The Screening Phase in Systematic Reviews: Can we speed up the process?

Igor Rožanc
University of Ljubljana
Faculty of Computer and Information Science
Večna pot 113, SI-1000 Ljubljana, SLOVENIA
igor.rozanc@fri.uni-lj.si

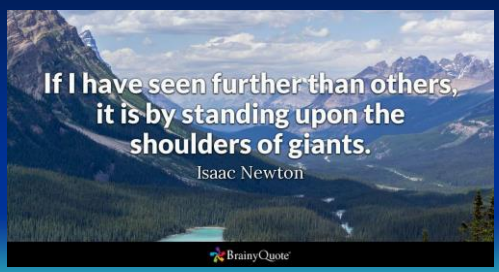
Index

- Introduction
- Background: Evidence Based Research in SE
  - empirical research methods, SLR, SMS
- Problem: effective article screening
  - manual screening issues
- Solution: automatic screening
  - automatic screening, process, tool prototype
- Experiment: SMS on process use in DS(M)L development
  - SMS protocol, article collection, experiment results
- Conclusion

Introduction

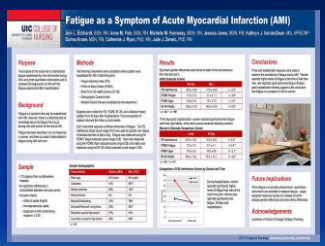
- New research starts with study of empirical evidence in research field
- Structured findings may be published as a review paper
- To be trustworthy, it must be well conduced (guidelines)
- Large portion of manual work is tedious and time-consuming
- Can we speed up the process?

Evidence Based Research



Example: EBR in Medicine

- Tradition in systematic research
- Large corpus of well-structured experience and research work,
- Clear **guidelines**



EBR in Software Engineering

- Diverse research methods/objectives/technologies/…
- Versatile and non-consistent work presentation
- Different guidelines

- Implement solutions from medicine!

Empirical Research Methods

- Zelkowitz and Wallace (1997)*: 12 types divided in 3 groups

1. **Observational methods:** performed during project development
   - Project Monitoring
   - Case study
   - Assertion
   - Field study

- * - M. V. Zelkowitz, D.Wallace, Experimental validation in software engineering, Information and Software Technology 39 (11) (1997).

Empirical Research Methods

2. **Historical methods**: on finished projects
   - Literature review
   - Legacy data
   - Lessons learned
   - Static analysis
3. **Controlled methods**: classical methods
   - Replicated experiment
   - Synthetic experiment
   - Dynamic analysis
   - Simulation

Empirical Research Methods

- Wieringa (2014)*: intended use prospective
  - Expert opinion
  - Single-case mechanism experiment
  - Technical action research
  - Statistical difference-making experiment
  - Observational case study
  - Meta-research method
  - Methods to collect data
  - Techniques to infer information from data

- * - R. Wieringa, Empirical research methods for technology validation: Scaling up to practice, Journal of Systems and Software 95 (2014)

Empirical Research Methods

- Kitchenham (2007)*: broad division

1. Primary study: empirical study of a specific research question
2. Secondary study: integration of several primary studies on specific research question
3. Tertiary study: review of secondary studies on (wider) research question

- * - B. A. Kitchenham, S. Charters, Guidelines for performing Systematic Literature reviews in Software Engineering, Version 2.3, Engineering 45 (4ve) (2007)

Systematic Literature Review SLR

SLR uses a well-defined methodology to identify, analyze and interpret all available evidence related to a specific research question in an unbiased and repeatable way

- Guidelines for SLR (Kitchengham*):
  1. Planning the review
  2. Conducing the review
  3. Reporting the review

* - B. A. Kitchenham, S. Charters, Guidelines for performing Systematic Literature reviews in Software Engineering, Version 2.3, Engineering 45 (4ve) (2007)

Systematic Literature Review SLR

1. Planning the review /1
   - Identification of the need
     - objectives, sources, inclusion/exclusion criteria, quality criteria, data extraction/composition, method to form conclusions from data
   - Commissioning a review
     - commissioning document – review questions, review methods, timetable and budget

