Cognitive Big Data Analysis: From the Literature to Public Policy

Ir.Rezzy Eko Caraka.,M.Sc(RES).,PhD

Researcher - BRIN Indonesia
Research Professor - CYUT Taiwan

Kenali target publisher anda

Dovepress

ELSEVIER

Springer

IEEE
Advancing Technology for Humanity

Taylor & Francis Group
an informa business

emerald PUBLISHING

BMC
Part of Springer Nature

cogent oa

WILEY
Publishers Since 1807

SAGE Publishing

ACS Publications
Most Trusted. Most Cited. Most Read.

nature publishing group

Aspek	**EMERALD**	**SPRINGER**	**DOVE Medical Press**	**COGENT (Taylor & Francis OA)**	**IEEE**	**Taylor & Francis (TNF)**	**BMC (BioMed Central)**
Fokus Utama	Manajemen, sosial, pendidikan, kebijakan	STEM, sains murni, teknik	Kedokteran, kesehatan, farmasi	Sosial, humaniora, multidisiplin	Teknik, komputasi, sinyal	Sosial & sains humaniora	Kedokteran, bioteknologi,
Nada Penulisan (Tone)	Semi-formal, aplikatif, komunikatif	Formal, objektif, metodologis	Klinis, teknikal, berbasis evidensi	Naratif, reflektif, diskursif	Sangat teknikal, netral, presisi tinggi	Teoretis-analitis, argumentatif	Objektif, empiris, berbasis data
Kalimat (Sentence Style)	Pendek (18–22 kata), aktif	Panjang (25–30 kata), pasif	Pendek-menengah, klinis	Menengah (20–24 kata), aktif	Pendek, padat, formulaic	Panjang (24–28 kata), naratif	Menengah (22–26 kata), formal
Readability (ARI)	10–12 (mudah dibaca)	14–16 (kompleks)	11–13 (moderate)	10–12 (mudah dibaca)	13–15 (teknikal)	12–14 (moderat-kompleks)	13–14 (moderat)
Dominant Verbs	*explores, provides, suggests, offers*	*demonstrates, proposes, reveals,*	*assesses, evaluates, observes, reports*	*examines, discusses, interprets, argues*	*proposes, implements, achieves, designs*	*examines, argues, investigates, theorizes*	*analyzes, assesses, reports, estimates*
Structure (IMRAD)	Fleksibel + Practical Implication	Ketat, ilmiah	IMRAD ketat + Ethics section	Fleksibel, sering tambah *Theoretical*	Sangat ketat + Equation-heavy	Standar + *Literature Contextualization*	IMRAD ketat + Ethics/Data Availability
Kutipan (Citation Style)	Harvard (Author, Year)	Numeric atau APA	Vancouver (numeric)	Harvard / APA	Numeric	APA / Harvard	Vancouver
Lexical Density (kompleksitas	Sedang (0.40–0.45)	Tinggi (0.55–0.60)	Menengah	Menengah	Tinggi (0.58)	Menengah-tinggi (0.50)	Menengah (0.48)
Dominant Noun Types	practice, innovation, management	model, data, algorithm, system	patient, trial, therapy	culture, identity, society	circuit, signal, algorithm	policy, framework, justice	genome, cell, outcome
Voice (Aktif/Pasif)	Aktif dominan (80%)	Pasif dominan (70%)	Campuran	Aktif (60%)	Pasif (75%)	Aktif (55%)	Campuran
Visuals & Tables	Moderate	Banyak (grafik, tabel)	Banyak (figure, table, flowchart)	Sedikit–moderate	Banyak sekali (diagram, formula)	Moderate	Banyak (grafik medis, table klinis)
Abstrak	Informative + applied	Informative + technical	Structured (Background–	Narrative informative	Structured concise	Thematic / conceptual	Structured medical
Kata transisi dominan	*furthermore, moreover,*	*hence, therefore, in contrast*	*in addition, in summary, as a result*	*similarly, however, in this context*	*thus, accordingly, overall*	*moreover, henceforth, however*	*specifically, therefore, notably*
Target Reader	Akademisi & praktisi	Peneliti sains & teknik	Dokter, klinisi	Akademisi sosial, multidisiplin	Engineer, ilmuwan komputer	Akademisi & pembuat kebijakan	Peneliti medis & epidemiolog
Ciri Khas Unik	Fokus pada *managerial implication*dan *applied*	Teknis, berbasis algoritma atau teori	Fokus *clinical outcome* dan *treatment efficacy*	*Inclusive tone*, terbuka untuk perspektif baru	Sangat formal dan matematis	Berbasis argumen & teori sosial	Berbasis *evidence* dan *open data*
Konklusi (Ending Style)	Implikasi praktis	Sintesis metodologis	Clinical impact	Implikasi sosial	Validation / future work	Theoretical implication	Clinical/public health recommendation

Why Does the Human Brain Need Data Cognition?

More than **5 million scientific papers** are published globally every year. The human brain, like a small bowl beneath a data waterfall, cannot process this volume unaided. **Cognitive Big Data analysis** acts as the intelligent filter: it does not merely count words, but extracts meaning, sentiment, and inter-concept relationships from the flood of literature.

The Problem: Information Overload

Researchers face an exponentially growing corpus. Reading even a fraction of relevant literature manually is no longer feasible for any single scholar or team.

The Solution: Cognitive Filtering

Computational methods — bibliometrics, NLP, topic modelling — serve as the intelligent sieve, surfacing patterns, clusters, and gaps that human reading alone would miss.

Understanding "Cognitive Analysis" in Text

Ordinary Data Mining

Counts word frequencies and co-occurrences. Tells you *how often* a term appears, but not *what it means* in context. Fast, but shallow.

Cognitive Analysis

Understands meaning, sentiment, and relational structure between concepts. Produces a *cognitive map* of a field rather than a word-count table.

Why It Matters

Policy-makers and editors need synthesis, not statistics. Cognitive analysis bridges raw data and actionable insight — the difference between a spreadsheet and a strategy.

Taxonomy of Academic Big Data

Academic Big Data is not a monolith. It comprises four distinct layers, each contributing a different dimension of knowledge to the cognitive analysis pipeline.

Bibliographic Metadata

Author names, affiliations, publication year, journal title, volume, and DOI. The skeleton of every record — essential for network construction and trend analysis.

Abstract Text

The richest source of semantic content. Abstracts encode problem statements, methodologies, and conclusions in condensed, structured prose.

Structured Citations

Reference lists reveal intellectual lineage. Co-citation and bibliographic coupling analyses depend entirely on the accuracy of this layer.

Author Networks

Co-authorship data maps collaboration patterns across institutions and countries, revealing research consortia and knowledge brokers.

The Role of the Elsevier Editor: What Do Reputable Journals Seek?

Beyond Beautiful Graphics

A visually stunning network map will not pass peer review if it lacks interpretive depth. Editors evaluate three core qualities above all else.

- **Novelty**: Does the paper advance knowledge, or merely confirm what is already known?
- **Critical Synthesis**: Are findings integrated into a coherent argument, or listed as disconnected observations?
- **Real-World Implications**: Does the work offer actionable insight for society, policy, or practice?

The Editor's Mental Checklist

When a manuscript lands on an editor's desk, the first questions are: "What is genuinely new here?" and "So what?" A paper that cannot answer both within its abstract is unlikely to proceed to review. Cognitive Big Data analysis, when executed rigorously, provides the evidence base to answer both questions with precision and confidence.

Ethics and Integrity in Academic Data

Preventing AI Hallucination

Large language models can fabricate citations, statistics, and author names. Every AI-assisted output must be cross-verified against primary sources before inclusion in any academic or policy document.

Repository Bias

Scopus and Web of Science over-represent English-language journals from the Global North. Researchers must acknowledge this bias explicitly and supplement with regional databases where relevant.

Methodological Transparency

Every step of the text analysis pipeline — from search query to final visualisation — must be documented in sufficient detail for independent replication. Opacity is a form of academic misconduct.

Data Attribution

Metadata exported from commercial databases carries licensing conditions. Always declare the source, export date, and any filters applied. Transparency protects both the researcher and the reader.

MODULE 2

Anatomy of a Search Query: Boolean Logic in Scopus and WoS

A well-constructed search query is the foundation of any rigorous literature review. Boolean operators control the logical relationships between search terms, determining which records are retrieved and which are excluded.

Core Operators

- **AND**: Both terms must appear. Narrows results.
- **OR**: Either term may appear. Broadens results.
- **NOT**: Excludes records containing the specified term.
- **Wildcards (* and ?)**: Match variable character strings (e.g., analyt* captures "analytics", "analytical", "analysis").

Example Scopus Query

```
TITLE-ABS-KEY(
  ("cognitive analytics" OR "big data analytics")
  AND
  ("policy report" OR "decision making")
)
```

This query searches titles, abstracts, and author keywords simultaneously, maximising recall while maintaining precision. Always test your query iteratively, checking a sample of retrieved records for relevance before full export.

Understanding the Major Repositories: Scopus vs WoS vs Lens.org

Scopus (Elsevier)

Largest abstract and citation database. Covers over 27,000 peer-reviewed journals. Strongest in natural sciences, engineering, and medicine. Excellent metadata completeness for author affiliations and keywords. Requires institutional subscription.

Web of Science (Clarivate)

Oldest and most prestigious citation index. Stricter journal selection criteria than Scopus. Particularly strong for impact factor calculations and Journal Citation Reports (JCR). Preferred by many funding bodies for evaluation purposes.

Lens.org

Open-access alternative aggregating Crossref, PubMed, and patent data. No subscription required. Useful for supplementing commercial databases, especially for grey literature and patent-linked research. Metadata quality is variable.

For maximum coverage in a systematic review, combine Scopus and WoS exports, then deduplicate. Use Lens.org as a supplementary check for open-access literature not indexed in the commercial databases.

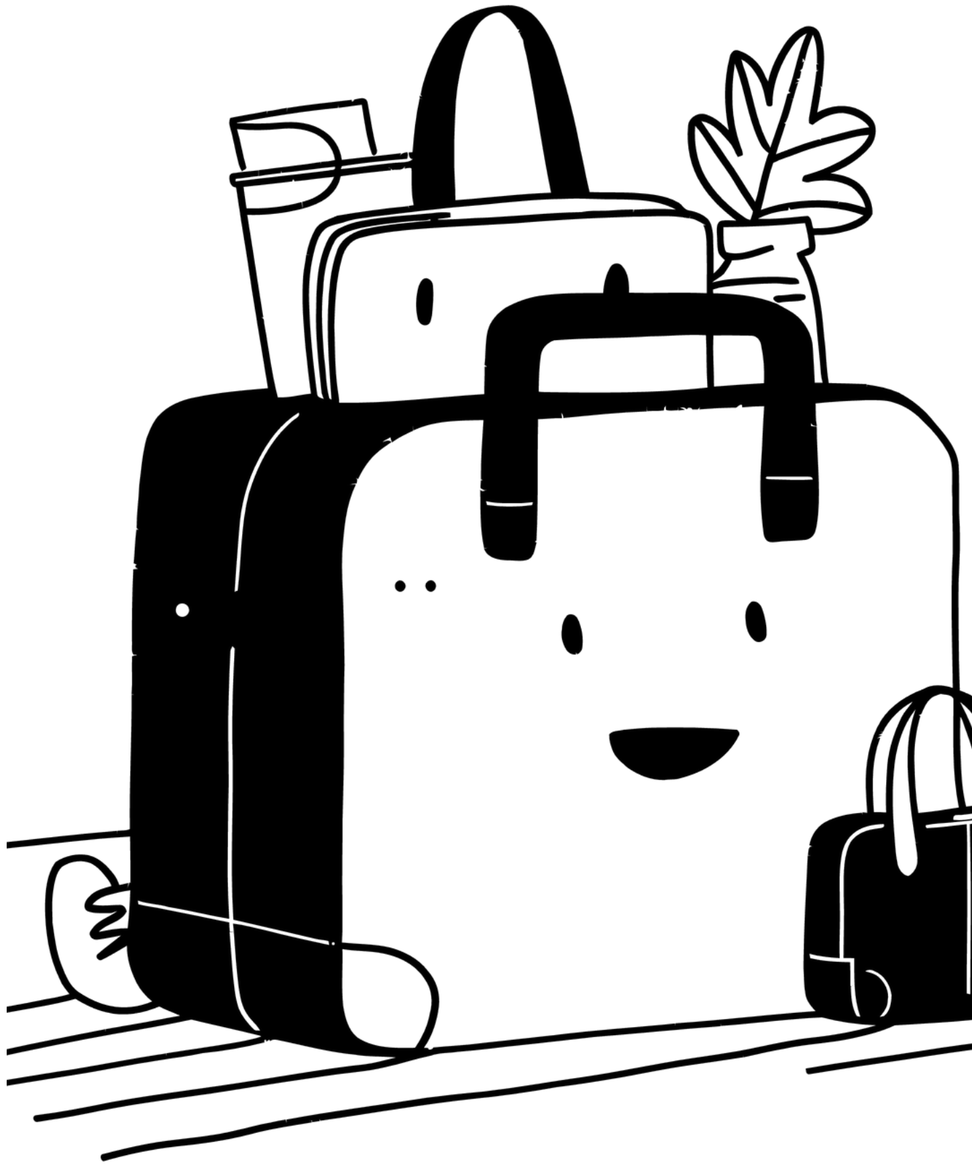
Extracting Metadata Files: BibTeX vs CSV

Why BibTeX (.bib) is Preferred

The .bib format preserves the full hierarchical structure of bibliographic records, including special characters (diacritics), nested author names, and field-level tagging. This structure is essential for downstream processing in R's bibliometrix package, which parses field tags directly.

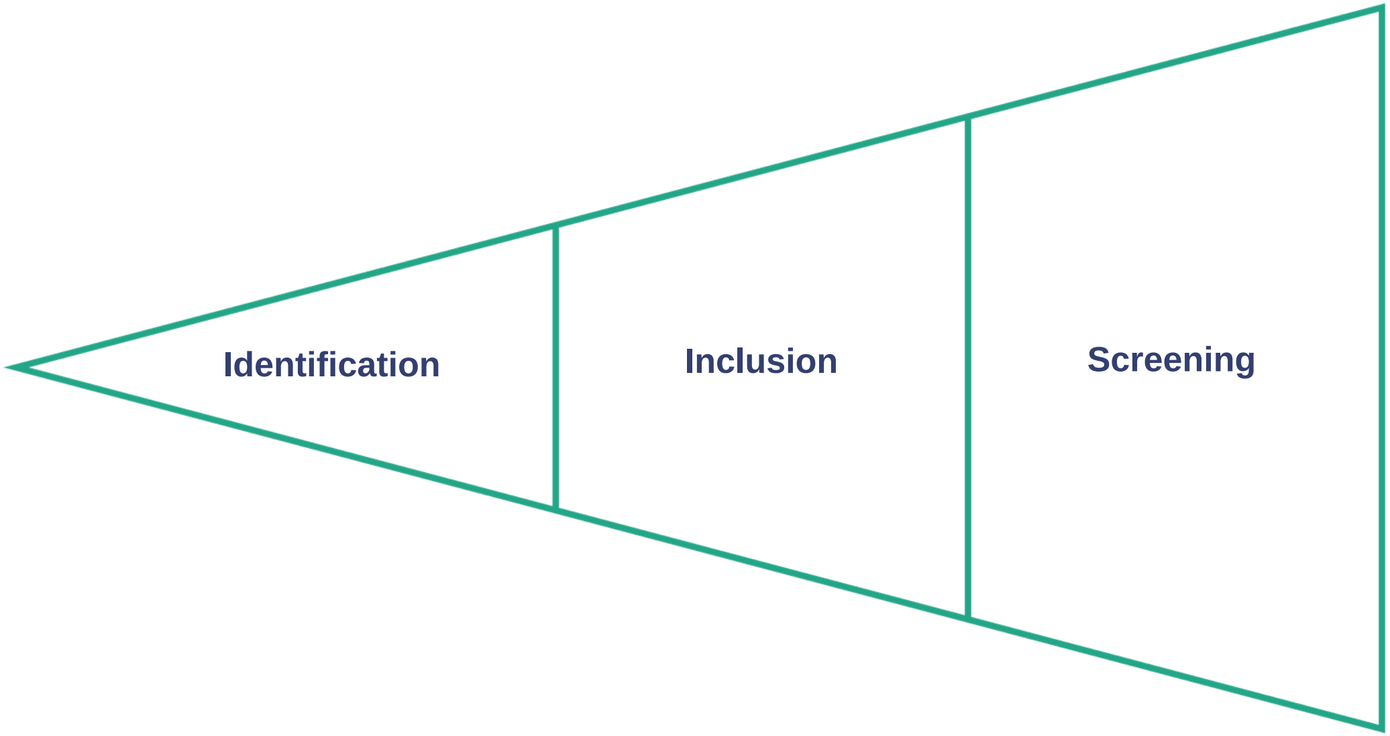
Limitations of CSV Export

Standard .csv exports flatten the record structure, often corrupting author name formatting, stripping diacritics, and merging multi-value fields (e.g., multiple authors or keywords) into a single cell. This creates significant preprocessing overhead and introduces errors into network analyses. Use CSV only when BibTeX is unavailable, and validate carefully before analysis.



