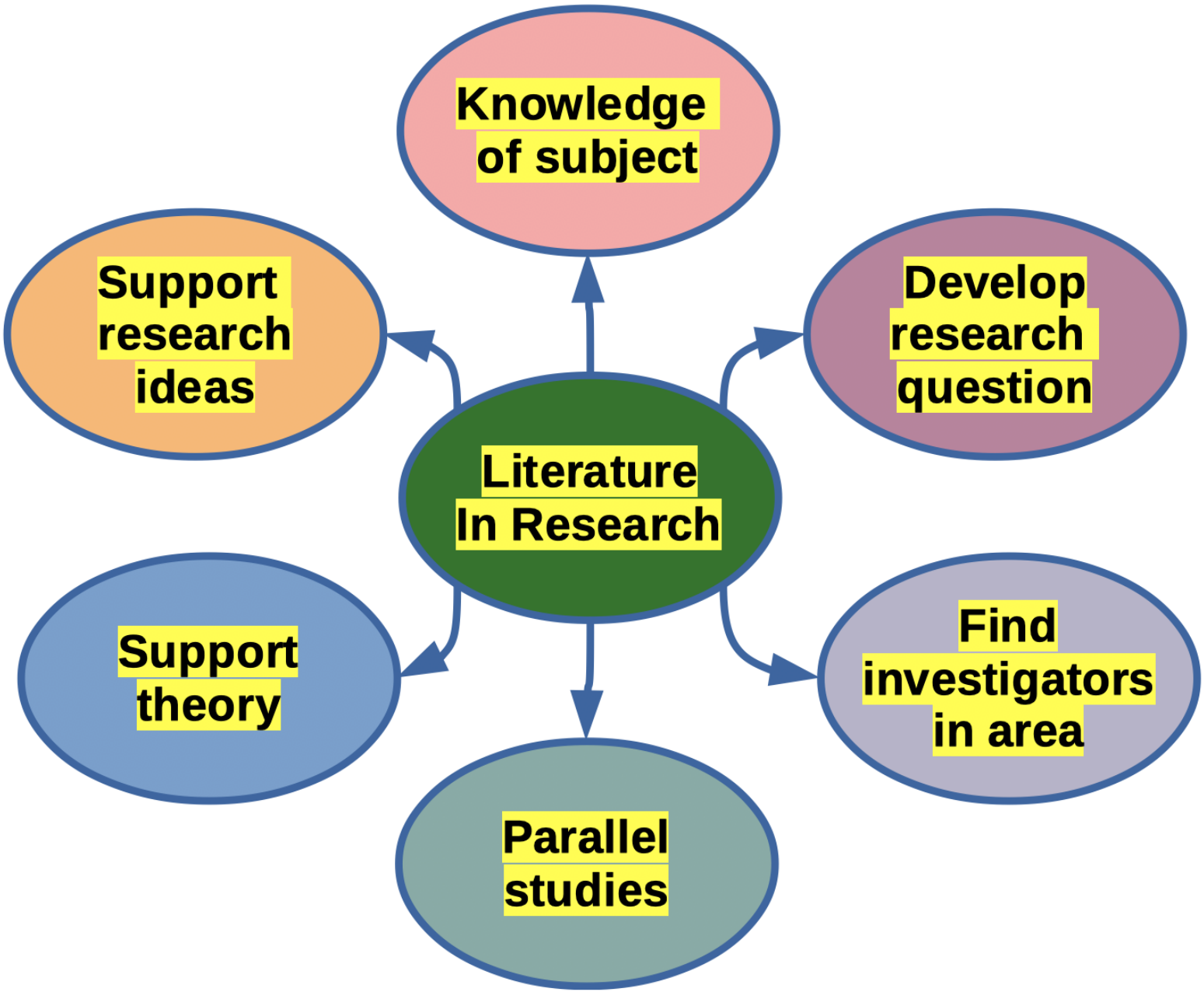
BeagleTM: An adaptable text mining method for relationship discovery in literature

