ICS 500: Research Methods and Experiment Design in Computing

Lecture

Systematic Literature Review

The slides are prepared by: Dr. Mahmood Niazi, Dr. Malak Baslyman, Dr. Hamoud Aljamaan & Dr. Mohammad Alshayeb

Lecture Objectives

- ✓ Evidence-based-software engineering
- ✓ What is a Systematic Review?
- ✓ Systematic Review Process
- ✓ SLR example

The problem

- Methods and tools used by software development organizations frequently lack sufficient evidence regarding their applicability, limitations, quality, prices, and inherent risks.
- Suppose you want to find a solution for a RESEARCH PROBLEM. How do you know that:
  - No one has already provided solutions
  - Evidence from the existing literature is correct
  - Methodology is sound
  - Data collected is reliable
  - Results are trustworthy

How do I know that the knowledge is true?

- There are three basic answers to this question:
  1. **Authority** - truth is given to us by someone more knowledgeable than ourselves
     - The authority is God
     - Surah 6: 97 وَهُوَ الَّذِىْ جَعَلَ لَـكُمُ النُّجُوْمَ لِتَهْتَدُوْا بِهَا فِىْ ظُلُمٰتِ الْبَرِّ وَالْبَحْرِ ؕ قَدْ فَصَّلْنَا الْاٰيٰتِ لِقَوْمٍ يَّعْلَمُوْنَ
     - He is the One who made for you the stars, so that you may be guided by them in darkness of the land and the sea. We have elaborated the signs for the people who know.

     - Human authority (the authority is a human expert)
     - Treatment without prevention is simply unsustainable (Dr XYZ)
     - Can also go wrong

How do I know that the knowledge is true?

2. Reason - what is true is that which can be proven using the rules of deductive logic

- Bachelors are unmarried men. Ahmad is unmarried. Therefore, Ahmad is a bachelor.

3. Experience

- Subjective experience
- Influenced by personal feelings, tastes, or opinions, e.g., the user interface of e-desk is user friendly and easy to use
- Empirical evidence
- Includes measurements or data collected through direct observation or experimentation, e.g., 98% of the participants of experiment found that the user interface of e-desk is user friendly and easy to use, i.e., the benchmarked tasks were performed within the allocated time and without any confusion and ambiguity.

The Evidence-Based Paradigm

- Evidence-Based Medicine (EBM) has changed research practices
  - Medical researchers found
    - Clinical judgment of experts can go wrong
    - These judgements should be based on some evidence
- Evidence-Based-Medicine: Integration of best research evidence with clinical expertise and patient values
- Evidence-Based-Software Engineering: Adapted from Evidence-Based Medicine
  - To provide the means by which current best evidence from research can be integrated with practical experience and human values in the decision making process regarding the development and maintenance of software.

Kitchenham et al., 2015

The steps of EBSE

EBSE is a process involving five steps:

- Converting a relevant problem or information need into an answerable question.
- **Searching the literature for the best available evidence to answer the question.**
- Critically appraising the evidence for its validity, impact, and applicability.
- Integrating the appraised evidence with practical experience and the values and circumstances of the customer to make decisions about practice.
- Evaluating performance and seeking ways to improve it.

Why Systematic Reviews?

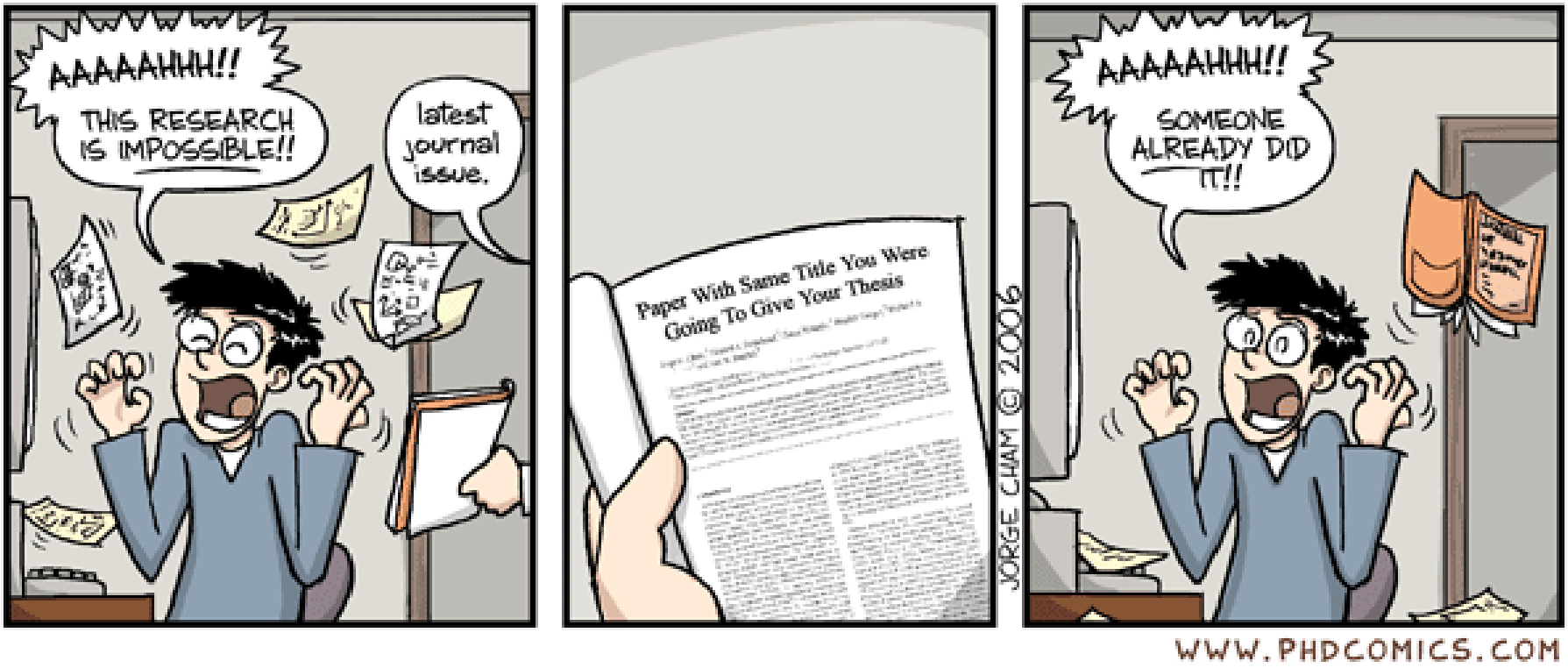
- Systematic review is an important part of evidence-based paradigm

- Systematic reviews aim to synthesize existing research
  - Fairly (evidence-based without bias)
  - Rigorously (according to a defined procedure or protocol)
  - Openly (ensuring that the review procedure is visible to and auditable by other researchers)

- Systematic reviews support repeatability and evolution

- Get a publication!
  - Get your work cited (SLRs are highly cited)
  - Help others to survey the field efficiently through your review

Why Do a Literature Review?



Why Do a Literature Review?

Vitamin C & the common cold

- It was commonly believed that very large doses of Vitamin C prevented the common cold - belief derived from a non-systematic review of literature by Nobel Laureate Linus Pauling (1986)
- An exhaustive systematic review of the literature by Knipschild & colleagues concluded that even mega doses of Vitamin C cannot prevent a cold
- Pauling's review missed 5 of the 'top 15' published studies (even though he was an expert)

What is a Systematic Review?

- Systematic (literature) review is a particular form of secondary study and aims to provide an objective and unbiased approach to finding relevant primary studies, and for extracting, aggregating and synthesizing the data from these.

