Systematic Review Automation with Information Retrieval

Harry Scells

Leipzig University

MANILA24 · 18.07.2024

Systematic Reviews

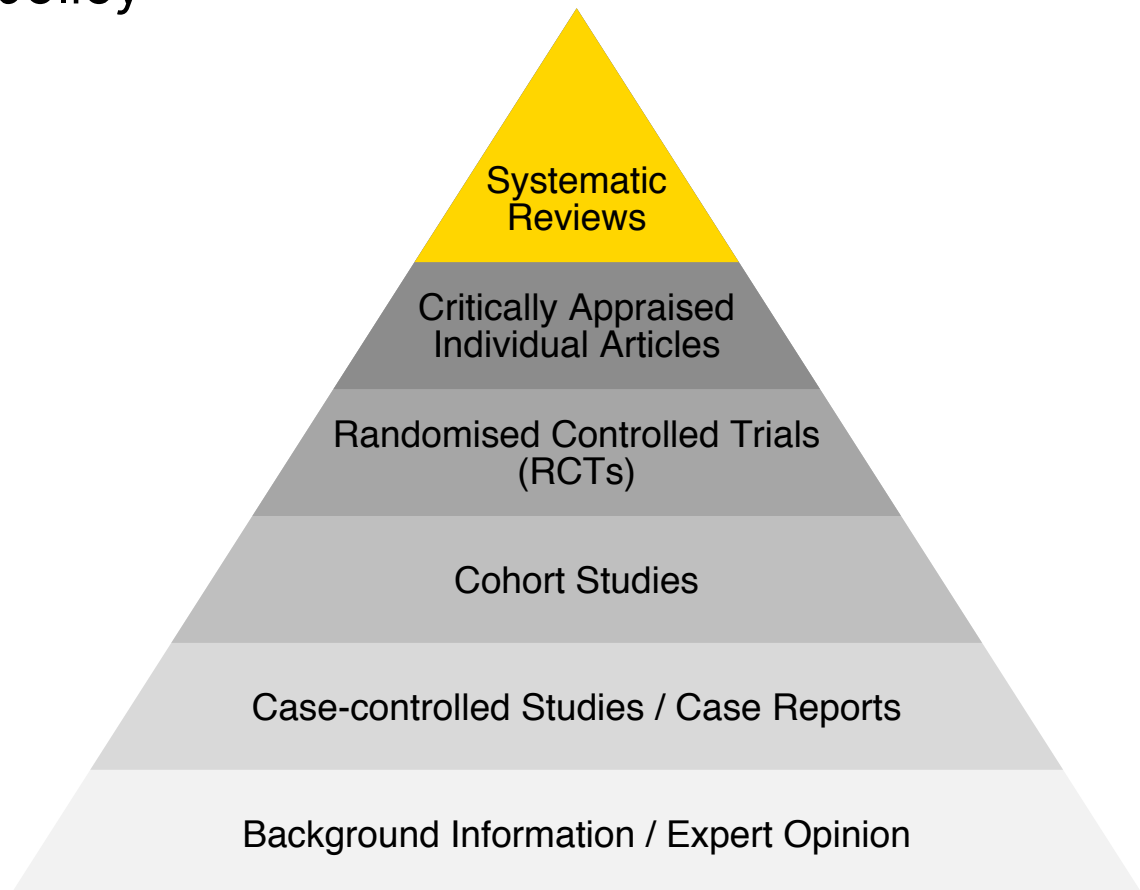
Automation

Summary

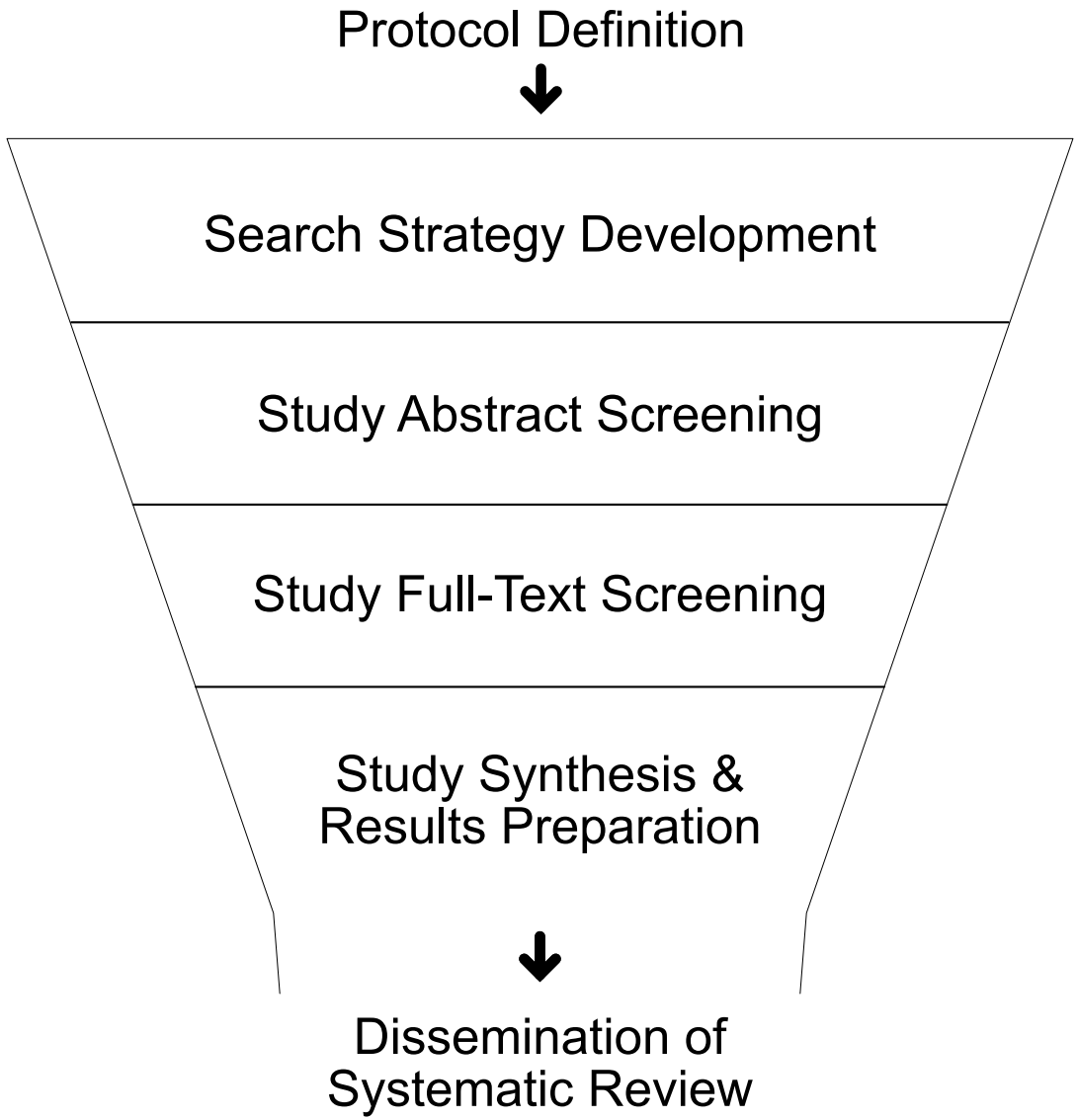
Overview

In medicine, systematic reviews:

