A Visual Approach to Query Formulation for Systematic Search

Tony Russell-Rose
UXlabs Ltd
3000 Cathedral Hill,
Guildford, Surrey, UK
tgr@uxlabs.co.uk

Jon Chamberlain
School of Computer Science and
Electronic Engineering,
University of Essex
Colchester, Essex, UK
jchamb@essex.ac.uk

Farhad Shokraneh
The Institute of Mental Health,
University of Nottingham
Nottingham, UK
farhadshokraneh@gmail.com

ABSTRACT

Knowledge workers (such as healthcare information professionals, patent agents and legal researchers) need to create and execute search strategies that are accurate, repeatable and transparent. The traditional solution offered by most database vendors is to use proprietary line-by-line 'query builders'. However, these offer limited support for error checking or query optimisation, and their output can often be compromised by errors and inefficiencies. Using the healthcare domain for context, we demonstrate a new approach to search strategy formulation in which concepts are expressed as objects on a two-dimensional canvas, and relationships are articulated using direct manipulation. This approach eliminates many sources of syntactic error, makes the query semantics more transparent, and offers new ways to optimise, save and share search strategies and best practices.

CCS CONCEPTS

• **Information systems** → **Query representation**; **Search interfaces**; *Expert search*; • **Computing methodologies** → *Representation of Boolean functions*;

KEYWORDS

healthcare information retrieval, boolean, systematic search, search visualisation, systematic review

ACM Reference Format:
Tony Russell-Rose, Jon Chamberlain, and Farhad Shokraneh. 2019. A Visual Approach to Query Formulation for Systematic Search. In *2019 Conference on Human Information Interaction and Retrieval (CHIIR '19), March 10–14, 2019, Glasgow, United Kingdom.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3295750.3298919

1 INTRODUCTION

Medical knowledge is growing so rapidly that it is difficult for healthcare professionals to keep up. As the volume of published studies increases each year [14], the gap between research knowledge and professional practice grows ever wider [2]. Healthcare information professionals play a key role in closing this gap by synthesising the complex, incomplete and at times conflicting findings of biomedical research into a form that can readily inform healthcare decision making [5]. The systematic literature review process relies on the painstaking and meticulous searching of multiple databases using complex Boolean search strategies that often consist of hundreds of keywords, operators and ontology terms [12]. An example from the CLEF 2017 eHealth Lab dataset[1] is shown in Figure 1 (titled 'Galactomannan detection for invasive aspergillosis in immunocompromised patients').

```
1 "Aspergillus"[MeSH]
2 "Aspergillosis"[MeSH]
3 "Pulmonary Aspergillosis"[MeSH]
4 aspergill*[tiab]
5 fungal infection[tw]
6 (invasive[tiab] AND fungal[tiab])
7 1 OR 2 OR 3 OR 4 OR 5 OR 6
8 "Serology"[MeSH]
9 Serology"[MeSH]
10 (serology[tiab] OR serodiagnosis[tiab] OR serologic[tiab])
11 8 OR 9 OR 10
12 "Immunoassay"[MeSH]
13 (immunoassay[tiab] OR immunoassays[tiab])
14 (immuno assay[tiab] OR immuno assays[tiab])
15 (ELISA[tiab] OR ELISAs[tiab] OR EIA[tiab] OR EIAs[tiab])
16 immunosorbent[tiab]
17 12 OR 13 OR 14 OR 15 OR 16
18 Platelia[tw]
19 "Mannans"[MeSH]
20 galactomannan[tw]
21 18 OR 19 OR 20
22 11 OR 17 OR 21
23 7 AND 22
```

Figure 1: An example Boolean search strategy

The choice of search strategy plays a vital role in ensuring that the review process is sufficiently exhaustive and that the outcome is not biased by easily accessible studies [10]. In addition, the strategy needs to be transparent and repeatable, so that others may replicate the methodology. However, systematic literature reviews can take years to complete [2], with new research findings may be published in the interim, leading to a lack of currency and potential for inaccuracy [23]. Moreover, there are often mistakes in search strategies reported in the literature that prevent them from being executed in their published form. In one study of 63 MEDLINE strategies, at least one error was detected in over 90%, including

[1] https://sites.google.com/site/clefehealth2017/task-2, accessed 10 October 2018.

