Scale. Fagel, Brussels. Fondation Atelier Marcel Hastir

First of all, I would like to thank the chairs for inviting me.

Designing empirical studies often requires balancing different needs and challenges. In this talk we review several recent empirical studies from the methodological point of view: what are the decision points considered by the researchers? What alternatives could have been considered? Many of these decisions are not inherently right or wrong but need to be explicitly discussed to appreciate the results and the limitations of the study. SIESTA participants are encouraged to bring their own research papers/preprints - during the session you will be asked to discuss your work with your neighbour and jointly reflect on explicit, and even more important implicit decisions embedded in their study designs and the impact of these decisions.

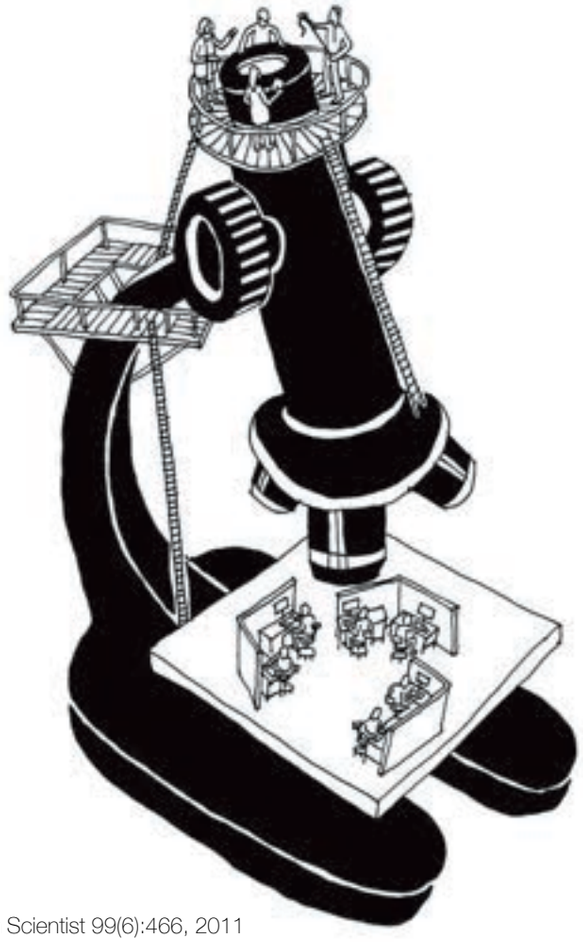


Illustration by Tom Dunne. American Scientist 99(6):466, 2011

Before discussing the trade-offs let us ensure that we are all on the same page when it comes to what the word “empirical” means. Empirical comes from εμπειρία (Greek) - experience. In the empirical world, **observation is the only source of knowledge, while reasoning or intuition can be source of belief or conjecture**. In software engineering researchers describe **how software is being developed** to suggest best practices how to develop better software (product) and how to develop software better (process).

So in this illustration by Tom Dunne, developers are placed on the stage of the microscope, while we as researchers are above, looking through the eye-piece. Notice that we have three objective lenses here: as researchers we select a right lens depending on what we want to study.

Scale. Fagel, Brussels. Fondation Atelier Marcel Hastir

When we are selecting the lens, the research method, we need to be mindful that each decision is associated with strengths and weaknesses and that we need to balance them.

Dan Bennett. Big dog sniffs small Dachshund at the beach. Wikicommons

And of course, these decisions should be taken both in large and in small. Decisions in large are related to such choices to our goals, do we want maximise generalisability of the study or our confidence in causal implications; decisions in small are often related to choices of datasets, tools and instruments.

When it comes to decisions in **large**, one of the oppositions we need to take into account is the opposition between targeting generality and targeting specificity.

This is a generic car, it is not a recognisable brand. Observations that we can derive from such a generic model can be expected to hold very **broadly** - for example, this care has four wheels and two wing mirrors; but they are also **shallow, not** necessarily **useful** and might **miss corner cases**. We do not know, for example, what would the maximal speed of this car.