Oliver BONHAM-CARTER
Allegheny College
Meadville, PA

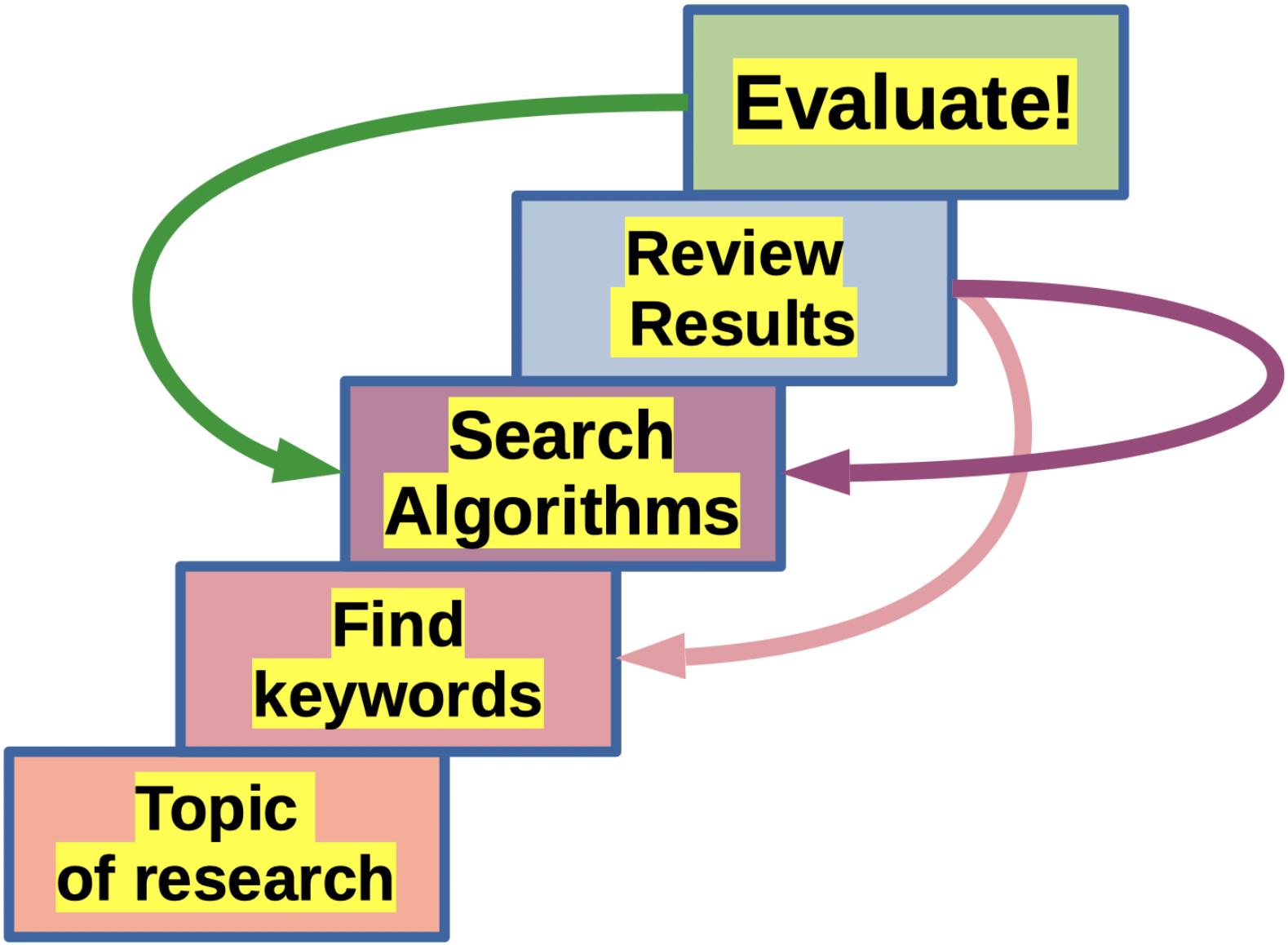
FICC 2020

6^{th} March

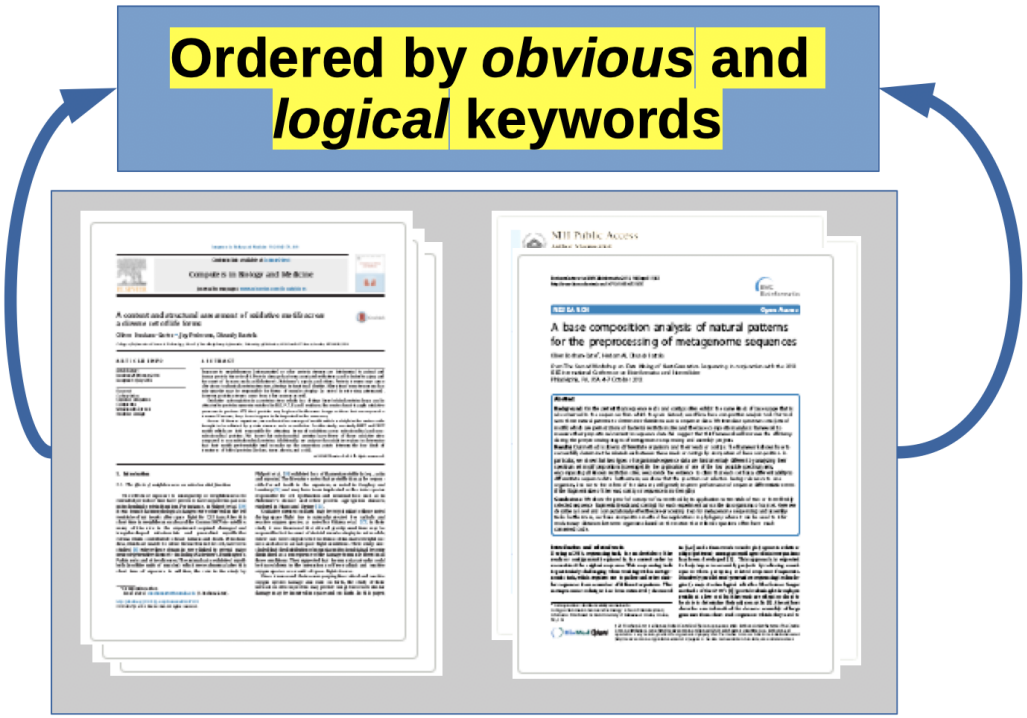
The Role of Literature in Research



Applying Literature in Research

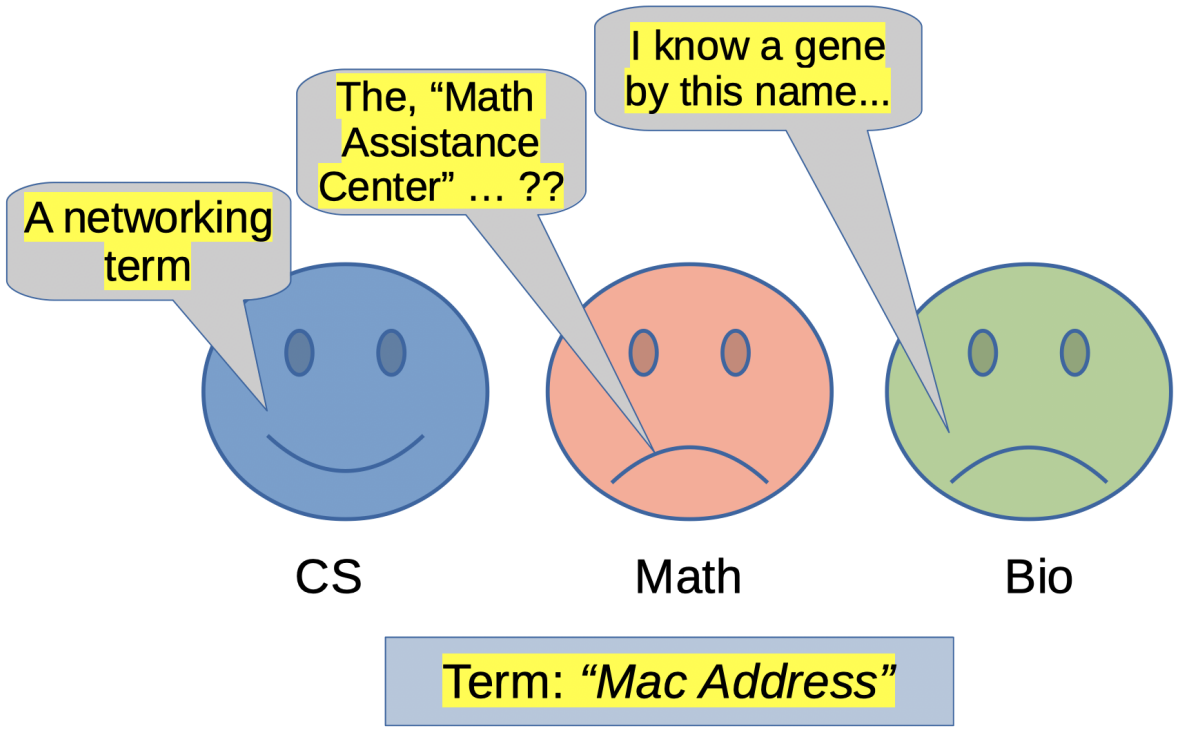


Function of Keywords



- To allow discovery by an **obvious** and **logical** connection to work
- To inform of scope
- To help diverse search algorithms organize articles by *same* criteria