Systematic Literature Review SLR

1. Planning the review /2
   - Specifying the research questions
     - SE research: effect, frequency or rate, cost and risk factors, impact on reliability, performance and cost, cost benefit analysis
     - PICOC (Population, Intervention, Comparison, Outcome, Context)

Systematic Literature Review SLR

1. Planning the review /3
   - Developing a review protocol
     - background, research questions, strategy, sources, selection criteria, selection procedures, quality assessment, data extraction strategy, synthesis of the extracted data, dissemination strategy, project timetable
   - Evaluating the review protocol

Systematic Literature Review SLR

2. Conducing the review /1
   - Identification of research
     - digital libraries (IEEExplore, ACM DL, Web Of Science, Google Scholar, Citeseer, Inspec, ScienceDirect, EI Compendex, SpringerLink, SCOPUS), references (snowballing), journals and conference proceedings, gray literature (technical and other reports), research registers, work of specific researchers
     - Bibliography management and document retrieval tools
     - Documented in enough detail (name, search strategy, date of search, years for DL, rationale)

Systematic Literature Review SLR

2. Conducing the review /2
   - Selection of primary studies
     - first screening by formal inclusion/exclusion criteria (language, participants or subjects, research design, sampling method), detailed screening
     - performed by more knowledgeable performers
     - list of excluded studies, documented agreements, consistency check (test-retest)

Systematic Literature Review SLR

2. Conducing the review /3
   - Quality assessment of the studies
     - apply detailed inclusion/exclusion criteria to minimize systematic error and maximize validity
     - initial and detailed quality assessments
     - quality checklists on generic and specific items, bias and validity problems

Systematic Literature Review SLR

2. Conducing the review /4
   - Data extraction and monitoring
     - data extraction forms (set of numerical values), piloted, more researchers, same method, consistency check, exclude duplicates
   - Data synthesis
     - descriptive or quantitative (statistical information), different outcomes (meta-analysis, forest plots)

Systematic Literature Review SLR

3. Reporting the review
   - Specifying dissemination mechanisms
     - academic (and non-academic) journals and/or conferences
   - Formatting the main report
     - technical report or Ph.D. thesis (one section)
   - Evaluating the report
     - peer reviewed

Systematic Mapping Study SMS

SMS is a (broad) review of primary studies in a specific topic area that applies the principle of clustering to identify the evidence available.

Main differences between SLR and SMS:

- SMS has broader, less concrete goal, it classifies items into clusters using statistical characteristics
- SMS uses more methods of data collection/extraction, but they don't require deep understanding of studies
- The number of studies is (much) larger in case of SMS

Systematic Mapping Study SMS

- Guidelines for SMS (Peterson*):

1. Need for the map
2. Study identification
3. Extraction and classification
4. Study validity and presentation

- * - K. Petersen, S. Vakkalanka, L. Kuzniarz, Guidelines for conducting systematic mapping studies in software engineering: An update, Information and Software Technology 64 (2015)

Systematic Mapping Study SMS

1. Need for the map
   - research questions (extent/range/nature of research activity, summarize/disseminate findings, identify research gap)
2. Study identification /1
   - Choosing the search strategy
     - different clusters/authors/years/publishers, start with reasonable number of articles, iterate (snowballing)

Systematic Mapping Study SMS

2. Study identification /2
   - Developing the search phase
     - PICOC too restrictive, use experts, iterative search
   - Evaluate the search
     - Criteria for stop search
   - Inclusion and exclusion phase
     - less restrictive, strategies (objective criteria, resolve disagreements, decision rules)
   - Quality assessment
     - less strict

Systematic Mapping Study SMS

3. Extraction and classification
   - Extraction and classification process
     - several researchers for same evidence, pilot set
   - Topic-independent classification
     - by the venue, research type or method
   - Topic-specific classification
     - classification scheme (keywords)

Systematic Mapping Study SMS

4. Study validity and presentation
   - validity threats (publication bias, poor extraction, researcher bias, low quality of studies, bad aggregation of results, low reliability of conclusions)
   - visualization (usualy bubble plots, bar plots, and pie charts, less line diagram, Venn diagram, heat map)

Systematic Mapping Study SMS

- Practical issues:
  - Large quantity of versatile evidence
  - The quality of studies differs a lot
  - Search engines are not designed for SMS
  - DLs does not contain everything
  - Inconsistent tools for search DLs

Problem: Effective Screening

- Goal 1: collect all (or as much as possible) POSSIBLY suitable studies using different approaches

↓ MANUAL SCREENING

- Goal 2: perform effective SLR/SMS analysis on ACTUALY suitable studies only

Why not suitable?

