Impact Analysis: Key Research Questions

Jinxi_Hu-48528608, Samarth_Grover-38220463

November 25, 2025

Environment Setup

```
library(readr)
library(igraph)
library(RColorBrewer)
library(ggplot2)
library(reshape2)
library(scales)
library(dplyr)
library(knitr)
library(wordcloud)
library(RColorBrewer)
set.seed(48528608)

# Create output directory
if (!dir.exists("plots")) {
  dir.create("plots", recursive = TRUE)
}
```

```
# Load data
nodes_connected <- read.csv("data/nodes_connected.csv")
edges_connected <- read.csv("data/edges_connected.csv")
nodes_all_raw <- read.csv("data/nodes.csv")

# Filter out invalid nodes from nodes_all
nodes_all <- nodes_all_raw %>%
  filter(
    !is.na(title) & trimws(title) != "",  # Valid title
    !is.na(local_id),                     # Valid ID
    !is.na(year) & year > 1900 & year <= 2025,  # Valid year range
    !is.na(citations) & citations >= 0,   # Valid citation count
    !is.na(references) & references >= 0, # Valid reference count
    !is.na(subtopic) & trimws(subtopic) != "",  # Valid subtopic
    !is.na(institution) & trimws(institution) != ""  # Valid institution
  )

# Create graph
graph <- graph_from_data_frame(edges_connected, vertices = nodes_connected, directed = TRUE)
graph <- simplify(graph, remove.multiple = TRUE, remove.loops = TRUE)
```

```r
# Calculate centrality metrics
V(graph)$in_degree <- degree(graph, mode="in")
V(graph)$out_degree <- degree(graph, mode="out")
V(graph)$pagerank <- page_rank(graph, directed = TRUE)$vector
V(graph)$betweenness <- betweenness(graph, directed = TRUE, normalized = TRUE)
V(graph)$closeness_in <- closeness(graph, mode = "in", normalized = TRUE)
V(graph)$eigenvector <- eigen_centrality(graph, directed = TRUE)$vector

# Add centrality metrics to nodes dataframe
nodes_analysis <- nodes_connected
nodes_analysis$in_degree <- V(graph)$in_degree
nodes_analysis$out_degree <- V(graph)$out_degree
nodes_analysis$pagerank <- V(graph)$pagerank
nodes_analysis$betweenness <- V(graph)$betweenness
nodes_analysis$closeness_in <- V(graph)$closeness_in
nodes_analysis$eigenvector <- V(graph)$eigenvector
```

1. Which paper is the most impactful?

```r
cat("=== Most Impactful Paper Analysis ===\n\n")
```

```
## === Most Impactful Paper Analysis ===
```

```r
# Find most impactful papers by different metrics
top_by_citations <- nodes_analysis[order(nodes_analysis$citations, decreasing = TRUE), ][1:10, ]
top_by_indegree <- nodes_analysis[order(nodes_analysis$in_degree, decreasing = TRUE), ][1:10, ]
top_by_pagerank <- nodes_analysis[order(nodes_analysis$pagerank, decreasing = TRUE), ][1:10, ]
top_by_betweenness <- nodes_analysis[order(nodes_analysis$betweenness, decreasing = TRUE), ][1:10, ]

# Create comprehensive impact score (normalized weighted average)
nodes_analysis$impact_score <- scale(nodes_analysis$citations)[,1] * 0.3 +
                               scale(nodes_analysis$in_degree)[,1] * 0.25 +
                               scale(nodes_analysis$pagerank)[,1] * 0.25 +
                               scale(nodes_analysis$betweenness)[,1] * 0.2

top_impact <- nodes_analysis[order(nodes_analysis$impact_score, decreasing = TRUE), ][1:15, ]

cat("Top 5 papers by citation count:\n")
```

```
## Top 5 papers by citation count:
```

```r
kable(top_by_citations[1:5, c("title", "year", "citations", "institution", "subtopic")])
```

	title	year	citations	institution	subtopic
166	Artificial intelligence in healthcare	2018	2383	Harvard University	Artificial Intelligence in Healthcare and Education

	title	year	citations	institution	subtopic
55	Large language models in medicine	2023	2361	University Of Cambridge	Artificial Intelligence in Healthcare and Education
2	AI in health and medicine	2022	1931	Harvard University	Artificial Intelligence in Healthcare and Education
233	Explainability for artificial intelligence in healthcare: a multidisciplinary perspective	2020	1394	Eth Zurich	Artificial Intelligence in Healthcare and Education
1	Foundation models for generalist medical artificial intelligence	2023	1155	Stanford University	Machine Learning in Healthcare

```r
cat("\nTop 5 papers by PageRank:\n")
```

```
##
## Top 5 papers by PageRank:
```

```r
kable(top_by_pagerank[1:5, c("title", "year", "pagerank", "citations", "institution", "subtopic")])
```

	title	year	pagerank	citations	institution	subtopic
166	Artificial intelligence in healthcare	2018	0.038062	2283	Harvard University	Artificial Intelligence in Healthcare and Education
30	Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: The STARD-AI Steering Group	2020	0.030402	128	Imperial College London	Artificial Intelligence in Healthcare and Education
542	Framing the challenges of artificial intelligence in medicine	2018	0.018571	212	Harvard University	Artificial Intelligence in Healthcare and Education
102	Artificial intelligence (AI) systems for interpreting complex medical datasets	2017	0.016426	696	Stanford University	Artificial Intelligence in Healthcare and Education
399	Potential Liability for Physicians Using Artificial Intelligence	2019	0.014795	478	University Of Michigan	Artificial Intelligence in Healthcare and Education

