EVIDENCE-BASED SOFTWARE ENGINEERING

AISHA UMAIR

aiu@mmmi.sdu.dk

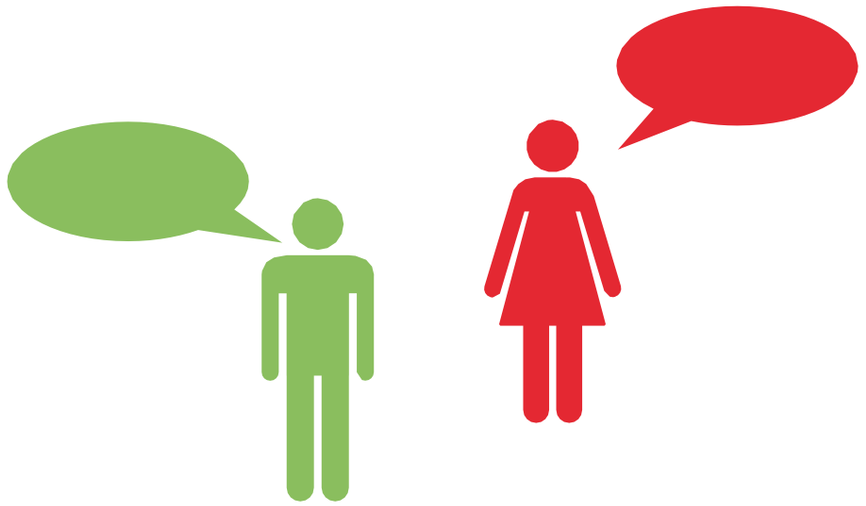
AGENDA

- Evidence-based software engineering
- Systematic Literature Review
- Compulsory Exercise (Deadline: 12th October)
- Systematic Literature Mapping **(Next Week)**
- Grey literature review and multivocal literature review **(Next Week)**
- Replication Study **(Next Week)**
- Guest Lecture on Information Retrieval **(Next Week)**

WHAT IS EVIDENCE ?

WHAT IS EVIDENCE?

Physical evidence



Witness Statements

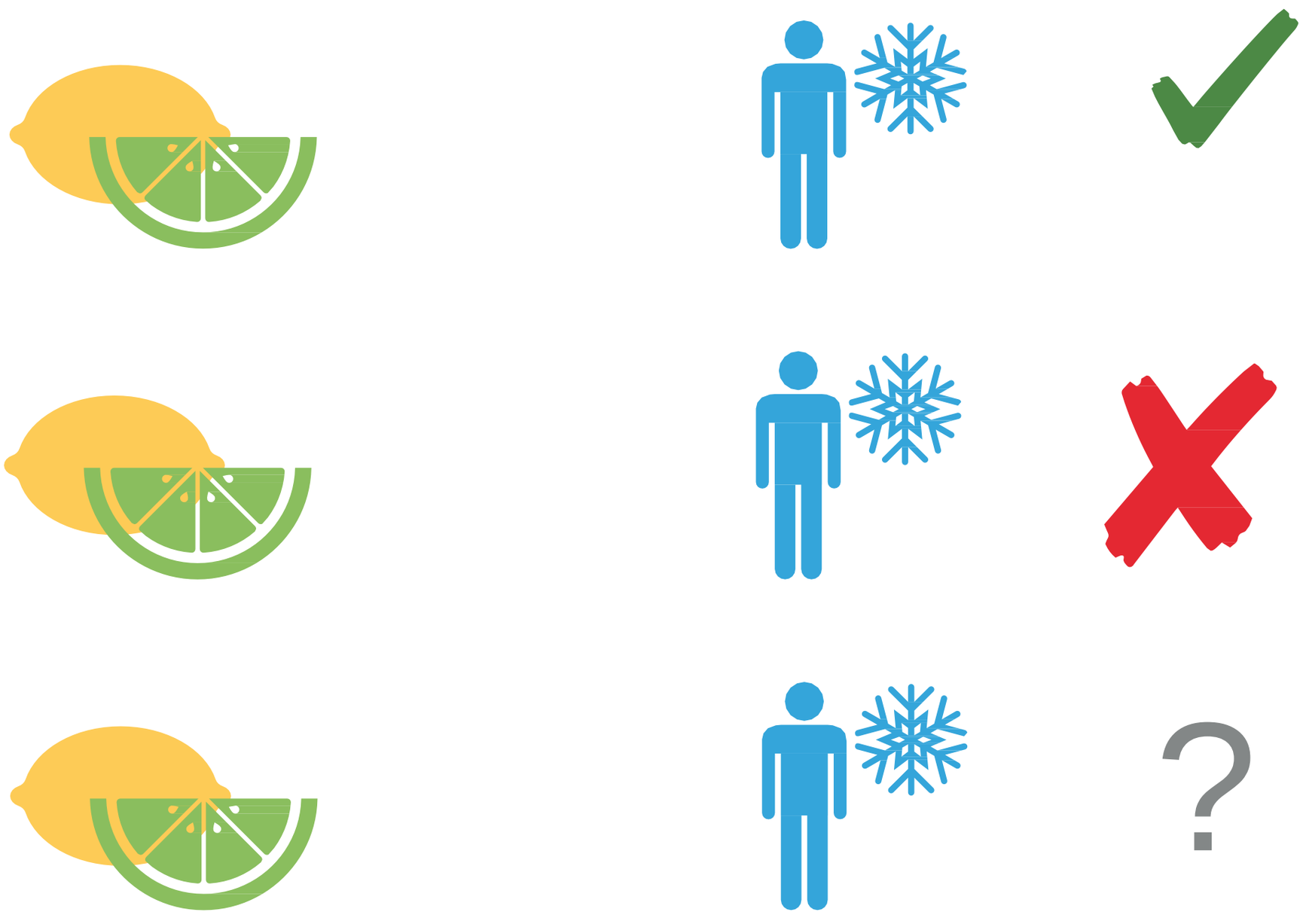
EVIDENCE-BASED MOVEMENT

- Originated in medicine, i.e. Evidence-based medicine (EBM), initiated by Dr. Archie Cochrane.

- Motivation was to find the most effective way to evaluate medical evidence.

EVIDENCE-BASED MOVEMENT

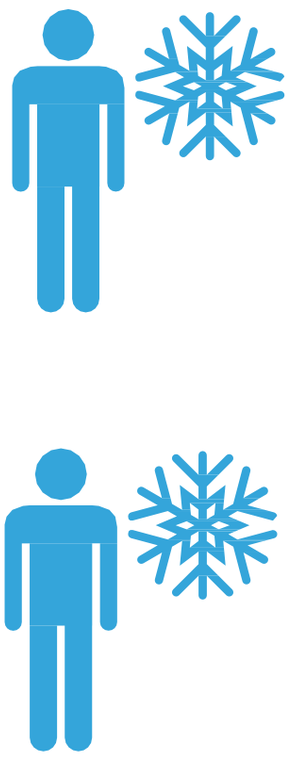
- Different clinical trials seem to have inconsistent results.

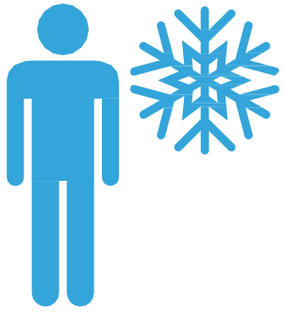


EVIDENCE-BASED MOVEMENT

- Different clinical trials seem to have inconsistent results.

AS A DOCTOR SHOULD I SUGGEST MY PATIENTS TO TAKE VITAMIN C WHEN THEY HAVE A COLD?

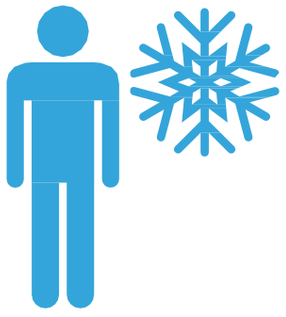


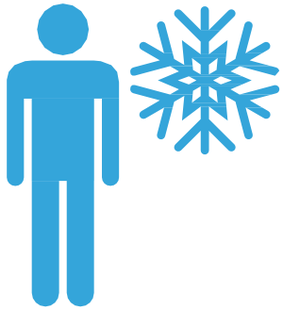


EVIDENCE-BASED MOVEMENT

- Different clinical trials seem to have inconsistent results.

THERE IS A NEED TO SYNTHESISE EVIDENCE FROM MULTIPLE STUDIES!





EVIDENCE-BASED SOFTWARE ENGINEERING

- The same motivation is voiced in the software engineering community.
  - Different studies seem to report different results.
    - Should my team adopt a specific approach or technology to improve productivity?
- Aggregating evidence can be achieved through the process of **Systematic Literature Review (SLR).**

SYSTEMATIC LITERATURE REVIEW (SLR)

KEY TERMS

- Primary Study - “An empirical study in which we directly make measurements about the objects of interest, whether by surveys, experiments, case studies, etc.” (Kitchenham et al., 2016).

- Secondary Study - “An analysis of a set of ***primary studies,*** and usually seeks to aggregate the results from these in order to provide stronger forms of evidence about a particular phenomenon”(Kitchenham et al., 2016).

Forms of secondary study:

- Systematic Literature Review (SLR)/ Meta Study/ Meta Analysis

- Systematic Literature Mapping/ Systematic Mapping Study (SM)

Kitchenham, B., Budgen, D., Brereton, P., 2016. Evidence-based software engineering and systematic review, CRC Press.