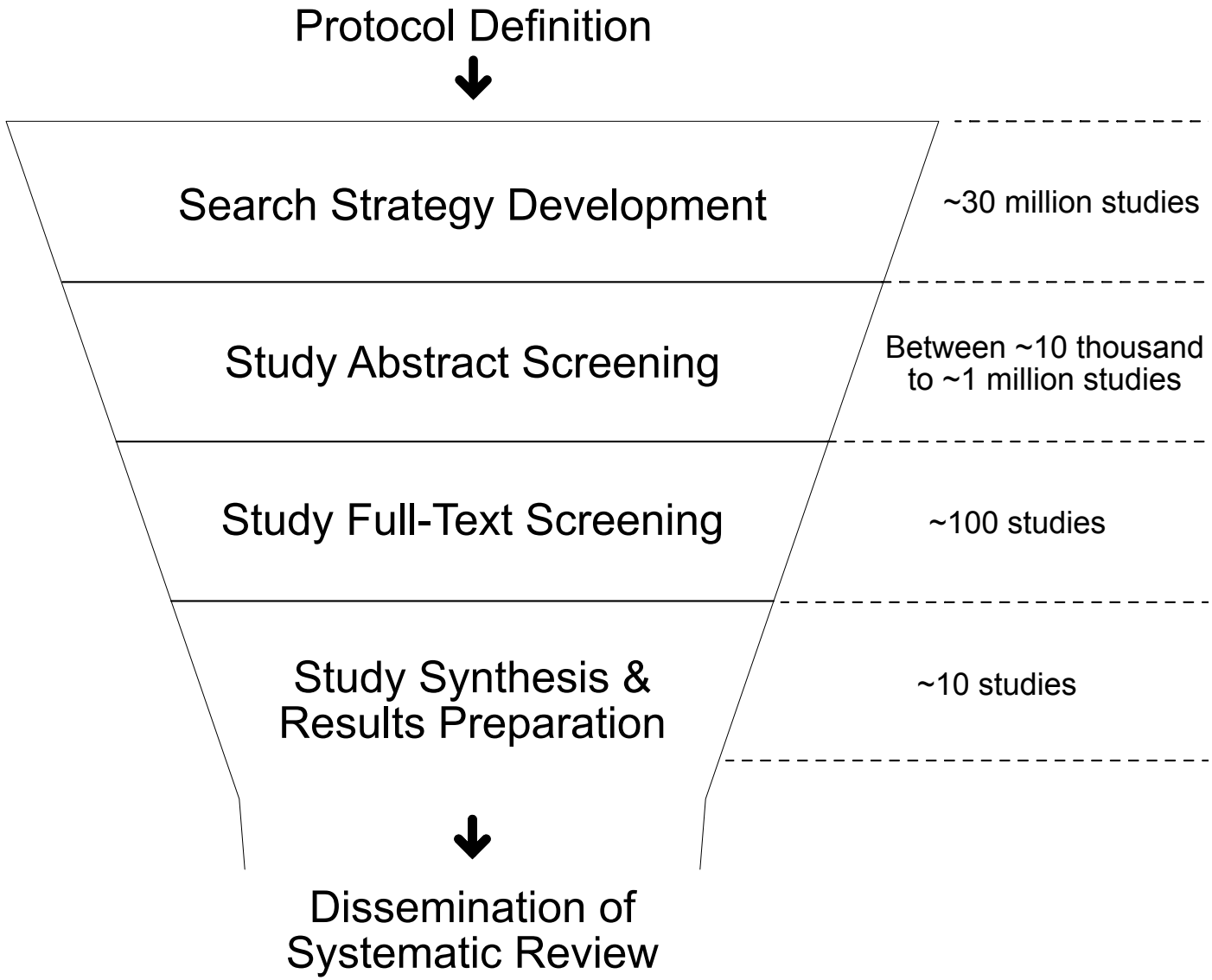
- **Guide** clinical decisions
- **Inform** practice and policy
- **Provide** evidence



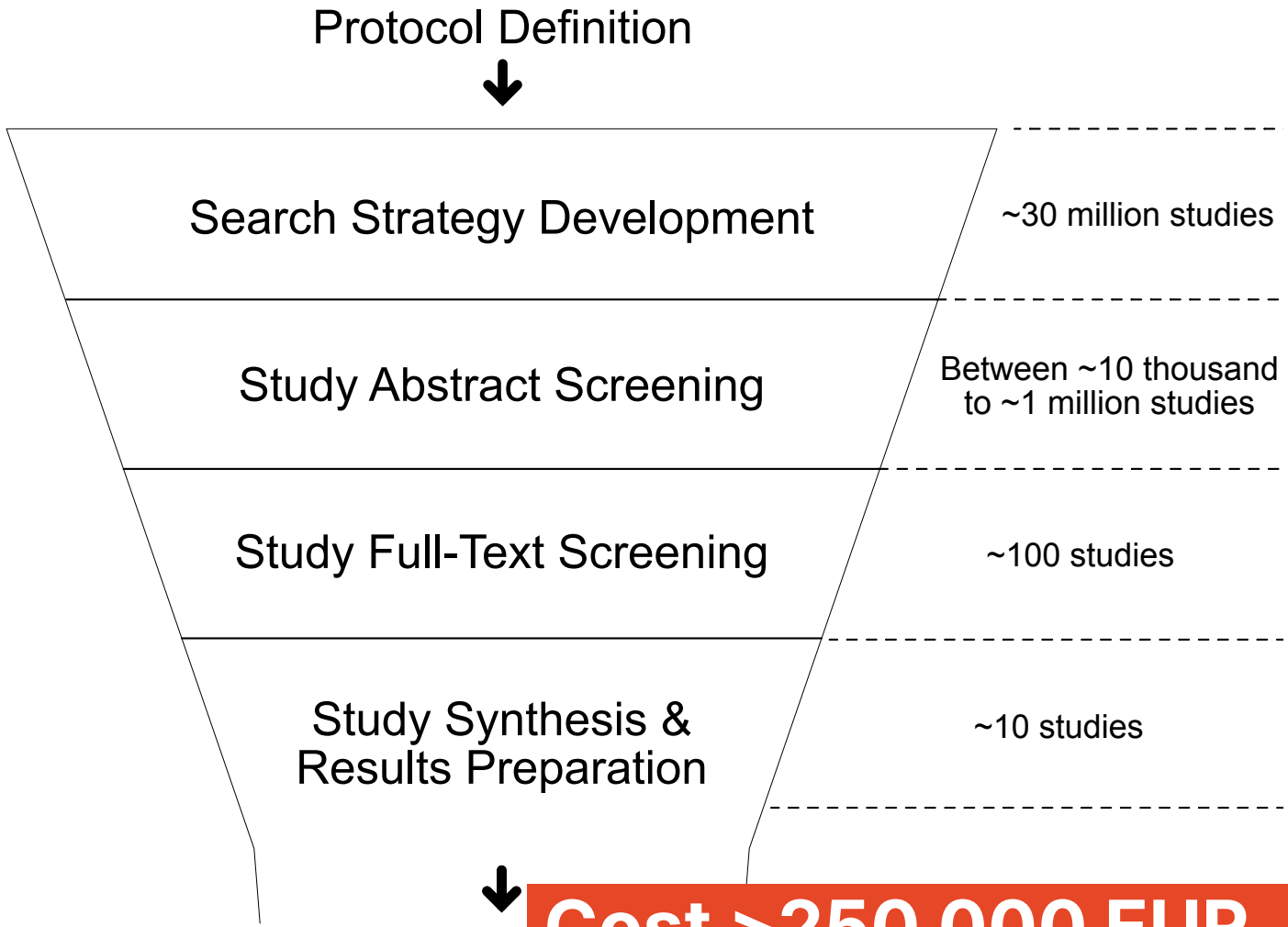
Systematic review creation is hard!