PRISMA 2020 Standards for Data Collection

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) protocol provides the gold standard for transparent, reproducible literature collection.



Key Stages

- **Identification**: Run queries across all selected databases; record total hits per source.
- **Screening**: Remove duplicates; screen titles and abstracts against inclusion criteria.
- **Eligibility**: Assess full texts; document reasons for exclusion.
- **Inclusion**: Final corpus ready for cognitive analysis.

Reference

Page, M. J., et al. (2021). *The PRISMA 2020 statement: an updated guideline for reporting systematic reviews.* BMJ, 372, n71. **https://doi.org/10.1136/bmj.n71**

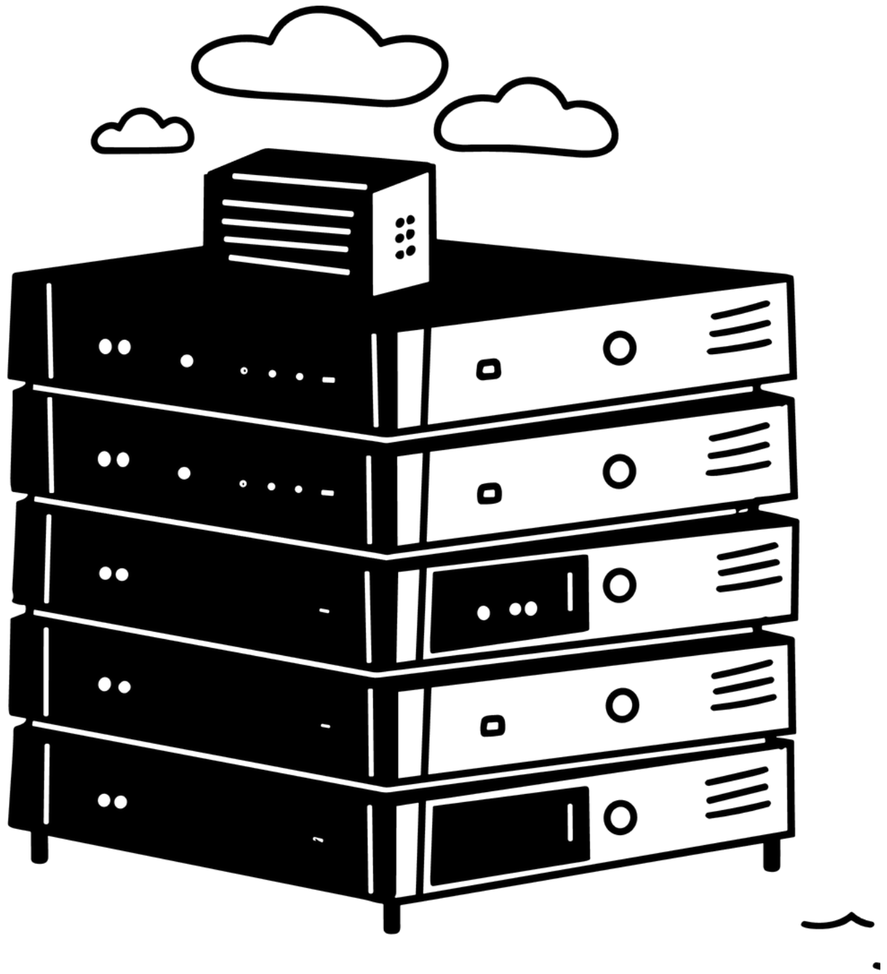
Handling Publication Bias and Predatory Journals

Scopus Quartile Filtering (Q1–Q4)

Scopus assigns journals to quartiles (Q1 = top 25% by CiteScore within their subject category). For high-quality cognitive analysis, restrict your corpus to Q1 and Q2 journals unless your research question specifically requires broader coverage. Q3 and Q4 journals carry higher risk of methodological inconsistency.

Scimago Journal Rank (SJR)

SJR weights citations by the prestige of the citing journal, providing a more nuanced quality signal than raw impact factor. Use SJR alongside quartile filtering to identify journals that are both productive and influential. The Scimago portal (**scimagojr.com**) is freely accessible and updated annually. Always document your quality thresholds in the Methods section of your report.



Converting and Merging Multi-Source Metadata

When combining exports from Scopus and Web of Science, deduplication is the critical first step. Duplicate records — the same article appearing in both databases — inflate frequency counts, distort network centrality measures, and introduce false co-citation links.

1

Export from Scopus

Download as BibTeX with all metadata fields selected. Record the query string, date, and total record count.

2

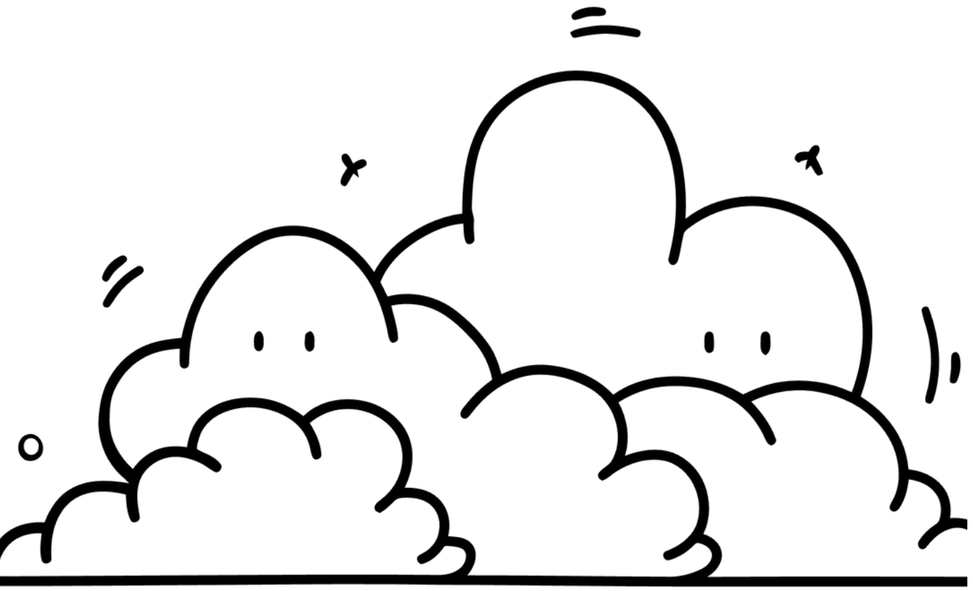
Export from WoS

Download as BibTeX (Plain Text format). Ensure "Full Record and Cited References" is selected for complete metadata.

3

Merge and Deduplicate

Use bibliometrix::mergeDbSources() in R, which identifies duplicates by DOI and title similarity, retaining the richer record from either source.



Academic APIs for Automated Data Retrieval

Elsevier API and pybliometrics

The Elsevier Developer Portal provides API keys granting programmatic access to Scopus metadata. The Python library pybliometrics wraps these endpoints, enabling scheduled, automated data refreshes — essential for living systematic reviews that must stay current with a rapidly evolving literature.

Practical Workflow

```
from pybliometrics.scopus import
ScopusSearch

query = 'TITLE-ABS-KEY("big
data" AND "policy")'
s = ScopusSearch(query)
results = s.results
```

Store your API key securely in an environment variable, never in the script itself. Set up a quarterly cron job to re-run the query and append new records to your master dataset, flagging additions for review.

Data Quality Checklist Before Analysis

Before any analytical step, audit your corpus systematically. Poor-quality input data produces misleading outputs — a principle sometimes called "garbage in, garbage out" that is especially consequential in network-based science mapping.

1 Author Name Completeness

Check the percentage of records with missing or malformed author fields. Rates above 5% require manual correction before co-authorship network analysis.

2 Affiliation Coverage

Affiliation data is essential for geographical mapping. Scopus typically provides better affiliation metadata than WoS. Flag records with missing affiliations for manual lookup.

3 Keyword Availability

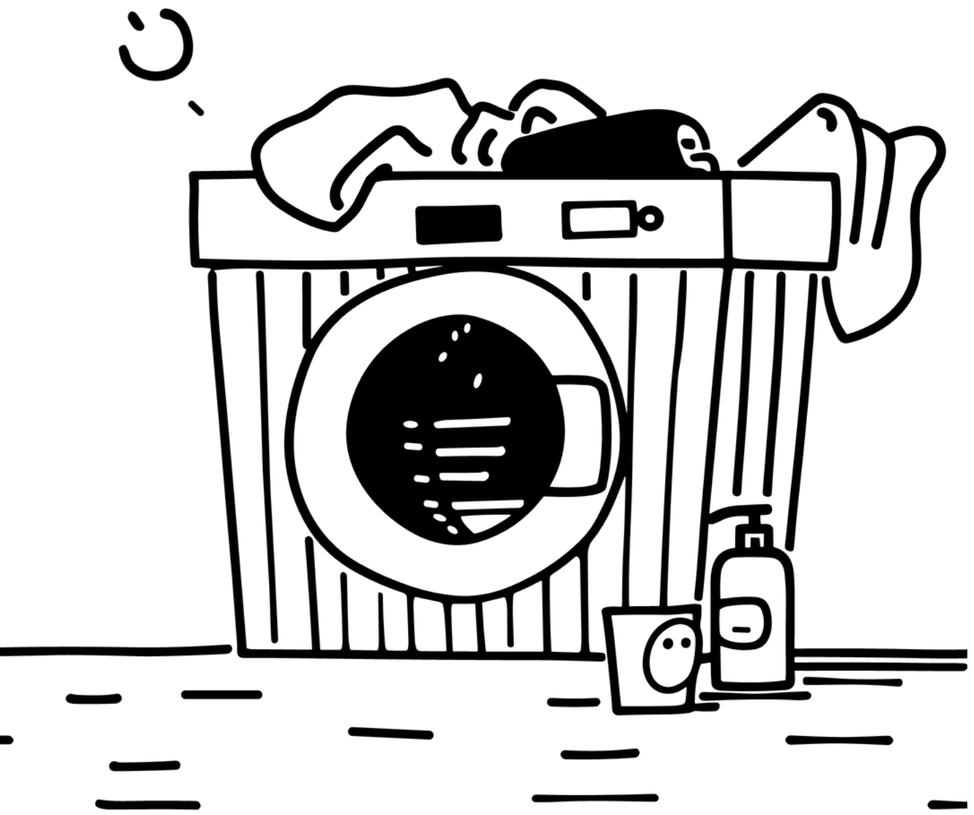
Author Keywords (DE field) and Keywords Plus (ID field) drive co-occurrence analysis. Corpora with fewer than 60% keyword-complete records will produce sparse, unreliable networks.

4 Abstract Presence

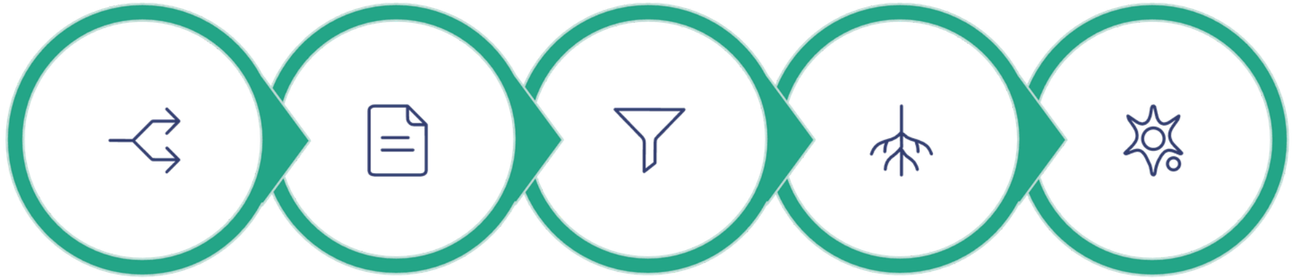
NLP and topic modelling require abstract text. Records without abstracts should be excluded from text-based analyses but may be retained for citation network work.

MODULE 3

Cognitive Text Preprocessing Workflow



Raw text from abstracts and titles cannot be fed directly into analytical algorithms. A structured preprocessing pipeline transforms unstructured prose into a clean, machine-readable format. Each step removes noise and standardises the input, improving the signal-to-noise ratio for all downstream analyses.



Stop-Words Specific to Academic Research

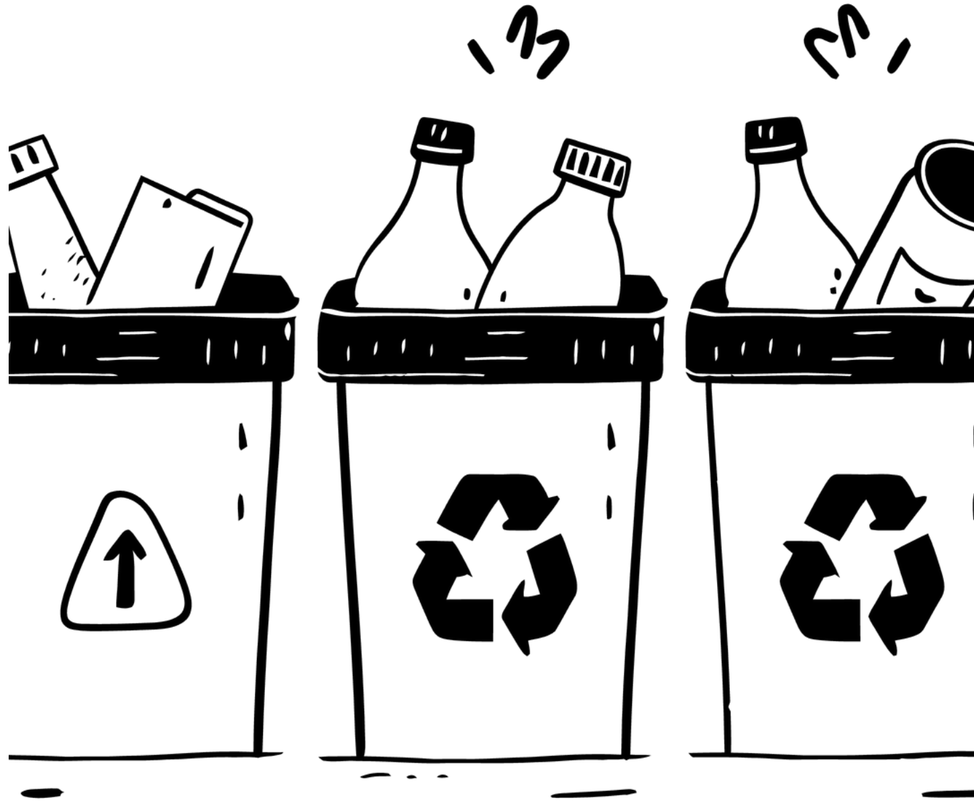
Why Generic Stop-Word Lists Are Insufficient

Standard stop-word lists (e.g., NLTK's English list) remove common function words like "the", "is", and "at". However, academic text contains a second layer of high-frequency but semantically empty terms that are specific to scholarly writing. These must be identified and removed separately.