SYSTEMATIC LITERATURE REVIEW

- SLR is “a form of secondary study that uses a well-defined methodology to identify, analyse and interpret all available evidence related to a specific **research question**.”
  - A form of research method
  - Could be qualitative or quantitative

- Guidelines:
  - Kitchenham, B. , Charters, S. , 2007. Guidelines for Performing Systematic Literature Reviews in Software Engineering. Technical Report, EBSE-2007-01. Keele University, UK.

SLR STEPS

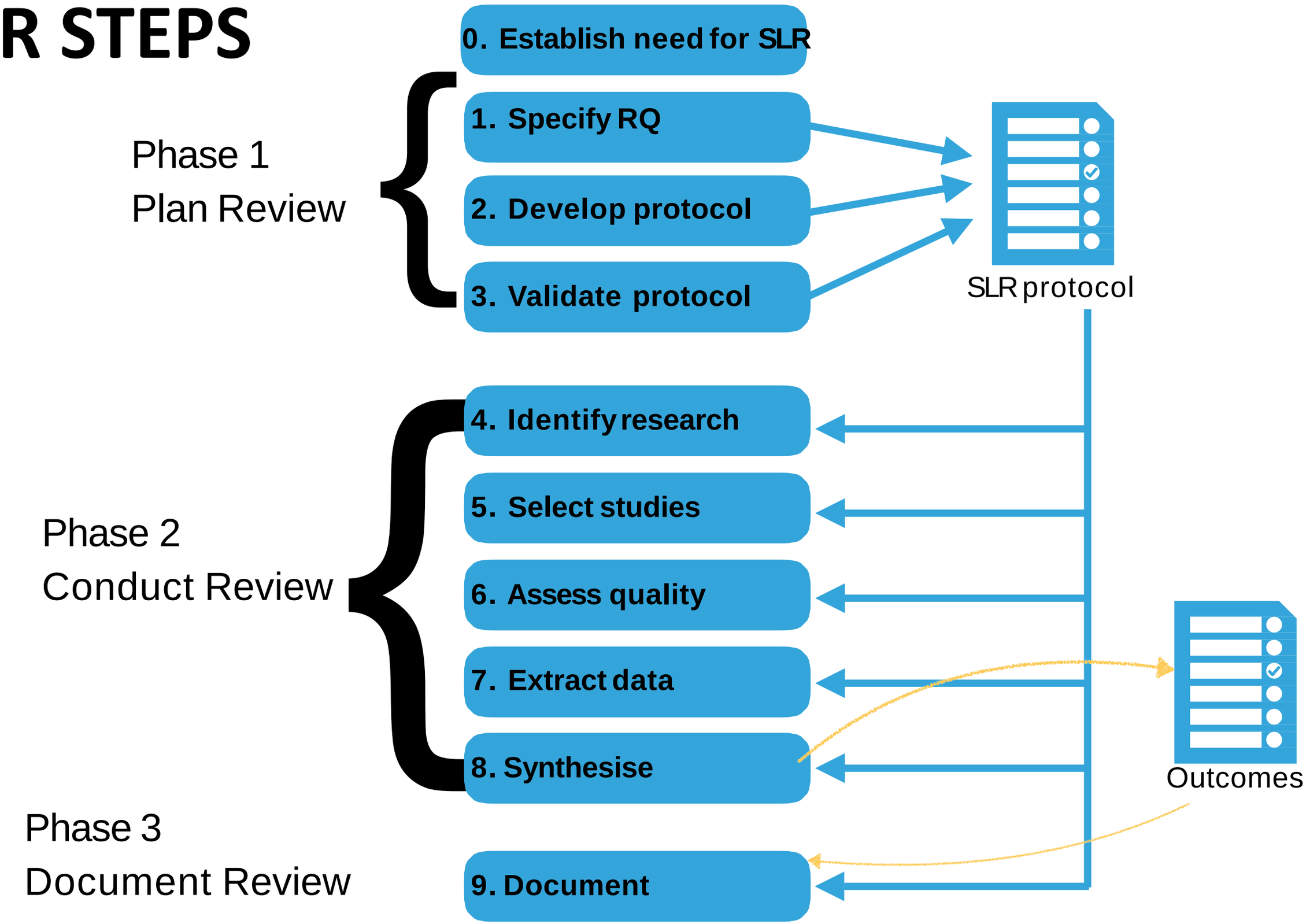


Image adapted from Kitchenham et al. (2016) Figure 4.1 page 40

Phase 1: Plan Review

Image adapted from Kitchenham et al. (2016) Figure 4.1 page 40

0. NEED FOR AN SLR (KITCHENHAM AND CHARTERS, 2007)

- To summarise the existing evidence concerning a treatment or technology.
  - E.g., benefits and limitations of pair-programming.
- To identify any gaps in current research in order to suggest areas for further investigation.
- To examine the extent to which empirical evidence supports/contradicts theoretical hypotheses, or even to assist the generation of new hypotheses

Kitchenham, B. , Charters, S. , 2007. Guidelines for Performing Systematic Literature Reviews in Software Engineering. Technical Report, EBSE-2007-01. Keele University, UK

1. RESEARCH QUESTION

- Quantitative
- Qualitative

1. RESEARCH QUESTION

- Quantitative

Comparison of two (or more) technologies to determine which one is more effective or efficient in a certain context.

PICOC - Population Intervention Comparison Outcome Context (Petticrew and Roberts)

- Population: roles (testers, developers), expertise, industry domain (telecommunication, automotive, etc.)
- Intervention: technology of interests, e.g., TDD, pair programming
- Comparison: test-last development, solo programming
- Outcome: quality (number of defects, time to complete a task)
- Context: Academia - industry, SME - Large corporation.

https://libguides.baylor.edu/evidencesynthesis/home

https://www.researchgate.net/publication/295863661_The_Effects_of_Test_Driven_Development_on_Internal_Quality_External_Quality_and_Productivity_A_systematic_review
https://www.researchgate.net/publication/256848134_Effects_of_Test-Driven_Development_A_Comparative_Analysis_of_Empirical_Studies

1. RESEARCH QUESTION

- Qualitative
  - Understanding the benefits, risk, impacts of adopting a specific technology.

EXAMPLE (QUALITATIVE)

Empirical studies of agile software development: A systematic review

An SLR by Dybå and Dingsøyr (2008) aims to "collect evidence on benefits and limitations of agile software development, understand the strength of evidence of the findings, practical and research implications.".

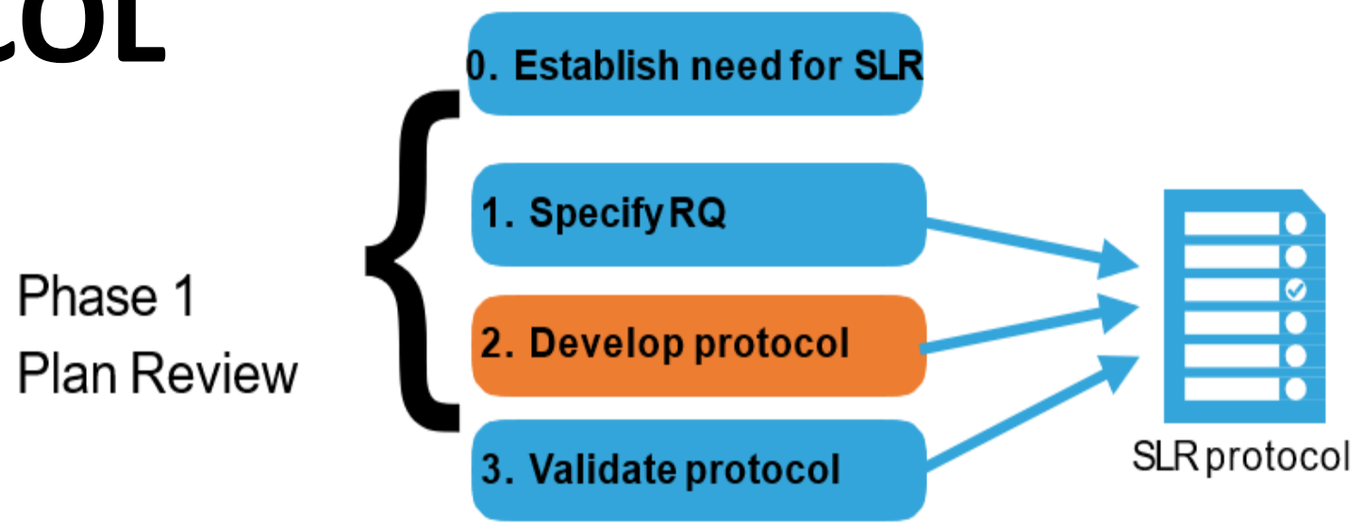
They formulated the following RQs:

1. What is currently known about the benefits and limitations of agile software development?

2. What is the strength of the evidence in support of these findings?

3. What are the implications of these studies for the software industry and the research community?

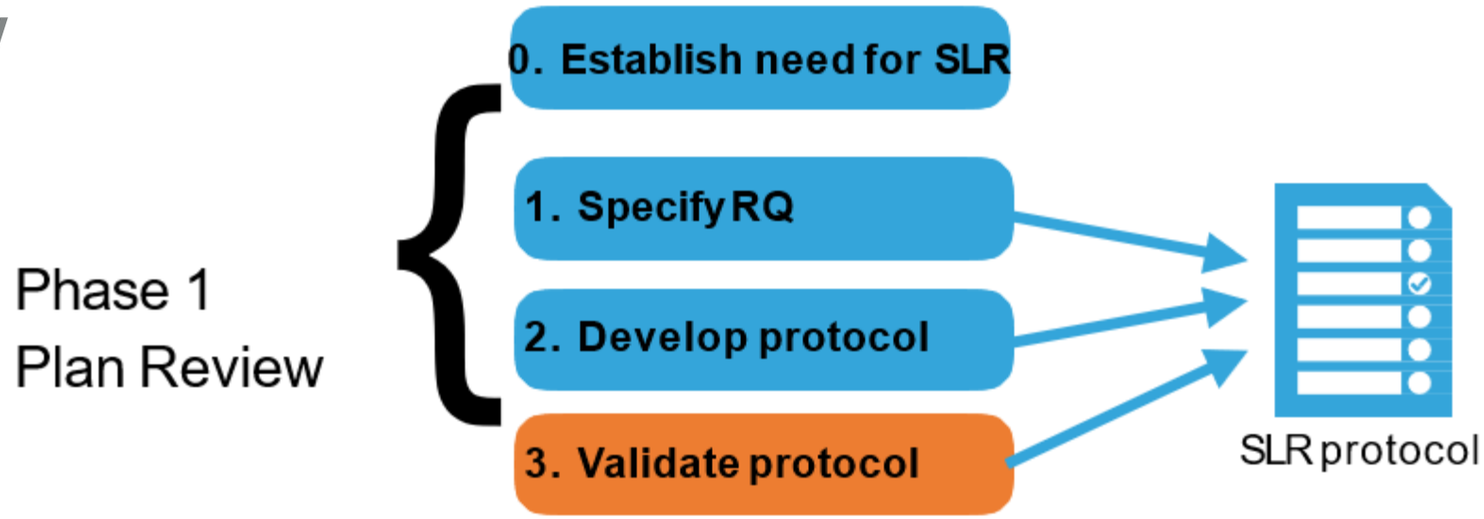
2. DEVELOP PROTOCOL



- Background.
- The research questions
- The search strategy
- Study selection criteria
- Study selection procedures
- Study quality assessment checklists and procedures
- Data extraction strategy
- Synthesis of the extracted data

Etc...

3. VALIDATE PROTOCOL



- Piloting
- Have a person with expertise and experience in SLR to review your protocol against a guideline.
  - Is the motivation clear?
  - Are the research questions clearly stated?
  - Is the search strategy clearly defined?
  - Are the inclusion/ exclusion criteria defined?

SLR STEPS

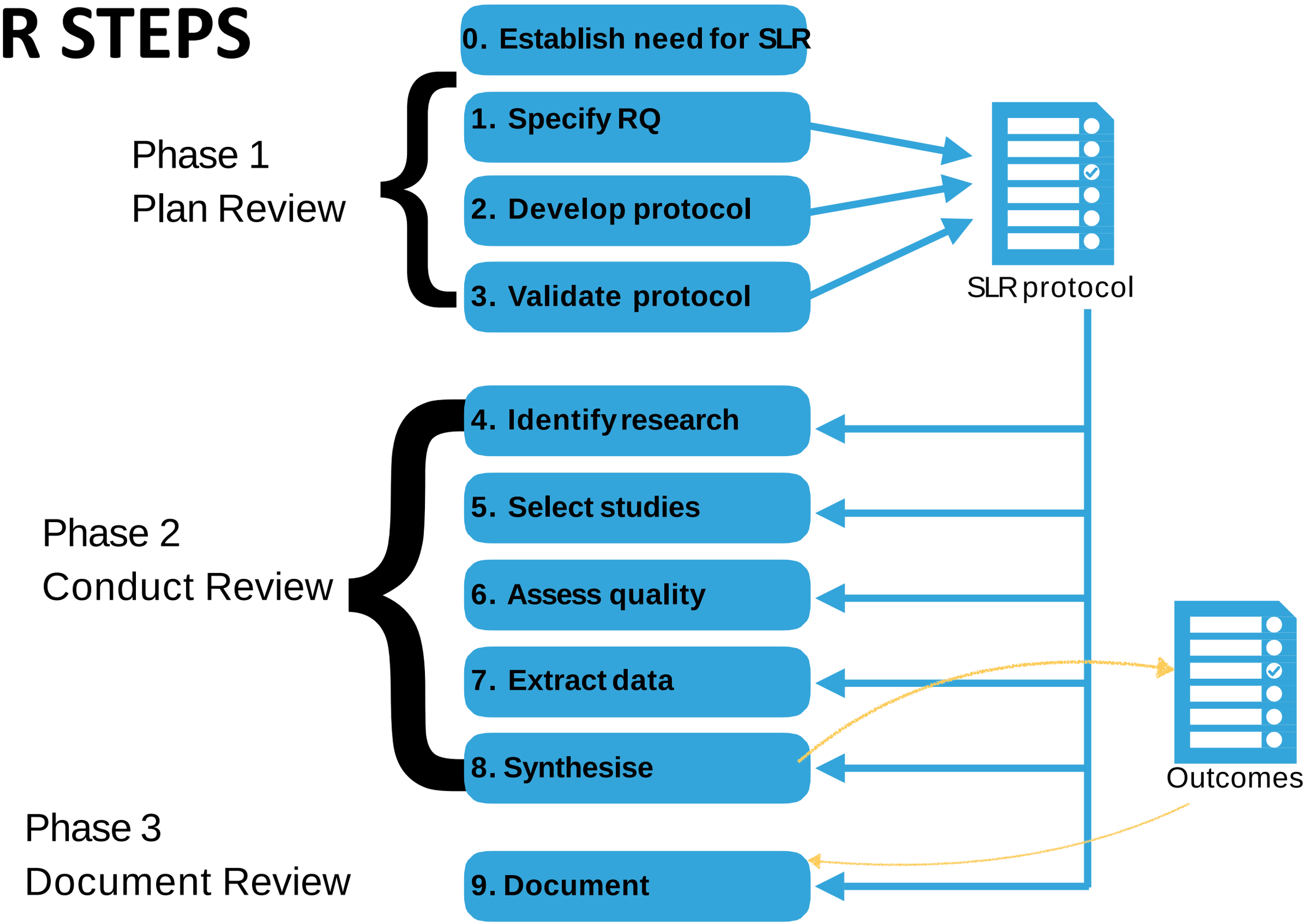


Image adapted from Kitchenham et al. (2016) Figure 4.1 page 40

Phase 2: Conduct Review

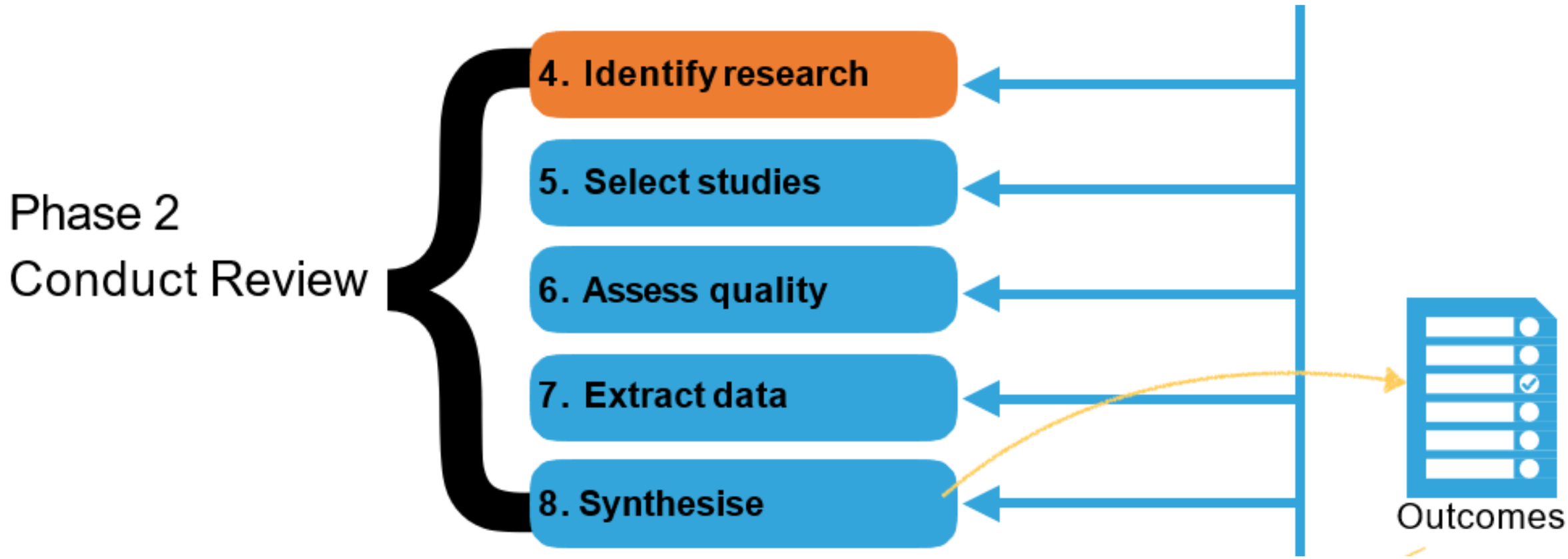


Image adapted from Kitchenham et al. (2016) Figure 4.1 page 40

4. IDENTIFY RESEARCH:

- Search for relevant primary studies that fulfil the goals of the SLR and answer its research questions.
- An iterative process.
- Before starting out on search, you need to have a basis of completeness.
  - To ensure that your strategy will lead to as many relevant primary studies as possible.