Klaas-Jan Stol, Brian Fitzgerald: The ABC of Software Engineering Research. ACM Trans. Softw. Eng. Methodol. 27(3): 11:1-11:51 (2018)

This a prototype vehicle designed in China. It still has two wind mirrors but it only has two wheels! We can answer more specific questions such as the maximal speed and derive more specific conclusions. However, these conclusions might only hold for this particular car!

Klaas-Jan Stol, Brian Fitzgerald: The ABC of Software Engineering Research. TOSEM 27(3): 11:1-11:51 (2018)
https://www.ecagroup.com/media-picture/5573-3162-1653-eca-group-driving-simulation-ef-car-14.jpg

This is not even a car but a car simulator.

CONTROLLABLE X REALISTIC?

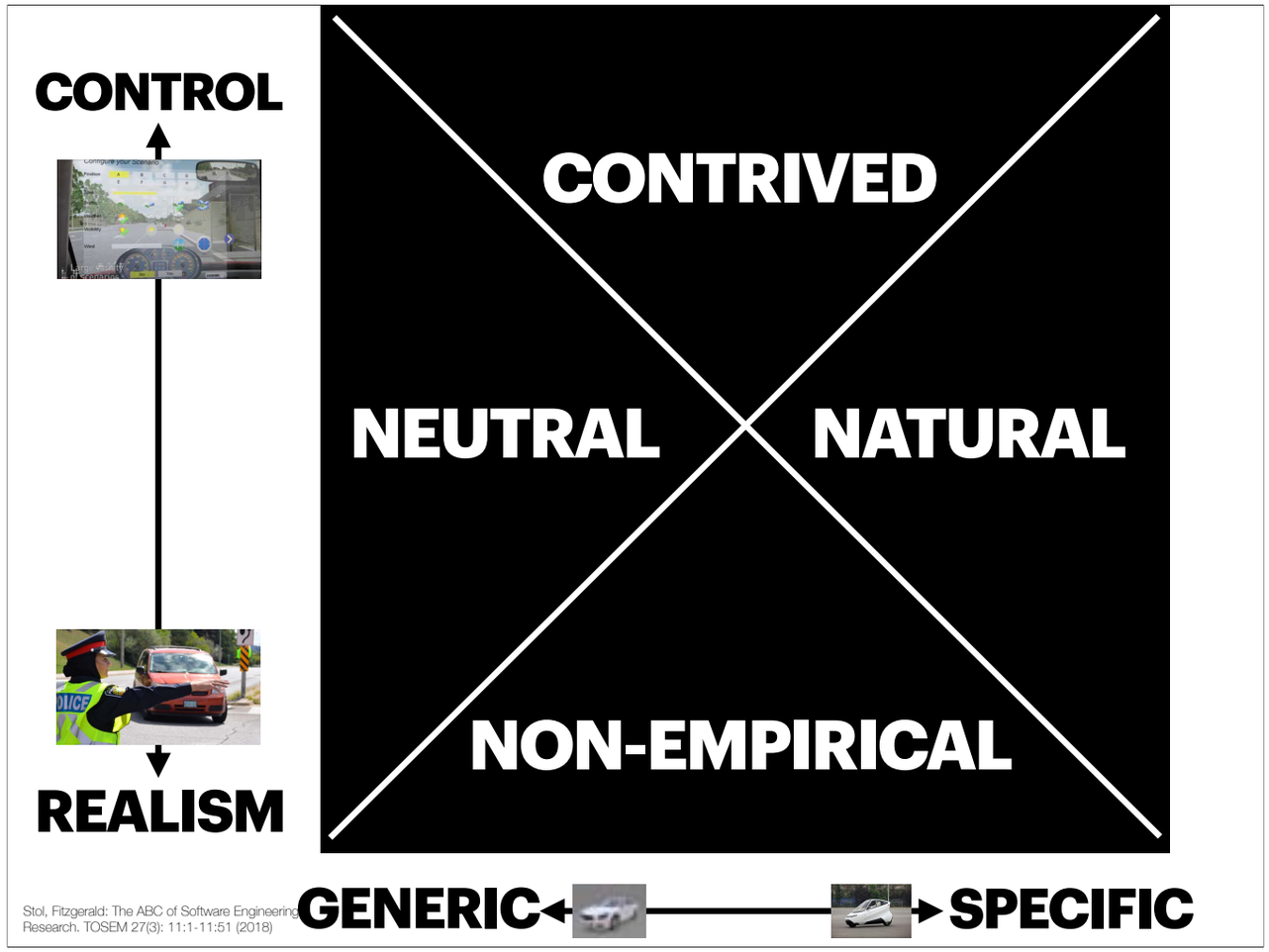
Klaas-Jan Stol, Brian Fitzgerald: The ABC of Software Engineering Research. TOSEM 27(3): 11:1-11:51 (2018)
https://www.ecagroup.com/media-picture/5573-3162-1653-eca-group-driving-simulation-ef-car-14.jpg