Why is systematic review creation hard?



Why is systematic review creation hard?



Systematic Reviews

Automation

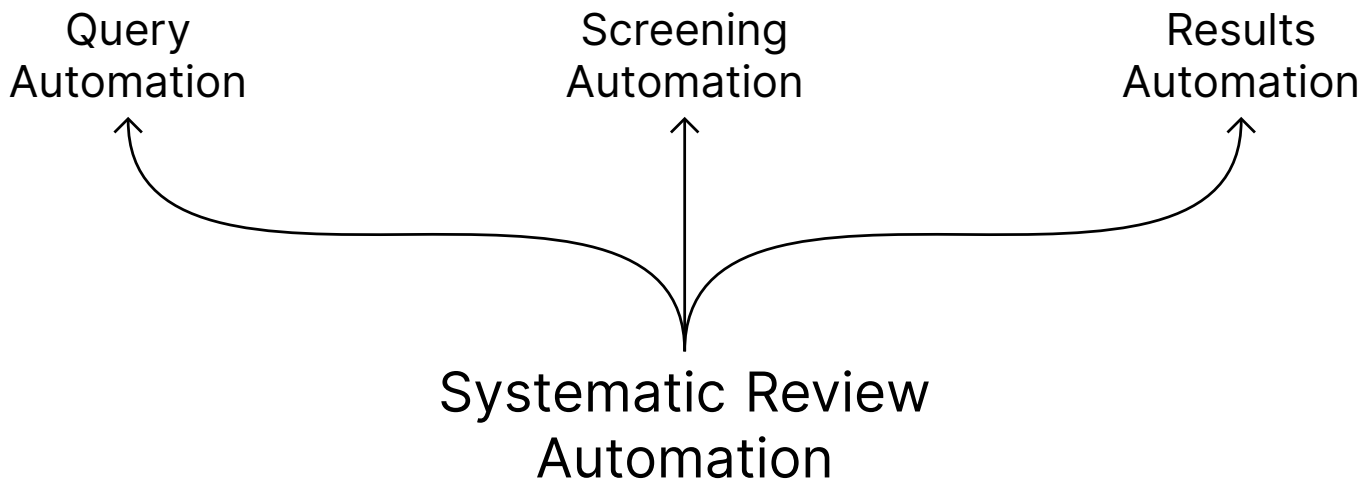
- Query Automation
- Screening Automation
- Results Automation