4. IDENTIFY RESEARCH

- Search strategies:
  - Automated search
  - Manual search
  - Snowballing

4. IDENTIFY RESEARCH

Automated search

- Use digital libraries, e.g., IEEE Xplore, ACM Digital library, etc.
- Use services like Scopus and Web of Science.

- Need to compose search strings.
  - Search string usually include keywords or phrases relevant to the topic with some boolean operator.
  - Composing search string is an iterative process.

SEARCH STRING

Details and hands-on exercise on formulating search string will be done next week (week 41) with a guest lecturer from SDU Bibliotek.

4. IDENTIFY RESEARCH

Manual search

- Manually searching primary studies from journals and conference proceedings.
- Identify leading software engineering journal and conferences.

✓ Some “well-known” and reputable conferences: ICSE, FSE/ESEC, ASE, RE, ESEM.

✓ Journal: TSE, ESEJ, TOSEM, IST, JSS, REJ, SoSym.

✓ List of conferences can be found here:

https://research.com/conference-rankings/computer-science/software-programming

4. IDENTIFY RESEARCH

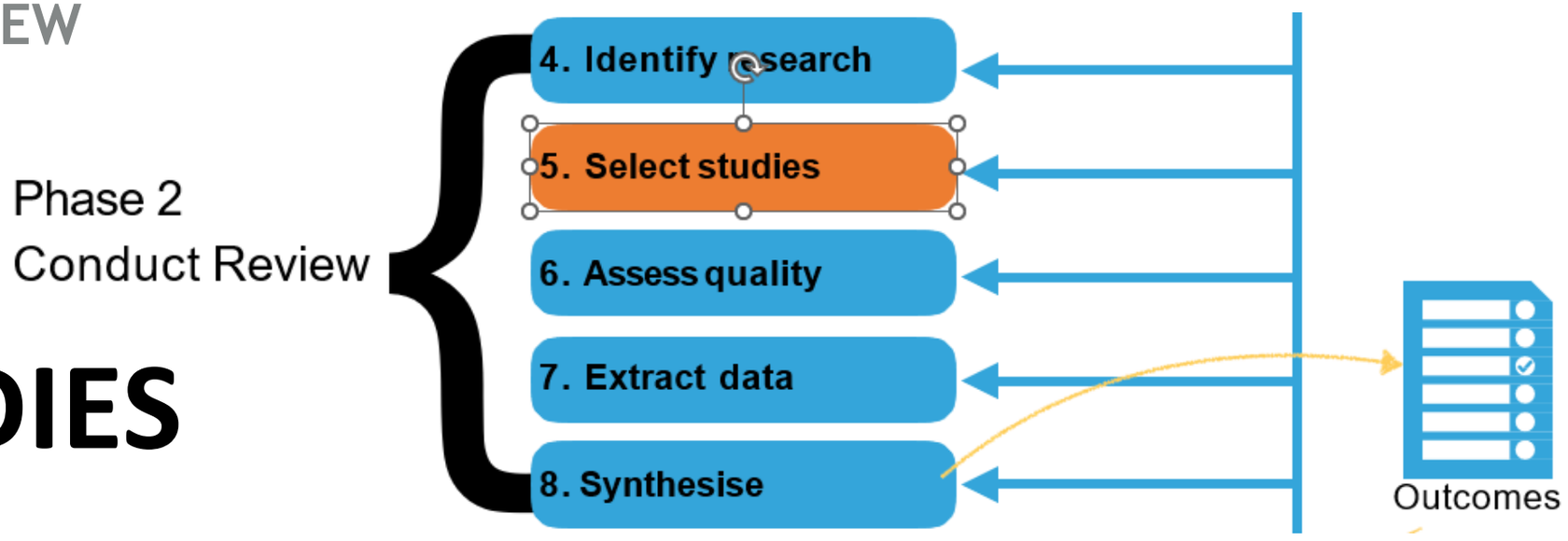
- Snowballing
- Citation analysis
  - Backward snowballing - where you look at the references of known papers.
  - Forward snowballing - where you look at other papers that cite known papers.
- Suitable where there are a small number of papers. New area.
- A study by Jalali and Wohlin (2012) shows that snowballing is more precise in identifying relevant primary studies than automated search.

S.Jalali and C. Wohlin, "Systematic literature studies: Database searches vs. backward snowballing," Proceedings of the 2012 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement, Lund, 2012, pp. 29-38

COMPLETENESS?

- It is not possible to include ALL relevant primary studies.
- A way to mitigate is to at least use two search strategies.
  - Automated + manual
  - Automated + snowball
- You can also find an expert in the area that can check for the completeness.

Break (15 min)



5. SELECT STUDIES

- Select primary studies to include in an SLR.
- Define inclusion/ exclusion criteria.
  - Generic inclusion criteria:
    - Papers written in a specific language(s), usually English.
  - Generic exclusion criteria:
    - Studies reported in more than one papers.
    - Papers not in relevant fields.
    - Non-peer reviewed papers.

EXAMPLE: Inclusion Criteria

Taken from Dybå and Dingsøyr (2008): aims to **"collect evidence on benefits and limitations of agile software development, understand the strength of evidence of the findings, practical and research implications."**

- Papers pertaining to Agile software development
- Must be an empirical study (case study, survey, experiment).
- Written in English
- Focus on Agile methods in general and not specific practices like such as pair programming, unit testing etc

Ref: Empirical studies of agile software development: A systematic review

5. SELECT STUDIES

- Usually conducted in a team, at least two people to minimise bias.

https://www.youtube.com/watch?v=fOR_8gkU3UE

  - Need to check level of agreement using Kappa analysis.

Kappa co-efficient	Strength of agreement
0 - 0.29	Poor
0.30 - 0.40	Fair
0.41 - 0.60	Moderate
0.61 - 0.80	Good
0.81 - 1.00	Very good

6. ASSESS QUALITY

- WHY?
  - Quality assessment is to determine the extent to which the results of an empirical study are valid and free (minimised) from bias.
  - The variation of quality needs to be considered when synthesising the results.

6. ASSESS QUALITY

- Quality criteria
  - Quality criteria could differ depending what types of studies you aim to include.
  - Generic study quality checklist: Dybå, T. , Dingsøyr, T. , 2008. **Empirical studies of agile software development: a systematic review**. Inf. Softw. Technol. 50 (91), 833–859 .

EXAMPLE (DYBÅ AND DINGSØYR, 2008)

	Criteria	Score (Yes-1/ No-0)
1	Is there an adequate description of the context in which the research was carried out?	
2	Is there a clear statement of the aims of the research?	
3	Was the research design appropriate to address the aims of the research?	
4	Is there a clear statement of findings?	

Generic study quality checklist: Dybå, T., Dingsøyr, T., 2008. Empirical studies of agile software development: a systematic review. Inf. Softw. Technol. 50 (91), 833–859.

RIGOUR AND RELEVANCE (IVARSSON AND GORSCHEK, 2011)

- Rigour - the precision or exactness of the research method used.
- Relevance - evaluation of academic relevance or research impact in industry.

M. Ivarsson, T. Gorschek, A method for evaluating rigor and industrial relevance of technology evaluations, Empir. Softw. Eng. 16 (3) (2011) 365–395.

RIGOUR EVALUATION

Aspect	Strong (1)	Medium (0.5)	Weak (0)
Context described	The context is described to the degree where a reader can understand and compare it to another context.	The context in which the study is performed is mentioned or presented in brief.	No description
Study design described	The study design is described in depth, with details of variables, treatment, sampling, etc.	The study design is briefly described.	No description
Validity discussed	The validity and limitation of the study is discussed in detail. Including mitigation strategies.	The validity of the study is mentioned but not described in detail.	No description

RELEVANCE EVALUATION

Aspect	Contribute to relevance (1)	Do not contribute to relevance (0)
Subjects	Industrial practitioners.	Students, researchers.
Context	Industrial setting	Laboratory setting
Scale	Real-industry project or application	Toy example
Research Method	The research method mentioned to be used in the evaluation is one that facilitates investigating real situation: (1) case study, (2) survey, (3) ethnography in industry.	The research method mentioned to be used in the evaluation does not lend itself to investigate real situations: (1) conceptual analysis, (2) laboratory experiment

6. ASSESS QUALITY

- Quality assessment can be done in a team. In that case consensus needs to be reached for the quality score of the study.
- Discussion.
- Or aggregation of the quality score.

7. EXTRACT DATA

- Extract the data needed to address the research questions.
- Could be numerical data or textual data.
- Supporting tools
  - Spreadsheet
  - Qualitative analysis tools, e.g., NVivo.

EXAMPLE

- An SLR by Dybå and Dingsøyr (2008) aims to "**collect evidence on benefits and limitations of agile software development, understand the strength of evidence of the findings, practical and research implications.**". They formulated the following RQs:

1. What is currently known about the benefits and limitations of agile software development?
2. What is the strength of the evidence in support of these findings?
3. What are the implications of these studies for the software industry and the research community?