An important advantage of the simulator is that we can control parameters that we would never be able to control in reality such as traffic, weather, visibility or wind, and understand how these parameters affect drivers' driving style. However, this is an unrealistic situation since drivers do not sit in a chair, or reality might be complex than a simulation.

REALISTIC BUT NOT CONTROLLABLE

Klaas-Jan Stol, Brian Fitzgerald: The ABC of Software Engineering Research. TOSEM 27(3): 11:1-11:51 (2018)
https://www.peelpolice.ca/en/resourcesGeneral/Homepage-Banner/officer-road-safety-direct-trafficedit.jpg

A very different scenario is on this photo. Here a police officer checks the behaviour the car driver. This is a real driver in the real context but one cannot control the factors that can affect the driver's behaviour such as weather, wind, visibility or traffic…

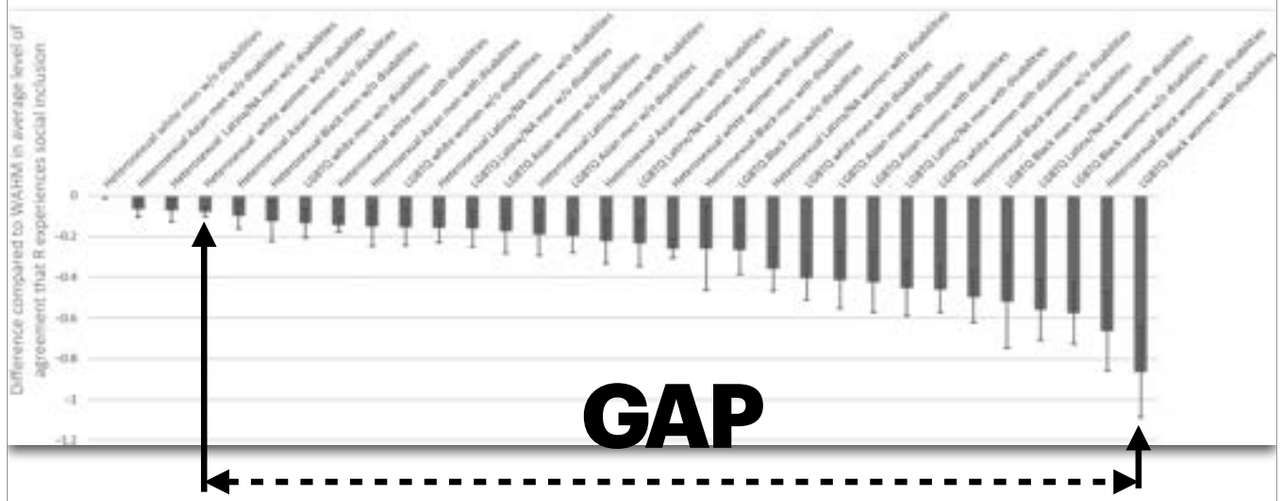


Here we have four quadrants of research settings: natural, contrived, neutral and non-empirical.
Natural: rich in context and no manipulations
Contrived: all under control but unrealistic
Neutral: everything is neutralised, individual differences are blurred
Non-empirical: not immediately based on observations

So, let us check an example. Not the English musician, singer, songwriter and record producer known as Example, but an example of an empirical study.

https://freerangestock.com/photos/142291/three-female-coworkers-talking-and-smiling-while-standing-in-the-workplace.html

We have been studying experiences of women in software development since 2013… However, as we have been exploring this domain we have started realising that it is not enough to look at women in software engineering as a homogenous group: obviously different women might have different experiences and we need to understand their experiences.