- Part of collected articles is not suitable:
  - written in wrong format or language
  - too poor article quality
  - duplicate of existing article
  - extended version existing article
  - not (enough) related to research question:
    - Search keywords problem

Issues of Manual Screening

- Screening must be performed in precise and consistent way
- Repeated on different levels of detail
- Replicated by more reviewers
- Unsatisfactory tools
- Limitations of manual work

- **Screening effort is HUGE!!!**

Practical consequences:

- Limited time = collect less articles:
  - Time limited search
  - Limit search to selected sources
  - Limit search questions
  - Deliberately omit part of corpus

↓

Non-selective article loss !!!

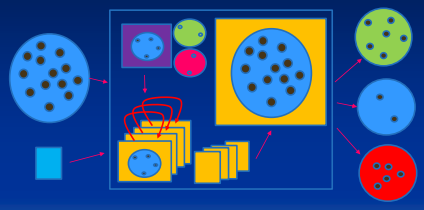
Solution: Automatic Screening

- Automatic replaces manual screening:
  - Less limitations:
    - more articles, not limited on title and abstract
  - Faster:
    - quick performance and repetition, enables gradual tuning
  - More demanding:
    - structured decision (inclusion/exclusion) criteria
    - additional article manipulation
  - Less control:
    - decision without subjective understanding

Basic approach

- Flexible:
  - used for different research fields, topics, questions
  - used (primarily) for SMSs and (possibly) SLRs
  - freely configurable
  - adaptable to different levels of manual involvement
  - use of standard formats + available support tools
- Consistent to guidelines:
  - performers with enough domain knowledge and screening experience
  - Input: set of collected articles + configuration
  - Operation: manual pilot + iterative with adaptive tuning of decision rules
  - Output: sets of included, excluded and possibly included (margin) articles

Idea



Decision-making approach

- Text statistic analysis:
  - Simulation of manual screening
  - Frequency of positive / negative keywords
  - Domain expert defines initial structure rules
  - Rules gradually tuned manually or automatically
  - Dependent of domain expert, but understandable
  - Simpler, yet effective approach
- AI machine learning approach:
  - Uses text statistic data as well
  - Rules derived from past positive/negative decisions
  - Use machine learning approach, expert is not needed
  - Complex, not controllable process
  - In practice, big learning set needed

Text statistic analysis

- Keywords:
  - well selected, crucial for decision making
  - defined by domain expert / screening performer
  - strict : loose analysis: different forms of same word: small/big caps, singular/plural, (,.!)
  - three types of keywords: required, positive, negative
- Groups:
  - synonyms and similar keywords define a group
  - rules may be applied on keywords or groups

Text statistic analysis

- Decision rule:
  - count the actual number of occurrences of a keyword/group in article
    - title + abstract or all text of article
  - compare with predefined criteria =
    (required / negative / positive) numbers of occurrences
  - (in case of more groups) use logical function (of group decisions) to decide

Criteria	Defined on	Number of occurrences	Decision
Required occurrences	keyword, group	less than	EXCLUDE
Negative occurrences	keyword, group	equal or more	EXCLUDE
Positive occurrences	keyword, group	equal or more	INCLUDE

Text statistic analysis

- Included articles:
  - more or equal than ALL required numbers* AND
  - less than ALL negative numbers* AND
  - more or equal than SOME positive number*.
- Excluded articles:
  - less than ANY required number* OR
  - equal or more than SOME negative number* OR
  - less than ALL positive numbers*.
- Possibly included (margin) articles:
  - all remaining articles.