EXAMPLE - APPENDIX C (DYBÅ AND DINGSØYR, 2008)
Data extraction form

Study description		
1.	Study identifier	Unique id for the study
2.	Date of data extraction	
3.	Bibliographic reference	Author, year, title, source
4.	Type of article	Journal article, conference paper, workshop paper, book section
5.	Study aims	What were the aims of the study?
6.	Objectives	What were the objectives?
7.	Design of study	Qualitative, quantitative (experiment, survey, case study, action research)
8.	Research hypothesis	Statement of hypotheses, if any
9.	Definition of agile software development given in study	Verbatim from the study
10.	Sample description	Size, students, professionals (age, education, experience)
11.	Setting of study	Industry, in-house/supplier, products and processes used
12.	Control group	Yes, no (number of groups, sample size)
13.	Data collection	How was the data obtained? (questionnaires, interviews, forms)
14.	Data analysis	How was the data analyzed? (qualitative, quantitative)
Study findings		
1.	Findings and conclusions	What were the findings and conclusions? (verbatim from the study)
2.	Validity	Limitations, threats to validity
3.	Relevance	Research, practice

EXAMPLE - APPENDIX C (DYBÅ AND DINGSØYR, 2008)

RQ1. Benefits and limitation

Study findings		
1.	Findings and conclusions	What were the findings and conclusions? (verbatim from the study)
2.	Validity	Limitations, threats to validity
3.	Relevance	Research, practice

EXAMPLE - APPENDIX C (DYBÅ AND DINGSØYR, 2008)

RQ1. Benefits and limitation

Study findings		
1.	Findings and conclusions	What were the findings and conclusions? (verbatim from the study)
2.	Validity	Limitations, threats to validity
3.	Relevance	Research, practice

RQ2. Strength of evidence

EXAMPLE - APPENDIX C (DYBÅ AND DINGSØYR, 2008)

RQ1. Benefits and limitation

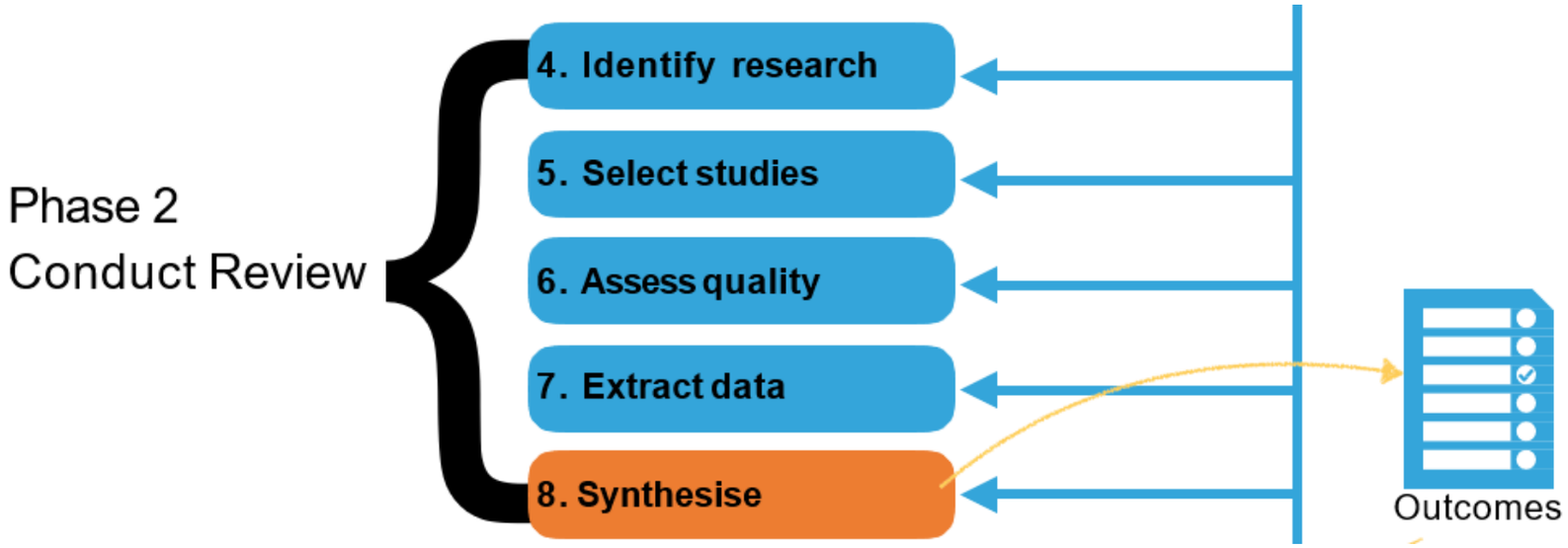
Study findings

1.	Findings and conclusions	What were the findings and conclusions? (verbatim from the study)
2.	Validity	Limitations, threats to validity
3.	Relevance	Research, practice

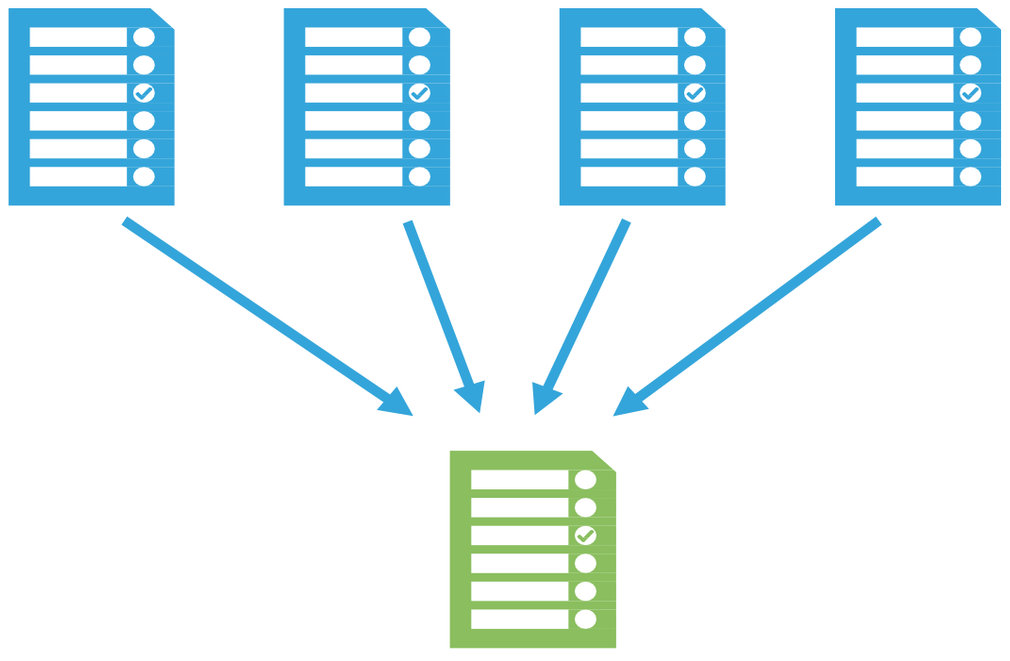
RQ2. Strength of evidence

RQ3. Implications to research and practice

8. SYNTHESISE



- Synthesis - to summarise, integrate, combine, and compare the results from primary studies of a common topic.



8. SYNTHESISE

- Synthesis - to combine the results from primary studies as a whole.
  - Aggregation (qualitative)
  - Meta-analysis (quantitative)

EXAMPLE (DYBÅ AND DINGSØYR, 2008)

4.7.2. Productivity

Four studies compared the productivity of agile teams with the productivity of teams using traditional development methods (S7, S10, S14, S32); see Table 14. Ilieva et al. (S10) compared the productivity of two similar projects, one of which used traditional methods and the other of which used XP. They measured the productivity for three iterations of each project. Overall, the results showed a 42% increase in productivity for the agile team. The increase in productivity was largest for the first iteration, while there was virtually no difference in productivity for the last iteration.

The case study by Layman et al. (S14) compared an old release developed with traditional methods with a new

Table 14
Comparisons of productivity

Study	$Productivity_{TRAD}$	$Productivity_{AGILE}$	Productivity gain (%)
S7	3 LOC/h[a]	13.1 LOC/h	337
S10	3.8 LOC/h	5.4 LOC/h	42
S14	300 LOC/month	440 LOC/month	46
S32	157 LOC/ engineer[b]	88 LOC/engineer	−44

[a] V-model.

[b] Comparisons were made between two one-semester courses; however, the actual hours worked by the members of the teams were not measured.

8. SYNTHESISE

- Quantitative synthesis

  The outcome is statistical information: correlation, descriptive statistics, etc.

- Different methods:
  - Vote-counting
  - Meta-analysis

VOTE COUNTING - EXAMPLE (MUNIR ET AL. 2014)

- Vote counting (at its simplest) involves counting how many primary studies show significant effects and how many do not.

Table 5
Outcome of studies (category A) with respect to variables.