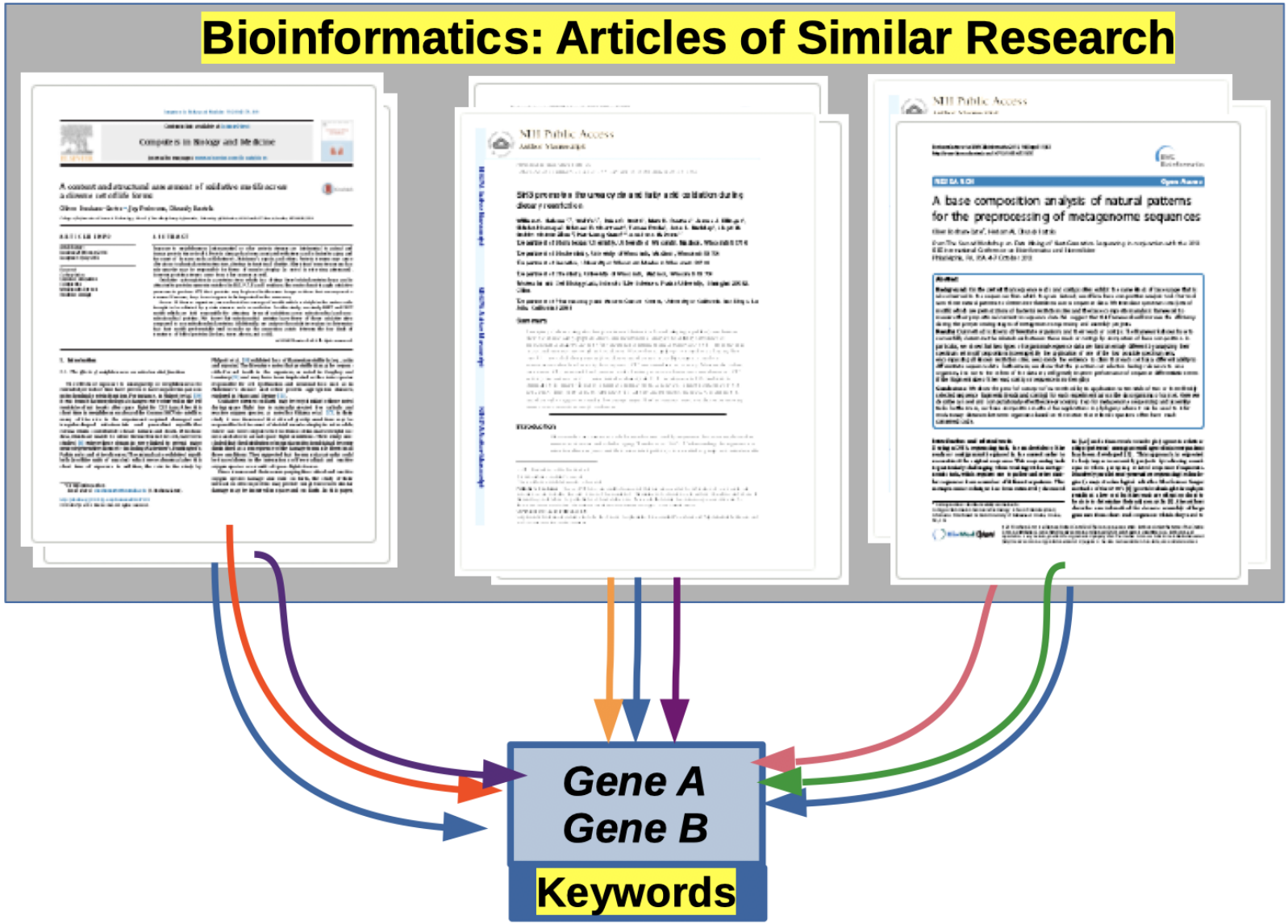
Keywords Often Fail



- Are often *vague*, describing little of article's contents
- Are chosen by researchers who may not know the vocabulary used by the rest of a community
- In multidisciplinary disciplines (i.e., bioinformatics): chosen keywords likely to have different meanings within own community

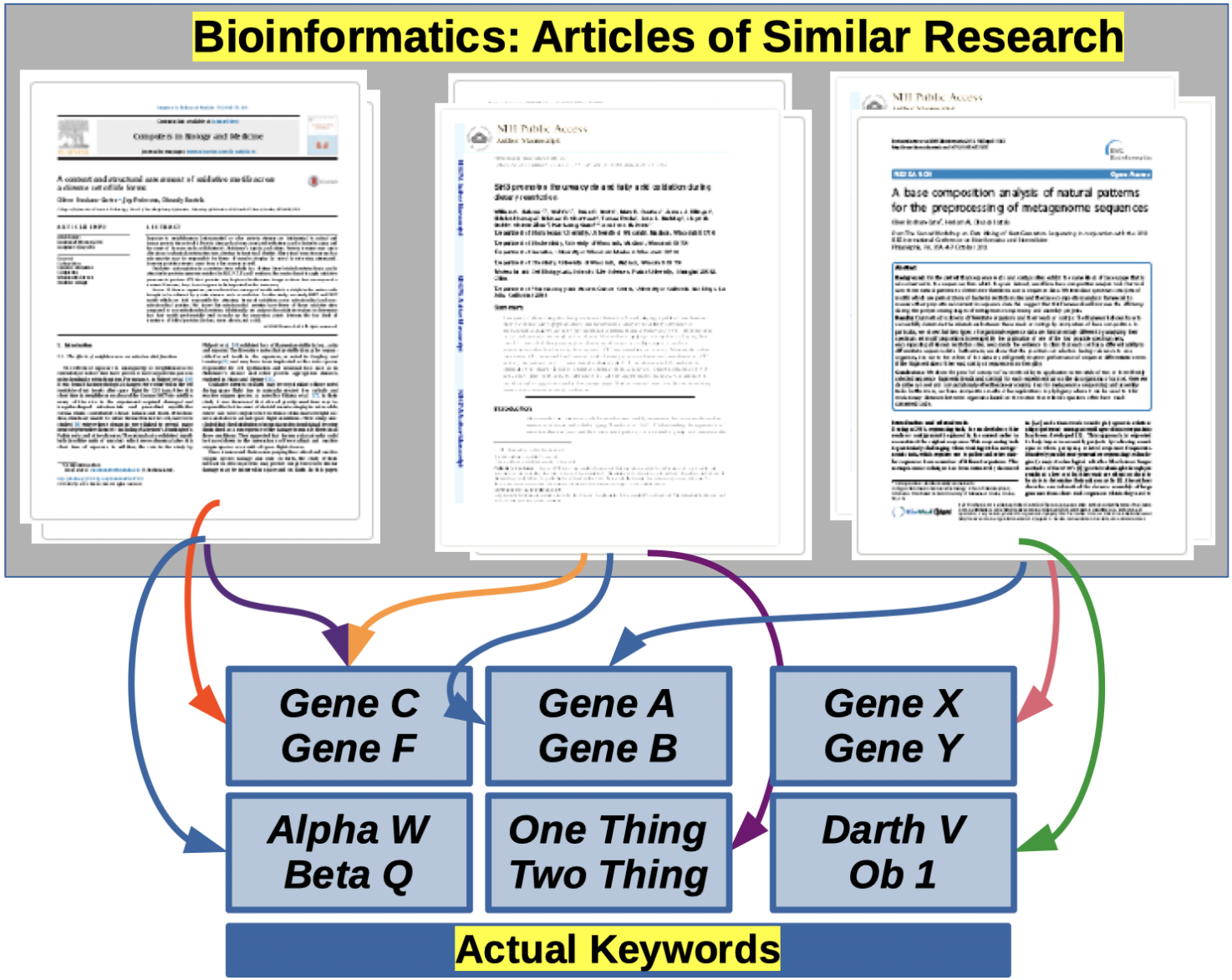
Expectations

The same keywords for all similar knowledge



Actuality

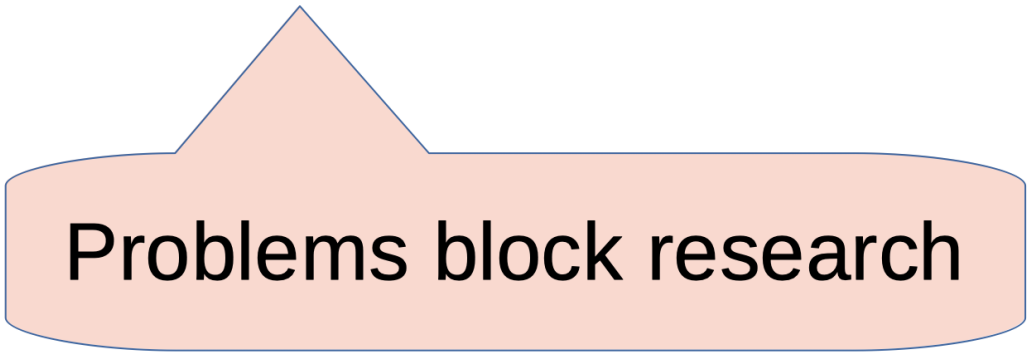
Articles have different keywords!