* = for all the keywords/groups

Example:

- Decision = In AND NOT Out

Keyword	Group	Required	Negative	Positive
DSL	In	2		5
language	In			5
digital	Out		2	
ontology	Out		2	
	In	2		5
	Out		2	

Article	DSL	language	digital	ontology	
Article 1	1	0	1	0	EXCLUDED
Article 2	2	0	7	0	EXCLUDED
Article 3	4	3	0	0	INCLUDED

Typical decision rules

- Initial (or too simple) set of rules:
  - No decision: ALL articles in POSSIBLY INCLUDED set
- Too strict (or complicated) set of rules:
  - One sided decision – ALL articles are EXCLUDED (INCLUDED)
- Good (gradually updated) set of rules:
  - Most articles are in correct set, none of suitable ones EXCLUDED, some in POSSIBLY INCLUDED set
- Optimal set of rules:
  - All articles in correct (INCLUDED or EXCLUDED) set

Tuning of decision rules

- Manual tuning requires:
  - knowledgeable and experienced performer
  - deep understanding of research topic
  - careful performance, enough iterations
  - quality check (manual screening results for a pilot set)
  - sensitive decision when to stop
- Automatic tuning:
  - based on manual screening decisions
  - includes assessment of current decision rules
  - several iterations

Automatic tuning

- Screening efficiency depends on:
  - number of articles for (pilot) manual screening
  - number of iterations for decision rules definition
- Automatic tuning of decision rules:
  - (initial) structure of rules is defined by expert
  - LOOP
    - new pilot set is automatically screened
    - decisions are assessed - compared with (correct) manual ones
    - rules (defined numbers of occurrences) are corrected

Automatic decision assessment

- Numeric marking = conformance of automatic and correct (manual) decisions
- (Subjective) principles:
  - same decisions are best, opposite worst
  - better to decide (INCLUDE, EXCLUDE) than not to
  - it is safer to INCLUDE than to EXCLUDE

Automatic screening	Referential screening		
	INCLUDE	POSSIBLY INCLUDE	EXCLUDE
INCLUDE	0	1	6
POSSIBLY INCLUDE	4	3	5
EXCLUDE	7	2	0

Example:

- Result : average mark

Article	Referential screening	Automatic screening	MARK
Article 1	include	include	0
Article 2	possibly include	exclude	2
Article 3	include	possibly include	4
Article 4	exclude	possibly include	5
Article 5	possibly include	possibly include	3
AVERAGE			2,8

Iterative tuning of decision rules

- Optimal set of decision rules = correct structure + consistent decisions with correct screening results
- Iterative correction of (required/positive/negative) numbers
- Manual decisions are used to set (better) criteria
- Practice: reviewers bias – use suitable improvement strategy

Adaptation:	Description:
none	the decision rules are fixed
delicate	all articles with more then X occurrences have the same decision
reasonable	same as before, but possibly included decisions are ignored
strong	same, also up to three articles with opposite decision in row are ignored
radical	same, but up to five articles with opposite decision in row are ignored
brutal	even more than five articles with opposite decision in row are ignored

Example:

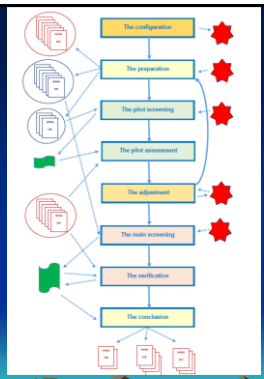
Article	DSL	language	digital	ontology	
Article 1	1	0	1	0	EXCLUDED
Article 2	2	0	7	0	INCLUDED
Article 3	4	3	0	0	EXCLUDED

- Initial:
  - rules: DSL/language/digital/ontology: 0,0,0*
  - decision: PI**, PI**, PI**
  - average mark: 4,67
- Tuning (strategy DELICATE):
  - DSL – 0,0,0* language – 0,0,0* digital – 7,0,0* ontology – 0,0,0*
  - decision: E**, I**, E**
  - average mark: 0,00

* - required number, positeve number, negative number
** - I = INCLUDED, PI= POSSIBLY INCLUDED, E=EXCLUDED,

Implementation

- Automatic screening process:
  - Adaptation of manual screening + specifics
  - Input: PDF (TXT) articles + configuration (paths, pilot size, decision rules)
  - Process: iterative screening on pilot set with manual/automatic + complete set of articles screening
  - Output: 3 separated sets of PDFs + JebRef lists
- Prototype tool:
  - process support,
  - execution of experiment