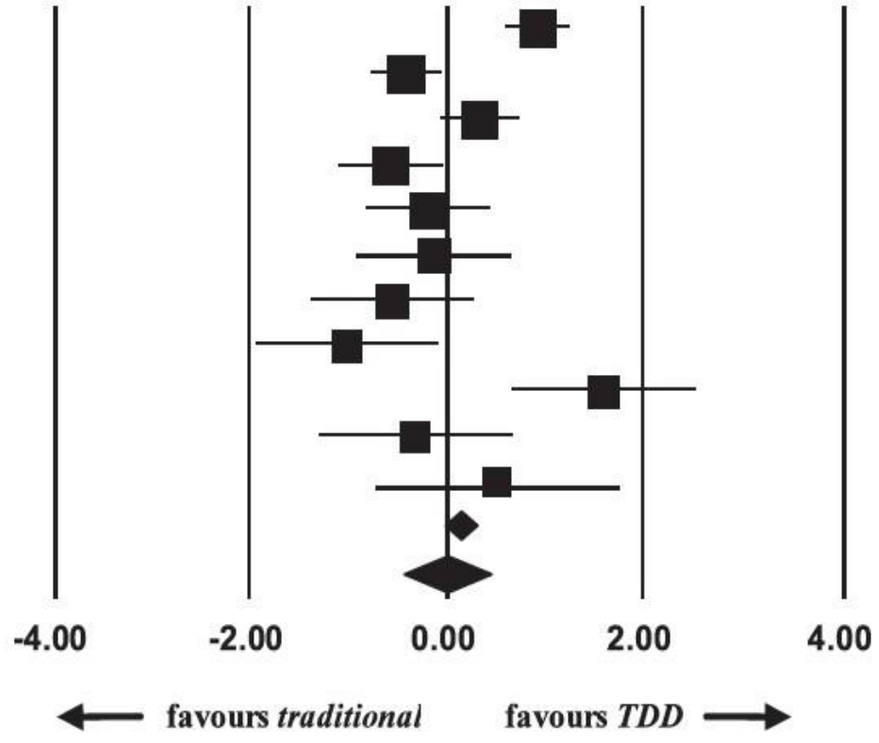
Variable	In favor of TDD	No diff.	Neg. for TDD
Productivity		[64]	[59,60]
External quality	[74,60,59,63,62,71,64]		
Complexity	[61,71]		
Developer opinion	[72,74]		
Total	11	1	2

Munir H.,, Moayyed, M., Petersen, K. 2014. Considering rigor and relevance when evaluating test driven development: A systematic review, Information and Software Technology:56(4), pp. 375-394

META-ANALYSIS - EXAMPLE (RAFIQUE AND MIŠIĆ, 2013)

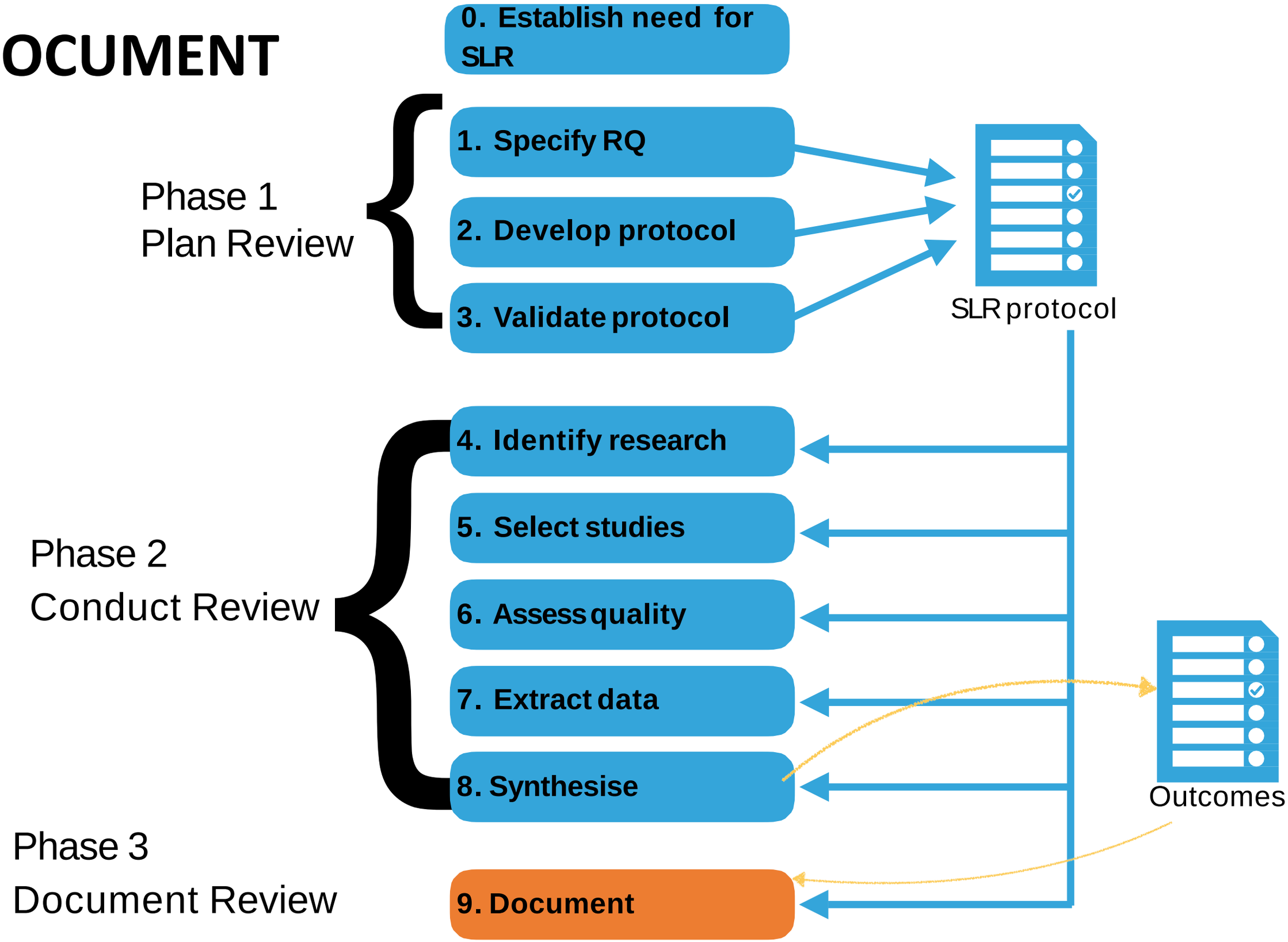
- Meta-analysis is the use of statistical methods to analyse data from primary studies that compares two different treatments.

Experiment	Hedges's *g*	95% CI		p-Value	Relative Weight Random	Relative Weight Fixed
Desai et al. [50]	0.907	0.588	1.225	0.000	11.26	25.81
Madeyski E2 [48]	-0.437	-0.781	-0.093	0.013	11.15	22.09
Slyngstad et al. [63]	0.309	-0.082	0.701	0.121	10.93	17.06
Madeyski E1 [48]	-0.595	-1.123	-0.067	0.027	10.22	9.37
Huang & Holcombe [52]	-0.217	-0.834	0.400	0.491	9.70	6.86
Erdogmus et al. [5]	-0.152	-0.928	0.625	0.702	8.74	4.33
Gupta & Jalote E2 [47]	-0.580	-1.402	0.242	0.167	8.47	3.87
Muller & Hagner [54]	-1.039	-1.961	-0.118	0.027	7.88	3.08
Gupta & Jalote E1 [47]	1.579	0.650	2.509	0.001	7.83	3.02
Pancur & Ciglaric E2,E4 [49]	-0.342	-1.314	0.630	0.490	7.59	2.76
Xu & Li [58]	0.495	-0.734	1.724	0.430	6.23	1.73
Overall – Fixed Effects	0.106	-0.056	0.267	0.201		
Overall – Random Effects	-0.010	-0.454	0.433	0.964		



Y.Rafique and V.B. Mišić, 2013. The Effects of Test-Driven Development on External Quality and Productivity: A Meta-Analysis. IEEE Transactions on Software Engineering:39(6), pp. 835-856

9. DOCUMENT


0. Establish need for SLR
Phase 1
Plan Review
1. Specify RQ
2. Develop protocol
3. Validate protocol
SLR protocol
Phase 2
Conduct Review
4. Identify research
5. Select studies
6. Assess quality
7. Extract data
8. Synthesise
Outcomes
Phase 3
Document Review
9. Document

9. DOCUMENT

- Be aware of the audience and format of the report.
- More on academic writing in week 47

Compulsory Exercise
Deadline: 12th October 2023

Conduct Systematic Literature Review: (SE Students)

- Look at two SLRs on **TDD** (uploaded on itslearning), just for an inspiration.
- Follow the SLR steps (taught in the lecture) to conduct systematic literature review on the topic from **Advanced Software Architecture** course

COMPARE TWO SLRS: (GD Students)

- Look at two SLRs on **Gamificatio**n (uploaded on itslearning)
- Discuss the following:
  - What is the main difference of aims between the two SLRs?
  - Can you trace the connection between the aims and the research questions/ hypotheses? How do they differ between the two SLRs?
  - Can you find connection between the aims/ research questions and the search strings? How do they differ between the two SLRs?
  - Can you trace the connection between the research questions/ hypotheses and the analysis method? Compare between the SLRs?
  - Can you trace the connection between the aims and results? How do they differ?
  - What is the main difference in the conclusions between the two SLRs?
  - As a student researcher/student developer, How would you use the results from each SLR?

NEXT WEEK

- Systematic Literature Mapping (12:15 – 13:00)
- Guest lecture: Dr. Evgenios Vlachos, Research Librarian, SDU Bibliotek.

Topic: Information retrieval: (13:15 – 15:15)

SYSTEMATIC LITERATURE MAPPING

SYSTEMATIC MAPPING

- Systematic mapping (SM) or mapping study is “a secondary study intended to identify and classify the set of publication on a topic”. (Kitchenham et al. 2016)
- SM is “designed to give an overview of a research area through classification and counting contributions in relation to the categories of that classification” (Petersen et al., 2015)

Kitchenham, B., Budgen, D., Brereton, P.,2016. Evidence-based software engineering and systematic review, CRC Press.

Petersen K., Vakkalanka, S., Kuzniarz, L. 2015. Guidelines for conducting systematic mapping studies in software engineering: An update. Information and Software Technology:64, pp.1-18.

SYSTEMATIC MAPPING

- Guidelines:
  - Petersen, K., Feldt, R., Mujtaba, S., Mattsson, M. 2008. Systematic mapping studies in software engineering. In Proceedings of the 12th international conference on Evaluation and Assessment in Software Engineering (EASE'08), pp. 68-77.
  - Petersen K., Vakkalanka, S., Kuzniarz, L. 2015. Guidelines for conducting systematic mapping studies in software engineering: An update. Information and Software Technology:64, pp.1-18.