CHIIR '19, March 10–14, 2019, Glasgow, United Kingdom

ACM ISBN 978-1-4503-6025-8/19/03.
https://doi.org/10.1145/3295750.3298919

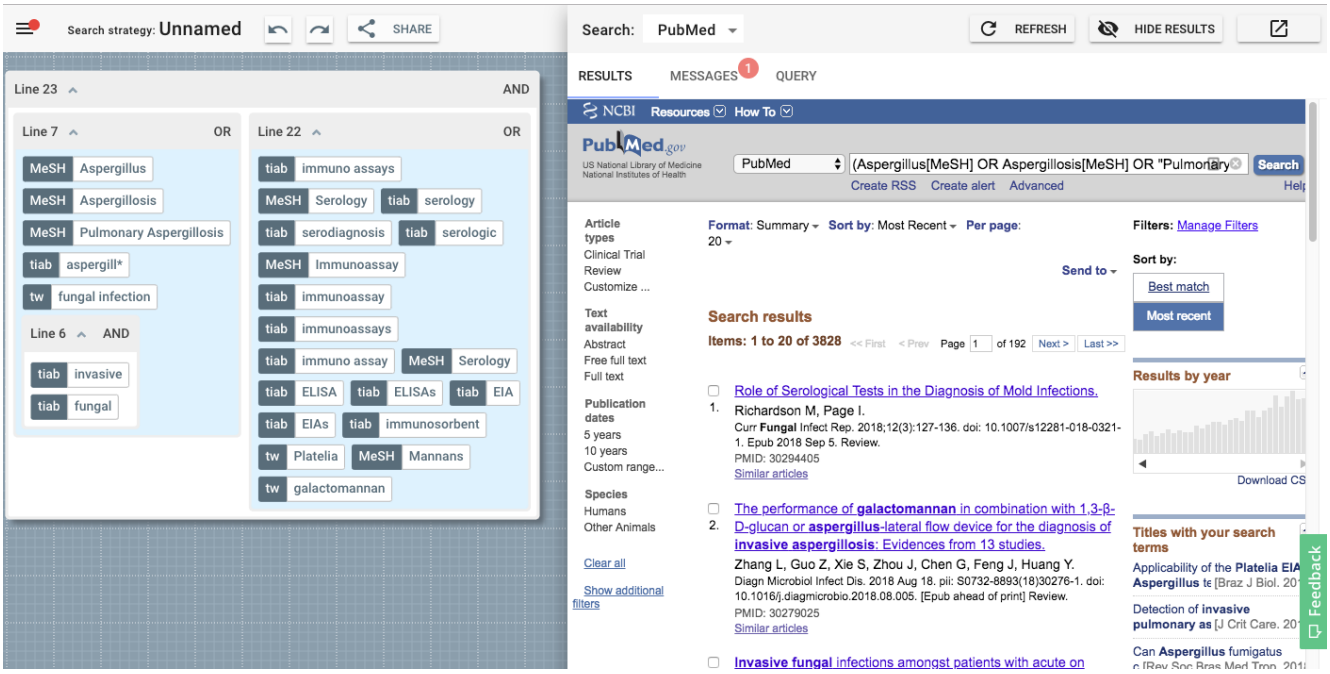

Figure 2: Screenshot of the 2dSearch application showing the query canvas (left) and the search results pane (right).

spelling errors, truncation errors, logical operator error, incorrect query line references, redundancy without rationale, etc. [5].

One study showed the average time spent on a systematic literature search was 26.9 hours (median: 18.5 hours), depending on the information professional's experience, most of which is spent on developing the search strategies and translating them into a compatible format for each search interface [4]. Unfortunately, these search strategies, as part of research methods, are getting lost in the process of reporting. This loss undermines the two principles of replicability and reproducibility of systematic reviews [7, 13, 15]. Although because of space limitation some journals have to accept the search strategies as external supplementary files, this practice is not strategic in terms of storage and data linkage [17], and these appendices are becoming inaccessible as the time passes [6]. Despite valuable efforts to save and share these search strategies in an online static database and update them [22], it is not widely resourced or updated so the information professionals end up creating and recreating search strategies for the same concepts. These requirements are considered valuable in multiple professional contexts in which complex Boolean search is a common task, such as media monitoring, recruitment (sourcing) [18], patent search and legal research [20].