Academic Stop-Words to Remove

- "study", "studies", "studied"
- "results", "findings", "outcomes"
- "analysis", "analyses", "analysed"
- "paper", "article", "research"
- "method", "approach", "framework"
- "propose", "proposed", "present"

These words appear in virtually every abstract and, if retained, will dominate topic models and keyword networks without adding discriminative information.



Synonym Merging: Building a Domain Thesaurus

Variant spellings, hyphenation differences, and disciplinary synonyms fragment what is conceptually a single term into multiple distinct tokens. This fragmentation artificially reduces the apparent frequency and network centrality of important concepts.

The Problem

"big data", "big-data", "large-scale data", "massive datasets", and "high-volume data" all refer to the same concept but will appear as separate nodes in a keyword network unless unified.

The Solution: Thesaurus File

In bibliometrix, create a thesaurus CSV with two columns: the variant form and the canonical replacement. Apply it using thesaurus() before network construction. This is one of the highest-impact preprocessing steps available.

Practical Tip

Build your thesaurus iteratively: run an initial keyword frequency table, identify the top 200 terms, then manually review for synonyms and variants. A domain expert's input at this stage is invaluable.

N-Gram Analysis: Capturing Meaningful Phrases

Unigram

Single tokens: "Data", "Analytics", "Policy". Fast to compute but loses phrase-level meaning. "Machine" and "Learning" treated as unrelated terms.

Bigram

Two-token sequences: "Big Data", "Machine Learning", "Policy Report". Captures the most common compound concepts in academic literature. The recommended default for keyword analysis.

Trigram

Three-token sequences: "Big Data Analytics", "Natural Language Processing". Captures highly specific technical phrases. Use selectively — trigrams are sparse and require large corpora to be statistically meaningful.

In practice, combine unigram and bigram analysis. Use unigrams for broad topic mapping and bigrams for identifying the specific compound concepts that define a research field's vocabulary.

Lemmatisation vs Stemming

Lemmatisation: Linguistically Aware

Lemmatisation reduces words to their dictionary base form (lemma) using grammatical knowledge. It preserves meaning: "better" becomes "good"; "running" becomes "run"; "studies" becomes "study". The output is always a real word. Preferred for academic text where precision matters. Requires a language model (e.g., spaCy's English pipeline).

Stemming: Fast but Crude

Stemming applies rule-based suffix stripping without linguistic awareness. "studying" becomes "studi"; "policies" becomes "polici". The output is often not a real word. Faster to compute but introduces ambiguity: "university" and "universal" may both stem to "univers", conflating unrelated concepts. Use stemming only when computational speed is the primary constraint.

Constructing the Document-Term Matrix (DTM)

The Document-Term Matrix is the numerical backbone of all text-based analyses. It transforms a corpus of unstructured text into a structured, machine-readable format that algorithms can process.

1

Rows = Documents

Each row represents one paper (or abstract) in the corpus. A corpus of 5,000 papers produces a matrix with 5,000 rows.

2

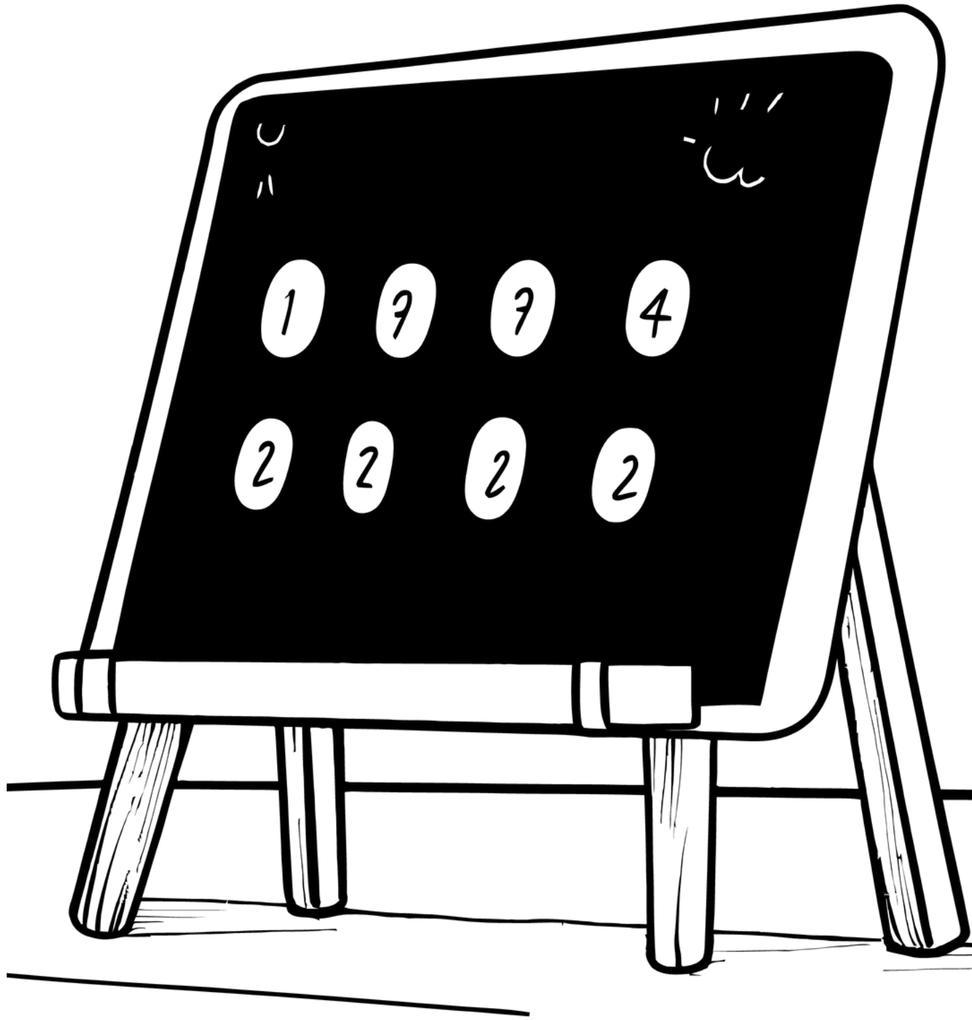
Columns = Terms

Each column represents one unique term (after preprocessing). A vocabulary of 10,000 terms produces 10,000 columns. Most cells are zero — the matrix is highly sparse.

3

Cell Values = Frequency

Each cell contains the count (or weighted score) of how often that term appears in that document. This numerical representation enables all downstream statistical and machine learning analyses.



Measuring Word Weight: TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) is the standard weighting scheme for identifying terms that are genuinely informative about a specific document, rather than merely common across the entire corpus.

The Formula

TF(t,d): How often term *t* appears in document *d*.
IDF(t): log(N / df(t)), where N is the total number of documents and df(t) is the number of documents containing term *t*.
TF-IDF(t,d) = TF(t,d) × IDF(t)
A term that appears frequently in one document but rarely across the corpus receives a high TF-IDF score — it is a strong signal of that document's specific topic.

Intuition: Diamond vs Gravel

Imagine weighing a rare diamond (a specific technical term like "bibliometric coupling") against a sack of gravel (a common word like "data"). The diamond is far more valuable precisely because of its rarity. TF-IDF formalises this intuition: a word that appears in every abstract is informationally worthless; a word that appears in only 3% of abstracts but dominates those 3% is a diagnostic signal of a distinct research cluster.

MODULE 4

Introduction to the bibliometrix Package in R

Why bibliometrix is the Gold Standard

The bibliometrix package provides a comprehensive, peer-reviewed framework for quantitative science mapping. It handles data import, cleaning, descriptive analysis, network construction, and visualisation within a single, reproducible R workflow. Its companion web application, biblioshiny(), makes the same analyses accessible through a point-and-click interface.

Key Reference

Aria, M., and Cuccurullo, C. (2017). *bibliometrix: An R-tool for comprehensive science mapping analysis.* Journal of Informetrics, 11(4), 959-975. https://doi.org/10.1016/j.joi.2017.08.007

This paper has itself accumulated thousands of citations, reflecting the package's adoption as the community standard for bibliometric research across disciplines.

Installation and Launching the biblioshiny Interface

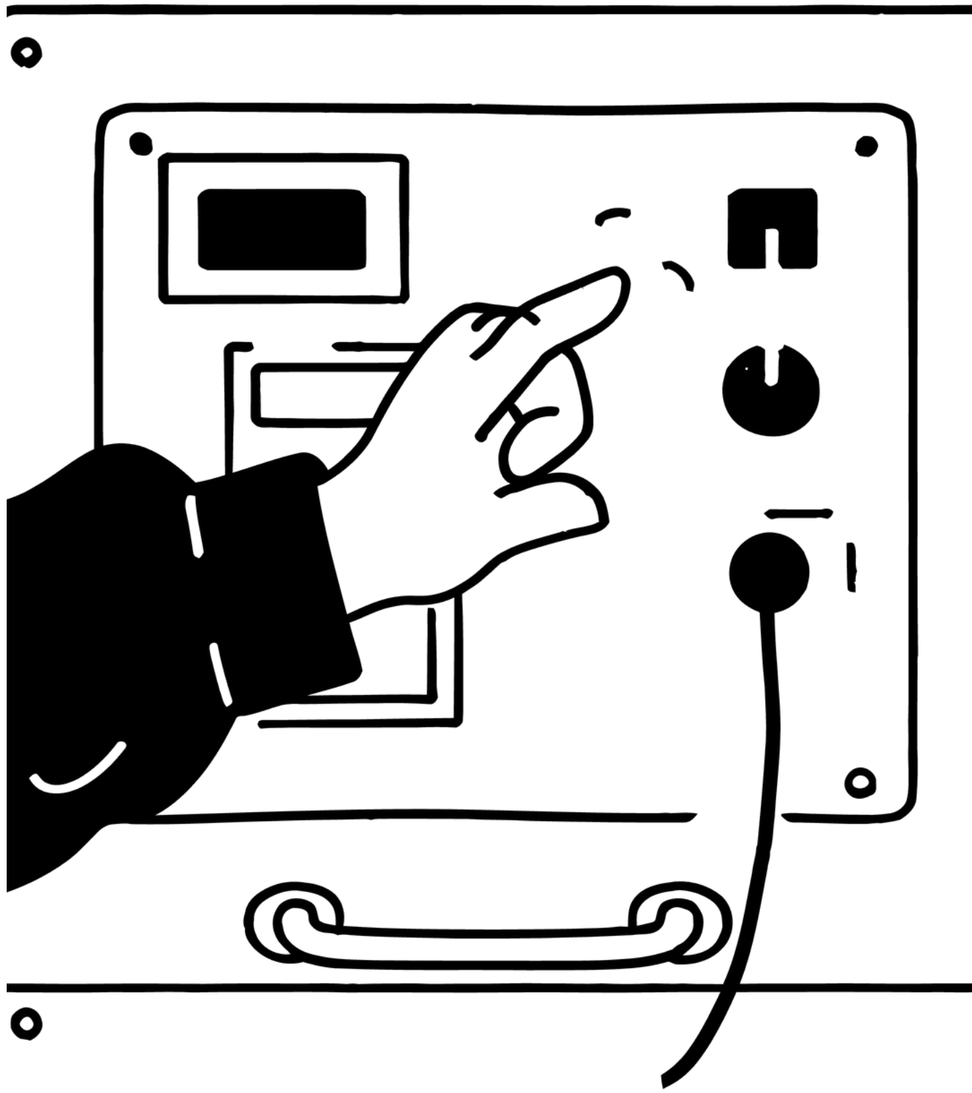
R Installation Script

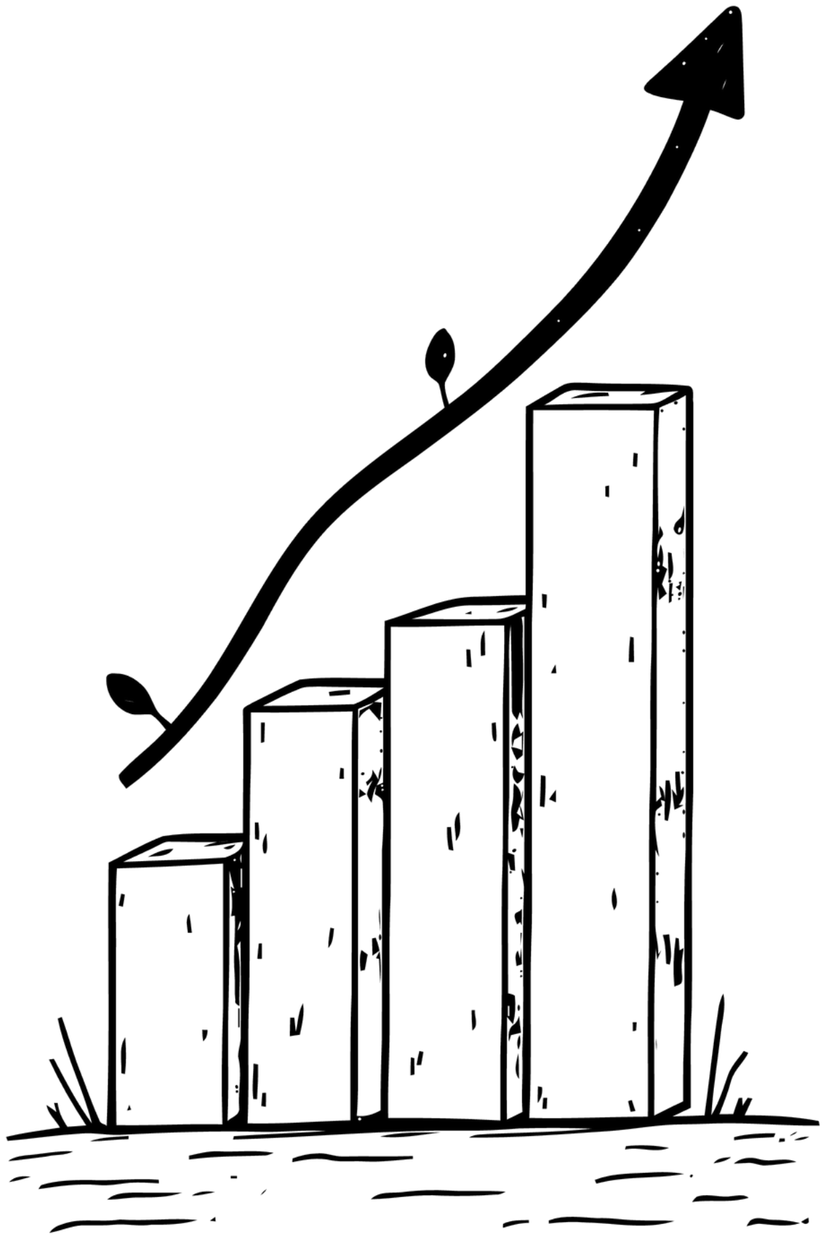
```
install.packages("bibliometrix")
library(bibliometrix)
biblioshiny()
```

Running biblioshiny() launches an interactive Shiny dashboard in your default web browser. No additional coding is required for most analyses — the interface guides you through data import, analysis selection, and export.

System Requirements

R version 4.0 or higher is recommended. The package has dependencies on ggplot2, igraph, Matrix, and stringr, all of which install automatically. For large corpora (more than 20,000 records), allocate at least 8 GB of RAM and consider running analyses in batch mode via script rather than the Shiny interface to avoid browser memory limits.





Descriptive Bibliographic Analysis

Before constructing networks, always begin with descriptive statistics. These provide the essential context for interpreting all subsequent analyses and are required in the Methods section of any bibliometric publication.

5M+

Papers Published Annually

Global scientific output, underscoring the necessity of computational analysis methods.