Kitchenham et al., 2015

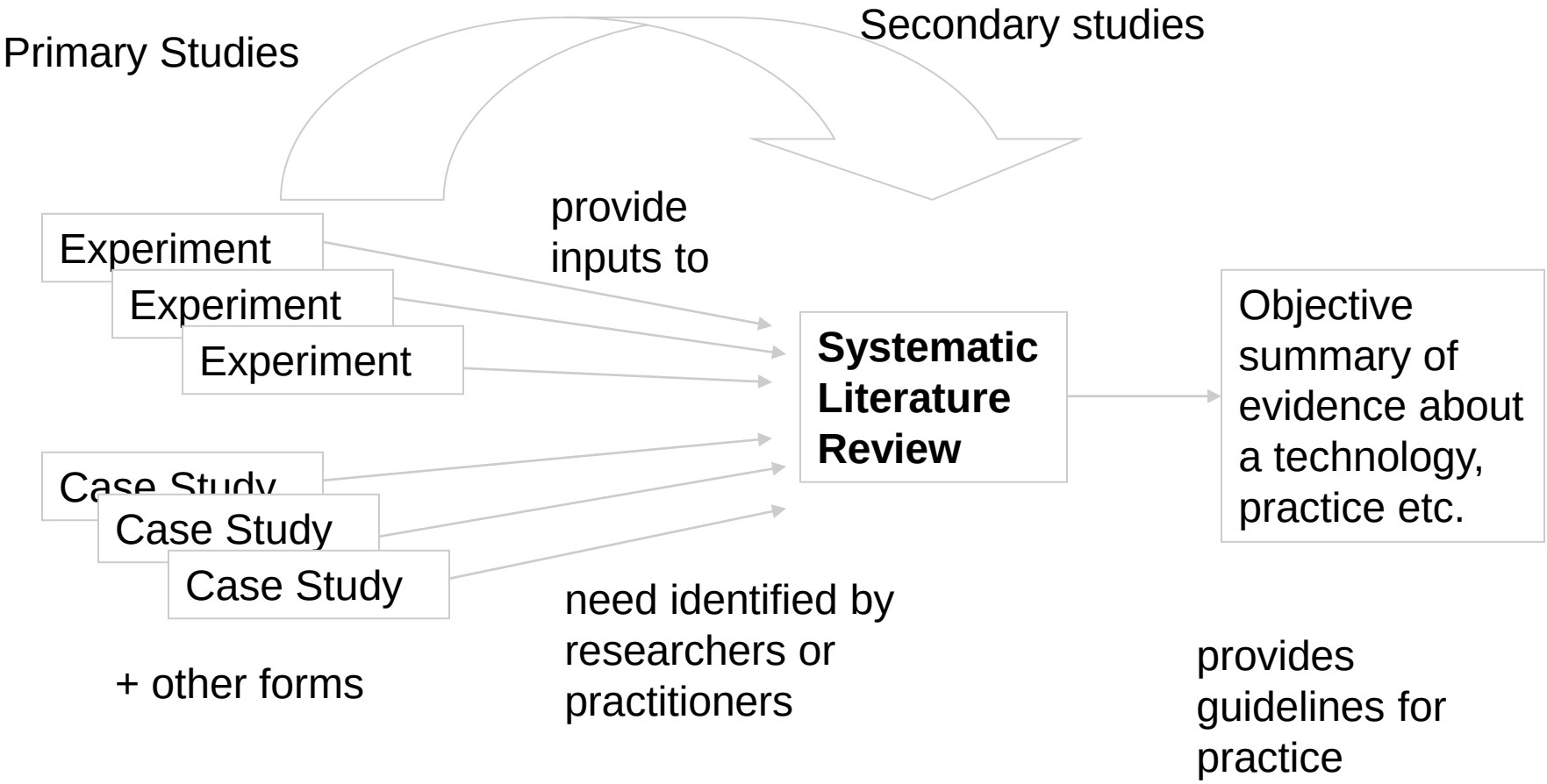
- A systematic review is a defined and methodical way to summarise the empirical evidence concerning a treatment or technology, to identify missing areas in current research or to provide background in order to justify new research.

- A systematic review proceeds by **identifying**, **assessing** and **analyzing** published primary studies relating to a specific research question.

- Systematic reviews require considerably **more effort** than conventional literature reviews but provide a much stronger basis for making claims about research questions.

Kitchenham and Charter, 2007

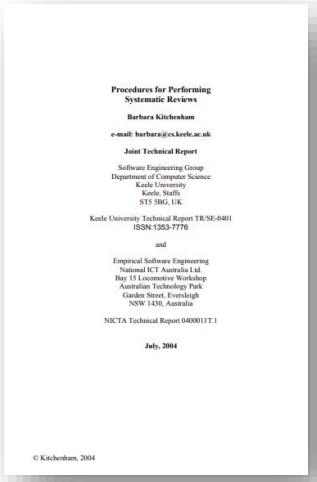
What is a Systematic Review?

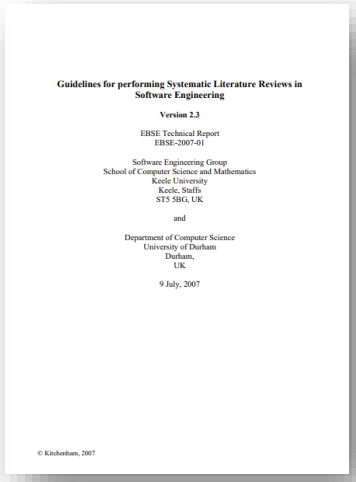


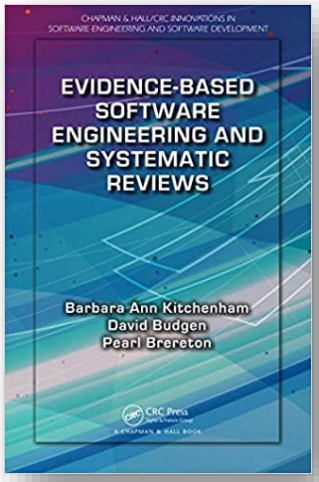
Kitchenham and Charter, 2007

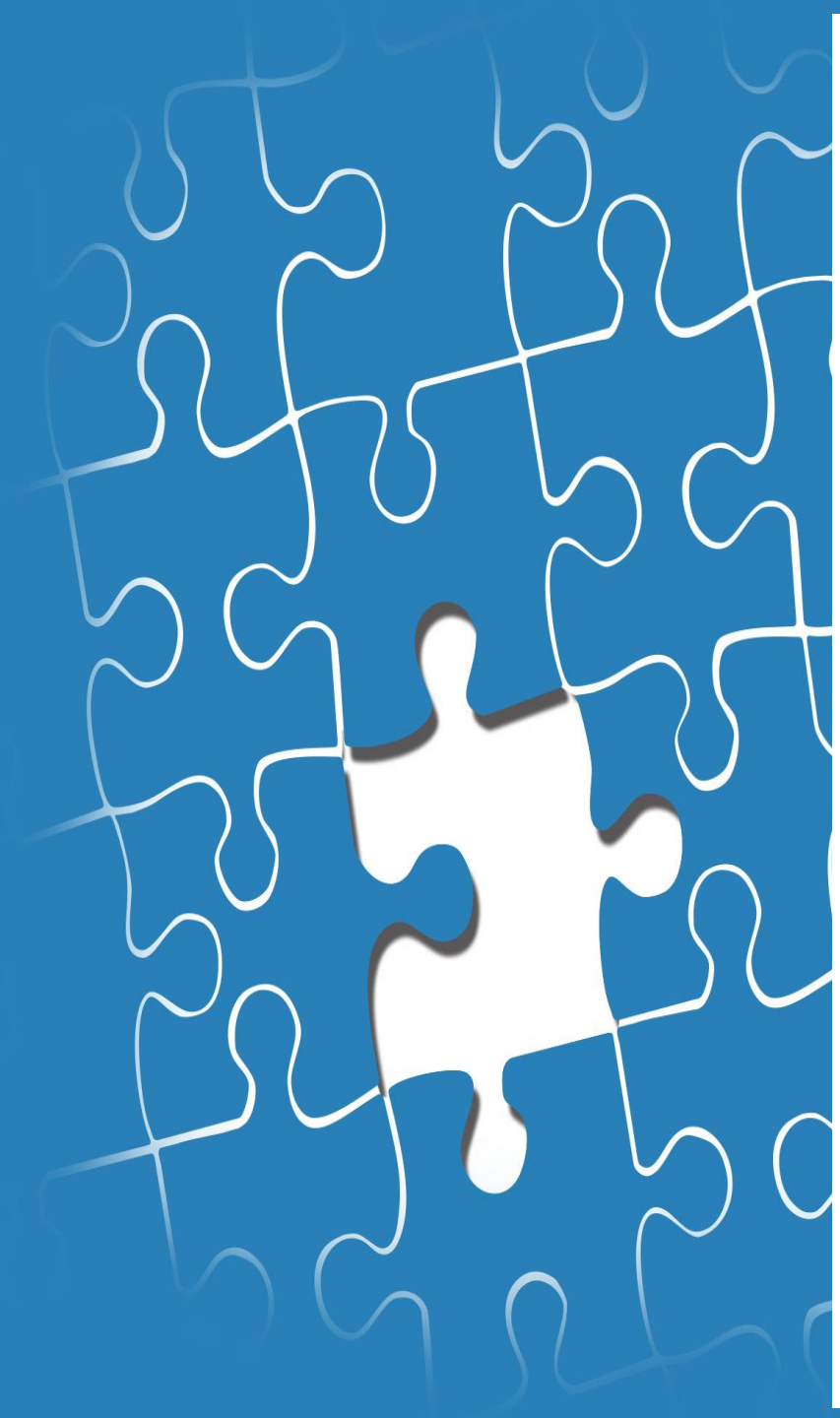
Guidelines

- SLR guidelines
  - B.A. Kitchenham, Procedures for Performing Systematic Reviews, Keele University Technical Report TR/SE-0401, ISSN:1353-7776
  - B.A. Kitchenham, S. Charters, Guidelines for performing systematic literature reviews in software engineering, Tech. Rep. EBSE-2007-01, Keele University and University of Durham, 2007.
  - B.A. Kitchenham, D. Budgen, P. Brereton, Evidence-Based Software Engineering and Systematic Reviews, CRC, 2015