Three Main Problems With Keywords

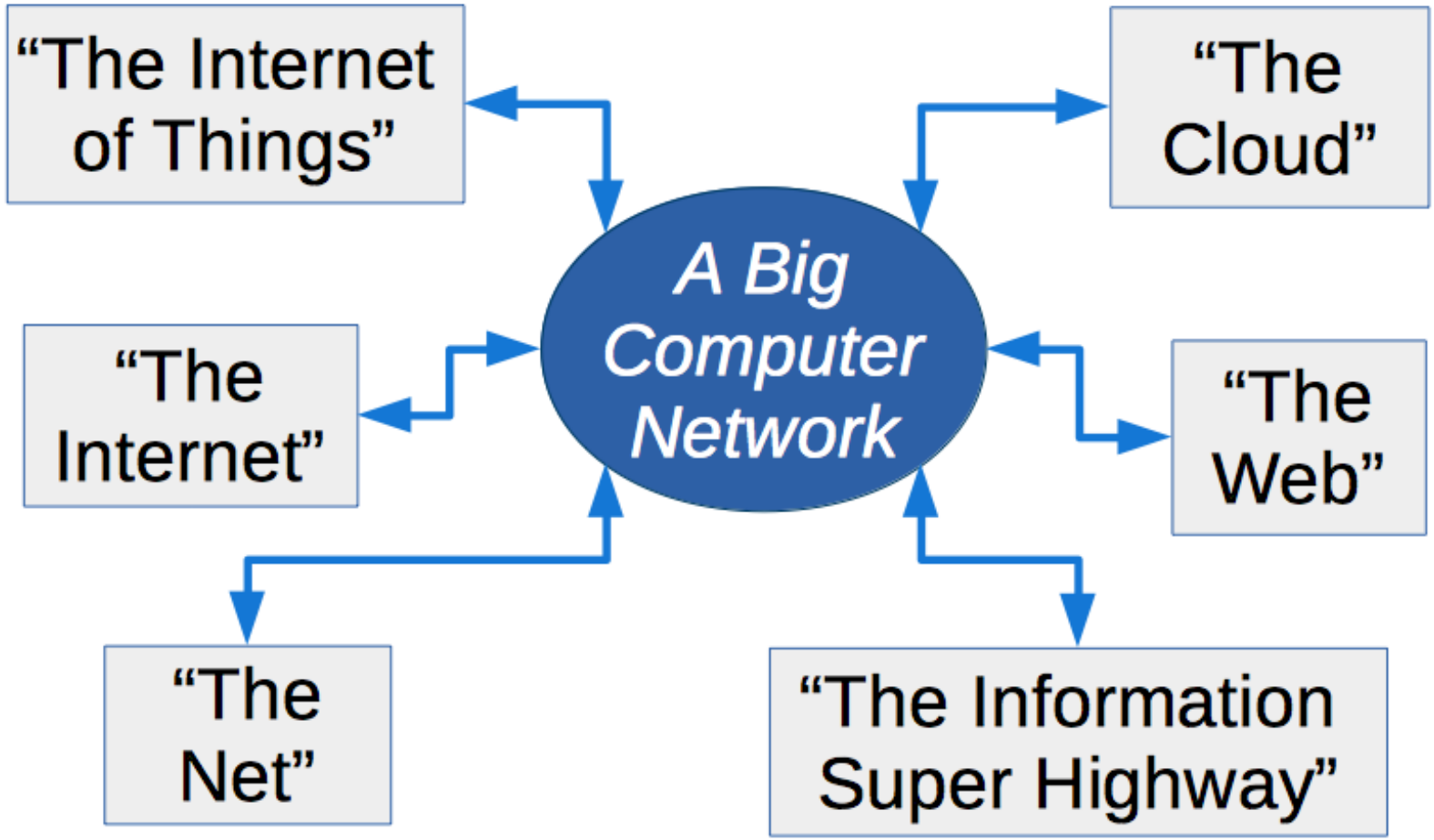
First

- No popular convention when creating keywords
- Investigators call-it *as they like* ...
- No fixed language: An investigator's keyword usage must be (already) known to find his/her own articles
- Articles of old-fashioned keywords appear outdated



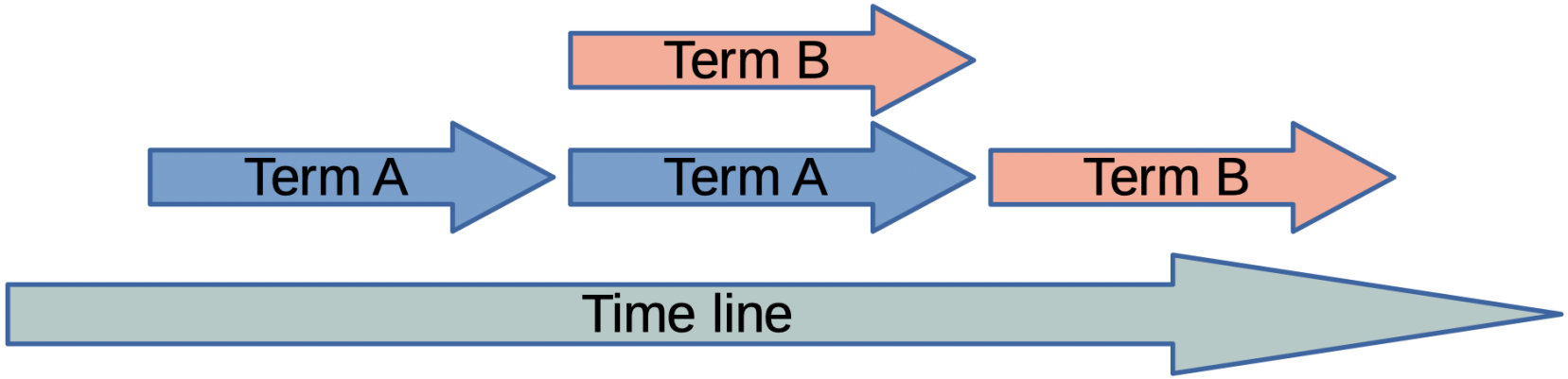
For Example

Different Terms, Same Interest Area



Three Main Problems With Keywords

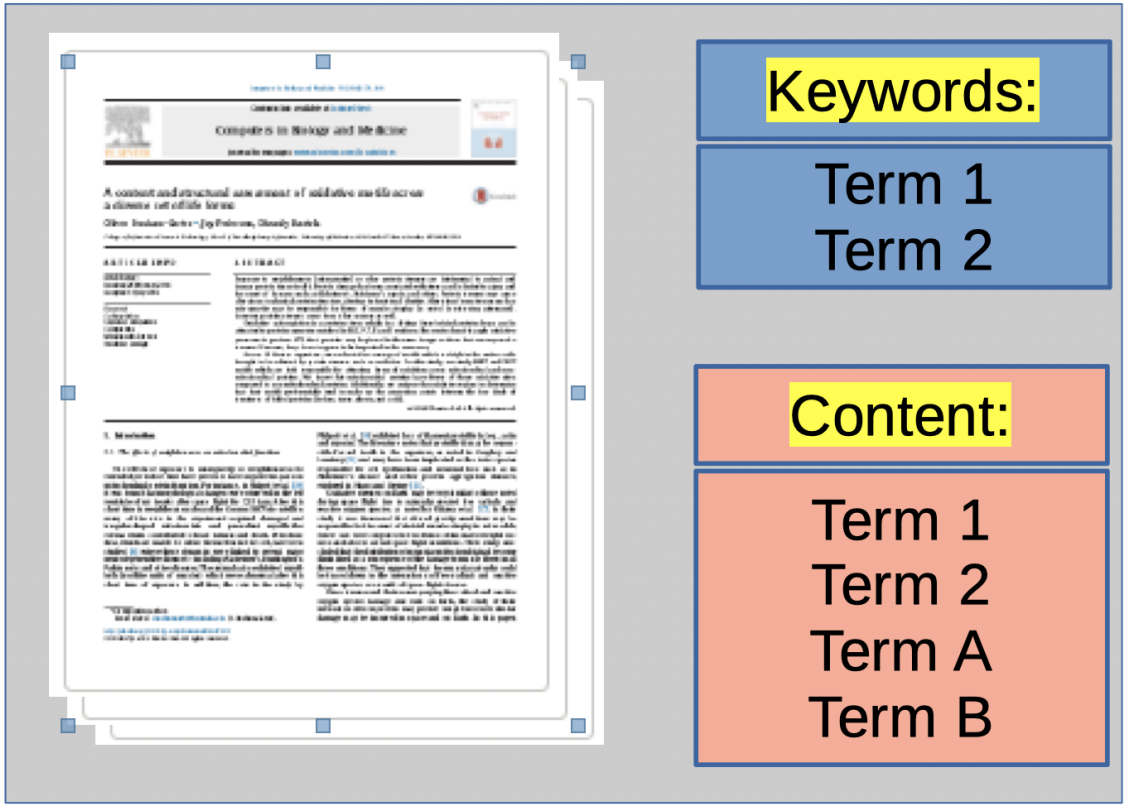
Second



- Language of discipline evolves in tandem with growth
- As keywords evolve, there is a gap created between older research and the contemporary work

Three Main Problems With Keywords

Third Problem



- Articles may contain information which is not indicated by their keywords
- One must be familiar with article to discover information
- This knowledge may be lost to community because search algorithms cannot locate it

Research Interest

Questions

- What relationships exist between works that follow similar themes?
- Can we find those relationships with a simple (low-tech) bag-of-words approach?
- What can we learn from these relationships

Method and Prototype Tool

- We developed a method and tool, *BeagleTM*, to apply a bag of words approach to finding relationships between papers.