SLR VS SM

	SLR	SM
Research Question	Aimed at aggregating evidence pertaining to the impact of an intervention.	Aimed at discovering research trends
Quality assessment	Must be performed	Nice to have
Outcome	Qualitative or quantitative synthesis.	Classification of primary studies

SM STEPS

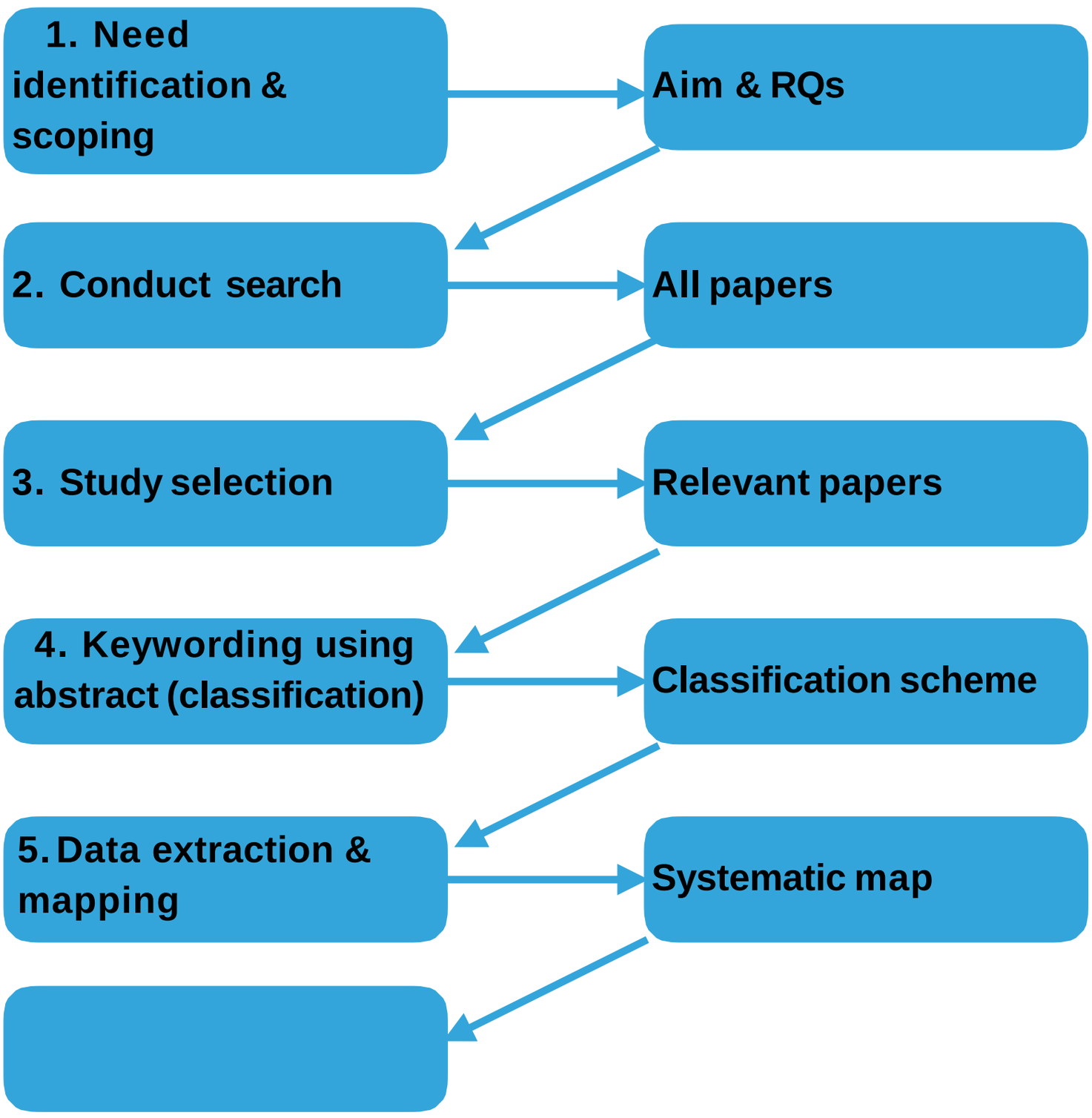


Image adapted from Petersen et al. (2008)

1. IDENTIFICATION & SCOPING

- Aimed at identifying trends and gap in the literature, for example:
  - Classify relevant literature in software product line variability (Mujtaba et al., 2008). Formulated RQs:
    - RQ1:What areas in software product line variability have been addressed and how many articles cover the different areas?
    - RQ2:What types of papers are published in the area and in particular what type of evaluation and novelty do they constitute?
- More examples see Petersen et al. (2015).

S. Mujtaba, K. Petersen, R. Feldt, M. Mattsson. 2008. Software product line variability: a systematic mapping study, School of Engineering, Blekinge

Petersen K., Vakkalanka, S., Kuzniarz, L. 2015. Guidelines for conducting systematic mapping studies in software engineering: An update. Information and Software Technology:64, pp. 1-18.

2. CONDUCT SEARCH

The same strategies as in an SLR.

- Manual search
- Automated search
- Snowball search

3. STUDY SELECTION

- Similar to an SLR.
- Generic:
  - Relevant to the focus area and within the scope
  - Within the domain i.e. software engineering
  - Language scope

4. CLASSIFICATION

- Topic-dependent
  - Use keywords use in the abstract or list of keywords mentioned in the primary studies.
  - Use existing classification from ISO standards or SWEBOK
  - Find other similar mapping studies
- Topic-independent
  - Venue and research method
  - Research type (Wieringa et al., 2006): evaluation, solution proposal, validation, etc…

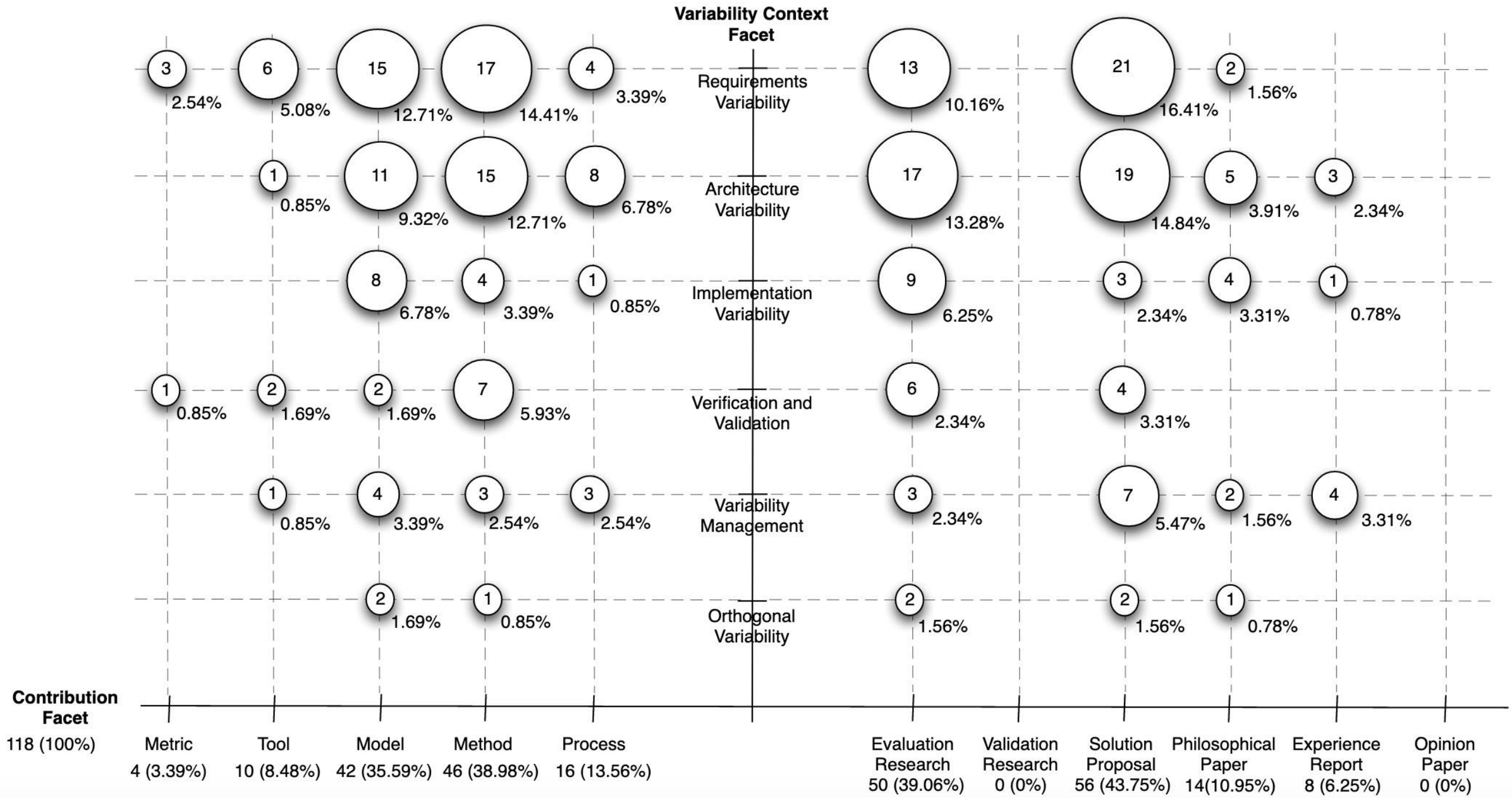
R.Wieringa, N. Maiden, N. Mead, C. Rolland, Requirements engineering paper classification and evaluation criteria: a proposal and a discussion, Requirements Eng. 11 (1) (2006) 102–107

5. DATA EXTRACTION AND MAPPING

- Once classifications are developed, data is extracted and the primary studies are mapped into the classification.
- Usually visualisation is used to show the mapping:
  - Bubble plots (popular and effective)
  - Histograms
  - Tables

What does it say ?????