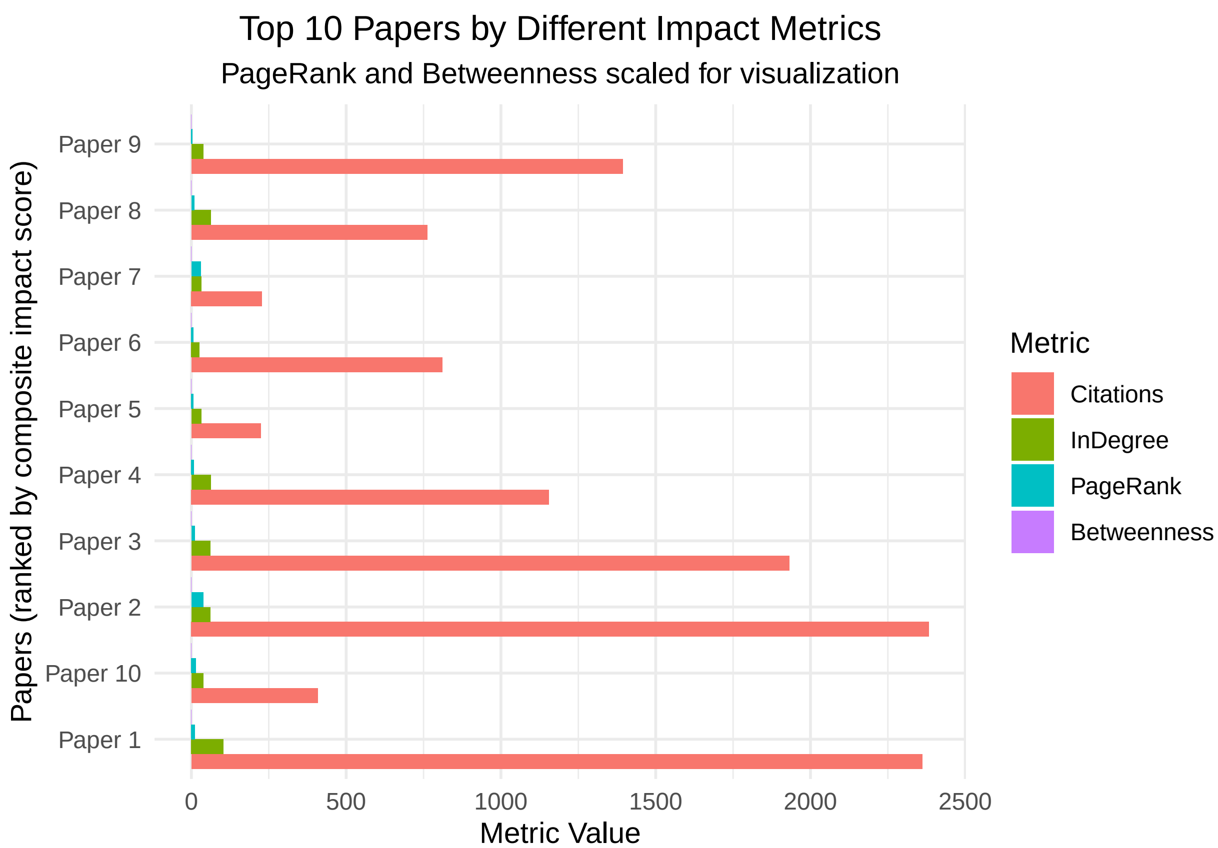
```r
cat("\nTop 5 papers by comprehensive impact score:\n")
```

```
##
## Top 5 papers by comprehensive impact score:
```

```r
kable(top_impact[1:5, c("title", "year", "impact_score", "citations", "in_degree", "institution", "subt
```

	title	year	impact_score	citations	in_degree	institution	subtopic
55	Large language models in medicine	2023	13.740089	2361	104	University Of Cambridge	Artificial Intelligence in Healthcare and Education
166	Artificial intelligence in healthcare	2018	13.559042	2383	61	Harvard University	Artificial Intelligence in Healthcare and Education
2	AI in health and medicine	2022	10.400098	1931	61	Harvard University	Artificial Intelligence in Healthcare and Education
1	Foundation models for generalist medical artificial intelligence	2023	7.953654	1155	63	Stanford University	Machine Learning in Healthcare
11	The shaky foundations of large language models and foundation models for electronic health records	2023	6.458862	2224	32	Stanford University	Machine Learning in Healthcare

```r
# Visualize impact metrics
impact_comparison <- data.frame(
  Paper = paste("Paper", 1:10),
  Title_Short = substr(top_impact$title[1:10], 1, 40),
  Citations = top_impact$citations[1:10],
  InDegree = top_impact$in_degree[1:10],
  PageRank = top_impact$pagerank[1:10] * 1000, # scaled for visualization
  Betweenness = top_impact$betweenness[1:10] * 100 # scaled for visualization
)

impact_melted <- reshape2::melt(impact_comparison[, c("Paper", "Citations", "InDegree", "PageRank", "Be
                                id.vars = "Paper")

p1 <- ggplot(impact_melted, aes(x = Paper, y = value, fill = variable)) +
  geom_bar(stat = "identity", position = "dodge") +
  coord_flip() +
  labs(title = "Top 10 Papers by Different Impact Metrics",
       subtitle = "PageRank and Betweenness scaled for visualization",
       x = "Papers (ranked by composite impact score)",
       y = "Metric Value",
       fill = "Metric") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5))

ggsave("plots/impact_metrics_comparison.png", plot = p1, width = 12, height = 8, dpi = 150)
print(p1)
```



2. Which is the oldest paper that's still very relevant until now?

```
cat("\n=== Oldest But Still Relevant Papers ===\n\n")
```

```
##
## === Oldest But Still Relevant Papers ===
```

```
# Filter older papers (e.g., before 1995) that still have high impact
old_papers <- nodes_analysis[nodes_analysis$year <= 1995 & !is.na(nodes_analysis$year), ]

if(nrow(old_papers) > 0) {
  # Sort by impact metrics
  old_papers_by_citations <- old_papers[order(old_papers$citations, decreasing = TRUE), ]
  old_papers_by_indegree <- old_papers[order(old_papers$in_degree, decreasing = TRUE), ]
  old_papers_by_pagerank <- old_papers[order(old_papers$pagerank, decreasing = TRUE), ]

  cat("High-impact papers published in 1995 or earlier (by citation count):\n")
  if(nrow(old_papers_by_citations) > 0) {
    kable(head(old_papers_by_citations[, c("title", "year", "citations", "in_degree", "institution", "s
  }

  # Visualize relationship between age and impact
```

```r
  yearly_impact <- nodes_analysis %>%
    filter(!is.na(year), year >= 1980, year <= 2020) %>%
    group_by(year) %>%
    summarise(
      avg_citations = mean(citations, na.rm = TRUE),
      avg_in_degree = mean(in_degree, na.rm = TRUE),
      paper_count = n(),
      max_citations = max(citations, na.rm = TRUE)
    )

  p2 <- ggplot(yearly_impact, aes(x = year)) +
    geom_line(aes(y = avg_citations, color = "Average Citations"), size = 1) +
    geom_line(aes(y = max_citations/10, color = "Max Citations (scaled)"), size = 1) +
    geom_bar(aes(y = paper_count), alpha = 0.3, stat = "identity") +
    scale_y_continuous(
      name = "Citations / Paper Count",
      sec.axis = sec_axis(~.*10, name = "Max Citations")
    ) +
    labs(title = "Research Impact Over Time",
         subtitle = "Average citations, maximum citations, and paper count by year",
         x = "Year",
         color = "Metric") +
    theme_minimal() +
    theme(plot.title = element_text(hjust = 0.5),
          plot.subtitle = element_text(hjust = 0.5))

  ggsave("plots/impact_over_time.png", plot = p2, width = 12, height = 6, dpi = 150)
  print(p2)
} else {
  cat("No papers found from 1995 or earlier\n")
}
```

```
## No papers found from 1995 or earlier
```

```r
# Examine earliest high-impact papers
earliest_high_impact <- nodes_analysis[nodes_analysis$citations >= quantile(nodes_analysis$citations, 0
                                      nodes_analysis$in_degree >= quantile(nodes_analysis$in_degree, 0.
earliest_high_impact <- earliest_high_impact[order(earliest_high_impact$year), ]

cat("\nEarliest high-impact papers (citations or in-degree in top 10%):\n")
```

```
##
## Earliest high-impact papers (citations or in-degree in top 10%):
```

```r
kable(head(earliest_high_impact[, c("title", "year", "citations", "in_degree", "institution", "subtopic
```