A Text Mining Approach By *BeagleTM*

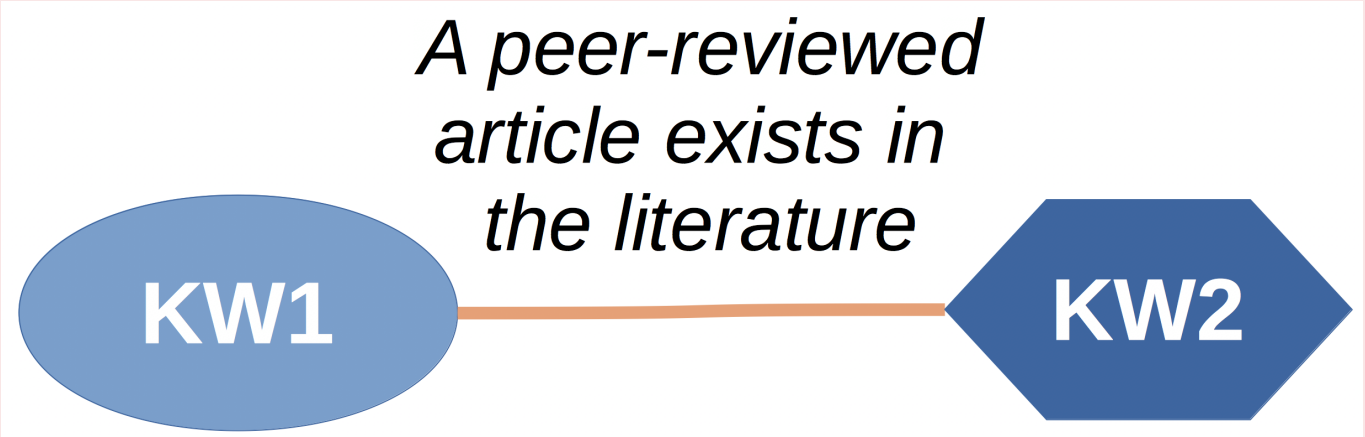
- *BeagleTM* processes a downloaded corpus of articles from NCBI's PubMed archive
  - Contents: at least 27 million articles from more than 7,000 journals
  - About 4 million of these articles are full text
- Tool is designed to show links (relationships) between terms, according to PubMed literature

- *BeagleTM* processes articles individually; arbitrarily large corpus' could be used
- Other tools unsuccessfully tried to load all corpus data into memory before completing analysis

Linked By Articles in the Literature

- Tool is designed to show links between terms according to discussion in PubMed literature
- Supervised: need to know keywords for analysis
  - For example: want to find links between genes **KW1** and **KW2**
- Relationship networks are drawn between articles according to a list of user-defined keywords

Simple Relationship Network



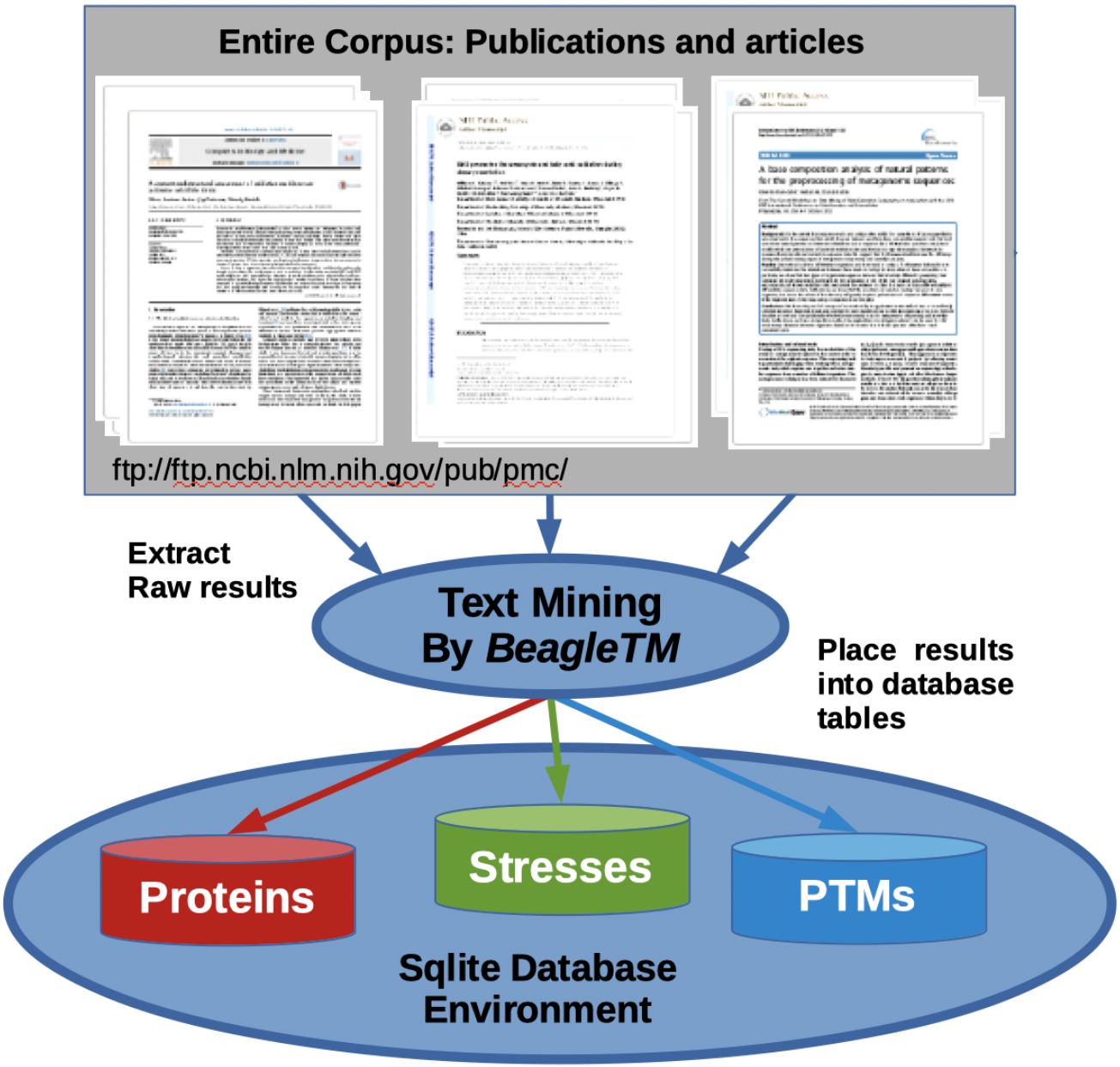
A Concentration on Abstracts

Why article abstracts?