Summary

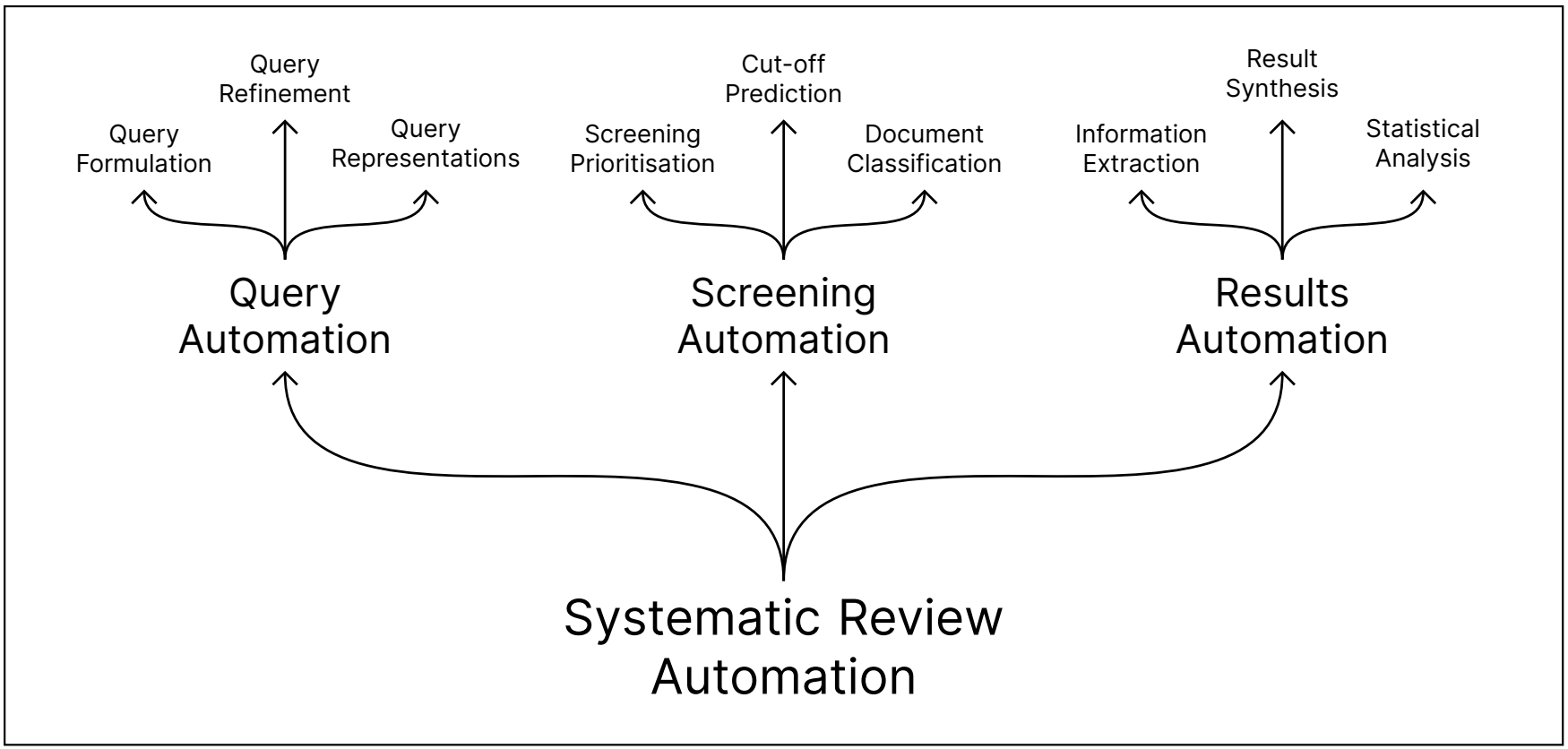
Overview

Systematic Review
Automation

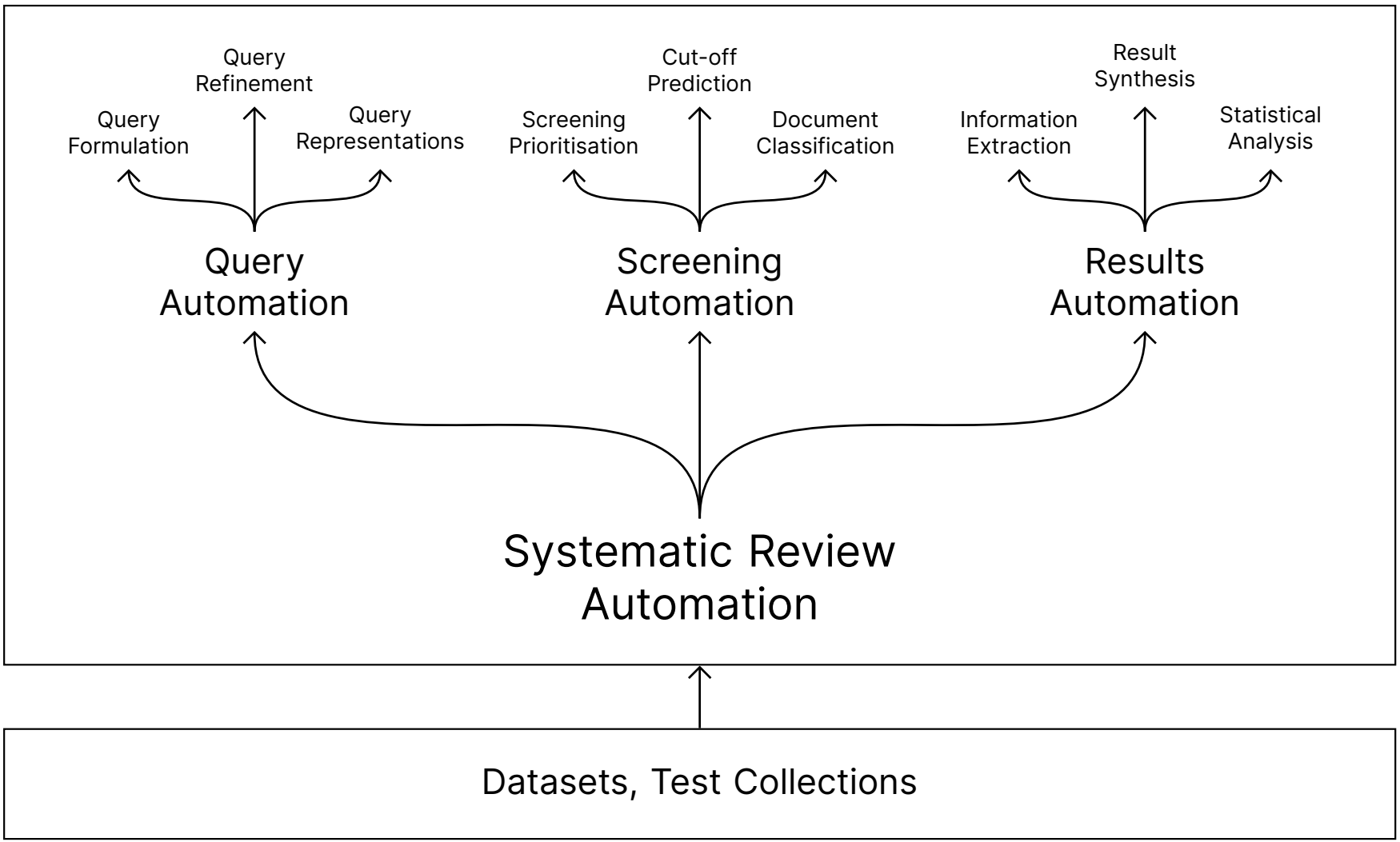
Overview


Query Automation
Screening Automation
Results Automation
Systematic Review Automation

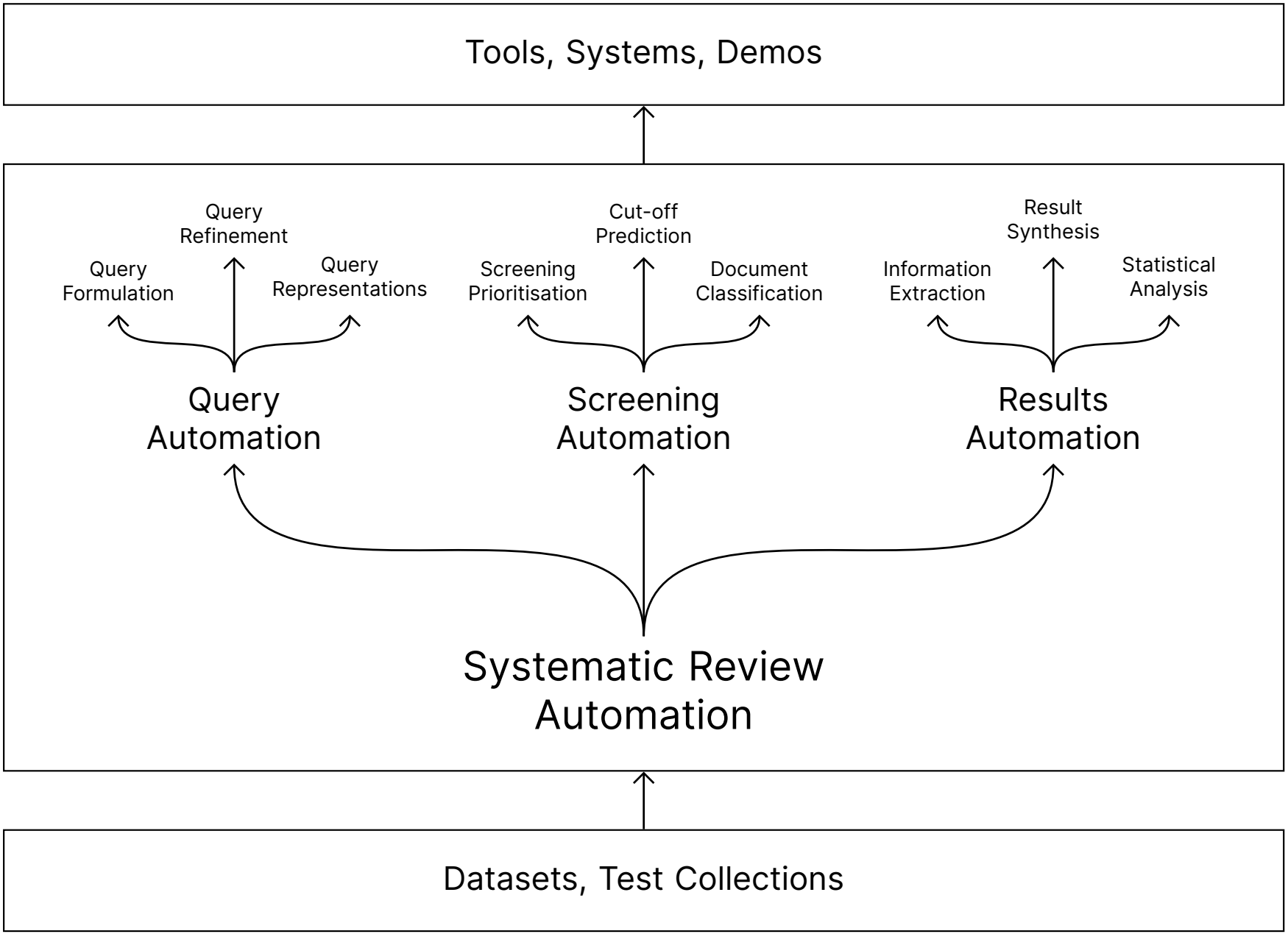
Overview


Query Refinement
Query Formulation
Query Representations
Cut-off Prediction
Screening Prioritisation
Document Classification
Result Synthesis
Information Extraction
Statistical Analysis
Query Automation
Screening Automation
Results Automation
Systematic Review Automation

Overview


Query Refinement
Query Formulation
Query Representations
Cut-off Prediction
Screening Prioritisation
Document Classification
Result Synthesis
Information Extraction
Statistical Analysis
Query Automation
Screening Automation
Results Automation
Systematic Review Automation
Datasets, Test Collections

