Lecture #1-1

⇒ Environment Setup for R

Introduction to Quantitative Research

Design Science Paradigm

Michael Fu

https://michaelfu1998-create.github.io/

Dr. Kla Tantithamthavorn

Senior Lecturer in Software Engineering

http://chakkrit.com @klainfo

Schedule *http://chakkrit.com/teaching/quantitative-research-methods*

Date	Location	Topics
11 November 2024	G.13 Woodside Building (20 Exhibition Walk)	Design Science Paradigm
15 November 2024	G.13 Woodside Building (20 Exhibition Walk)	Statistical Analysis
25 November 2024	G.13 Woodside Building (20 Exhibition Walk)	Modern Regression Analysis &
29 November 2024	G.13 Woodside Building (20 Exhibition Walk)	ML Quality Assurance

Environment Setup for Lecture #2

1. Download and install R and RStudio

https://posit.co/download/rstudio-desktop/

2. Create a working dir on your machine

3. Download the "data" folder to your working dir

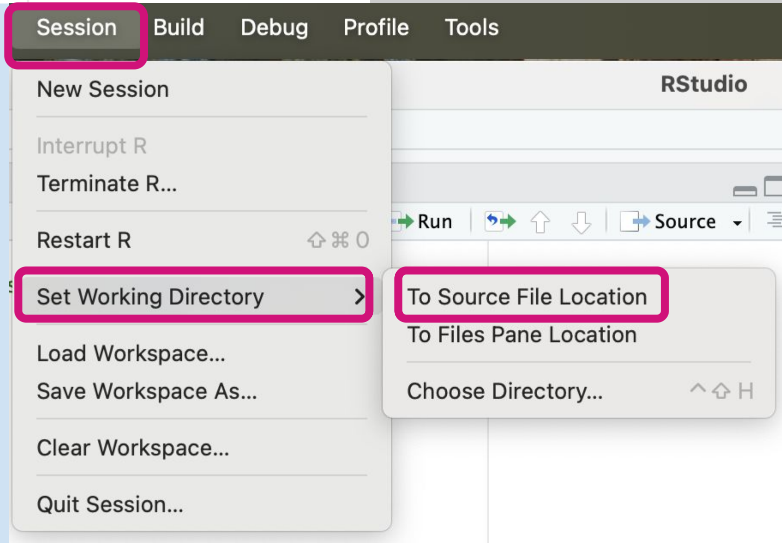
https://shorturl.at/lck0S

Read csv from file and print the first three rows

```r
# load csv from file
df <- read.csv("./data/house_price_data.csv", header=TRUE)
# print the first three rows
head(df, 3)
```

Expected output

```
> df <- read.csv("./data/house_price_data.csv", header=TRUE)
>
> head(df, 3)
  Id MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape LandContour Utilities LotConfig
1  1         60       RL          65    8450   Pave  <NA>      Reg         Lvl    AllPub    Inside
2  2         20       RL          80    9600   Pave  <NA>      Reg         Lvl    AllPub       FR2
3  3         60       RL          68   11250   Pave  <NA>      IR1         Lvl    AllPub    Inside
  LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle OverallQual OverallCond YearBuilt
1       Gtl      CollgCr       Norm       Norm     1Fam     2Story           7           5      2003
2       Gtl      Veenker      Feedr       Norm     1Fam     1Story           6           8      1976
3       Gtl      CollgCr       Norm       Norm     1Fam     2Story           7           5      2001
  YearRemodAdd RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType MasVnrArea ExterQual ExterCond
1         2003     Gable  CompShg     VinylSd     VinylSd    BrkFace        196        Gd        TA
2         1976     Gable  CompShg     MetalSd     MetalSd       None          0        TA        TA
3         2002     Gable  CompShg     VinylSd     VinylSd    BrkFace        162        Gd        TA
```