	title	year	citations	n_depts	institution	subtopic
7	How to Train good Word Embeddings for Biomedical NLP	2016	351	1	University Of Cambridge	Natural Language Processing Techniques
414	What This Computer Needs Is a Physician	2017	414	13	Stanford University	Artificial Intelligence in Healthcare and Education
521	A Report on the 2017 Native Language Identification Shared Task	2017	158	1	Macquarie University	Natural Language Processing Techniques
594	Findings of the VarDial Evaluation Campaign 2017	2017	144	3	University Of Cologne	Natural Language Processing Techniques
19	With an eye to AI and autonomous diagnosis	2018	267	17	Moorfields Eye Hospital Nhs Foundation Trust	Machine Learning in Healthcare
26	Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (AI) and autonomous robotic surgery	2018	371	8	Hospital Das Clínicas Da Faculdade De Medicina Da Universidade De São Paulo	Artificial Intelligence in Healthcare and Education
166	Artificial intelligence in healthcare	2018	2383	61	Harvard University	Artificial Intelligence in Healthcare and Education
245	Artificial Intelligence in Surgery: Promises and Perils	2018	1145	27	Massachusetts General Hospital	Artificial Intelligence in Healthcare and Education
351	Artificial intelligence-enabled healthcare delivery	2018	523	8	Deakin University	Artificial Intelligence in Healthcare and Education
390	Artificial intelligence powers digital medicine	2018	436	8	Stanford University	Artificial Intelligence in Healthcare and Education

3. What subtopics have the highest concentration of research?

```
cat("\n=== Research Subtopic Concentration Analysis ===\n\n")
```

```
##
## === Research Subtopic Concentration Analysis ===
```

```r
# Analyze distribution of research subtopics
subtopic_stats <- nodes_all %>%
  group_by(subtopic) %>%
  summarise(
    paper_count = n(),
    avg_citations = mean(citations, na.rm = TRUE),
    total_citations = sum(citations, na.rm = TRUE),
    unique_institutions = n_distinct(institution, na.rm = TRUE),
    year_span = ifelse(all(is.na(year)), 0, max(year, na.rm = TRUE) - min(year, na.rm = TRUE)),
    latest_year = ifelse(all(is.na(year)), NA, max(year, na.rm = TRUE))
  ) %>%
  arrange(desc(paper_count))

# Calculate concentration index (Herfindahl-Hirschman Index)
total_papers <- nrow(nodes_all)
subtopic_stats$concentration_index <- (subtopic_stats$paper_count / total_papers)^2

cat("Research subtopics ranked by paper count:\n")
```

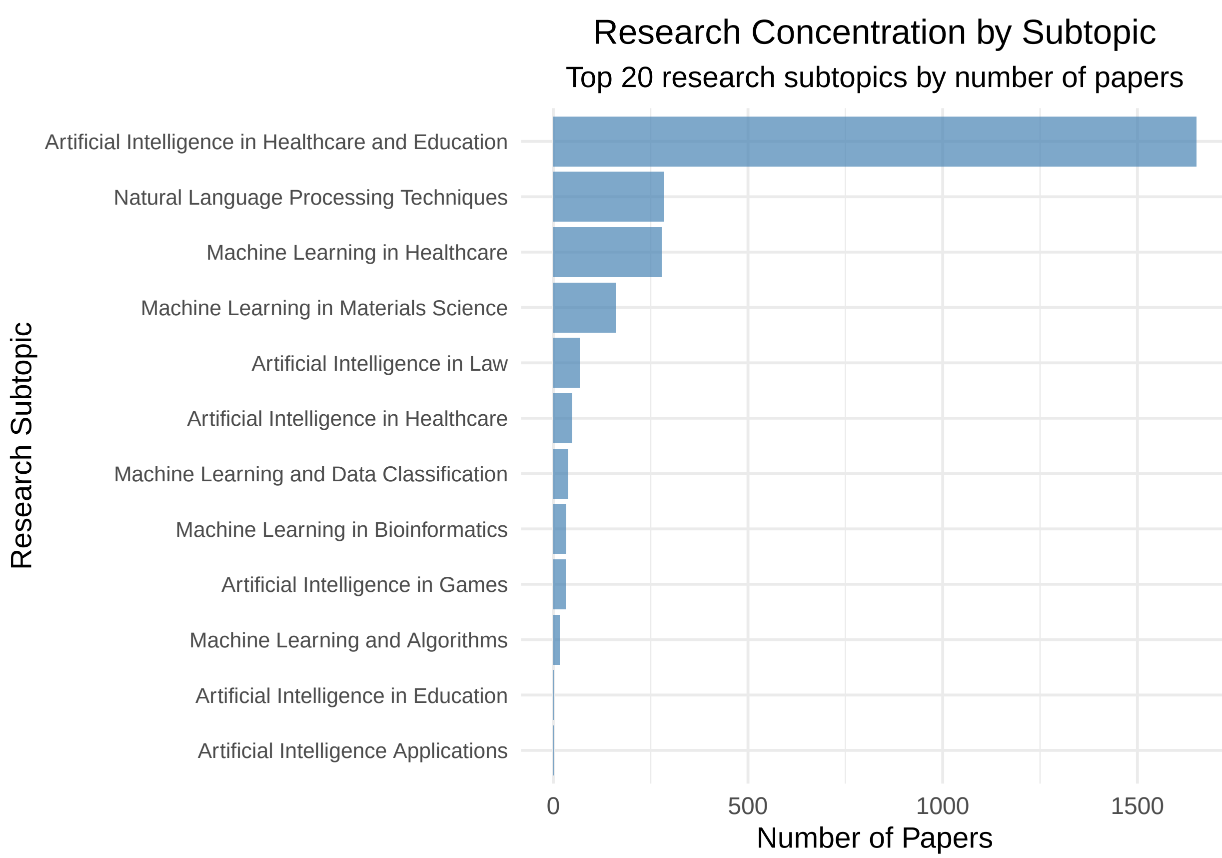
```
## Research subtopics ranked by paper count:
```

```r
kable(head(subtopic_stats, 15))
```