EXAMPLE (MUJTABA ET AL., 2008)



S. Mujtaba, K. Petersen, R. Feldt, M. Mattsson. 2008. Software product line variability: a systematic mapping study, School of Engineering, Blekinge

6. DOCUMENT

- Similar to SLR.

TERTIARY STUDY

TERTIARY STUDY

- A review of secondary studies, e.g., systematic reviews or systematic mapping.
- It can be done when there are considerable number of secondary studies.
- The motivation:
  - Mapping of existing secondary studies.
  - Aggregation of findings in secondary studies.
  - Evaluation of existing secondary studies.

TERTIARY STUDY

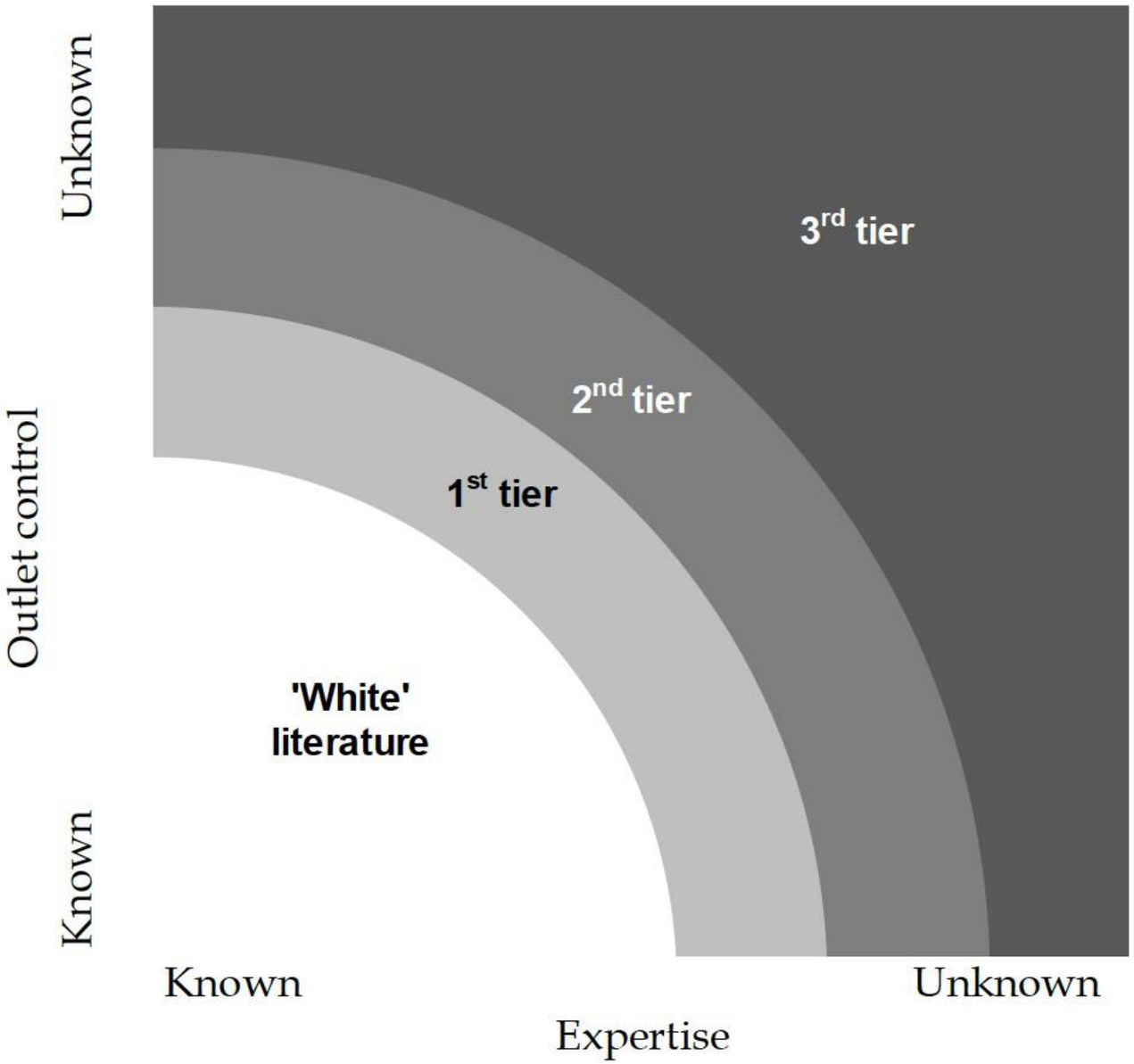
- Same process and steps as an SLR or SM.

GREY LITERATURE
&
MULTIVOCAL LITERATURE REVIEW

GREY LITERATURE REVIEW

- Grey Literature (GL) - non peer-reviewed papers or articles. Including videos,
- According to Cochrane handbook for systematic reviews of intervention, GL is “literature that is not formally published in sources such as books or journal articles”
  - **OBS!** Not all books are peer-reviewed. Anyone can write and publish a book.
- It is done when research is not extensively done on the topic.
- More importantly practitioners are more likely to read grey literature.
- Common in other disciplines such as education and nursing.

"SHADES" OF GREY LITERATURE



3rd tier GL:
Low outlet control/ Low credibility:
Blogs, emails, tweets

2nd tier GL:
Moderate outlet control/ Moderate credibility:
Annual reports, news articles, presentations, videos, Q/A sites (such as StackOverflow), Wiki articles

1st tier GL:
High outlet control/ High credibility:
Books, magazines, government reports, white papers

Image source: R.J. Adams, P.Smart, and A. S. Huff, "Shades of Grey: Guidelines for Working with the Grey Literature in Systematic Reviews for Management and Organizational Studies," International Journal of Management Reviews, 2016.

MULTIVOCAL LITERATURE REVIEW

- Peer review + grey literature = Multivocal Literature Review (MLR)
- Multivocal literature review is “a form of a Systematic Literature Review (SLR) which includes the grey literature (e.g., blog posts, videos and white papers) in addition to the published (formal) literature (e.g., journal and conference papers).” (Garousi et al., 2019)
- Guideline:
  - Garousi, V., Felderer, M., Mäntylä, M.V. 2019. Guidelines for including grey literature and conducting multivocal literature reviews in software engineering, Information and Software Technology (106), pp. 101-121.

MLR STEPS

- Essentially, has the same steps as a traditional SLR.
  - Except in the search and quality assessment steps.

Search:

  - Search engines like Google.
  - Specialised websites:
    - Archives: arxiv.org, opengrey.eu
    - Q&A: Stackoverflow
    - Special interests: AgileAlliance (www.agilealliance.org), ISTQB (testing www.istqb.org), ISPMA(software product management), etc.

Quality assessment:

  - See Garousi et al. (2019) **(self Study)**

REPLICATION STUDY

Definitions

A study that deliberately repeats a previous study (the "original study") to determine whether its results can be reproduced

(Carver et al., 2013)

- To *reproduce* means to repeat the original study's data analysis on the original study's data.
- To *replicate* means to repeat a study by collecting new data and repeating the original study's analysis on the new data.

https://sigsoft.org/EmpiricalStandards/docs/?standard=Replication

What is the Main Purpose of Replication Studies?

- To establish the credibility of scientific studies
- To increase the validity of originally conducted research
- To generalize the original study

https://sigsoft.org/EmpiricalStandards/docs/?standard=Replication

Essential Attributes

- Discusses the motivation for conducting the replication
- Defines the type of the replication by:

✓ **methodological similarity** (exact, methodological, conceptual) (Dennis and Valacich, 2015)

- Exact replication: same research questions, same method, same context.
- Methodological replication: same research questions, same method, different context.
- Conceptual replication: same research questions, different method, different context.

✓ **overlap** (partial, complete, extended) (Carver, 2010)

- Partial replication: addresses a subset of the original research questions.
- Complete replication: addresses each of the original research questions.
- Extended replication: addresses each of the original research questions and additional ones.

✓ **participants** (internal, external, mixed) (da Silva et al., 2012)

- Internal replication: the replicating team is the same as the original study's team.
- External replication: the replicating team is different from the original study's team.
- Mixed replication: overlaps exist between the replicating team and the original study's team.

https://sigsoft.org/EmpiricalStandards/docs/?standard=Replication

Jeffrey C. Carver. 2010. Towards reporting guidelines for experimental replications: A proposal.

Essential Attributes

- describes the original study including:
  - ✓ research questions
  - ✓ design
  - ✓ participants (their number and any relevant characteristics)
  - ✓ artifacts
  - ✓ context variables
  - ✓ major findings
- Describes and justifies any differences from the original study (design, participants, artifacts, procedures, data collection, or analysis)
- Compares the results of the replication to the results of the original study
- Differentiates between results that are consistent and inconsistent with the original study

https://sigsoft.org/EmpiricalStandards/docs/?standard=Replication

Jeffrey C. Carver. 2010. Towards reporting guidelines for experimental replications: A proposal.