Lecture #1-2

~~Environment Setup for R~~

⇒ Introduction to Quantitative Research

Design Science Paradigm

Michael Fu

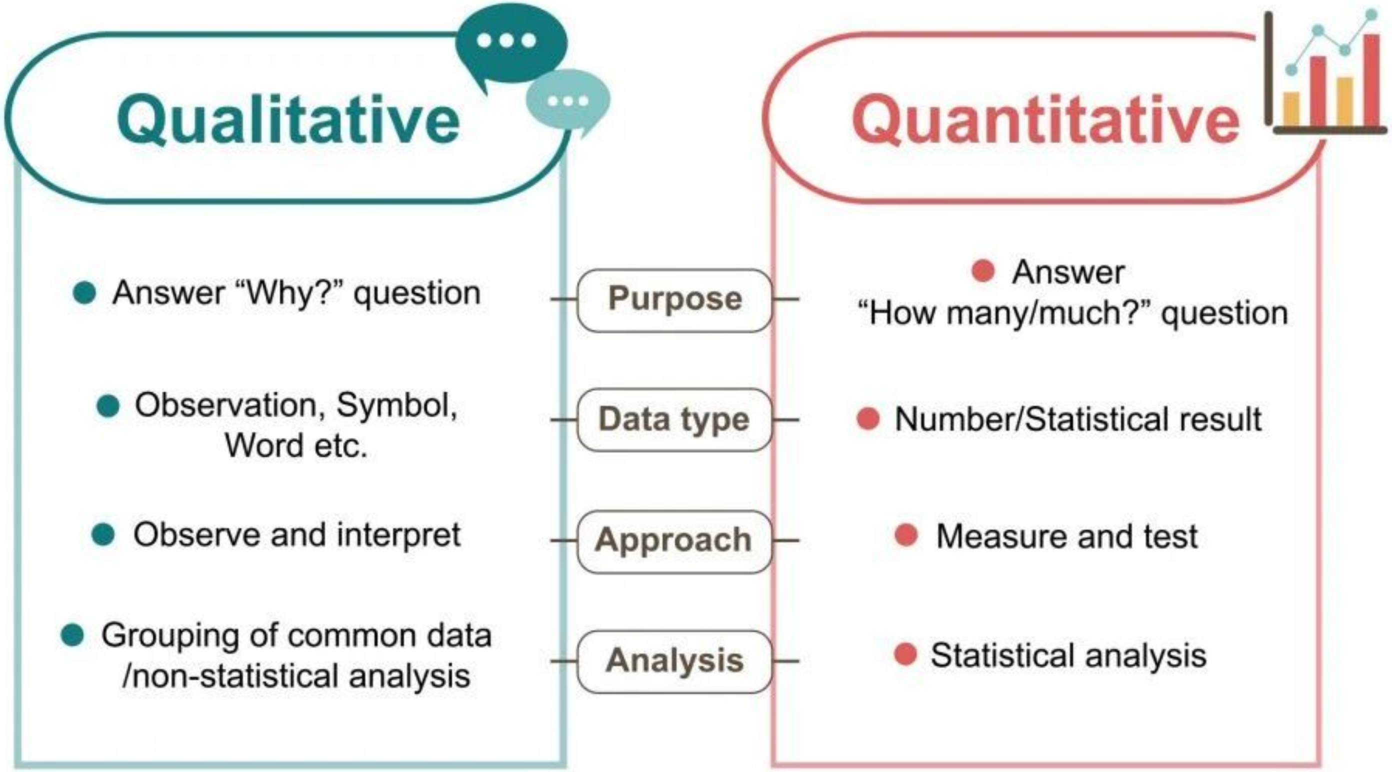
https://michaelfu1998-create.github.io/

Dr. Kla Tantithamthavorn

Senior Lecturer in Software Engineering

http://chakkrit.com @klainfo

Types of Research Design: Qual’ vs Quant’



What Is Quantitative Research? | Definition, Uses & Methods

- Quantitative research is the process of **collecting and analyzing** numerical data. It can be used to find patterns and averages, make predictions, test causal relationships, and generalize results to wider populations.
- Quantitative research is widely used in the natural and social sciences: biology, chemistry, psychology, economics, sociology, marketing, etc. including software engineering and data science

Three types of quantitative methods

- **In descriptive research,** you simply seek an overall summary of your study variables.
- **In correlational research,** you investigate relationships between your study variables.
- **In experimental research,** you systematically examine whether there is a cause-and-effect relationship between variables.

Survey is also a Quant! (e.g., likert scales)

Research method	How to use	Example
Experiment	Control or manipulate an independent variable to measure its effect on a dependent variable.	To test whether an intervention can reduce procrastination in college students, you give equal-sized groups either a procrastination intervention or a comparable task. You compare self-ratings of procrastination behaviors between the groups after the intervention.
Survey	Ask questions of a group of people in-person, over-the-phone or online.	You distribute questionnaires with rating scales to first-year international college students to investigate their experiences of culture shock.

Quantitative research question examples

- What is the demographic makeup of Singapore in 2020? **Descriptive**
- How has the average temperature changed globally over the last century? **Descriptive**
- Does environmental pollution affect the prevalence of honey bees? **Correlation**
- Does working from home increase productivity for people with long commutes? **Experimental**

In SE:

- Do developers productivity increase when using Copilot/ChatGPT? **Experimental**
- How many vulnerabilities that ChatGPT can find? **Descriptive**
- Does Copilot generate vulnerable code? **Experimental**

Quantitative data analysis

- Once data is collected, you may need to process it before it can be analyzed. For example, survey and test data may need to be transformed from words to numbers. Then, you can use statistical analysis to answer your research questions.
- **Descriptive statistics** will give you a summary of your data and include measures of averages and variability. You can also use graphs, scatter plots and frequency tables to visualize your data and check for any trends or outliers.
- **Using inferential statistics**, you can make predictions or generalizations based on your data. You can test your hypothesis or use your sample data to estimate the population parameter.

Examples of descriptive and inferential statistics

- You hypothesize that first-year college students procrastinate more than fourth-year college students. You collect data on procrastination levels of the two groups using 7-point self-rating scales.
- **First,** you use descriptive statistics to get a summary of the data. You find the mean (average) and the mode (most frequent rating) of procrastination of the two groups, and plot the data to see if there are any outliers.
- **Next,** you perform inferential statistics to test your hypothesis. Using a t-test to compare the mean ratings of the two groups, you find a significant difference and support for your hypothesis.

Advantages of quantitative research

- **Replication:** Repeating the study is possible because of standardized data collection protocols and tangible definitions of abstract concepts.
- **Direct comparisons of results:** The study can be reproduced in other cultural settings, times or with different groups of participants. Results can be compared statistically.
- **Large samples:** Data from large samples can be processed and analyzed using reliable and consistent procedures through quantitative data analysis.
- **Hypothesis testing:** Using formalized and established hypothesis testing procedures means that you have to carefully consider and report your research variables, predictions, data collection and testing methods before coming to a conclusion.

Disadvantages of quantitative research

- **Superficiality**: Using precise and restrictive operational definitions may **inadequately represent complex concepts.** For example, the concept of mood may be represented with just a number in quantitative research, but explained with elaboration in qualitative research.
- **Narrow focus:** Predetermined variables and measurement procedures can mean that you ignore other relevant observations.
- **Structural bias:** Despite standardized procedures, structural biases can still affect quantitative research. **Missing data, imprecise measurements or inappropriate sampling methods** are biases that can lead to the wrong conclusions.
- **Lack of context:** Quantitative research often uses unnatural settings like laboratories or fails to consider historical and cultural contexts that may affect data collection and results.

Lecture #1-3

~~Environment Setup for R~~

~~Intro to Quantitative Research Methods~~

⇒ Design Science Paradigm

Michael Fu

https://michaelfu1998-create.github.io/

Dr. Kla Tantithamthavorn

Senior Lecturer in Software Engineering

http://chakkrit.com @klainfo

Research Paradigms

Formal Science

- Building internally consistent systems of knowledge (e.g., philosophy, mathematics)
- Empirically void, as it does not related to any empirical observation or validation