subtopic	paper_count	avg_citations	total_citations	unique_institutions	year_span	latest_year	concentration_index
Artificial Intelligence in Healthcare and Education	1651	41.11326	67878	544	9	2025	0.4001411
Natural Language Processing Techniques	284	18.12324	5147	112	10	2025	0.0118401
Machine Learning in Healthcare	278	30.91007	8593	119	10	2025	0.0113451
Machine Learning in Materials Science	161	34.04348	5481	69	8	2025	0.0038051
Artificial Intelligence in Law	67	17.35821	1163	32	7	2025	0.0006590
Artificial Intelligence in Healthcare	48	10.06250	483	44	9	2025	0.0003382
Machine Learning and Data Classification	38	11.42105	434	20	7	2025	0.0002120
Machine Learning in Bioinformatics	33	22.75758	751	24	5	2025	0.0001599
Artificial Intelligence in Games	32	21.21875	679	22	9	2025	0.0001503
Machine Learning and Algorithms	16	4.62500	74	14	7	2025	0.0000376
Artificial Intelligence Applications	1	56.00000	56	1	0	2019	0.0000001
Artificial Intelligence in Education	1	0.00000	0	1	0	2025	0.0000001

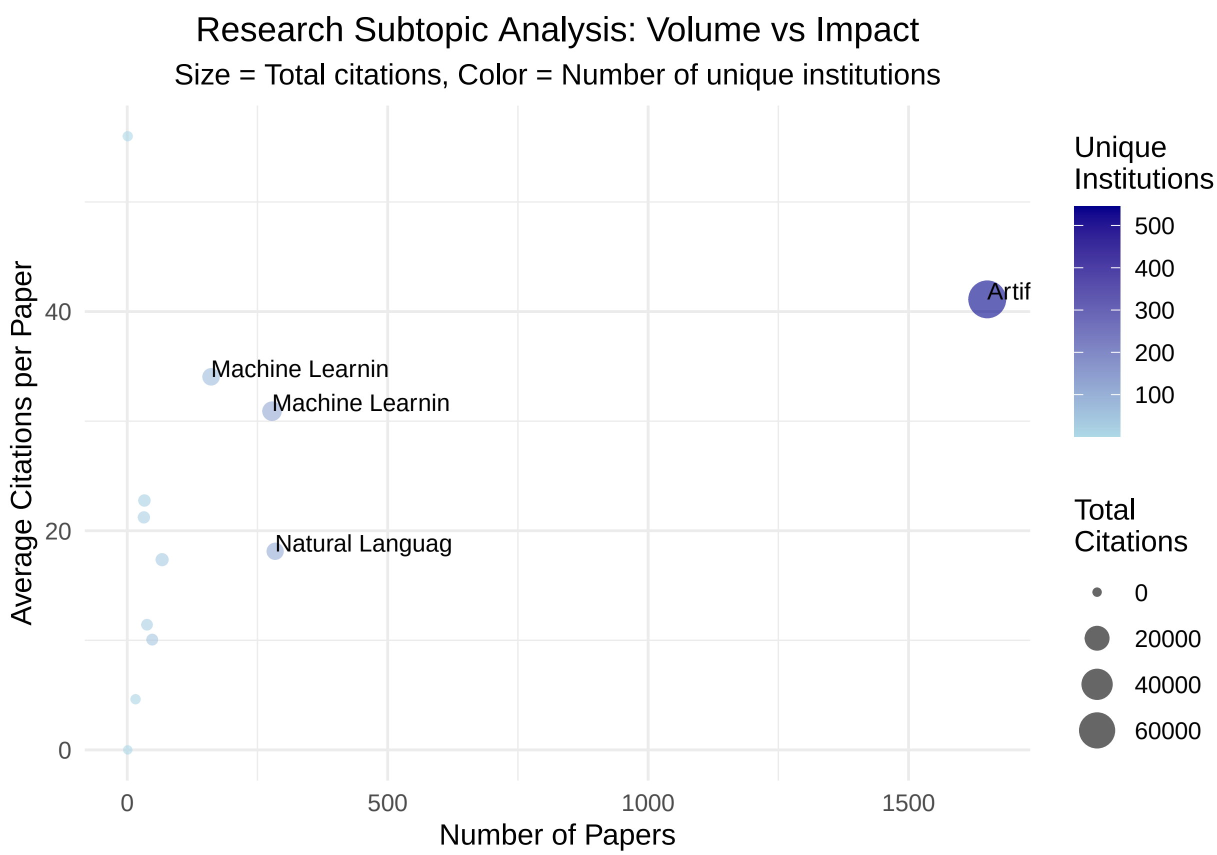
```r
# Visualize research concentration
p3 <- ggplot(head(subtopic_stats, 20), aes(x = reorder(subtopic, paper_count), y = paper_count)) +
  geom_bar(stat = "identity", fill = "steelblue", alpha = 0.7) +
```

```r
  coord_flip() +
  labs(title = "Research Concentration by Subtopic",
       subtitle = "Top 20 research subtopics by number of papers",
       x = "Research Subtopic",
       y = "Number of Papers") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5),
        axis.text.y = element_text(size = 8))

ggsave("plots/research_concentration_by_subtopic.png", plot = p3, width = 12, height = 8, dpi = 150)
print(p3)
```




```r
# Impact vs volume scatter plot
p4 <- ggplot(subtopic_stats, aes(x = paper_count, y = avg_citations)) +
  geom_point(aes(size = total_citations, color = unique_institutions), alpha = 0.6) +
  geom_text(aes(label = ifelse(paper_count > 100 | avg_citations > quantile(avg_citations, 0.9, na.rm =
                               substr(subtopic, 1, 15), ""),
            hjust = 0, vjust = 0, size = 3, check_overlap = TRUE) +
  scale_color_gradient(low = "lightblue", high = "darkblue", name = "Unique\nInstitutions") +
  scale_size_continuous(name = "Total\nCitations") +
  labs(title = "Research Subtopic Analysis: Volume vs Impact",
       subtitle = "Size = Total citations, Color = Number of unique institutions",
       x = "Number of Papers",
       y = "Average Citations per Paper") +
```

```r
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5))

ggsave("plots/subtopic_volume_vs_impact.png", plot = p4, width = 12, height = 8, dpi = 150)
print(p4)
```




```r
# Word cloud showing main research areas
if(nrow(subtopic_stats) > 0) {
  png("plots/research_subtopics_wordcloud.png", width = 800, height = 600, res = 150)
  wordcloud(words = subtopic_stats$subtopic,
            freq = subtopic_stats$paper_count,
            min.freq = 1,
            max.words = 100,
            random.order = FALSE,
            rot.perc = 0.35,
            colors = brewer.pal(8, "Dark2"))
  dev.off()
}
```

```
## pdf
##   2
```