- Abstracts are about 255 words concerning contents
- Well chosen words; all details in abstract are likely important features in article
- Direct language; no sarcasm, misleading statements, etc.
- An idea in abstract ought to be supported by discussion
- Likely free of extraneous text that could confuse other text miners

Overview of Method

To build relationships in Bioinformatics



Article Meta-Data

PubMed articles


```
<PubmedArticle>
    <PMID Version="1">26448611</PMID>
        <Journal>
            <Title>PloS one</Title>
            <ISOAbbreviation>PLoS ONE</ISOAbbreviation>
        </Journal>
        <ArticleTitle>MsrA Overexpression Targeted to the Mitochondria, but
        Not Cytosol, Preserves Insulin Sensitivity in Diet-Induced Obese
        Mice.</ArticleTitle>
        <Abstract>
          <AbstractText>There is growing evidence that oxidative stress
          plays an integral role in the processes by which obesity causes
          type 2 diabetes. We previously identified that mice lacking the
          protein oxidation repair enzyme methionine sulfoxide reductase A
          (MsrA) are particularly prone to obesity-induced insulin
          resistance suggesting an unrecognized role for this protein in
          metabolic regulation. The goals of this study were to test whether
          increasing the expression of MsrA in mice can protect against
          obesity-induced metabolic dysfunction and to elucidate the
          potential underlying mechanisms. ... Our data suggest that
          identification of targets that maintain and regulate the integrity
          of the mitochondrial proteome, particular against oxidative
          damage, may play essential roles in the protection against
          metabolic disease.
        </AbstractText>
      </Abstract>
</PubmedArticle>
```

PMID

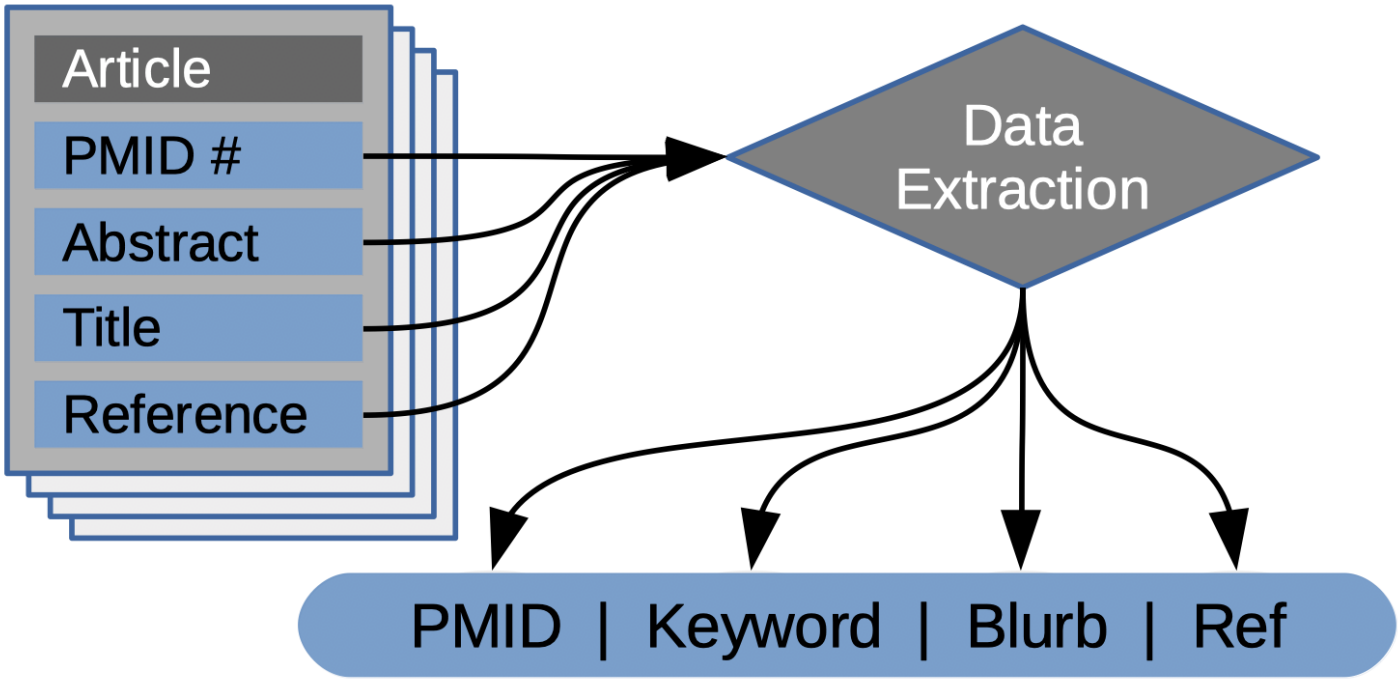
J. Title

A. Title

Abstract

For Each Article

Parse our defined keywords



- Supervised: The keywords for an analysis must be defined

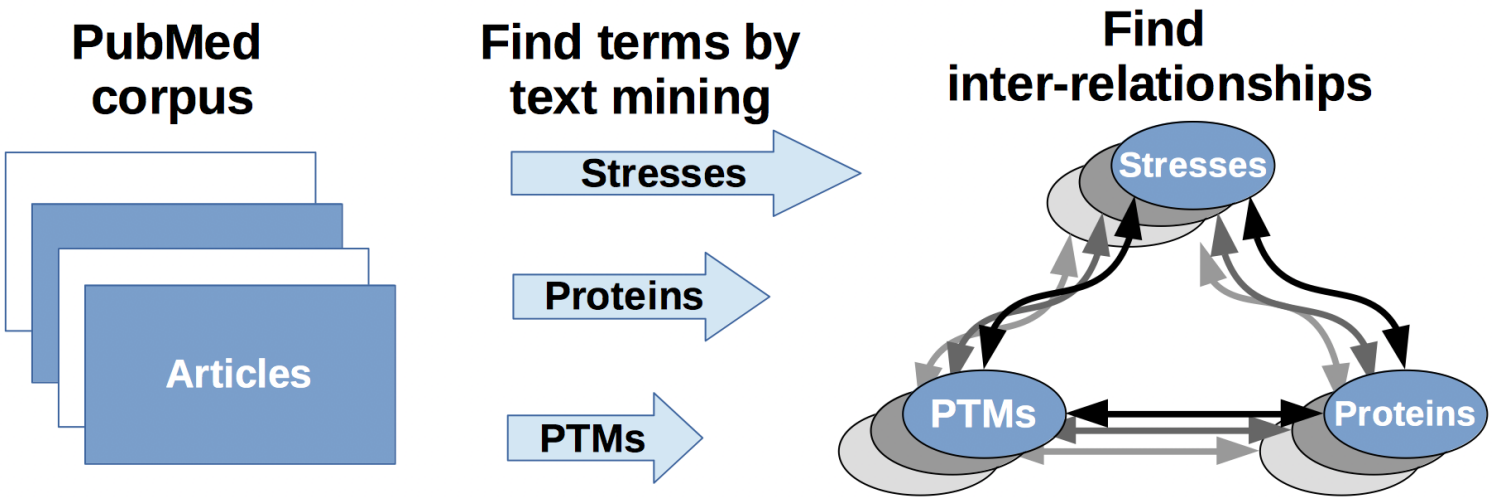
Database Support
Basic SQL

```
CREATE TABLE Functional (
  pmid varchar PRIMARY KEY,
  funct varchar NOT NULL,
  count integer NOT NULL,
  blurb text NOT NULL,
  journal text NOT NULL );
```

```
CREATE TABLE Stress (
   pmid varchar PRIMARY KEY,
   stress varchar NOT NULL,
   count integer NOT NULL,
   blurb text NOT NULL,
   journal text NOT NULL);
```

- SQL code to create two of the tables in our database: all tables similar.
- The PMID (and keyword) of articles containing the user-selected keywords, was recorded in an SQL base and PMID numbers were queried to build relationship networks