Explanatory Science

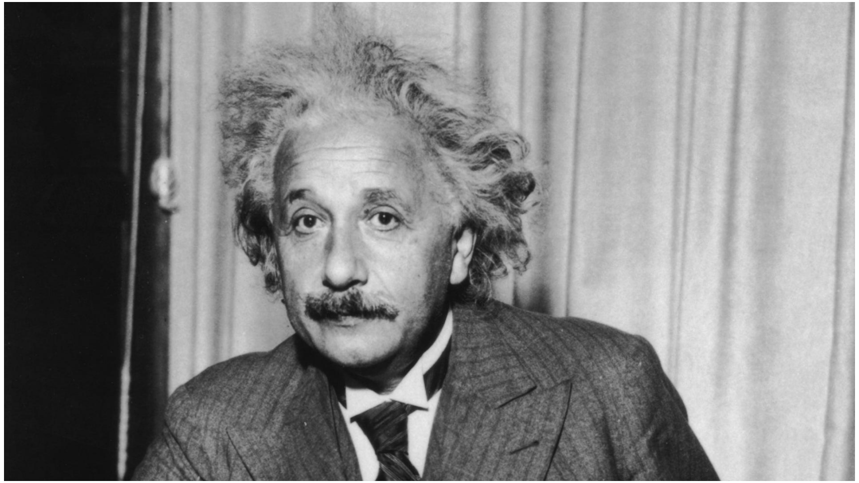
- Describe and explain phenomena that exist, without intervention (e.g., natural science, social science)

Design Science (Today's Focus)

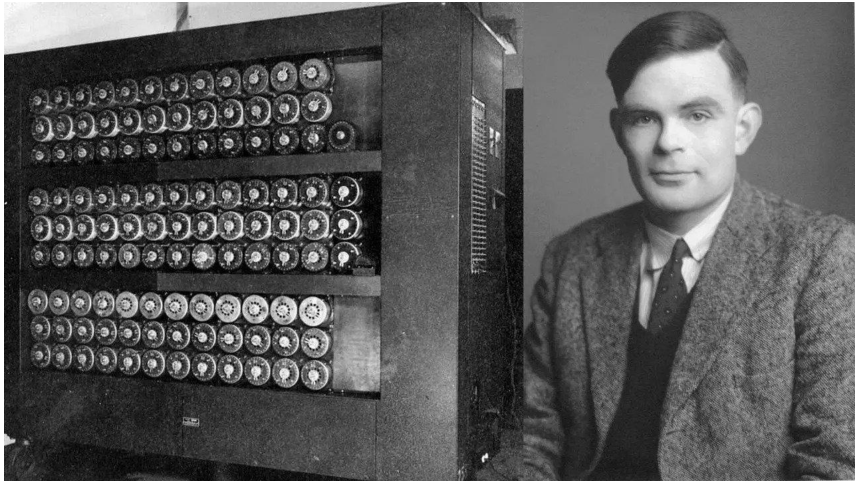
- Understand and improve human-made designs in an area of practice
  (e.g., engineering science, medical science)
- In SE point of view, design science is a research paradigm that helps frame research and aims to improve an area of practice.

Activity 1 - Formal, Explanatory, Design Science?

Albert Einstein

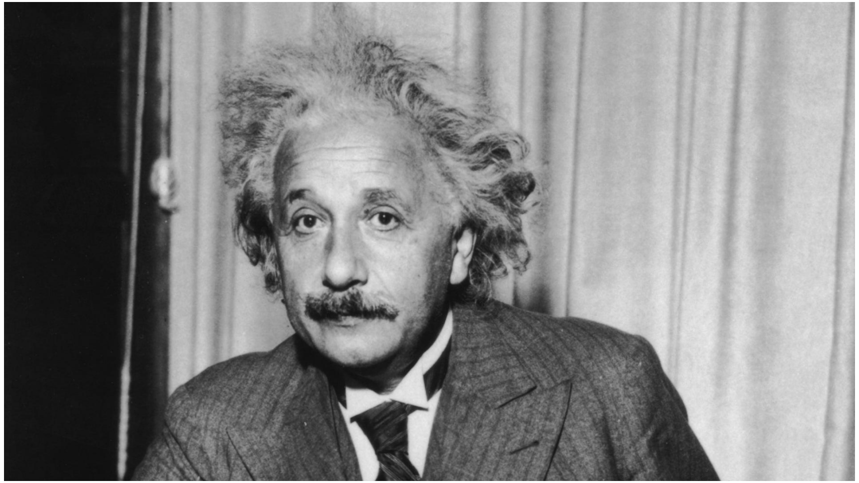


Alan Turing



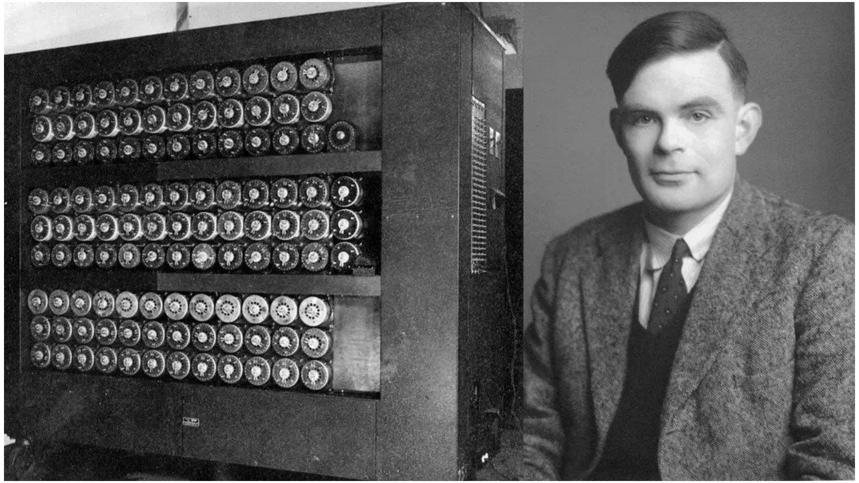
Or… you can share the researcher you admire