4. Which country/institution has the most research output?

```r
cat("\n=== Research Output Analysis ===\n\n")
```

```
##
## === Research Output Analysis ===
```

```r
# Extract country information (assuming it can be inferred from institution names)
# This needs to be adjusted based on actual data format
nodes_all$country <- NA  # needs to be filled based on actual situation

# Analyze by institution
institution_stats <- nodes_all %>%
  filter(!is.na(institution) & institution != "") %>%
  group_by(institution) %>%
  summarise(
    paper_count = n(),
    avg_citations = mean(citations, na.rm = TRUE),
    total_citations = sum(citations, na.rm = TRUE),
    latest_year = ifelse(all(is.na(year)), NA, max(year, na.rm = TRUE)),
    research_areas = n_distinct(subtopic, na.rm = TRUE)
  ) %>%
  arrange(desc(paper_count))

cat("Institutions with highest research output (top 20):\n")
```

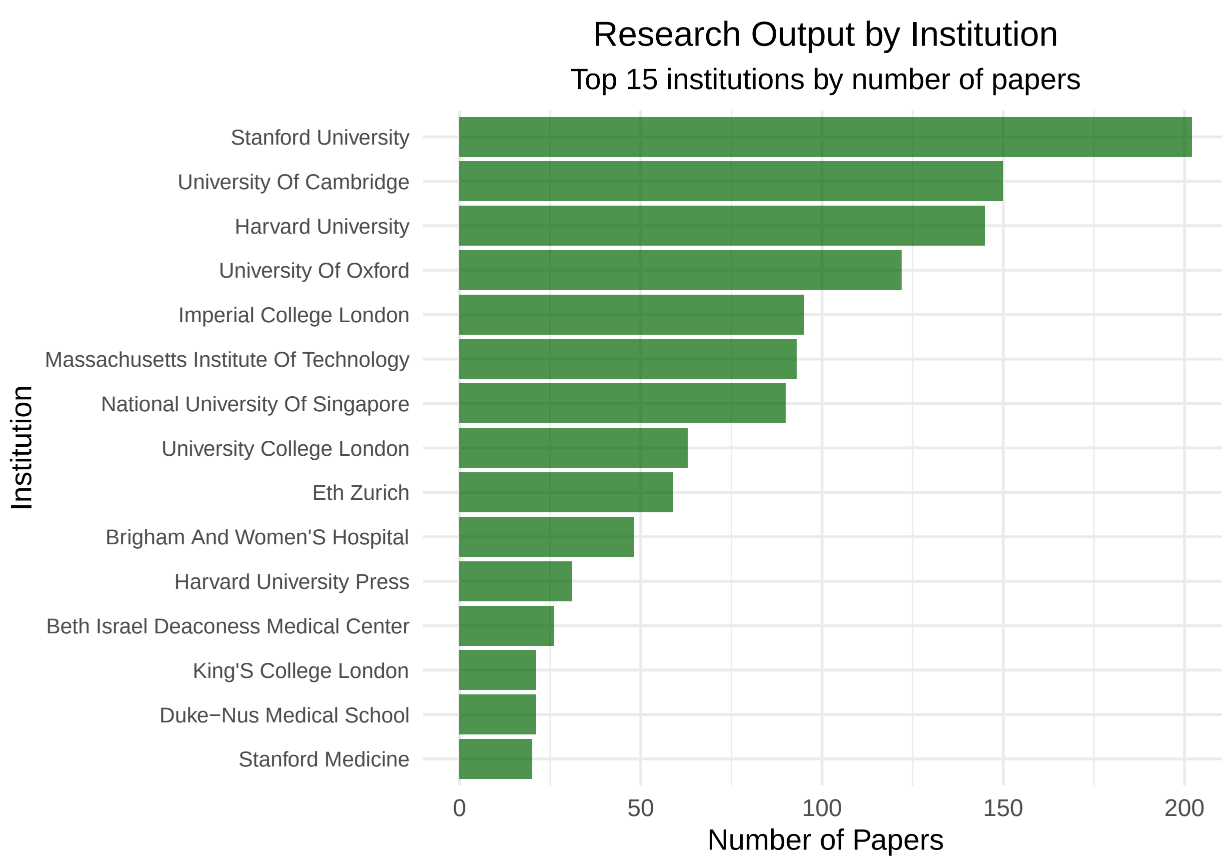
```
## Institutions with highest research output (top 20):
```

```r
kable(head(institution_stats, 20))
```

institution	paper_count	avg_citations	total_citations	latest_year	research_areas
Stanford University	202	39.20297	7919	2025	9
University Of Cambridge	150	38.24667	5737	2025	10
Harvard University	145	64.82069	9399	2025	7
University Of Oxford	122	35.02459	4273	2025	11
Imperial College London	95	26.20000	2489	2025	8
Massachusetts Institute Of Technology	93	30.70968	2856	2025	9
National University Of Singapore	90	29.56667	2661	2025	9
University College London	63	12.09524	762	2025	9
Eth Zurich	59	39.28814	2318	2025	9
Brigham And Women'S Hospital	48	28.37500	1362	2025	3
Harvard University Press	31	30.58065	948	2025	7
Beth Israel Deaconess Medical Center	26	26.80769	697	2025	2
Duke-Nus Medical School	21	30.19048	634	2025	2
King'S College London	21	22.09524	464	2025	4
Stanford Medicine	20	50.10000	1002	2025	4
University Of Toronto	14	109.92857	1539	2025	2