Phase 1
Plan Review

1. Specify RQ(s)
2. Develop Protocol
3. Validate Protocol

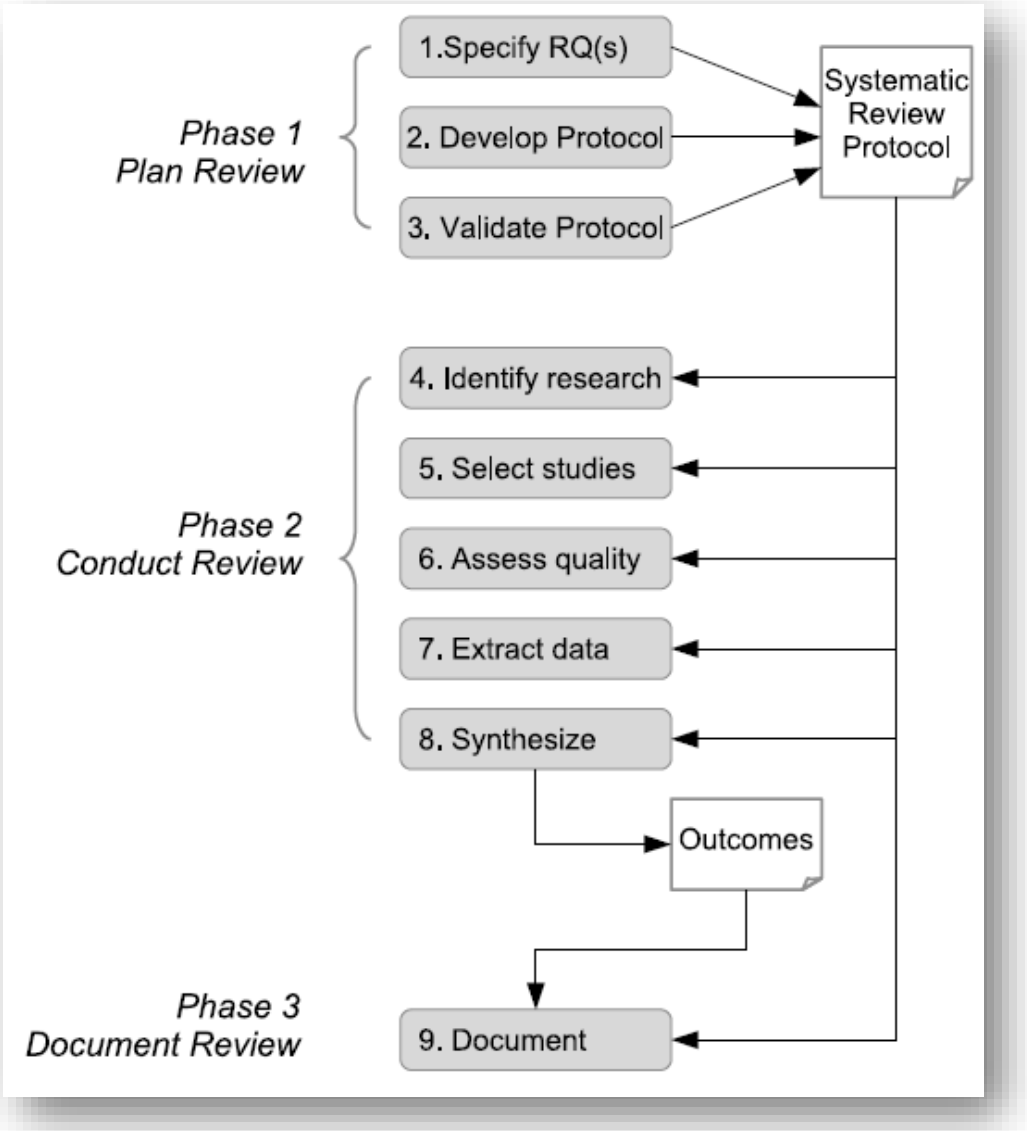
Phase 2
Conduct Review

4. Identify research
5. Select studies
6. Assess quality

Phase 3
Document Review

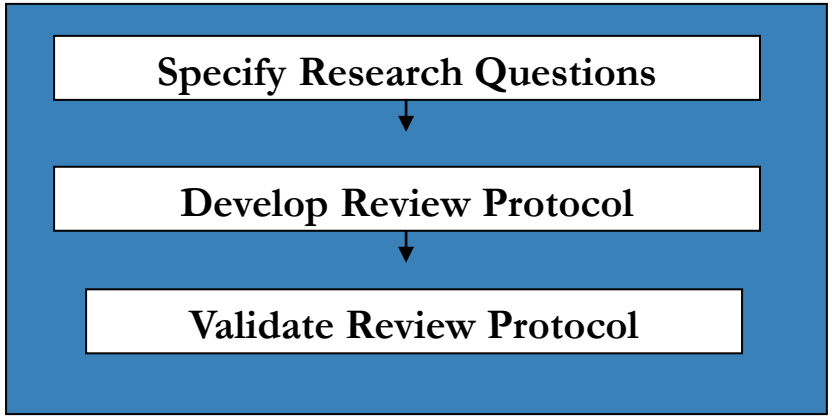
7. Extract data
8. Synthesize
9. Document

Systematic Review Process



Kitchenham et al., 2015

Systematic Review Process - Planning



Research Question(s)

- Starting point of standard SLRs
- Determines relevant primary studies
- Guides the search process
- Helps specification of
  - Search strings
  - Inclusion/exclusion criteria
  - Data collection process

Research Question(s) Examples 1

- RQ1. What factors are important for establishing trust in offshore software outsourcing relationships?
- RQ2. What factors are important for maintaining and strengthening trust in offshore software outsourcing relationships?
- RQ3. What factors do offshore software outsourcing vendors need to address in order to have a positive impact on software outsourcing clients?

Research Question(s) Examples 2

	Research Question	Motivation
RQ1	*What is the status of design and code smells detection?* RQ1.1: What are the statistics of the PSes in each year? RQ1.2: What are the statistics of the PSes in each database?	• Capture the smell detection research trend over years. • Identify the scholarly databases with most publications related to this problem.
RQ2	*What approaches and methods are used by the PSes to detect design and code smells?* RQ2.1: What are the categories of detection techniques? RQ2.2: What are the techniques investigated in each PS? RQ2.3: What is the input type used by each detection technique? RQ2.4: What are the intermediate representations used by each detection technique? RQ2.5: What are the performance measures used? RQ2.6: What are the targeted programing languages?	• Detail every approach features and characteristics to present deeper comparison between all the proposed techniques • Extract the most widely used approaches on smell detection. • Group the smell detection techniques based on some common features. • List the main input type, transformations and targeted programming languages in each classified technique. • Summarize the different performance measures adopted by each detection technique.
RQ3	*What are the targeted software development levels (code and design)?* What are the percentages of code and design levels? What are the targeted design diagrams?	• Find out the levels that received the most and least attention in the detection technique to give more focus on the level with least attention. • Identify the research gap based on these two levels.
RQ4	*What are the targeted design and code smells?*	• Summarize the list of design and code smells investigated mostly by researchers to focus on other smells in future research.
RQ5	*What are the metrics and quality attributes used in each PS?* What is the ratio of PSes dealing with quality attributes? What is the list of quality attributes in each PS? What is the ratio of PSes using software metrics to detect bad smells fully or partially? What is the list of software metrics used by each PS?	• Quality attributes and their corresponding software metrics represent a major component of the software quality assessment process. • Enumerate, the most used software metrics in the detection of bad smells and their corresponding quality attributes, to explore their effect in this process.
RQ6	*What are the techniques implemented into a software tool?* What is the detection automation percentage? What is the ratio of PSes that proposed a detection tool along with their detection technique? What is the list of the proposed tools? What are the ratios of standalone and plugin detection tools.?	• For software managers and developers, the tools represent the main concern as they help them in exploring the software product status. • Thus, we investigate the ratio of proposed detection techniques implemented as a tool.
RQ7	*What is the current status of the validation and datasets?* RQ7.1: What is the ratio of validation of empirical/experimental techniques versus case studies? RQ7.2: What is the available information and type of the validation dataset? RQ7.3: What is the validation dataset size and programing language? RQ7.4: What are the validation dataset project names and programing languages?	• The use of a suitable validation dataset is crucial to validate the detection process and its underlying methods. • This question addresses the validation strategies adopted by the researchers. The answer will result in identifying the validation datasets and their properties. • Such findings will enable equal-foot comparison and benchmarking between all detection techniques considered.