Cech. The intersectional privilege of white able-bodied heterosexual men in STEM. Science Advances 8(24), 2022

Using survey data of U.S. STEM professionals (N = 25,324), this study of Erin Cech examines whether white able-bodied heterosexual men (WAHM) are uniquely privileged in STEM. The results show that WAHM experience better treatment and rewards in STEM compared with members of all 31 other intersectional gender, race, sexual identity, and disability status categories. This figure shows that WAHM experience more social inclusion; similar figures show higher professional respect, career opportunities, salaries and persistence intentions (compared to STEM professionals in 31 other intersectional groups). https://www.science.org/doi/10.1126/sciadv.abo1558

What **age- and gender-specific experiences** have veteran software developers of marginalized genders had in their careers?

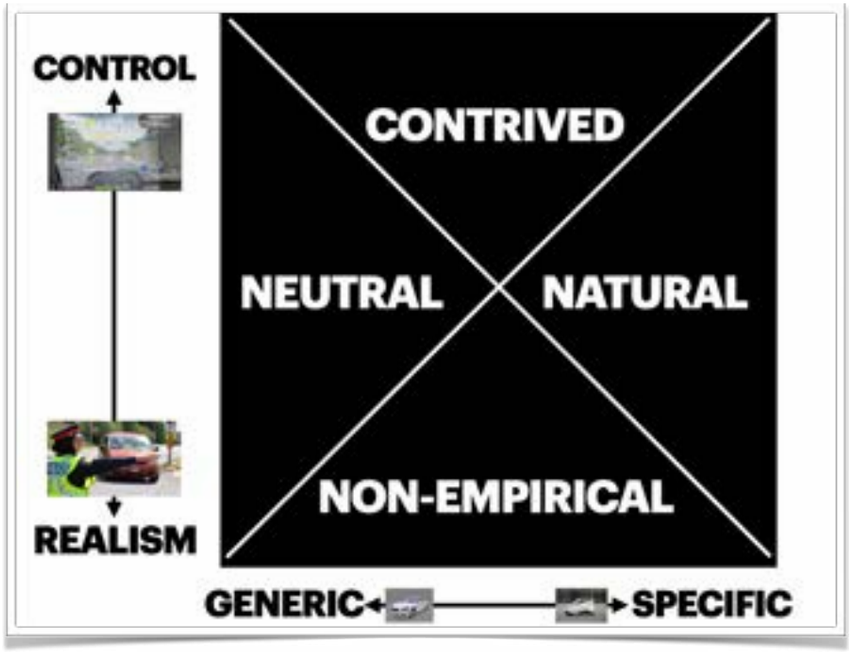
Sterre van Breukelen, Ann Barcomb, Sebastian Baltes, Alexander Serebrenik: "STILL AROUND": Experiences and Survival Strategies of Veteran Women Software Developers. ICSE 2023: 1148-1160

Gender and age are the most studied diversity axes in social software engineering research.

- Hence, our research questions were:…
- In our **research questions and sampling approach, we included all marginalized genders**, but our participants identified as women plus one non-binary person who identified as woman for most of their career.
- Hence our results focus on women, while our research questions are broader.

RQ2. What strategies have veteran software developers of marginalized genders adopted that they perceive as contributing to their survival in software engineering?

What **age- and gender-specific experiences** have veteran software developers of marginalized genders had in their careers?