3

Core Metrics

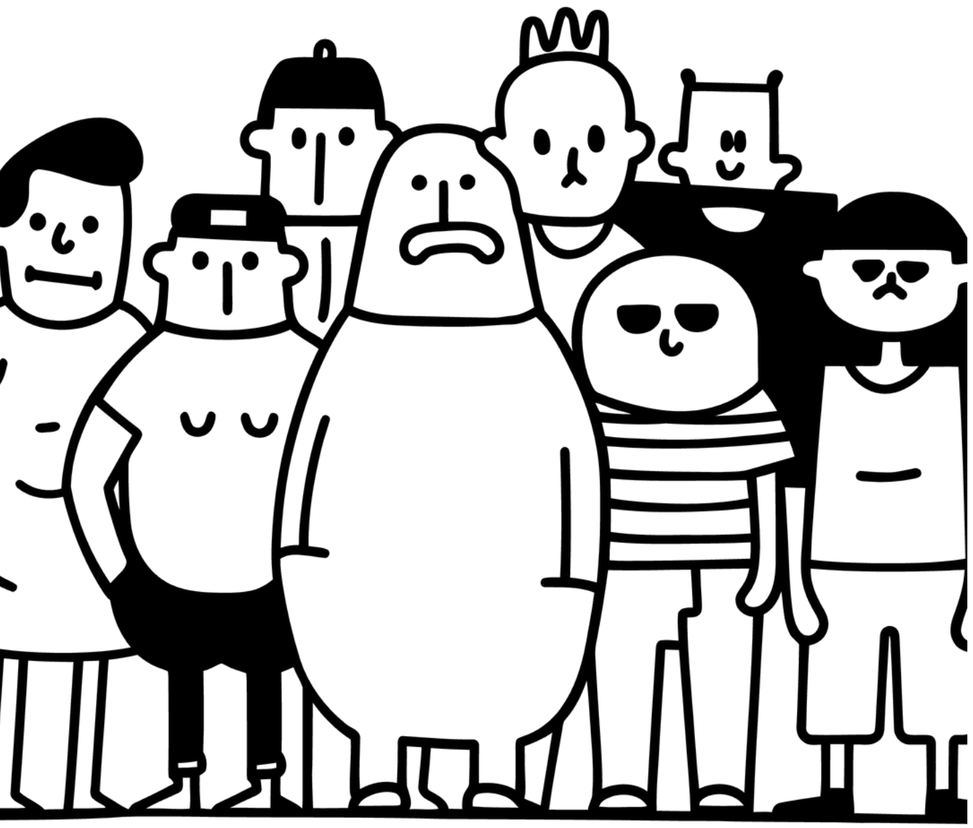
Annual Scientific Production, Average Citations per Article, and International Co-authorship Rate are the essential descriptive outputs.

300

DPI Minimum

Resolution required for publication-quality figure exports from bibliometrix analyses.

The summary(results) function in bibliometrix produces a structured report covering all key descriptive metrics. Review these before proceeding to network analysis — anomalies in the descriptive statistics often signal data quality issues that must be resolved first.



Lotka's Law and Author Productivity

The Law

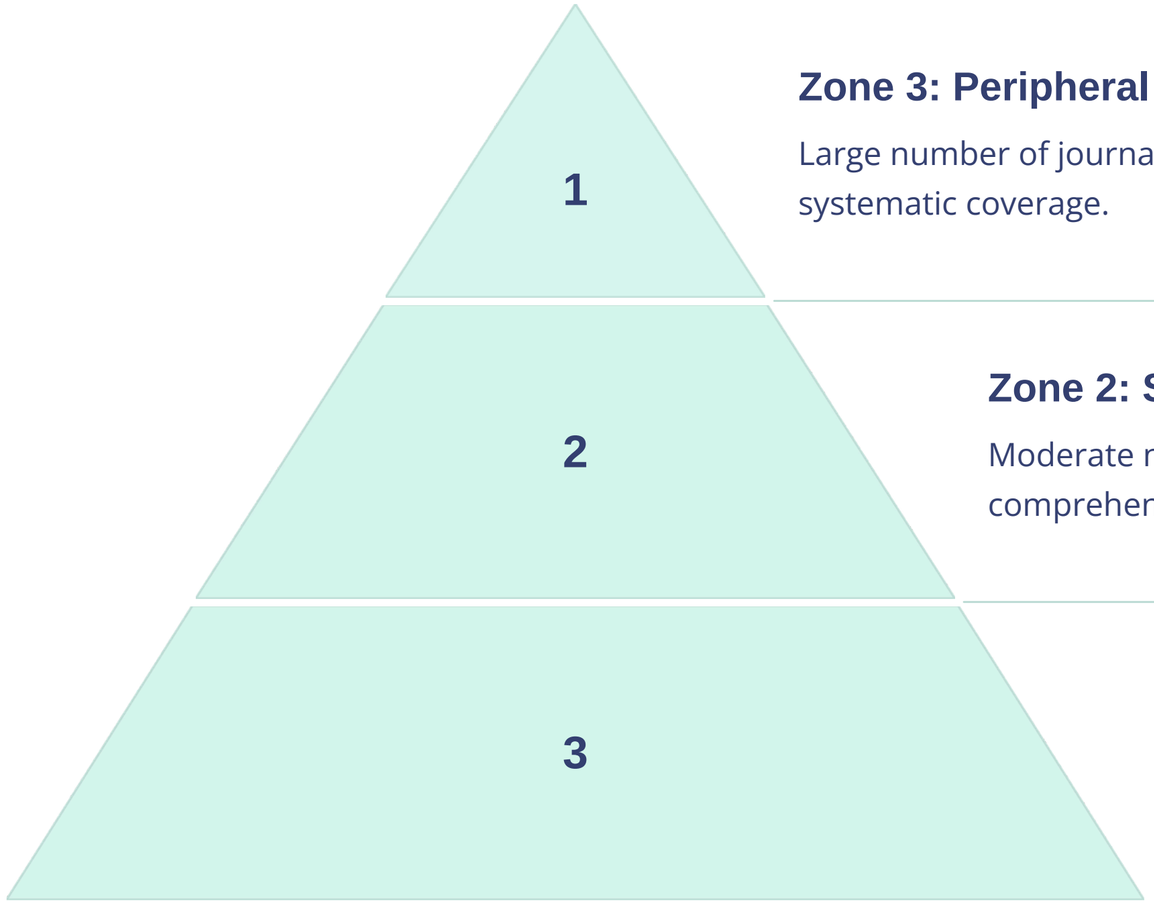
Lotka's Law states that the number of authors producing *n* papers is approximately $1/n^2$ of those producing one paper. In practice, a small elite of highly productive researchers generates the majority of publications in any field. Typically, around 10% of authors account for 50% or more of total output.

Analytical Implications

Identifying the core productive authors in your corpus is not merely a bibliometric curiosity — it reveals the intellectual leadership of a field. These authors' most-cited works are the papers most likely to define the theoretical foundations and methodological standards that shape all subsequent research. In a policy context, their institutional affiliations indicate where the most influential expertise is concentrated.

Bradford's Law: Identifying Core Journals

Bradford's Law of Scattering describes how scientific literature on a given topic is distributed across journals. It divides journals into three zones of decreasing productivity, each containing roughly equal numbers of relevant articles.



Zone 3: Peripheral

Large number of journals, each contributing very few relevant articles. Diminishing returns for systematic coverage.

Zone 2: Secondary

Moderate number of journals with moderate productivity. Important for comprehensive coverage but not essential for core reading.

Zone 1: Core

A small nucleus of journals producing the majority of relevant articles. These are the essential venues for both reading and publishing in the field.

Keyword Co-occurrence Analysis

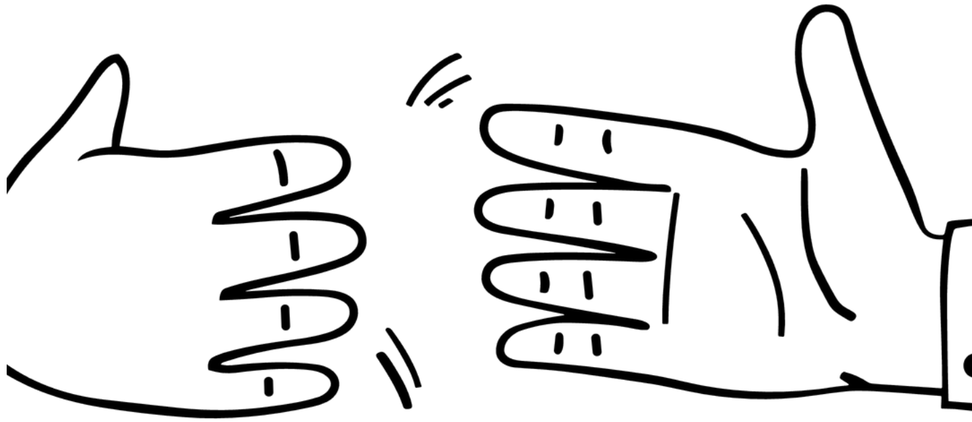
What It Measures

Keyword co-occurrence analysis identifies pairs of keywords that appear together within the same article more frequently than chance would predict. High co-occurrence indicates a strong conceptual relationship between the two terms — they tend to be discussed together because they are intellectually linked.

How to Run It in bibliometrix

```
NetMatrix <- biblioNetwork(M,
  analysis = "co-occurrences",
  network = "keywords",
  sep = ";")

net <- networkPlot(NetMatrix,
  normalize = "association",
  n = 50,
  Title = "Keyword Co-
occurrences",
  type = "fruchterman",
  size = TRUE)
```

The resulting network reveals the conceptual architecture of a research field — which ideas cluster together, which bridge different clusters, and which are isolated outliers.



Structural Evolution: Thematic Evolution Maps

Thematic evolution maps track how the conceptual structure of a research field changes over time. By dividing the corpus into time periods and comparing the keyword networks across periods, we can observe the emergence, growth, decline, and transformation of research themes.

1

2000–2010

Dominant themes: "Data Mining", "Knowledge Discovery", "Information Retrieval". Foundational computational methods applied to structured databases.

2

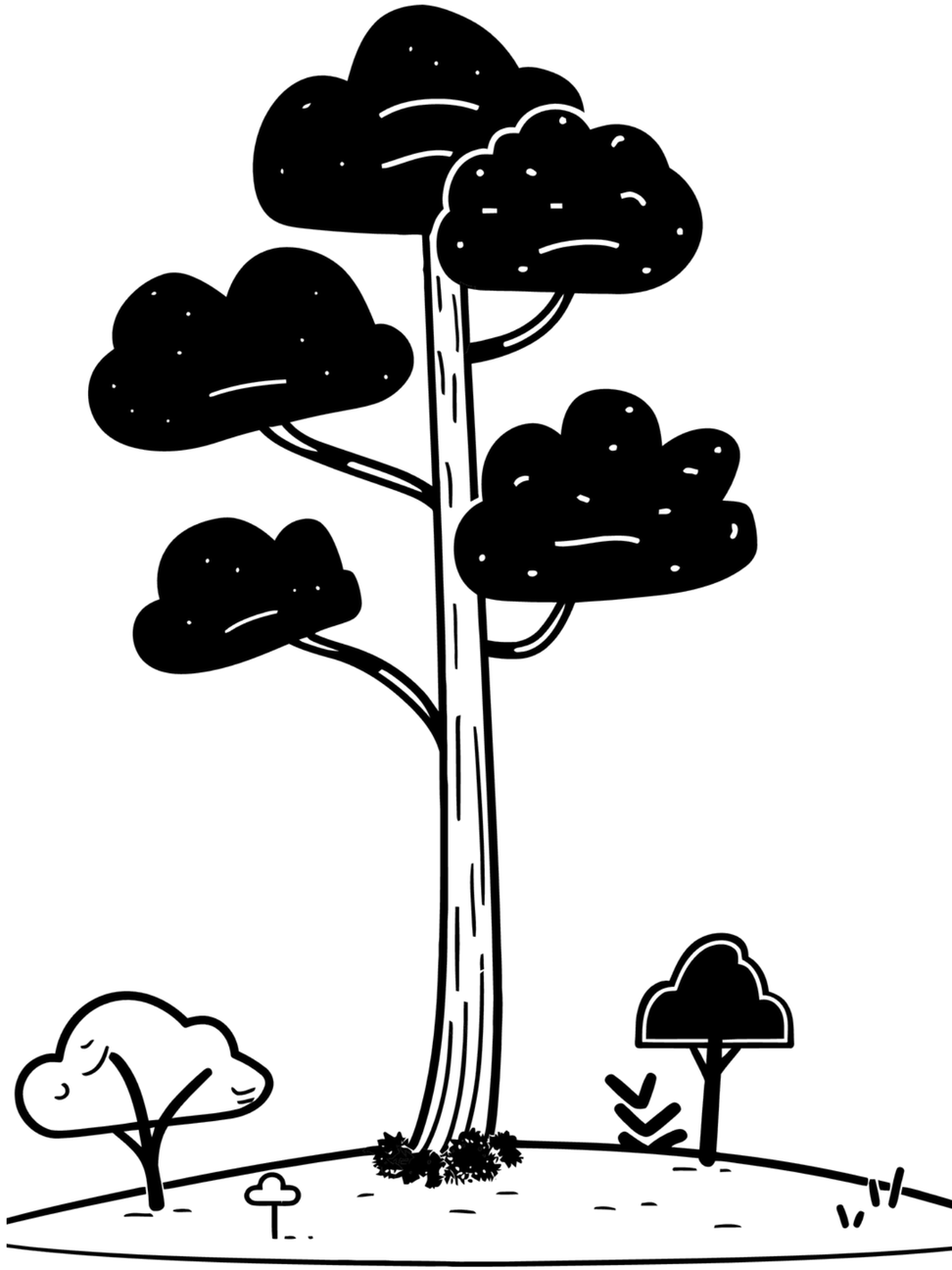
2010–2018

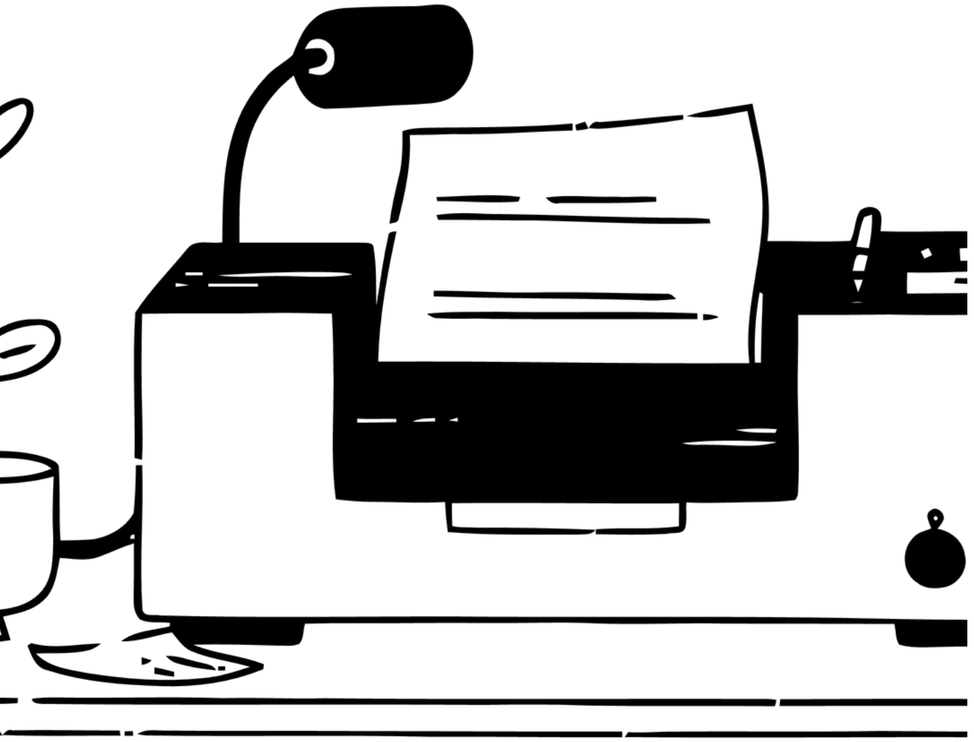
Emergence of "Big Data", "Cloud Computing", "Social Media Analytics". Scale and unstructured data become central concerns.