AbuHassan et al., 2020

What is a Review Protocol?

- A systematic review protocol specifies the methods that will be used to undertake a specific systematic review.

- Predefined protocol
  - Reduces researcher bias
    - Selection of papers driven by researcher expectations
    - Changing the research question to fit the results of the searches
  - Support reproducibility
  - Good practice for any empirical study

Template for a Systematic Review Protocol

Template for a Systematic Review Protocol

1. Change Record

This should be a list or table summarizing the main updates and changes embodied in each version of the protocol and (where appropriate), the reasons for these.

2. Background

a) explain why there is a need for a study on this topic
b) specify the main research question being addressed by this study
c) specify any additional research questions that will be addressed
d) if extending previous research on the topic, explain why a new study is needed

3. Search Process

a) specify and justify basic strategy: manual search, automated search, or mixed
b) for automated searches, specify search terms and compounds of these and record results of any prototyping of the search strings
c) for automated searches, identify resources to be used (specifying the digital libraries and search engines)
d) for manual searches, identify the journals and conferences to be searched
e) specify the time period to be covered by the review and any reasons for your choice
f) identify any ancillary search procedures, for example, asking leading researchers or research groups, or accessing their web sites; or checking reference lists of primary studies
g) specify how the search process is to be evaluated (for example, against a known subset of papers; or against the results from a previous systematic review)

4. Primary Study Selection Process

a) identify the *inclusion* criteria for primary studies
b) identify the *exclusion* criteria
c) define how selection will be undertaken (roles of reviewers)
d) define how agreement among reviewers will be evaluated
e) define how any differences between reviewers will be resolved

5. Study Quality Assessment Process

a) specify the quality checklists to be used
b) specify how the checklist will be evaluated (if a new checklist has been developed)
c) define how agreement among data extractors will be evaluated
d) define how any differences between data extractors will be resolved
e) identify the procedures to use for applying the checklists, such as details inclusion/exclusion, partitioning the primary studies during aggregation or meta-analysis, and explaining the results of primary studies

6. Data Extraction Process

a) design data extraction form (and check via a dry run)
b) specify the strategy for extracting and recording the data (for example, paper form, on-line. Form or database)
c) identify how the data extraction process is to be undertaken and validated, particularly any data that require numerical calculations, or are subjective

7. Data Synthesis Process

a) specify the form of analysis/synthesis to be used (for example, narrative, tabulation, meta-analysis)
b) discuss how the synthesis will be validated

8. Study Limitations

a) assess the threats to validity (construct, internal, external), particularly constraints on the search process and deviations from standard practice
b) specify residual validity issues including potential conflicts of interest that are inherent in the context of the study, rather than arising from the plan

9. Reporting

a) identify target audience, relationship to other studies, planned publications, authors of the publications
b) agree in advance who will be included in the list of authors and whose assistance will reported in the acknowledgements section.

10. Schedule

Provide time estimates for all of the major steps.

Kitchenham et al., 2015

Protocol Contents -1/2

- Background
  - Rationale for survey
- Research question
  - The most important activity during protocol is to formulate the research questions
- Strategy to find primary studies
  - Search terms, resources, databases, journals, conferences

Protocol Contents – 2/2

- Quality assessment criteria
  - Criteria used to evaluate quality of primary sources
- Data extraction strategy
  - How information will be extracted from primary sources
  - Where data will be stored
- Procedures for data synthesis
  - Means of summarising data
- Threats/limitations
- Timetable for research

Review Protocol Validation

- The protocol is a critical element of any systematic review

- Researchers must agree a procedure for reviewing the protocol. A group of independent experts should be asked to review the protocol. The same experts can later be asked to review the final report
  - Experts in SE technology
  - Experts in empirical methods

- PhD students should present their protocol to their supervisors for review and criticism.

Review Protocol Validation

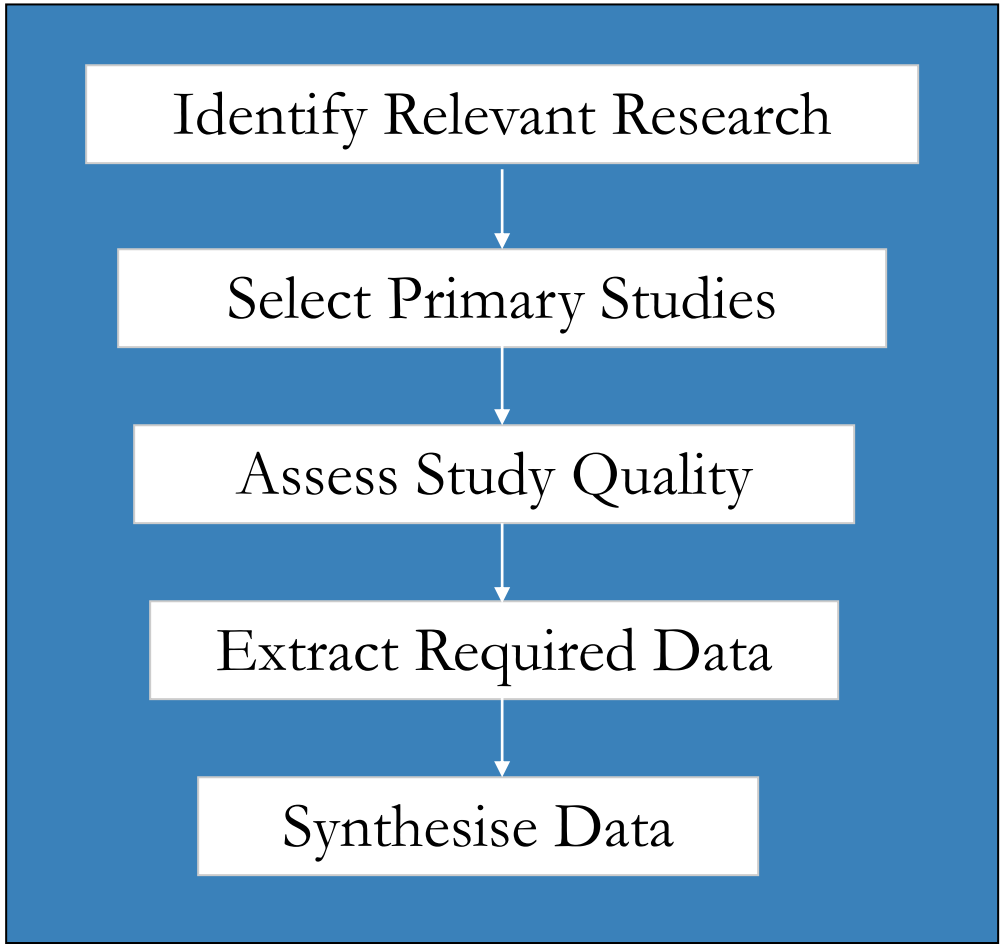
- Protocol elements should be tested during protocol construction

- Expect to revise/refine
  - Questions
  - Search Strategies
  - Data extraction forms
  - Data aggregation process

- Protocols are often long documents
  - 20 pages or more
  - Independent review helps
  - Basis for final report

- Sample protocol:
  - https://www.crd.york.ac.uk/PROSPEROFILES/3611_STRATEGY_20130031.pdf
  - https://madeyski.e-informatyka.pl/download/slr/EquivalentMutantsSLRProtocol.pdf