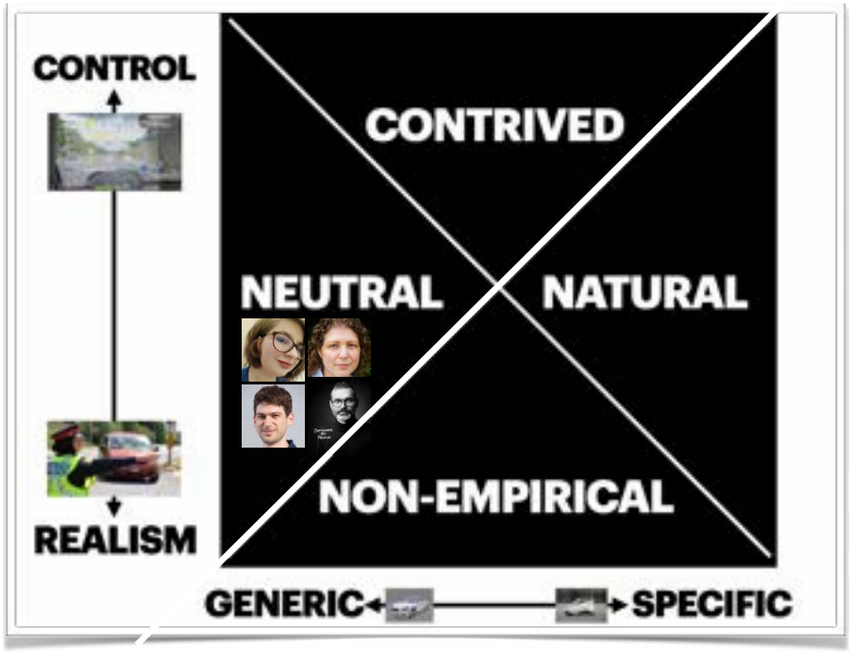


Klaas-Jan Stol, Brian Fitzgerald: The ABC of Software Engineering Research. TOSEM 27(3): 11:1-11:51 (2018)

Sterre van Breukelen, Ann Barcomb, Sebastian Baltes, Alexander Serebrenik: "STILL AROUND": Experiences and Survival Strategies of Veteran Women Software Developers. ICSE 2023: 1148-1160

Question to you: which quadrant would you choose to conduct such a study and why?
I expect: natural (within one specific context) or neutral (across contexts when we are trying to balance or eliminate differences between the contexts)

What **age- and gender-specific experiences** have veteran software developers of marginalized genders had in their careers?

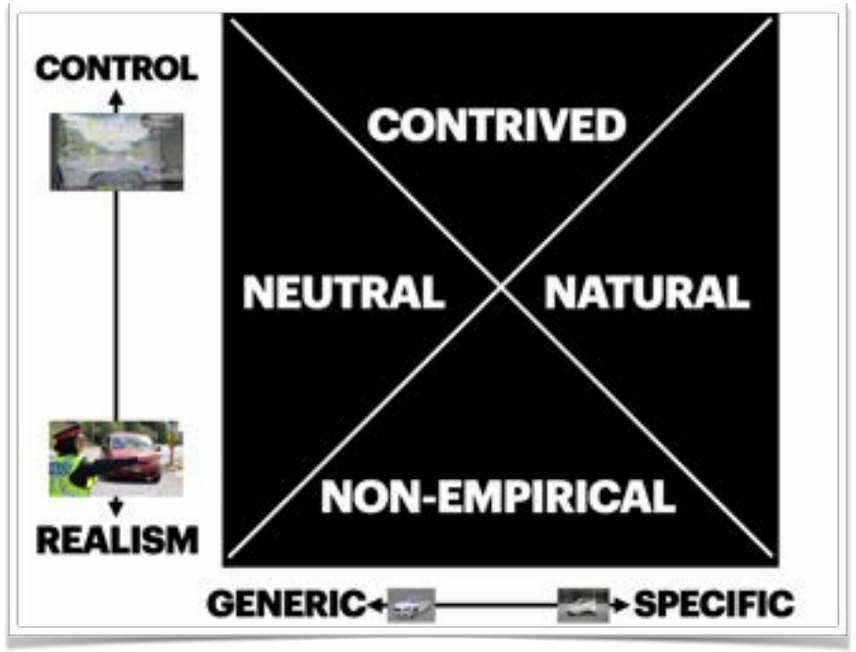


Klaas-Jan Stol, Brian Fitzgerald: The ABC of Software Engineering Research. TOSEM 27(3): 11:1-11:51 (2018)

Sterre van Breukelen, Ann Barcomb, Sebastian Baltes, Alexander Serebrenik: "STILL AROUND": Experiences and Survival Strategies of Veteran Women Software Developers. ICSE 2023: 1148-1160