Our proposed approach (called 2dSearch[2]) offers a radical alternative to conventional 'advanced search'. Instead of entering Boolean strings into one-dimensional search boxes, queries are formulated by combining objects on a two-dimensional canvas. The use of a visual approach eliminates many sources of syntactic error, makes the query semantics more transparent, and offers further opportunities for query refinement and optimisation.

[2]https://2dsearch.com

In this paper, we describe related work and explain how 2dSearch supports and extends their key design principles and insights. We then explore the application in more detail, reviewing the support it offers for search strategy formulation, optimisation, sharing and re-use. Finally, we discuss some of the implementation details along with opportunities for commercial impact and further work.

2 RELATED WORK

The application of data visualisation to search query formulation can offer significant benefits, such as fewer zero-hit queries, improved query comprehension and better support for exploration of an unfamiliar database [9]. An early example of such an approach is that of Anick et al. [1], who developed a system that could parse natural language queries and represent them as movable tiles on a two-dimensional canvas. The user was able to re-arrange the tiles to reformulate the expression, and to activate or deactivate alternative elements to optimise the query. In addition, the system offered support for integration with thesauri and it displayed the number of hits in the lower left corner of each tile.

In subsequent work, Fishkin and Stone [8] investigated the application of direct manipulation techniques to the problem of database query formulation, using a system of 'lenses' to refine and filter the data. Lenses could be combined by stacking them and applying a suitable operator, or combined to create compound lenses, and hence support the encapsulation of queries of arbitrary complexity. Influential work by Jones [11] proposed an approach in which concepts are expressed using a Venn diagram notation combined with integrated query result previews. Queries could be formulated

by overlapping objects within the workspace to create intersections and disjunctions, and subsets could be selected to facilitate execution of subcomponents of an overall query.

A further example is that of Yi et al. [24], who developed a system based around a 'dust and magnet' metaphor, in which dimensions of interest within the data could be represented as 'magnets' on a visual canvas, and the relationships between points in the data could be understood by observing the effect of the 'magnetic forces' on individual 'data particles'. More recently, Nitsche and Nürnberger [16] developed a system based around a radial interface in which queries and results could be integrated and collectively manipulated. The concept utilised a pseudo-desktop metaphor in which objects of interest clustered toward the centre. Query objects could be entered directly onto this canvas, and their proximity to the centre and to other objects was used as a relevance cue, influencing the selection and position of search results.

2dSearch adopts and extends many of the design principles and insights embodied in this work, for example:

- Boolean expressions can be formulated as a objects on a canvas, and arranged by direct manipulation;
- Query elements can be individually invoked or interrogated to facilitate exploration;
- By nesting aggregate structures, it is possible to create queries of arbitrary complexity;
- Interaction and animation can be used to communicate meaning and structure;
- Real-time feedback is fundamental to effective query optimisation.

3 DESIGN CONCEPT

3.1 Query formulation

At the heart of 2dSearch is a graphical editor which allows the user to formulate search strategies using a visual framework in which concepts are expressed as objects on a two-dimensional canvas. Concepts can be simple keywords or attribute:value pairs representing controlled vocabulary terms (e.g. Mesh terms) or database-specific search operators (e.g. field codes and other commands). They can be combined using Boolean (and other) operators to form higher-level groups and then iteratively nested to create expressions of arbitrary complexity.

The application itself consists of two panes (see Figure 2): a query canvas on the left and a search results pane on the right (which can be resized or detached in a separate tab or window). The canvas itself can be resized or zoomed, and features an 'overview' widget which allows the user to view or navigate to elements that may be outside the current viewport. Adopting design cues from Google's Material Design language[3], a sliding menu is offered on the left, providing file I/O and other options. This is complemented by a navigation bar across the top which provides support for common document-level functions such as naming and sharing search strategies.

Although 2dSearch supports the creation of complete strategies from a blank canvas, its function and value are most readily understood by reference to an example (i.e. text-based) search strategy,

[3]https://material.io

Figure 3: Visualising a text-based search strategy

such as that shown in Figure 1. A trained professional may be able to mentally 'parse' the sequence of commands shown and interpret the general approach, but without associated documentation it is difficult to understand exactly what the searcher intended. Moreover, it is difficult to optimise, debug or re-use strategies expressed in this form.