3

2018–2025

Dominance of "Deep Learning", "Natural Language Processing", "Large Language Models". Contextual understanding replaces frequency-based methods.





Exporting bibliometrix Results for Reporting

Descriptive Tables

Export summary statistics as CSV or Excel files. Include all key metrics: annual production, top authors, top journals, top keywords, and citation counts. These tables form the backbone of the Results section.

High-Resolution Figures

Export all network visualisations and trend plots at a minimum of 300 DPI in TIFF or PDF format. Most Elsevier journals require vector-format figures (PDF or EPS) for final submission.

Network Structure Files

Export adjacency matrices and edge lists in GraphML or GML format for further analysis in VOSviewer or Gephi. These files preserve the full network topology for advanced visualisation.

MODULE 5

Why Use VOSviewer?

The VOS Principle

VOSviewer implements the **Visualisation of Similarities (VOS)** mapping technique. Items (keywords, authors, or journals) are positioned in two-dimensional space such that the distance between any two items reflects their similarity: items that co-occur frequently are placed close together; items that rarely co-occur are placed far apart. The result is an intuitive spatial representation of the intellectual structure of a research field.

Key Reference

van Eck, N. J., and Waltman, L. (2010). *Software survey: VOSviewer, a computer program for bibliometric mapping.* Scientometrics, 84(2), 523-538. **https://doi.org/10.1007/s11192-009-0146-3**

VOSviewer is freely available at **www.vosviewer.com** and runs on Windows, macOS, and Linux. It accepts input directly from Scopus, WoS, and bibliometrix network exports.

Reading Colour, Node Size, and Line Thickness in Networks

Node Size = Frequency

Larger nodes represent keywords, authors, or journals that appear more frequently in the corpus. The biggest nodes are the most prominent concepts or contributors in the field — but prominence does not always equal importance.

Line Thickness = Link Strength

Thicker lines between two nodes indicate stronger co-occurrence or co-citation relationships. A thick line between two keywords means they are discussed together very frequently — a strong conceptual bond.

Colour = Cluster / Topic

Nodes of the same colour belong to the same cluster — a group of items that are more strongly connected to each other than to items in other clusters. Each colour represents a distinct "island of thought" within the field.

Clustering Techniques: Louvain and VOS Algorithms

The Louvain Algorithm

The Louvain method optimises **modularity** — a measure of how well a network divides into distinct communities. It works by iteratively moving nodes between clusters to maximise the density of connections within clusters relative to connections between clusters. The result is a partition of the network into "islands of thought" that are internally coherent and externally distinct.

VOS Clustering

VOSviewer's native clustering algorithm uses a resolution parameter (default = 1.0) to control cluster granularity. Higher resolution values produce more, smaller clusters; lower values produce fewer, larger clusters. There is no single "correct" resolution — the optimal setting depends on the research question and the size of the corpus. Always test multiple resolution values and select the one that produces the most interpretable and substantively meaningful partition.

Overlay Visualisation: Mapping Temporal Trends

The overlay visualisation in VOSviewer colours nodes by the average publication year of the documents in which they appear. This transforms the static network map into a temporal map of research evolution.

Blue Nodes: Established Topics

Keywords with low average publication years (e.g., 2005-2012) represent foundational, well-established concepts. These are the bedrock of the field — important but no longer at the frontier. Publishing in these areas requires a very strong novelty argument.

Yellow/Green Nodes: Emerging Topics

Keywords with high average publication years (e.g., 2020-2025) represent cutting-edge research directions. These are the areas where the field is currently growing fastest — high opportunity for novel contributions, but also higher uncertainty about long-term relevance.

Strategic Implication

Position your research at the intersection of established (blue) foundational concepts and emerging (yellow) frontier topics. This combination signals both credibility and novelty — exactly what editors and reviewers seek.

Density Visualisation: Finding Research Black Holes and Gold Mines

High-Density Areas: Saturated Fields

Bright, dense regions of the density map indicate topics that have attracted enormous research attention. These areas are intellectually rich but highly competitive. New contributions must offer exceptional novelty to stand out. For a policy report, these areas provide the most robust evidence base.

Low-Density Areas: Research Gaps

Dark, sparse regions represent topics that exist in the literature but have received relatively little attention. These are the "gold mines" — areas where a well-executed study can make a disproportionate contribution to knowledge. Identifying these gaps is one of the most valuable outputs of a bibliometric analysis, directly informing the "Future Research Directions" section of any academic paper or policy report.

Co-Citation vs Direct Citation Analysis

Direct Citation

Paper A directly cites Paper B. A direct citation link indicates that the author of A explicitly acknowledged B as relevant to their work. Direct citation networks reveal the intellectual lineage and influence hierarchy of a field — which papers are most frequently cited, and by whom.

Co-Citation

Papers B and C are both cited by Paper A (and many other papers). Co-citation does not require B and C to cite each other — they are linked by the fact that other authors consistently reference them together. Co-citation networks reveal the foundational literature clusters that define the intellectual structure of a field, independent of chronological citation chains.

Bibliographic Coupling

Papers A and B share many of the same references. Bibliographic coupling links papers that draw on the same intellectual foundations, even if they do not cite each other. Useful for identifying contemporary research fronts — clusters of recent papers addressing the same problem from a shared theoretical base.

Integrating VOSviewer with Gephi for Advanced Graph Analysis

Betweenness Centrality

Measures how often a node lies on the shortest path between other nodes. High betweenness centrality identifies **bridge keywords** — concepts that connect otherwise separate research clusters. These interdisciplinary bridges are often the most fertile ground for novel research, as they link communities that do not yet communicate directly.

Eigenvector Centrality

Measures a node's influence based not just on how many connections it has, but on the importance of those connections. A node connected to many highly central nodes scores higher than one connected to many peripheral nodes. In author networks, eigenvector centrality identifies the scholars whose work is most embedded in the intellectual core of the field — the true thought leaders, not merely the most prolific publishers.

Avoiding False Implications from Network Graphs

⚠ A visually beautiful network graph is not evidence of anything. Without rigorous contextual interpretation, it is decoration, not analysis.

Proximity is Not Causation

Two keywords appearing close together in a VOSviewer map means they co-occur frequently — not that one causes the other, or that they are theoretically related in any meaningful way. The researcher must provide the interpretive argument.

Cluster Labels Require Expert Judgement

Algorithms assign nodes to clusters based on connection patterns, not meaning. The researcher must read the actual papers in each cluster to assign an accurate, substantive label. Never label a cluster based solely on its most frequent keyword.

The Editor's Warning

Manuscripts that present network graphs without substantive interpretation of what the patterns mean for theory, methodology, or practice will be rejected at desk review. The graph is the starting point of the analysis, not its conclusion.

MODULE 6

What is Topic Modelling? Latent Dirichlet Allocation (LDA)

The Core Idea

LDA treats each document as a mixture of topics, and each topic as a probability distribution over words. Imagine every scientific abstract as a recipe: it might be 60% "machine learning", 30% "healthcare policy", and 10% "data governance". LDA discovers these latent recipes automatically from the patterns of word co-occurrence across the corpus.

Key Reference

Blei, D. M. (2012). *Probabilistic topic models.* Communications of the ACM, 55(4), 77-84. **https://doi.org/10.1145/2133806.2133826**

LDA is an unsupervised method — it requires no labelled training data. The researcher specifies the number of topics (K) and the algorithm infers the topic structure from the data. Selecting K is one of the most consequential methodological decisions in topic modelling.

Evolution from LDA to BERTopic: Contextual Embeddings

LDA (2003)

Bag-of-words model. Treats each word independently, ignoring word order and context. "Bank" in "river bank" and "bank account" are treated identically. Fast and interpretable, but semantically shallow.

Word2Vec / GloVe (2013-2014)

Distributed word representations. Words with similar contexts receive similar vector representations. Captures some semantic relationships but still operates at the word level, not the sentence level.

BERT (2018)

Bidirectional Encoder Representations from Transformers. Generates context-sensitive embeddings: the same word receives different vector representations depending on its surrounding context. Understands meaning, not just frequency.

BERTopic (2022)

Combines BERT embeddings with UMAP dimensionality reduction and HDBSCAN clustering to produce coherent, contextually meaningful topics. Significantly outperforms LDA on short texts such as abstracts.

Determining the Optimal Number of Topics: Topic Coherence Score

The Problem

In LDA, the researcher must specify K (the number of topics) before running the model. Too few topics produce broad, undifferentiated clusters that obscure meaningful distinctions. Too many topics produce fragmented, overlapping clusters that are difficult to interpret. There is no universally correct K — it depends on the corpus size, the breadth of the research field, and the analytical purpose.

The Coherence Score (Cv)

The coherence score measures the semantic similarity of the top words within each topic. High coherence indicates that the top words of a topic tend to co-occur in the same documents — they form a meaningful, interpretable cluster. Run models for K = 5, 10, 15, 20, 25, and 30. Plot coherence scores against K. Select the K value at the "elbow" of the curve — the point where additional topics no longer substantially improve coherence. Always supplement statistical selection with qualitative review of the top words in each topic.

Sentiment and Tone Analysis in Scientific Articles

Scientific writing is not emotionally neutral. Researchers express cognitive stances toward the technologies, methods, and findings they discuss — and these stances carry important information about the state of a field.

Optimistic Tone

Language expressing enthusiasm, potential, and promise: "transformative", "unprecedented", "breakthrough". Common in early-stage technology literature. Signals emerging fields with high investment and low critical scrutiny.

Cautious Tone

Language expressing qualification, limitation, and conditionality: "however", "limitations", "further research is needed". Signals a maturing field where initial enthusiasm is being tempered by empirical evidence.

Critical Tone

Language expressing scepticism, contradiction, or failure: "failed to replicate", "overstated", "methodological flaws". Signals a field undergoing paradigm revision or replication crisis. Highly valuable for policy reports that must assess the reliability of evidence.

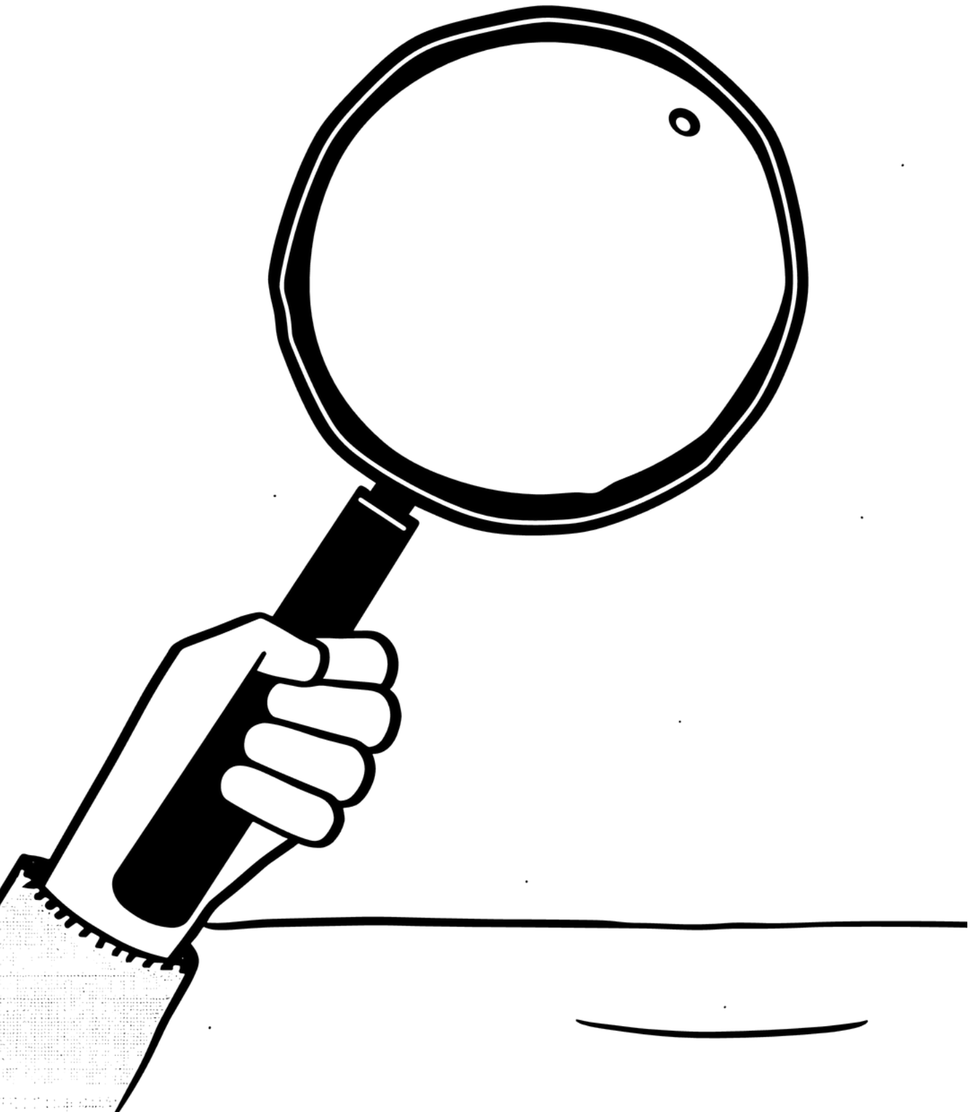
Distributed Cognitive Mapping: Word Embeddings and Word2Vec

The Vector Space Concept

Word embeddings represent each word as a dense numerical vector in a high-dimensional space. Words with similar meanings occupy similar regions of this space. The geometry of the space encodes semantic relationships: the vector distance between "king" and "queen" is similar to the distance between "man" and "woman".

Analogical Reasoning with Vectors

The famous example: **King - Man + Woman = Queen**. In the academic policy domain, similar analogical reasoning applies: **"Policy" - "Government" + "Technology" ≈ "E-Governance"**. This arithmetic of meaning allows researchers to explore conceptual relationships that would be invisible to keyword frequency analysis. Word2Vec, GloVe, and FastText are the standard tools for generating these embeddings from academic corpora.



Extracting Author Intent: Intent Mining

Every scientific abstract follows an implicit rhetorical structure. Intent mining identifies which sentences serve which communicative function — enabling automated extraction of problem statements, methodological choices, and policy recommendations at scale.

1

Problem Statement

Sentences identifying the gap, challenge, or question the paper addresses. Typically in the first 1-2 sentences of the abstract. Key signal words: "lack", "gap", "challenge", "unclear", "limited understanding".

2

Methodology

Sentences describing the data, methods, and analytical approach. Key signal words: "we collected", "using", "applied", "dataset", "sample", "model".

3

Policy Recommendation

Sentences translating findings into actionable guidance. Key signal words: "recommend", "suggest", "policy-makers should", "implications for", "call for". These sentences are the most valuable for policy report synthesis.

Cognitive Validation by Domain Experts: Human-in-the-Loop

Why AI/NLP Results Must Be Validated

Topic models and sentiment classifiers are trained on statistical patterns, not domain knowledge. They can produce clusters that are statistically coherent but semantically nonsensical to a subject expert. A topic labelled "data governance" by the algorithm might, on closer inspection, conflate three distinct research streams that a domain expert would immediately distinguish.

The Validation Protocol

After generating topics or sentiment labels, present the top 10-20 words per topic to at least two independent domain experts. Ask them to: (1) assign a meaningful label to each topic; (2) flag any topics that appear to conflate distinct concepts; (3) identify any important concepts that appear to be missing. Resolve disagreements through discussion. Document the validation process in the Methods section. This human-in-the-loop step is not optional — it is the difference between a publishable analysis and a computational exercise.