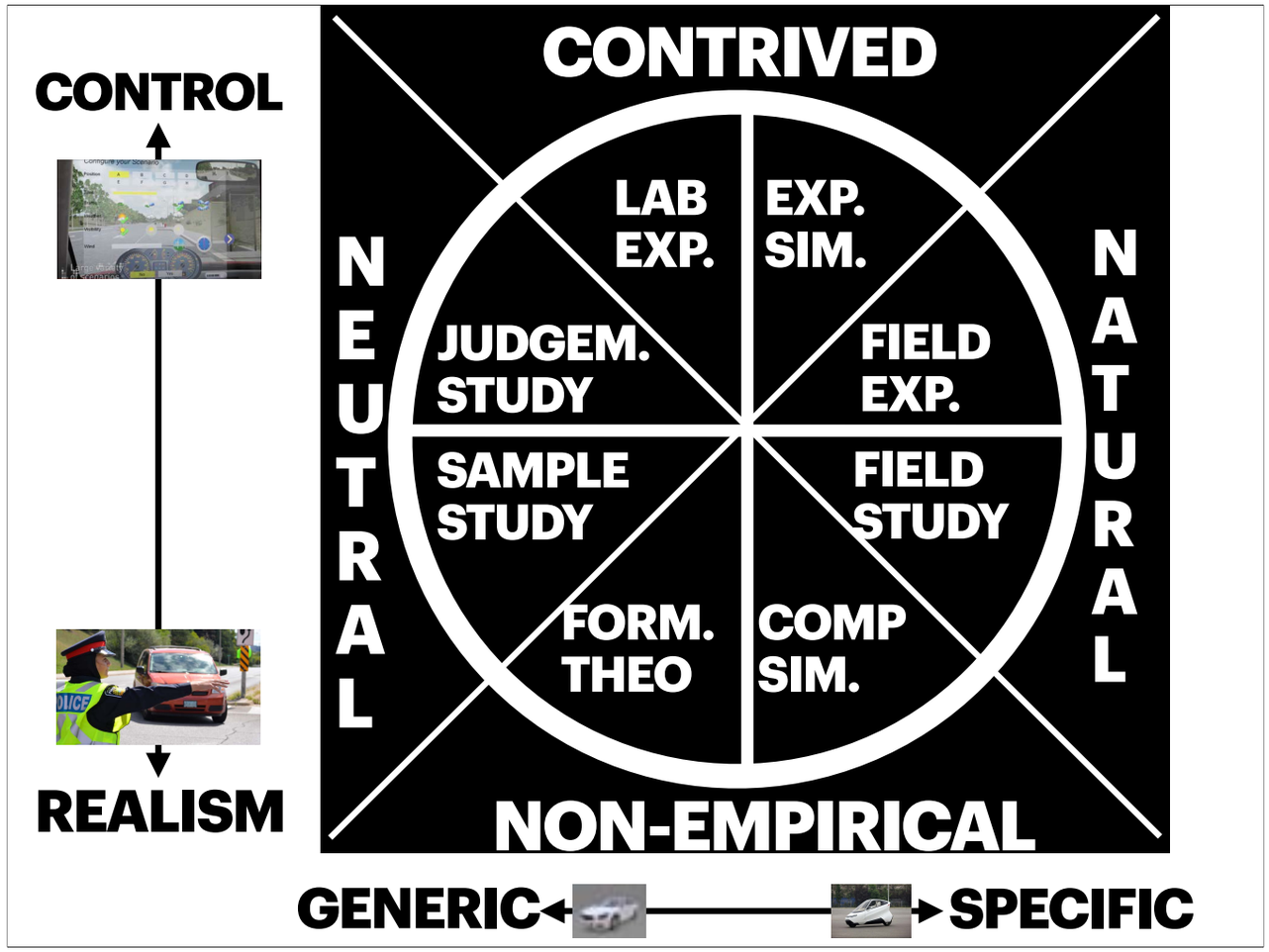
We have chosen a neutral strategy because we wanted to paint a broad picture and understand diverse experiences.

What quadrant was chosen in the paper you have brought today?