However, when this strategy is opened using 2dSearch, its structure becomes much more apparent (see Figure 3). It can be seen that the overall expression consists of a conjunction of two disjunctions (Lines 7 and 22), the first of which articulates variations on the fungal infection concept, while the latter contains various nested disjunctions to capture the diagnostic test (serology) and associated procedures. Evidently, the line numbers themselves are somewhat arbitrary in this context, having served an original purpose analogous to that of line numbering in first generation BASIC. However, by displaying them as nested groups with transparent structure, 2dSearch offers support for abstraction, in which lower-level details can be hidden and higher-level structure revealed. Moreover, it is now possible to give meaningful names to sub groups, so that they can be saved and re-used as modular components.

Although visualisation of search strategies in this manner can offer immediate utility, the true value of the approach is not so much in the *information* design, but in the *interaction* design. For example, to edit the expression, the user can move terms from one block to another using direct manipulation, and create new groups simply by combining terms. They can also cut, copy, delete, and lasso multiple objects. If they want to understand the effect of one block in isolation, they can execute it individually. Conversely, if they want to remove one element from consideration, they can temporarily disable it. It is also possible edit the content inline, interchanging Mesh terms with keywords and field tags as required.

In each case, the effects of each editing operation are displayed in real time in the adjacent search results pane.

3.2 Query execution

2dSearch functions as a meta-search engine, so is in principle agnostic of any particular search technology or platform. In practice however, to execute a given query and retrieve results, the semantics of the canvas content must be mapped to the API of the underlying database. This is achieved via an abstraction layer or set of 'adapters' for common search platforms such as Bing, Google, PubMed, etc. These are user selectable in the interface via a drop-down control.

Search platforms vary widely in the extent to which they support complex querying. Google, for example, is limited in its support for Boolean querying, so an adapter for this platform has been provided more for reasons of familiarity and proof of concept than long-term utility. Conversely, the true value of approaches such as 2dSearch becomes more apparent when coupled with databases that offer more sophisticated search functionality, such as PubMed and other specialist databases.

It is common for healthcare information professionals to want to search more than one database, particularly when undertaking a systematic literature review [19]. In practice, this requires a process of 'translation' of the search strategy to match the syntax of the target database and the search operators it supports. For a relatively simple query this may not be a major undertaking, particularly if such operators form a relatively small proportion of the overall search strategy. However, the user still has to understand which elements are platform-specific, identify the closest equivalent in the other database and manually edit their query, all of which is laborious and time consuming [3].

Since 2dSearch uses a visual framework that is database-agnostic, it raises the prospect of a universal language for search strategy formulation, in which information needs can be articulated in a generic manner, and the task of mapping to the semantics of an underlying database can be delegated to platform-specific adapters. Evidently, this is an ambitious goal, since some search strategies will always require human judgement for accurate translation [3]. However, a significant proportion of the translation process is routine in nature and is thus amenable to some form of automated support [19].

2dSearch already provides elementary support for search strategy translation in the form of a 'Messages' tab on the results pane. This serves a similar purpose to a console or messages pane in a software IDE, alerting the user to compilation issues and offering advice, fixes and workarounds. For example, if the user tries to execute a query string using Bing containing operators specific to Google, an alert is shown and the unknown operators are listed in the 'Messages' tab. In due course, this mechanism could be extended to offer a greater degree of interactive support for the translation of strategies across databases. Moreover, 2dSearch also offers the potential for search strategy optimisation through the elimination of redundant structure (eg. spurious brackets or duplicate elements) and comparison of canonical representations.

4 IMPLEMENTATION

2dSearch is implemented as a web app using Vue.js[4] and other Javascript libraries. User authentication is provided via Auth0[5] and persistence of search strategies and other user data is implemented using MongoDB.[6] Query suggestions are provided via an NLP services API which utilises various Python NLP libraries (for user authentication, word embedding, keyword extraction, etc.) and SPARQL endpoints (for linked open data ontology lookup)[21]. The NLP API is deployed via Digital Ocean[7], and the production instance of the 2dSearch client is deployed via Heroku.[8]

2dSearch offers utility to anyone who needs to create search strategies that are comprehensive, repeatable, and transparent. This includes applications in digital libraries, healthcare, legal research, media monitoring, patent search and recruitment/sourcing [19]. In practice, many of these would entail the development of additional, specialist adapters to support custom database integration. However, the profession for whom the benefit may be greatest is (arguably) healthcare information professionals, due to the strict governance and reporting requirements of systematic literature review. In this context, our current integration with PubMed offers immediate utility, particularly with its ability to visualise, save and share components as reusable, executable building blocks, and the potential for semi-automated strategy translation and optimisation.