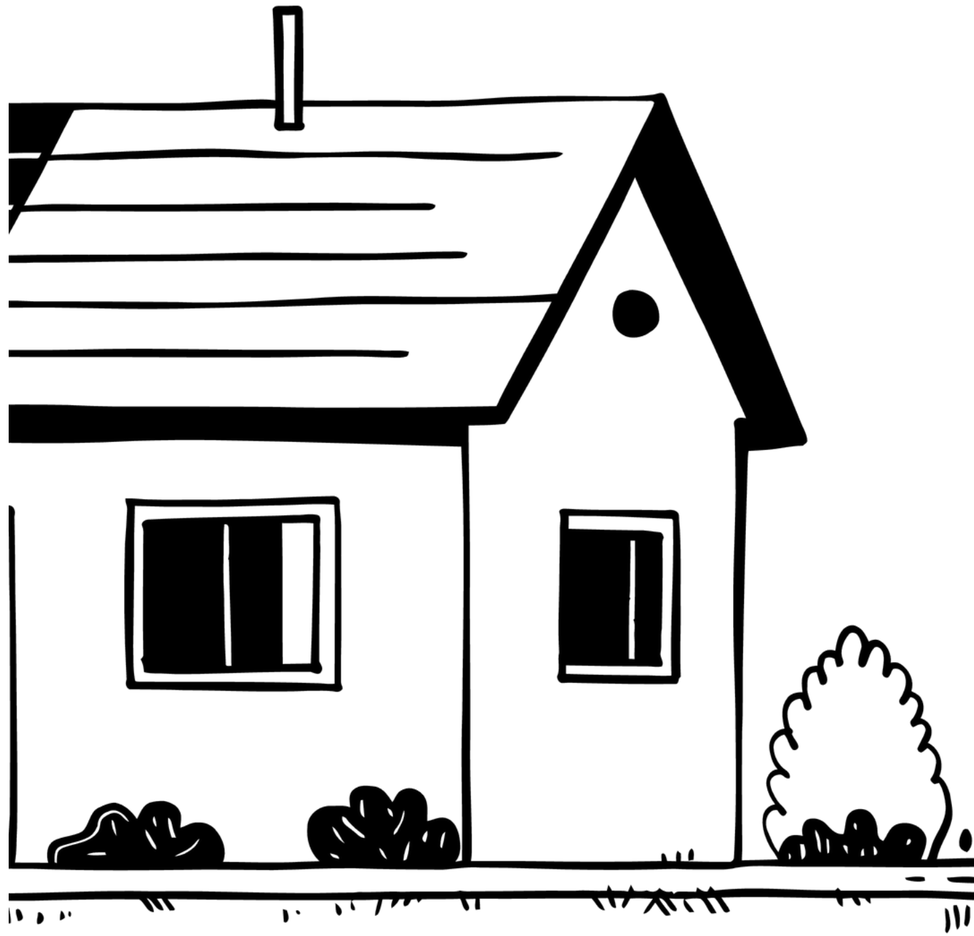
Summary of Cognitive NLP Tools: R vs Python

R Ecosystem

- tm: Core text mining framework. DTM construction, preprocessing, TF-IDF weighting.
- topicmodels: LDA implementation. Standard for academic bibliometric studies.
- stm: Structural Topic Model. Incorporates document-level metadata (e.g., publication year, country) as covariates, enabling richer topic analysis.
- quanteda: High-performance text analysis. Excellent for large corpora.

Python Ecosystem

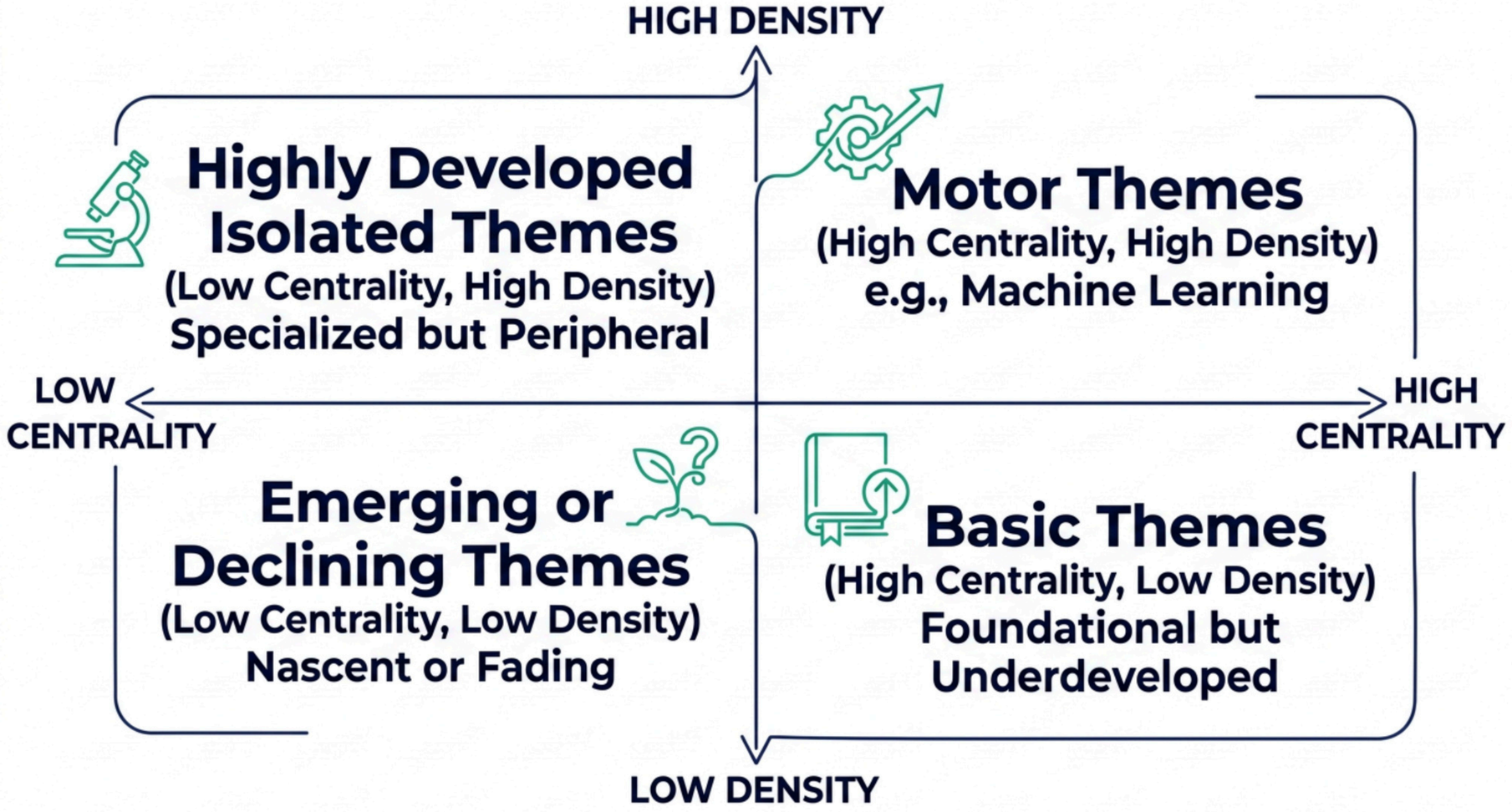
- spaCy: Industrial-strength NLP. Tokenisation, lemmatisation, named entity recognition. Fastest preprocessing pipeline available.
- gensim: LDA, Word2Vec, and Doc2Vec implementations. Scalable to very large corpora.
- BERTopic: State-of-the-art contextual topic modelling. Best performance on short texts (abstracts). Requires GPU for large-scale use.
- transformers (HuggingFace): Access to all major pre-trained language models including BERT, RoBERTa, and SciBERT (trained specifically on scientific text).



MODULE 7

Building the Strategic Diagram: Callon's Thematic Map

The strategic diagram, developed by Callon and colleagues, positions research themes in a two-dimensional space defined by two axes: **Centrality** (horizontal, measuring a theme's importance to the field) and **Density** (vertical, measuring a theme's internal development).



Motor themes (Quadrant 1) are the engine of a research field — well-developed and central to the mainstream discourse. Basic themes (Quadrant 4) are important but underdeveloped, representing opportunities for methodological or theoretical deepening. Emerging themes (Quadrant 3) are the frontier — high risk, high reward.

Identifying Research Gaps: The Cognitive Gap Matrix

Theoretical Gap

A concept or relationship that existing theory does not adequately explain. Identified when the literature consistently raises a question without providing a satisfactory theoretical answer. Example: "How does cognitive load affect the adoption of AI-assisted policy tools?" — a question present in the literature but without a unifying theoretical framework.

Methodological Gap

A limitation in the research methods used to study a phenomenon. Identified when the literature relies predominantly on one method (e.g., survey data) and lacks validation through complementary approaches (e.g., experimental or longitudinal designs). Methodological gaps are often the most tractable — they can be addressed by applying existing methods to existing questions.

Contextual/Empirical Gap

A lack of evidence from specific geographic regions, populations, or institutional contexts. Particularly common in policy research, where findings from the Global North are frequently generalised to the Global South without empirical validation. Addressing contextual gaps is both scientifically valuable and ethically important.

Qualitative vs Quantitative Synthesis: Integrating Meta-Analysis

The Quantitative Layer

Bibliometric maps, citation counts, keyword frequencies, and topic model outputs provide the quantitative skeleton of the synthesis. They reveal patterns, trends, and structures that are invisible to manual reading. But numbers alone do not constitute an argument — they require interpretation.

The Qualitative Layer

Close reading of the most-cited papers in each cluster provides the substantive content that gives meaning to the quantitative patterns. For each major cluster identified in the network analysis, read the top 5-10 papers by citation count. Extract the key theoretical claims, methodological innovations, and empirical findings. Weave these textual insights into the narrative of the synthesis, using the quantitative data as evidence and the qualitative reading as interpretation. The result is a synthesis that is both rigorous and readable.

Affiliation and Cross-Country Collaboration Mapping

Geographical Mapping in bibliometrix

The countryCollaboration() function in bibliometrix generates a country-level collaboration network, revealing which nations are most active in a research field and which international partnerships are most productive. This analysis is essential for any policy report addressing global challenges.

The Global North-South Divide

In most academic fields, research output is heavily concentrated in a small number of high-income countries (USA, UK, China, Germany, Netherlands). This concentration creates a systematic bias: the problems studied, the methods used, and the solutions proposed reflect the contexts of the Global North. For policy reports addressing developing country contexts, this bias must be explicitly acknowledged and, where possible, corrected by supplementing the mainstream literature with regional databases and grey literature from the Global South.

Research Leadership Dynamics: Top Authors and Institutions

Identifying Thought Leaders

The top 10-20 authors by total citations in a corpus are typically the field's intellectual leaders. Their most-cited papers define the theoretical vocabulary and methodological standards that all subsequent work must engage with. In a policy context, these are the experts whose testimony carries the most weight.

Institutional Consortia

Co-authorship analysis at the institutional level reveals the research consortia that drive knowledge production. Institutions with high betweenness centrality in the co-authorship network are the knowledge brokers — they connect otherwise separate research communities and facilitate the cross-pollination of ideas.

Dissemination Patterns

Tracking how ideas from leading authors spread through the citation network reveals the mechanisms of knowledge diffusion. Papers that are widely cited across multiple clusters are the "bridges" of the field — they translate specialised findings into broadly applicable insights.

Constructing a Novel Conceptual Framework

From Synthesis to Framework

A novel conceptual framework is not invented from scratch — it is synthesised from the patterns, gaps, and relationships identified across hundreds of papers. The framework makes explicit what the literature implies: the key constructs, their relationships, and the boundary conditions under which those relationships hold.

The Architecture of a Framework

- **Constructs**: The key concepts identified through topic modelling and keyword analysis.
- **Relationships**: The directional links between constructs, grounded in the theoretical arguments of the most-cited papers.
- **Boundary Conditions**: The contextual factors (e.g., institutional capacity, data availability) that moderate the relationships.
- **Propositions**: Testable statements derived from the framework that future empirical research can validate or refute.

Cognitive Synthesis Checklist Before Writing the Report

Before moving from analysis to writing, verify that your synthesis is complete and defensible. A checklist prevents the most common errors in translating bibliometric findings into policy-relevant narratives.

1 Have You Answered "So What?"

Every finding must be connected to a consequence. "The literature shows three dominant clusters" is not a finding — "The dominance of Cluster A indicates that policymakers have systematically neglected the role of institutional capacity in technology adoption" is a finding.

2 Have You Answered "Now What?"

Every "So What?" must generate an "Now What?" — a concrete implication for research, practice, or policy. If your synthesis cannot generate actionable implications, it is not yet complete.

3 Have You Identified the Gaps?

A synthesis without gap identification is a summary, not a contribution. Explicitly state the theoretical, methodological, and contextual gaps that your analysis has revealed, and explain why addressing them matters.

4 Is Your Framework Novel?

Compare your proposed conceptual framework against the existing frameworks in the literature. Articulate precisely what is new: a new construct, a new relationship, a new boundary condition, or a new synthesis of previously separate streams.

MODULE 8

The Fundamental Difference: Academic Manuscript vs Policy Report

Academic Manuscript

- Primary audience: peer researchers and journal editors
- Focus: methodological rigour, theoretical contribution, literature positioning
- Length: 8,000-12,000 words with full references
- Success metric: acceptance in a high-impact journal
- Timeline: months to years from submission to publication

Policy Report

- Primary audience: decision-makers, civil servants, and public stakeholders
- Focus: actionable solutions, implementation feasibility, public impact
- Length: 2-10 pages with an executive summary of one page or less
- Success metric: influencing a decision or allocating a resource
- Timeline: weeks — policy windows close fast

Reference: Groom, B., and Turvey, C. (2021). *Making policy-relevant scientific research count.* Environmental and Resource Economics, 79(2), 205-226. https://doi.org/10.1007/s10640-021-00562-x

Anatomy of a Standard Policy Report: Elsevier and World Bank Structure

1

1. Executive Summary

One page maximum. The entire report distilled into its most essential message. Written last, placed first. Must stand alone — many decision-makers will read nothing else.

2

2. Context and Problem

Why does this issue matter now? What is the cost of inaction? Establish urgency without alarmism. Ground the problem in data, not assertion.

3

3. Data-Driven Evidence

Present the key findings from the cognitive analysis. Translate complex visualisations into clear, accessible insights. Every claim must be traceable to a specific finding.

4

4. Policy Options and Trade-offs

Present 2-4 distinct policy options with their respective costs, benefits, risks, and feasibility assessments. Never present only one option — decision-makers need choice.

5

5. Actionable Recommendations

Specific, time-bound, resource-costed recommendations. Assign responsibility to named institutions. Define success metrics. This is the section that determines whether the report influences action.

Writing the Perfect One-Page Executive Summary

The Hook-Evidence-Action Formula

Hook: Open with the single most striking finding or the most urgent problem. One sentence. Make the reader stop scrolling. Example: "Across 12,000 peer-reviewed papers, one finding is unambiguous: current data governance frameworks are failing the communities they are designed to protect."
Evidence: Two to three sentences summarising the strongest evidence. Specific numbers, not vague claims. "Analysis of 10 years of publication data reveals a 340% increase in AI policy research, yet fewer than 8% of studies include empirical data from low-income countries."
Action: One to two sentences stating the single most important recommendation. Specific, not aspirational. "The Ministry of Digital Affairs should mandate open data sharing protocols for all publicly funded AI research by Q2 2026."

The Three-Minute Rule

A senior policy-maker has, on average, three minutes to decide whether a report is worth reading in full. The executive summary must earn that decision. Test your executive summary by reading it aloud — if it takes more than three minutes, it is too long. If a colleague cannot summarise its main point after one reading, it is too complex. Revise until both tests are passed.

Translating Complex Network Graphs into Clear Policy Visuals

The Simplification Imperative

A VOSviewer map with 500 nodes and 2,000 edges is analytically rich but communicatively useless for a policy audience. The researcher's task is to extract the 4-6 most policy-relevant insights from the full network and represent them in a format that a non-specialist can understand in 30 seconds.

The Simplification Protocol

- Identify the 3-5 largest clusters in the network. These represent the major research themes.
- Assign each cluster a plain-language label based on its top keywords and most-cited papers.
- Identify the 2-3 most policy-relevant findings within each cluster.
- Represent the clusters as a simple diagram (e.g., four labelled boxes or a hub-and-spoke model) with brief annotations.
- Include a footnote: "Simplified from a full network analysis of [N] keywords across [N] papers. Full technical report available on request."

This approach preserves validity while maximising communicative impact — the hallmark of excellent science communication.