Albert Einstein



Formal Science + Explanatory Science

Alan Turing

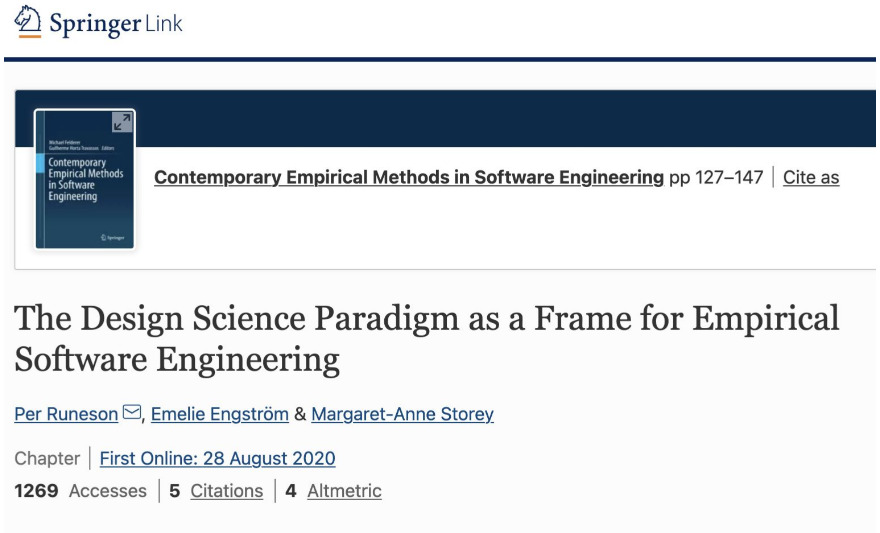


Formal Science + Design Science

No strict answer, just how you interpret it…

Design Science Paradigm

- A framework for articulating and communicating prescriptive software engineering research contributions.
- To embrace problem conceptualization, solution (or artifact) design, and validation of solution proposals, with recommendations for practice phrased as technological rules.



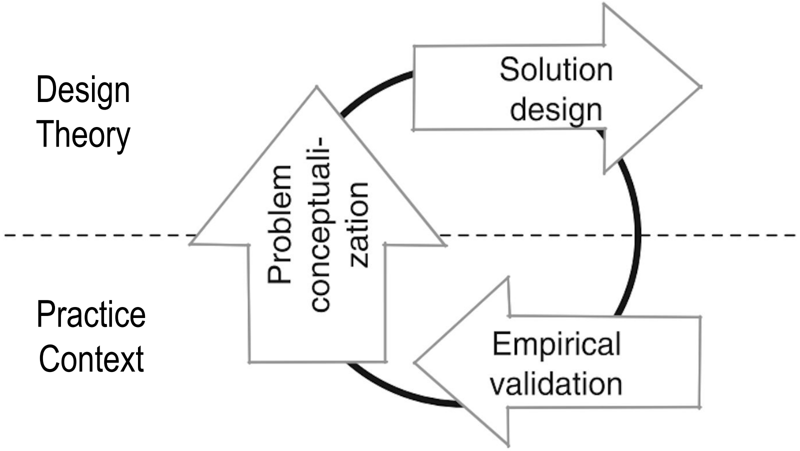
Reference: Runeson, P., Engström, E., & Storey, M. A. (2020). The design science paradigm as a frame for empirical software engineering. In Contemporary empirical methods in software engineering (pp. 127-147). Springer.

Design Science in Software Engineering

A research paradigm that helps frame research and aims to improve an area of practice, i.e., software engineering.

Design science addresses problems by
studying specific problem instances in practice

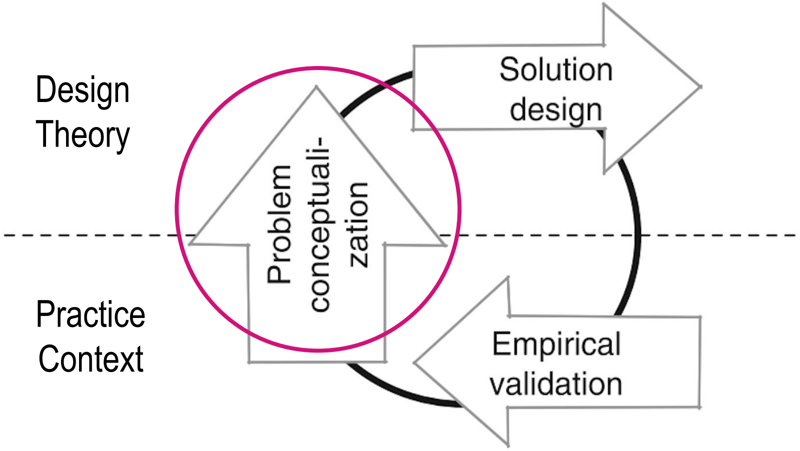
which constitute the research contexts,
where different **research activities**, e.g., conceptualization, solution design, and validation take place.



Research activities diagram

*A single study or research paper **may or may not** contain all the constituents of the design science paradigm

Problem Conceptualization | Describing the Problem

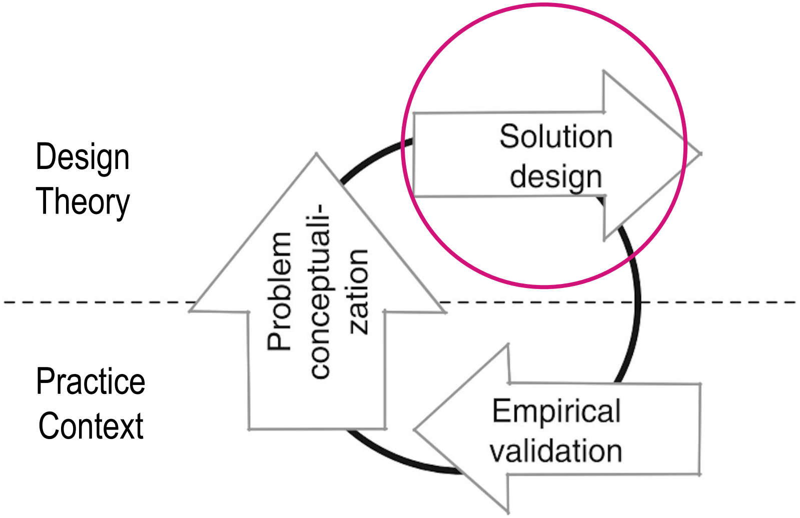


Research activities diagram

Design science research aims to address real practice problems, and thus problem conceptualization is a core constituent of the research.

Understanding a general problem in terms of one or more concrete problem instances is a basis for understanding how this general problem may be solved.