5 SUMMARY AND FURTHER WORK

2dSearch is a framework for search strategy formulation in which queries are expressed by combining objects on a two-dimensional canvas. Transforming logical structure into visual layout provides a more direct mapping between the underlying semantics and the physical appearance. This helps to eliminate syntax errors, makes the query semantics more transparent and offers new ways to optimise, save and share search strategies.

We currently provide adapters for Google, Google Scholar, Bing and PubMed, the latter of which offers immediate utility to anyone wishing to search MEDLINE in a systematic manner. In due course, other adapters will be provided, but in the short term we would hope to engage in a formal, user-centric evaluation of 2dSearch, particularly in relation to traditional 'line-by line' query builders. We are currently engaging in an outreach programme with the healthcare information community and we welcome feedback of any sort. We hope to work with domain experts in building repositories of curated (or user generated) content in the form of best practice examples and templates.

Adopting a database-agnostic approach presents challenges, but it also offers the prospect of a universal framework for search strategy formulation in which information needs can be articulated in a generic manner and the task of mapping to the semantics of an underlying database can be delegated to platform-specific adapters. If that transpires to be a practicable proposition, then such a development could have profound implications for the way in which professional search skills are taught, learnt and applied.

[4] https://vuejs.org
[5] https://auth0.com
[6] https://www.mongodb.com
[7] https://www.digitalocean.com
[8] https://www.heroku.com

REFERENCES

[1] P. G. Anick, J. D. Brennan, R. A. Flynn, D. R. Hanssen, B. Alvey, and J. M. Robbins. 1990. A Direct Manipulation Interface for Boolean Information Retrieval via Natural Language Query. In *Proceedings of the 13th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '90)*. ACM, New York, NY, USA, 135–150. https://doi.org/10.1145/96749.98015

[2] H. Bastian, P. Glasziou, and I. Chalmers. 2010. Seventy-Five Trials and Eleven Systematic Reviews a Day: How Will We Ever Keep Up? *PLoS Med* 7, 9 (2010). https://doi.org/10.1371/journal.pmed.1000326

[3] W. M. Bramer, M. L. Rethlefsen, J. Kleijnen, and O. H. Franco. 2017. Optimal Database Combinations for Literature Searches in Systematic Reviews: A Prospective Exploratory Study. *Systematic Reviews* 6, 245 (2017). https://doi.org/10.1186/s13643-017-0644-y

[4] K. Bullers, A. M. Howard, A. Hanson, W. D. Kearns, J. J. Orriola, R. L. Polo, and K. A. Sakmar. 2018. It Takes Longer Than You Think: Librarian Time Spent on Systematic Review Tasks. *Journal of the Medical Library Association (JMLA)* 106, 2 (2018), 198–207. https://doi.org/10.5195/jmla.2018.323

[5] J. H. Elliott, T. Turner, O. Clavisi, J. Thomas, J. P. T. Higgins, C. Mavergames, and R. L. Gruen. 2014. Living Systematic Reviews: An Emerging Opportunity to Narrow the Evidence-Practice Gap. *PLoS Med* 11, 2 (18 Feb. 2014). https://doi.org/10.1371/journal.pmed.1001603

[6] E. Evangelou, T. A. Trikalinos, and J. P. Ioannidis. 2005. Unavailability of Online Supplementary Scientific Information From Articles Published in Major Journals. *FASEB Journal: Official publication of the Federation of American Societies for Experimental Biology* 19, 14 (2005), 1943–4. https://doi.org/10.1096/fj.05-4784lsf

[7] C. M. Faggion Jr, R. Huivin, L. Aranda, N. Pandis, and M. Alarcon. 2018. The Search and Selection for Primary Studies in Systematic Reviews Published in Dental Journals Indexed in MEDLINE Was Not Fully Reproducible. *Journal of Clinical Epidemiology* 98 (2018), 53–61. https://doi.org/10.1016/j.jclinepi.2018.02.011

[8] K. Fishkin and M. C. Stone. 1995. Enhanced Dynamic Queries via Movable Filters. ACM Press, 415–420.

[9] Joseph H. Goldberg and Uday N. Gajendar. 2008. Graphical condition builder for facilitating database queries. *U.S. Patent No. 7,383,513. 3* (2008).