Relationship Networks



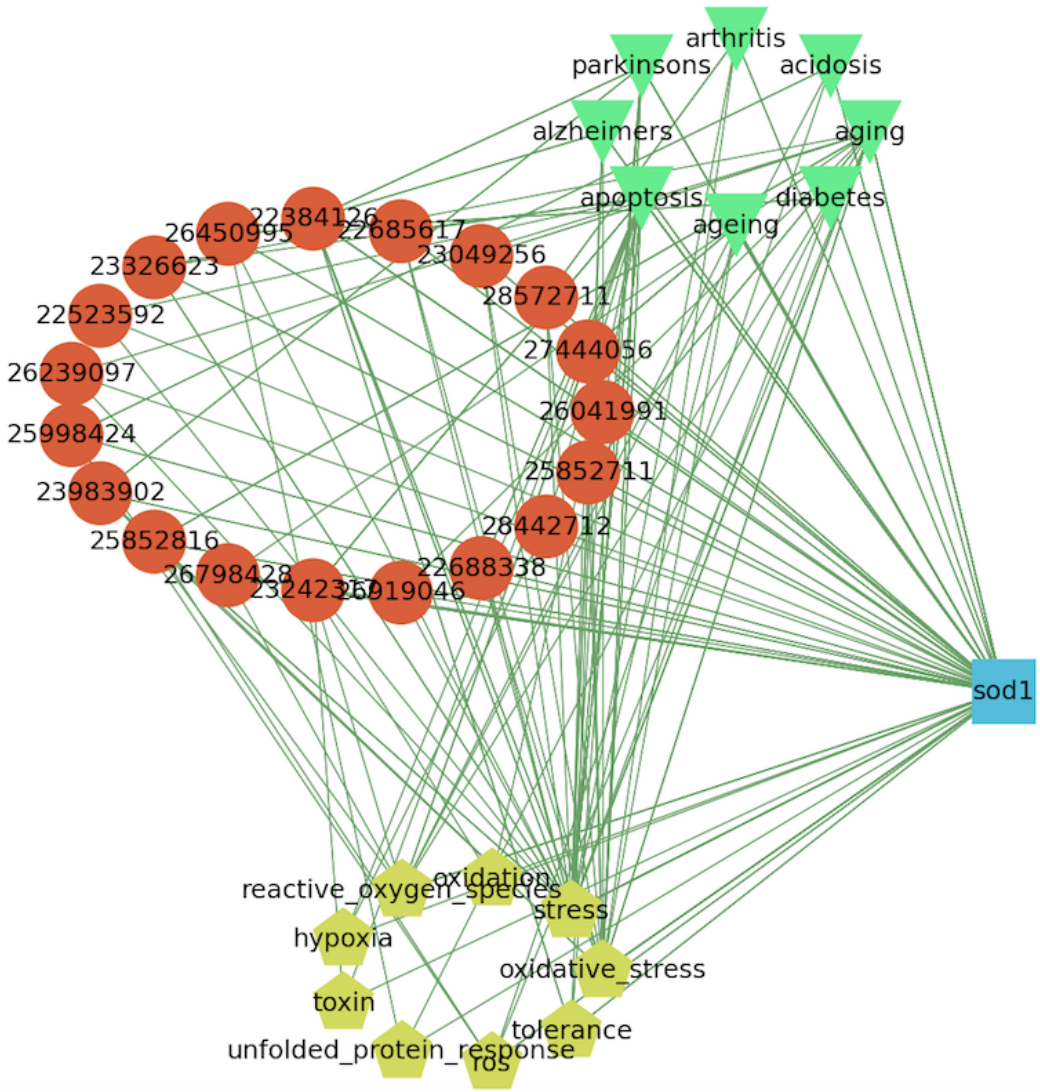
- We provide the terms *Stress*, *Proteins*, and *PTMs* (post-translational modifications) from bioinformatics
- An edge means that at least one study exists to connect two keyword nodes
- Terms that are guilty by association: Found studies are likely to connect the relevant terms

Our Keywords

Rubric	Sample	Total keywords
Diseases-specific	acidosis, ageing, Alzheimer's, apoptosis, arthritis, Crohn's, diabetes, obesity, Parkinson's and others	46
Mt Gene Symbols	oat, pc, opa1, cs, mut, msra, phb, sod1, mtor, aldh2 and others	619
PTMs (general types)	acetylation, glycosylation, methylation, oxidation, phosphorylation and others	35
Stresses	hypoxia, oxidation, oxidative stress, ROS (reactive oxygen species), tolerance, toxin, unfolded protein response and others	47

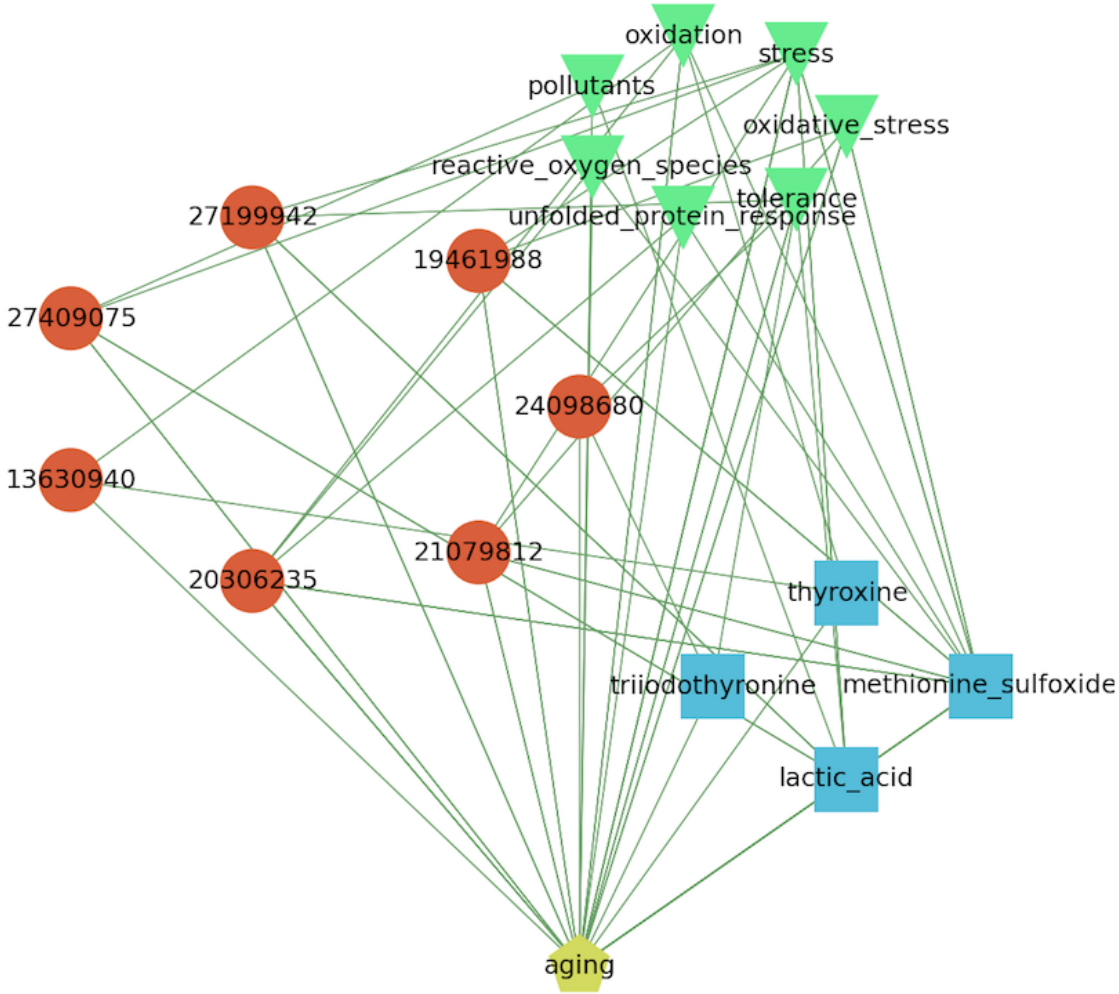
- Keyword for my own research: wanted to focus diseases and build any possible relationship networks using these other keywords, according to the literature