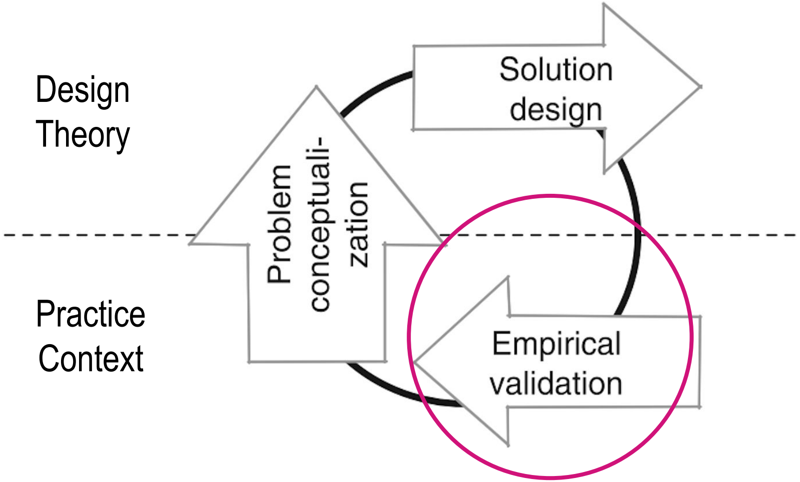
Solution Design | Mapping a problem to a general solution



Research activities diagram

During the exploration of a specific problem instance, it may become clearer what the core of the problem is, thus focusing the potential **solution design** to these areas.

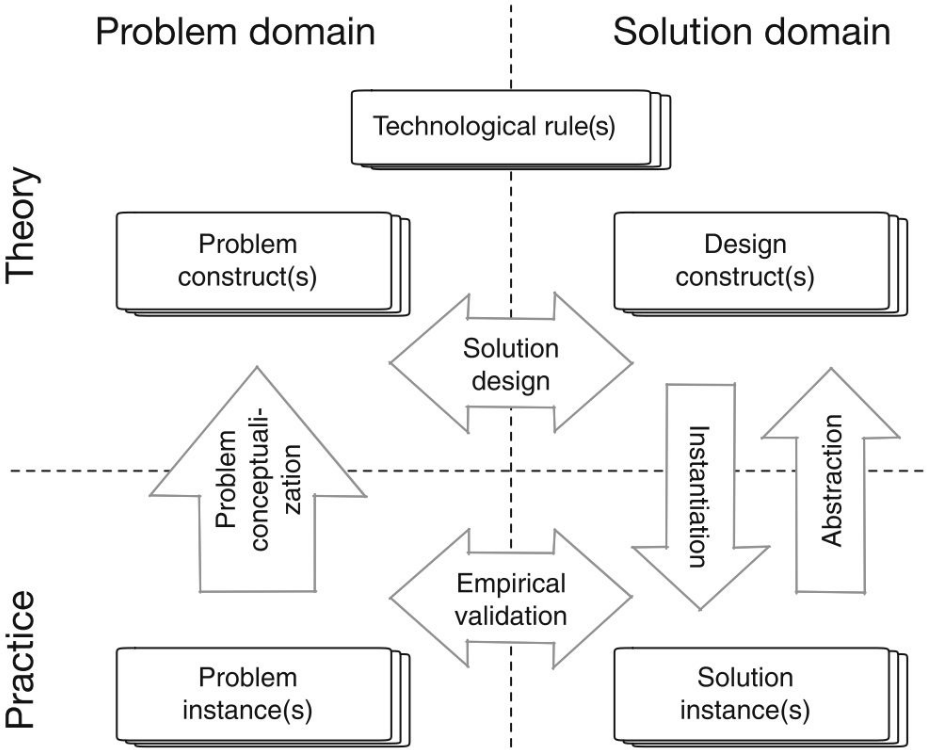
Empirical Validation | How well the implemented solution addresses the problem



Research activities diagram

The primary goal of empirical validation is to **assess** whether the solution proposal is feasible for the problem instance.

Model of Design Science in Software Engineering



Design Science Contributions

Theoretical contribution

- Generalization and scope definition
- Technological rule(s) and the corresponding constructs

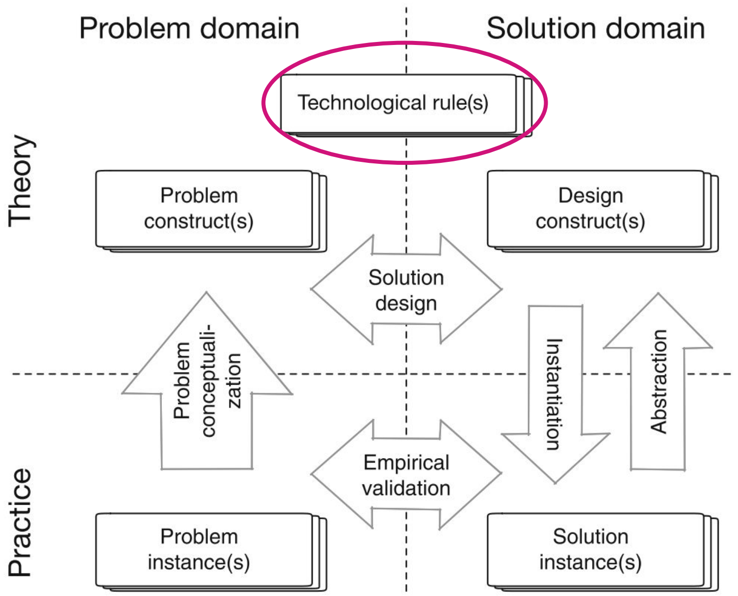
Practical contribution

- Actual problem solving
- Problem and Solution instance(s)

Note. **Arrows** indicate knowledge creating activities that can be performed by researchers/practitioners

Technological Rules | Capturing generalized knowledge about mapping problems and solutions

How to achieve specific outcomes (effects) in a given context by applying certain interventions (e.g., tools, practices, techniques)



Technological rules aim to help software engineers design customized solutions to their specific problems.

