

Mining ETDs for Trends in Graduate Research

CNI Fall 2020 Virtual Membership
Meeting

Bill Ingram, Virginia Tech
November 12, 2020



Opening Books and the National Corpus of Graduate Research

IMLS National Leadership Grants for Libraries

<https://www.ims.gov/grants/awarded/lg-37-19-0078-19>

Investigating innovative ways machine learning and natural language processing can be applied to the national corpus of electronic theses and dissertations in order to extract knowledge, bibliographic and scientific data, and facilitate its identification, discovery, and reuse.

Research Areas:

1. Document analysis, information extraction
2. Adding value through automatic classification and summarization
3. User services – building better digital libraries

ETD Research Team at Virginia Tech and Old Dominion University

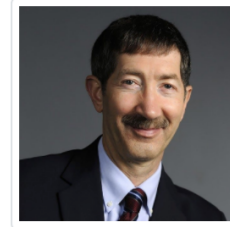
Bill Ingram



Principle Investigator

Assistant Dean and Director of IT,
University Libraries, Virginia Tech

Dr. Edward A. Fox



Co-PI

Professor, Computer Science, Virginia
Tech

Dr. Jian Wu



Co-PI

Professor, Computer Science, Old
Dominion University

Bipasha Banerjee



Graduate Assistant

Ph.D. Candidate, Computer Science,
Virginia Tech

Muntabir Choudhury



Graduate Research Assistant

Ph.D. Student, Computer Science,
Old Dominion University

And several students past and present:

Richard D. Pates, Jr., Adheesh Sunil Juvekar, Eman Abdelrahman,
Fatimah Alotaibi, Himarsha Jayanetti, Palakh Mignonne Jude,
Sampanna Kahu, John Aromando

Last year at CNI Fall 2019



William Ingram

**Virginia Polytechnic Institute and
State University**

Assistant Dean, University Libraries

3.3.1 Bringing Computational Access to Book-length Documents Via an ETD Pilot

Thanks to CNI, my talk led to an introduction and conversation with the chief strategy officer at ProQuest, which led to a collaboration and the opportunity for our team to pilot the new ProQuest TDM Studio.

TDM Studio Expert



John Dillon

Product Manager
TDM Studio, ProQuest

PQDT Expert



Austin Mclean

Sr. Director Academic Relations
Dissertations & Theses, ProQuest

Outline



Overview of the TDM Studio



Research Question



Describing the Data



Methodology



Results



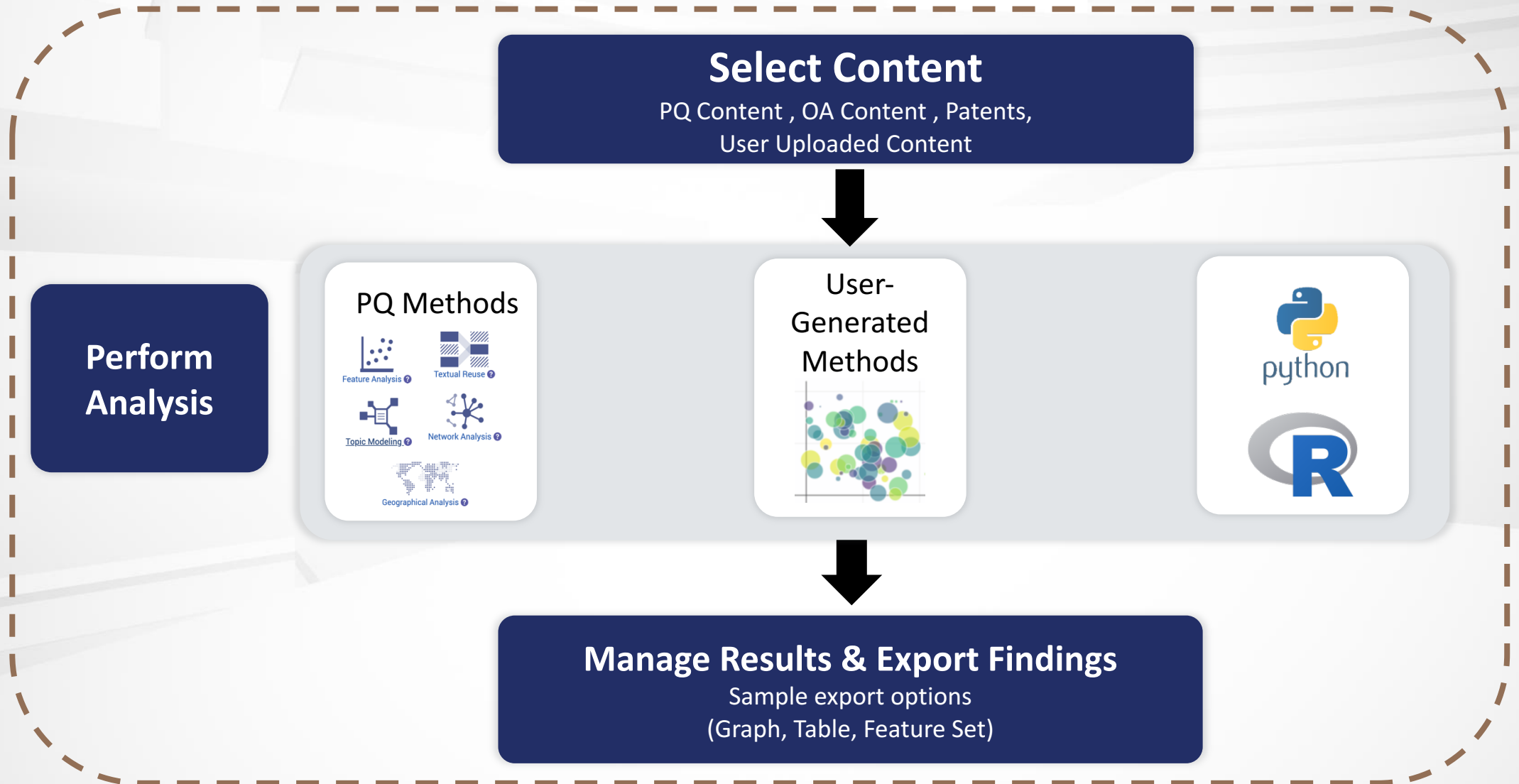
Discussion

Disclaimer

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of any organization.

TDM Studio and the Research Workflow

TDM Studio streamlines the process—making research more efficient & productive





BI

Bill Ingram

waingram@vt.edu

virginiatech0 Virginia Polytechnic Institute and State University

Open Jupyter Notebook

Running

Toggle to restart your environment

Datasets (Using 9 of 10)

Create datasets of up to 2,000,000 documents each

+ Create New Dataset

Select Publication Titles

Select ProQuest Databases

NAME	DESCRIPTION	QUERY	LOCATION	STATUS	D
2019and20		Show ▾	tdm-ale- data/133/corus/2019and20/	Completed	

- Web interface for selecting data
- Access to material from your library's subscriptions
- Content rights cleared for TDM
- Build datasets up to 2,000,000 documents

133 | Create New Dataset

● Find Content

○ Refine Content

○ Create Dataset

< Choose Databases (2 of 118)

 × 🔍

- Dissertations & Theses @ Virginia Polytechnic Institute and State University** This database gives access to the dissertations and theses produced by students at your institution. 📖 Full text
- ProQuest Dissertations & Theses Global** ProQuest Dissertations & Theses (PQDT) Global is the world's most comprehensive collection of dissertations and theses from around the world, offering millions of works from thousands of universities...

Items per page: 20 ▾ 1 – 2 of 2 |< < > >|

1 Database selected

Next: Refine Content

Enter search here...