Constructing the Policy Options and Trade-off Matrix

Decision-makers need structured comparisons, not narrative descriptions. The trade-off matrix presents policy options in a format that enables direct comparison across multiple evaluative dimensions.

Policy Option	Cost	Impact	Risk	Feasibility
Option A: Centralised Data Repository	High (infrastructure investment)	High (system-wide)	Medium (data security)	Medium (18-24 months)
Option B: Federated Data Sharing	Medium (coordination costs)	Medium (sector-specific)	Low (distributed risk)	High (6-12 months)
Option C: Open Data Mandate	Low (regulatory only)	High (long-term)	High (compliance resistance)	Low (requires legislation)

Formulating SMART Recommendations

Vague recommendations are the most common failure mode of policy reports. "The government should improve data infrastructure" is not a recommendation — it is a wish. SMART recommendations are the difference between a report that influences action and one that gathers dust.

Specific

Name the institution, the action, and the resource. Not "improve data systems" but "the Ministry of Digital Affairs should procure and deploy a national bibliometric analysis platform."

Measurable

Define the success metric. Not "increase research capacity" but "increase the number of R-certified data analysts in public sector institutions from 12 to 150."

Time-Bound

Specify the deadline. Example: "Ministry X allocates 5% of its annual research budget to install R-server infrastructure in regional offices by Q3 2026." A recommendation without a deadline is a suggestion.

Risk Assessment and Policy Mitigation

Unintended Consequences

Every data-driven policy recommendation carries the risk of unintended consequences — outcomes that the analysis did not predict and the policy-maker did not intend. A mandatory open data policy might inadvertently expose sensitive personal information. A centralised data repository might create a single point of failure for critical national infrastructure. Identifying these risks before implementation is far less costly than discovering them after.

The Mitigation Framework

- **Identify**: For each recommendation, brainstorm the three most plausible unintended consequences.
- **Assess**: Rate each consequence by probability (low/medium/high) and severity (low/medium/high).
- **Mitigate**: For high-probability or high-severity risks, propose a specific mitigation measure.
- **Monitor**: Define the early warning indicators that would signal a risk is materialising, and specify the monitoring mechanism.

Measuring Policy ROI: Return on Investment

Why ROI Matters for Policy Reports

Decision-makers operate under resource constraints. A recommendation that cannot demonstrate its expected return on investment — in terms of cost savings, efficiency gains, or measurable social outcomes — will struggle to compete for budget allocation against recommendations that can. Quantifying the expected ROI of data-driven policy is therefore not an optional add-on; it is a core component of a persuasive policy report.

Estimating Policy ROI

For each major recommendation, estimate: (1) the implementation cost (one-time and recurring); (2) the expected efficiency gain or cost saving (with a time horizon); (3) the net present value of the recommendation over a 5-year period; and (4) the break-even point. Where precise figures are unavailable, use ranges and sensitivity analysis. Always state your assumptions explicitly — a transparent estimate with acknowledged uncertainty is more credible than a precise figure with hidden assumptions.

PART II
R Software

R Software

https://www.r-project.org

[Home]

Download

CRAN

R Project

About R
Logo
Contributors
What's New?
Reporting Bugs
Conferences
Search
Get Involved: Mailing Lists
Developer Pages
R Blog

R Foundation

Foundation
Board

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To **download R**, please choose your preferred CRAN mirror.

If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

News

- **R version 4.1.1 (Kick Things)** has been released on 2021-08-10.
- **R version 4.0.5 (Shake and Throw)** was released on 2021-03-31.
- Thanks to the organisers of useR! 2020 for a successful online conference. Recorded tutorials and talks from the conference are available on the R Consortium YouTube channel.
- You can support the R Foundation with a renewable subscription as a supporting member

News via Twitter

The R Foundation

n has received a donation from the
SS Contributor Fund. Thank you

e.indeedeng.io/FOSS-Contribut…

naspaclust: Nature-Inspired Spatial Clustering

Implement and enhance the performance of spatial fuzzy clustering using Fuzzy Geographically Weighted Clustering with various optimization algorithms, mainly from Xin She Yang (2014) <ISBN:9780124167438> with book entitled Nature-Inspired Optimization Algorithms. The optimization algorithm is useful to tackle the disadvantages of clustering inconsistency when using the traditional approach. The distance measurements option is also provided in order to increase the quality of clustering results. The Fuzzy Geographically Weighted Clustering with nature inspired optimisation algorithm was firstly developed by Arie Wahyu Wijayanto and Ayu Purwarianti (2014) <doi:10.1109/CITSM.2014.7042178> using Artificial Bee Colony algorithm.

Version:	0.2.1
Depends:	R (≥ 3.5.0)
Imports:	Rdpack, rdist, stabledist, beepr
Suggests:	ppclust, spatialClust, cluster, ggplot2
Published:	2021-07-07
Author:	Bahrul Ilmi Nasution [aut, cre], Robert Kurniawan [aut], Rezzy Eko Caraka [aut]
Maintainer:	Bahrul Ilmi Nasution <bahrulnst at gmail.com>
License:	GPL-3
NeedsCompilation:	no
CRAN checks:	naspaclust results

Downloads:

Reference manual:	naspaclust.pdf
Package source:	naspaclust_0.2.1.tar.gz
Windows binaries:	r-devel: naspaclust_0.2.1.zip, r-release: naspaclust_0.2.1.zip, r-oldrel: naspaclust_0.2.1.zip
macOS binaries:	r-release (arm64): naspaclust_0.2.1.tgz, r-release (x86_64): naspaclust_0.2.1.tgz, r-oldrel: naspaclust_0.2.1.tgz
Old sources:	naspaclust archive

Linking:

Please use the canonical form https://CRAN.R-project.org/package=naspaclust to link to this page.

Install

- Go to your browser and search for R 4.1.1 windows
- Choose Download R4.1.1 for Windows
- Click, Download R4.1.1 for windows (64 megabytes, 32/64 bit)
- Then, save the file and run it after download is complete.
- Click next in all the popups that appear then finish.
- Open the installed r

https://www.r-project.org

- Go to your browser and search for R 4.1.1 mac
- Choose R for Mac OS X
- In the page that opens, click old in the second line
- Scroll down the page and choose R4.1.1.pkg
- Then, save the file and run it after download is complete.
- Click continue and then install
- Open the installed r

Install R Studio

- Open browser and search for r studio download
- Choose Download RStudio –RStudio
- Scroll down the page that opens and go to Installers
- Choose the download file according to your operating system
- Save the file and run it after download is complete.
- Click Next and then Finish
- Open R studio

Install

R Basic

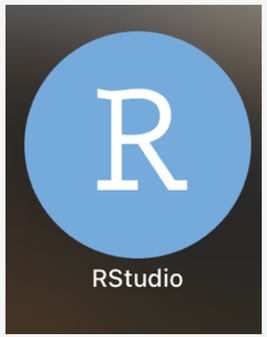
R Console

~

Help Search

```
R version 4.1.1 (2021-08-10) -- "Kick Things"
Copyright (C) 2021 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin17.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

During startup - Warning messages:
1: Setting LC_CTYPE failed, using "C"
2: Setting LC_COLLATE failed, using "C"
3: Setting LC_TIME failed, using "C"
4: Setting LC_MESSAGES failed, using "C"
5: Setting LC_MONETARY failed, using "C"
[R.app GUI 1.77 (7985) x86_64-apple-darwin17.0]

WARNING: You're using a non-UTF8 locale, therefore only ASCII characters will work.
Please read R for Mac OS X FAQ (see Help) section 9 and adjust your system preferences accordingly.
[Workspace restored from /Users/caraka/.RData]
[History restored from /Users/caraka/.Rapp.history]
```

R Studio



nn classification.R | dasar visualisasi.R | gwlr.R | cooc.R | boosting.R

```
1  library(rsample)      # data splitting
2  library(gbm)          # basic implementation
3  library(xgboost)      # a faster implementation of gbm
4  library(caret)        # an aggregator package for performing many machine learning models
5  library(h2o)          # a java-based platform
6  library(pdp)          # model visualization
7  library(ggplot2)      # model visualization
8  library(lime)         # model visualization
9
10
11 # Create training (70%) and test (30%) sets for the AmesHousing::make_ames() data.
12 # Use set.seed for reproducibility
13
14 set.seed(123)
15 ames_split <- initial_split(AmesHousing::make_ames(), prop = .7)
16 ames_train <- training(ames_split)
17 ames_test  <- testing(ames_split)
18
19
20 # for reproducibility
21 set.seed(123)
22
23 # train GBM model
24 gbm.fit <- gbm(
25   formula = Sale_Price ~ .,
26   distribution = "gaussian",
27   data = ames_train,
28   n.trees = 10000,
29   interaction.depth = 1,
30   shrinkage = 0.001,
31   cv.folds = 5,
32   n.cores = NULL, # will use all cores by default
33   verbose = FALSE
34 )
35
36 # print results
```

27:21 (Top Level) R Script

Console | Jobs

~/Desktop/File/DS UNIV INDO/UI/

```
+   formula = Sale_Price ~ .,
+   distribution = "gaussian",
+   data = ames_train,
+   n.trees = 10000,
+   interaction.depth = 1,
+   shrinkage = 0.001,
+   cv.folds = 5,
+   n.cores = NULL, # will use all cores by default
+   verbose = FALSE
+ )
>
```

Environment | History | Connections | Tutorial

Import Dataset | List

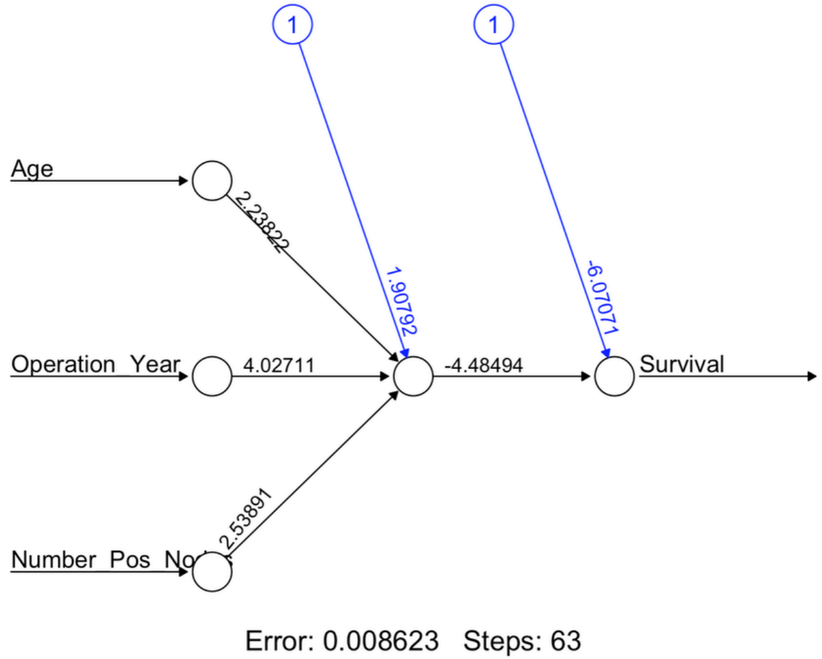
R | Global Environment

Data

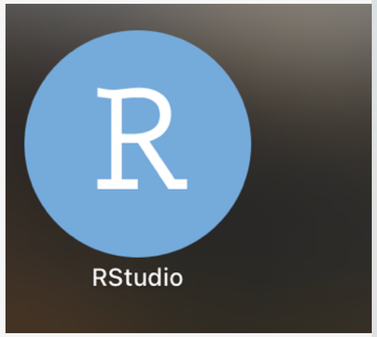
ames_split	Large mc_split (4 elements, 1.2 MB)
ames_test	879 obs. of 81 variables
ames_train	2051 obs. of 81 variables
BMA	List of 13
BPM	List of 13
c1	List of 2
c2	List of 2
c3	List of 12
Class_NN_ICs	12 obs. of 3 variables
coef.indosat	List of 11
data.gwlr	Formal class SpatialPointsDataFrame
dataku	38 obs. of 6 variables

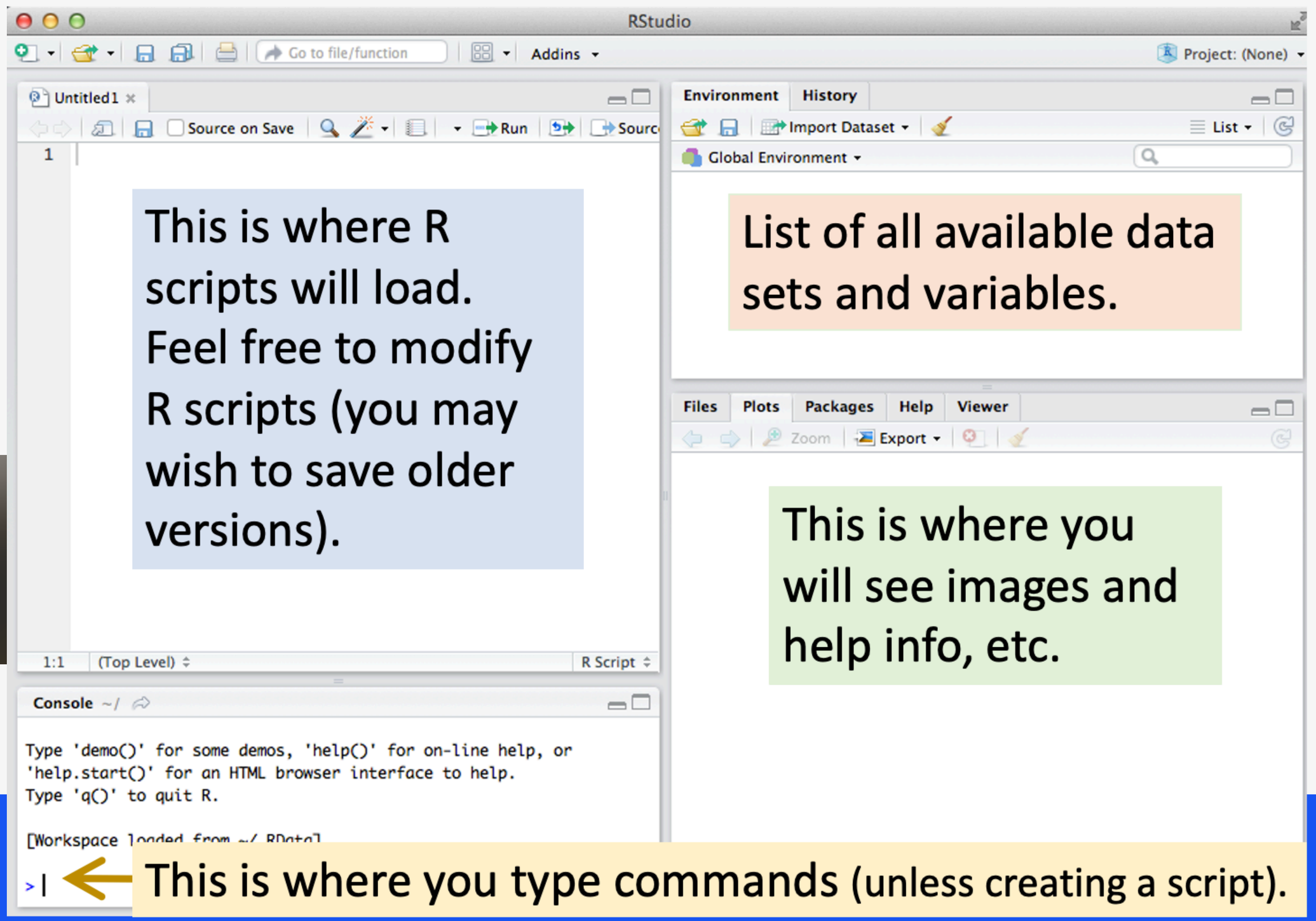
Files | Plots | Packages | Help | Viewer