Overview


Tools, Systems, Demos
Query Refinement
Query Formulation
Query Representations
Cut-off Prediction
Screening Prioritisation
Document Classification
Result Synthesis
Information Extraction
Statistical Analysis
Query Automation
Screening Automation
Results Automation
Systematic Review Automation
Datasets, Test Collections

Systematic Reviews

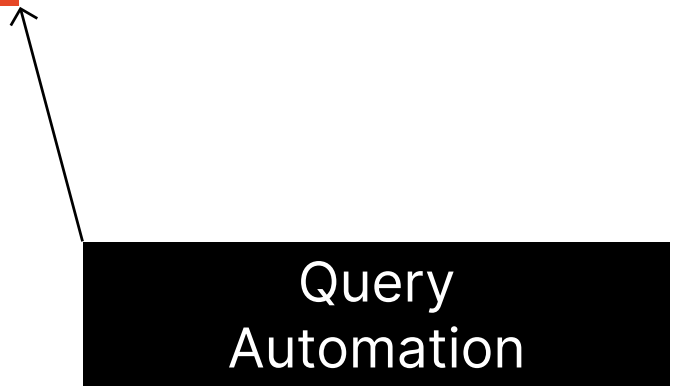
Automation

- Query Automation
- Screening Automation
- Results Automation

Summary

Query Formulation

- Formulation via LLM prompting; [Wang et al. 2023b]
- Transformer-based query formulation; [Wang et al. 2023a]
- Automitise human approaches; [Scells et al. 2021]

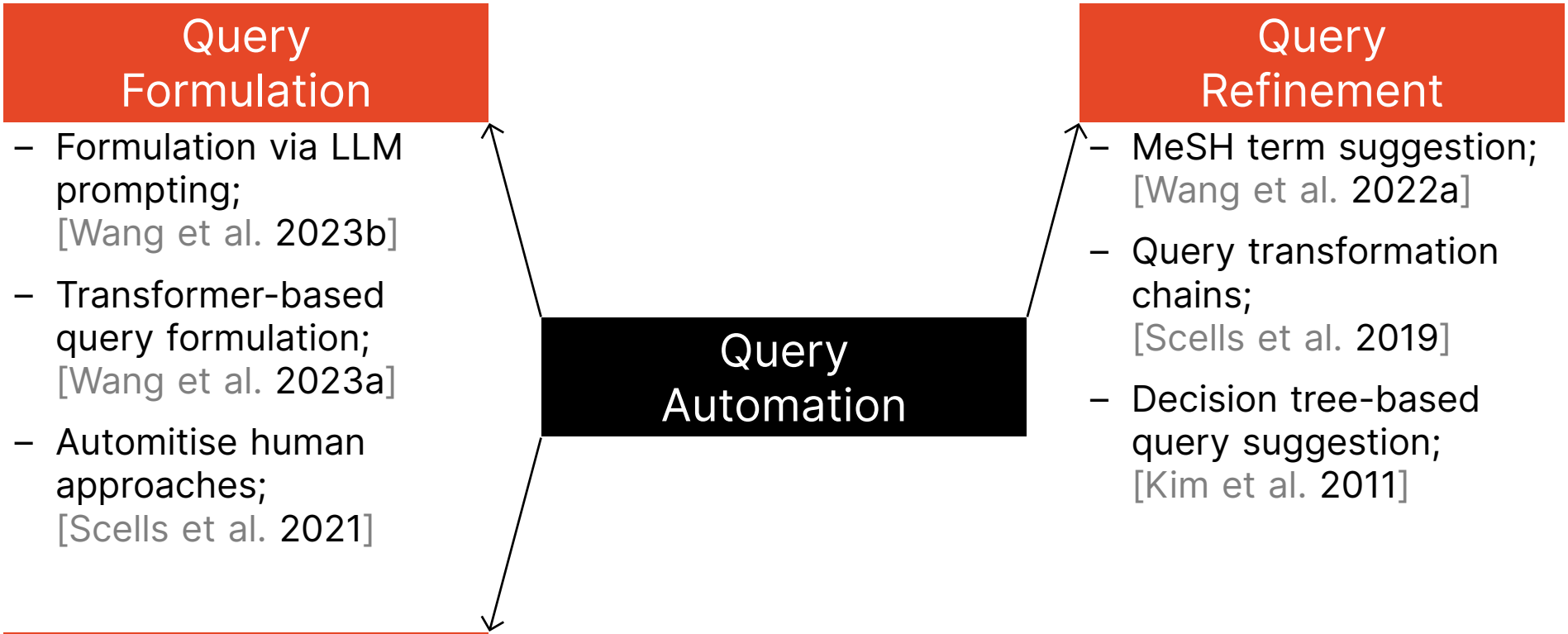


Query Formulation

- Formulation via LLM prompting; [Wang et al. 2023b]
- Transformer-based query formulation; [Wang et al. 2023a]
- Automitise human approaches; [Scells et al. 2021]

Query Refinement

- MeSH term suggestion; [Wang et al. 2022a]
- Query transformation chains; [Scells et al. 2019]
- Decision tree-based query suggestion; [Kim et al. 2011]


Query Formulation
– Formulation via LLM prompting; [Wang et al. 2023b]
– Transformer-based query formulation; [Wang et al. 2023a]
– Automitise human approaches; [Scells et al. 2021]
Query Refinement
– MeSH term suggestion; [Wang et al. 2022a]
– Query transformation chains; [Scells et al. 2019]
– Decision tree-based query suggestion; [Kim et al. 2011]
Query Automation
Query Representations
– Probabilistic operators; [Scells et al. 2023]
– Rank fusion scoring; [Scells et al. 2020]
– Query by document; [Lee and Sun 2018]

Query Formulation

- Formulation via LLM prompting; [Wang et al. 2023b]
- Transformer-based query formulation; [Wang et al. 2023a]
- Automitise human approaches; [Scells et al. 2021]

Query Refinement

- MeSH term suggestion; [Wang et al. 2022a]
- Query transformation chains; [Scells et al. 2019]
- Decision tree-based query suggestion; [Kim et al. 2011]

Query Automation

Query Representations

- Probabilistic operators; [Scells et al. 2023]
- Rank fusion scoring; [Scells et al. 2020]
- Query by document; [Lee and Sun 2018]

Tools

- Understandability (searchrefiner); [Scells and Zuccon 2018]
- Formulation (2dsearch); [Russell-Rose and Gooch 2018]

Systematic Reviews

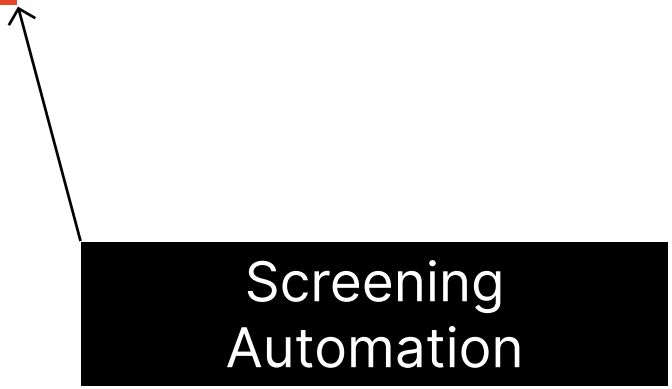
Automation

- Query Automation
- Screening Automation
- Results Automation

Summary

Screening Prioritisation

- Zero-shot LLMs for screening prioritisation; [Wang et al. 2024]
- Transformer-based screening prioritisation; [Wang et al. 2022b]
- Neural screening prioritisation; [Kusa et al. 2022]