A Technological rule frames the research outcome in terms of **effects of interventions**, rather than in terms of a solution to a problem. It could be expressed in the form:

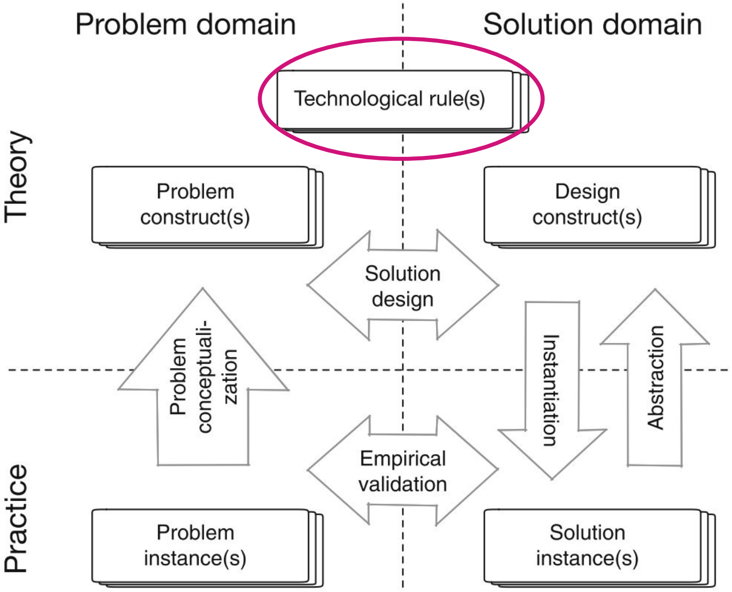
To achieve (avoid) <effect> in <context> (do not) apply <intervention>

- <Effect>: The desired outcome you want to achieve by using a particular intervention.
- <Context>: The specific situation or environment where the intervention will be applied.
- <Intervention>: The tool/practice/technique/framework/guidelines that you apply to achieve the effect.

Note. an **intervention can also be negative**, meaning it can advise against using a certain tool, practice, or technique in a specific context.

Technological Rules | Capturing generalized knowledge about mapping problems and solutions

How to achieve specific outcomes (effects) in a given context by applying certain interventions (e.g., tools, practices, techniques)



Technological rules aim to help software engineers design customized solutions to their specific problems.

A Technological rule frames the research outcome in terms of **effects of interventions**, rather than in terms of a solution to a problem. It could be expressed in the form:

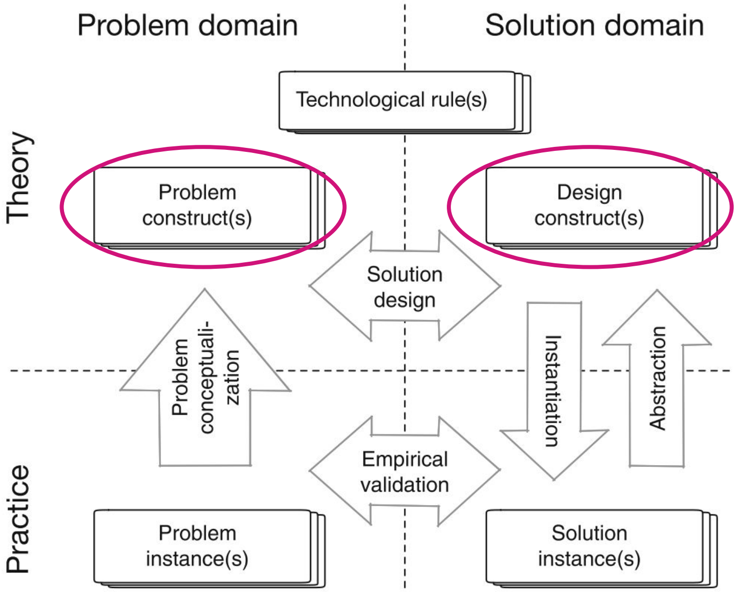
To achieve (avoid) <effect> in <context> (do not) apply <intervention>

- Effect: Avoiding slow performance.
- Context: A small application.
- Intervention: Not using MySQL.

To avoid slow performance in a small application, do not use a database management system (DBMS) like MySQL.

Note. an **intervention can also be negative**, meaning it can advise against using a certain tool, practice, or technique in a specific context.

Problem and Design constructs | Constructs on which we build technological rules

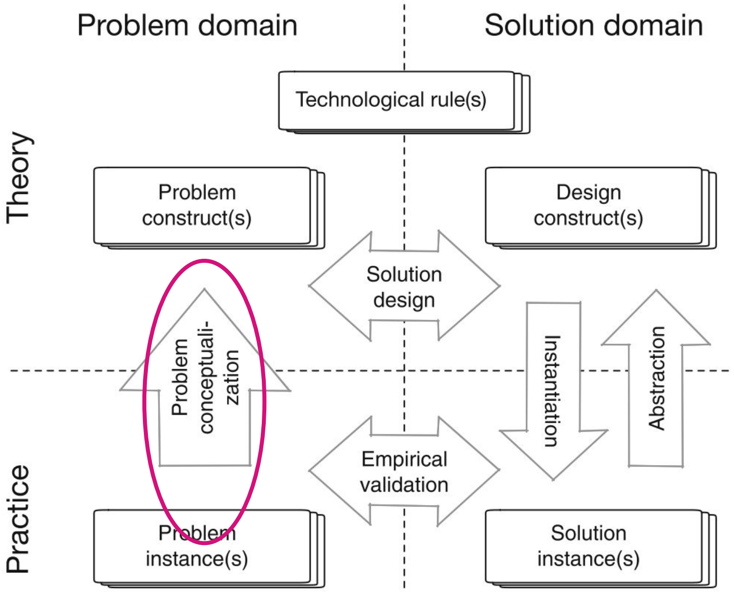


A Construct is the conceptualization of the problem domain and the solution domain.

Problem constructs: Outcome of problem conceptualization, matching corresponding constructs of an envisioned solution.

Solution Constructs: tools, techniques, or frameworks (like taxonomies) that help you organize your thinking and select an appropriate solution to the problem.