[10] P. Hemingway and N. Brereton. 2009. What is a Systematic Review? http://www.bandolier.org.uk/painres/download/whatis/Syst-review.pdf

[11] S. Jones. 1998. Graphical Query Specification and Dynamic Result Previews for a Digital Library. In *Proceedings of the 11th Annual ACM Symposium on User Interface Software and Technology (UIST '98)*. ACM, New York, NY, USA, 143–151. https://doi.org/10.1145/288392.288595

[12] S. Karimi, S. Pohl, F. Scholer, L. Cavedon, and J. Zobel. 2010. Boolean Versus Ranked Querying for Biomedical Systematic Reviews. *BMC Med Inform Decis Mak* 10, 58 (2010). https://doi.org/10.1186/1472-6947-10-58

[13] J. B. Koffel and M. L. Rethlefsen. 2016. Reproducibility of Search Strategies Is Poor in Systematic Reviews Published in High-Impact Pediatrics, Cardiology and Surgery Journals: A Cross-Sectional Study. *PloS One* 11, 9 (2016). https://doi.org/10.1371/journal.pone.0163309

[14] Z. Lu. 2011. PubMed and Beyond: A Survey of Web Tools for Searching Biomedical Literature. *Database* 2011 (2011). https://doi.org/10.1093/database/baq036

[15] M. M. Mullins, J. B. DeLuca, N. Crepaz, and C. M. Lyles. 2014. Reporting Quality of Search Methods in Systematic Reviews of HIV Behavioral Interventions (2000-2010): Are The Searches Clearly Explained, Systematic and Reproducible? *Research Synthesis Methods* 5, 2 (2014), 116–30. https://doi.org/10.1002/jrsm.1098

[16] M. Nitsche and A. Nürnberger. 2006. QUEST: Querying Complex Information by Direct Manipulation. *In: Yamamoto S. (eds) Human Interface and the Management of Information. Information and Interaction Design. HIMI 2013. Lecture Notes in Computer Science* 8016 (2006).

[17] A. Price, S. Schroter, M. Clarke, and H. McAneney. 2018. Role of Supplementary Material in Biomedical Journal Articles: Surveys of Authors, Reviewers and Readers. *BMJ Open* 8, 9 (2018). https://doi.org/10.1136/bmjopen-2018-021753

[18] T. Russell-Rose and J. Chamberlain. 2016. Searching for talent: The information retrieval challenges of recruitment professionals. *Business Information Review* 33, 1 (2016), 40–48. https://doi.org/10.1177/0266382116631849

[19] T. Russell-Rose and J. Chamberlain. 2017. Expert Search Strategies: The Information Retrieval Practices of Healthcare Information Professionals. *JMIR Med Inform* 5, 4 (02 Oct 2017), e33. https://doi.org/10.2196/medinform.7680

[20] T. Russell-Rose, J. Chamberlain, and L. Azzopardi. 2018. Information Retrieval in the Workplace: A Comparison of Professional Search Practices. *Information Processing & Management* 54, 6 (2018), 1042–1057. https://doi.org/10.1016/j.ipm.2018.07.003

[21] T. Russell-Rose and P. Gooch. 2018. 2dSearch: A Visual Approach to Search Strategy Formulation. In *Proceedings of DESIRES: Design of Experimental Search Information REtrieval Systems* (28-31 August 2018) *(DESIRES 2018)*.

[22] A. A. Saleh, M. A. Ratajeski, and J. Ladue. 2014. Development of a Web-based Repository for Sharing Biomedical Terminology From Systematic Review Searches: A Case Study. *Medical Reference Services Quarterly* 33, 2 (2014), 167–78. https://doi.org/10.1080/02763869.2014.897518

[23] H. Tang and J. H. K. Ng. 2006. Googling for a Diagnosis—Use of Google as a Diagnostic Aid: Internet Based Study. *BMJ* 333, 7579 (2006), 1143–1145. https://doi.org/10.1136/bmj.39003.640567.AE

[24] J. S. Yi, R. Melton, J. Stasko, and J. A. Jacko. 2005. Dust & Magnet: Multivariate Information Visualization Using a Magnet Metaphor. *Information Visualization* 4, 4 (Oct. 2005), 239–256. https://doi.org/10.1057/palgrave.ivs.9500099