Screening Automation

Screening Prioritisation

- Zero-shot LLMs for screening prioritisation; [Wang et al. 2024]
- Transformer-based screening prioritisation; [Wang et al. 2022b]
- Neural screening prioritisation; [Kusa et al. 2022]

Cut-off Prediction

- Point processes; [Stevenson and Hezam 2024]
- Reinforcement learning; [Hezam and Stevenson 2024]
- Geometric-based methods; [Cormack and Grossman 2016]

Screening Automation

Screening Prioritisation

- Zero-shot LLMs for screening prioritisation; [Wang et al. 2024]
- Transformer-based screening prioritisation; [Wang et al. 2022b]
- Neural screening prioritisation; [Kusa et al. 2022]

Cut-off Prediction

- Point processes; [Stevenson and Hezam 2024]
- Reinforcement learning; [Hezam and Stevenson 2024]
- Geometric-based methods; [Cormack and Grossman 2016]

Document Classification

- Transformer-based active learning; [Mao et al. 2024]
- Neural first-stage screening; [Wallace et al. 2010]
- Classical active learning; [Cormack and Grossman 2015]

Screening Automation

Screening Prioritisation

- Zero-shot LLMs for screening prioritisation; [Wang et al. 2024]
- Transformer-based screening prioritisation; [Wang et al. 2022b]
- Neural screening prioritisation; [Kusa et al. 2022]

Cut-off Prediction

- Point processes; [Stevenson and Hezam 2024]
- Reinforcement learning; [Hezam and Stevenson 2024]
- Geometric-based methods; [Cormack and Grossman 2016]

Document Classification

- Transformer-based active learning; [Mao et al. 2024]
- Neural first-stage screening; [Wallace et al. 2010]
- Classical active learning; [Cormack and Grossman 2015]

Tools

- EPPI-Reviewer; [Tsou et al. 2020]
- Rayyan; [Ouzzani et al. 2016]

Systematic Reviews

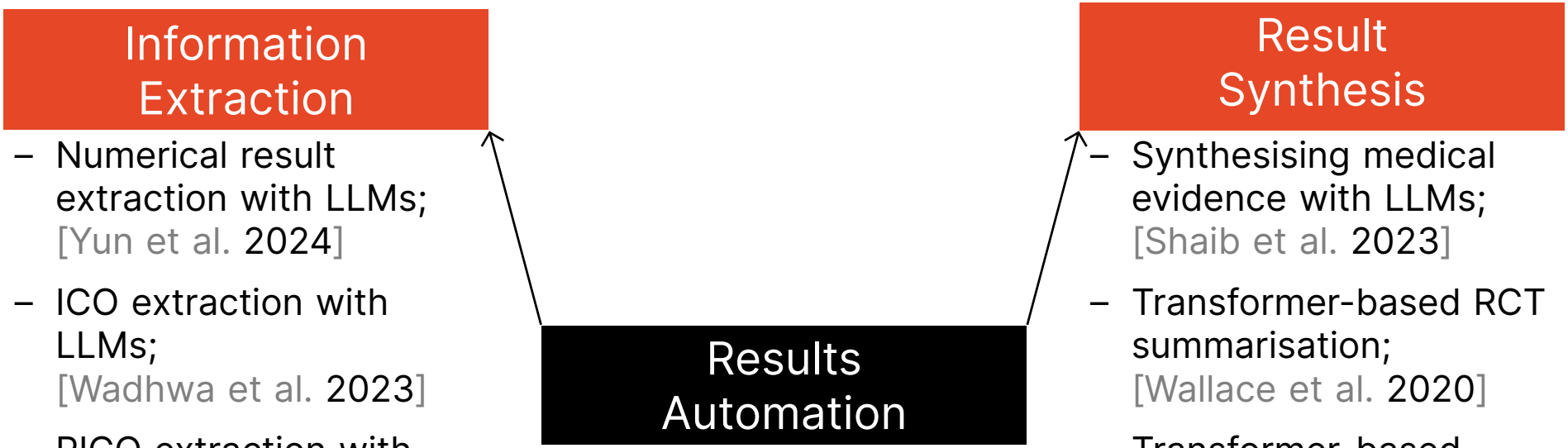
Automation

- Query Automation
- Screening Automation
- Results Automation

Summary

Information Extraction

- Numerical result extraction with LLMs; [Yun et al. 2024]
- ICO extraction with LLMs; [Wadhwa et al. 2023]
- PICO extraction with distant supervision; [Wallace et al. 2016]


Information Extraction
– Numerical result extraction with LLMs; [Yun et al. 2024]
– ICO extraction with LLMs; [Wadhwa et al. 2023]
– PICO extraction with distant supervision; [Wallace et al. 2016]
Results Automation
Result Synthesis
– Synthesising medical evidence with LLMs; [Shaib et al. 2023]
– Transformer-based RCT summarisation; [Wallace et al. 2020]
– Transformer-based plain language summarisation; [Devaraj et al. 2021]

Results Automation

Information Extraction

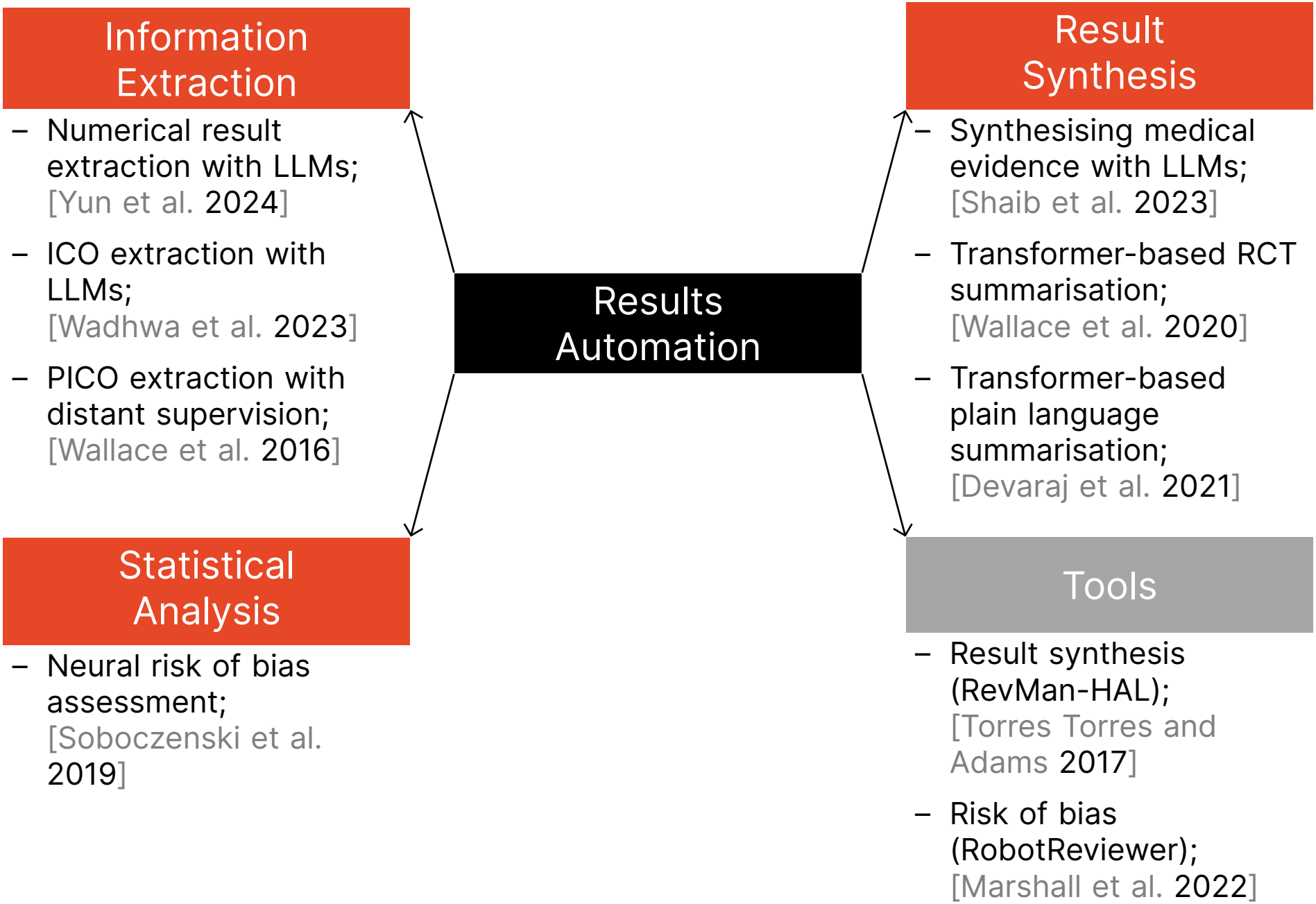
- Numerical result extraction with LLMs; [Yun et al. 2024]
- ICO extraction with LLMs; [Wadhwa et al. 2023]
- PICO extraction with distant supervision; [Wallace et al. 2016]

