MOLIERE

Automatic Biomedical Hypothesis Generation System

Justin Sybrandt

Clemson University
School of Computing

Michael Shtutman

University of South Carolina
College of Pharmacy

Ilya Safro

Clemson University
School of Computing

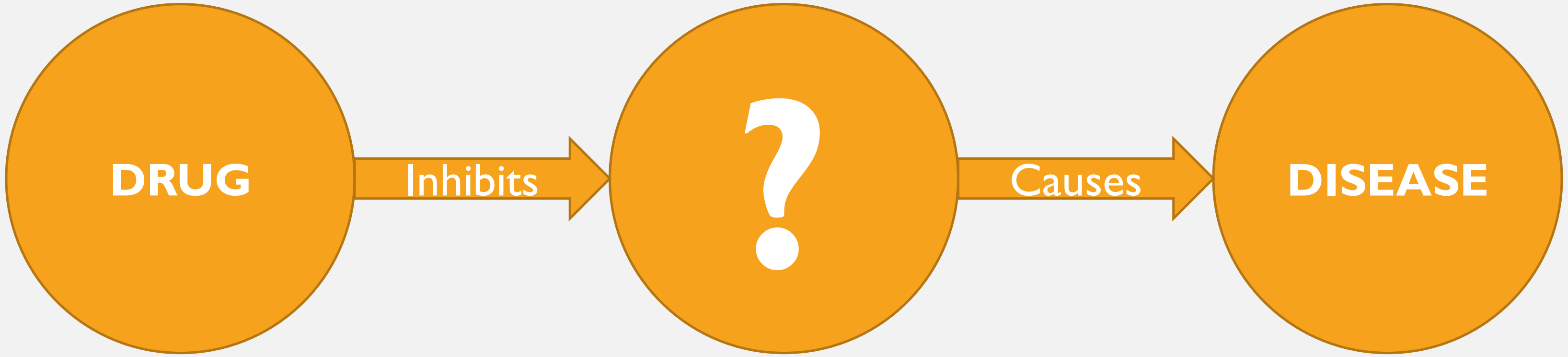
UNDISCOVERED PUBLIC KNOWLEDGE

- Proposed by Swanson in 1986
- The set of public knowledge is too large
- Contains implicit connections

HUMAN LIMITS

- No person can read everything
  - 2,000 – 4,000 new papers ***daily***
- People Specialize
  - Limits knowledge sharing across disciplines
- Bias and Inconsistent Assumptions

WHAT IS A HYPOTHESIS?

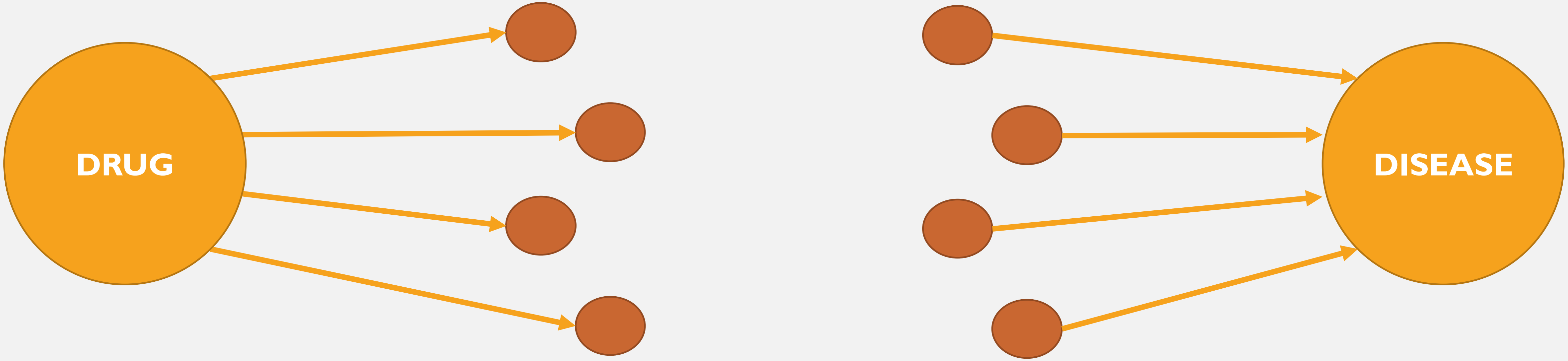


- An idea or explanation for something that is based on known facts but has not yet been proven
  - (Definition of "hypothesis" from the Cambridge Academic Content Dictionary © Cambridge University Press)

AUTOMATIC
HYPOTHESIS GENERATION

Network of Medical Objects

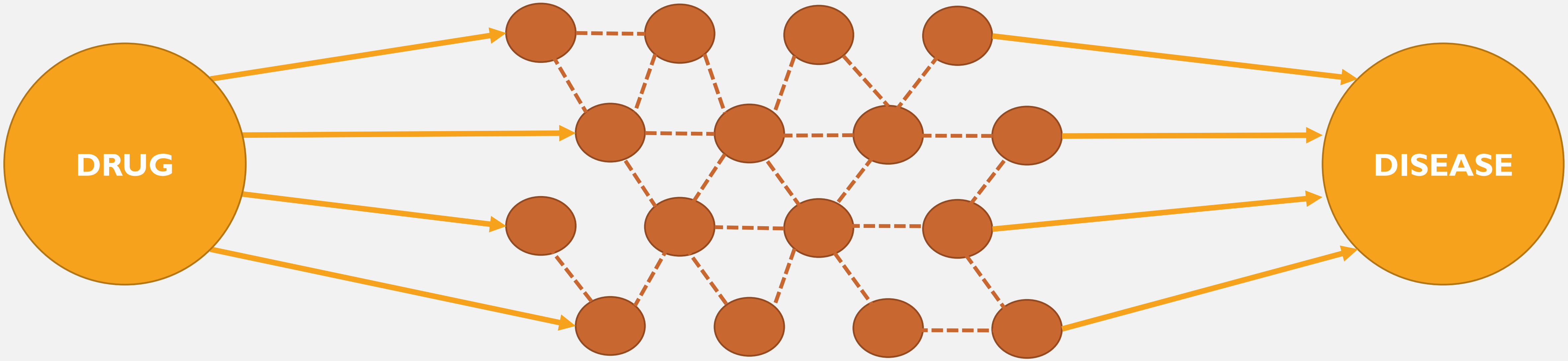
Connections Not Well Studied



AUTOMATIC HYPOTHESIS GENERATION

Network of Medical Objects

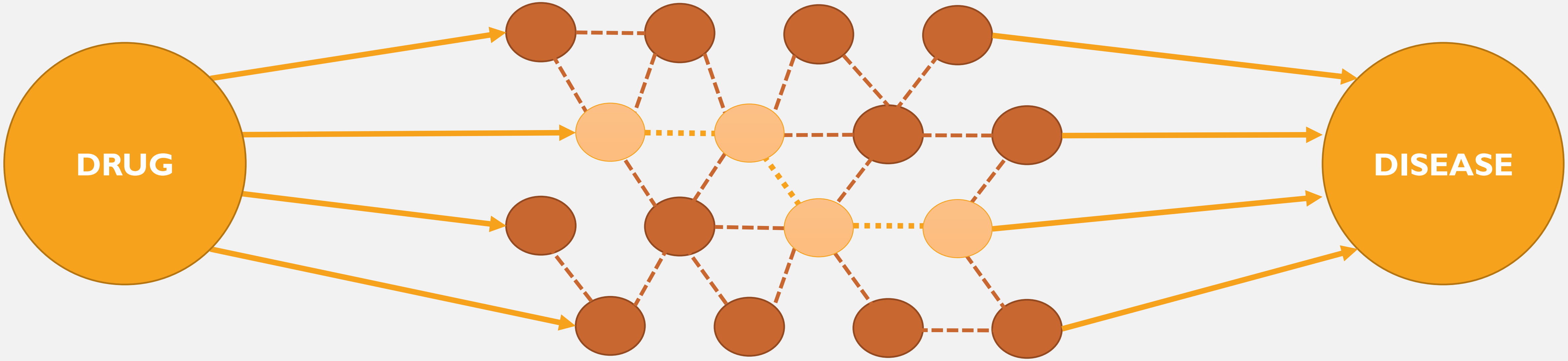
Connections Not Well Studied



AUTOMATIC
HYPOTHESIS GENERATION

Network of Medical Objects

Connections Not Well Studied



RELATED APPROACHES

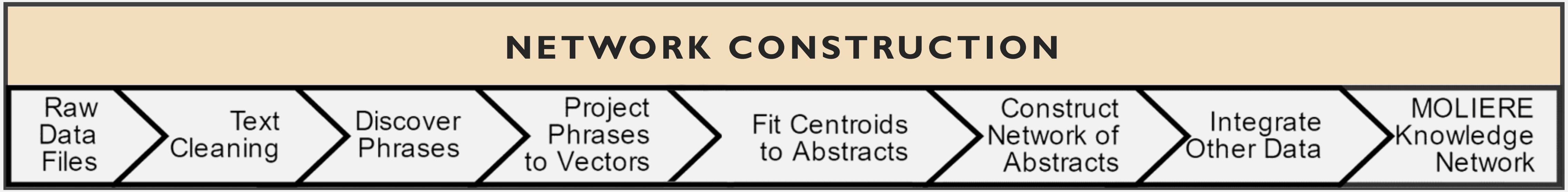
Methodology	Highlights
ARROWSMITH	• One of the first hypothesis generation systems. • Found link between Fish Oil and Raynaud's Disease.
DiseaseConnect	• Finds connections between genes and diseases. • Displays information in an interactive prompt.
BioLDA	• Constructs high quality topic models aided by domain-specific information. • Identified link between Venlafaxine and HTR1A.

RELATED APPROACHES

Methodology	Limitation	Reference
ARROWSMITH	Limited Document Set	Neil R Smalheiser and Don R Swanson. 1998.
DiseaseConnect	Limited Document Set	Chun-Chi Liu et al. 2014.
BioLDA	Limited Vocabulary	Huijun Wang et al. 2011.

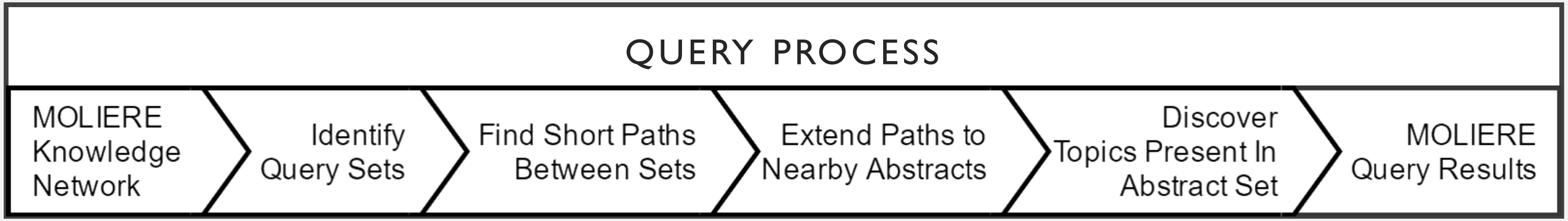
NEW APPROACH

- Construct a large network
- Identify meaningful paths
- Extend paths to neighboring nodes
- Mine neighborhoods for important topics



NETWORK CONSTRUCTION

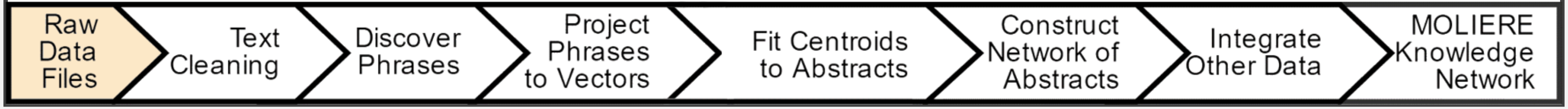
Raw Data Files → Text Cleaning → Discover Phrases → Project Phrases to Vectors → Fit Centroids to Abstracts → Construct Network of Abstracts → Integrate Other Data → MOLIERE Knowledge Network



QUERY PROCESS

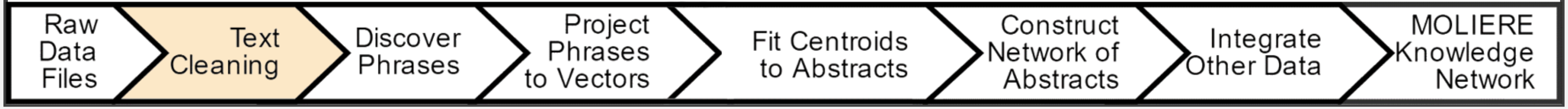
MOLIERE Knowledge Network → Identify Query Sets → Find Short Paths Between Sets → Extend Paths to Nearby Abstracts → Discover Topics Present In Abstract Set → MOLIERE Query Results

NETWORK CONSTRUCTION


Raw Data Files
Text Cleaning
Discover Phrases
Project Phrases to Vectors
Fit Centroids to Abstracts
Construct Network of Abstracts
Integrate Other Data
MOLIERE Knowledge Network

- National Library of Medicine (NLM)
- National Center for Biotechnology Information (NCBI)
- MEDLINE
  - 25 Million Documents
  - Titles and Abstracts

NETWORK CONSTRUCTION


Raw Data Files

Text Cleaning

Discover Phrases

Project Phrases to Vectors

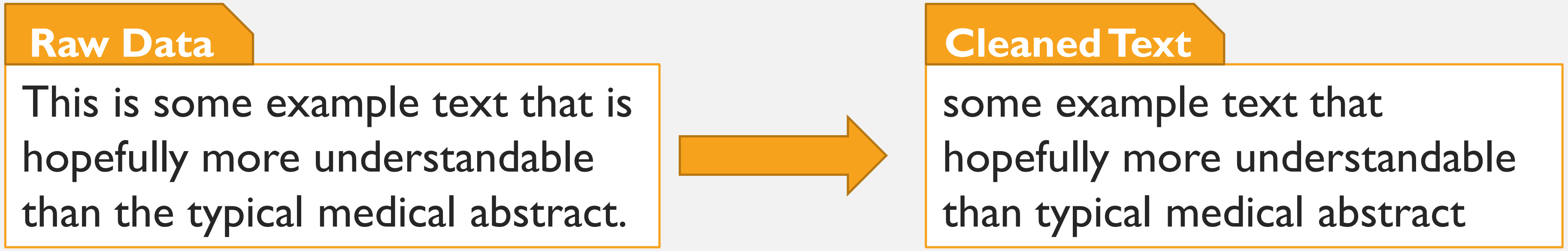
Fit Centroids to Abstracts

Construct Network of Abstracts

Integrate Other Data

MOLIERE Knowledge Network

- SPECALIST NLP TOOLSET
- Natural Language ToolKit (NLTK)

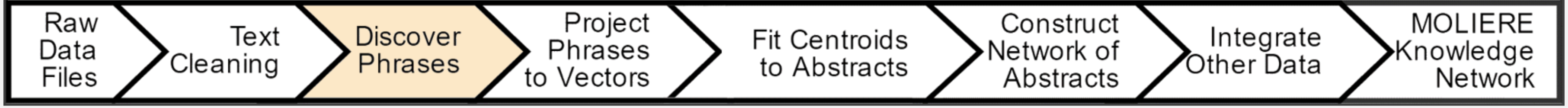

Raw Data

This is some example text that is hopefully more understandable than the typical medical abstract.

Cleaned Text

some example text that hopefully more understandable than typical medical abstract

NETWORK CONSTRUCTION


Raw Data Files
Text Cleaning
Discover Phrases
Project Phrases to Vectors
Fit Centroids to Abstracts
Construct Network of Abstracts
Integrate Other Data
MOLIERE Knowledge Network

- Topical Pattern Mining
  - Groups together common phrases
  - Creates 2,3,…,n-grams

Cleaned Text

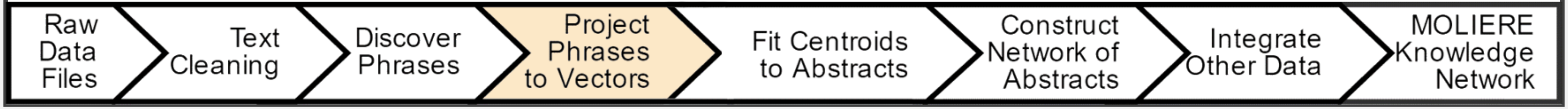
some example text that hopefully more understandable than typical medical abstract

→

Phrases

example text hopefully **more understandable** typical **medical abstract**

NETWORK CONSTRUCTION



Raw Data Files

Text Cleaning

Discover Phrases

Project Phrases to Vectors

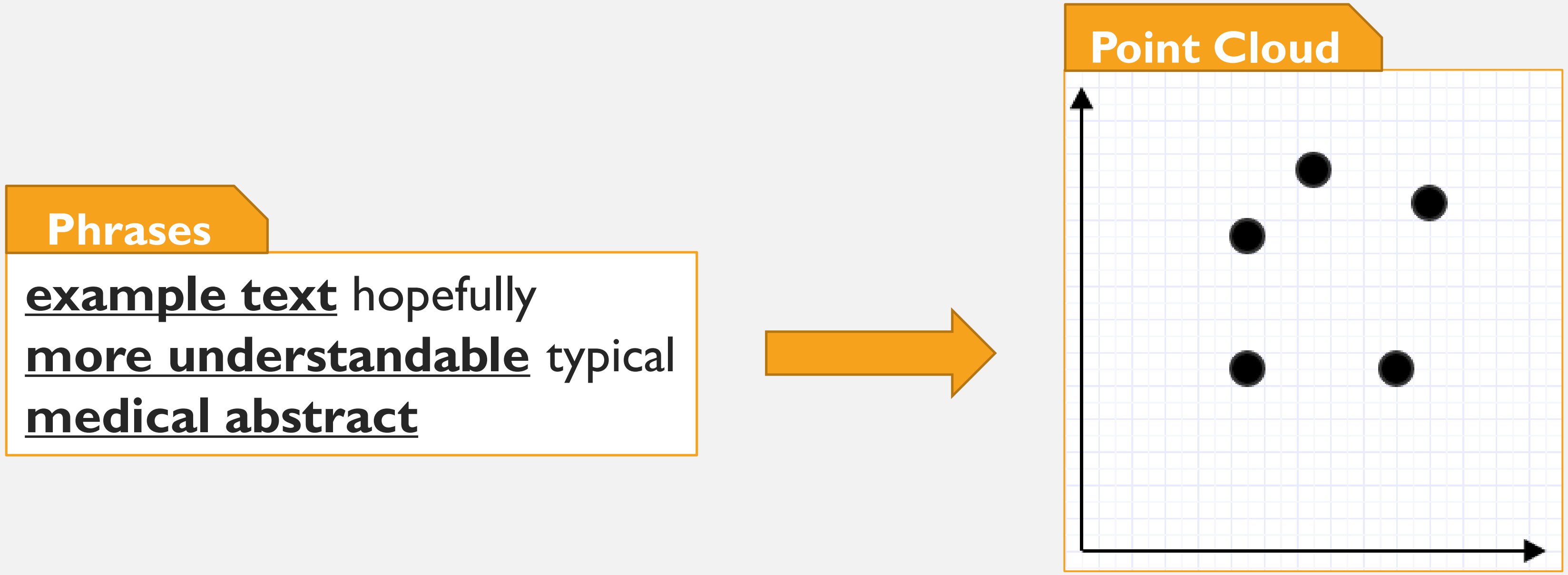
Fit Centroids to Abstracts

Construct Network of Abstracts

Integrate Other Data

MOLIERE Knowledge Network

- FastText: Projects phrases into real valued vector space
- Long word embedding composed from subwords

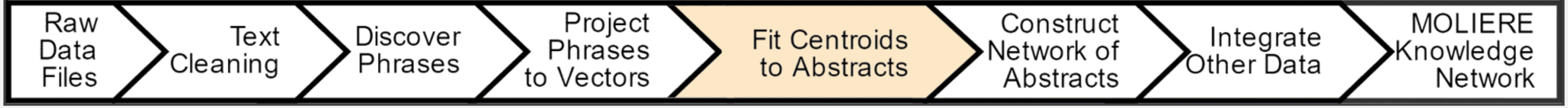


Phrases

<u>example text</u> hopefully **<u>more understandable</u>** typical **<u>medical abstract</u>**

Point Cloud

NETWORK CONSTRUCTION



Raw Data Files

Text Cleaning

Discover Phrases

Project Phrases to Vectors

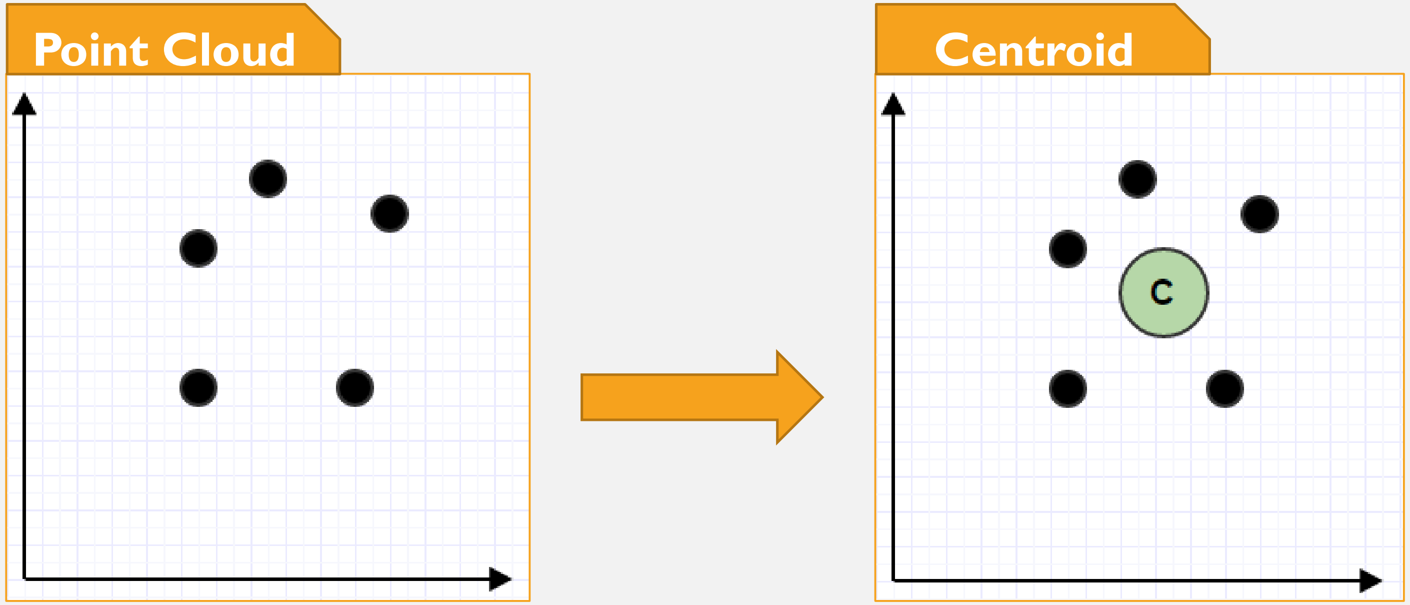
Fit Centroids to Abstracts

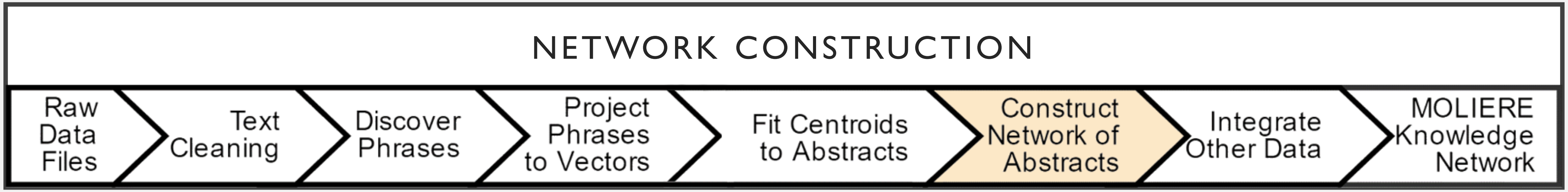
Construct Network of Abstracts

Integrate Other Data

MOLIERE Knowledge Network

- Embed documents by averaging over point clouds





NETWORK CONSTRUCTION

Raw Data Files

Text Cleaning

Discover Phrases

Project Phrases to Vectors

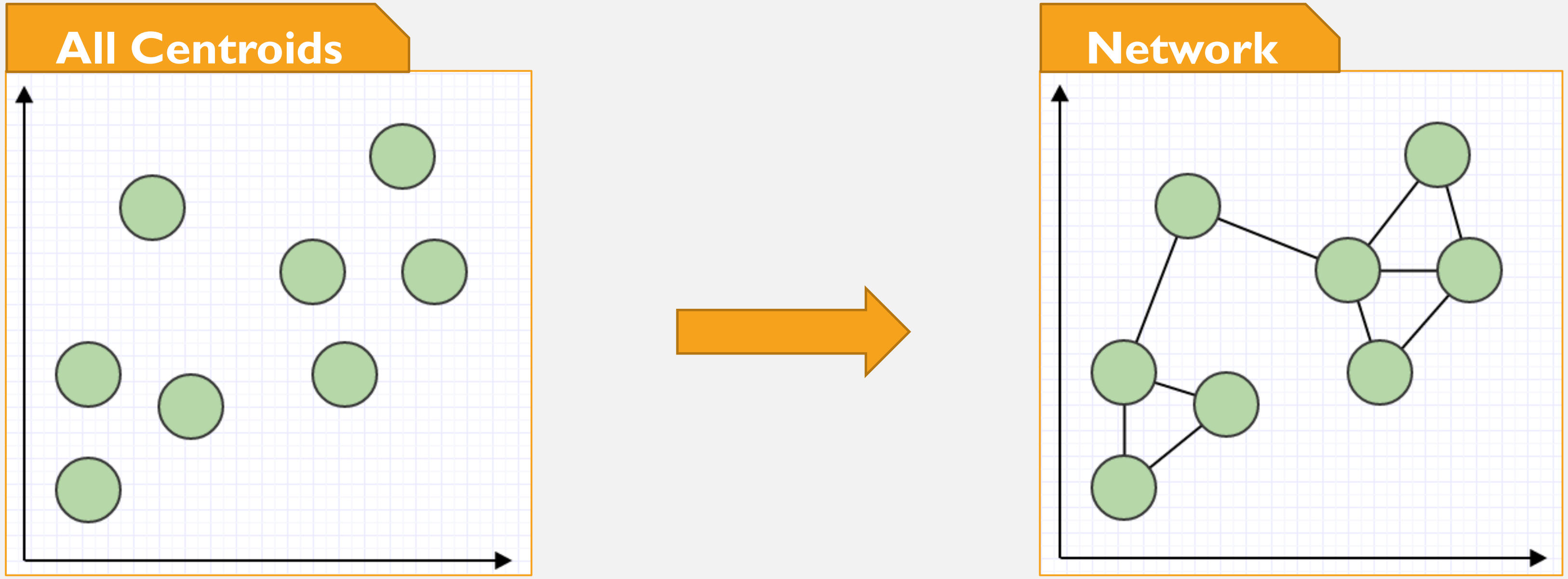
Fit Centroids to Abstracts

Construct Network of Abstracts

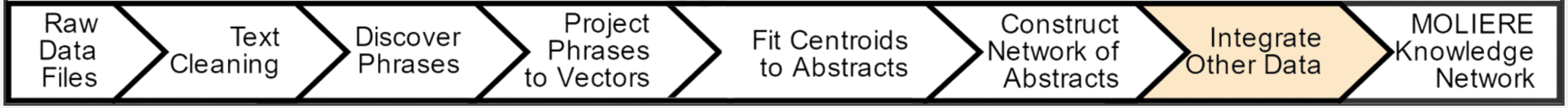
Integrate Other Data

MOLIERE Knowledge Network

- Construct KNN
    - Fast Library for Approximate Nearest Neighbors



NETWORK CONSTRUCTION



Raw Data Files

Text Cleaning

Discover Phrases

Project Phrases to Vectors

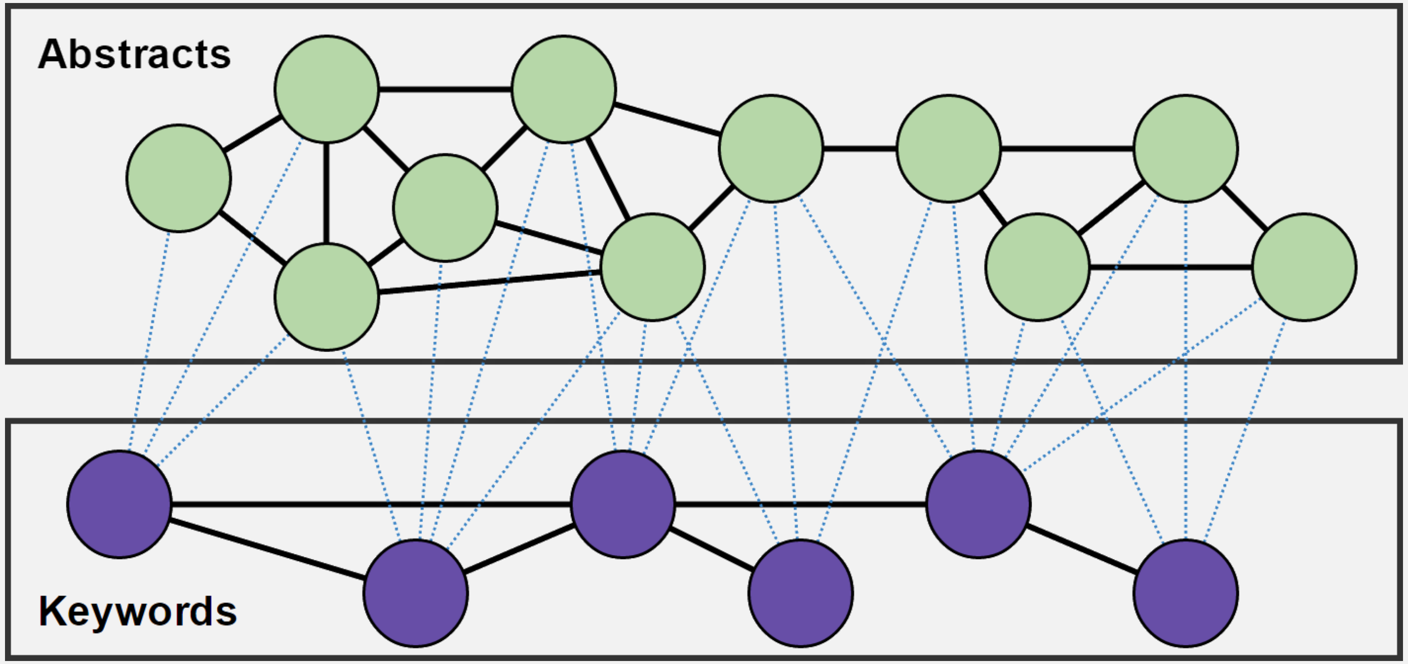
Fit Centroids to Abstracts

Construct Network of Abstracts

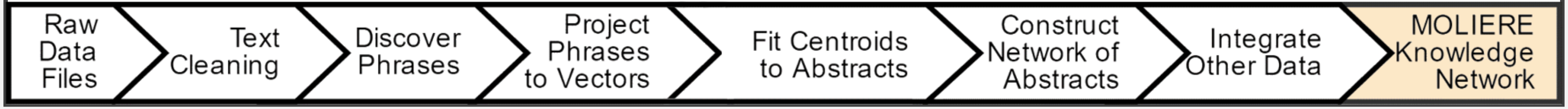
Integrate Other Data

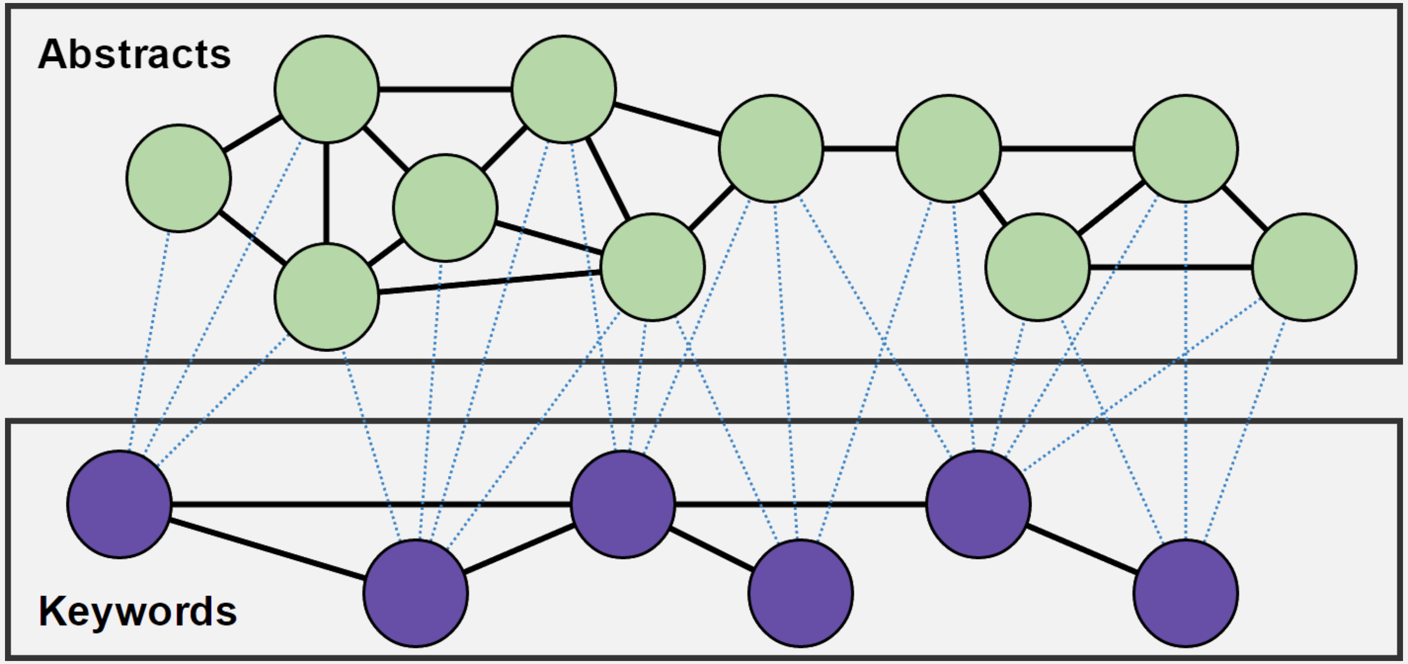
MOLIERE Knowledge Network

- UMLS Metathesarus
  - Curated keyword network
  - 2 Million Nodes
  - Superset of MESH

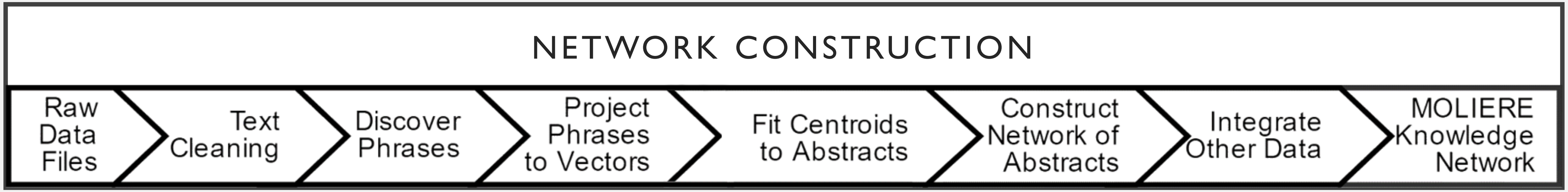


NETWORK CONSTRUCTION


Raw Data Files
Text Cleaning
Discover Phrases
Project Phrases to Vectors
Fit Centroids to Abstracts
Construct Network of Abstracts
Integrate Other Data
MOLIERE Knowledge Network

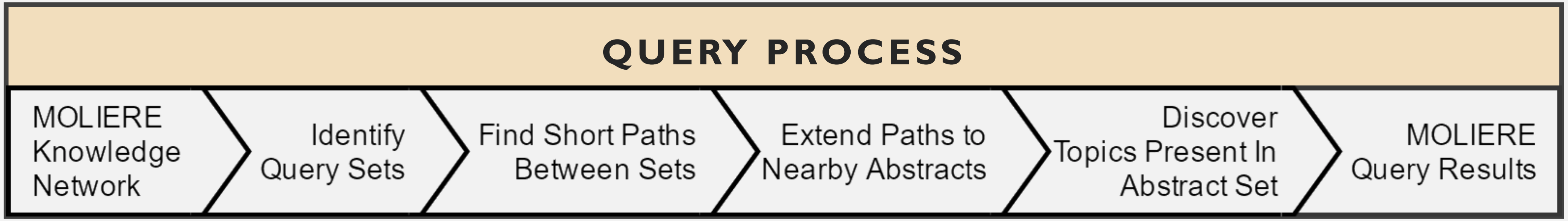


- Edge weight ~ Distance
- Inv. TF-IDF cross-layer edges
- Edges normalized [0,1]



NETWORK CONSTRUCTION

Raw Data Files → Text Cleaning → Discover Phrases → Project Phrases to Vectors → Fit Centroids to Abstracts → Construct Network of Abstracts → Integrate Other Data → MOLIERE Knowledge Network

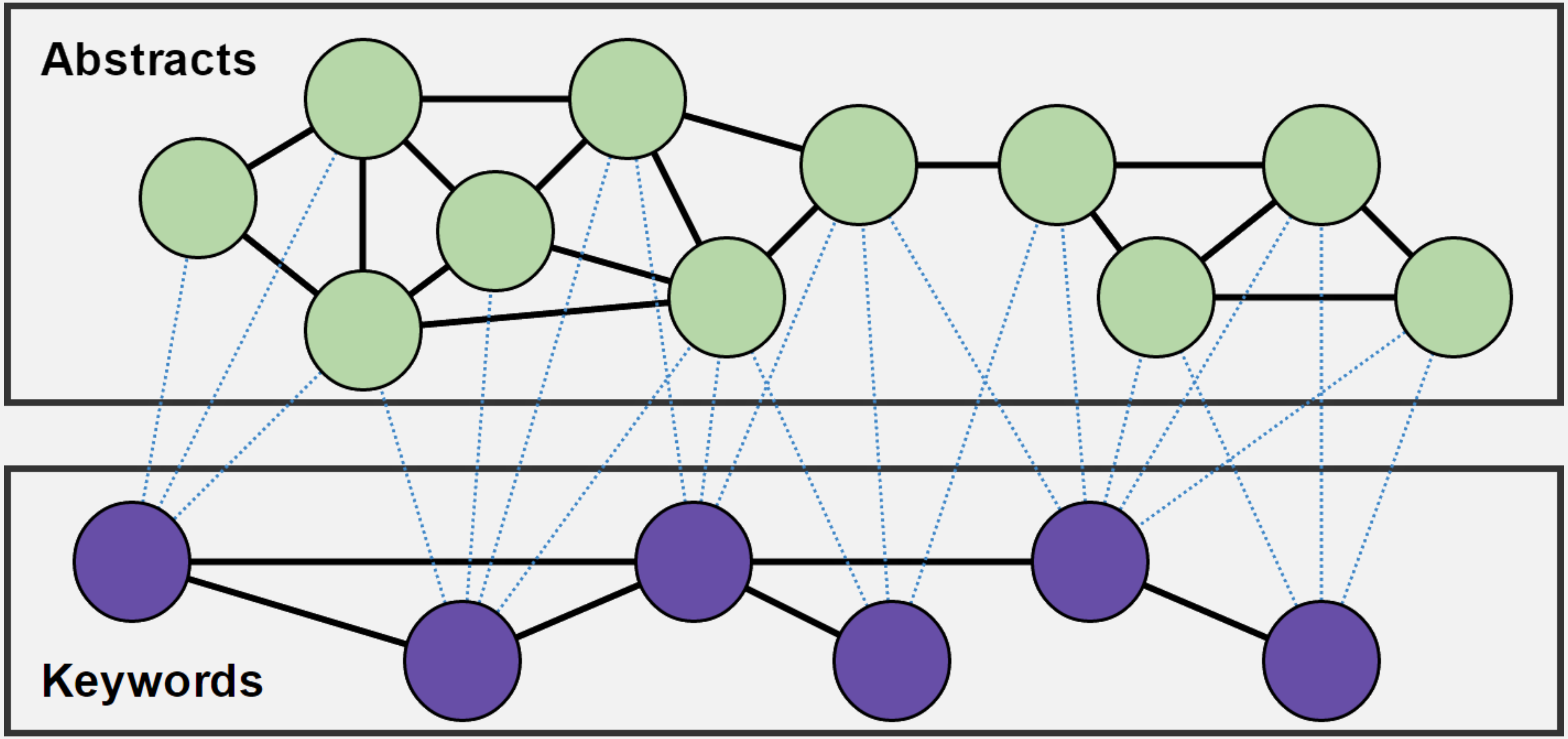


QUERY PROCESS

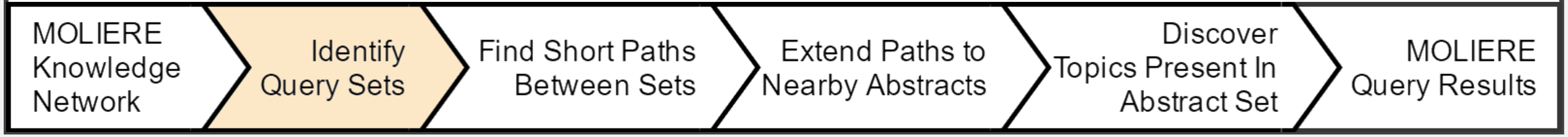
MOLIERE Knowledge Network → Identify Query Sets → Find Short Paths Between Sets → Extend Paths to Nearby Abstracts → Discover Topics Present In Abstract Set → MOLIERE Query Results

QUERY PROCESS

MOLIERE Knowledge Network → Identify Query Sets → Find Short Paths Between Sets → Extend Paths to Nearby Abstracts → Discover Topics Present In Abstract Set → MOLIERE Query Results



QUERY PROCESS


MOLIERE Knowledge Network

Identify Query Sets

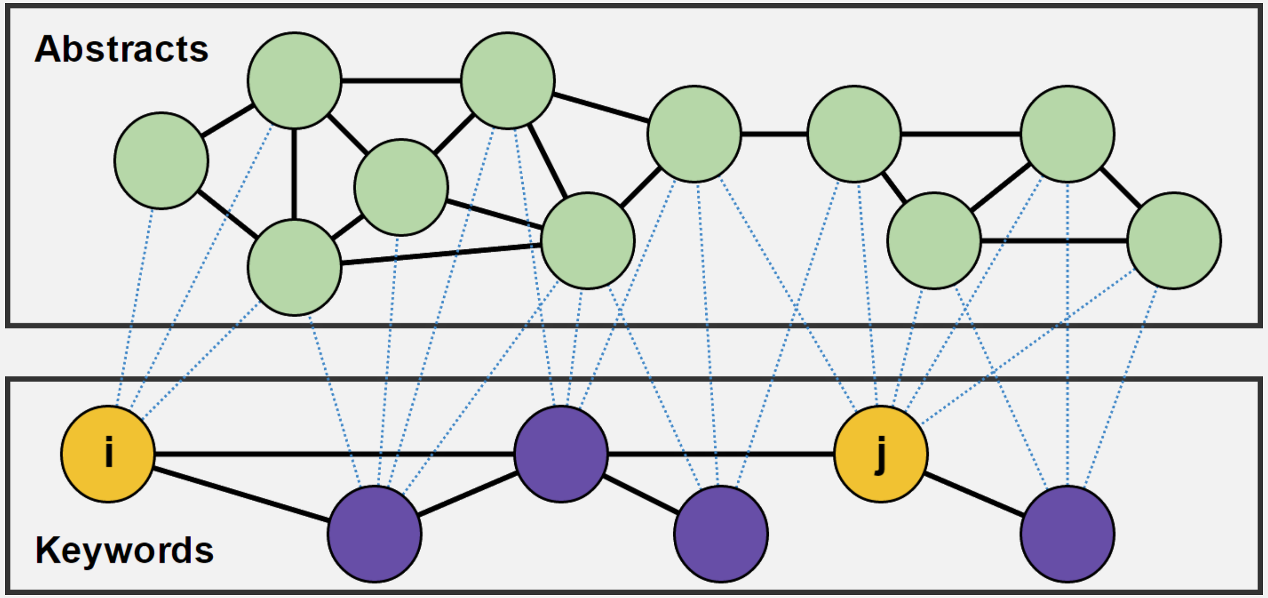
Find Short Paths Between Sets

Extend Paths to Nearby Abstracts

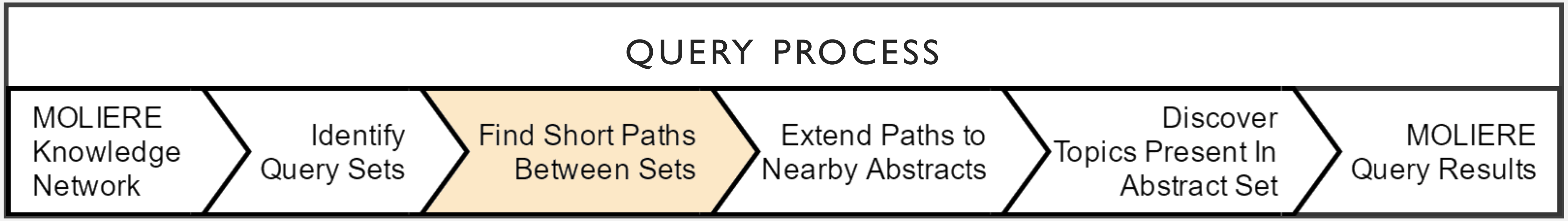
Discover Topics Present In Abstract Set

MOLIERE Query Results

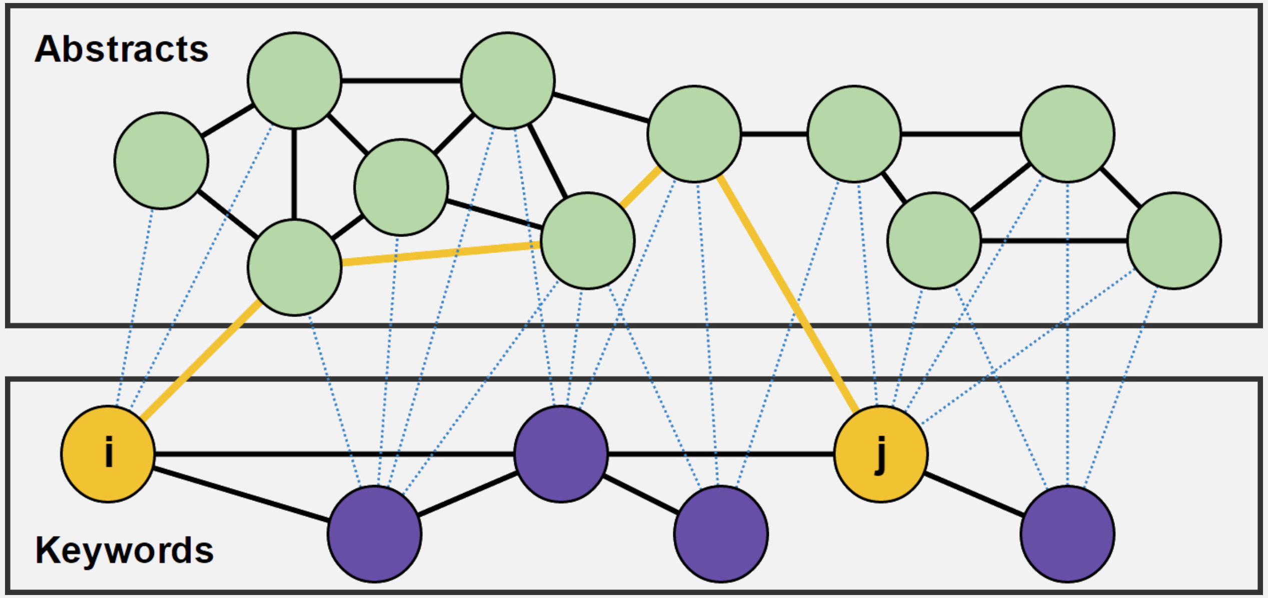
- User selects two nodes
  - Restrained to keywords
- Can generalize to two sets



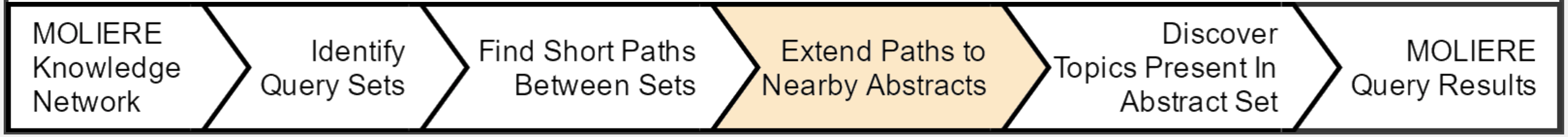
QUERY PROCESS


MOLIERE Knowledge Network
Identify Query Sets
Find Short Paths Between Sets
Extend Paths to Nearby Abstracts
Discover Topics Present In Abstract Set
MOLIERE Query Results

- Identify shortest path between query sets



QUERY PROCESS



MOLIERE Knowledge Network

Identify Query Sets

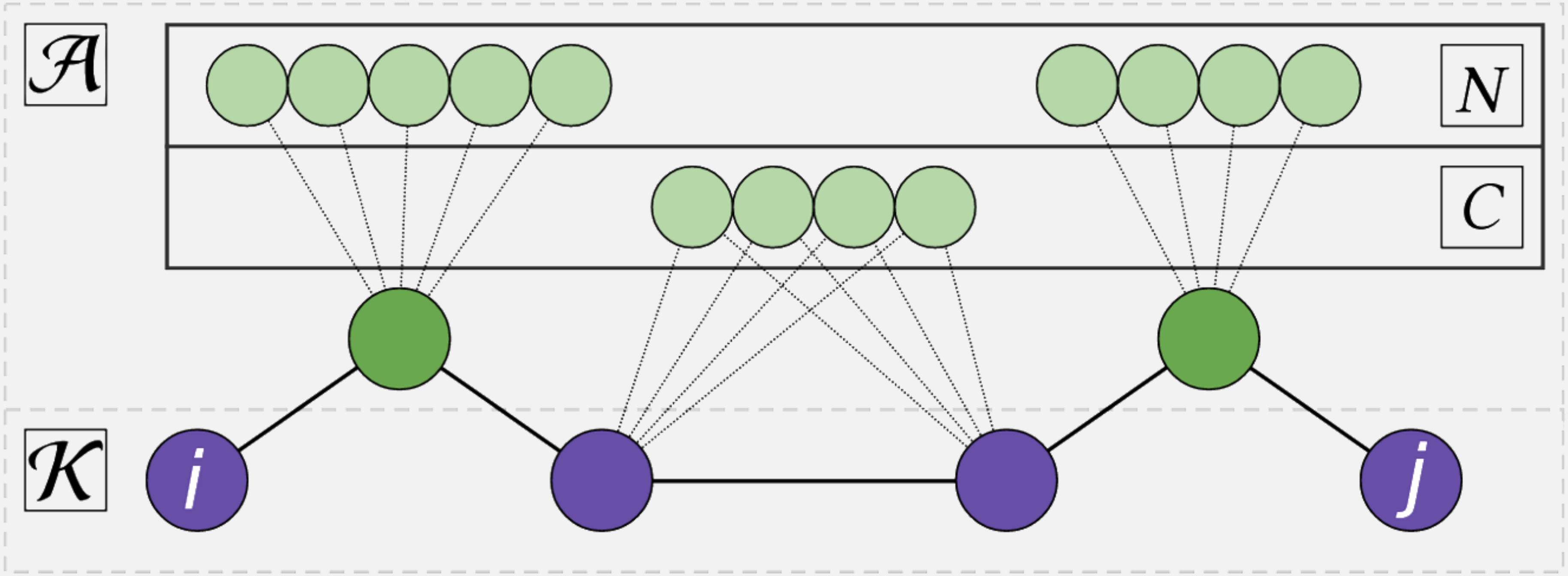
Find Short Paths Between Sets

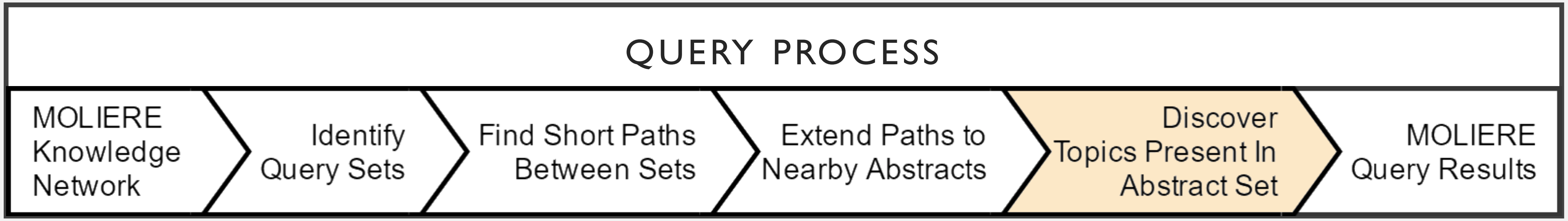
Extend Paths to Nearby Abstracts

Discover Topics Present In Abstract Set

MOLIERE Query Results

- *N*: Abstracts close to those in original path
- *C*: Abstracts which share path-adjacent keywords



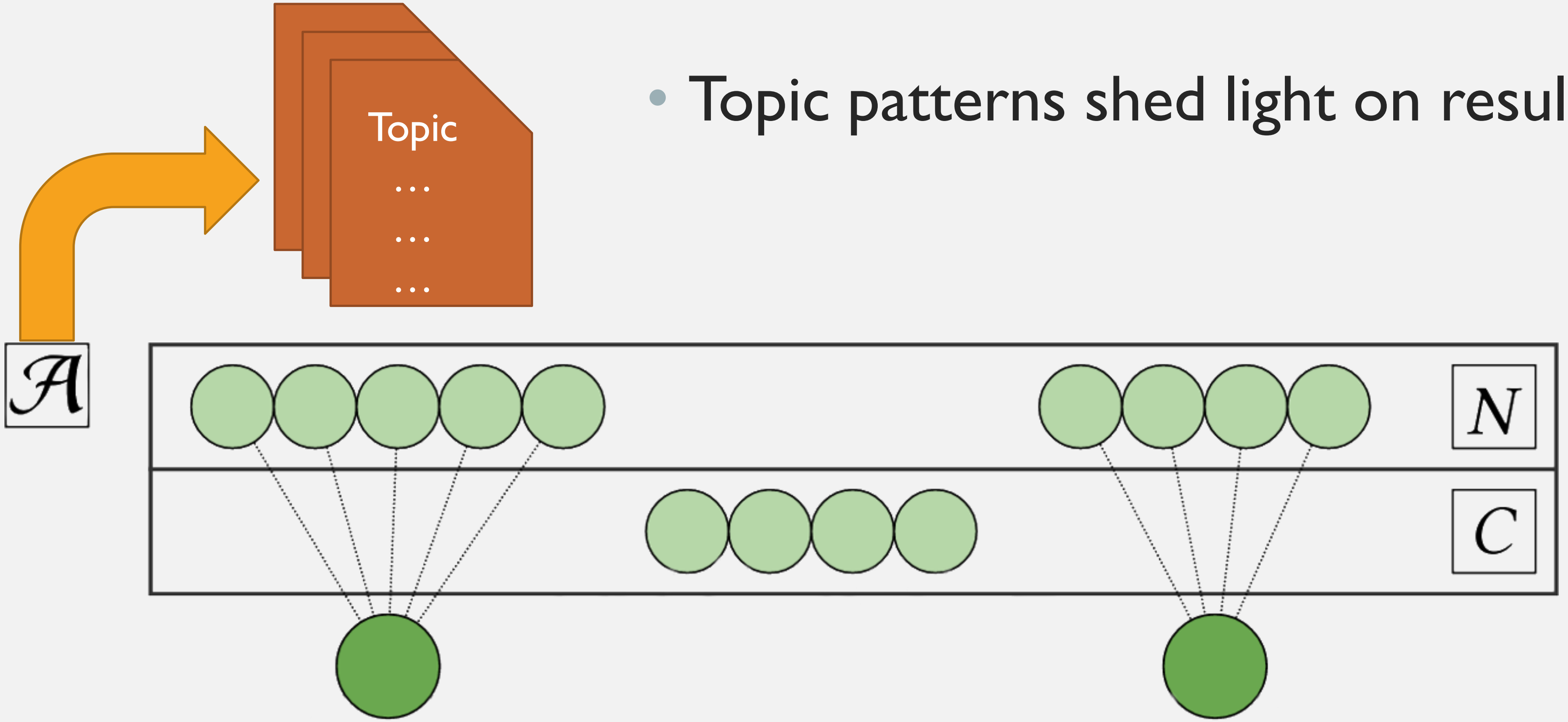


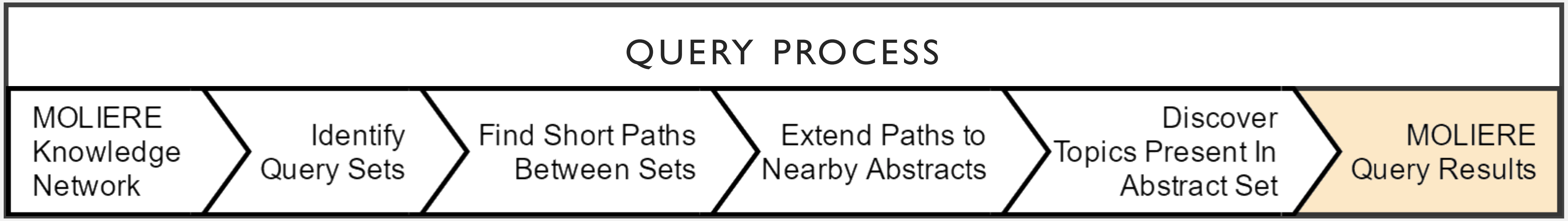
QUERY PROCESS

MOLIERE Knowledge Network → Identify Query Sets → Find Short Paths Between Sets → Extend Paths to Nearby Abstracts → Discover Topics Present In Abstract Set → MOLIERE Query Results

- PLDA+: Identifies topics present in a set of text.

- Topic patterns shed light on results




QUERY PROCESS
MOLIERE Knowledge Network
Identify Query Sets
Find Short Paths Between Sets
Extend Paths to Nearby Abstracts
Discover Topics Present In Abstract Set
MOLIERE Query Results

- Hypothesis represented as a topic model
- Shown: Venlaflaxine vs. HTR1A

TOPIC: 0

antidepressant_drugs
milnacipran
org
selected
ht

TOPIC: 1

increase
reduced
treatment
dorsal_raphe_nucleus
effect

TOPIC: 2

rats
sert
ht_receptor
escitalopram
potency

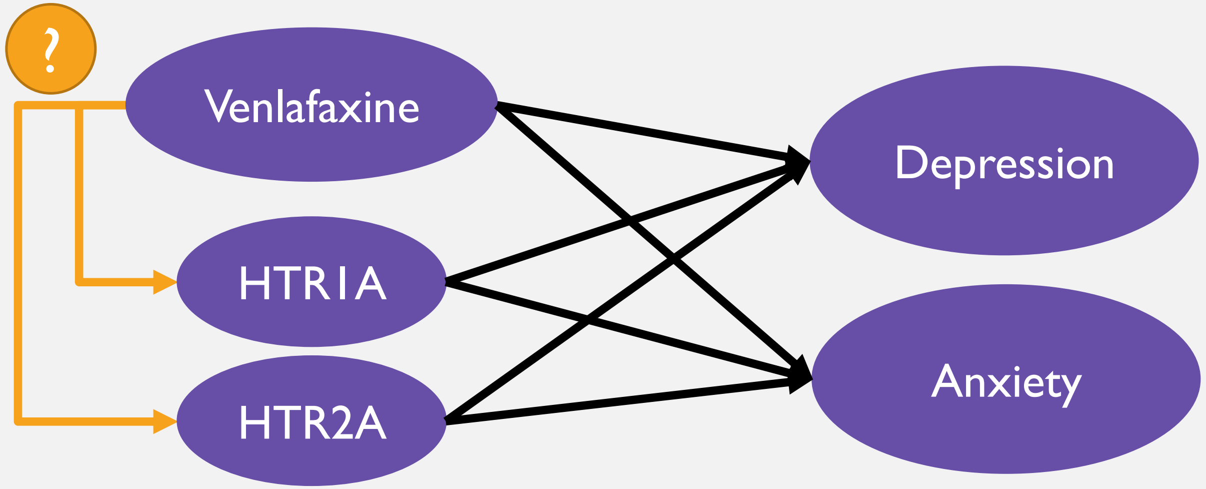
RESULTS

VENLAFAXINE

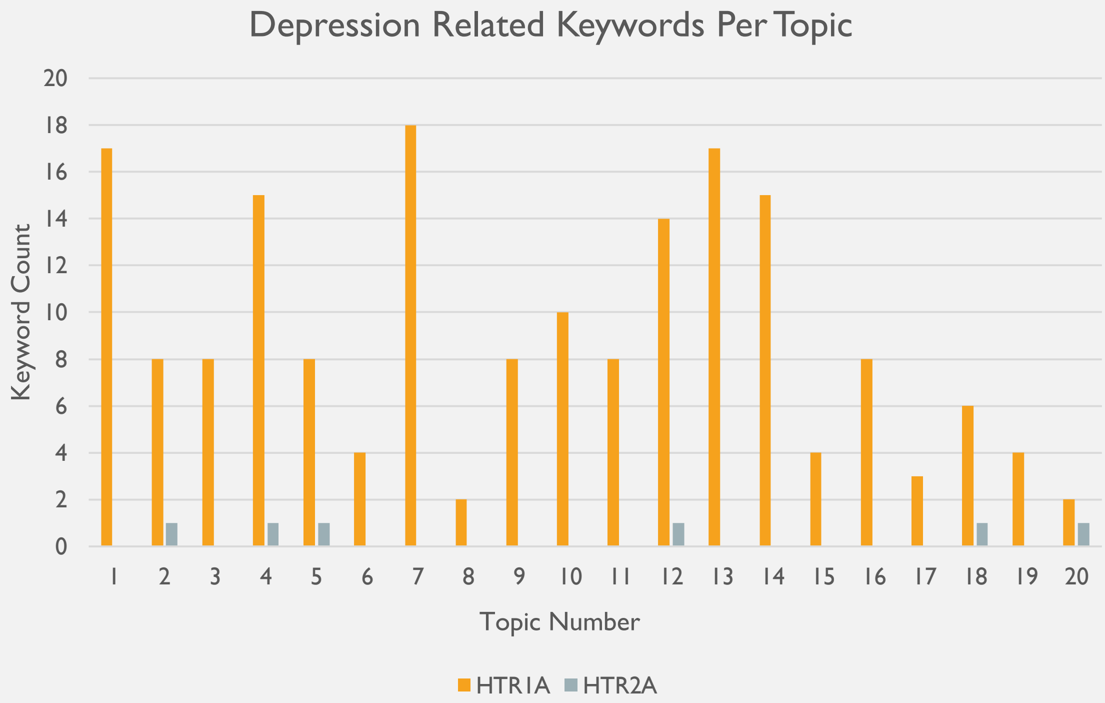
DRUG REPURPOSING

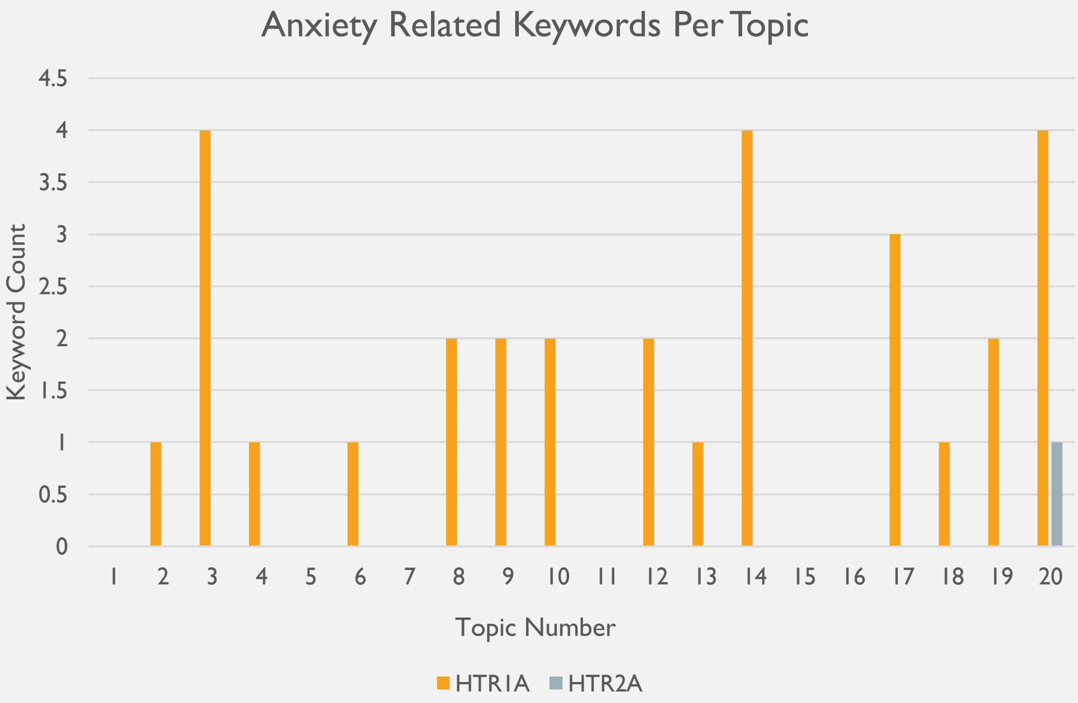
VENLAFAXINE EXAMPLE

- Venlafaxine:
  - Treats depression / anxiety
- HTR[12]A:
  - Linked to depression / anxiety
- No paper linked these concepts



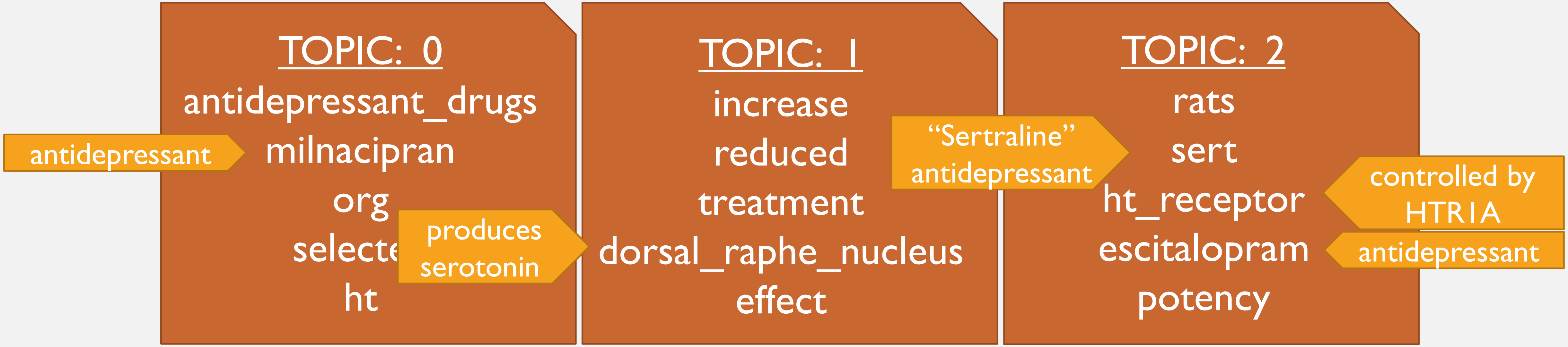
VENLAFAXINE RESULTS





VENLAFAXINE RESULTS

- Hypothesis represented as a topic model
- Shown: Venlaflaxine vs. HTR1A



TOPIC: 0

antidepressant_drugs

milnacipran

org

selecte

ht

TOPIC: 1

increase

reduced

treatment

dorsal_raphe_nucleus

effect

TOPIC: 2

rats

sert

ht_receptor

escitalopram

potency

antidepressant

produces serotonin

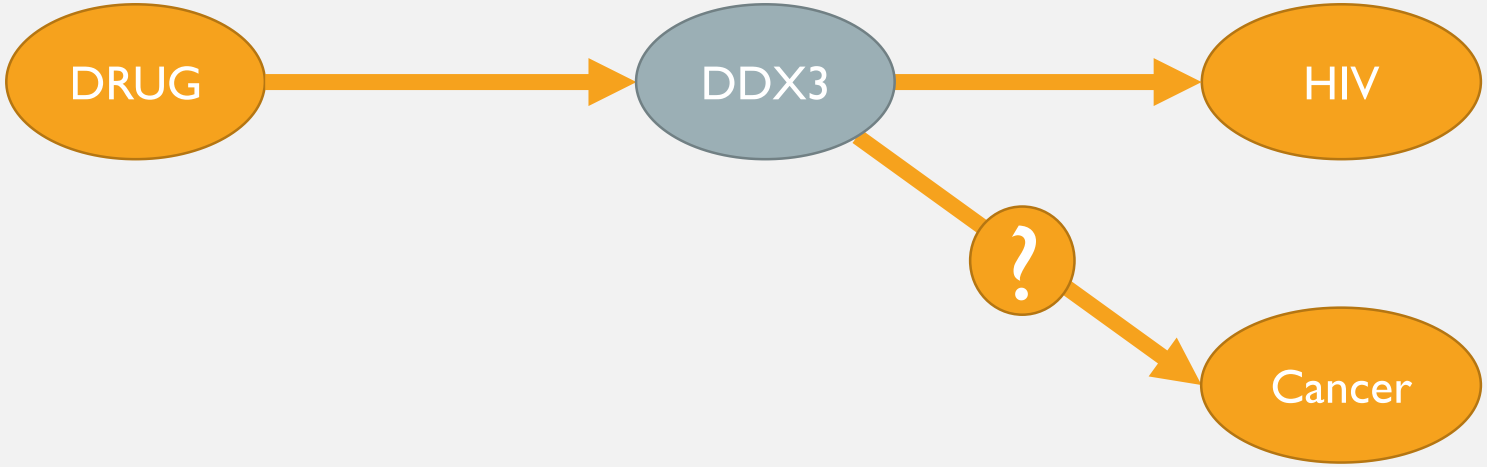
"Sertraline" antidepressant

controlled by HTR1A

antidepressant

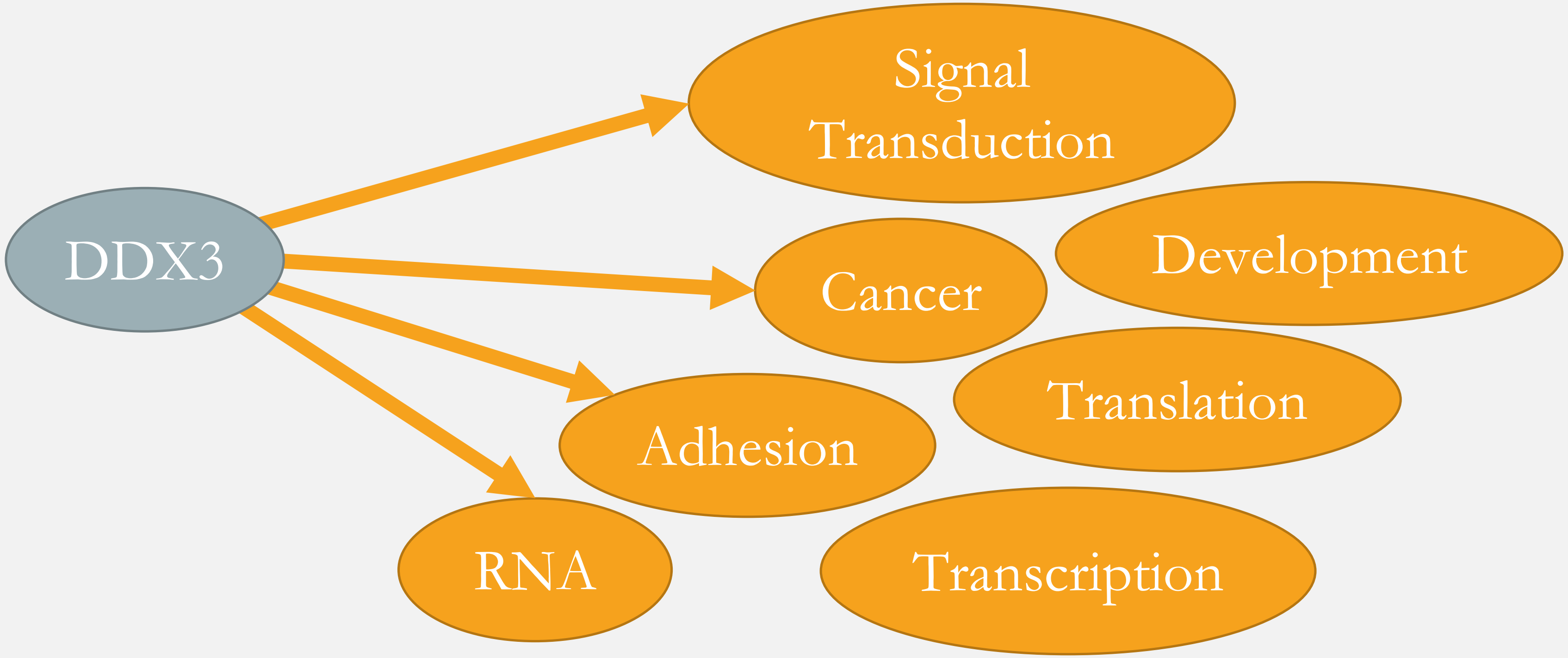
DRUG REPURPOSING EXAMPLE

- Drugs can be modified to treat new diseases
- Decreases drug development time and costs



DRUG REPURPOSING EXPERIMENT

- Ran nearly 10,000 queries involving DDX3:



DRUG REPURPOSING EXPERIMENT

- Ran nearly 10,000 queries involving DDX3:
- Expecting:

Cell – Cell Adhesion

WNT Signaling Pathways

Cell – Matrix Adhesion

DRUG REPURPOSING RESULTS

Cell – Cell Adhesion	cell-cell adhesion regulation of cell-cell adhesion cell-adhesion molecules
WNT Signaling Pathways	signal-transduction associated kinases cell adhesion kinase
Cell – Matrix Adhesion	substrate adhesion RGD cell adhesion domain cell adhesion factor focal adhesion kinase

APPLICATIONS

- Drug Repurposing
- Extensions to new domains
  - Patents, Economics, etc.
- Coping with Deadlines

OPEN RESEARCH QUESTIONS

- Result Interpretation
- System Verification
- Automatic Network Tuning
- Streaming Network Reconstruction
- Inclusion of Additional Data Sources

THANK YOU

J. Sybrandt, M. Shtutman, I. Safro "MOLIERE: Automatic Biomedical Hypothesis Generation System"

Code and Data: **https://people.cs.clemson.edu/~isafro/software.html**

Email: **JSYBRAN@CLEMSON.EDU**