Conducting the Review



Identify Relevant Research: *Search Terms*

- Choose 4-5 papers in the field under study
  - Choose highly cited papers
  - Or recommended by experts in the domain

- Collect data to be used in the research query
  - Keywords (and indexed keywords in the database)
  - Terminologies
  - Journals/ conferences it came from

- Choosing relevant search terms benefit from:
  - Preliminary searches
  - Trial searchers using various combinations of search terms
  - Reviews of research results
  - Consultations with experts in the field

Identify Relevant Research: *Search Queries (1/2)*

1. Use the Research Questions for the derivation of major terms (RQ1. What factors are important for establishing trust in offshore software outsourcing relationships?)
   - offshore software outsourcing relationships, factors, establishing, trust

2. For these major terms, find the alternative spellings and synonyms
   - **Trust**: (Trust OR trustworthy OR trustworthiness OR trusted OR reliance OR reputation OR satisfaction OR reliable OR reliability OR best performance OR "expectations match" OR responsiveness OR "good track record")
   - **Factors**: (factors OR drivers OR motivators OR elements OR characteristics OR parameters)

Identify Relevant Research: *Search Queries (2/2)*

3. Verify the keywords in any relevant paper

4. Use of Boolean Operators for conjunction if the database allows, in such a way, to use 'OR' operator for the concatenation of alternative spellings and synonyms whereas 'AND' for the concatenation of major term
   - ((Trust OR trustworthy OR trustworthiness OR trusted OR reliance OR reputation OR satisfaction OR reliable OR reliability OR best performance OR "expectations match" OR responsiveness OR "good track record")
   - AND
   - (Factors OR drivers OR motivators OR elements OR characteristics OR parameters))

Potential Search Databases

- Scopus (http://www.scopus.com/)
- Web of Science (https://webofknowledge.com/)
- IEEE Xplore (http://ieeexplore.ieee.org/)
- ACM Digital Library (http://dl.acm.org/)
- SpringerLink (http://www.springerlink.com/)
- Google Scholar (http://scholar.google.com/)
- Elsevier (https://www.elsevier.com)

Select Primary Studies: *Selection Criteria*

- Define study selection criteria
  - Selection criteria determine whether a piece of literature found by the search terms is included or excluded from the systematic review
  - The selection criteria should be decided during the protocol definition.
  - Inclusion and exclusion criteria should be based on the research question. They should be piloted to ensure that they can be reliably interpreted and that they classify studies correctly.
  - Selection criteria could be identified according to:
    - Geographic location of study/ date
    - Type of publication
    - Relevance to interests
    - Language
    - Peer reviewed
    - Etc.

Selection Process

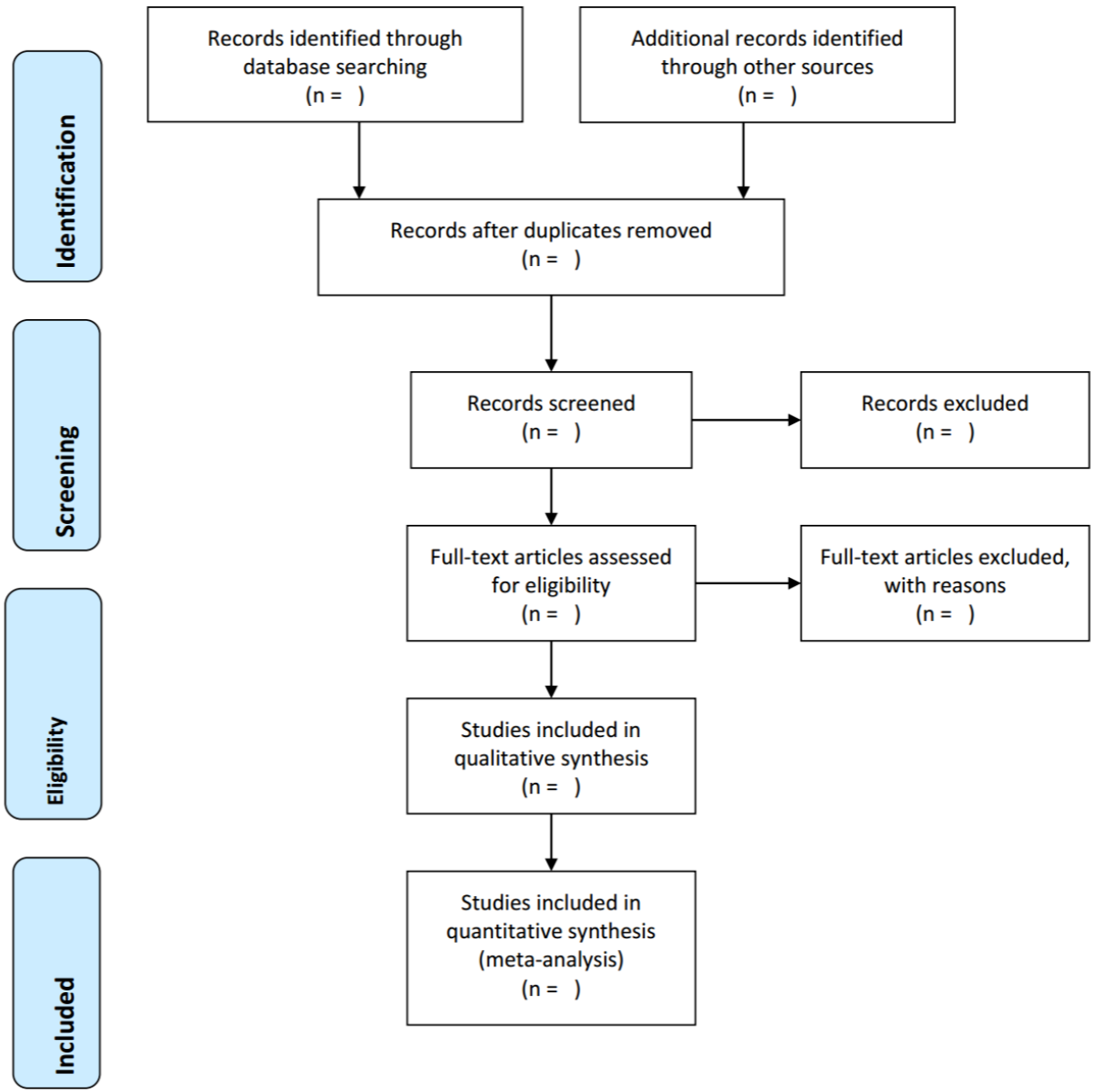
- Selection process
  - Study selection is a multistage process
  - The first phase uses all publications found by the search results archive, and is based on reading the titles and abstracts of publications, to see if they can be trivially excluded.
  - The second phase uses all publications selected during the first phase, and is based on a more thorough reading to identify that each publication contain information relevant to the selection criteria for this study.
  - Two or more researchers read full text.
  - Any disagreement between researchers must be resolved.
  - Maintain list of excluded studies with reasons.

Summary of Selection Results

PRISMA 2009 Flow Diagram

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). *P*referred *R*eporting *I*tems for *S*ystematic Reviews and *M*eta-*A*nalyses: The PRISMA Statement. PLoS Med 6(6): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit www.prisma-statement.org.



Snowballing

- Manual searching of selected journals and conference proceedings
  - Backwards snowballing: Checking papers that *are cited* in the papers included in a review
  - Forwards snowballing: Checking papers that *cite* the papers included in a review.

- The search strategy will aim to achieve an acceptable level of *completeness*

Quality assessment

- In addition, to general inclusion exclusion criteria, it is generally considered important to assess the "quality" of primary studies.