Screening process

1. Configuration
   - File paths (PDF, TXT), pilot parameters, decision rules (keywords, groups, numbers), XML configuration
2. Preparation
   - Tool (JebRef) initialization, import PDF articles, automatic TXT creation
3. Pilot screening
   - Pilot environment set up, article selection, screening, results (directory, list)

Screening process

4. Pilot assessment
   - Manual screening results insertion, configuration assessment
5. Adjustment
   - Iterative reconfiguration (automatic or manual), possible pilot reset, results saved separately
6. Main screening
   - Complete screening (best configuration), results saved, possibility of repetition with adjusted configuration

Screening process

7. Verification
   - Quality verification (aggregated values)
8. Conclusion
   - All sets of articles accordingly copied, appropriate exports prepared and ZIP archived (economical or full)

Tool prototype

- CONFIGURATOR:
  - Describes configuration, defines file locations, pilot screenings, groups and keywords, decision rules
- PERFORMER:
  - Initial file manipulation, performs pilot and main screenings
- MARKER:
  - Performs assessment of screening results, upgrades decision rules

Experiment: SMS on examining diferent processes while developing a DS(M)L

- Aspects:
  - **The process for the DS(M)L development**
  - The role of the development approach
  - The role of the end user
  - The accompanying tools
  - The development of accompanying tools

SMS protocol

- The research questions
  - RQ 1: Does the development of a DS(M)L follow a defined process? Can it be recognized as the utilization of the specified process?
    - RQ 1.1: Which parts of the process are used? More specifically, is it possible to recognize at least the main phases of the analysis and the design?
  - RQ 2: Which engineering principles are used while developing a DS(M)L?
    - RQ 2.1: How important are the agile principles in the development of a DS(M)L?

SMS protocol

- The research questions
  - RQ 3: What is the role of the end user in the development of a DS(M)L?
  - RQ 4: Is the DS(M)L development actually supported by a specific tool?
    - RQ 4.1: Which kind of tool was developed to support DS(M)L use?

SMS protocol

- The sources
  - Four digital libraries:
    - Science Direct,
    - IEEE eXplore,
    - ACM DL
    - Web of Science
  - Repository of a previous SMS on DSL (U Mb).
  - Manual search
    - snowballing
    - non-systematic

SMS protocol

- The inclusion/exclusion criteria
  - **English language,**
  - **research field: Computer Science,**
  - **published: 2006 – 2016,**
  - **full PDF available,**
  - **journals and conference proceedings only.**
  - DL search:
    - **DL AND (PR OR AD)**
    - **DL = (model-driven engineering OR domain-specific language OR**
    - **domain-specific modeling language OR MDE OR DSL OR DSML)**
    - **PR = (process OR approach OR development)**
    - **AD = (analysis AND design)**

Article collection

Source:	Type:	All found:	Excluded:	Used:
Science Direct	digital library	93	19	74
IEEE eXplore	digital library	365	47	318
ACM DL	digital library	382	42	340
Web of Science	digital library	408	67	341
SMS on DSL's [9]	SMS research	256	15	241
Snowballing	articles reference	24	1	23
Other	manual search	13	0	13
Total		1541	191	1350

- Excluded: non existing PDF, duplicates, poor quality

Manual screening

	Included:	Possibly included:	Excluded:
Number	375	266	699
Percentage	27,78%	19,70%	51,77%

- Two phases:
  - Quick screening of title and abstract (all)
  - Detailed screening of entire text (Included and Possibly Included only)

Experiment

- Randomized experiment
- Complete PDF articles downloaded
- Two initial sets of articles:
  - BIG (n=1350)
  - SMALL (n=76)
- Randomized selection of pilot articles
- More (5) repetitions – average final result

Experiment outcome

- Two measures for quality of decision rules :
- **average mark =**

  Σ of all marks / number of all articles

- **percentage of decisions taken =**

  100 * (number of INCLUDED and EXCLUDED articles / number of all articles)

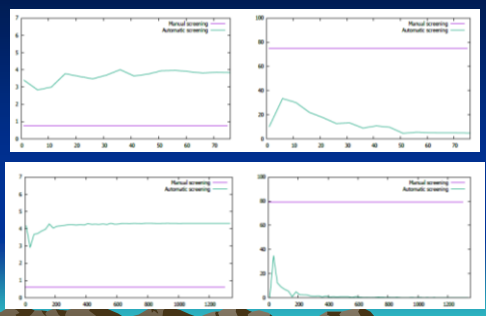
Experiment goals

- Four issues:
  1. The (optimal) size of pilot screening: how many articles should be manually screened to define efficient decision rules?
  2. The improvement strategy: What strategy of decision rules improvement is the most adequate one? What is the effect of using more radical approaches?
  3. The number and the type of different keywords: how many keywords are needed for a delicate enough decision? What is the effect of positive and negative keywords on the efficiency of decision rules?
  4. The grouping of keywords: Does the complexity of the decision rules structure and the use of logical expressions increase the quality of decisions taken?