Problem Conceptualization | Describing the problem



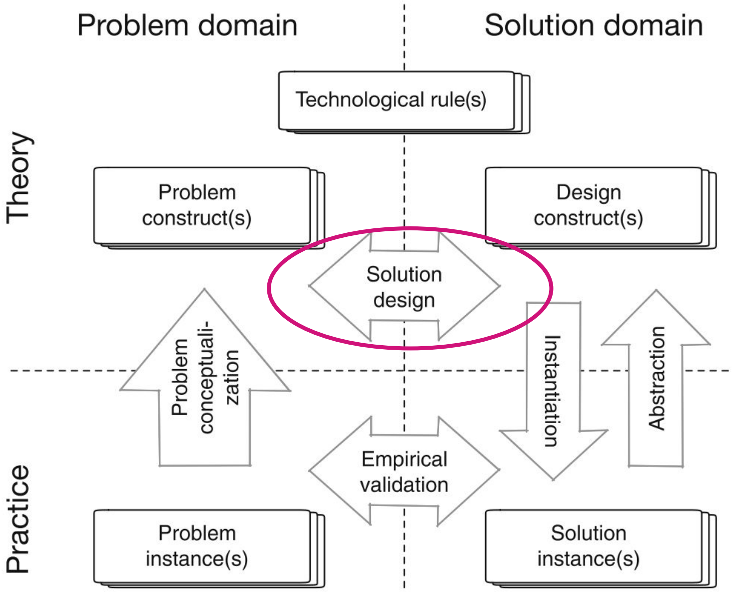
Problem conceptualization is when practitioners, often with researchers, thoroughly **define and understand a problem** within a specific context, such as a case study or real-world observation. This process may **involve considering possible solutions to better frame the problem.**

Outcome: problem constructs, matching corresponding constructs of an envisioned solution, e.g.,

- A taxonomy that is used to articulate a technological rule or classify a set of technological rules in a research review
- A conceptual model or a conceptualization approach that helps describe a problem in terms of an envisioned solution.

The problem conceptualization is tightly connected to the solution design and cannot be performed in isolation.

Solution Design | Selecting the solution for the problem

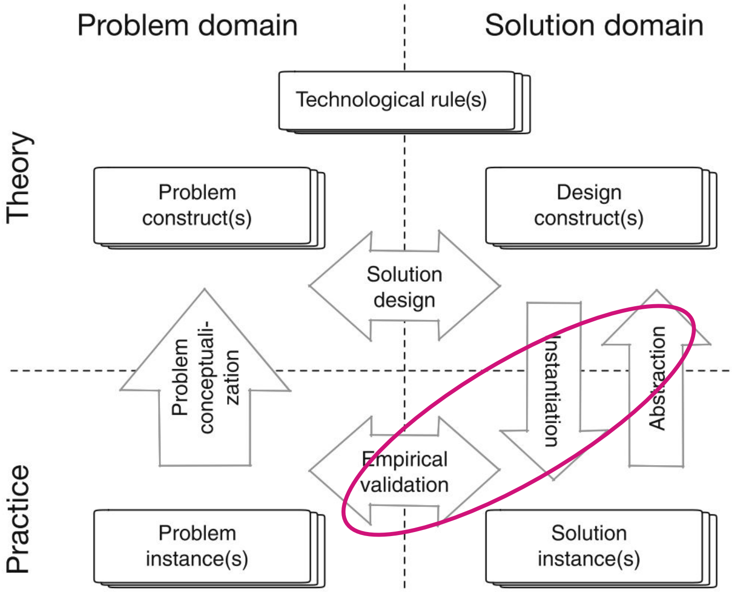


Solution Design refers to the activity of mapping the problem to a general solution. Solution design could be done in several abstraction levels from stakeholder level to implementation details level.

Outcome: Solutions (design knowledge) to the problems. The solutions could be for different abstraction levels, e.g., from the stakeholder's problem description to the level of implementation details (such as choice of algorithm).

Validation, Instantiation, and Abstraction |

The processes to validate technological rule(s).



1) **Abstraction** refers to the activity of identifying the key design decisions for a defined scope of validity of a solution;
2) **Instantiation** refers to the activity of implementing the solution in context
3) **Empirical validation** refers to an evaluation of how well the implemented solution addresses the problem.

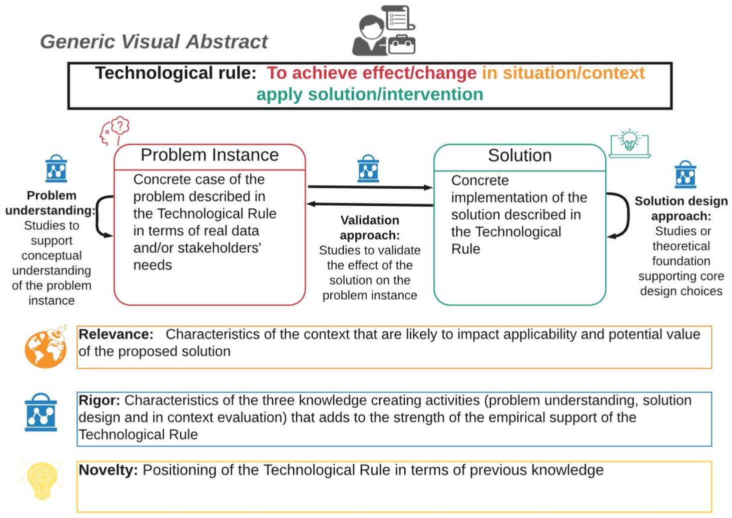
Technological rule must be instantiated before validation, preferably in multiple cases of problem-solution pairs with mixed-methods research to for a better validity strength.

In "*To achieve <effect> in <context> apply <intervention>*,"
- intervention is the object of validation study,
- context refers to where the research is conducted,
- and the expected effect defines the validation criteria.

Design Science Framework in Software Engineering

This visual abstract template can be used as a tool to analyze design science constructs in software engineering research.

The template aims to (1) capture the key takeaway from a research study to help researchers assess the research contribution, (2) build knowledge iteratively, and (3) communicate research to practitioners.



This template covers

- theoretical contribution in terms of *technological rule*
- Instantiation in terms of *problem-solution pair*
- empirical/theoretical support for *problem conceptualization* and *solution design*
- *relevance of the research*
- *rigor of the research activities*
- *a statement about what makes the technological rule novel*

Technological rule: To achieve more effective assignment of bugs to teams in large scale industrial contexts use ensemble-based learning to automate bug assignment

Related work quantifies the scale of the problem in real projects

Problem Instance

Labour-intensive and error-prone bug assignment in two companies from the telecom and the automation domain respectively

Validation approach: Application of solution to bug data

Solution

Stacked generalization (SG), combining several classifiers, automates bug assignment

Applied state-of-the art ensemble learner

Relevance: Problem observed in real projects: Eclipse Platform (Anvik and Murphy 2011), the Mozilla foundation (Bhattacharya et al. 2012), and at Ericsson (Jonsson et al. 2012). Evaluated on data from Telecom and Automation domains.