- What is quality in this context?
  - Clear articulation of objectives and study design
  - Extent to which primary study
    - Minimises bias: Tendency to produce results that depart systematically from the "true" results
    - Maximises internal validity: Extent to which design & conduct of experiment prevent systematic error.
    - Maximises external validity: Extent to which results are applicable outside the study context

Data Extraction

- The data extraction forms must be designed to collect all the information needed to address the review questions and the study quality criteria.
  - Reference details
  - Information to address research question
  - Quality data

- Pilot data extraction forms when constructing protocol

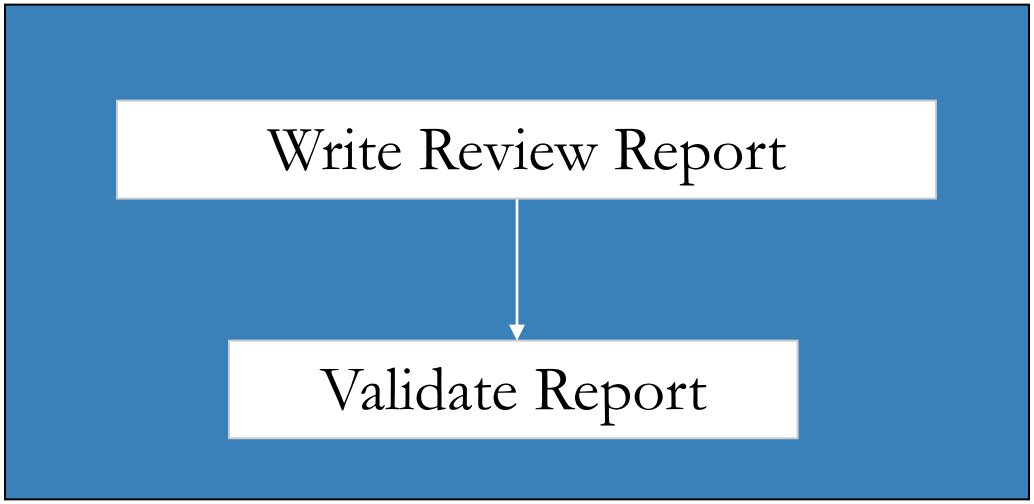
- Several researchers should extract data from each study
  - Disagreements must be resolved

Data Synthesis

- Data synthesis involves collating and summarising the results of the included primary studies.
  - In accordance with research questions
  - Structure tables to highlight similarities and differences between studies
  - To indicate possible causes of differences
  - Produces new knowledge, futuristic research trends, gaps, etc.

- The data synthesis activities should be specified in the review protocol. However, some issues cannot be resolved until the data is actually analysed.

Documenting the Review



Review Report – 1/3

- Content
  - Title, Authorship
  - Executive summary or structured abstract
  - Background
    - Why the review is needed
  - Review Methods
    - Data sources & search strategy
    - Criteria selection
    - Quality assessment
    - Data Extraction
    - Data Synthesis

Review Report - 2/3

- Contents continued
  - Included & excluded studies
  - Results
    - Findings
  - Discussion
    - Principal findings
    - Strengths & weaknesses
  - Conclusions & Recommendations
  - Acknowledgements
  - Conflicts of Interest
  - References & appendices

Review Report – 3/3

- Review reports should be reviewed
  - Expert panel if possible
- Should establish means of publicising results
  - Journal papers
  - Web publication
- Should have mechanism for responding to comments

Mapping Studies

- A form of secondary study intends to identify and classify the set of publications on a topic.
- The goal of a mapping study is to survey the available knowledge about a topic.
- Used to identify ‘evidence gaps’ where more primary studies are needed as well as ‘evidence clusters’ where it may be practical to perform a systematic review.

Systematic Reviews vs. Systematic Mapping

Systematic Mapping Review

- Provide an overview of a research area, and identify the amount, the type of research and results available
- Map the frequencies of publication over time to see trends
- Identify forums and relevant authors in which research in the area has been published
- Use – when you are beginning and are not sure about the area you want to do your research
- Advantages – it can be published, and may be a chapter of your thesis
- Effort - High

Systematic Literature Review

- Provide a status of the research being done on a specific research area, and identify the efforts being done to answer the open questions of this research area
- Identify authors that are working or have worked with the topic and at what point their research is published
- Use – when you already know your research area
- Advantages – if done properly can be published in a journal, and it will definitely be a thesis chapter
- Effort – VERY HIGH

Maíra Marques Samary, How to choose: Survey? Systematic Mapping Review? Systematic Literature Review? Snowballing?

Similarities and Differences Between Systematic Literature Reviews and Systematic Mapping (1/3)

- Focus of the review
  - SLR is to aggregate primary studies in terms of its results and investigate whether these results are consistent or contradictory.
  - SMs provide a broader view of a research topic and identify both clusters (group of studies related to a same theme – may be suitable to undertake an SLR on this theme).
- Research Questions
  - SLRs focus on very narrow questions
  - SMs focus on broad questions, SMs have an objective to find and classify primary studies in subtopics
- Methods for searching
  - Both SLRs and SMs attempt to be exhaustive in finding all relevant studies
  - SLRs generally look at one type of evidence, i.e., empirical studies,
  - SMs may include many different types of primary research, i.e., empirical studies, technical reports, theses, among others.

Napoleão, B.M., Felizardo, K.R., de Souza, E.F., & Vijaykumar, N. (2017). Practical similarities and differences between Systematic Literature Reviews and Systematic Mappings: a tertiary study, The 29th International Conference on Software Engineering and Knowledge Engineering, pp. 85-90, DOI: http://dx.doi.org/10.18293/SEKE2017-069

Similarities and Differences Between Systematic Literature Reviews and Systematic Mapping (2/3)

- Methods for selecting
  - The scope of an SM is broader and the analysis and synthesis more general than in an SLR.
  - SMs involve more studies to be selected while SLRs involve fewer studies, but they should be analyzed in a greater depth.
- Methods for data extraction
  - For undertaking an SM the data to be collected from the primary studies could include: (i) bibliographic information on the publications in which the primary studies are reported, e.g. Journal title, publication year; and (ii) basic data to describe what research has been done and how, e.g. country of the study, technique used, among others.
  - For an SLR the data one collects from the primary studies could include: (i) detailed data on the methods and results of each study; (ii) a structured description of each study; (iii) the results (findings) of each study.

Napoleão, B.M., Felizardo, K.R., de Souza, E.F., & Vijaykumar, N. (2017). Practical similarities and differences between Systematic Literature Reviews and Systematic Mappings: a tertiary study, The 29th International Conference on Software Engineering and Knowledge Engineering, pp. 85-90, DOI: http://dx.doi.org/10.18293/SEKE2017-069

Similarities and Differences Between Systematic Literature Reviews and Systematic Mapping (3/3)

- Synthesis
  - SMs simply describe basic details about each primary study and variables can be used in coding the studies. The description may include, e.g., methods used, geographical distribution of the studies, year of the publication. SMs tabulate primary studies into categories. SMs may synthesize all, or part of the research studies described in a map.
  - In SLRs the main interest is a full synthesis of results.

Napoleão, B.M., Felizardo, K.R., de Souza, E.F., & Vijaykumar, N. (2017). Practical similarities and differences between Systematic Literature Reviews and Systematic Mappings: a tertiary study, The 29th International Conference on Software Engineering and Knowledge Engineering, pp. 85-90, DOI: http://dx.doi.org/10.18293/SEKE2017-069

SLR example

SLR Example

- Performed an SLR on ML models published in the period from 1 January 1991 to 31 December 2010.

Information and Software Technology 54 (2012) 41–59

Contents lists available at SciVerse ScienceDirect

Information and Software Technology

journal homepage: www.elsevier.com/locate/infsof

Systematic literature review of machine learning based software development effort estimation models

Jianfeng Wen [a,*], Shixian Li [a], Zhiyong Lin [b], Yong Hu [c], Changqin Huang [d]

[a] Department of Computer Science, Sun Yat-sen University, Guangzhou, China
[b] Department of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, China
[c] Institute of Business Intelligence and Knowledge Discovery, Department of E-commerce, Guangdong University of Foreign Studies, Sun Yat-sen University, Guangzhou, China
[d] Engineering Research Center of Computer Network and Information Systems, South China Normal University, Guangzhou, China

ARTICLE INFO

Article history:
Received 28 October 2010
Received in revised form 8 August 2011
Accepted 8 September 2011
Available online 16 September 2011

Keywords:
Software effort estimation
Machine learning
Systematic literature review

ABSTRACT

Context: Software development effort estimation (SDEE) is the process of predicting the effort required to develop a software system. In order to improve estimation accuracy, many researchers have proposed machine learning (ML) based SDEE models (ML models) since 1990s. However, there has been no attempt to analyze the empirical evidence on ML models in a systematic way.
Objective: This research aims to systematically analyze ML models from four aspects: type of ML technique, estimation accuracy, model comparison, and estimation context.
Method: We performed a systematic literature review of empirical studies on ML model published in the last two decades (1991–2010).
Results: We have identified 84 primary studies relevant to the objective of this research. After investigating these studies, we found that eight types of ML techniques have been employed in SDEE models. Overall speaking, the estimation accuracy of these ML models is close to the acceptable level and is better than that of non-ML models. Furthermore, different ML models have different strengths and weaknesses and thus favor different estimation contexts.
Conclusion: ML models are promising in the field of SDEE. However, the application of ML models in industry is still limited, so that more effort and incentives are needed to facilitate the application of ML models. To this end, based on the findings of this review, we provide recommendations for researchers as well as guidelines for practitioners.