1. Pilot set size

- **Goal:**
  - find appropriate pilot size using fixed setting
- **Operation:**
  - perform screening on different sizes of randomized pilot sets
- **Setting:**
  - Size: SMALL (n=76), BIG (n=1350), gradually increasing pilot size
  - Rules: simple structure (3 groups,11 keywords, simple logic)
  - Strategy: 5 iterations, reasonable.
- **Result:**
  - average mark
  - percentage of decisions taken

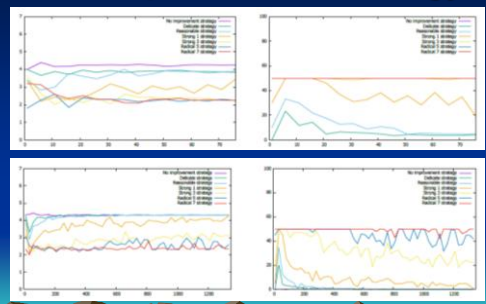
1. Pilot set size



2. Improvement strategy

- **Goal:**
  - find appropriate improvement strategy using fixed setting
- **Operation:**
  - perform screening on different sizes of randomized pilot sets using different improvement strategies
- **Setting:**
  - Size: SMALL (n=76), BIG (n=1350), gradually increasing pilot size
  - Rules: simple structure (3 groups, 11 keywords, simple logic)
  - Strategy: 5 iterations, none, delicate, reasonable, strong (1,3), radical (5,7)
- **Result:**
  - average mark
  - the percentage of decisions taken

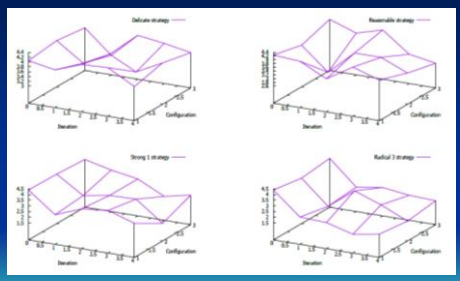
2. Improvement strategy



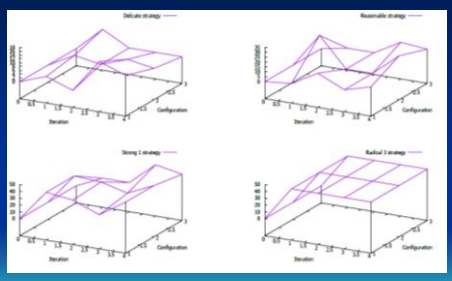
3. Number and type of keywords

- **Goal:**
  - investigate impact of different number of positive (and negative) keywords using fixed setting
- **Operation:**
  - perform screening using different decision rules and different improvement strategies
- **Setting:**
  - Size: SMALL (n=76, pilot=20), BIG (n=1350, pilot=135)
  - Rules: simple structure (3 groups, simple logic),11 (+), 62 (+), 71 (+ -) keywords
  - Strategy: 5 iterations, delicate, reasonable, strong, radical
- **Result:**
  - laverage mark
  - percentage of decisions taken