Sod1 (protein) Relationship to Stress and Disease



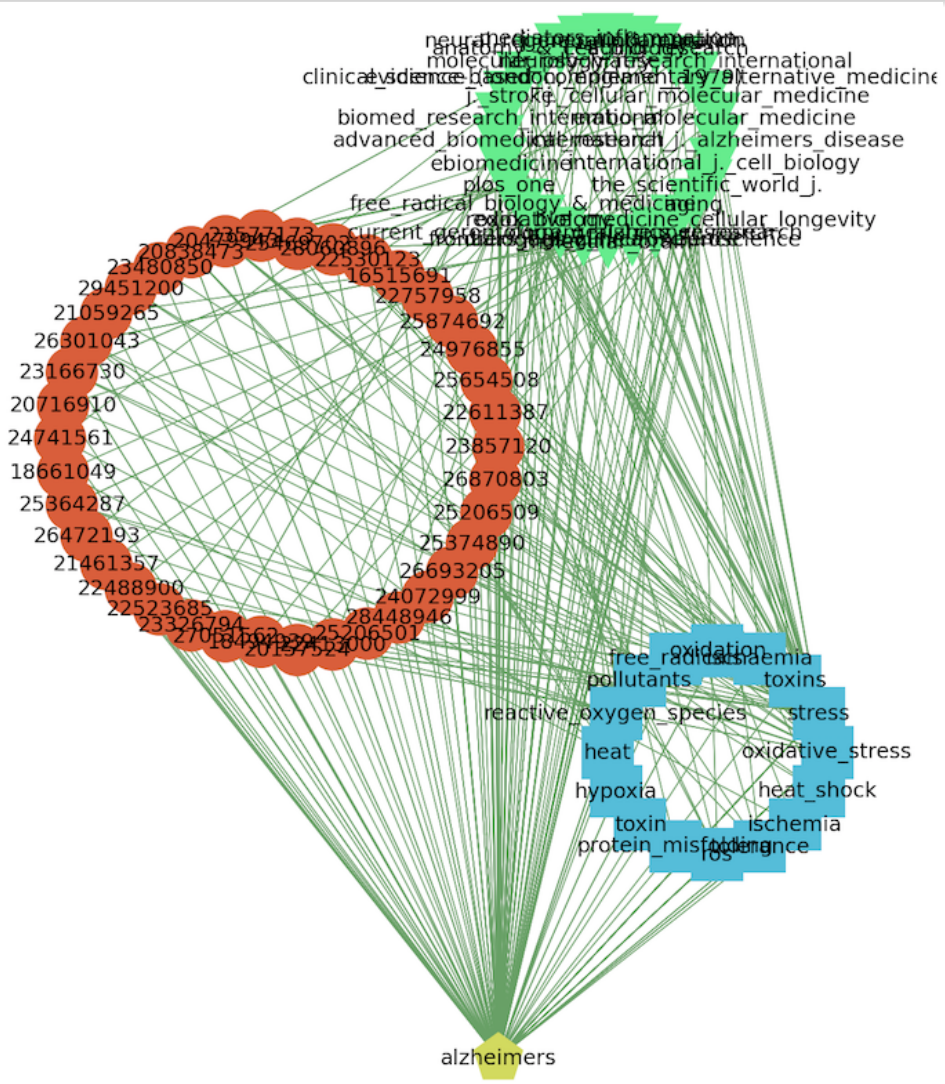
The circles and pentagons denote the PMIDs and stresses, the triangles denote the disorders that have documented relationships to the other nodes. All edges denote that terms are connected by at least one common article.

Aging Relationship to Stress-type and PTMs



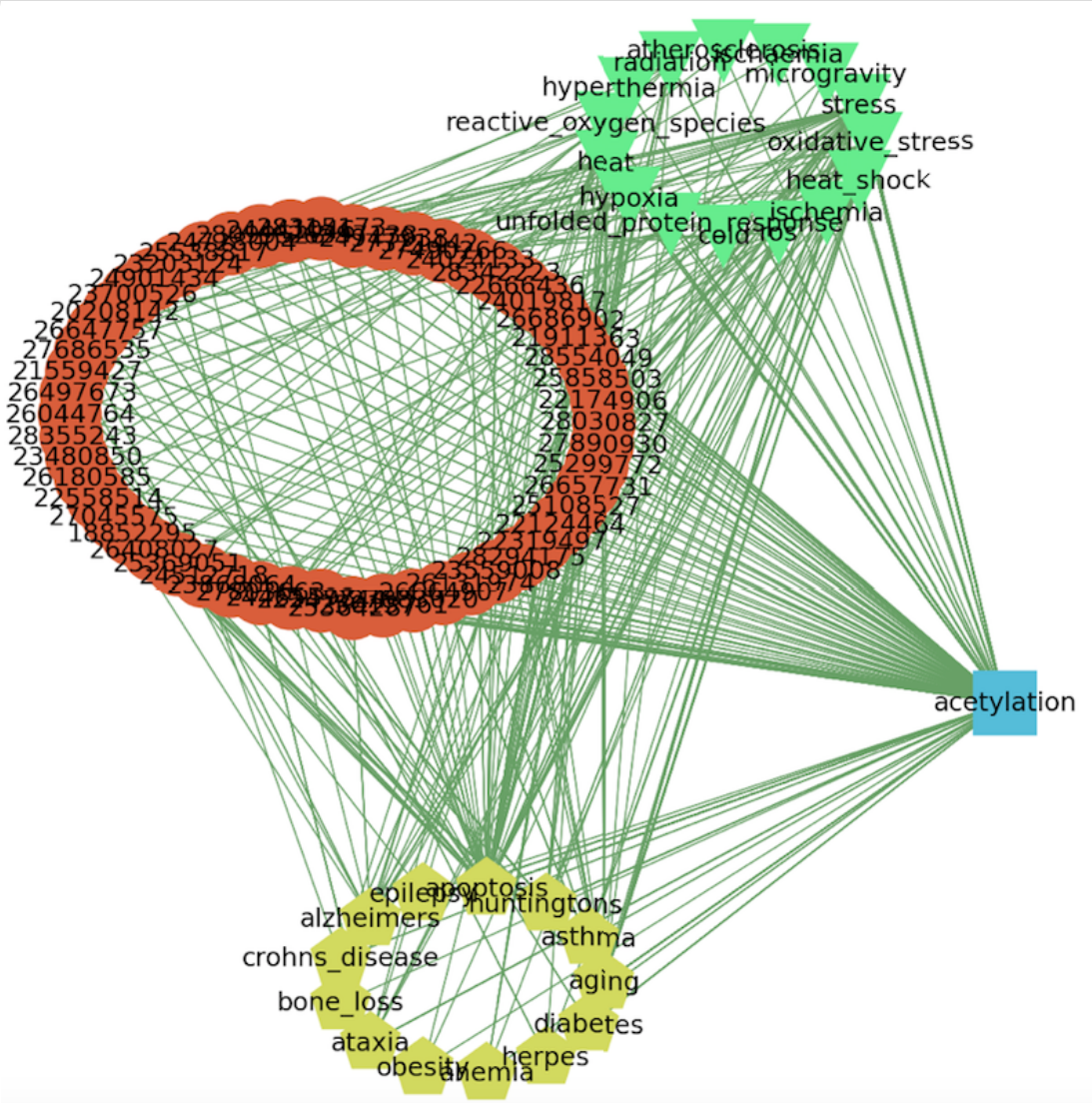
The red circles represent the PMID numbers for PubMed articles, the blue squares indicate PTMs, the green triangles denote the stress-factors and the mustard pentagons correspond to the ailment by name, to which all elements are related by the literature.

Alzheimer's Relationship to Stress and Disease



The red circles represent the PMID numbers for PubMed articles, the blue squares indicate stress-factors, the green triangles denote the journal names, and the mustard pentagon correspond to the ailment by name

Acetylation's Relationship to PTMs and Disease



The red circles represent the PMID numbers for PubMed articles, the blue square indicates a PTM, the green triangles denote stresses, and the mustard pentagon correspond to the ailment by name.

Conclusions

- *BeagleTM* is a supervised method that couples text mining with database query to extract relationship networks from the literature
- Networks inform of studies sharing common themes (since they have the same actors who play roles)
- An edge means that at least one study exists to connect two keyword nodes
- The actual story behind the edge must be explained by returning to the original sources; the model cannot explain the relationship

Thank You! Questions?

obonhamcarter@allegheny.edu

Professional Page:
https://www.cs.allegheny.edu/sites/obonhamcarter/