Contents

* Corresponding author. Tel.: +86 20 34022695.
E-mail address: wjfsysu@gmail.com (J. Wen).

0950-5849/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.infsof.2011.09.002

Abstract

A B S T R A C T

Context: Software development effort estimation (SDEE) is the process of predicting the effort required to develop a software system. In order to improve estimation accuracy, many researchers have proposed machine learning (ML) based SDEE models (ML models) since 1990s. However, there has been no attempt to analyze the empirical evidence on ML models in a systematic way.

Objective: This research aims to systematically analyze ML models from four aspects: type of ML technique, estimation accuracy, model comparison, and estimation context.

Method: We performed a systematic literature review of empirical studies on ML model published in the last two decades (1991–2010).

Results: We have identified 84 primary studies relevant to the objective of this research. After investigating these studies, we found that eight types of ML techniques have been employed in SDEE models. Overall speaking, the estimation accuracy of these ML models is close to the acceptable level and is better than that of non-ML models. Furthermore, different ML models have different strengths and weaknesses and thus favor different estimation contexts.

Conclusion: ML models are promising in the field of SDEE. However, the application of ML models in industry is still limited, so that more effort and incentives are needed to facilitate the application of ML models. To this end, based on the findings of this review, we provide recommendations for researchers as well as guidelines for practitioners.

Review Protocol

Research Questions

2.1. Research questions

This SLR aims to summarize and clarify the empirical evidence on ML based SDEE models. Towards this aim, five research questions (RQs) were raised as follows.

(1) *RQ1: Which ML techniques have been used for SDEE?*
RQ1 aims at identifying the ML techniques that have been used to estimate software development effort. Practitioners can take the identified ML techniques as candidate solutions in their practice. For ML techniques that have not yet been employed in SDEE, researchers can explore the possibility of using them as potential feasible solutions.

(2) *RQ2: What is the overall estimation accuracy of ML models?*
RQ2 is concerned with the estimation accuracy of ML models. Estimation accuracy is the primary performance metric for ML models. This question focuses on the following four aspects of estimation accuracy: accuracy metric, accuracy value, data set for model construction, and model validation method.

(3) *RQ3: Do ML models outperform non-ML models?*
In most of the existing studies, the proposed ML models are compared with conventional non-ML models in terms of estimation accuracy. RQ3 therefore aims to verify whether ML models are superior to non-ML models.

(4) *RQ4: Are there any ML models that distinctly outperform other ML models?*
The evidence of comparisons between different ML models can be synthesized to determine which ML models consistently outperform other ML models. Thus, RQ4 aims to identify the ML models with relatively excellent performance.

(5) *RQ5: What are the favorable estimation contexts of ML models?*
RQ5 aims at identifying the strengths and weaknesses of different ML models. With fully understanding the characteristics of the candidate ML models, practitioners can make a rational decision on choosing the ML models that favor the focused estimation contexts.

Search Strategy

- Search Terms

> software AND (effort OR cost OR costs) AND (estimat* OR predict*) AND (learning OR "data mining" OR "artificial intelligence" OR "pattern recognition" OR analogy OR "case based reasoning" OR "nearest neighbo*" OR "decision tree*" OR "regression tree*" OR "classification tree*" OR "neural net*" OR "genetic programming" OR "genetic algorithm*" OR "bayesian belief network*" OR "bayesian net*" OR "association rule*" OR "support vector machine*" OR "support vector regression").

- Resources

> The literature resources we used to search for primary studies include six electronic databases (*IEEE Xplore*, *ACM Digital Library*, *ScienceDirect*, *Web of Science*, *EI Compendex*, and *Google Scholar*) and one online bibliographic library (*BESTweb*). Some other impor-

Search Strategy

- Search process

  - Search phase 1: Search the six electronic databases separately and then gather the returned papers together with those from BESTweb to form a set of candidate papers.
  - Search phase 2: Scan the reference lists of the relevant papers to find extra relevant papers and then, if any, add them into the set.

Study selection

- Selection phase 1: Apply the inclusion and exclusion criteria (defined below) to the candidate papers so as to identify the relevant papers, which provide potential data for answering the research questions.
- Selection phase 2: Apply the quality assessment criteria (defined in the next section) to the relevant papers so as to select the papers with acceptable quality, which are eventually used for data extraction.

Study selection

Inclusion criteria:

- Using ML technique to estimate development effort.
- Using ML technique to preprocess modeling data.
- Using hybrid model that employs at least two ML techniques or combines ML technique with non-ML technique (e.g., combining with statistics method, fuzzy set, or rough set) to estimate development effort.
- Comparative study that compares different ML models or compares ML model with non-ML model.
- For study that has both conference version and journal version, only the journal version will be included.
- For duplicate publications of the same study, only the most complete and newest one will be included.

Exclusion criteria:

- Estimating software size, schedule, or time only, but without estimating effort.
- Estimating maintenance effort or testing effort.
- Addressing software project control and planning issues (e.g., scheduling, staff allocation, development process).
- Review papers will be excluded.

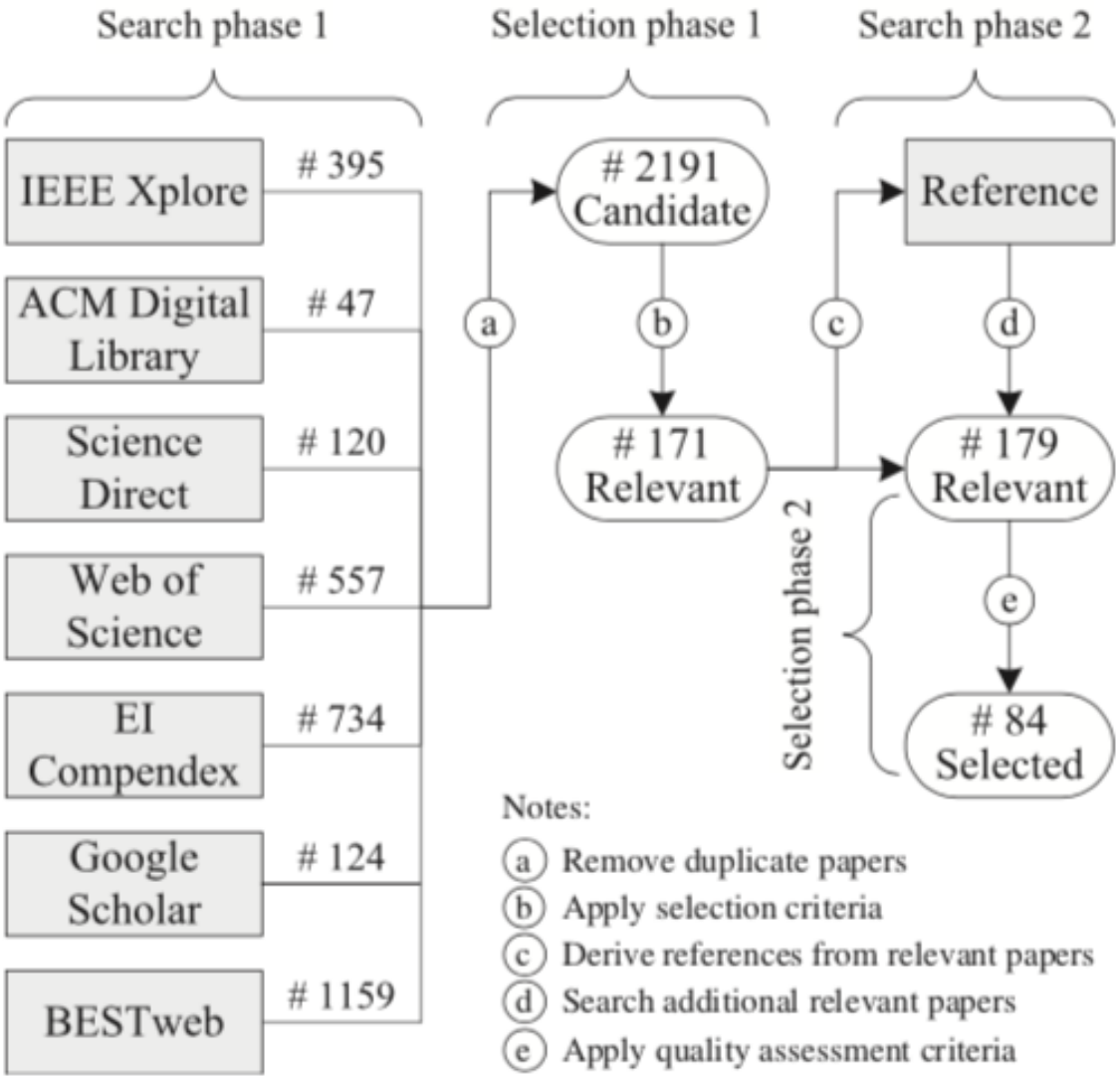
Study quality assessment

Quality assessment questions.