Statistical Analysis

- Neural risk of bias assessment; [Soboczenski et al. 2019]

Result Synthesis

- Synthesising medical evidence with LLMs; [Shaib et al. 2023]
- Transformer-based RCT summarisation; [Wallace et al. 2020]
- Transformer-based plain language summarisation; [Devaraj et al. 2021]


Information Extraction
– Numerical result extraction with LLMs; [Yun et al. 2024]
– ICO extraction with LLMs; [Wadhwa et al. 2023]
– PICO extraction with distant supervision; [Wallace et al. 2016]
Result Synthesis
– Synthesising medical evidence with LLMs; [Shaib et al. 2023]
– Transformer-based RCT summarisation; [Wallace et al. 2020]
– Transformer-based plain language summarisation; [Devaraj et al. 2021]
Results Automation
Statistical Analysis
– Neural risk of bias assessment; [Soboczenski et al. 2019]
Tools
– Result synthesis (RevMan-HAL); [Torres Torres and Adams 2017]
– Risk of bias (RobotReviewer); [Marshall et al. 2022]

Systematic Reviews

Automation

Summary

Next Steps

- Many effective methods to automate all areas of systematic reviews

Next Steps

- Many effective methods to automate all areas of systematic reviews
- In medicine: explosion of methods due to shared tasks and datasets
  - CLEF Technology Assisted Reviews
    [https://github.com/CLEF-TAR/tar]
  - CSMeD: Meta-dataset for systematic review automation evaluation
    [Kusa et al. 2023]

Next Steps

- Many effective methods to automate all areas of systematic reviews
- In medicine: explosion of methods due to shared tasks and datasets
  - CLEF Technology Assisted Reviews
    [https://github.com/CLEF-TAR/tar]
  - CSMeD: Meta-dataset for systematic review automation evaluation
    [Kusa et al. 2023]
- In climate science: need standardised datasets and test collections
  - Medicine ➔ PubMed; freely downloadable, open data
    [https://pubmed.ncbi.nlm.nih.gov/download/]
  - Climate Science ➔ OpenAlex? 250M scholarly works, CC0 license
    [https://openalex.org/]

Next Steps

- Many effective methods to automate all areas of systematic reviews
- In medicine: explosion of methods due to shared tasks and datasets
  - CLEF Technology Assisted Reviews
    [https://github.com/CLEF-TAR/tar]
  - CSMeD: Meta-dataset for systematic review automation evaluation
    [Kusa et al. 2023]
- In climate science: need standardised datasets and test collections
  - Medicine ➔ PubMed; freely downloadable, open data
    [https://pubmed.ncbi.nlm.nih.gov/download/]
  - Climate Science ➔ OpenAlex? 250M scholarly works, CC0 license
    [https://openalex.org/]

Stay in touch

https://scells.me/manila24 @hscells harry.scells@uni-leipzig.de

References I

Cormack, Gordon V. and Maura R. Grossman (2015). "Autonomy and Reliability of Continuous Active Learning for Technology-Assisted Review". In: *CoRR* abs/1504.06868. arXiv: 1504.06868.

Cormack, Gordon V. and Maura R. Grossman (2016). "Engineering Quality and Reliability in Technology-Assisted Review". In: *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*. ACM, pp. 75–84. DOI: 10.1145/2911451.2911510.

Devaraj, Ashwin, Iain James Marshall, Byron C. Wallace, and Junyi Jessy Li (2021). "Paragraph-level Simplification of Medical Texts". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*. Association for Computational Linguistics, pp. 4972–4984. DOI: 10.18653/V1/2021.NAACL-MAIN.395.

Hezam, Reem Bin and Mark Stevenson (2024). "RLStop: A Reinforcement Learning Stopping Method for TAR". In: *CoRR* abs/2405.02525. DOI: 10.48550/ARXIV.2405.02525. arXiv: 2405.02525.

Kim, Youngho, Jangwon Seo, and W. Bruce Croft (2011). "Automatic boolean query suggestion for professional search". In: *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*. ACM, pp. 825–834. DOI: 10.1145/2009916.2010026.

Kusa, Wojciech, Allan Hanbury, and Petr Knoth (2022). "Automation of Citation Screening for Systematic Literature Reviews Using Neural Networks: A Replicability Study". In: *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part I*. Vol. 13185. Lecture Notes in Computer Science. Springer, pp. 584–598. DOI: 10.1007/978-3-030-99736-6_39.

Kusa, Wojciech, Óscar E. Mendoza, Matthias Samwald, Petr Knoth, and Allan Hanbury (2023). "CSMeD: Bridging the Dataset Gap in Automated Citation Screening for Systematic Literature Reviews". In: *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Lee, Grace E. and Aixin Sun (2018). "Seed-driven Document Ranking for Systematic Reviews in Evidence-Based Medicine". In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*. ACM, pp. 455–464. DOI: 10.1145/3209978.3209994.

Lee, Grace E. and Aixin Sun (2022). "Towards Reducing Manual Workload in Technology-Assisted Reviews: Estimating Ranking Performance". In: *CoRR* abs/2201.05648. arXiv: 2201.05648.

Mao, Xinyu, Bevan Koopman, and Guido Zuccon (2024). "A Reproducibility Study of Goldilocks: Just-Right Tuning of BERT for TAR". In: *Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part IV*. Vol. 14611. Lecture Notes in Computer Science. Springer, pp. 132–146. DOI: 10.1007/978-3-031-56066-8_13.

Marshall, Iain, Joël Kuiper, byron wallace, Derek, Sebastián Gálvez, Edward Banner, Frank, and Arash Joorabchi (July 2022). *ijmarshall/robotreviewer: RobotReviewer v0.7*. Version 0.7. DOI: 10.5281/zenodo.6855718.