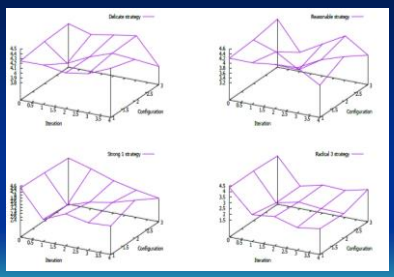
3. Number and type of keywords



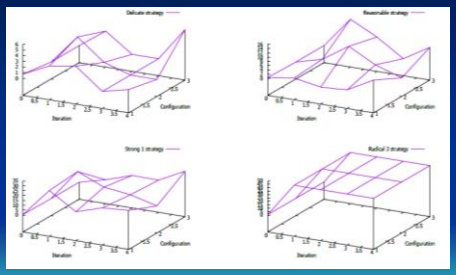
3. Number and type of keywords



3. Number and type of keywords



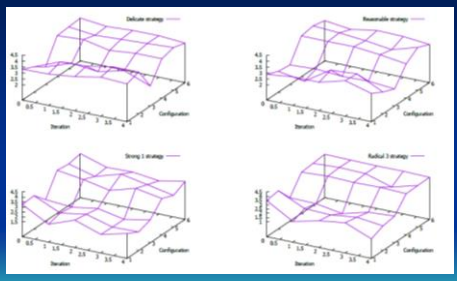
3. Number and type of keywords



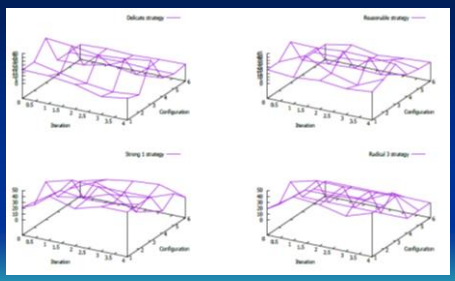
4. Structure of decision rules

- **Goal:**
  - investigate impact of different keyword grouping and logical function complexity
- **Operation:**
  - perform screening using different decision rules and different improvement strategies
- **Setting:**
  - Size: SMALL (n=76, pilot=20), BIG (n=1350, pilot=135)
  - Rules: 11/3 (+) , 62/11 (+), 71/11 (+ -) with simple / complex logic
  - Strategy: 5 iterations, delicate, reasonable, strong, radical
- **Result:**
  - average mark
  - percentage of decisions taken

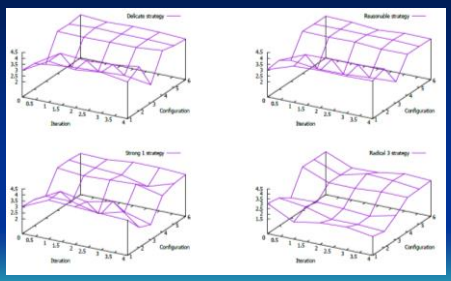
4. Structure of decision rules



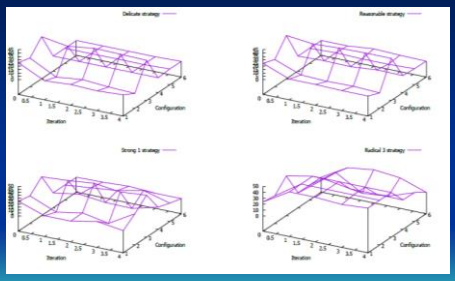
4. Structure of decision rules



4. Structure of decision rules



4. Structure of decision rules



Experiment observations

1. Experiment proved practical usability of our automatic screening approach
2. It can be adapted to a variety of different research goals
3. Detailed insight into research topic is a MUST
4. Experienced reviewer is needed to efficiently perform automatic screening
5. Approach enables notable savings in screening time

Conclusion

- Automatic screening approach is defined, which is:
  - **rigorously designed** to be consistent with the strict SR guidelines
  - implements a **specific combination of carefully selected principles**
  - follows a **highly adjustable screening process**
  - **operational and successful in practice**.

Future work

Main directions:

- **apply text statistic analysis in other SMS phases!!**
- extensive testing using additional criteria
- construction of an efficient (user friendly) tool to support the process
- implementation of additions/corrections of proposed process
- use of proposed approach to effeciently perform SMSs.

The reward for work well done is the opportunity to do more.

Jonas Salk

PICTUREQUOTES.COM

Thank you !

Questions ?

IGOR ROŽANC
Faculty of Computer and Information Science
University of Ljubljana, SLOVENIA
E-mail: igor.rozanc@fri.uni-lj.si

Reference

Igor Rožanc, Marjan Mernik:
The Screening Phase in Systematic Reviews: Can we speed up the process?
Advances in Computers 123 (1) (2021) 116-191.
ISSN 0065-2458