No.	Question
QA1	Are the aims of the research clearly defined?
QA2	Is the estimation context adequately described?
QA3	Are the estimation methods well defined and deliberate?
QA4	Is the experimental design appropriate and justifiable?
QA5[a]	Is the experiment applied on sufficient project data sets?
QA6	Is the estimation accuracy measured and reported?
QA7	Is the proposed estimation method compared with other methods?
QA8	Are the findings of study clearly stated and supported by reporting results?
QA9	Are the limitations of study analyzed explicitly?
QA10	Does the study add value to academia or industry community?

[a] "Yes" (Sufficient): two or more data sets; "Partly" (Partly sufficient): only one data set; "No" (Insufficient): no data set.

Search and selection process



Data extraction

The form of data extraction card.

Data extractor
Data checker
Study identifier
Year of publication
Name of authors
Source
Article title
Type of study (experiment, case study, or survey)
RQ1: Which ML techniques have been used for SDEE?
ML techniques used to estimate software development effort
RQ2: What is the overall estimation accuracy of ML models?
Data sets used in experiments
Validation methods used in experiments
Metrics used to measure estimation accuracy
Estimation accuracy values
RQ3. Do ML models outperform non-ML models?
Non-ML models that this ML model compares with
Rank of the ML and non-ML models regarding estimation accuracy
Degree of improvement in estimation accuracy
RQ4. Are there any ML models that distinctly outperform other ML models?
Other ML models that this ML model compares with
Rank of the ML models regarding estimation accuracy
Degree of improvement in estimation accuracy
RQ5. What are the favorable estimation contexts of ML models?
Strengths of ML technique in terms of SDEE
Weaknesses of ML technique in terms of SDEE

Data synthesis

- The data extracted in this review include both **quantitative** data (e.g., values of estimation accuracy) and **qualitative** data (e.g., strengths and weaknesses of ML techniques).

For the data pertaining to RQ1 and RQ2, we used *narrative synthesis* method. That is, the data were tabulated in a manner consistent with the questions. Some visualization tools, including bar chart, pie chart, and box plot, were also used to enhance the

For the data pertaining to RQ1 and RQ2, we used *narrative synthesis* method. That is, the data were tabulated in a manner consistent with the questions. Some visualization tools, including bar chart, pie chart, and box plot, were also used to enhance the

Results

Publication venues and distribution of selected studies.

Publication venue	Type	# Of studies	Percent
Information and Software Technology (IST)	Journal	14	16
IEEE Transactions on Software Engineering (TSE)	Journal	12	14
Journal of Systems and Software (JSS)	Journal	9	11
Empirical Software Engineering (EMSE)	Journal	9	11
Expert Systems with Applications	Journal	4	5
International Conference on Predictive Models in Software Engineering (PROMISE)	Conference	4	5
International Software Metrics Symposium (METRICS)	Conference	4	5
International Symposium on Empirical Software Engineering and Measurement (ESEM)	Conference	3	4
International Conference on Tools with Artificial Intelligence (ICTAI)	Conference	3	4
International Conference on Software Engineering (ICSE)	Conference	2	2
Journal of Software Maintenance and Evolution: Research and Practice	Journal	2	2
Others		18	21
Total		84	100

Results

3.2. Types of ML techniques (RQ1)

From the selected studies, we identified eight types of ML techniques that had been applied to estimate software development effort. They are listed as follows.

- Case-Based Reasoning (CBR)
- Artificial Neural Networks (ANN)
- Decision Trees (DT)
- Bayesian Networks (BN)
- Support Vector Regression (SVR)
- Genetic Algorithms (GA)
- Genetic Programming (GP)
- Association Rules (AR)

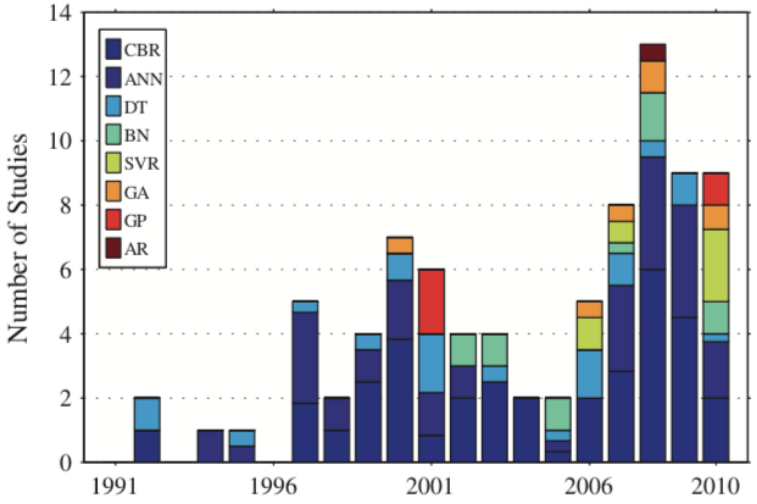
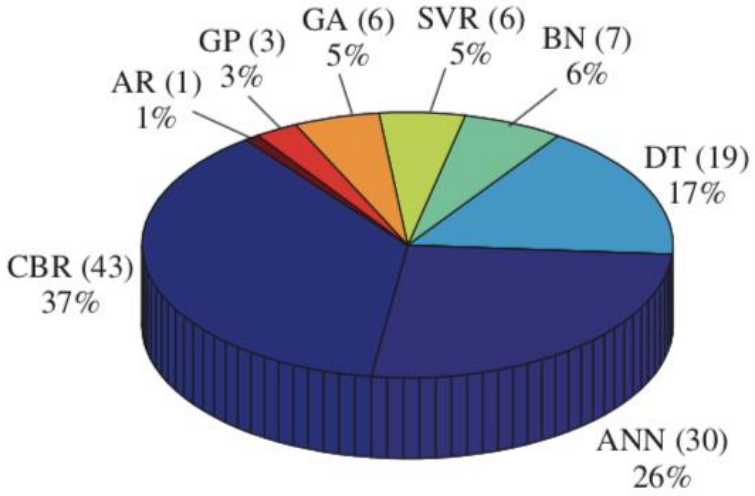


Fig. 4. Distribution of the studies over publication year.



Results

3.6. Favorable estimation contexts of ML models (RQ5)

Given the estimation contexts of an SDEE task, we may concern ourselves with selecting ML models appropriate for the contexts. Such concern can be addressed by investigating the candidate ML models (or, more precisely, the ML techniques) from their characteristics, which are mainly reflected by the strengths and weaknesses of the ML techniques. To this aim, in this review we extracted the strengths and weaknesses of ML techniques and synthesized them based on the type of ML technique. Since different

Favorable/unfavorable estimation contexts of ML models ("√" means favorable, "×" means unfavorable, "–" means not mentioned).

	Small data set	Outliers	Categorical features	Missing values
CBR	√	√	×	×
ANN	×	√	×[a](S68)	–
DT	–	√	√	–
BN	√	√[a](S78)	√[a](S78)	√[a](S67)

[a] Supported by only one study (the study ID is shown in parentheses).

References

- B.A. Kitchenham, Procedures for Performing Systematic Reviews, Keele University Technical Report TR/SE-0401, ISSN:1353-7776
- B.A. Kitchenham, S. Charters, Guidelines for performing systematic literature reviews in software engineering, Tech. Rep. EBSE-2007-01, Keele University and University of Durham, 2007.
- B.A. Kitchenham, D. Budgen, P. Brereton, Evidence-Based Software Engineering and Systematic Reviews, CRC, 2015