Ouzzani, Mourad, Hossam Hammady, Zbys Fedorowicz, and Ahmed Elmagarmid (2016). "Rayyan—a web and mobile app for systematic reviews". In: *Systematic reviews* 5, pp. 1–10. DOI: https://doi.org/10.1186/s13643-016-0384-4.

Russell-Rose, Tony and Phil Gooch (2018). "2dSearch: A Visual Approach to Search Strategy Formulation". In: *Proceedings of the First Biennial Conference on Design of Experimental Search & Information Retrieval Systems, Bertinoro, Italy, August 28-31, 2018*. Vol. 2167. CEUR Workshop Proceedings. CEUR-WS.org, pp. 90–96.

References II

Scells, Harrisen, Leif Azzopardi, Guido Zuccon, and Bevan Koopman (2018). "Query Variation Performance Prediction for Systematic Reviews". In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*. ACM, pp. 1089–1092. DOI: 10.1145/3209978.3210078.

Scells, Harrisen, Ferdinand Schlatt, and Martin Potthast (2023). "Smooth Operators for Effective Systematic Review Queries". In: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*. ACM, pp. 580–590. DOI: 10.1145/3539618.3591768.

Scells, Harrisen and Guido Zuccon (2018). "searchrefiner: A Query Visualisation and Understanding Tool for Systematic Reviews". In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*. ACM, pp. 1939–1942. DOI: 10.1145/3269206.3269215.

Scells, Harrisen, Guido Zuccon, and Bevan Koopman (2019). "Automatic Boolean Query Refinement for Systematic Review Literature Search". In: *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*. ACM, pp. 1646–1656. DOI: 10.1145/3308558.3313544.

Scells, Harrisen, Guido Zuccon, and Bevan Koopman (2020). "You Can Teach an Old Dog New Tricks: Rank Fusion applied to Coordination Level Matching for Ranking in Systematic Reviews". In: *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part I*. Vol. 12035. Lecture Notes in Computer Science. Springer, pp. 399–414. DOI: 10.1007/978-3-030-45439-5_27.

Scells, Harrisen, Guido Zuccon, and Bevan Koopman (2021). "A comparison of automatic Boolean query formulation for systematic reviews". In: *Inf. Retr. J.* 24.1, pp. 3–28. DOI: 10.1007/S10791-020-09381-1.

Shaib, Chantal, Millicent L. Li, Sebastian Joseph, Iain James Marshall, Junyi Jessy Li, and Byron C. Wallace (2023). "Summarizing, Simplifying, and Synthesizing Medical Evidence using GPT-3 (with Varying Success)". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*. Association for Computational Linguistics, pp. 1387–1407. DOI: 10.18653/V1/2023.ACL-SHORT.119.

Soboczenski, Frank, Thomas A. Trikalinos, Joël Kuiper, Randolph G. Bias, Byron C. Wallace, and Iain James Marshall (2019). "Machine learning to help researchers evaluate biases in clinical trials: a prospective, randomized user study". In: *BMC Medical Informatics Decis. Mak.* 19.1, 96:1–96:12. DOI: 10.1186/S12911-019-0814-Z.

Stevenson, Mark and Reem Bin Hezam (2024). "Stopping Methods for Technology-assisted Reviews Based on Point Processes". In: *ACM Trans. Inf. Syst.* 42.3, 73:1–73:37. DOI: 10.1145/3631990.

Torres Torres, Mercedes and Clive E Adams (2017). "RevManHAL: towards automatic text generation in systematic reviews". In: *Systematic reviews* 6, pp. 1–7. DOI: https://doi.org/10.1186/s13643-017-0421-y.

Tsou, Amy Y, Jonathan R Treadwell, Eileen Erinoff, and Karen Schoelles (2020). "Machine learning for screening prioritization in systematic reviews: comparative performance of Abstrackr and EPPI-Reviewer". In: *Systematic reviews* 9, pp. 1–14. DOI: https://doi.org/10.1186/s13643-020-01324-7.

Wadhwa, Somin, Jay DeYoung, Benjamin E. Nye, Silvio Amir, and Byron C. Wallace (2023). "Jointly Extracting Interventions, Outcomes, and Findings from RCT Reports with LLMs". In: *Machine Learning for Healthcare Conference, MLHC 2023, 11-12 August 2023, New York, USA*. Vol. 219. Proceedings of Machine Learning Research. PMLR, pp. 754–771.

References III

Wallace, Byron C., Joël Kuiper, Aakash Sharma, Mingxi (Brian) Zhu, and Iain James Marshall (2016). "Extracting PICO Sentences from Clinical Trial Reports using Supervised Distant Supervision". In: *J. Mach. Learn. Res.* 17, 132:1–132:25.

Wallace, Byron C., Sayantan Saha, Frank Soboczenski, and Iain James Marshall (2020). "Generating (Factual?) Narrative Summaries of RCTs: Experiments with Neural Multi-Document Summarization". In: *CoRR* abs/2008.11293. arXiv: 2008.11293.

Wallace, Byron C., Thomas A. Trikalinos, Joseph Lau, Carla E. Brodley, and Christopher H. Schmid (2010). "Semi-automated screening of biomedical citations for systematic reviews". In: *BMC Bioinform.* 11, p. 55. DOI: 10.1186/1471-2105-11-55.

Wang, Shuai, Harrisen Scells, Bevan Koopman, Martin Potthast, and Guido Zuccon (2023a). "Generating Natural Language Queries for More Effective Systematic Review Screening Prioritisation". In: *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, SIGIR-AP 2023, Beijing, China, November 26-28, 2023*. ACM, pp. 73–83. DOI: 10.1145/3624918.3625322.

Wang, Shuai, Harrisen Scells, Bevan Koopman, and Guido Zuccon (2022a). "Automated MeSH term suggestion for effective query formulation in systematic reviews literature search". In: *Intell. Syst. Appl.* 16, p. 200141. DOI: 10.1016/J.ISWA.2022.200141.

Wang, Shuai, Harrisen Scells, Bevan Koopman, and Guido Zuccon (2022b). "Neural Rankers for Effective Screening Prioritisation in Medical Systematic Review Literature Search". In: *Proceedings of the 26th Australasian Document Computing Symposium, ADCS 2022, Adelaide, SA, Australia, December 15-16, 2022*. ACM, 4:1–4:10. DOI: 10.1145/3572960.3572980.

Wang, Shuai, Harrisen Scells, Bevan Koopman, and Guido Zuccon (2023b). "Can ChatGPT Write a Good Boolean Query for Systematic Review Literature Search?" In: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*. ACM, pp. 1426–1436. DOI: 10.1145/3539618.3591703.

Wang, Shuai, Harrisen Scells, Shengyao Zhuang, Martin Potthast, Bevan Koopman, and Guido Zuccon (2024). "Zero-Shot Generative Large Language Models for Systematic Review Screening Automation". In: *Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part I*. Vol. 14608. Lecture Notes in Computer Science. Springer, pp. 403–420. DOI: 10.1007/978-3-031-56027-9_25.

Yun, Hye Sun, David Pogrebitskiy, Iain James Marshall, and Byron C. Wallace (2024). "Automatically Extracting Numerical Results from Randomized Controlled Trials with Large Language Models". In: *CoRR* abs/2405.01686. DOI: 10.48550/ARXIV.2405.01686. arXiv: 2405.01686.