Zoom | Export | Publish



R Studio


R
RStudio

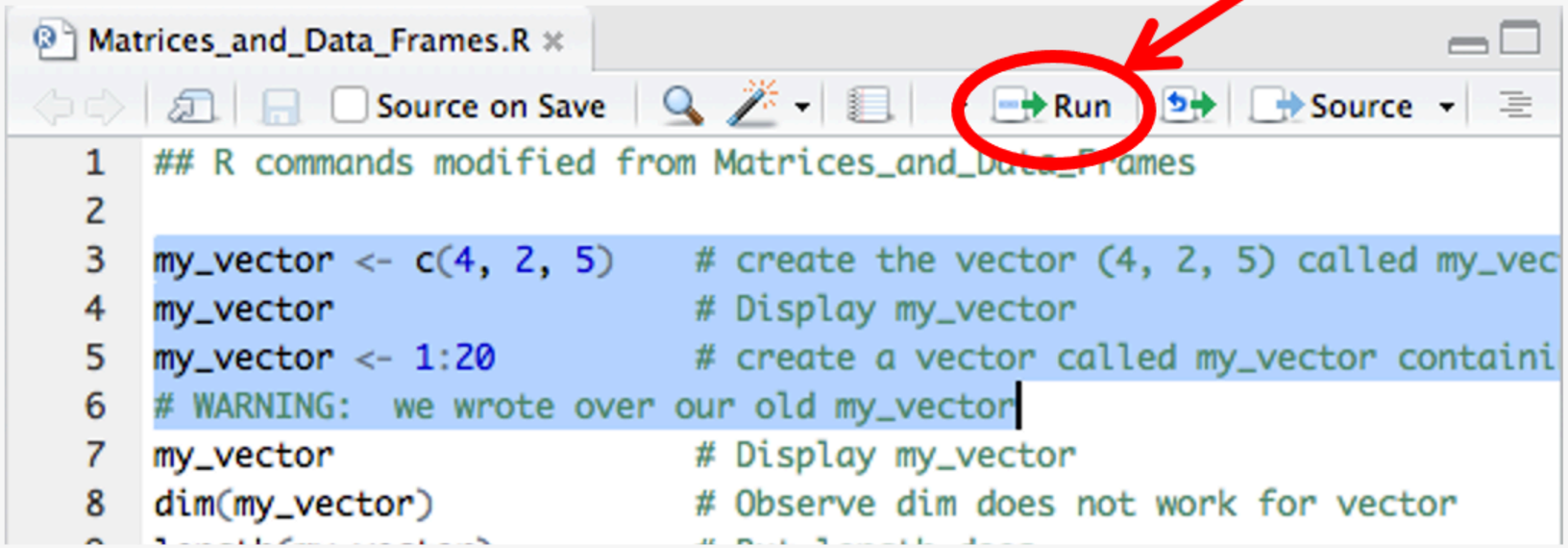

RStudio
Go to file/function
Addins
Project: (None)
Untitled1
Source on Save
Run
Source
1
This is where R scripts will load. Feel free to modify R scripts (you may wish to save older versions).
1:1
(Top Level)
R Script
Environment
History
Import Dataset
List
Global Environment
List of all available data sets and variables.
Files
Plots
Packages
Help
Viewer
Zoom
Export
This is where you will see images and help info, etc.
Console ~/
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
[Workspace loaded from ~/.RData]
>
This is where you type commands (unless creating a script).

To run an R script:

1.) Select the section you want to run and click run

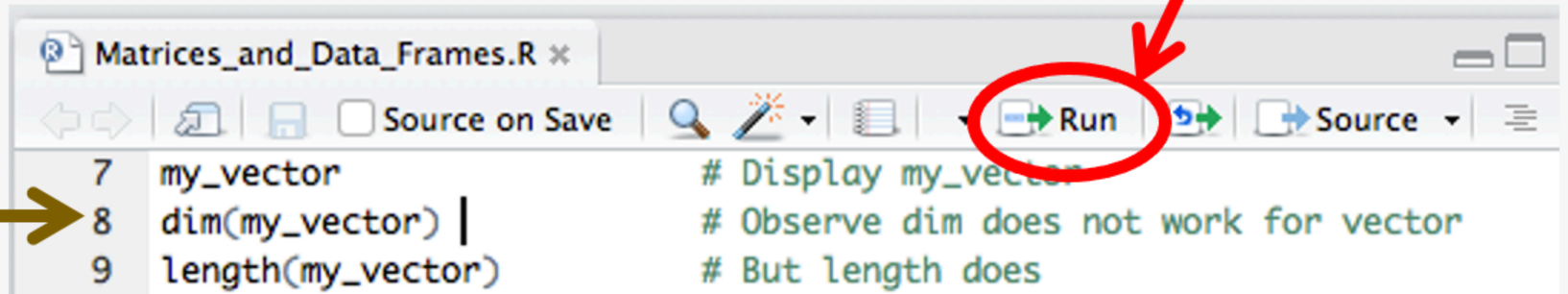
R Studio

Run blue highlighted portion



2.) Click on the line you want to run and click run.
Continue clicking run to run the code line by line.

Run line 8

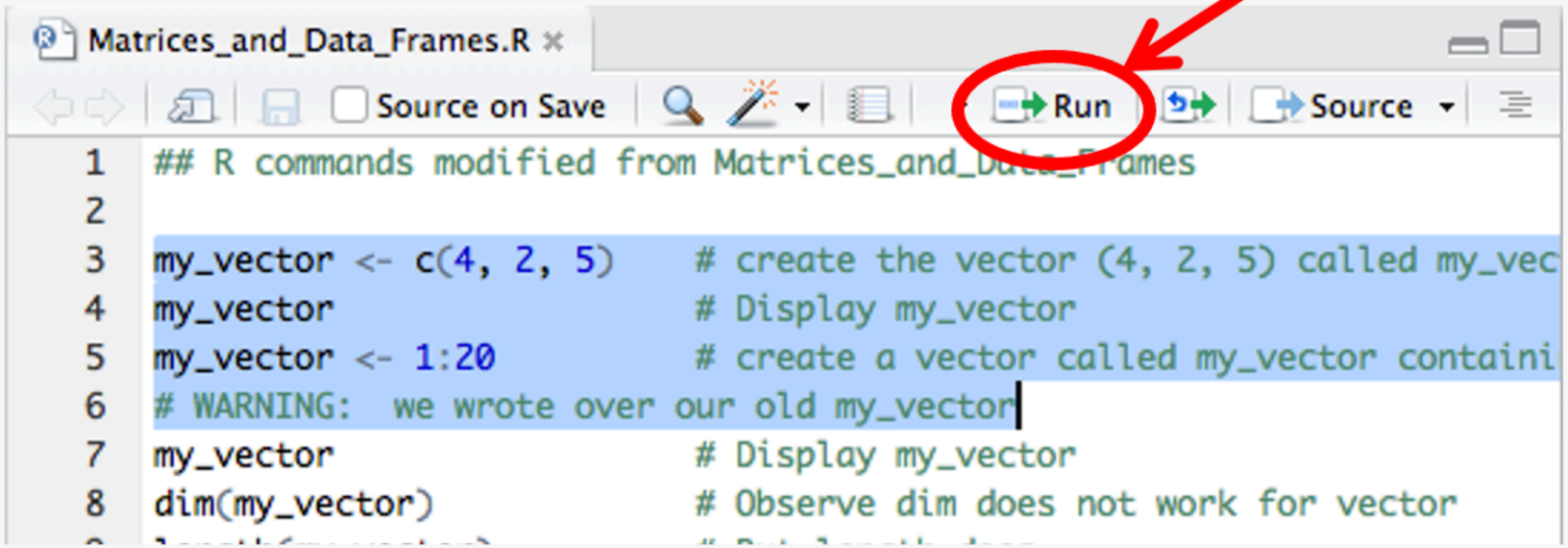


To run an R script:

1.) Select the section you want to run and click run

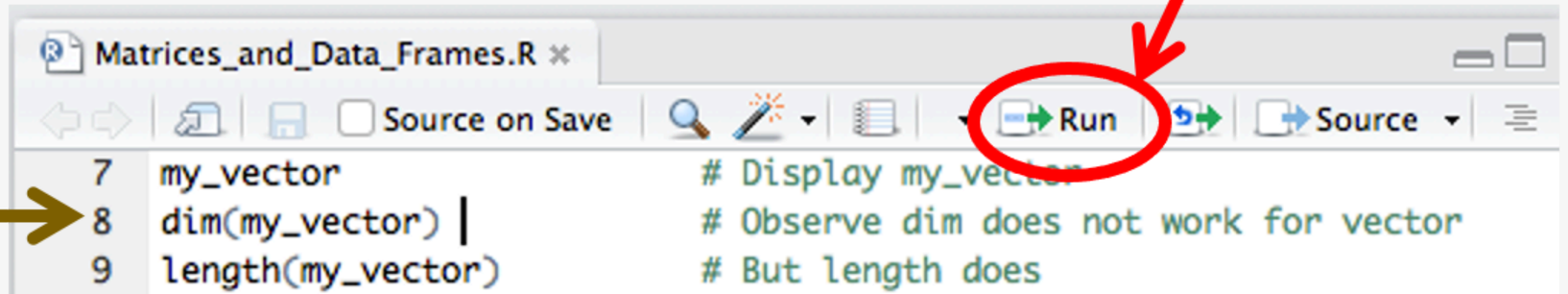
R Studio

Run blue highlighted portion

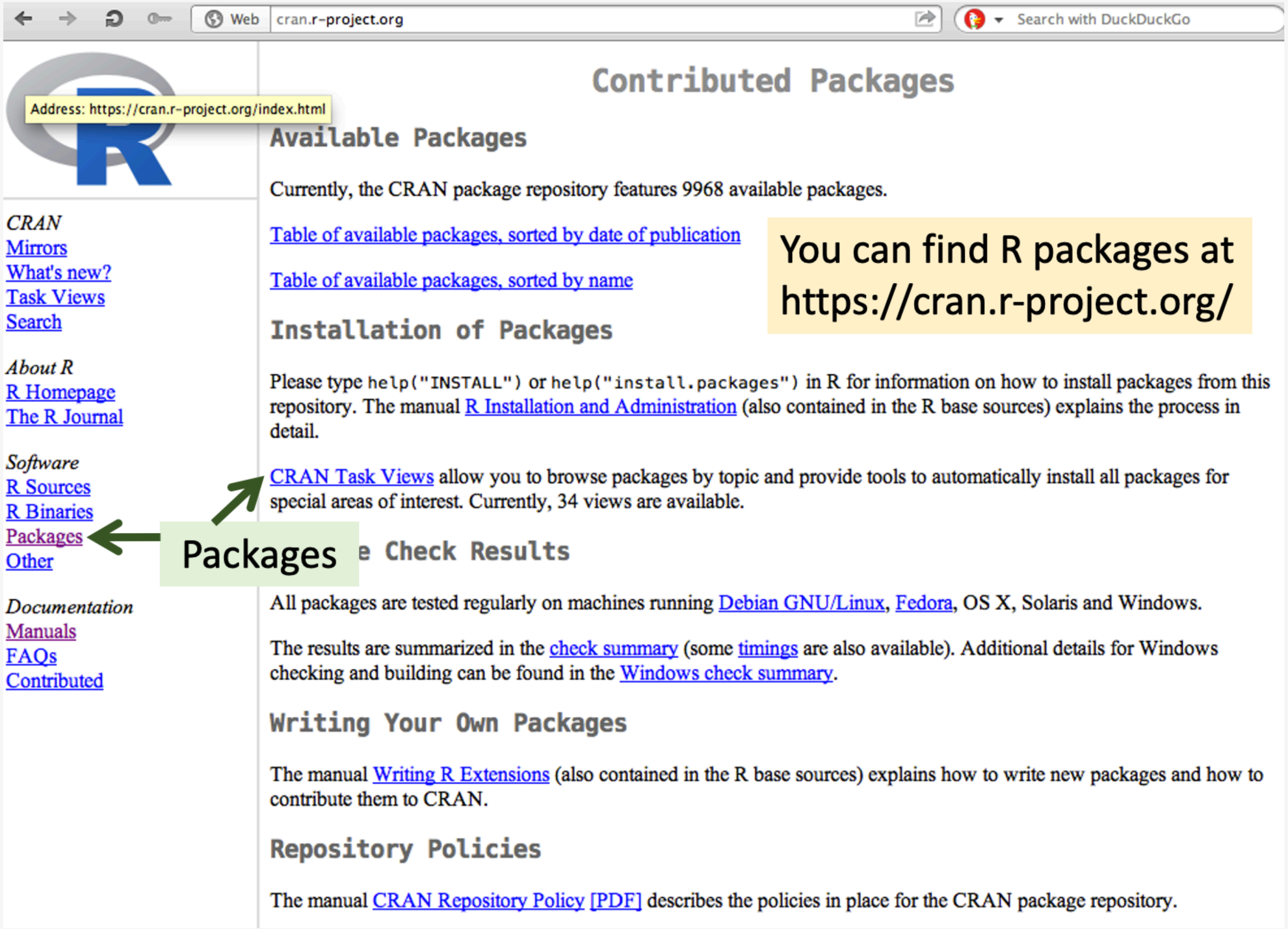


2.) Click on the line you want to run and click run.
Continue clicking run to run the code line by line.

Run line 8



R Studio


Web
cran.r-project.org
Search with DuckDuckGo
Address: https://cran.r-project.org/index.html
CRAN
Mirrors
What's new?
Task Views
Search
About R
R Homepage
The R Journal
Software
R Sources
R Binaries
Packages
Other
Documentation
Manuals
FAQs
Contributed
Packages
Contributed Packages
Available Packages
Currently, the CRAN package repository features 9968 available packages.
Table of available packages, sorted by date of publication
Table of available packages, sorted by name
You can find R packages at https://cran.r-project.org/
Installation of Packages
Please type help("INSTALL") or help("install.packages") in R for information on how to install packages from this repository. The manual R Installation and Administration (also contained in the R base sources) explains the process in detail.
CRAN Task Views allow you to browse packages by topic and provide tools to automatically install all packages for special areas of interest. Currently, 34 views are available.
e Check Results
All packages are tested regularly on machines running Debian GNU/Linux, Fedora, OS X, Solaris and Windows.
The results are summarized in the check summary (some timings are also available). Additional details for Windows checking and building can be found in the Windows check summary.
Writing Your Own Packages
The manual Writing R Extensions (also contained in the R base sources) explains how to write new packages and how to contribute them to CRAN.
Repository Policies
The manual CRAN Repository Policy [PDF] describes the policies in place for the CRAN package repository.

naspaclust: Nature-Inspired Spatial Clustering

Implement and enhance the performance of spatial fuzzy clustering using Fuzzy Geographically Weighted Clustering with various optimization algorithms, mainl
Optimization Algorithms. The optimization algorithm is useful to tackle the disadvantages of clustering inconsistency when using the traditional approach. The di
results. The Fuzzy Geographically Weighted Clustering with nature inspired optimisation algorithm was firstly developed by Arie Wahyu Wijayanto and Ayu Pur

Version:	0.2.1
Depends:	R (≥ 3.5.0)
Imports:	Rdpack, rdist, stabledist, beepr
Suggests:	ppclust, spatialClust, cluster, ggplot2
Published:	2021-07-07
Author:	Bahrul Ilmi Nasution [aut, cre], Robert Kurniawan [aut], Rezzy Eko Caraka [aut]
Maintainer:	Bahrul Ilmi Nasution <bahrulnst at gmail.com>
License:	GPL-3
NeedsCompilation:	no
CRAN checks:	naspaclust results

Downloads:

Reference manual:	naspaclust.pdf
Package source:	naspaclust_0.2.1.tar.gz
Windows binaries:	r-devel: naspaclust_0.2.1.zip, r-release: naspaclust_0.2.1.zip, r-oldrel: naspaclust_0.2.1.zip
macOS binaries:	r-release (arm64): naspaclust_0.2.1.tgz, r-release (x86_64): naspaclust_0.2.1.tgz, r-oldrel: naspaclust_0.2.1.tgz
Old sources:	naspaclust archive

Linking:

Please use the canonical form `https://CRAN.R-project.org/package=naspaclust` to link to this page.