Rigor: Evaluated in 5 real projects across 2 companies/domains, on 50 k bug reports, using K-fold cross-validation and sliding window validation.

Novelty: Precision in automated bug assignment on par with manual (50-89%), which makes it useful in practice, saving cost and time. SG consistently outperforms individual classifiers. When training SG, aim for at least 2,000 bug reports in the training set. Relying only on K-fold cross-validation is not enough to evaluate automated bug assignment.

Automated bug assignment: Ensemble-based machine learning in large scale industrial contexts

Jonsson et al. 2016

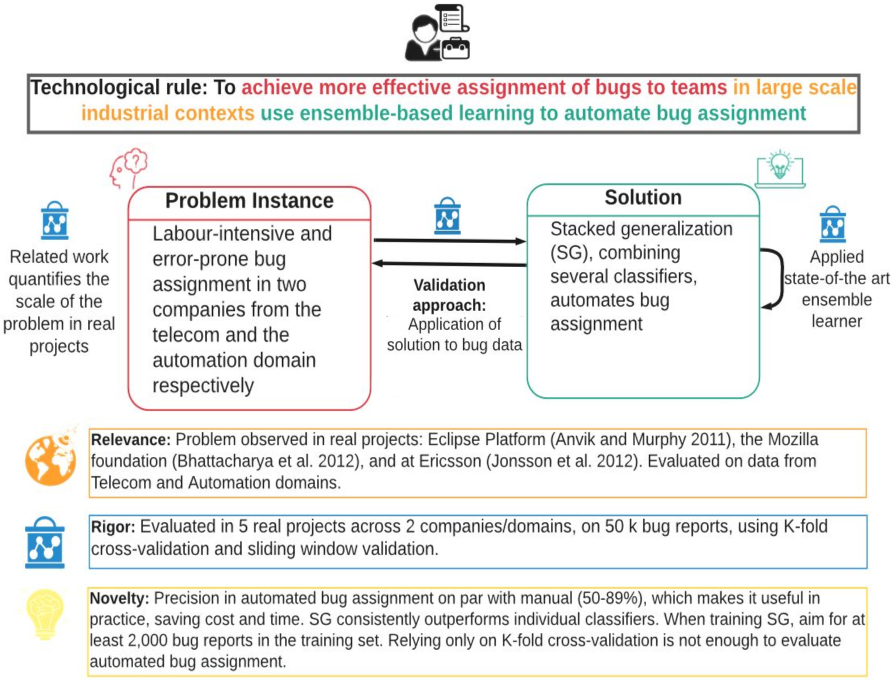
Conference/Journal Paper Review

MONASH University MONASH INFORMATION TECHNOLOGY

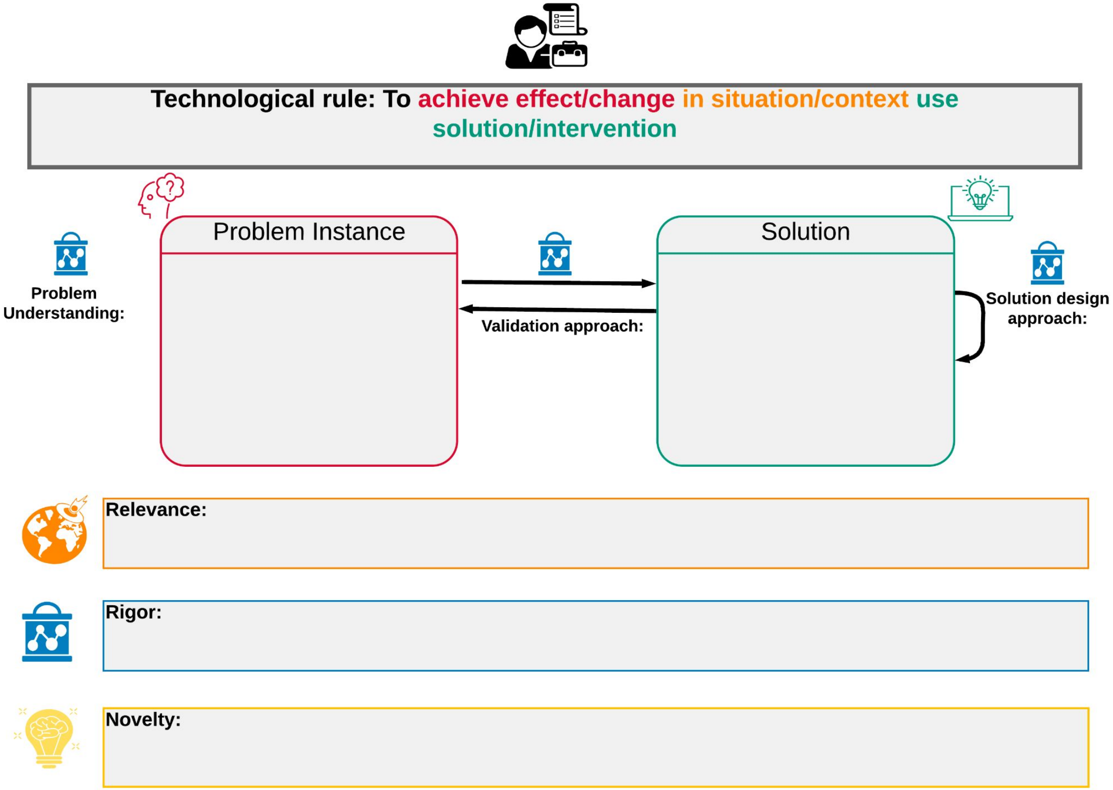
[Paper Summary]

…

- **Rigor: …**
- **Relevance: …**
- **Novelty: …**
- **Presentation: …**
- **Verifiability & Transparency: …**



Activity 2 https://shorturl.at/xjeh1



Paper Title: