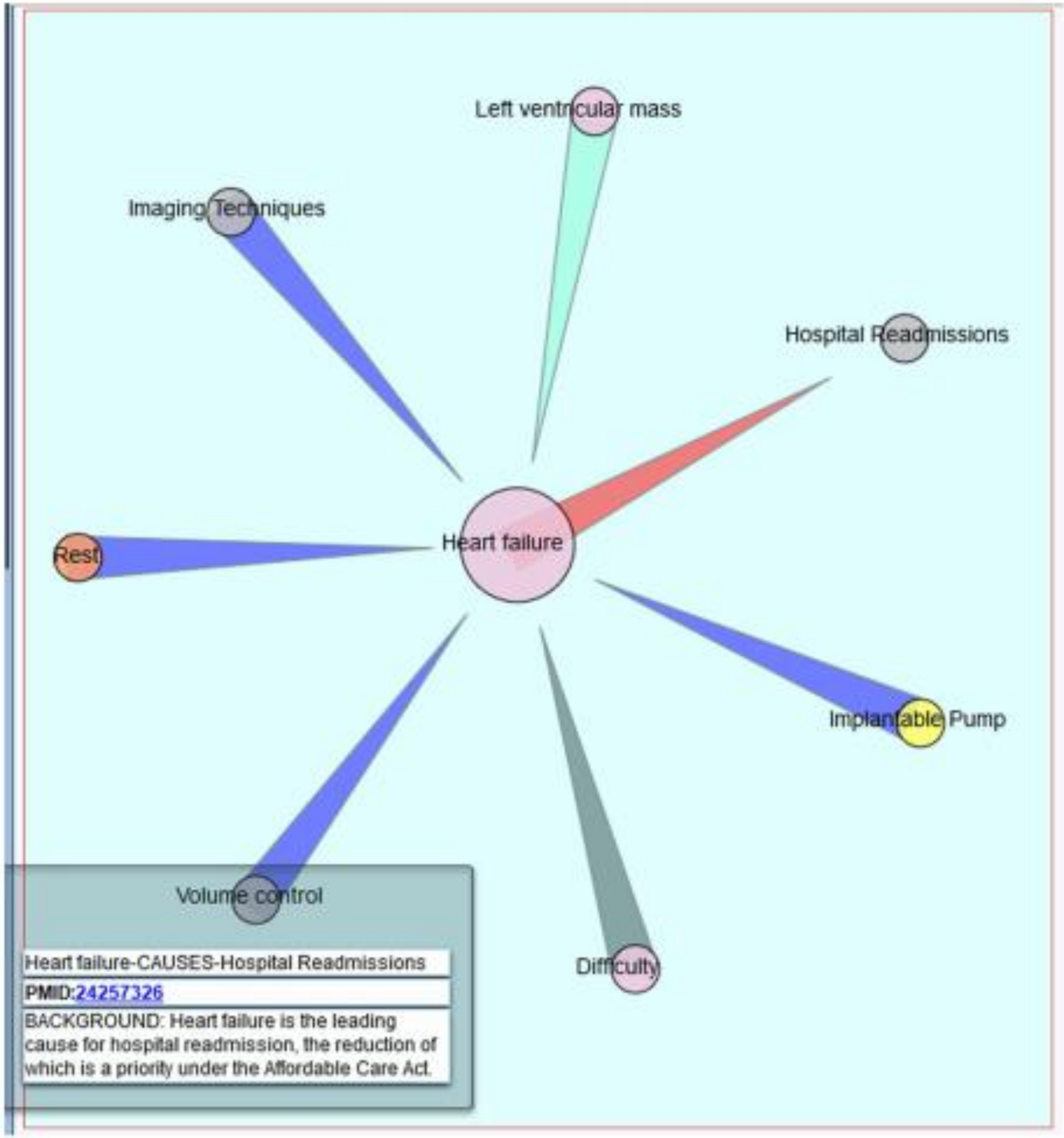
Using concept vectors, relation vectors, and vector operations we can arrive at a conclusion



"Finding Schizophrenia's Prozac Emergent Relational Similarity in Predication Space" - Cohen - 2011

Discovery Browsing

- Spark
  - A framework for 'Serendipitous Knowledge Discovery'



Workman, T. Elizabeth, et al. "Spark, an application based on Serendipitous Knowledge Discovery." *Journal of biomedical informatics* 60 (2016): 23-37.

Discovery Browsing

- Semantic MEDLINE

Kilicoglu, Halil, et al. "Semantic MEDLINE: a web application for managing the results of PubMed Searches." *Proceedings of the third international symposium for semantic mining in biomedicine*. Vol. 2008. 2008.

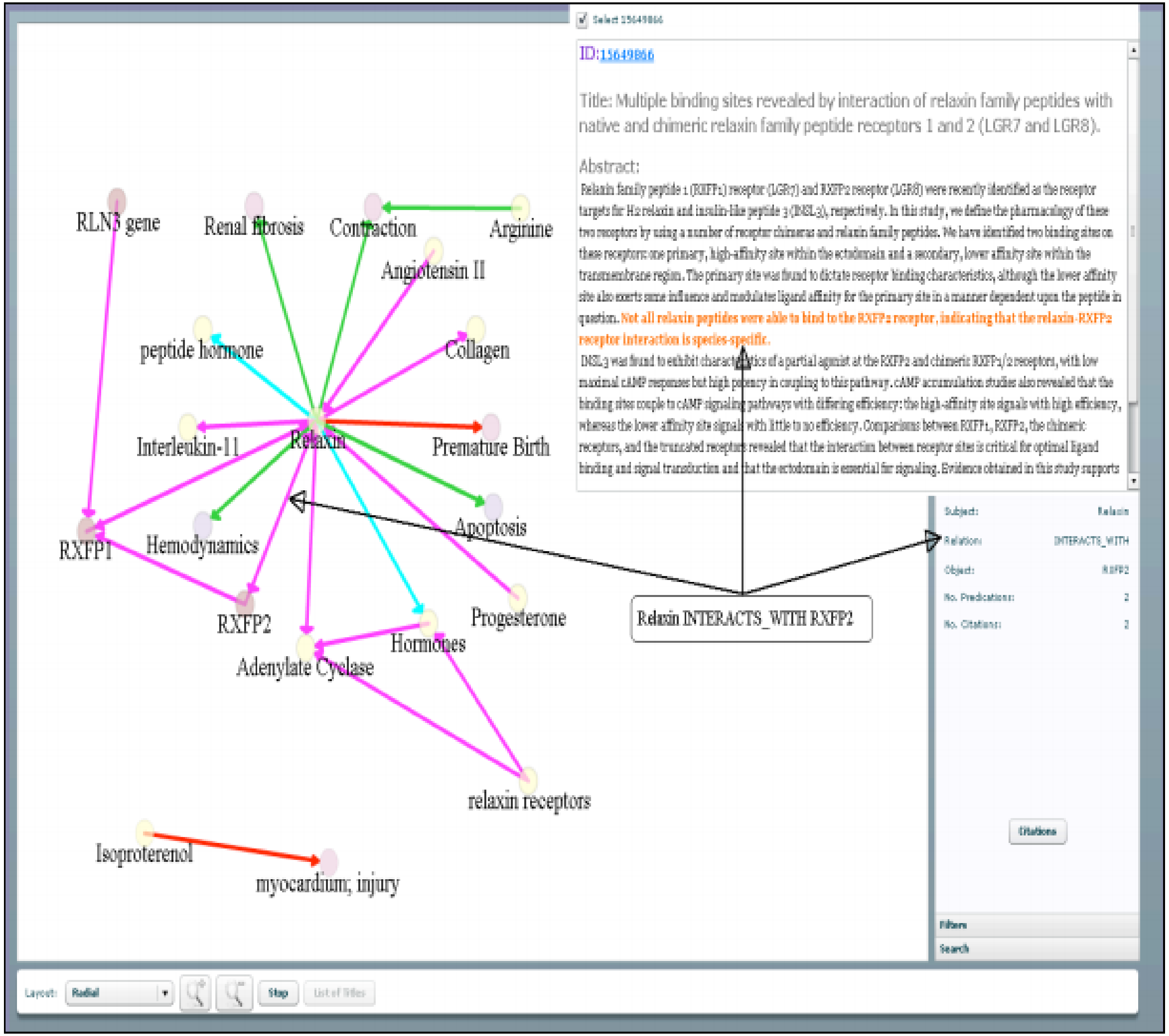
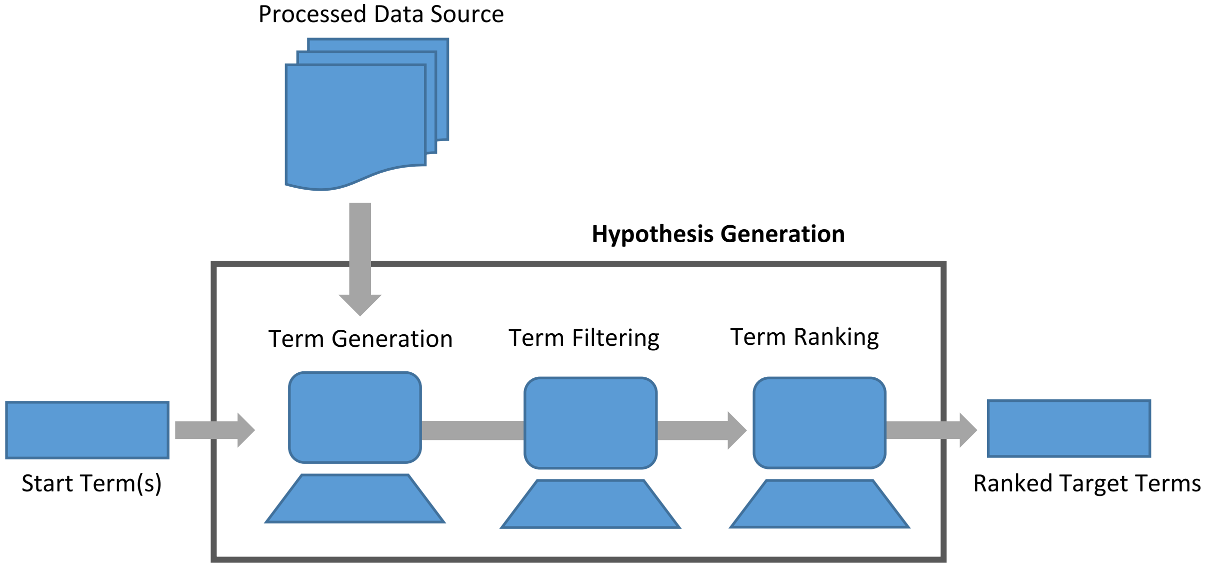


Fig. 3. Visualizing summarization results for Relaxin search, with Relaxin INTERACTS_WITH RXFP2 relation highlighted.

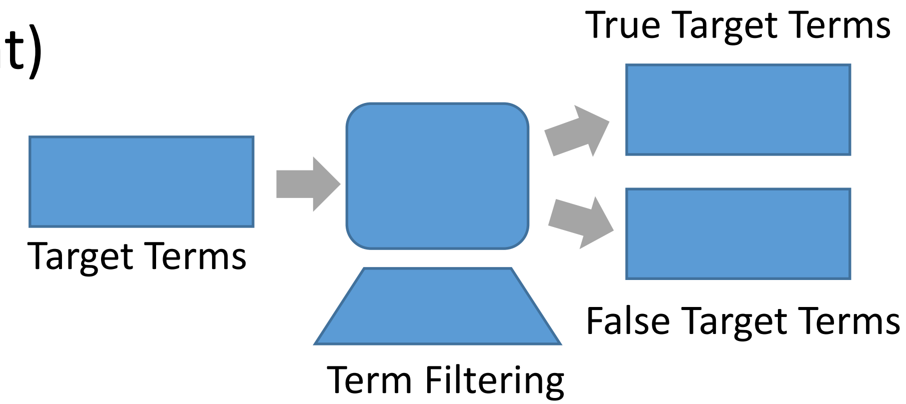
Hypothesis Generation

1. Term Generation
2. Term Filtering
3. Term Ranking



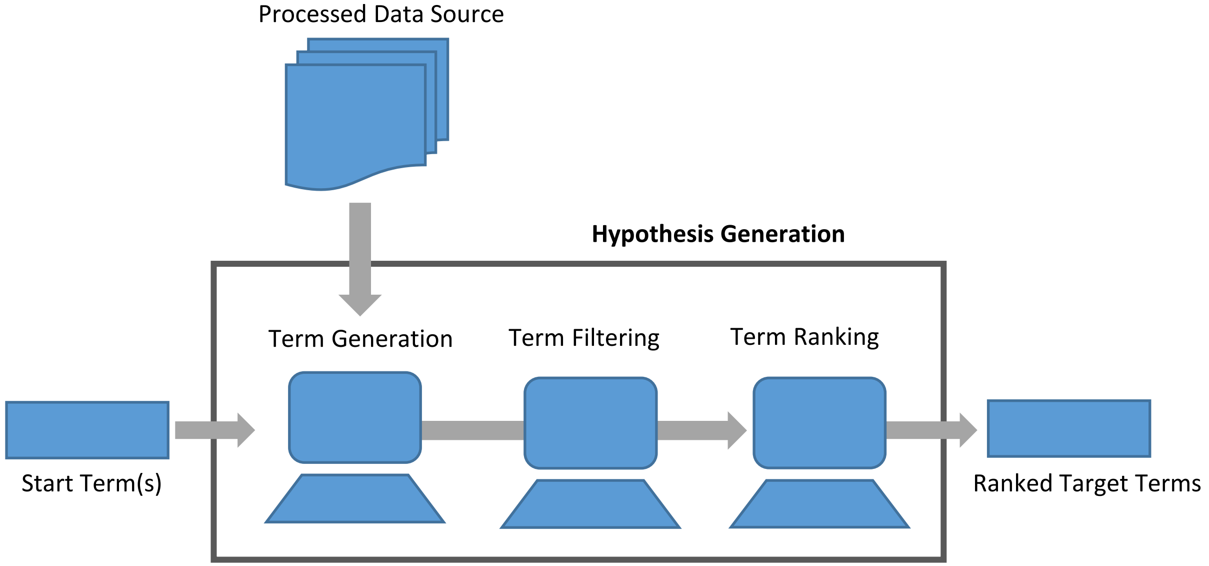
Term Filtering

- Term generation typically over-generates target terms
- Term filtering removes uninteresting or untrue terms
- Examples:
  - Term occurrence rate (too frequent or infrequent)
  - UMLS hierarchy to remove terms that are:
    - too broad or too similar to the start term
    - Not the desired semantic type
  - Information retrieval metrics and thresholds
    - Term Frequency-Inverse Document Frequency



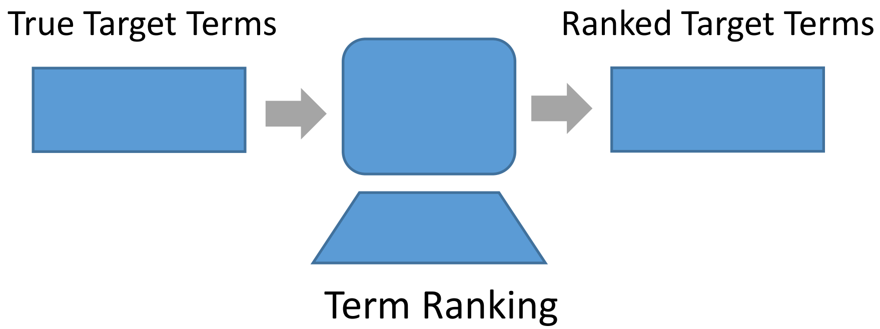
Hypothesis Generation

1. Term Generation
2. Term Filtering
3. Term Ranking



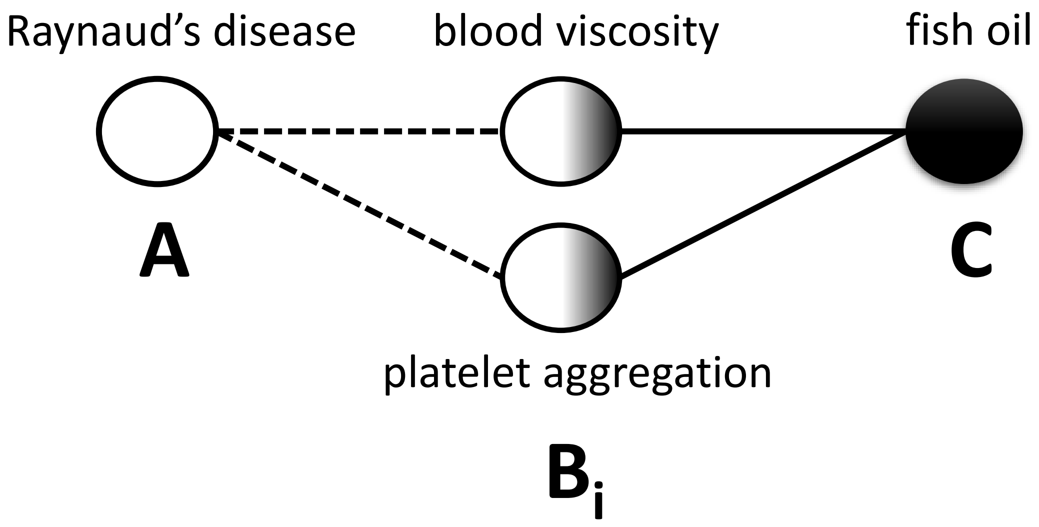
Term Ranking

- Ranks hypotheses based on their interestingness
- Too many Target Terms:
  - **51,931** target terms when replicating Raynaud's Disease – Fish Oil discovery
  - Small world problem
- Information Retrieval ranks don't work
  - Rely on direct co-occurrences



Method 1: Linking Term Count (LTC)[1]

- The best performing target term ranking measure
- The count of unique shared linking terms



count (B) = 2 = **LTC**

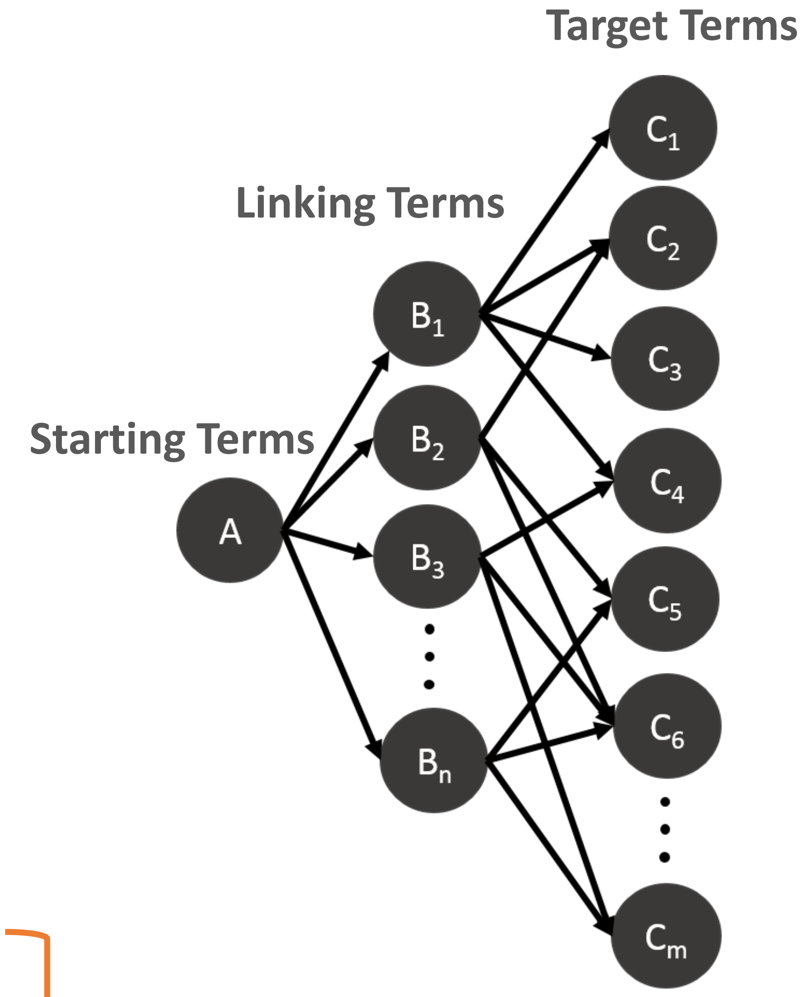
[1]Swanson, Don R., and Neil R. Smalheiser. "An interactive system for finding complementary literatures: a stimulus to scientific discovery." *Artificial intelligence* 91.2 (1997): 183-203.

Term Ranking Methods

Term Co-occurrence	
Gordon and Lindsay [51]	Relative Frequency*
Hristovski, et al. [89]	Confidence*
Hristovski, et al. [54]	Support
Swanson, et al. [90]	Literature Cohesiveness (COH)
Cole and Bruza [71]	Odds-Ratio
Stegmann and Grohmann [91]	Equivalence Index
Measures of Independence	
Yetisgen-Yildiz and Pratt [53, 88]	Z-Score
Wren, et al. [87]	Mutual Information Measure (MIM)
Cole and Bruza [71]	Log Likelihood (ll)
Semantic Predication	
Hristovski, et al. [92]	Predication Frequency
Wilkowski, et al. [74]	Degree Centrality
Cameron, et al. [93]	Intra-Cluster Predication Similarity
Nearest Neighbor Search	
Gordon and Dumais [64]	Cosine Distance
Bruza, et al. [66]	Euclidean Distance
Bruza, et al. [66]	Information Flow
Implicit Term	
Hristovski, et al. [89]	$X \rightarrow Z$ Support
Wren, et al. [87]	Average Mutual Information Measure (AMIM)
Wren, et al. [87]	Minimum Mutual Information Measure (MMIM)
Wren, et al. [87]	Average Minimum Weight (AMW)
Swanson and Smalheiser [48]	Linking Term Count (LTC)
Yetisgen-Yildiz and Pratt [88]	Linking Term Count-Average Minimum Weight (LTC-AMW)
Rastegar, et al. [14]	Predicate Independence/Interdependence

Designed for Ranking Implicit Terms

* Confidence and relative frequency are equivalent



Association Metrics

Collect Co-occurrences

traditional

Populate Contingency Table
(observed values)

	stop	¬ stop	totals
smoking	2955	75020	77975
¬ smoking	308792	2712312165	2712620957
totals	311747	2712387185	2712698932

Calculate Association Measure

$$G^2 = 2 * \sum_{i,j=1}^{2} n_{ij} * log(\frac{n_{ij}}{m_{ij}})$$

Association Score

Contingency Table of Observed Values

n_{11} = count of “stop smoking”

n_{1p} = count of “stop <anything>”

n_{p1} = count of “<anything> smoking”

n_{pp} = count of “<anything> <anything>”

- Other values can be computed from these four

  n_{12} = “stop <anything but smoking>”
  = $n_{1p} - n_{11}$

  n_{p2} = “<anything> <anything but smoking>”
  = $n_{pp} - n_{p1}$

  etc..

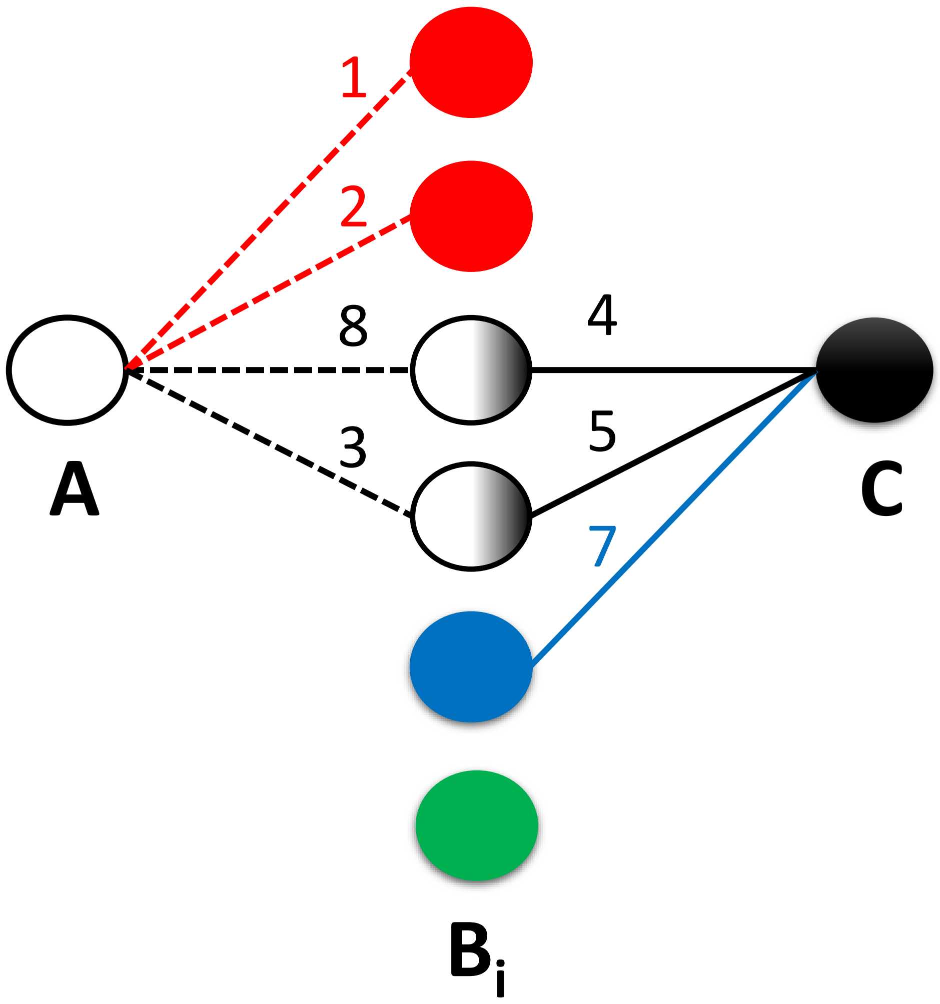
	stop	¬ stop	totals
smoking	2955	75020	77975
¬ smoking	308792	2712312165	2712620957
totals	311747	2712387185	2712698932

	Y	$\overline{Y}$	totals
X	$n_{11} = XY$	$n_{12} = X\overline{Y}$	$n_{1p} = X*$
$\overline{X}$	$n_{21} = \overline{X}Y$	$n_{22} = \overline{XY}$	$n_{2p} = \overline{X}*$
totals	$n_{p1} = *Y$	$n_{p2} = *\overline{Y}$	$n_{pp} = **$

Indirect Ranking Measures

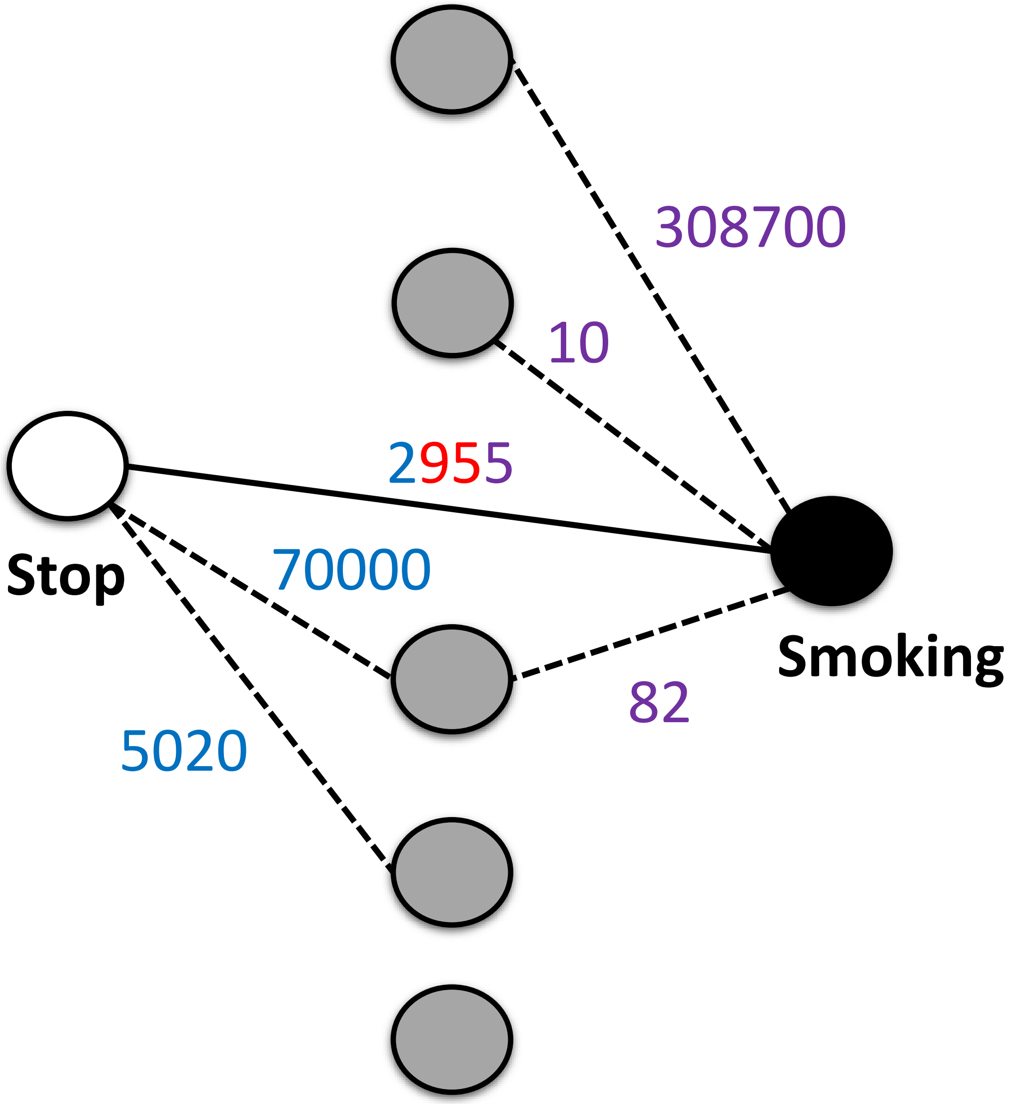
They don't take into account the whole picture

- Terms that only A co-occurs with
- Terms that only C co-occurs with
- Terms that co-occur with neither



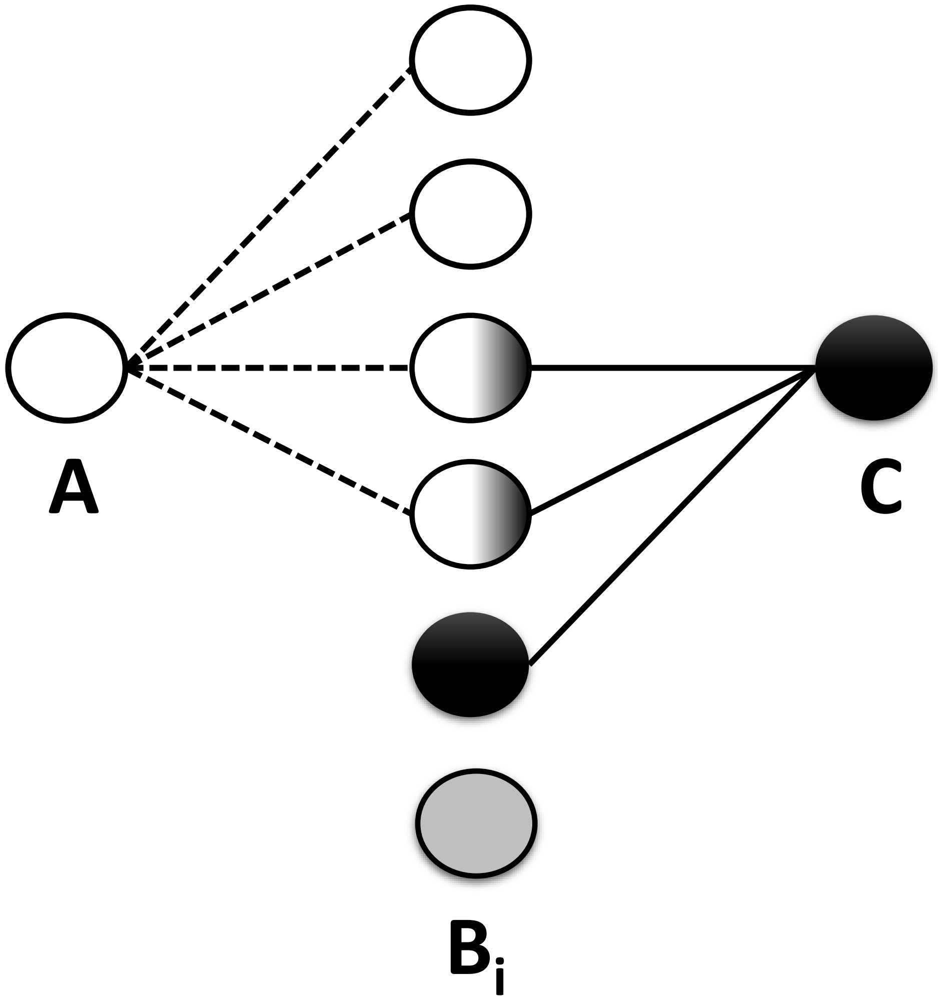
Contingency Table to Co-occurrence Graph

	stop	¬ stop	totals
smoking	2955	75020	77975
¬ smoking	308792	2712312165	2712620957
totals	311747	2712387185	2712698932



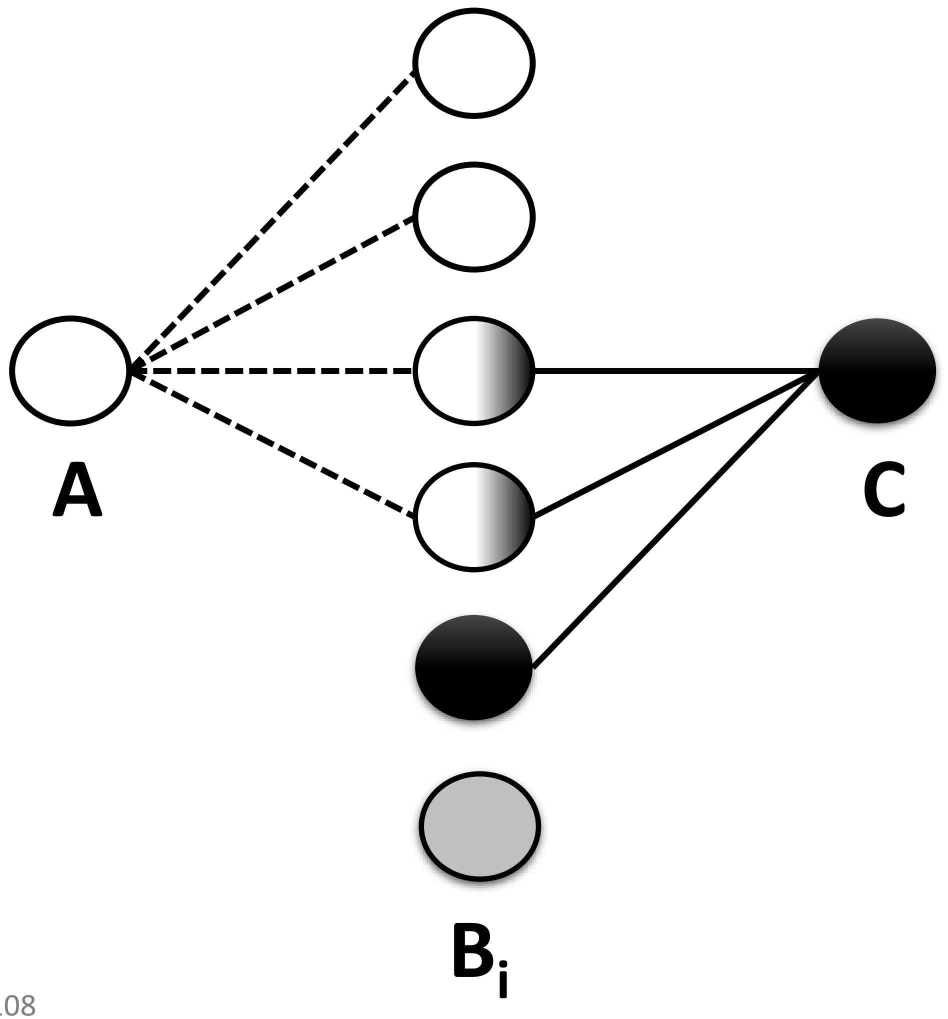
Collect co-occurrence information on implicit relations

- Linking term association
- Minimum weight association

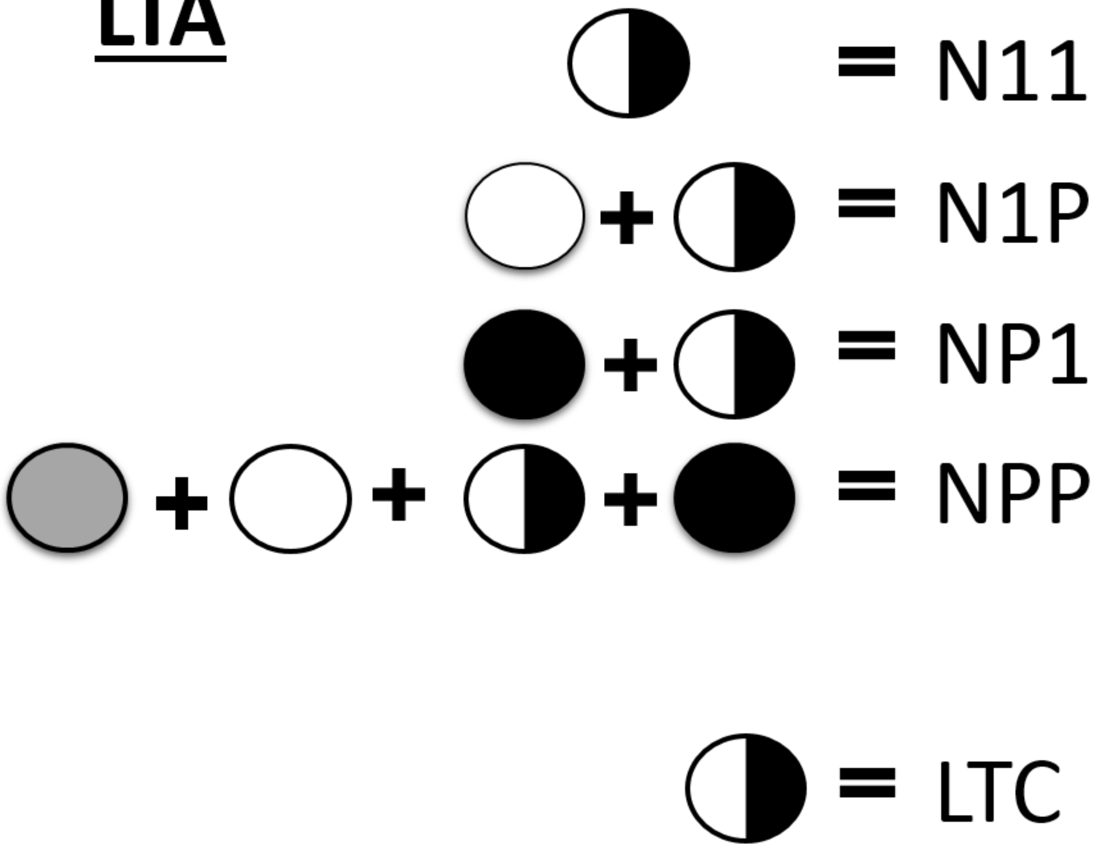


Metric 1: Linking Term Association (LTA)

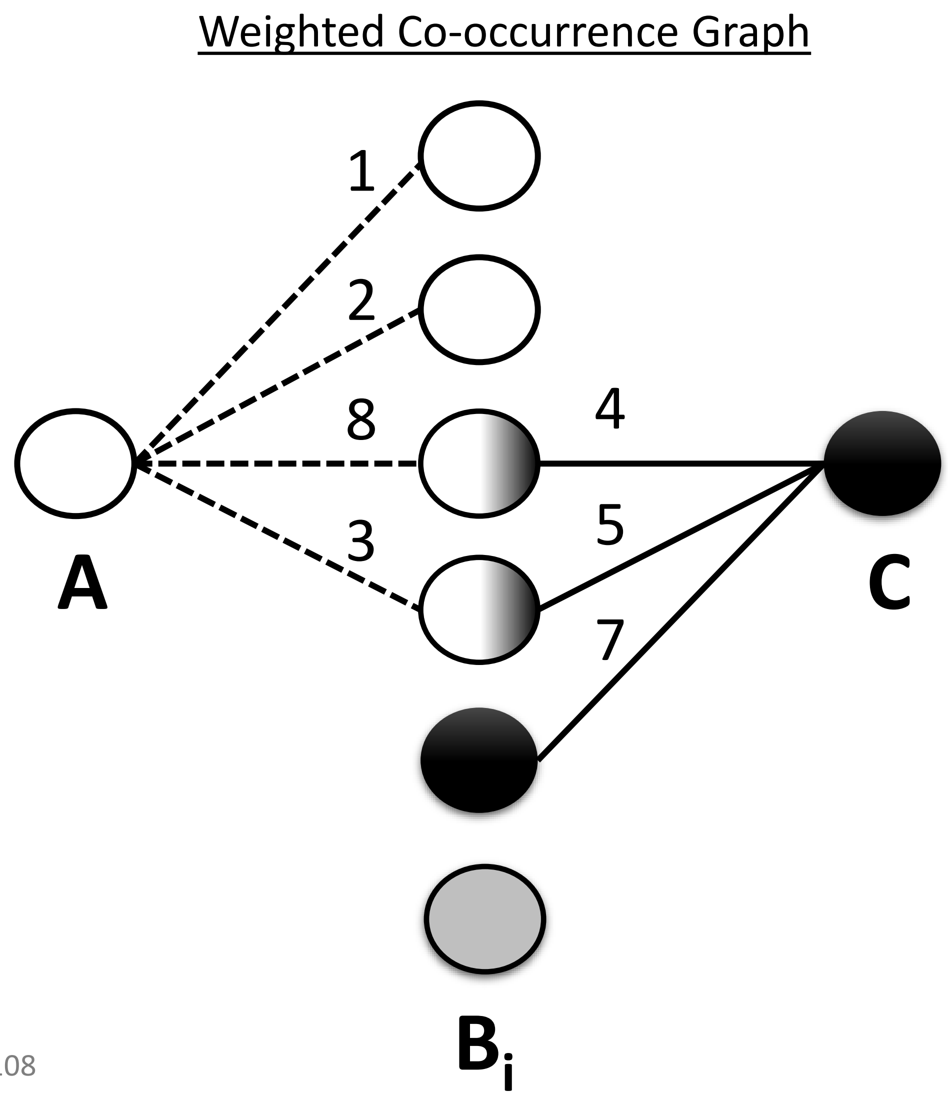
Unweighted Co-occurrence Graph



LTA



Metric 2: Minimum Weight Association (MWA)



MWA

$$\min(8,4) + \min(3,5) = N11$$

$$1 + 2 + 8 = N1P$$

$$4 + 5 + 7 = NP1$$

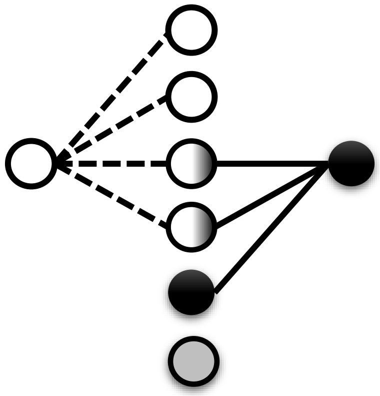
$$|C| = NPP$$

Association Metrics

Collect Co-occurrences

traditional

LTA or MWA



Populate Contingency Table (observed values)

	stop	¬ stop	totals
smoking	2955	75020	77975
¬ smoking	308792	2712312165	2712620957
totals	311747	2712387185	2712698932

Calculate Expected Values

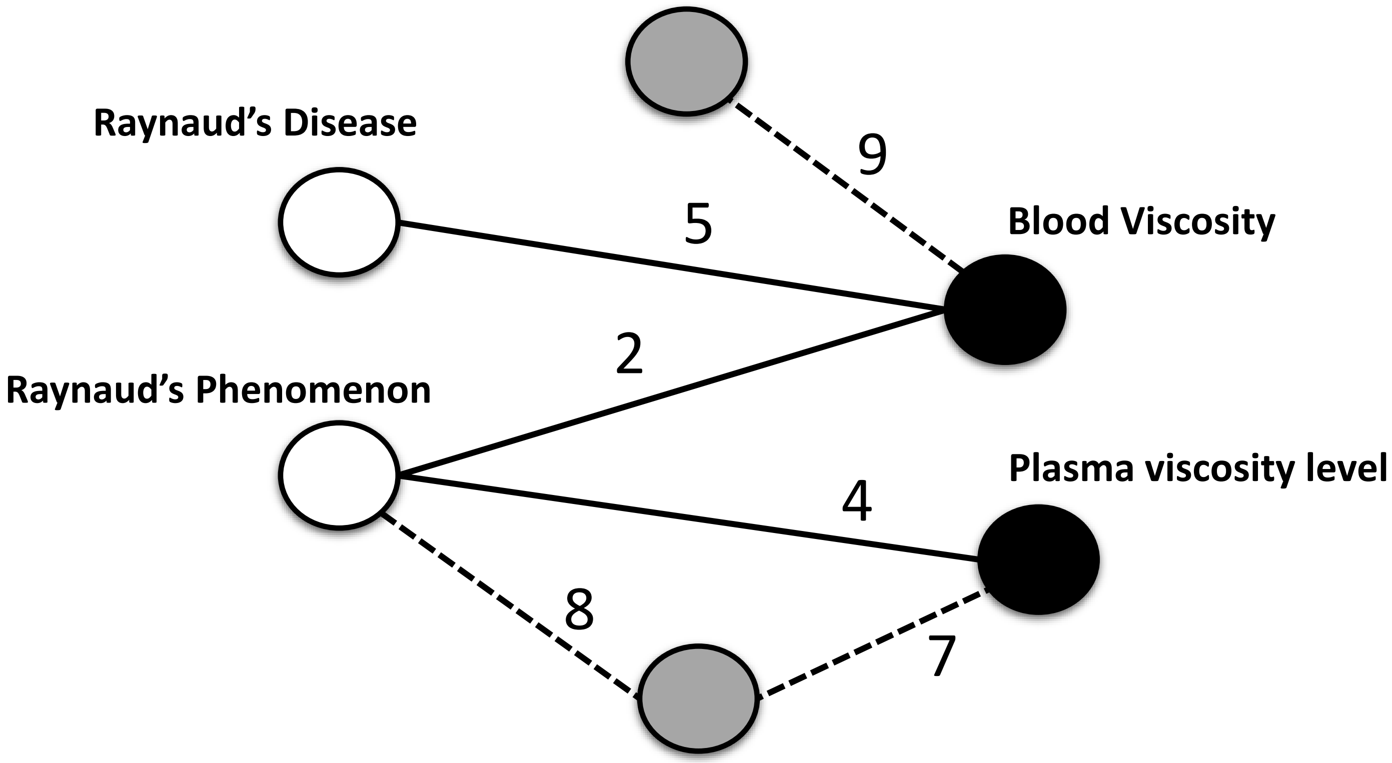
	stop	¬ stop	totals
smoking	9	77966	77975
¬ smoking	311738	2712309219	2712620957
totals	311747	2712387185	2712698932

Calculate Association Measure

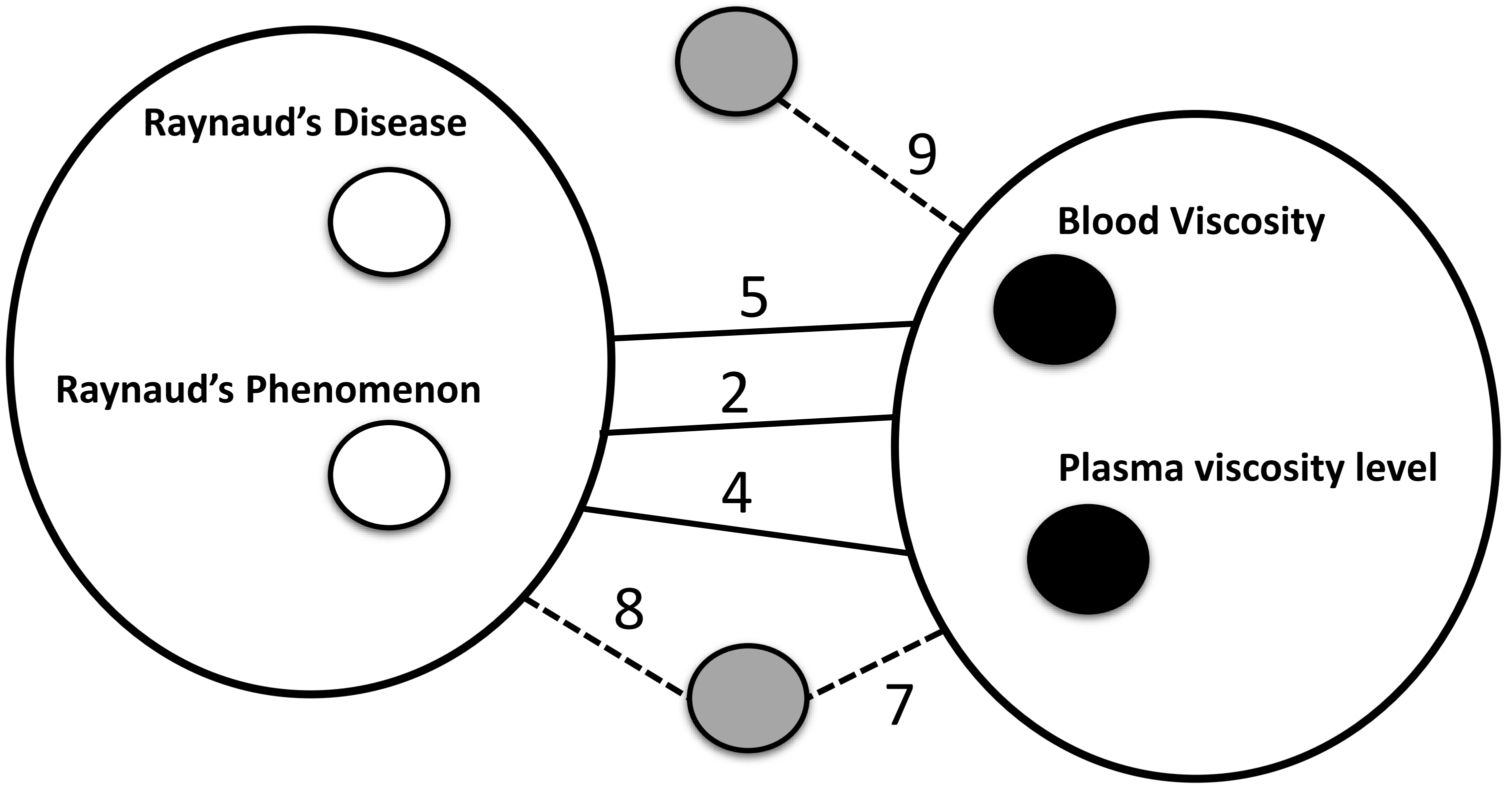
$$G^2 = 2 * \sum_{i,j=1}^{2} n_{ij} * log(\frac{n_{ij}}{m_{ij}})$$

Association Score

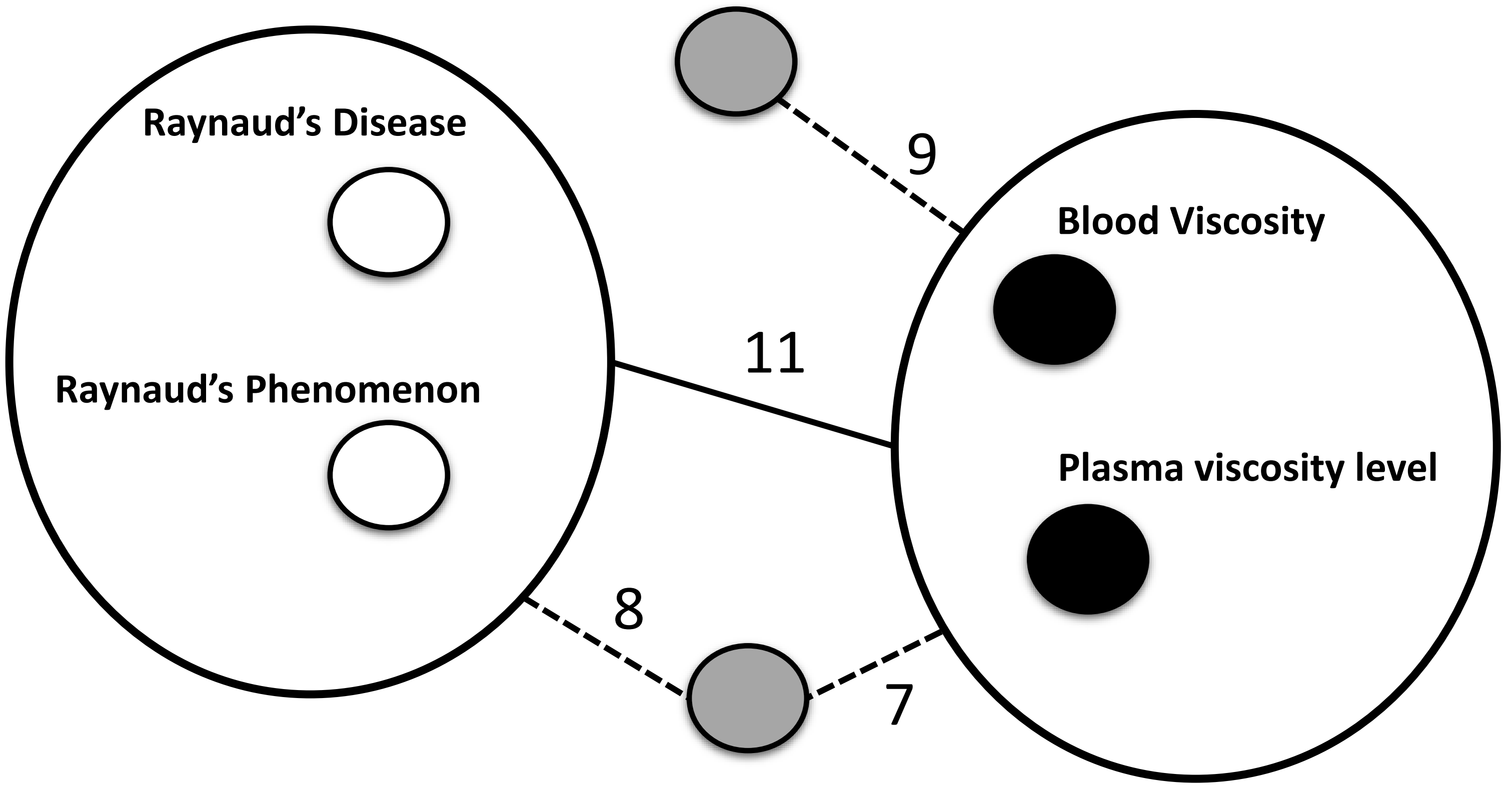
Set Associations



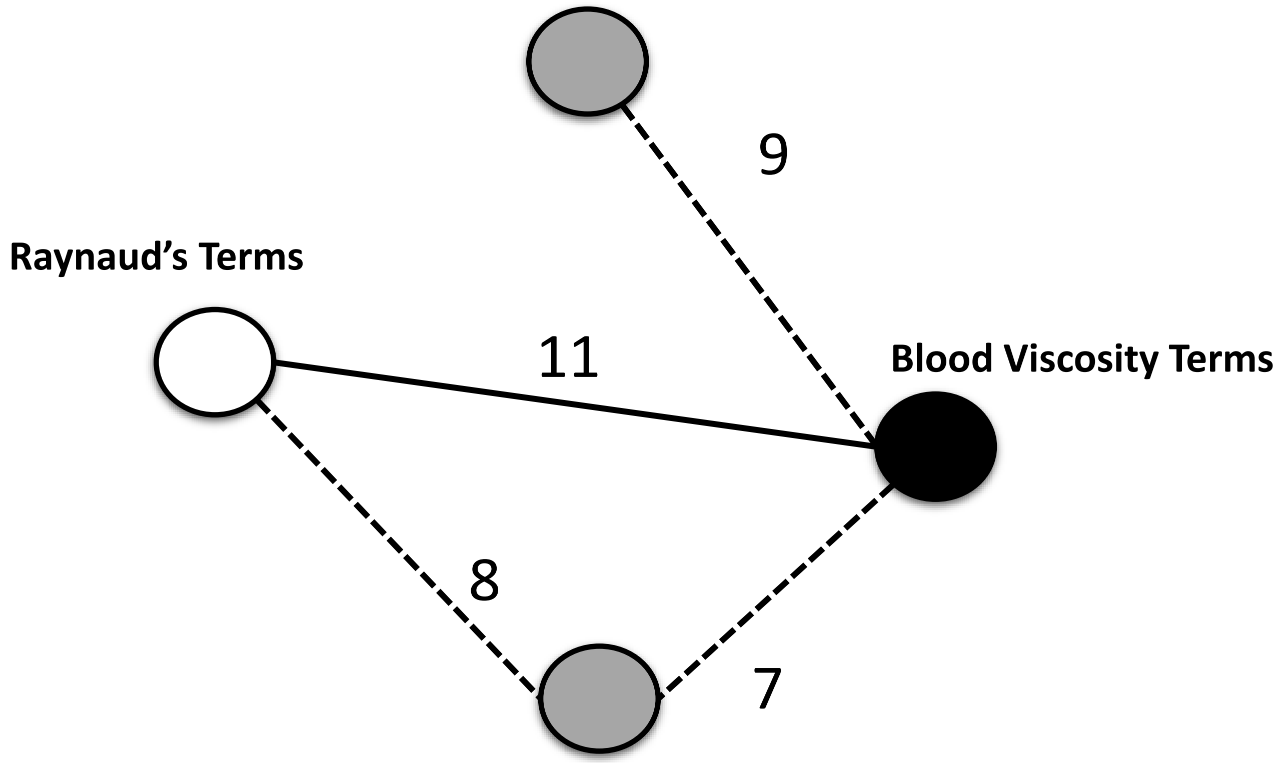
Set Associations



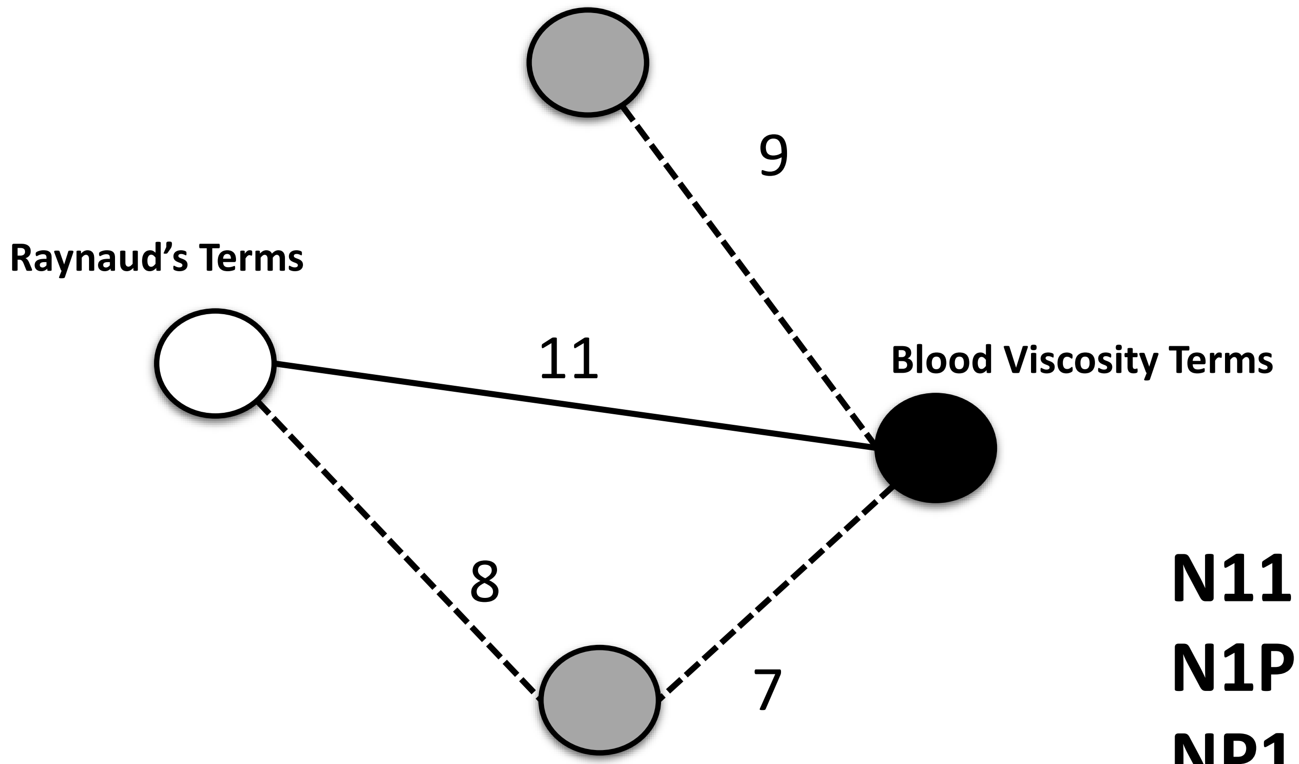
Set Associations



Set Associations



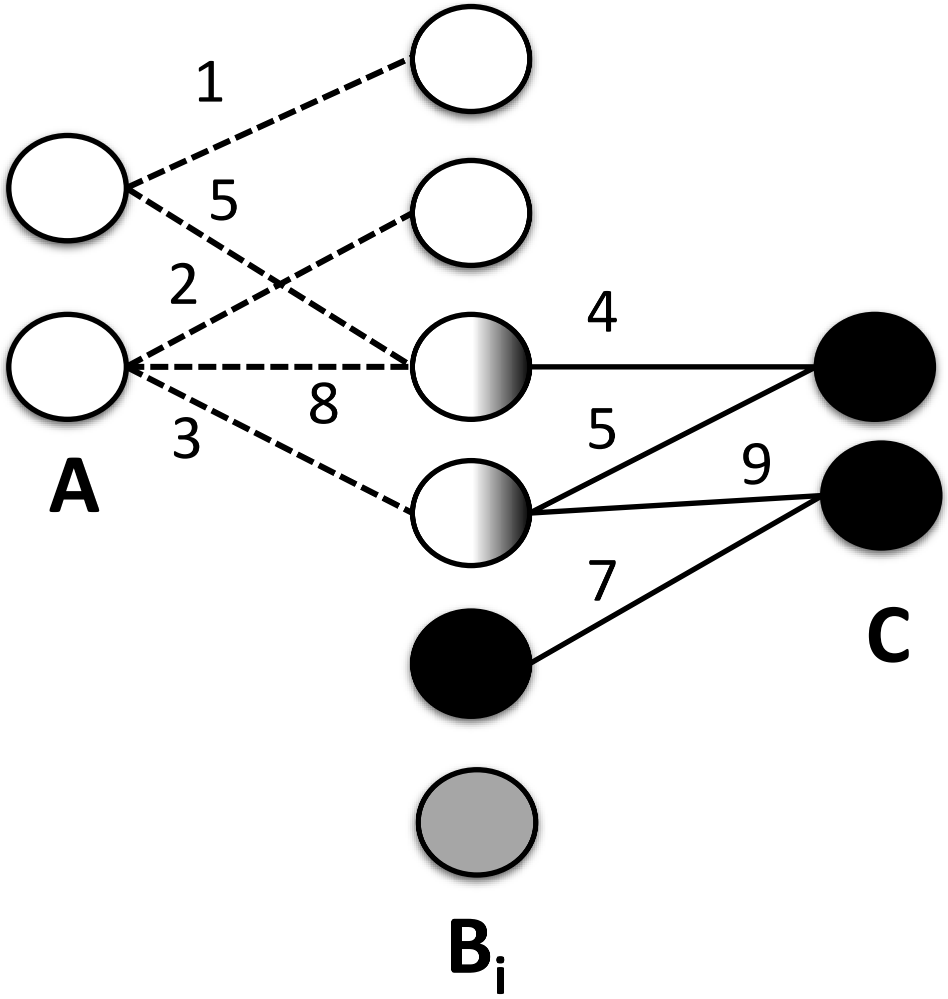
Set Associations



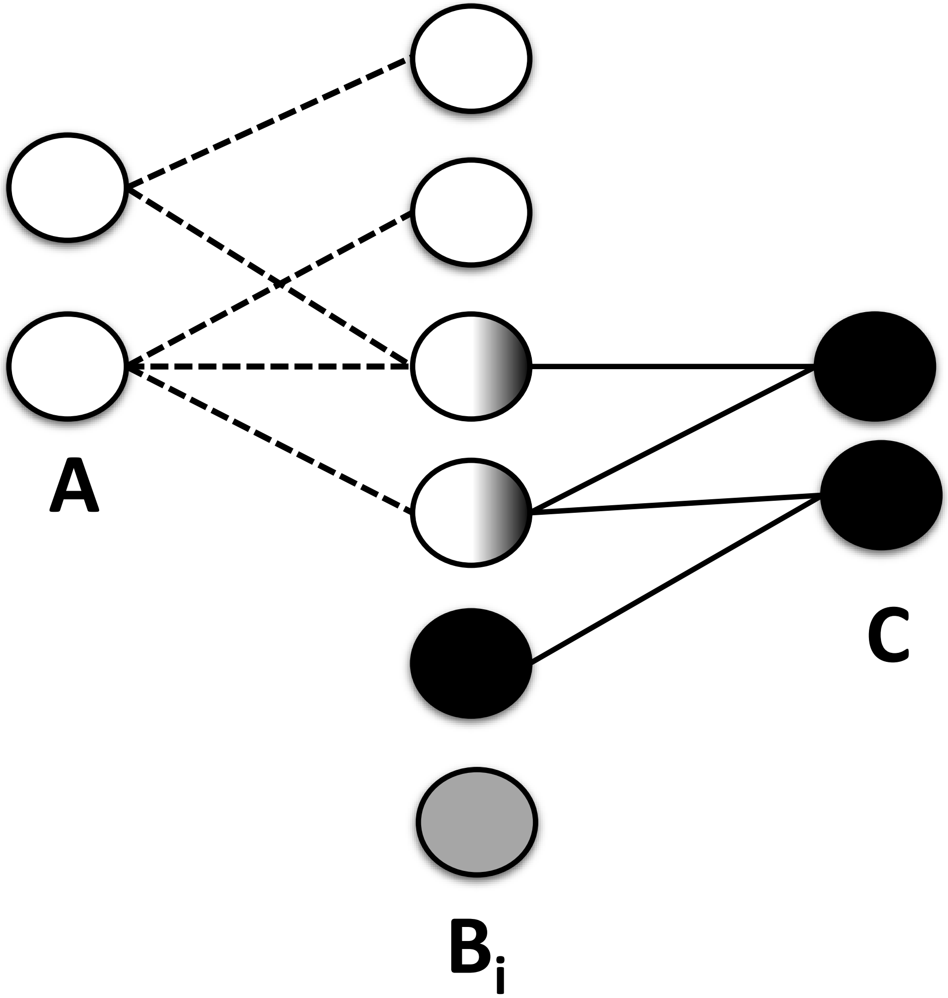
N11 = 11
N1P = 8
NP1 = 9 + 7
NPP = |C|

N11 = sum of A and B N11s
N1P = sum of A N1Ps
NP1 = sum of B NP1s
NPP = all co-occurrences

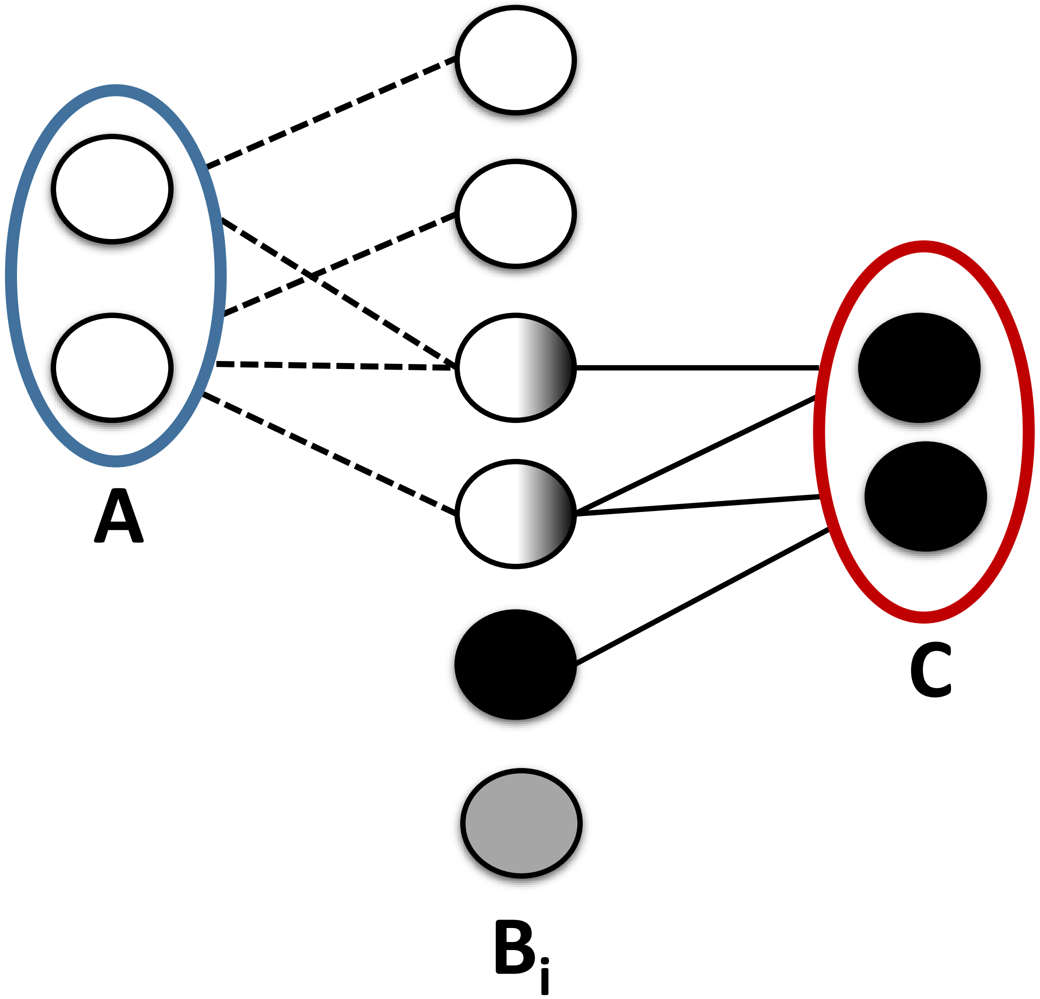
Scenario



LTA Set Modifications

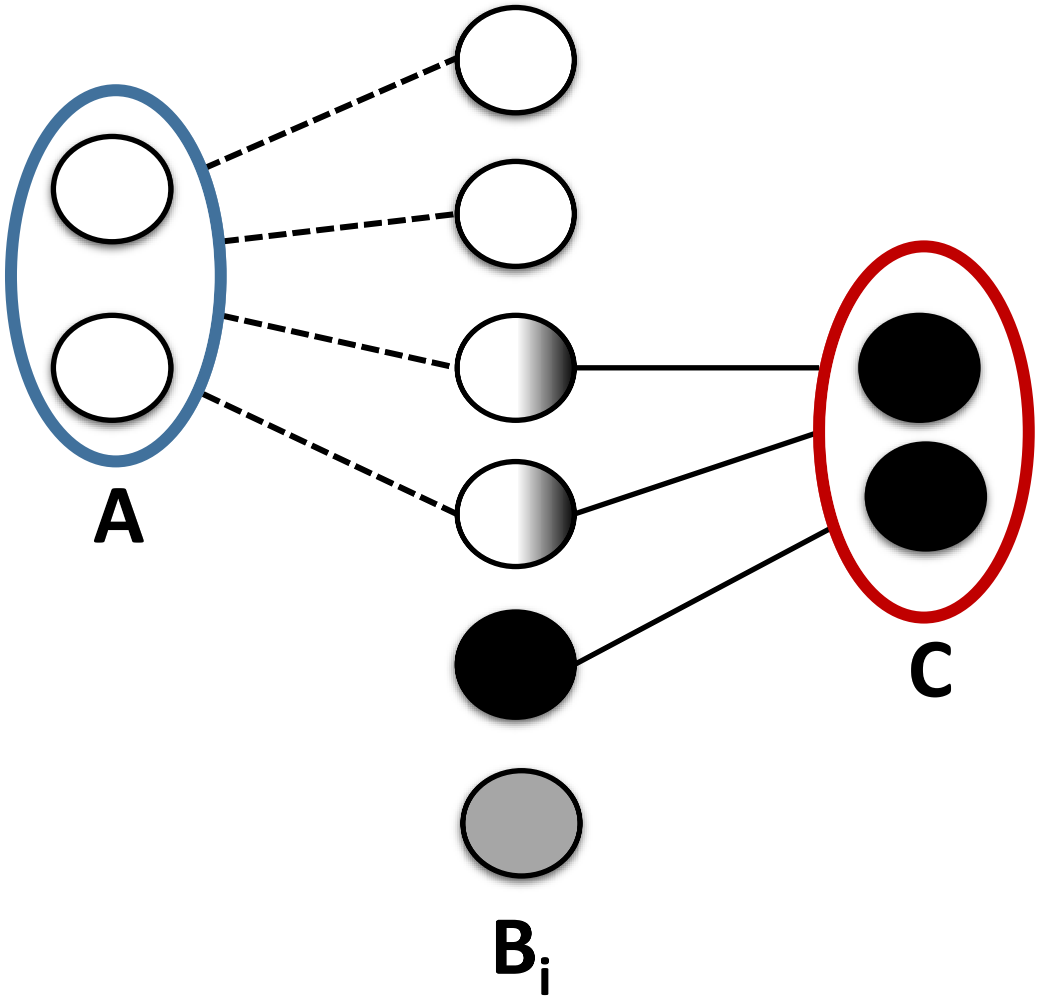


LTA Set Modifications



1) Group A and C terms

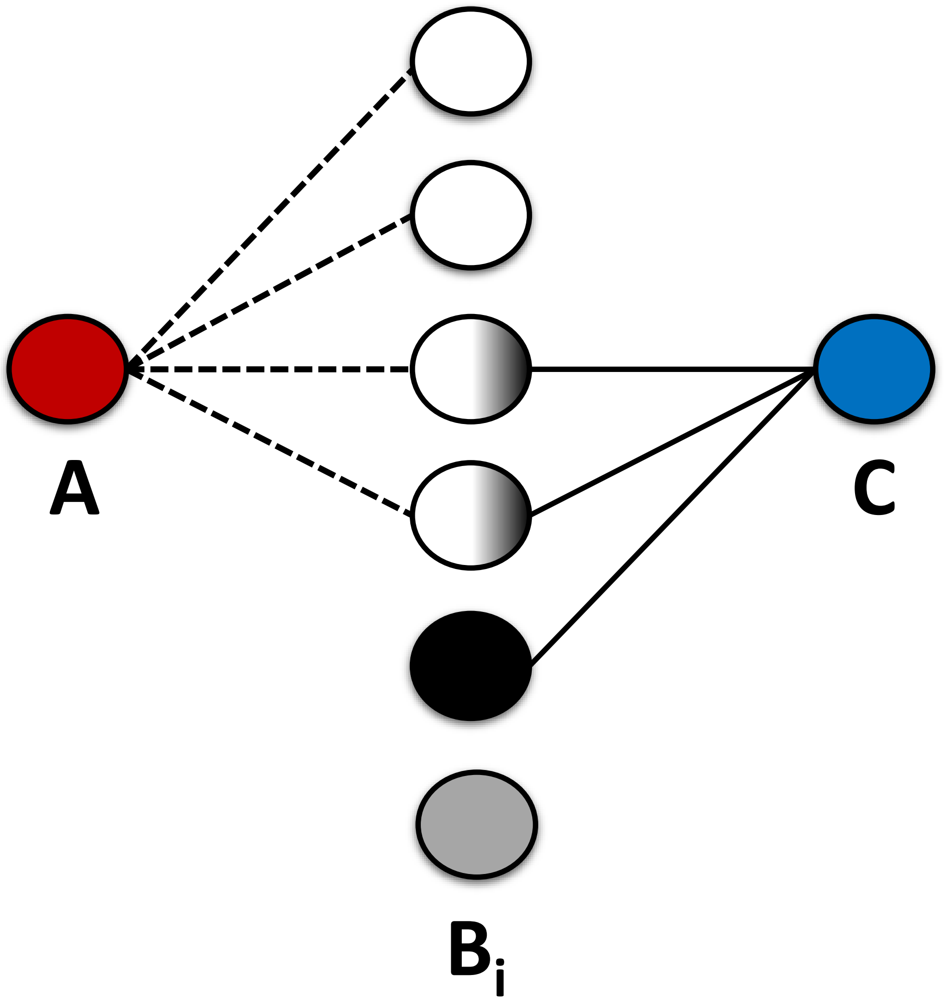
LTA Set Modifications



2) Collapse Edges

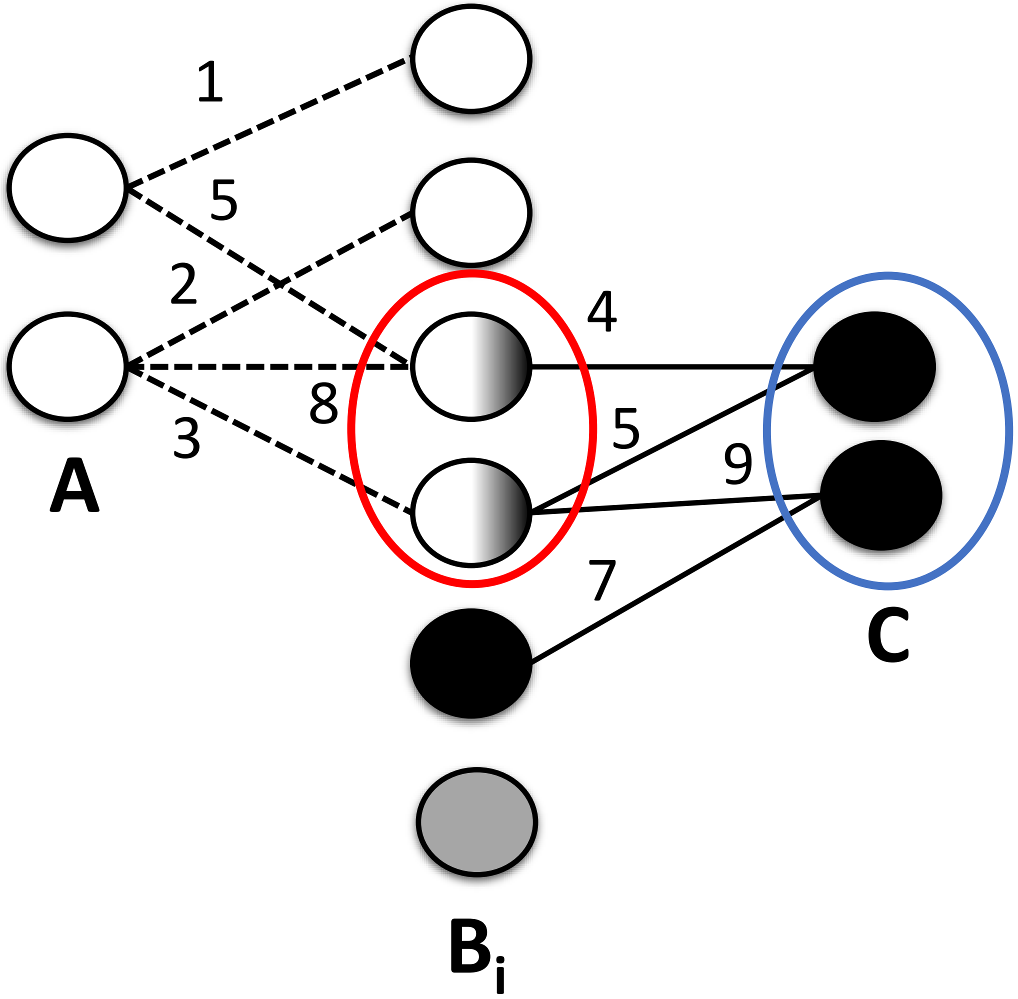
- Keep only the unique among sets

LTA Set Modifications

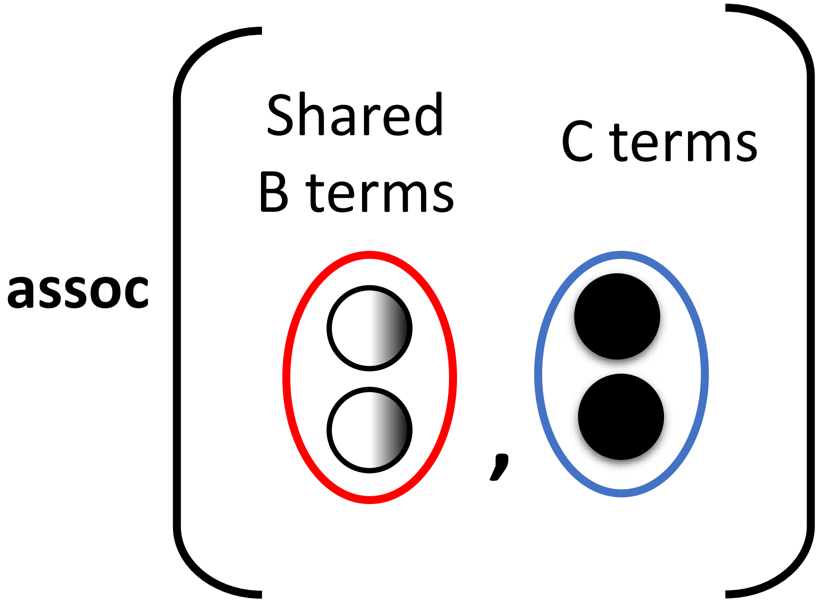


- **N11** = count of unique shared linking terms = LTC
- **N1P** = count of unique terms that set A co-occurs with
- **NP1** = count of unique terms that set C co-occurs with
- **NPP** = all possible unique terms (vocabulary size)

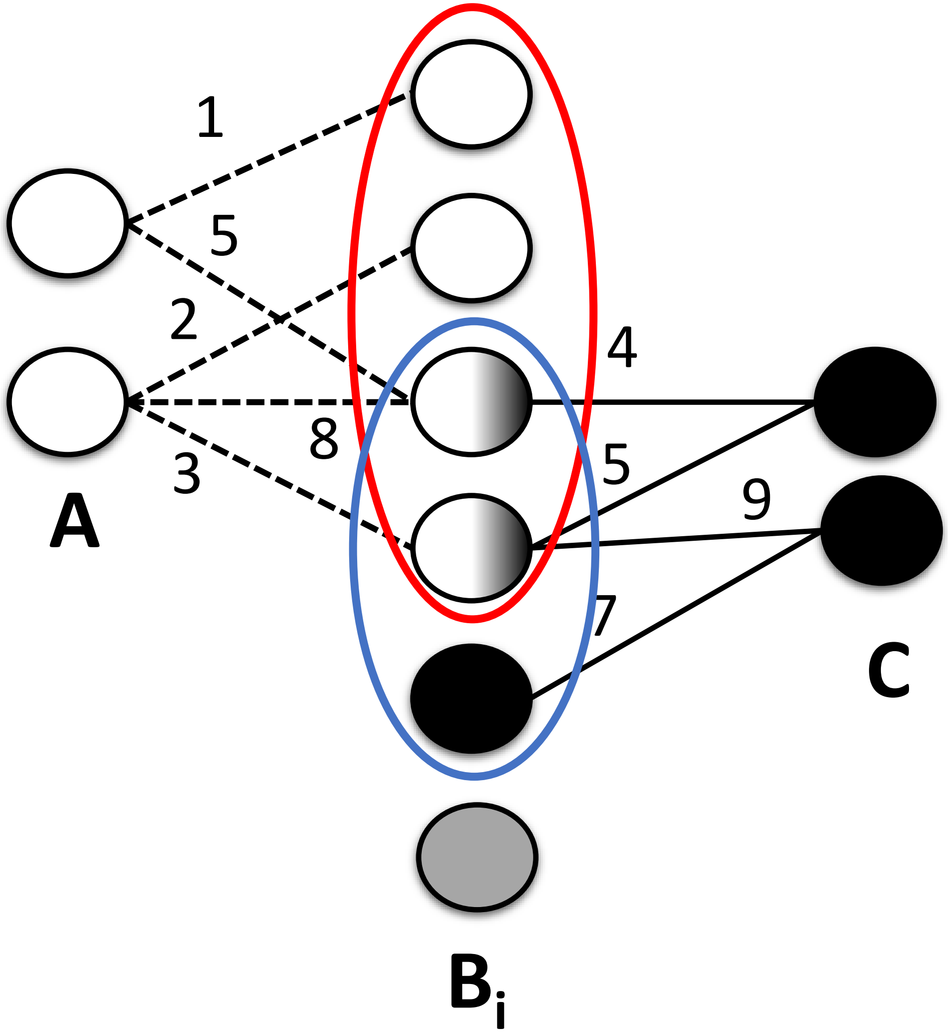
Shared B to C Set Association



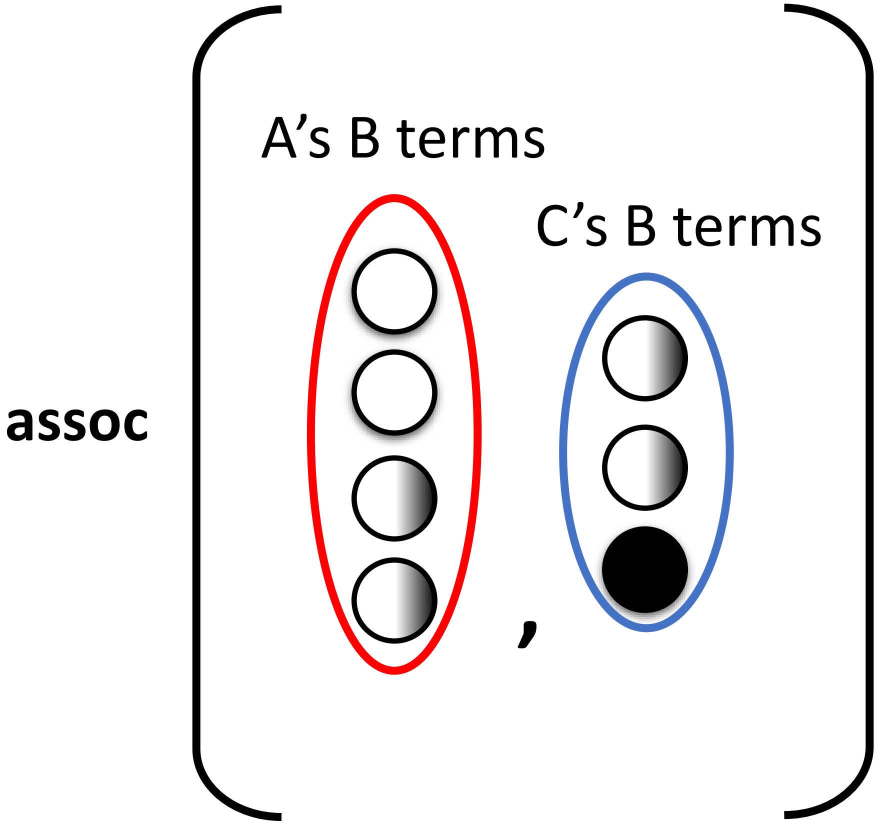
- Set association between shared B terms and C terms



Shared B Set Association



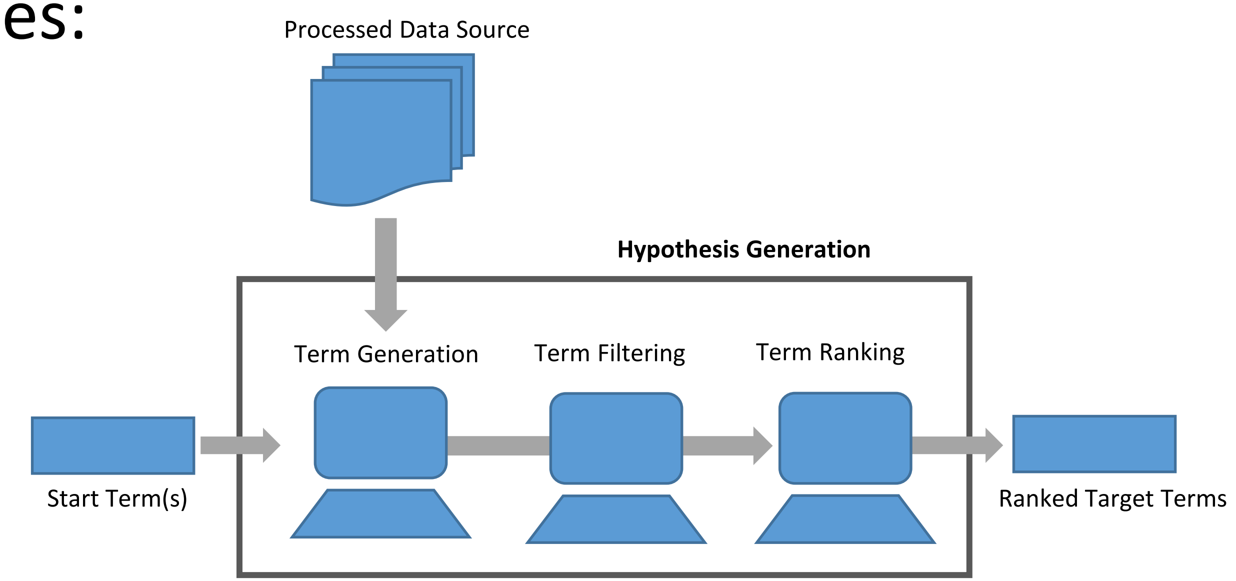
- Set association between A's B terms and C's B terms



Hypothesis Generation

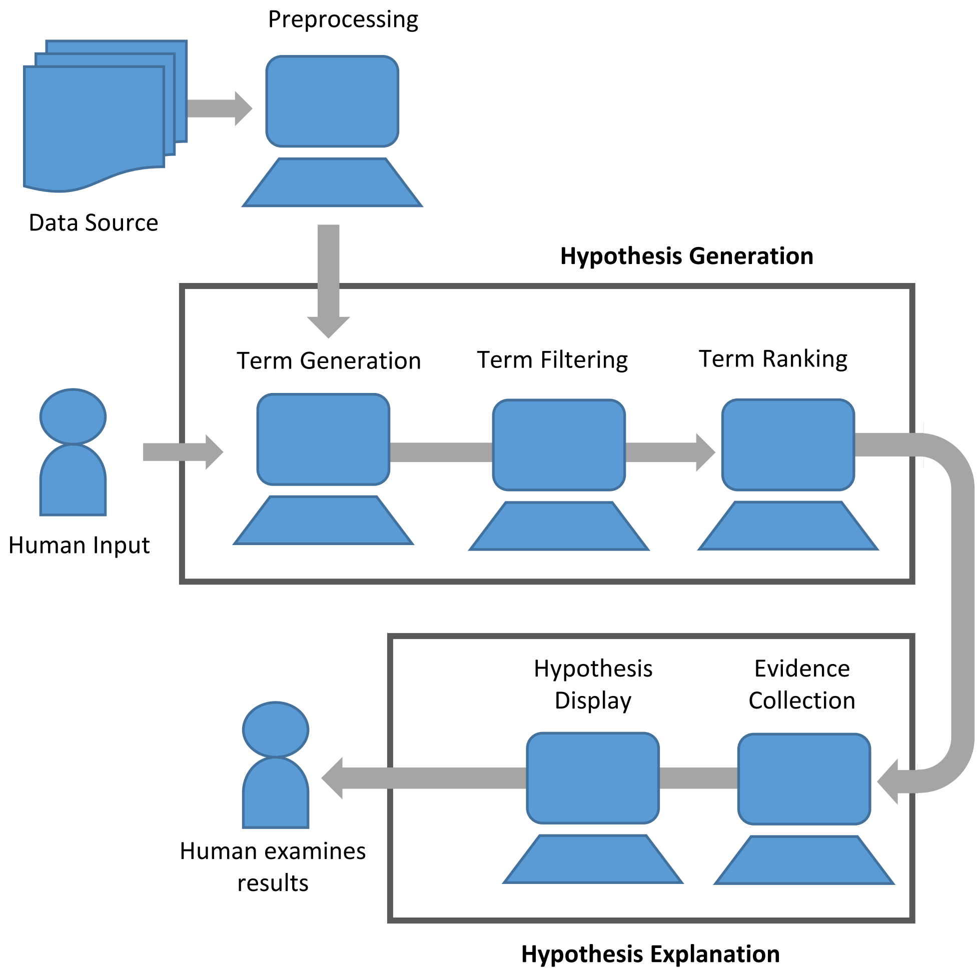
The whole process of generating, filtering, and ranking hypotheses:

1. Term Generation
2. Term Filtering
3. Term Ranking



System Components

1. Data Source
2. Preprocessing
3. Hypothesis Generation
4. Hypothesis Explanation



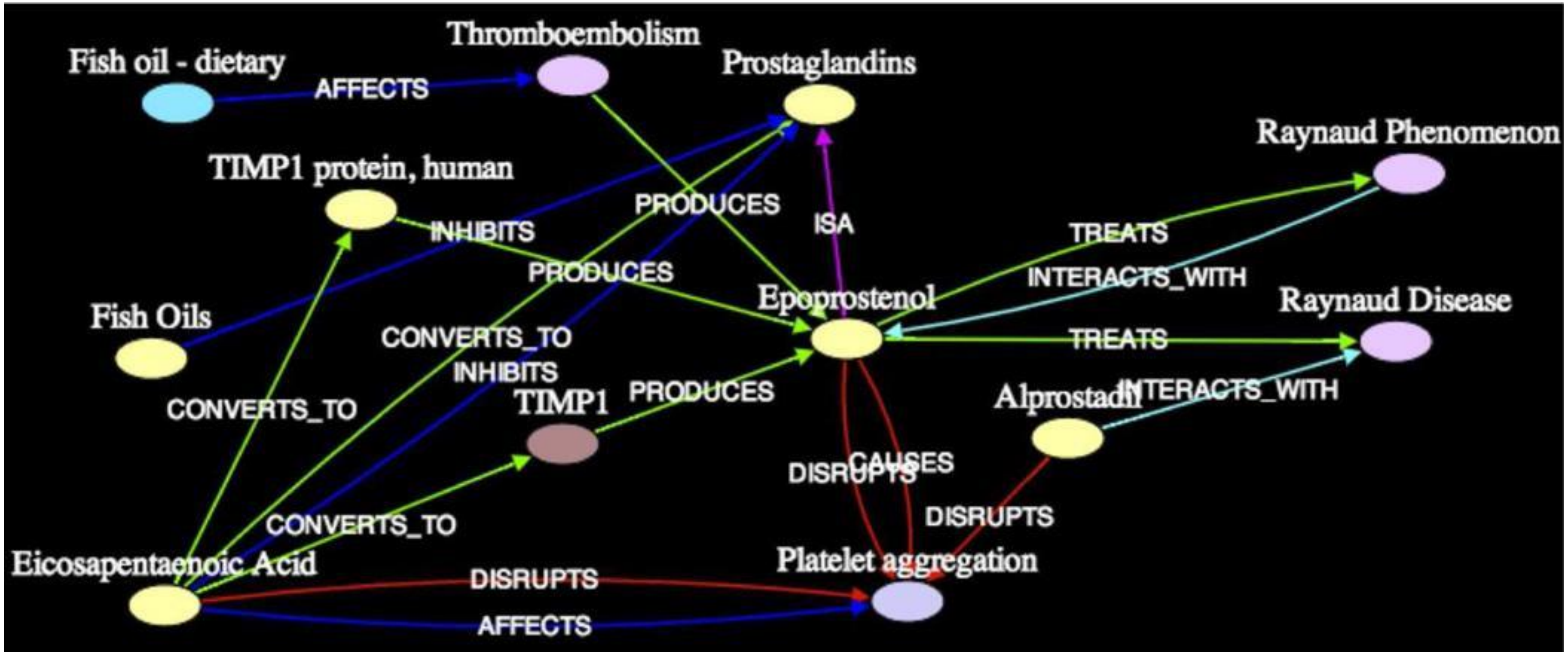
Hypothesis Explanation

- Explaining the reasons behind a hypothesis
- How the results of LBD are displayed

Hypothesis Explanation

- A target term list
- The articles in which the terms co-occur
- The relationship chain linking the terms

More Complex Graph Analysis



D. Cameron, R. Kavuluru, T. C. Rindflesch, A. P. Sheth, K. Thirunarayan, O. Bodenreider, Context-driven automatic subgraph creation for literaturebased discovery, Journal of Biomedical Informatics 54 (2015) 141–157.

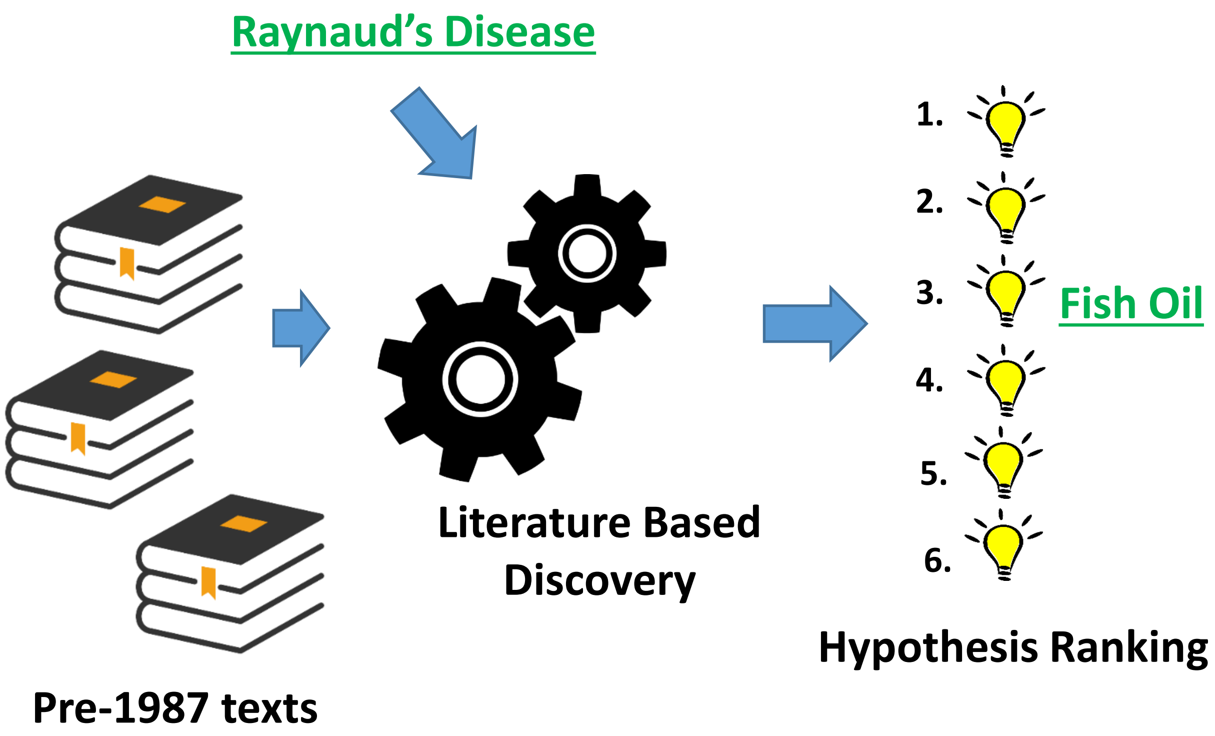
Evaluation

Evaluation

- Discovery Replication
- Time Slicing Analysis
- Expert Evaluation
- User Studies
- Link Prediction

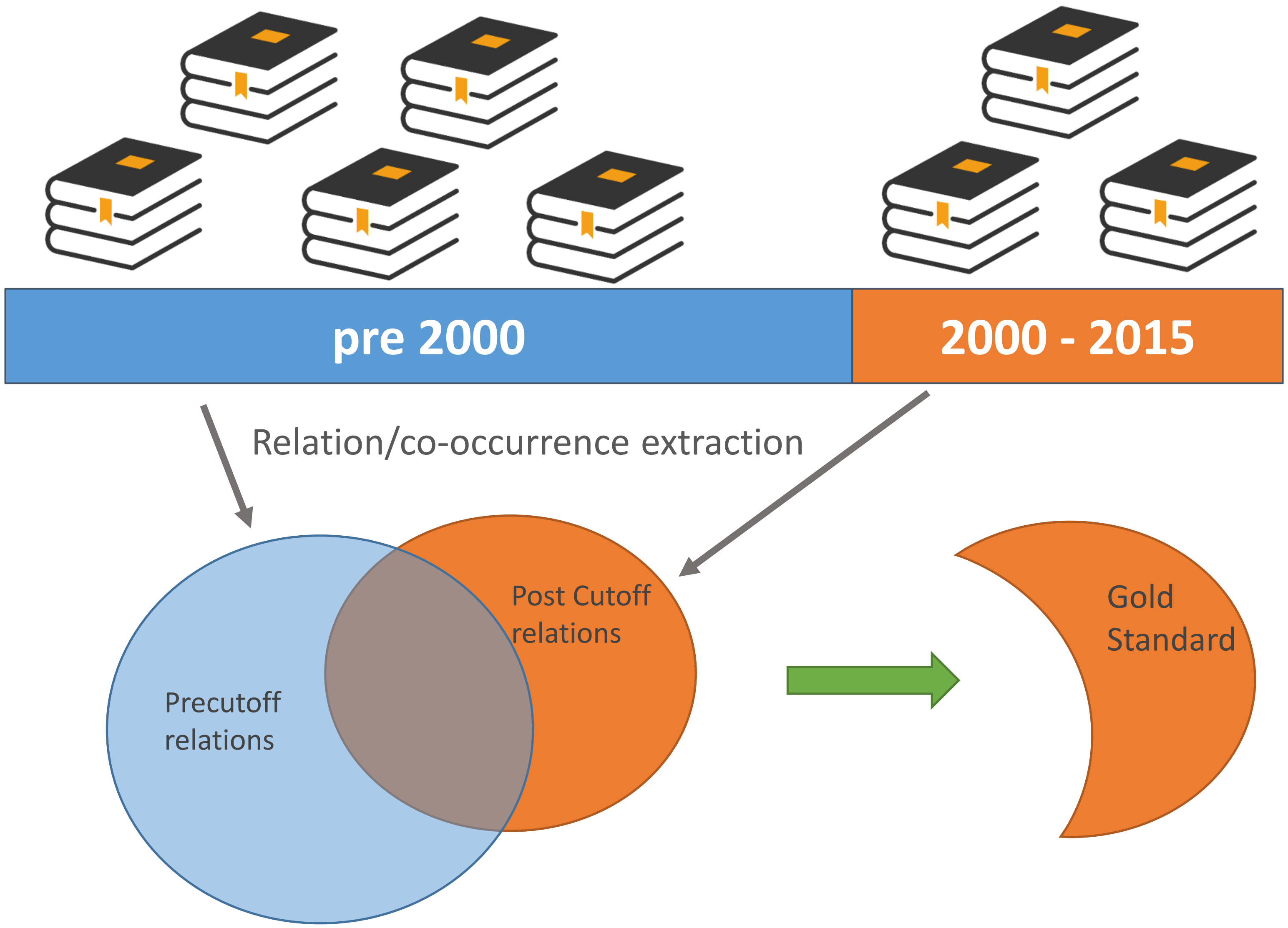
Discovery Replication

- Reproduce a discovery from the past
- Raynaud's Disease – Fish Oil made in **1987**

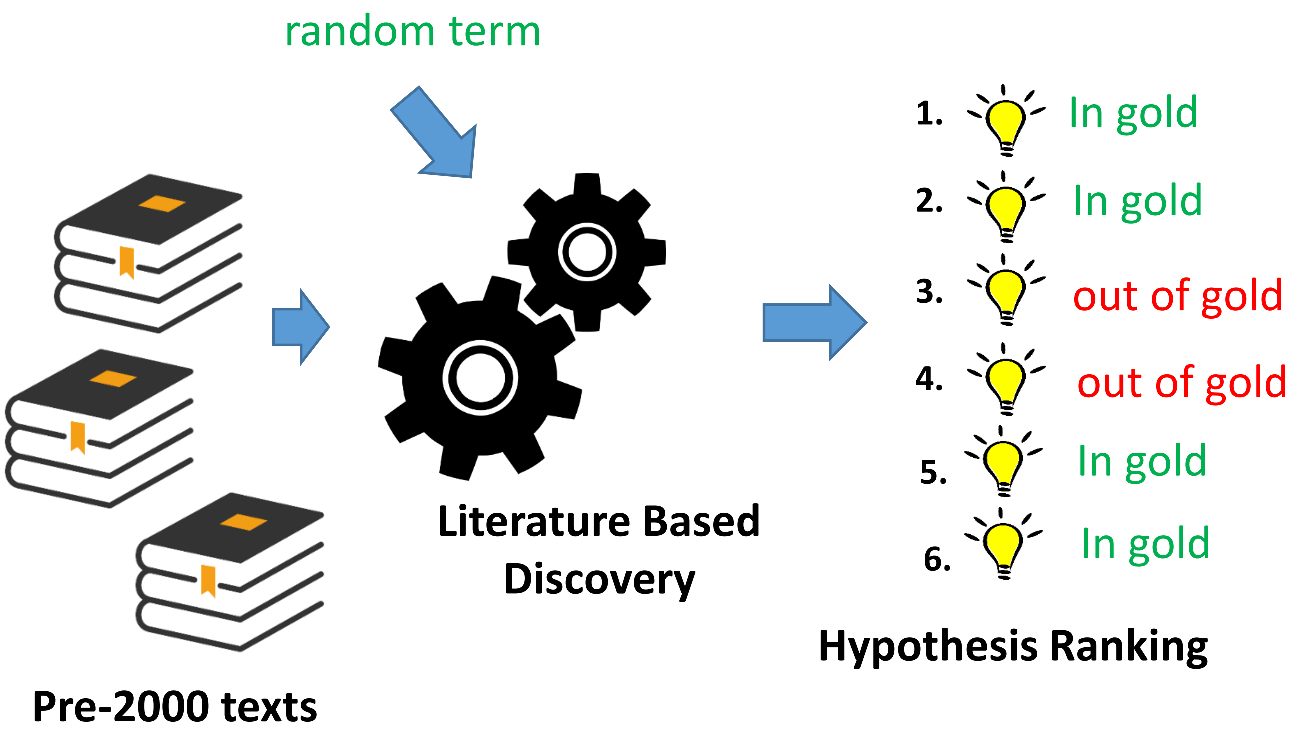


My system reproduced Raynaud's Disease – Fish Oil, and fish oil was ranked third

Time Slicing



Time Slicing



My system produced 4 hypotheses in the gold standard and 2 out of the gold standard

Precision = 66%
Recall = 10% *(made up)*

Average Precision = ~82%
(1/1 + 2/2 + 0 + 0 + 3/5 + 4/6)/4

Expert Evaluation

1. Using the system to propose new discoveries
2. Validating those discoveries
    - Expert Vetting
        - Publication and therefore peer review/scrutiny
        - Expert testimonial: A co-author (usually a doctor) attest to the discovery's validity
    - Support through other means such as micro-array analysis
    - Support through testing:
        - In-Vitro
        - In-Vivo
        - Clinical Trials

Expert Evaluation

- Strengths:
  - Proves that an LBD system works
  - Exposes LBD in that community
- Weaknesses:
  - Not quantitative
  - Not informative

Expert Evaluation

- Strengths:
  - Proves that an LBD system works
  - Exposes LBD in that community
- Weaknesses:
  - Not quantitative
  - Not informative

Is this really evaluation?

User Studies

- User studies determine what users like and dislike about a system
- Determine how a system is used, and how it can be improved
- People use an LBD system, and:
  - Complete questionnaires
  - Are monitored
  - Are interviewed

- Example:
  - Smalheiser, et al. perform a five year study involving a group of 120 voluntary researchers. Researchers filled out notebooks describing their use of Arrowsmith; weekly phone calls were made to monitor their progress. This, combined with “unsolicited” suggestions from web users were used to improve the web-interface, guide development of their system, and discovered novel ways the system was being used.

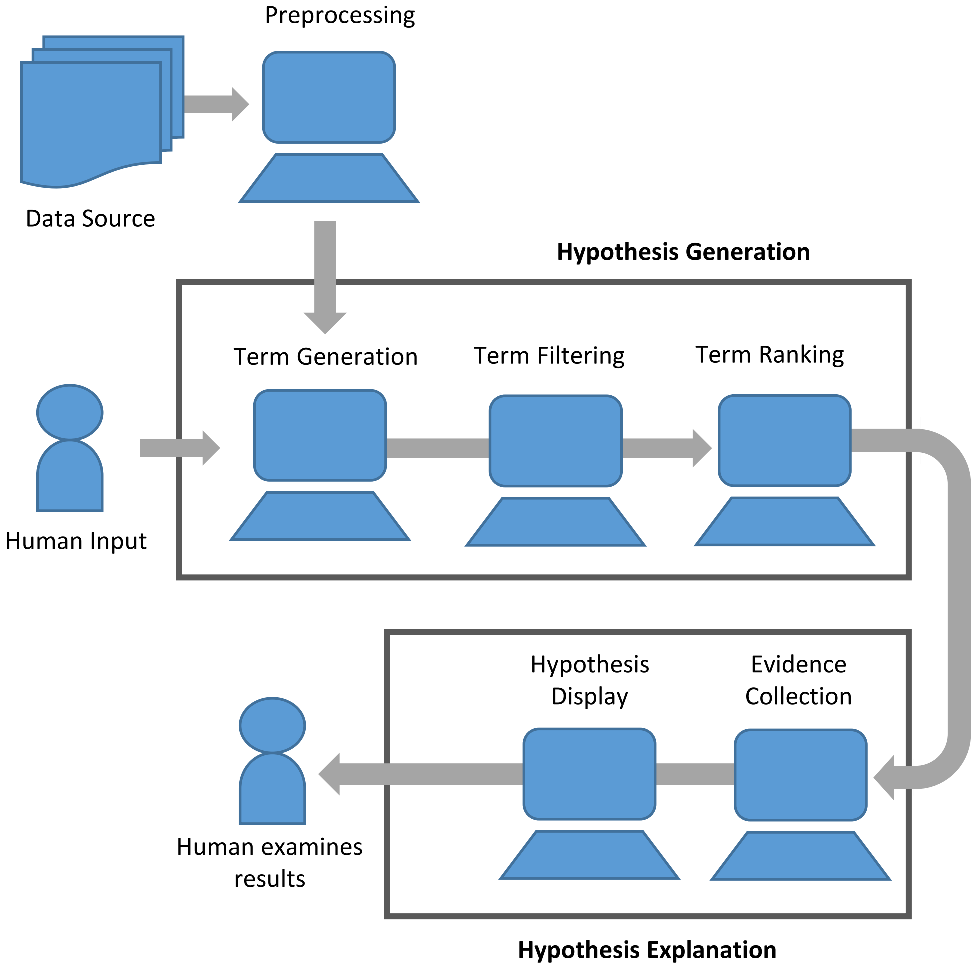
N. R. Smalheiser, V. I. Torvik, A. Bischoff-Grethe, L. B. Burhans, 965 M. Gabriel, R. Homayouni, A. Kashef, M. E. Martone, G. A. Perkins, D. L. Price, Collaborative development of the arrowsmith two node search interface designed for laboratory investigators, Journal of biomedical discovery and collaboration 1 (1) (2006) 8.

User Studies

- Strengths:
  - User studies are critical for understanding how LBD systems are actually being used.
  - User studies ensure the LBD tools we develop are both usable and useful
- Weaknesses:
  - Subjective
  - Not quantitative
  - Not automated or replicable.

Review and Example System

- Preprocessing:
  - Convert MEDLINE to SemMedDB
- Term Generation
  - ABC Linking of relationships
- Term Filtering
  - Remove terms that occur in more than 150,000 articles
  - Remove terms for which the start-target LTC > 1000
  - Keep only ‘Disease’ semantic types
- Term ranking
  - Use cosine distance between start and target co-occurrence vectors
- Evidence Collection
  - Find ABC relationship paths linking the terms
- Hypothesis Display
  - Visually display the top 10 ranked discoveries and the relationships linking them



Questions?

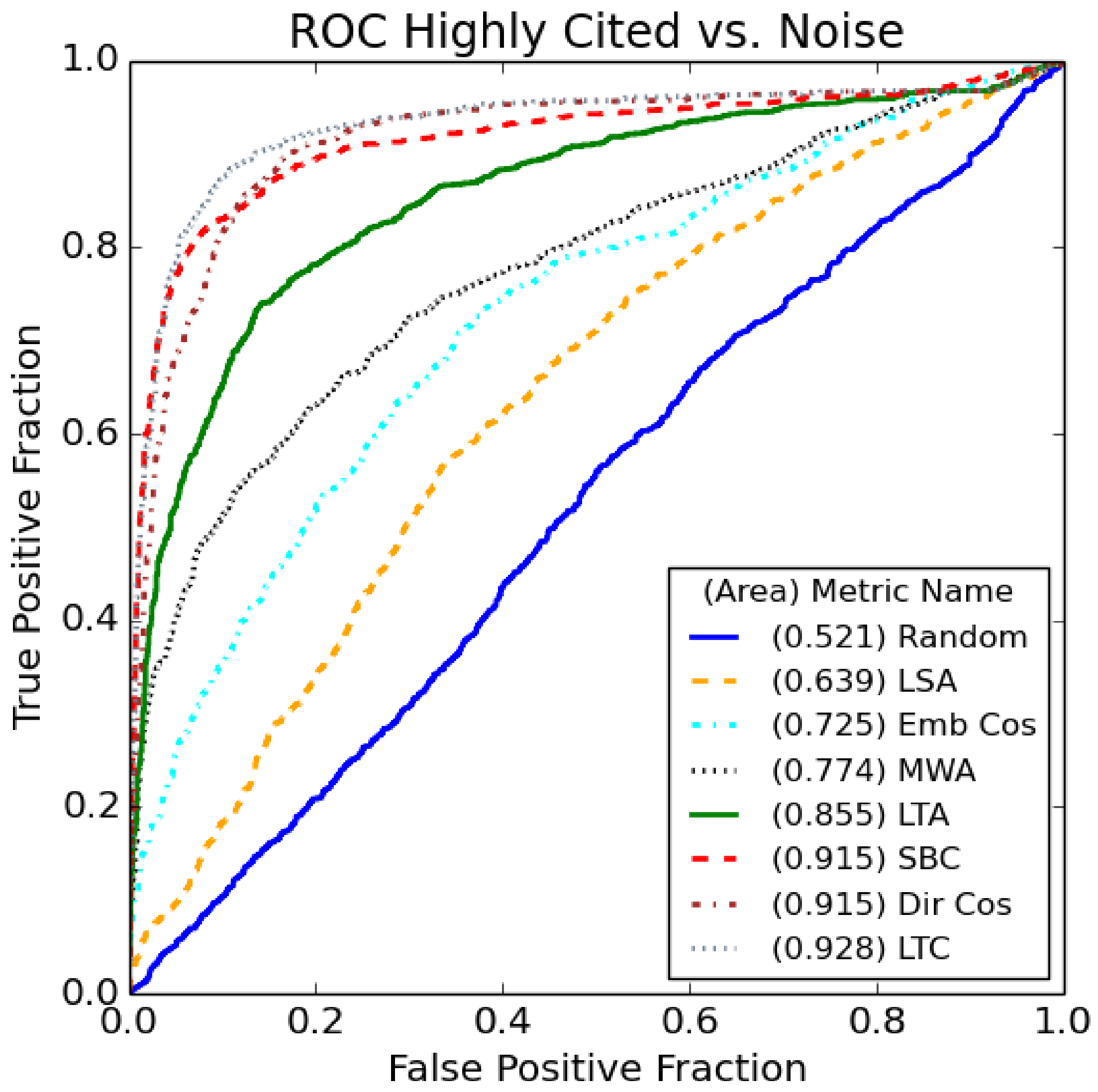
Backup slides

Link Prediction

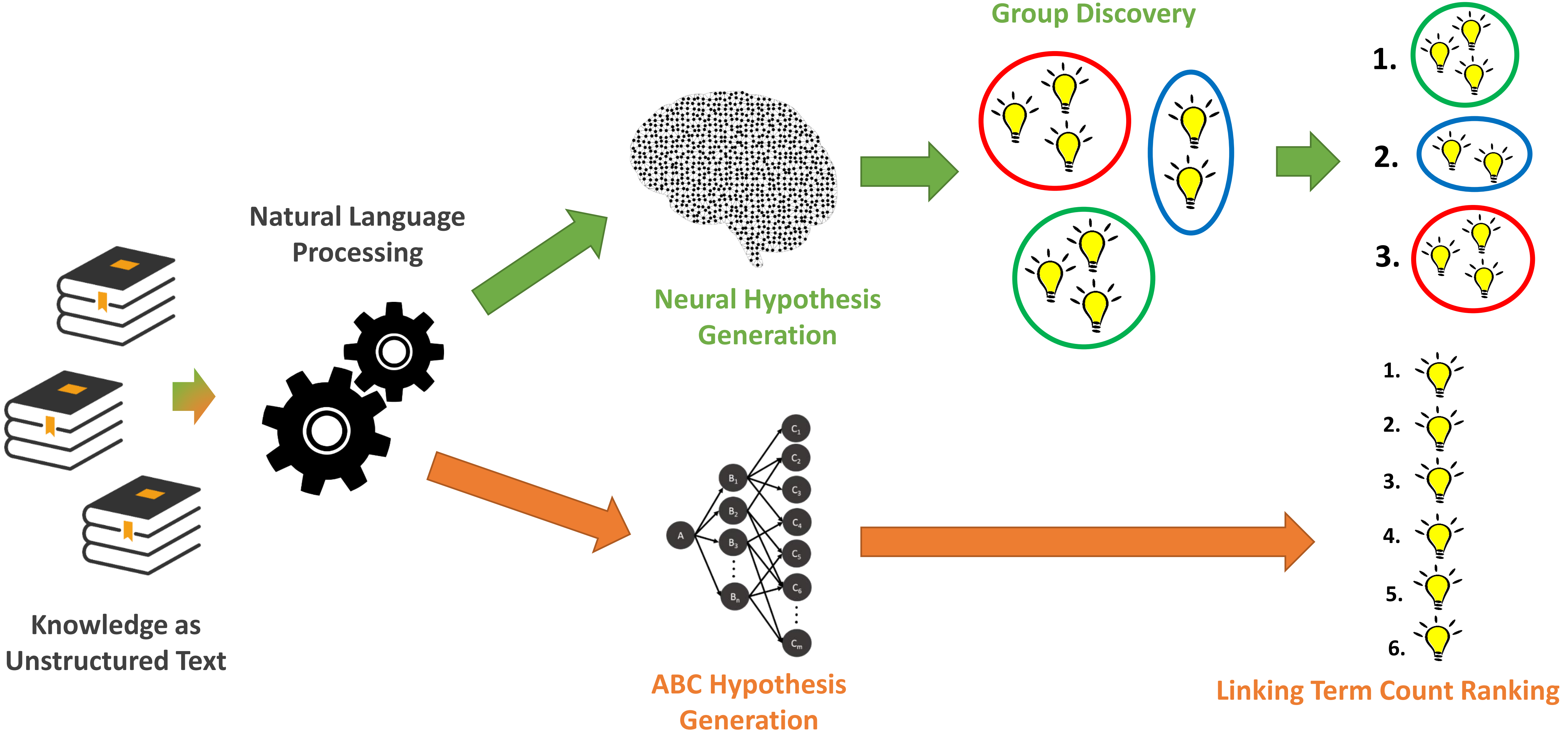
- Relax the true/false discovery assumptions.
  - They don’t look at whether a system can make true, actual discoveries but rather if it can or can’t predict links in a graph
- Evaluation with ROC curves
- Receiver operating characteristic (ROC) curves:
  - Plot the tradeoff between true and false positive rates
  - typically to evaluate the performance of binary classifiers
  - They require a dataset with true and false samples
  - Generated by varying some parameter (typically a threshold) and calculating the true and false positive rates at different values of this parameter.
  - The area under the ROC curve (AUROC) can be calculated to provide a single number to summarize a system’s performance

Link Prediction

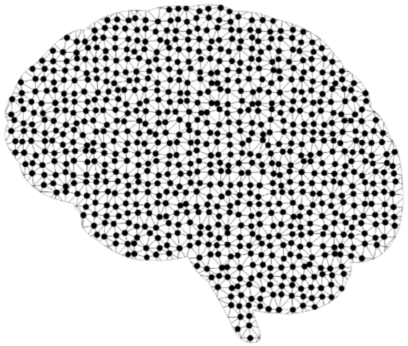
- Link prediction to compare different target term ranking algorithms
- SemMedDB
- Time sliced:
  - Training: 1975-2009
  - True samples are term pairs in 2010+, not in training
  - False are term pairs that do not occur in either dataset



Proposed vs Baseline Systems

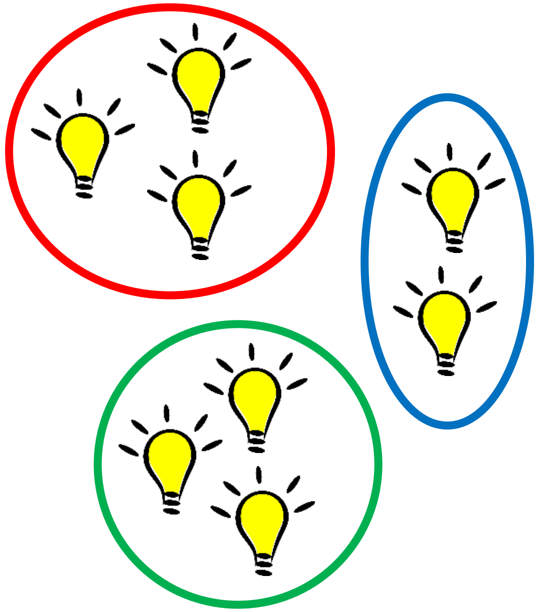


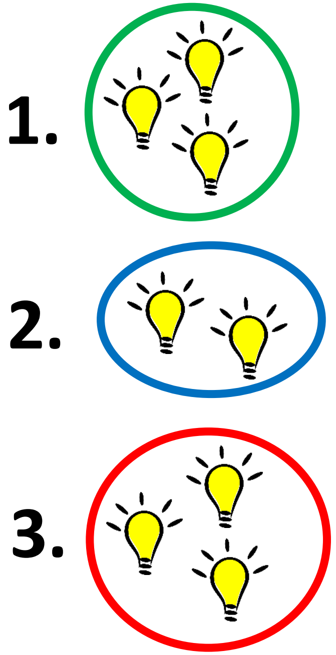
Three Primary Innovations (proposed works)



Neural Hypothesis Generation

Automatic Functional Group Discovery

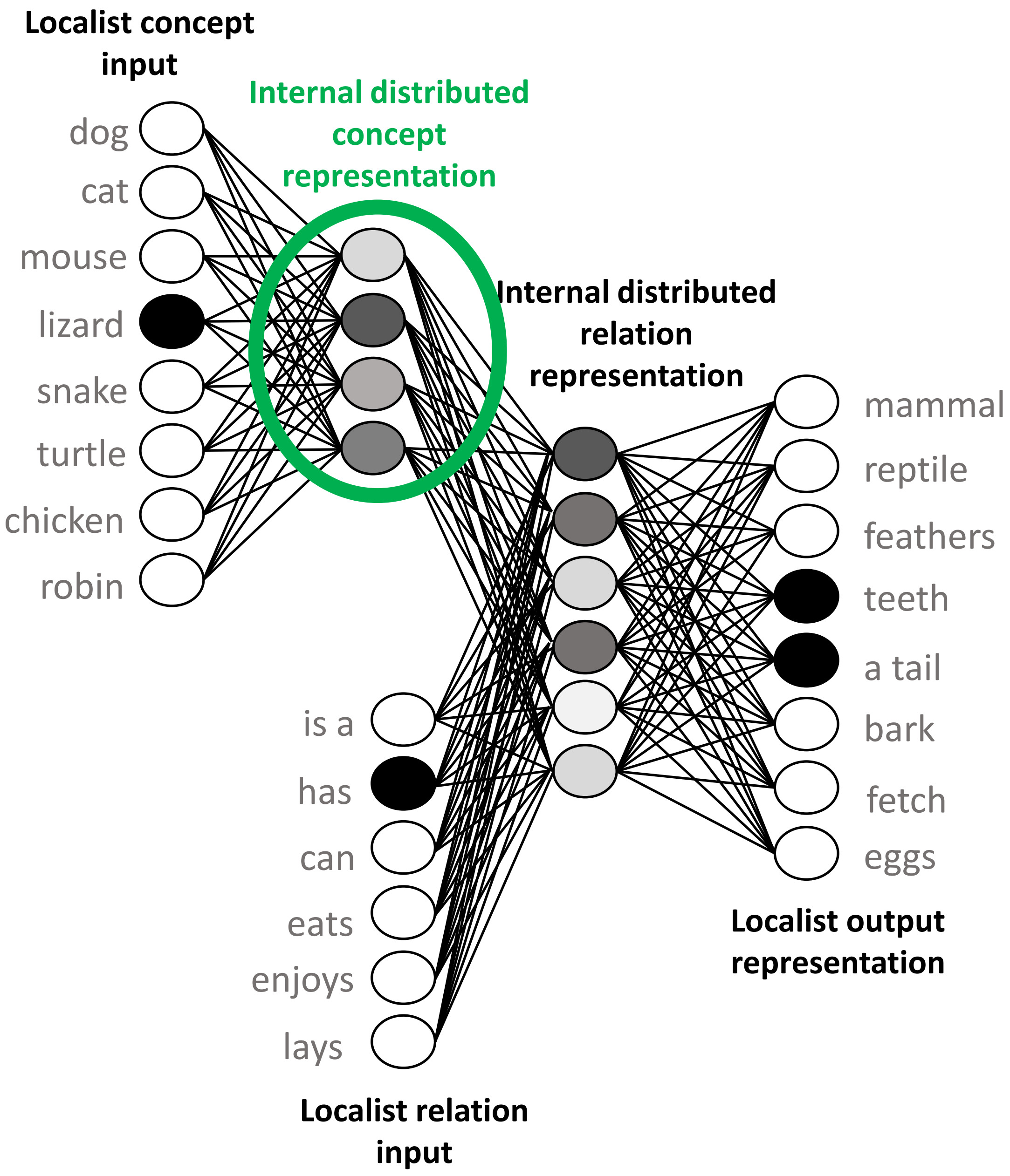




Implicit Association Ranking of Sets

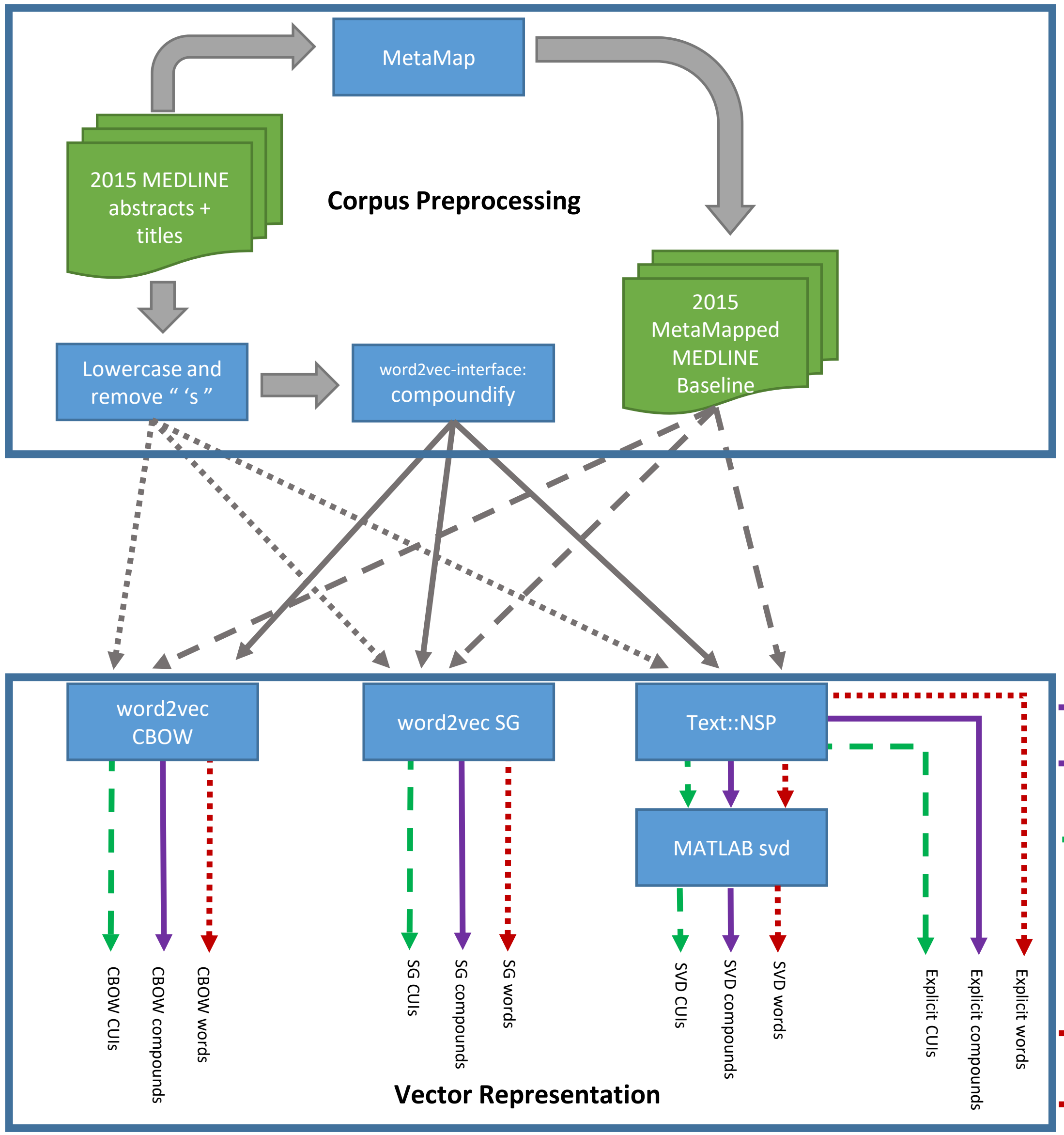
Rumelhart Network

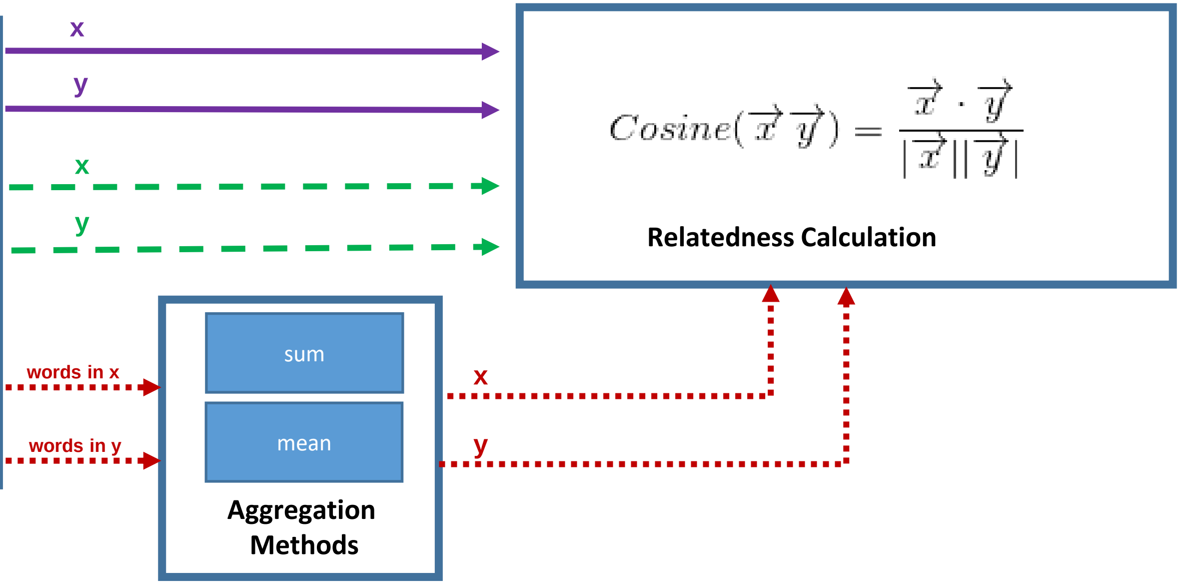
- Language Learning:
  - Monitor the internal representations as learning progressed
- It generalizes as it learns
- Learns internal representations hierarchically
  - All birds can fly
  - later learns: that most birds can fly, but ostriches can't
- Happens through similarities between the internal distributed representations



Vector representations of multi-word terms[1]

Method	UMNSRS		MiniMayoSRS		
	Sim.	Rel.	Phys.	Cod.	avg
CBOW words (ours)	0.61 (n=396)	0.69 (n=374)	0.82 (n=29)	0.82 (n=29)	0.82
CBOW comp. (ours)	0.65 (n=393)	0.70 (n=373)	0.80 (n=29)	0.78 (n=28)	0.79
CBOW cuis (ours)	0.60 (n=413)	0.73 (n=388)	0.77 (n=29)	0.83 (n=29)	0.80
Sajadi, et al. [150]	0.39 (n=566)	0.39 (n=597)	-	-	0.8
Pakhomov, et al. [149]	0.62 (n=449)	0.58 (n=458)	-	-	-
Muneeb, et al. [147]	0.52 (n=462)	0.45 (n=465)	-	-	-
Chiu, et al. [148]	0.65 (n=?)	0.60 (n=?)	-	-	-

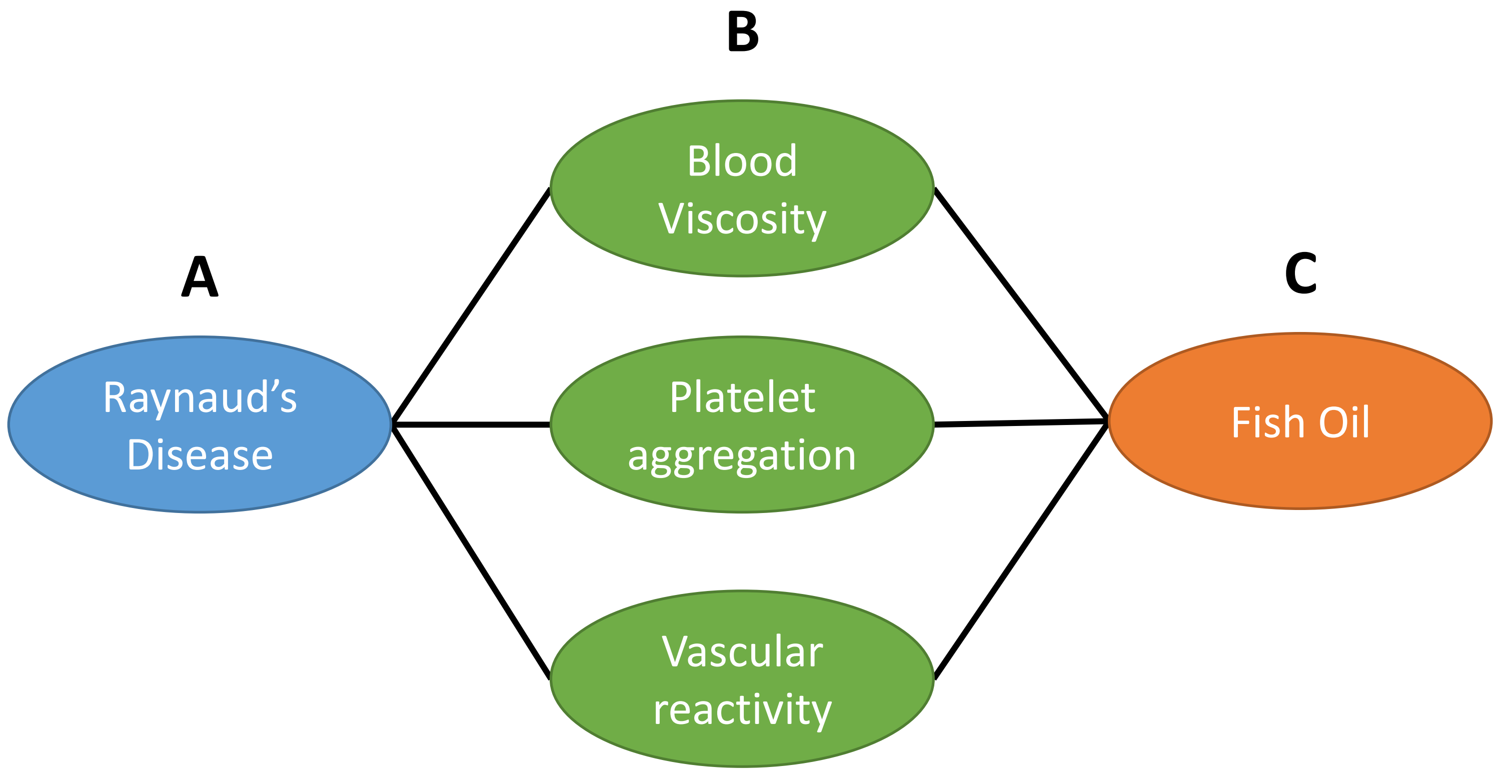




 [1]**Sam Henry**, Clint Cuffy, and Bridget T. McInnes. "Vector representations of multi-word terms for semantic relatedness." *Journal of biomedical informatics* 77 (2018): 111-119.

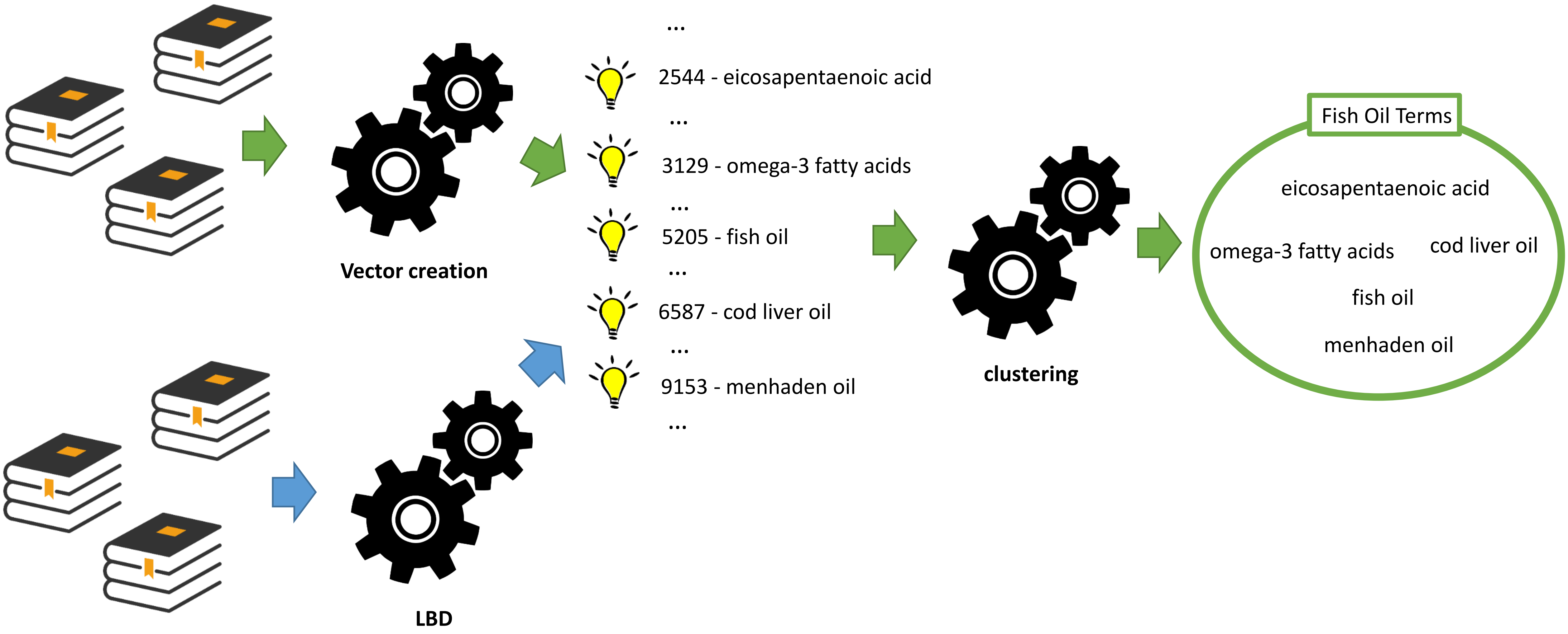
Functional Pathways

- Doctors found how fish oil treats Raynaud's Disease
- Weeber, et al. manually replicated this with LBD[1]

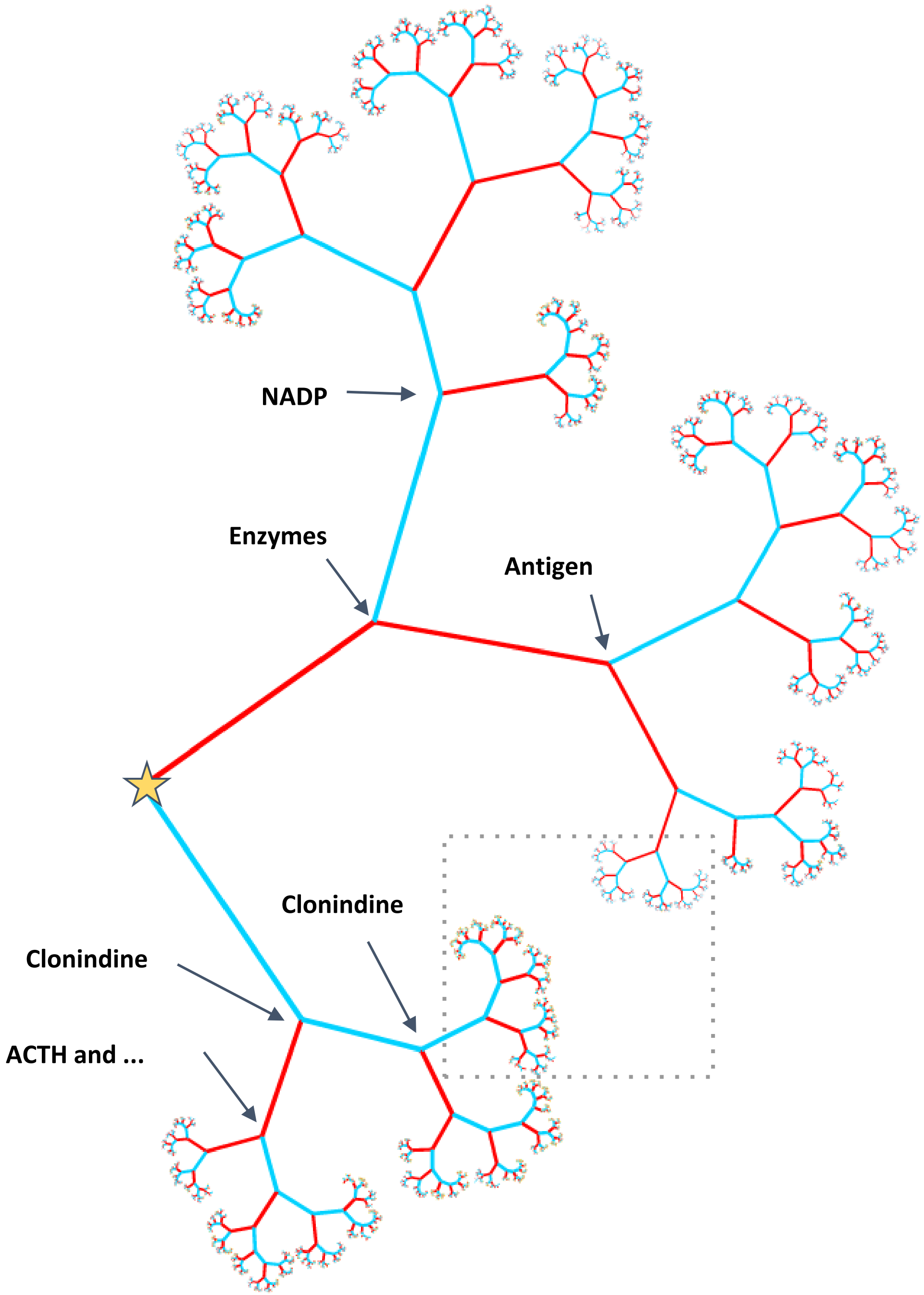


[1]Weeber, Marc, et al. "Using concepts in literature-based discovery: Simulating Swanson's Raynaud–fish oil and migraine–magnesium discoveries." *Journal of the Association for Information Science and Technology* 52.7 (2001): 548-557.

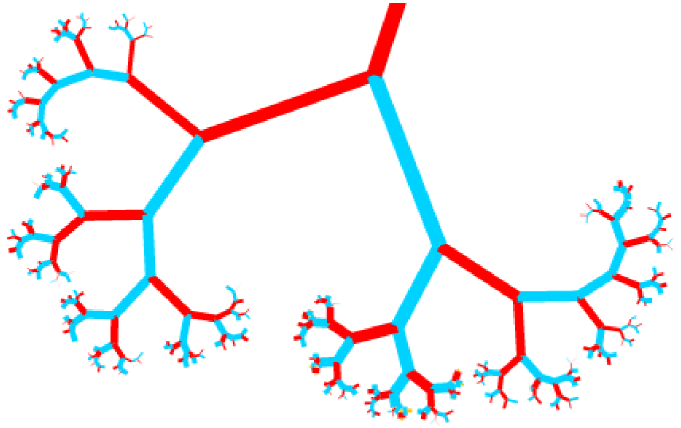
Hierarchical Clustering

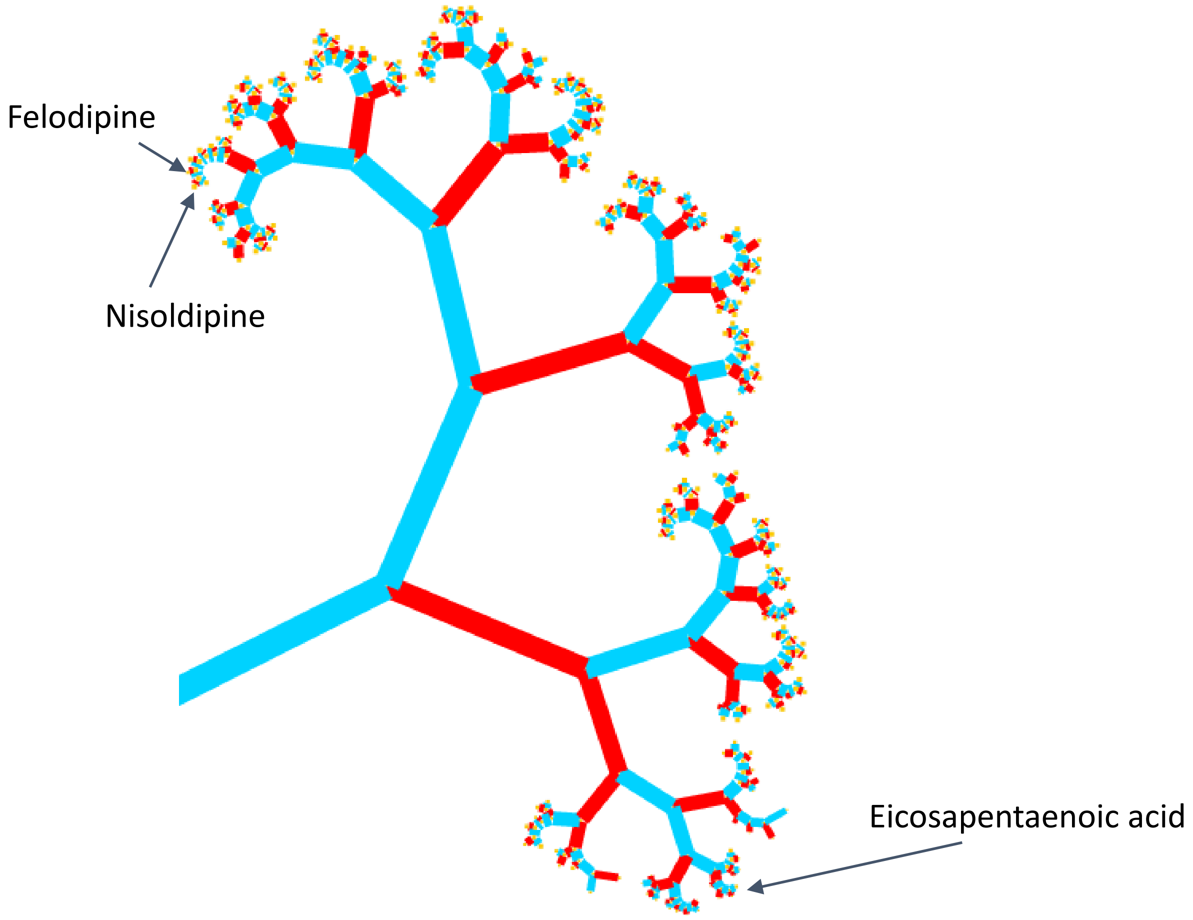


Interactive Exploratory Interface



Interactive Exploratory Interface

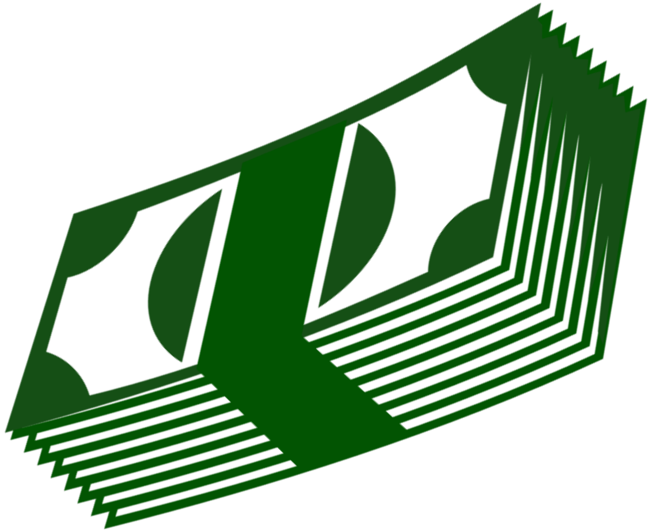




In Summary, LBD is:

- What?
  - Automated methods to find connections between disjoint fields
- Why?
  - There are millions of new scholarly publications per year
  - Increased specialization and narrowing of disciplines
  - Can guide research and produce new knowledge
- How?
  - Through text mining and natural language processing

Motivation



LBD Applications

Drug Discovery

Drug Repurposing

Adverse Event Prediction

LBD provides a better understanding of

drug mechanisms, interactions, and **side effects**

Drug Discovery

Expensive

- Costs $500 million to $2 billion to develop new drugs

Time Consuming

- Takes 10 – 15 years

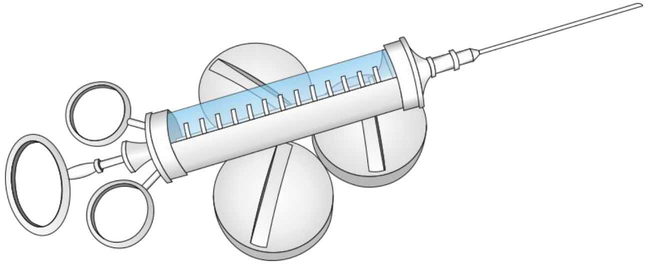
Difficult

- Success rate is less than 10%

Drug Repurposing

On the rise

- ~30% of newly approved drugs



Approved Drugs Exist

- ~4,000 drugs approved for human use

Saves Money

- may reduce costs by 50%

Adverse Event Prediction

Caused by:

- Normal use, misuse, discontinuation of medication

Common

- ADEs account for ~12% of all emergency room visits

Deadly

- ADEs kill people