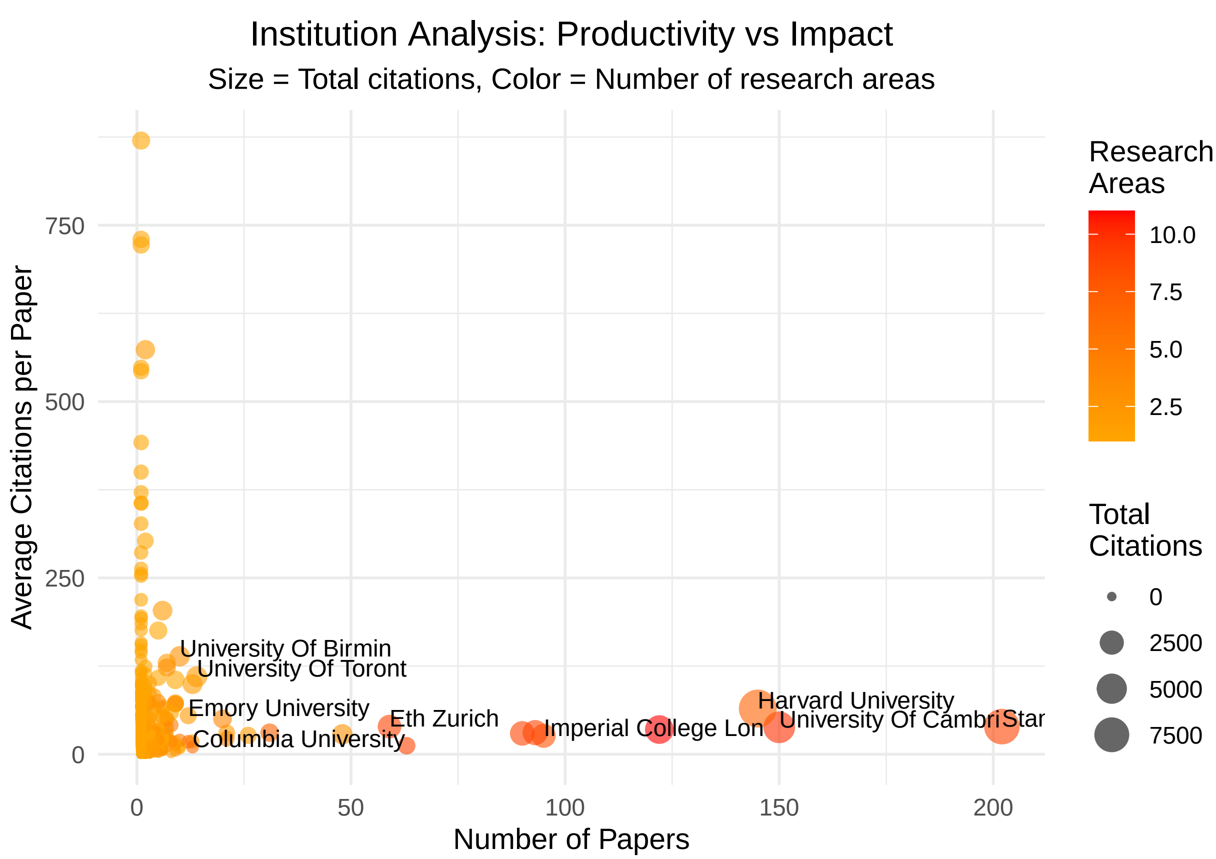
institution	paper_count	avg_citations	total_citations	latest_year	research_areas
Boston Children'S Hospital	13	18.30769	238	2025	3
Columbia University	13	10.23077	133	2025	6
Massachusetts General Hospital	13	99.46154	1293	2025	2
Emory University	12	54.58333	655	2025	2

```r
# Visualize institutional output
p5 <- ggplot(head(institution_stats, 15), aes(x = reorder(institution, paper_count), y = paper_count)) +
  geom_bar(stat = "identity", fill = "darkgreen", alpha = 0.7) +
  coord_flip() +
  labs(title = "Research Output by Institution",
       subtitle = "Top 15 institutions by number of papers",
       x = "Institution",
       y = "Number of Papers") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5),
        axis.text.y = element_text(size = 8))

ggsave("plots/research_output_by_institution.png", plot = p5, width = 12, height = 8, dpi = 150)
print(p5)
```




```r
# Institutional impact analysis
p6 <- ggplot(institution_stats, aes(x = paper_count, y = avg_citations)) +
```

```r
  geom_point(aes(size = total_citations, color = research_areas), alpha = 0.6) +
  geom_text(aes(label = ifelse(paper_count > quantile(paper_count, 0.95, na.rm = TRUE) |
                               avg_citations > quantile(avg_citations, 0.95, na.rm = TRUE),
                               substr(institution, 1, 20), "")),
            hjust = 0, vjust = 0, size = 3, check_overlap = TRUE) +
  scale_color_gradient(low = "orange", high = "red", name = "Research\nAreas") +
  scale_size_continuous(name = "Total\nCitations") +
  labs(title = "Institution Analysis: Productivity vs Impact",
       subtitle = "Size = Total citations, Color = Number of research areas",
       x = "Number of Papers",
       y = "Average Citations per Paper") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5))

ggsave("plots/institution_productivity_vs_impact.png", plot = p6, width = 12, height = 8, dpi = 150)
print(p6)
```



5. Where is the newer research headed?

```r
cat("\n=== Research Trend Analysis ===\n\n")
```

```
##
```

```
## === Research Trend Analysis ===
```

```r
# Analyze recent research trends
recent_years <- nodes_all %>%
  filter(!is.na(year), year >= 2015) %>%
  group_by(year, subtopic) %>%
  summarise(paper_count = n(), .groups = 'drop') %>%
  group_by(year) %>%
  mutate(year_total = sum(paper_count),
         percentage = paper_count / year_total * 100) %>%
  ungroup()

# Find fastest growing research areas
growth_analysis <- recent_years %>%
  group_by(subtopic) %>%
  summarise(
    total_papers = sum(paper_count),
    years_active = n_distinct(year),
    latest_year_papers = sum(paper_count[year == max(year)]),
    earliest_year_papers = sum(paper_count[year == min(year)]),
    .groups = 'drop'
  ) %>%
  filter(years_active >= 3, total_papers >= 10) %>%
  mutate(growth_rate = (latest_year_papers - earliest_year_papers) / earliest_year_papers * 100) %>%
  arrange(desc(growth_rate))

cat("Fastest growing research areas (since 2015):\n")
```

```
## Fastest growing research areas (since 2015):
```

```r
kable(head(growth_analysis[growth_analysis$growth_rate > 0, ], 15))
```

subtopic	total_papers	years_active	latest_year_papers	earliest_year_papers	growth_rate
Artificial Intelligence in Healthcare and Education	1651	10	402	2	20000
Machine Learning in Healthcare	278	9	54	1	5300
Machine Learning in Materials Science	161	9	28	1	2700
Artificial Intelligence in Healthcare	48	8	14	1	1300
Natural Language Processing Techniques	284	11	30	4	650
Artificial Intelligence in Law	67	8	8	2	300
Machine Learning in Bioinformatics	33	6	7	2	250
Machine Learning and Algorithms	16	7	2	1	100
Artificial Intelligence in Games	32	9	3	2	50

```r
# Emerging research areas (appeared only in recent years)
emerging_topics <- nodes_all %>%
  filter(!is.na(year)) %>%
  group_by(subtopic) %>%
  summarise(
```