Selected Databases

[Change](#)



ProQuest Dissertations & Theses Global

[X Clear All Filters](#)

Limit to

Full text

Date Published

1/1/2000  to 12/31/2019 

Source Type

Dissertations & Theses (5097347)

Document Type

5,097,347 documents

Discovery of the Hemifusion Trigger in Lassa Virus Entry

ProQuest Dissertations and Theses, 2022

Mechanisms of CNS Viral Reservoir Formation and Reseeding by HIV+ Mature Monocytes: Potential Therapies for NeuroAids in the Art-Era

ProQuest Dissertations and Theses, 2022

Mind the Gap: A Crosswalk Analysis of California Teacher Preparation Standards and Public K-12 Local Teacher Evaluations

ProQuest Dissertations and Theses, 2022

Plasma Cell Activation in Systemic Lupus Erythematosus

ProQuest Dissertations and Theses, 2022

Examining the Relationship Between Perceived Parenting Style and Attachment in Deaf Adults (A Replication Study)

ProQuest Dissertations and Theses, 2022

The Role of Emotional Intelligence in Shaping Experiences of High School Leadership

ProQuest Dissertations and Theses, 2022

Consequences of Post-deployment Health Assessment Exclusion in U.S. Navy Personnel

ProQuest Dissertations and Theses, 2022

An Examination of Three Emerging Interventions and Proposed Best Practices to Treat Latina Women Who Have Experienced Intimate Partner

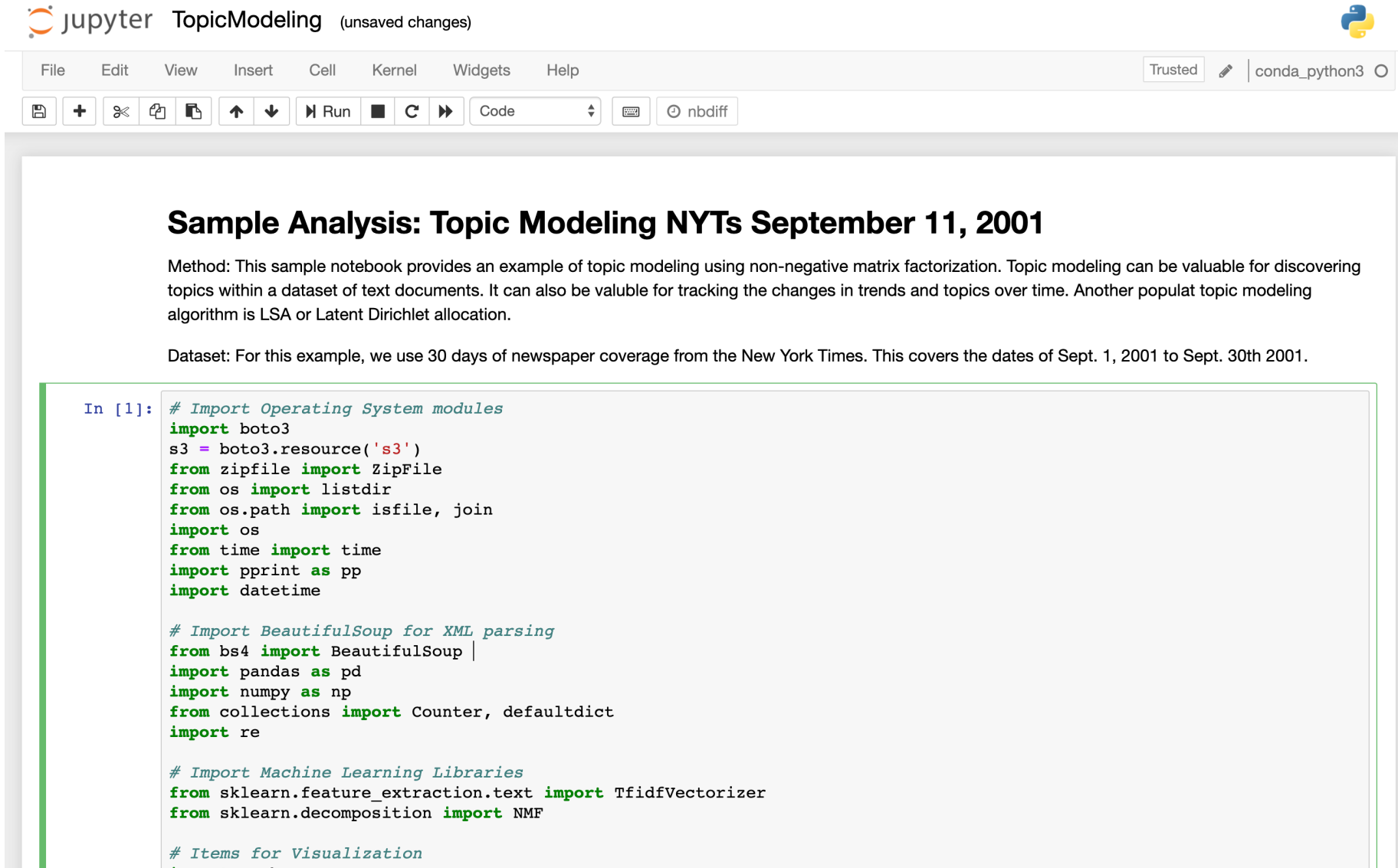
5,097,347 Documents




Please refine your results below 2,000,000 documents to proceed.

[Next: Review Dataset](#)

TDM Studio Workbench – Jupyter notebooks for R and Python



jupyter TopicModeling (unsaved changes) 

File Edit View Insert Cell Kernel Widgets Help Trusted | conda_python3

Code nbdiff

Sample Analysis: Topic Modeling NYTs September 11, 2001

Method: This sample notebook provides an example of topic modeling using non-negative matrix factorization. Topic modeling can be valuable for discovering topics within a dataset of text documents. It can also be valuable for tracking the changes in trends and topics over time. Another popular topic modeling algorithm is LSA or Latent Dirichlet allocation.

Dataset: For this example, we use 30 days of newspaper coverage from the New York Times. This covers the dates of Sept. 1, 2001 to Sept. 30th 2001.

```
In [1]: # Import Operating System modules
import boto3
s3 = boto3.resource('s3')
from zipfile import ZipFile
from os import listdir
from os.path import isfile, join
import os
from time import time
import pprint as pp
import datetime

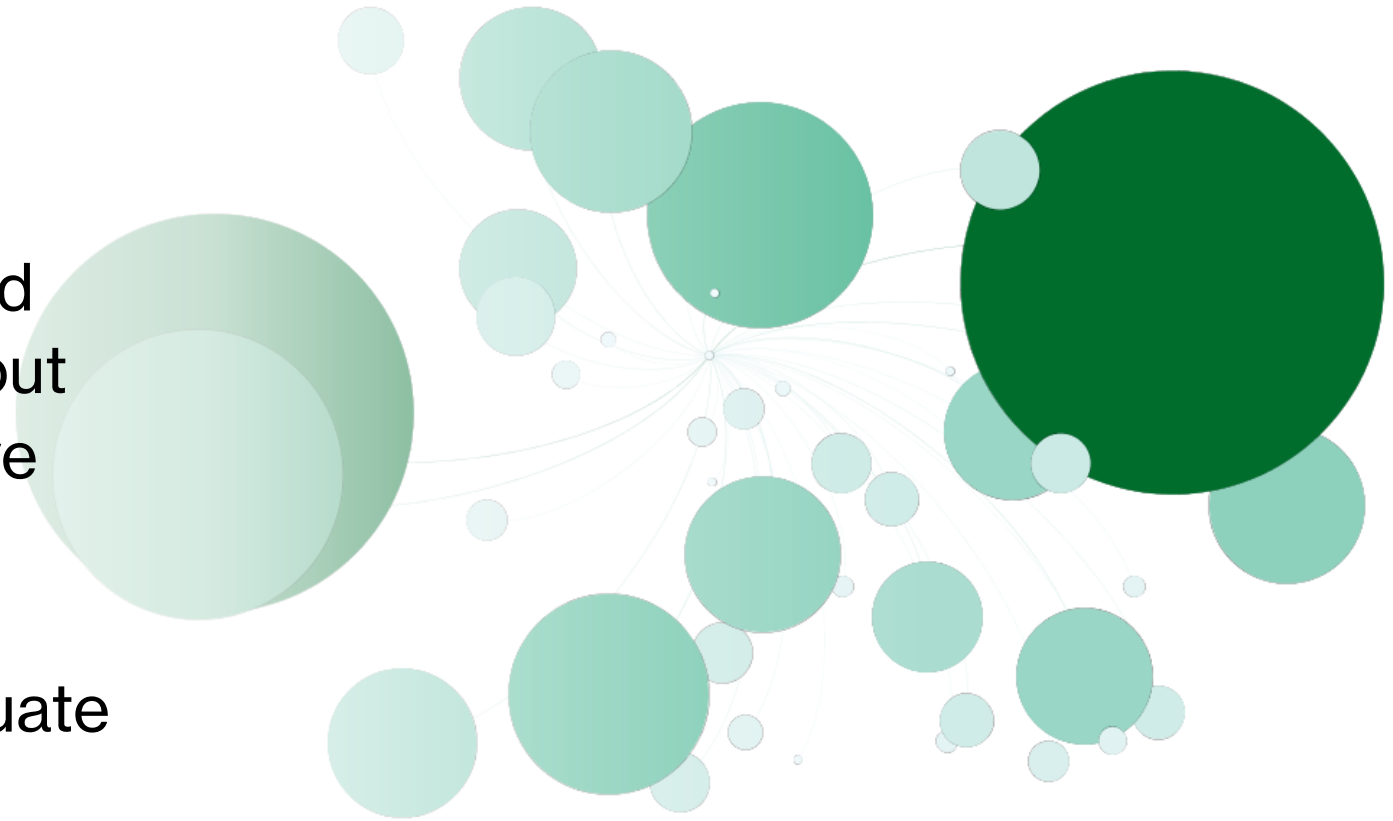
# Import BeautifulSoup for XML parsing
from bs4 import BeautifulSoup |
import pandas as pd
import numpy as np
from collections import Counter, defaultdict
import re

# Import Machine Learning Libraries
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.decomposition import NMF

# Items for Visualization
```

Research Question

What can we learn through text and data mining of the ETD corpus about how graduate **research topics** have **evolved**, how different topics and disciplines overlap, and how has **interdisciplinarity** evolved in graduate research?



Data and Feature Set

- Roughly 1.3 million ETDs from 2000-2018
- Full-text XML files with metadata
- Only about 600,000 with department metadata
- Extract features
 - Title
 - Abstract
 - Department
 - Year of publication
- Organize them in batches by years and majors (departments)
- Intuition: top terms in title and abstract will indicate research topic

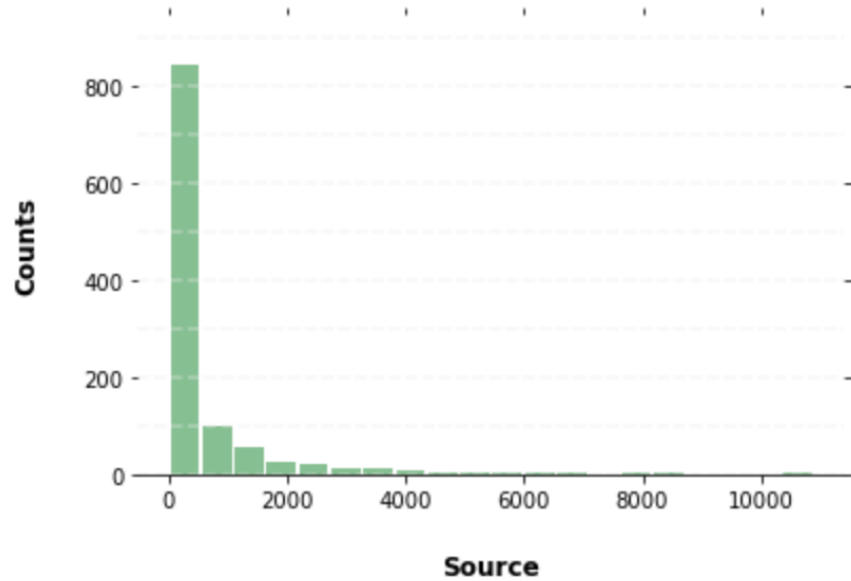
Psychology	22262
Chemistry	17645
Education	16625
Mechanical Engineering	11642
Computer Science	11012
English	10219
Physics	9580
Electrical Engineering	9196
School of Education	7756
Electrical and Computer Engineering	7738
History	7575
Mathematics	7357
Biology	7285
Economics	7175
Department not provided	6457
Civil Engineering	5664
Nursing	5567
Sociology	5527
Educational Leadership	5183
Anthropology	5173
Music	4732

Distribution of top 20 depts

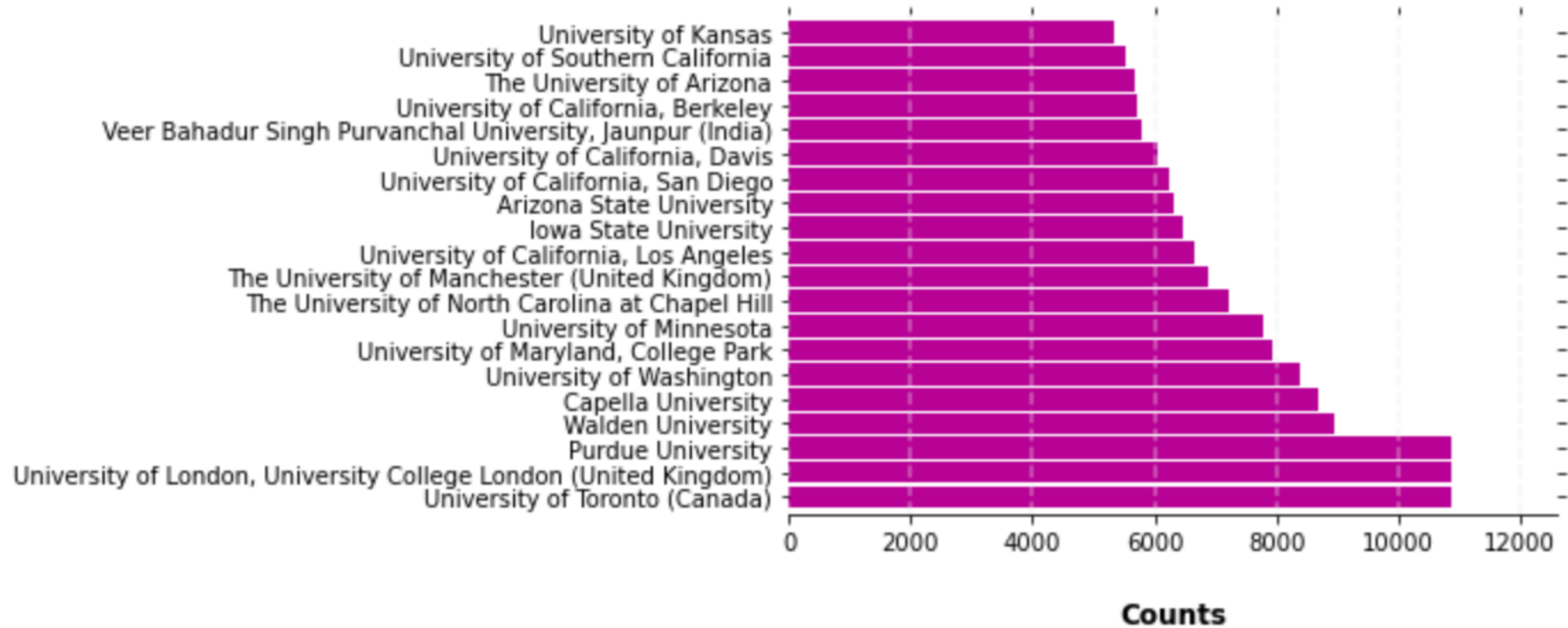
Data Sources – Top 20 Institutions

	School Name	counts			
0	University of Toronto (Canada)	10886	10	University of California, Los Angeles	6669
1	University of London, University College Londo...	10870	11	Iowa State University	6477
2	Purdue University	10855	12	Arizona State University	6322
3	Walden University	8954	13	University of California, San Diego	6227
4	Capella University	8702	14	University of California, Davis	6048
5	University of Washington	8391	15	Veer Bahadur Singh Purvanchal University, Jaun...	5777
6	University of Maryland, College Park	7948	16	University of California, Berkeley	5708
7	University of Minnesota	7767	17	The University of Arizona	5684
8	The University of North Carolina at Chapel Hill	7205	18	University of Southern California	5532
9	The University of Manchester (United Kingdom)	6880	19	University of Kansas	5321

Distribution of Data Sources



Source (top 20 only)



Methodology

- Determine research focus using TF-IDF (term frequency–inverse document frequency) to calculate the most important two- and three-word phrases in the corpus.
- Initially this looked promising, but the method yielded too many irrelevant phrases.

user interface
digital libraries
web services
digital library
requirements generation
web service
data assimilation
software development
software engineering
usability engineering
information visualization
requirements engineering
usability evaluation
data mining
virtual environments

Top phrases from computer science and biology

cell cycle
organic matter
eps production
wild type
gene expression
land use
leaf breakdown
corticosterone levels
quorum sensing
headwater streams
xenopus laevis
cell wall
mine drainage
acid mine drainage
acid mine

results show
gene expression
large scale
recent years
results indicate
also provides
goal research
response time
future work
commonly used
novel approach
experimental data
life cycle
high level
multiple sources

Common irrelevant phrases

Methodology

- **Wikifier** [1] is a named entity recognition tool to disambiguate terms using Wikipedia.

Essays on Investment and Exports of Multinational Firms in South Korea This dissertation consists of three essays on investment and exports of multinational firms in South Korea. The first chapter examines firm-level evidence that banking crises in source countries affect investment decisions of foreign multinationals. Using firm-level data on annual financial statements with information on banking crises and foreign-owned companies in South Korea from 1994 to 2013, I find that an increase in foreign shareholding decreases the investment rate of foreign multinationals during banking crises in source countries. Firm characteristics and financial vulnerabilities alter the impact of the banking crisis on investment decision of foreign multinationals.



Essays on Investment and Exports of [Multinational Firms in South Korea](#) This **dissertation** consists of three essays on investment and exports of **multinational firms in South Korea** . The first chapter examines firm-level evidence that **banking** crises in source countries affect investment decisions of foreign **multinationals**. Using firm-level data on annual **financial statements** with information on **banking** crises and foreign-owned companies in [South Korea](#) from **1994** to **2013**, I find that an increase in foreign shareholding decreases the investment rate of foreign **multinationals** during **banking** crises in source countries. Firm characteristics and financial **vulnerabilities** alter the impact of the **banking crisis** on **investment decision** of foreign **multinationals**. In the case of non-chaebol, non-exporter firms, or less financially sensitive industries, there exists

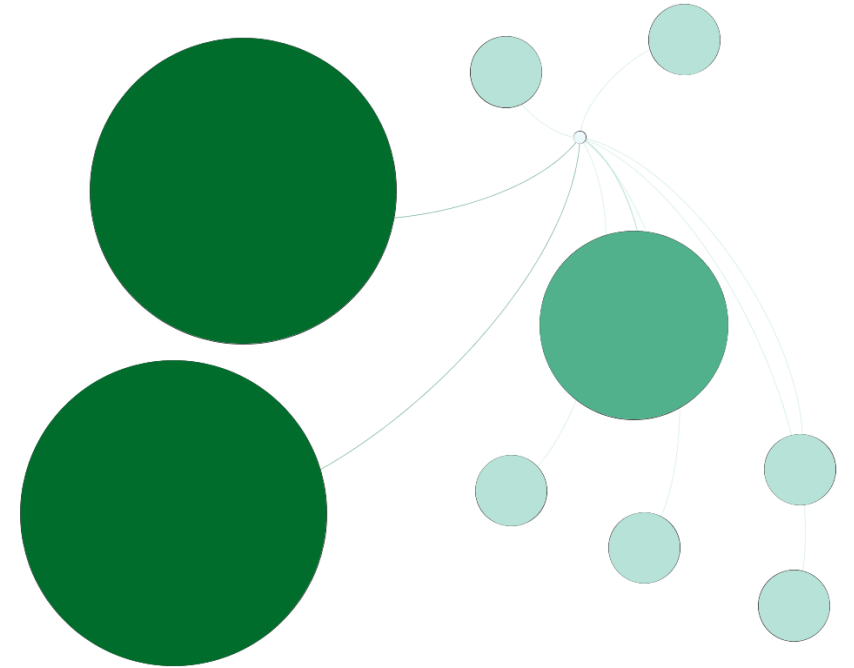
[1] Lev Ratinov and Dan Roth and Doug Downey and Mike Anderson, [Local and Global Algorithms for Disambiguation to Wikipedia](#) ACL (2011)

Methodology

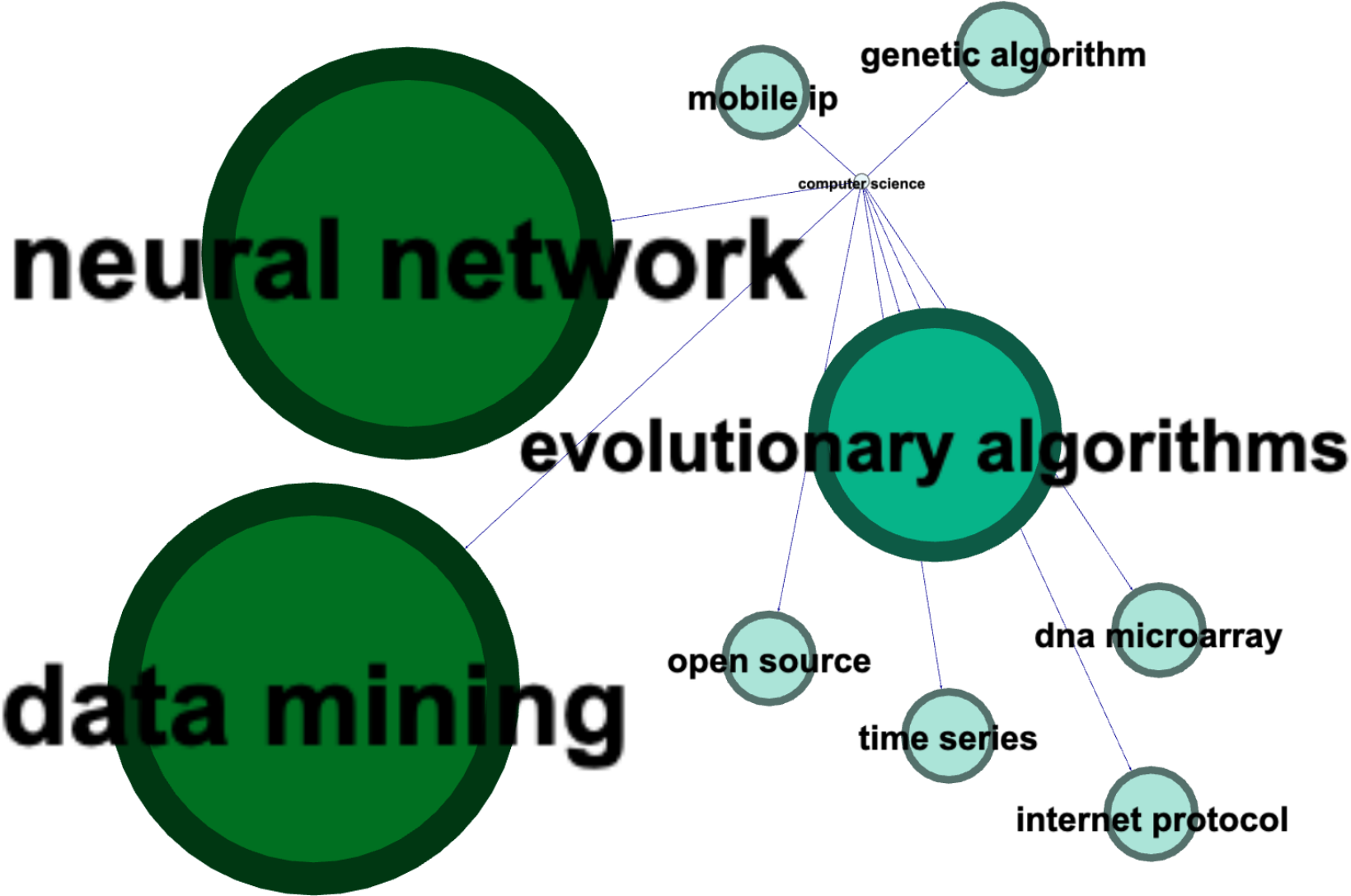
Topic Identification	For every document in each department batch, use Wikifier to identify terms from Wikipedia – these terms will represent research topics for that department or major.
Calculate Document Frequency	Calculate the document frequency of each term for a time interval [2001-2005, 2006-2009, 2010-2013, 2014-2018].
Plot Results	Plot the terms with highest document frequency for a department and time interval.
Compare Plots	Compare plots of other time intervals for the same department to see the evolution of research topics within that department or major.
Repeat	Repeat these steps for other departments.
Plot Multiple Departments	See how research topics are shared across departments and how this interdisciplinarity evolves over time.

Results

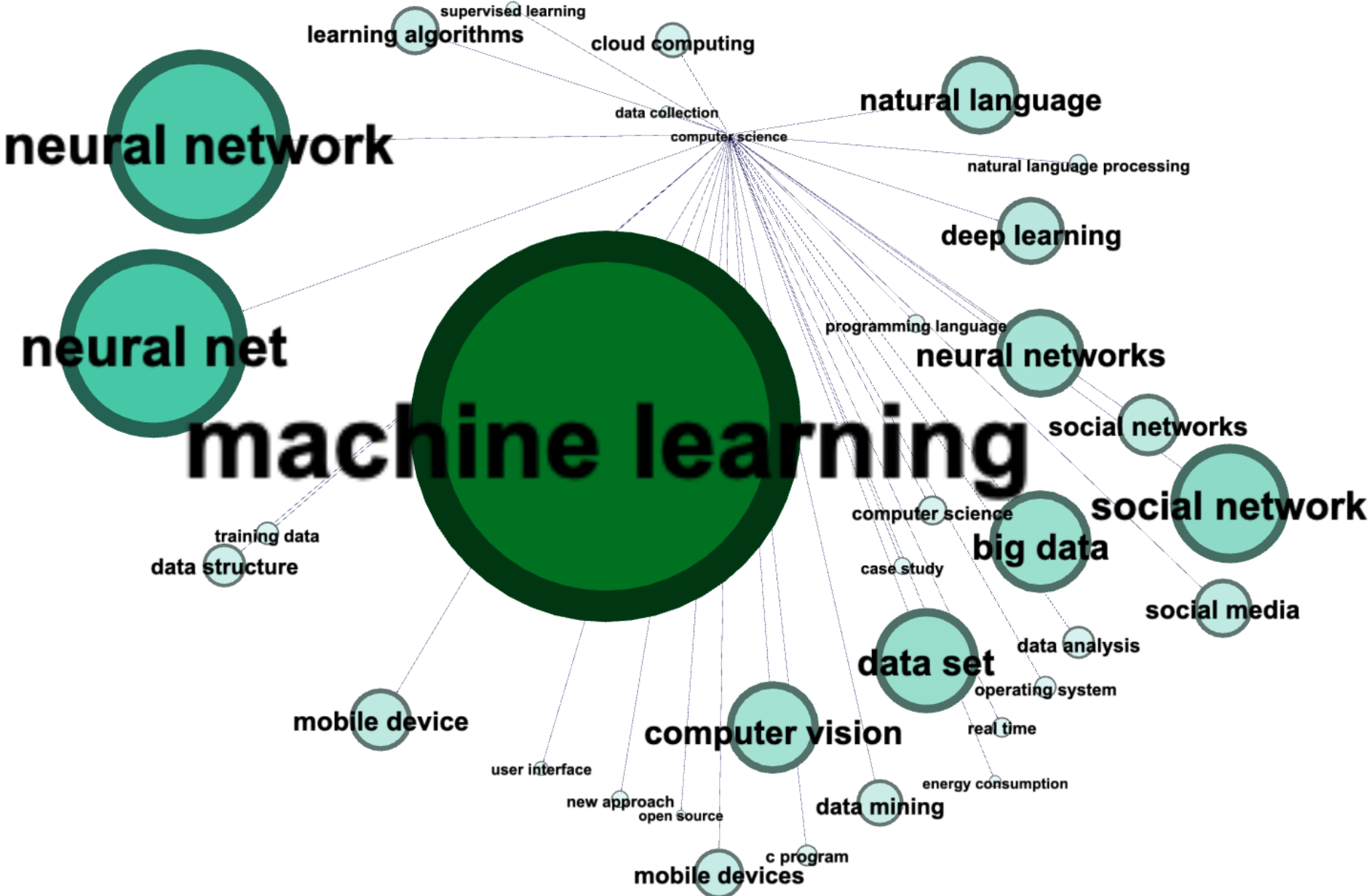
The following graphs show the top terms from different disciplines, mainly **computer science** and **biology**, for different years. We also show the intersection of terms from both disciplines.



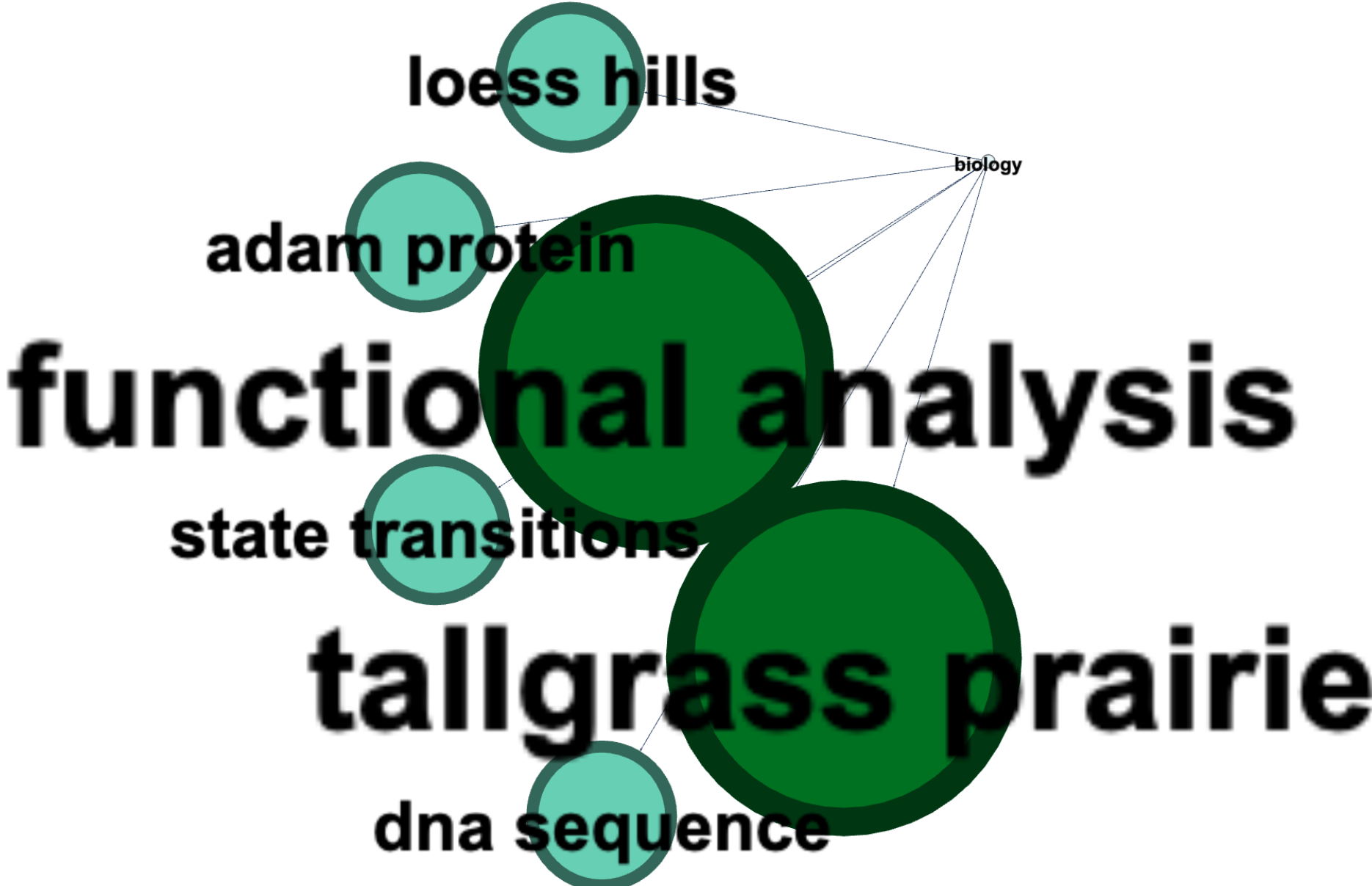
Computer Science 2001-2005



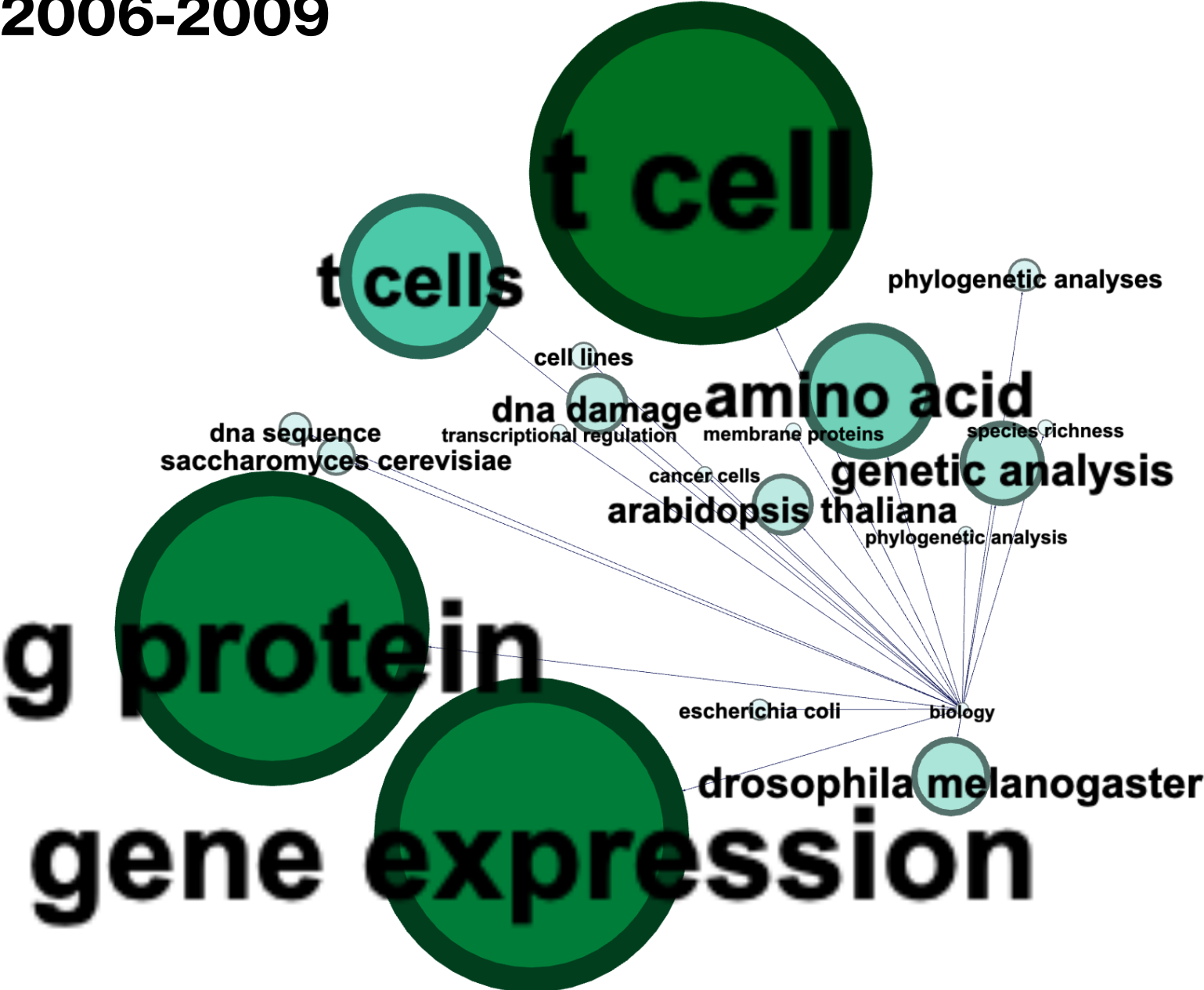
Computer Science 2014-2018



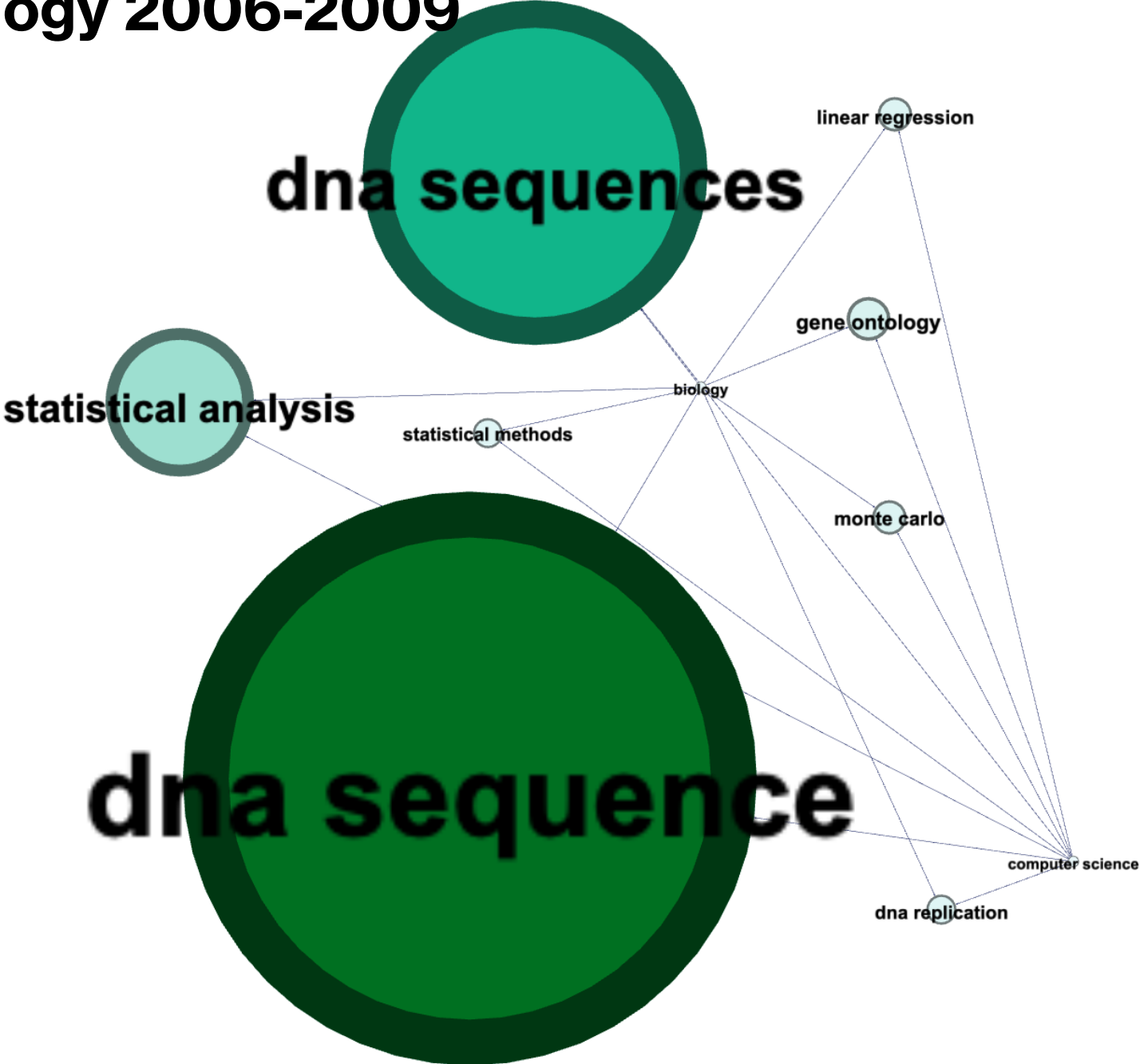
Biology 2001-2005



Biology 2006-2009



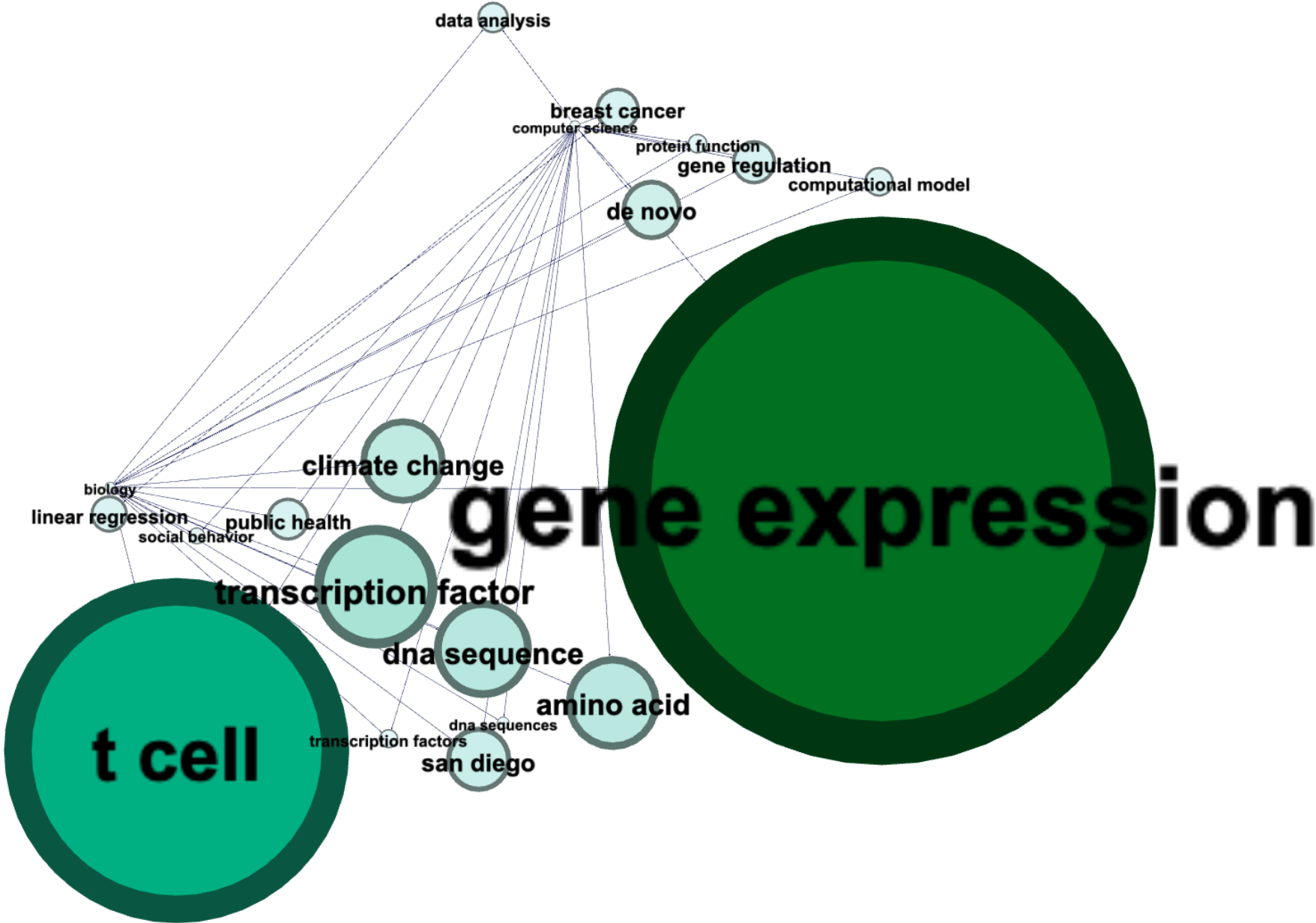
CS and Biology 2006-2009



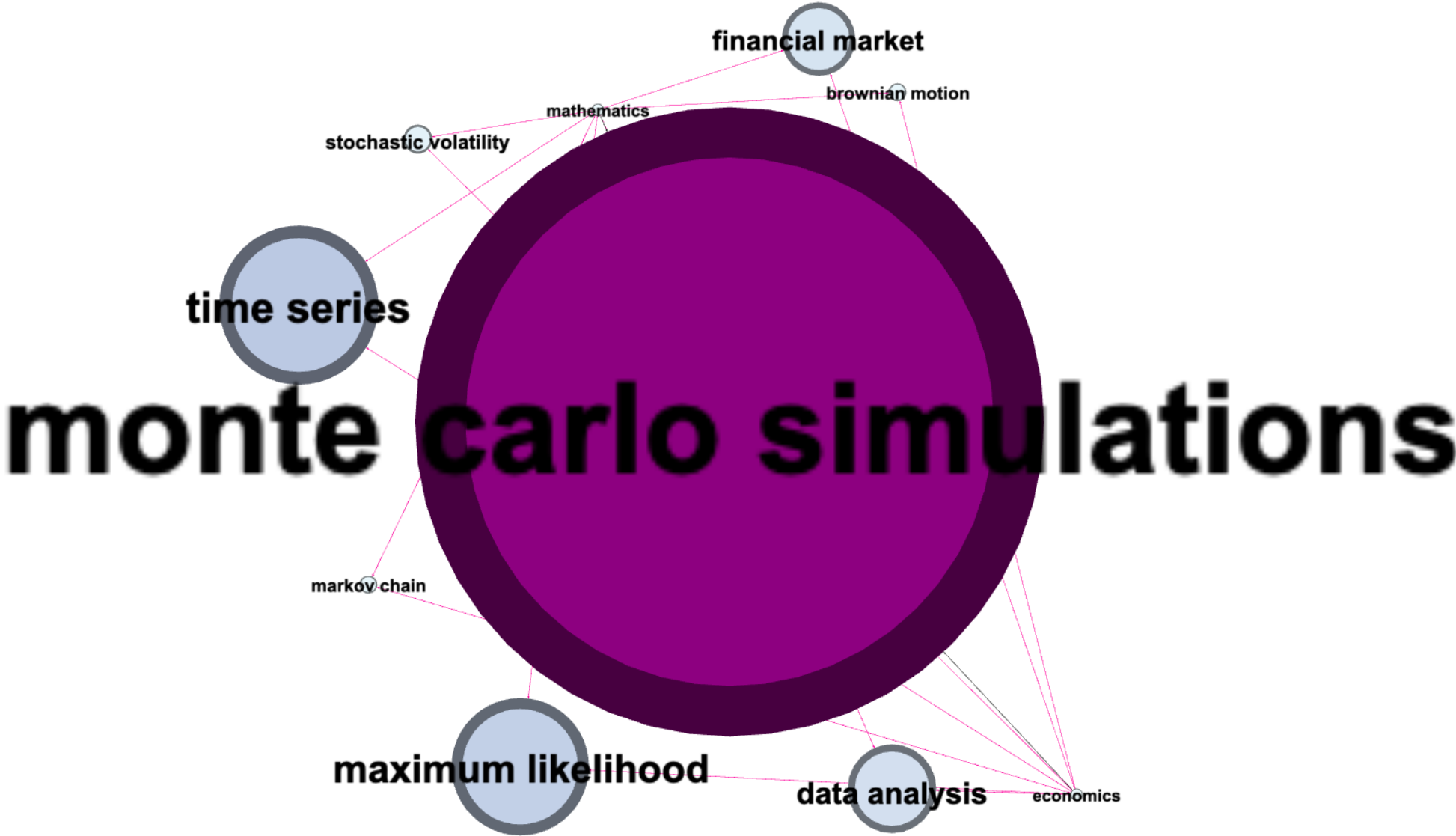
CS and Biology 2010-2013



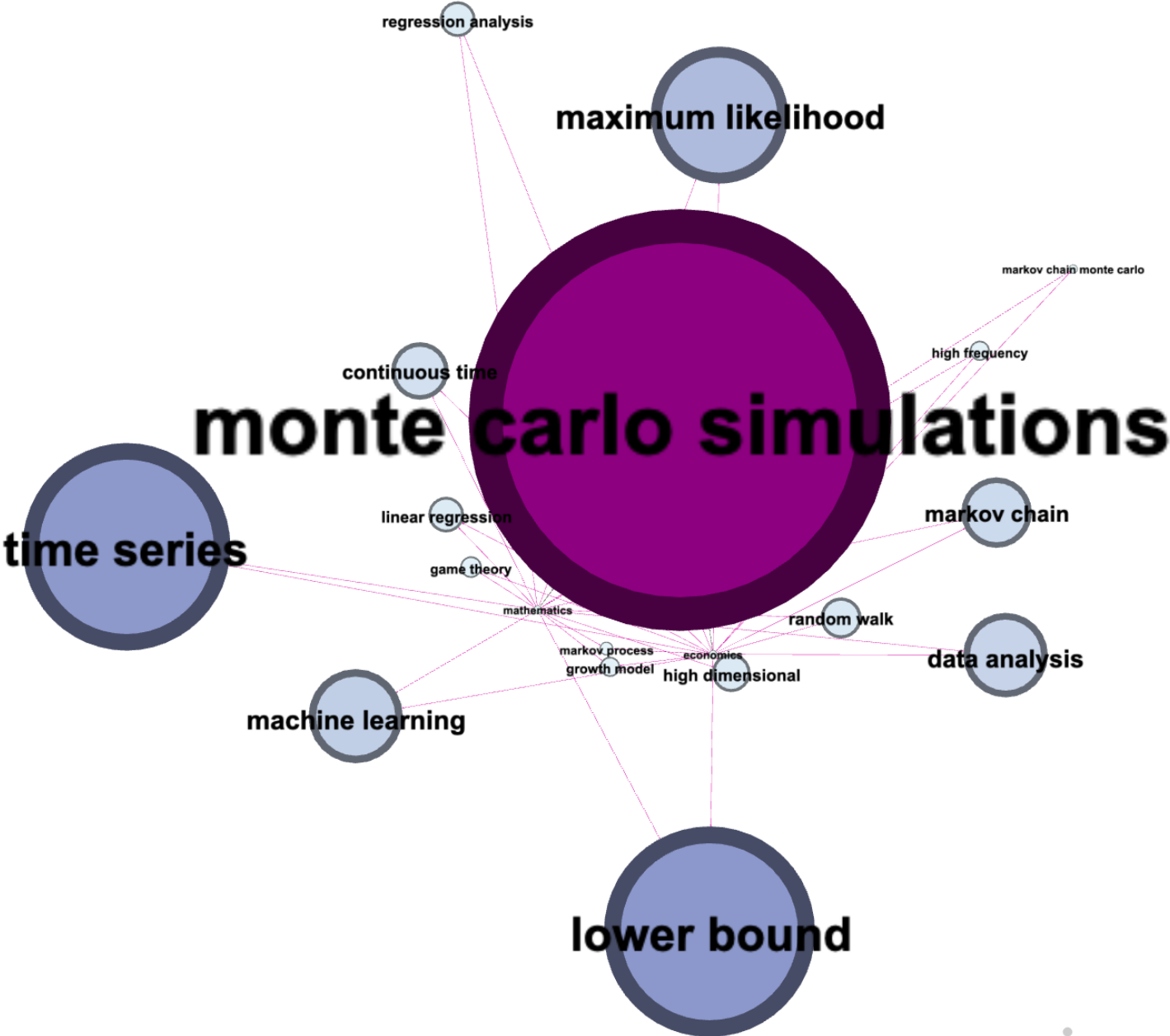
CS and Biology 2014-2018



Economics and Mathematics 2010-2013



Economics and Mathematics 2014-2018



Discussion

Back to our research question:

What can we learn through text and data mining of the ETD corpus about how graduate research topics have evolved, how different topics and disciplines overlap, and how has interdisciplinarity evolved in graduate research?

We've shown that it is possible to determine the research focus of ETDs using the Illinois Wikifier method for concept disambiguation.

Graphing the document frequency of these research topics allows us to visualize the relative importance of these topics within and across academic disciplines and their evolution over time.



QUESTIONS?



Support was made in part by the Institute of Museum and Library Services for grant [LG-37-19-0078-198](#).



This project was supported in part by ProQuest, which provided access to TDM Studio. The university subscribes to the dataset, ProQuest Dissertations & Theses (PQDT).

Thank you!

<https://opening-etds.github.io>



For more information about ProQuest's TDM Studio or ProQuest Dissertations & Theses Global contact:

John.Dillon@proquest.com or Austin.Mclean@proquest.com



Where can I learn more about TDM Studio?

Visit: <https://about.proquest.com/products-services/TDM-Studio.html>



Where can I learn more about PQDT Global?

Visit: dissertations.com



How can I trial TDM Studio?

Email: TDMStudio@proquest.com



How can I subscribe to PQDT Global?

Inquire at:
<https://about.proquest.com/contact/contact-sales/>