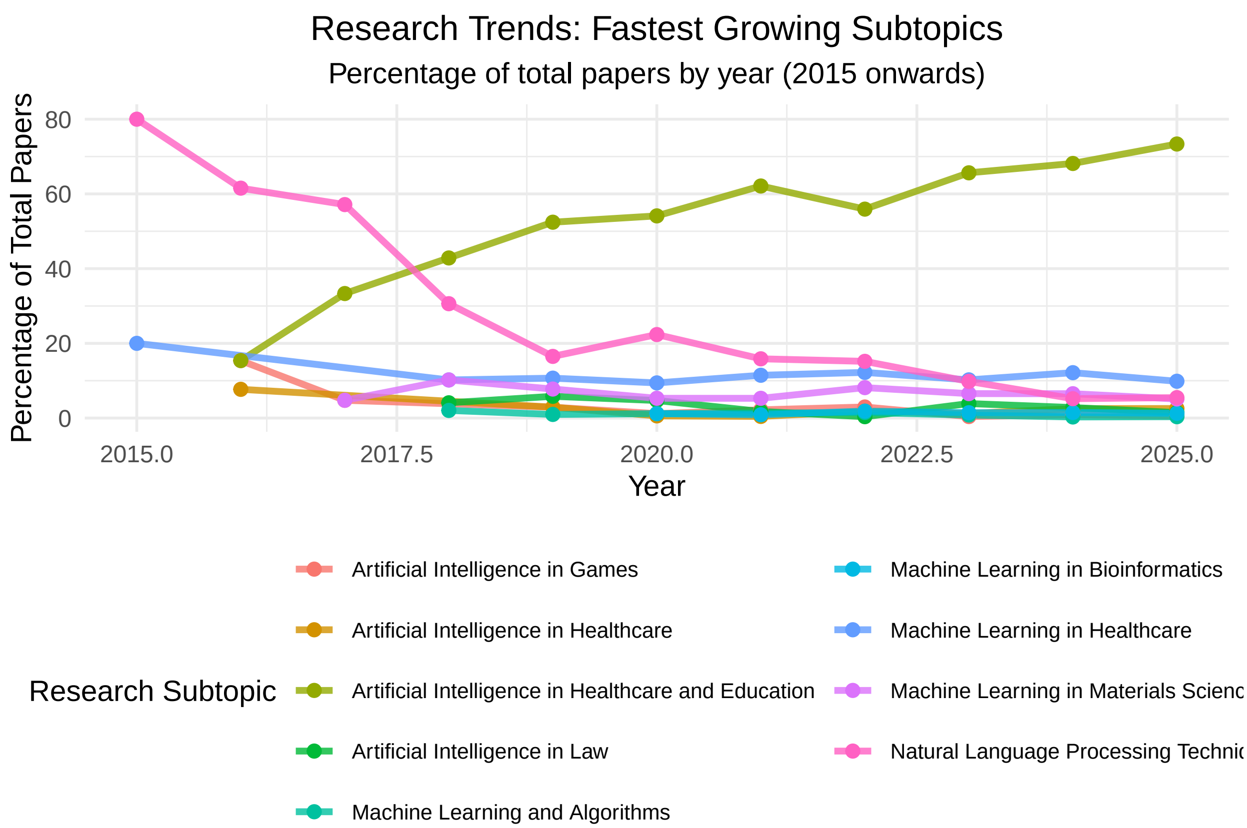
```r
    first_appearance = min(year, na.rm = TRUE),
    paper_count = n(),
    avg_citations = mean(citations, na.rm = TRUE),
    .groups = 'drop'
  ) %>%
  filter(first_appearance >= 2018, paper_count >= 5) %>%
  arrange(desc(paper_count))

cat("\nEmerging research areas (first appeared in 2018 or later with significant scale):\n")
```

```
##
## Emerging research areas (first appeared in 2018 or later with significant scale):
```

```r
kable(emerging_topics)
```

subtopic	first_appearance	paper_count	avg_citations
Artificial Intelligence in Law	2018	67	17.35821
Machine Learning and Data Classification	2018	38	11.42105
Machine Learning in Bioinformatics	2020	33	22.75758
Machine Learning and Algorithms	2018	16	4.62500

```r
# Visualize research trends
trending_topics <- head(growth_analysis[growth_analysis$growth_rate > 0, ], 10)$subtopic

trend_data <- recent_years %>%
  filter(subtopic %in% trending_topics) %>%
  group_by(year) %>%
  mutate(year_total = sum(paper_count)) %>%
  ungroup() %>%
  mutate(percentage = paper_count / year_total * 100)

p7 <- ggplot(trend_data, aes(x = year, y = percentage, color = subtopic)) +
  geom_line(size = 1.2, alpha = 0.8) +
  geom_point(size = 2) +
  labs(title = "Research Trends: Fastest Growing Subtopics",
       subtitle = "Percentage of total papers by year (2015 onwards)",
       x = "Year",
       y = "Percentage of Total Papers",
       color = "Research Subtopic") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5),
        legend.position = "bottom",
        legend.text = element_text(size = 8)) +
  guides(color = guide_legend(ncol = 2))

ggsave("plots/research_trends_over_time.png", plot = p7, width = 14, height = 10, dpi = 150)
print(p7)
```

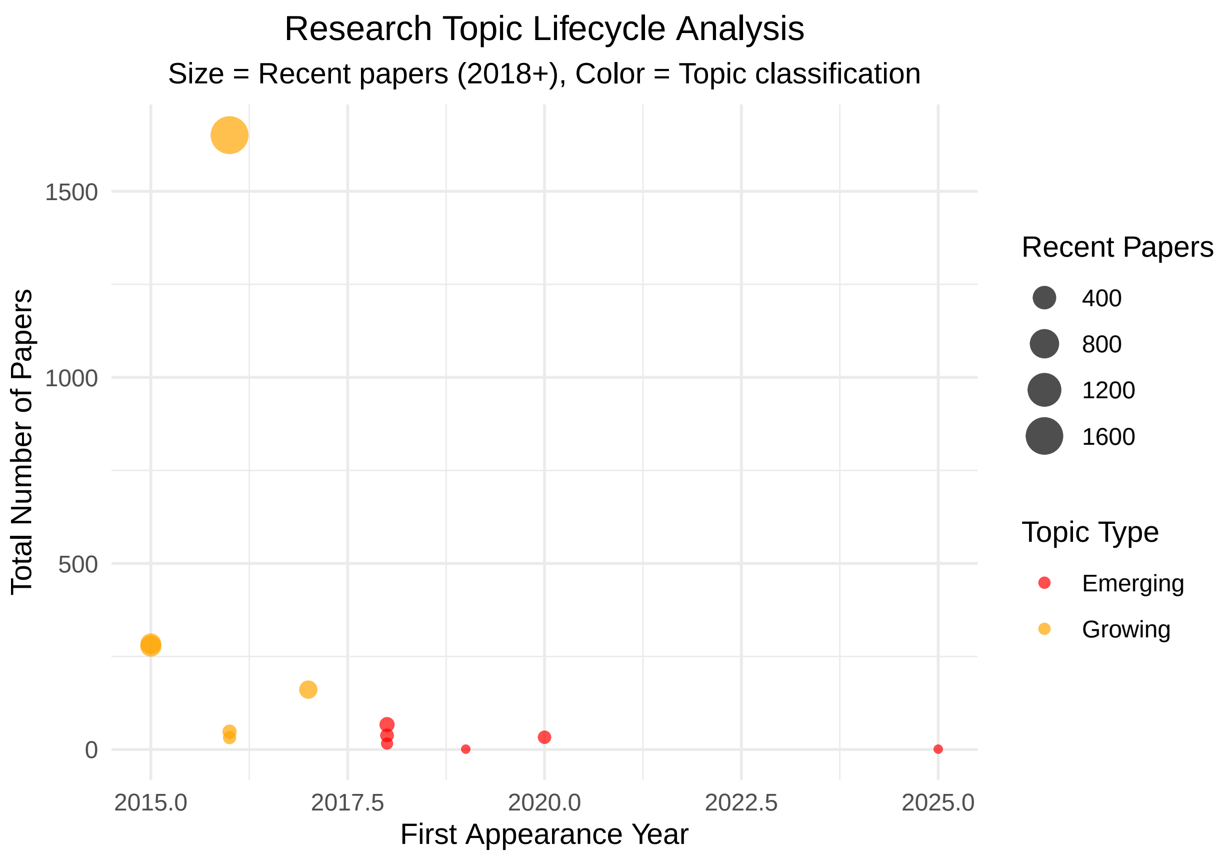



```r
# Annual research topic word cloud (last 5 years)
recent_subtopics <- nodes_all %>%
  filter(!is.na(year), year >= 2019) %>%
  count(subtopic, sort = TRUE) %>%
  filter(!is.na(subtopic), subtopic != "")

if(nrow(recent_subtopics) > 0) {
  png("plots/recent_research_trends_wordcloud.png", width = 800, height = 600, res = 150)
  wordcloud(words = recent_subtopics$subtopic,
            freq = recent_subtopics$n,
            min.freq = 1,
            max.words = 80,
            random.order = FALSE,
            rot.perc = 0.35,
            colors = brewer.pal(8, "Set2"))
  dev.off()
}
```

```
## pdf
##   2
```

```r
# Popular vs emerging topics comparison
topic_classification <- nodes_all %>%
  filter(!is.na(year), !is.na(subtopic)) %>%
  group_by(subtopic) %>%
```

```r
  summarise(
    first_year = min(year),
    last_year = max(year),
    total_papers = n(),
    recent_papers = sum(year >= 2018),
    .groups = 'drop'
  ) %>%
  mutate(
    topic_type = case_when(
      first_year >= 2018 ~ "Emerging",
      recent_papers / total_papers >= 0.5 ~ "Growing",
      total_papers >= 50 ~ "Established",
      TRUE ~ "Declining"
    )
  )

p8 <- ggplot(topic_classification, aes(x = first_year, y = total_papers, color = topic_type)) +
  geom_point(aes(size = recent_papers), alpha = 0.7) +
  scale_color_manual(values = c("Emerging" = "red", "Growing" = "orange",
                                "Established" = "blue", "Declining" = "gray")) +
  labs(title = "Research Topic Lifecycle Analysis",
       subtitle = "Size = Recent papers (2018+), Color = Topic classification",
       x = "First Appearance Year",
       y = "Total Number of Papers",
       color = "Topic Type",
       size = "Recent Papers") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5))

ggsave("plots/research_topic_lifecycle.png", plot = p8, width = 12, height = 8, dpi = 150)
print(p8)
```



Summary

```r
cat("=== Analysis Summary ===\n\n")
```

```
## === Analysis Summary ===
```

```r
cat("1. Most impactful paper:\n")
```

```
## 1. Most impactful paper:
```

```r
cat("   - Highest comprehensive impact score:", top_impact$title[1], "\n")
```

```
##    - Highest comprehensive impact score: Large language models in medicine
```

```r
cat("   - Publication year:", top_impact$year[1], "\n")
```

```
##    - Publication year: 2023
```

```r
cat("   - Citation count:", top_impact$citations[1], "\n\n")
```

```
##    - Citation count: 2361
```

```r
if(nrow(earliest_high_impact) > 0) {
  cat("2. Oldest but still relevant paper:\n")
  cat("   - Earliest high-impact paper:", earliest_high_impact$title[1], "\n")
  cat("   - Publication year:", earliest_high_impact$year[1], "\n\n")
}
```

```
## 2. Oldest but still relevant paper:
##    - Earliest high-impact paper: How to Train good Word Embeddings for Biomedical NLP
##    - Publication year: 2016
```

```r
cat("3. Subtopics with highest research concentration:\n")
```

```
## 3. Subtopics with highest research concentration:
```

```r
for(i in 1:5) {
  cat("   -", subtopic_stats$subtopic[i], "(", subtopic_stats$paper_count[i], "papers)\n")
}
```

```
##    - Artificial Intelligence in Healthcare and Education ( 1651 papers)
##    - Natural Language Processing Techniques ( 284 papers)
##    - Machine Learning in Healthcare ( 278 papers)
##    - Machine Learning in Materials Science ( 161 papers)
##    - Artificial Intelligence in Law ( 67 papers)
```

```r
cat("\n4. Institutions with most research output:\n")
```

```
##
## 4. Institutions with most research output:
```

```r
for(i in 1:5) {
  cat("   -", institution_stats$institution[i], "(", institution_stats$paper_count[i], "papers)\n")
}
```

```
##    - Stanford University ( 202 papers)
##    - University Of Cambridge ( 150 papers)
##    - Harvard University ( 145 papers)
##    - University Of Oxford ( 122 papers)
##    - Imperial College London ( 95 papers)
```

```r
cat("\n5. Emerging research trends:\n")
```

```
##
## 5. Emerging research trends:
```

```r
if(nrow(emerging_topics) > 0) {
  for(i in 1:min(5, nrow(emerging_topics))) {
    cat("   -", emerging_topics$subtopic[i], "(first appeared:", emerging_topics$first_appearance[i], ")
  }
}
```

```
##    - Artificial Intelligence in Law (first appeared: 2018 )
##    - Machine Learning and Data Classification (first appeared: 2018 )
##    - Machine Learning in Bioinformatics (first appeared: 2020 )
##    - Machine Learning and Algorithms (first appeared: 2018 )
```

```r
cat("\nAll analysis charts have been saved to the plots/ folder.\n")
```

```
##
## All analysis charts have been saved to the plots/ folder.
```