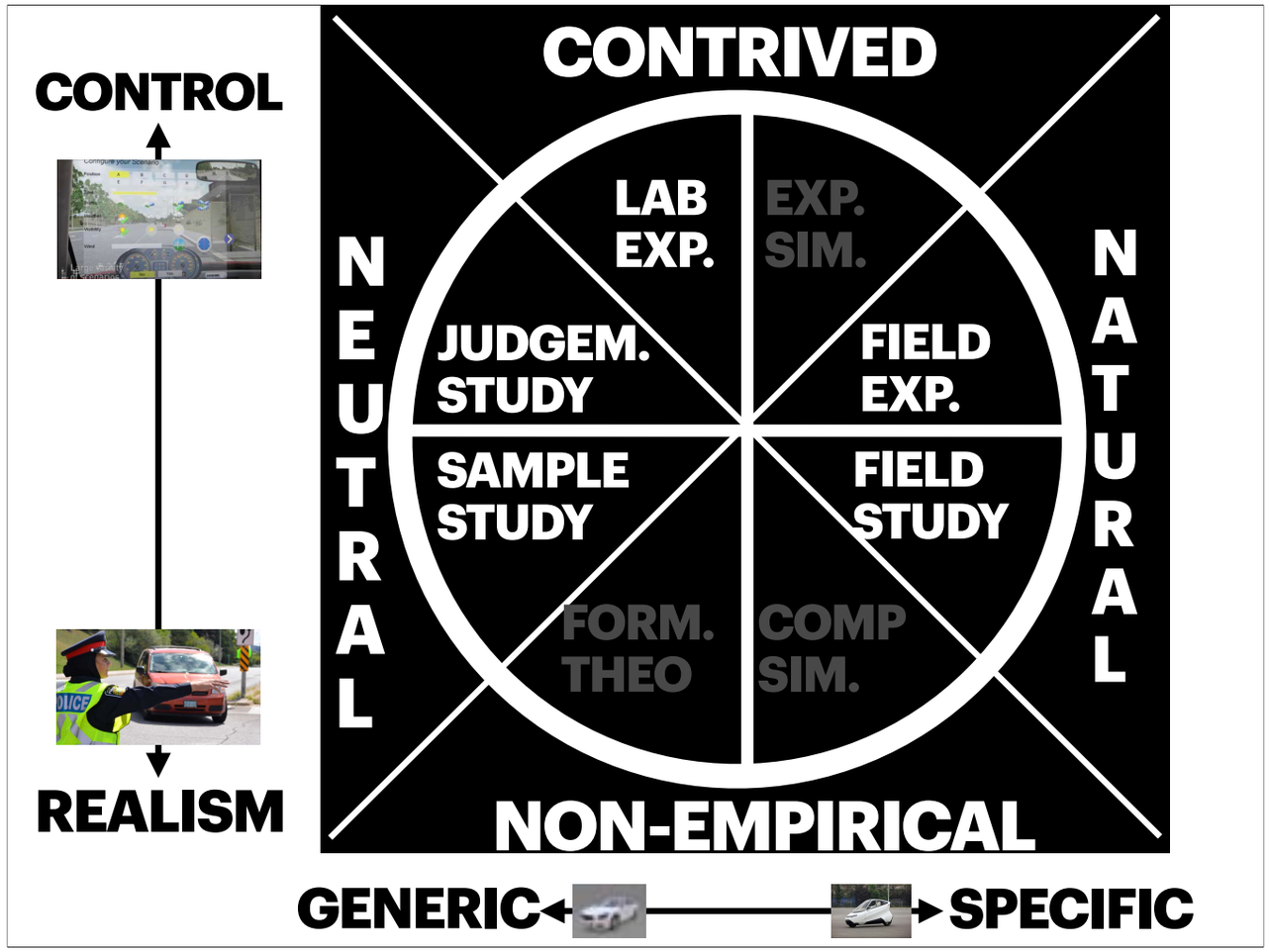


Klaas-Jan Stol, Brian Fitzgerald: The ABC of Software Engineering Research. TOSEM 27(3): 11:1-11:51 (2018)

Question to you: which quadrant would you choose to conduct such a study and why?
I expect: natural (within one specific context) or neutral (across contexts when we are trying to balance or eliminate differences between the contexts)



In general, Stol and Fitzgerald have identified two research strategies for each quadrant, and normally when I am teaching empirical methods we discuss all these strategies.

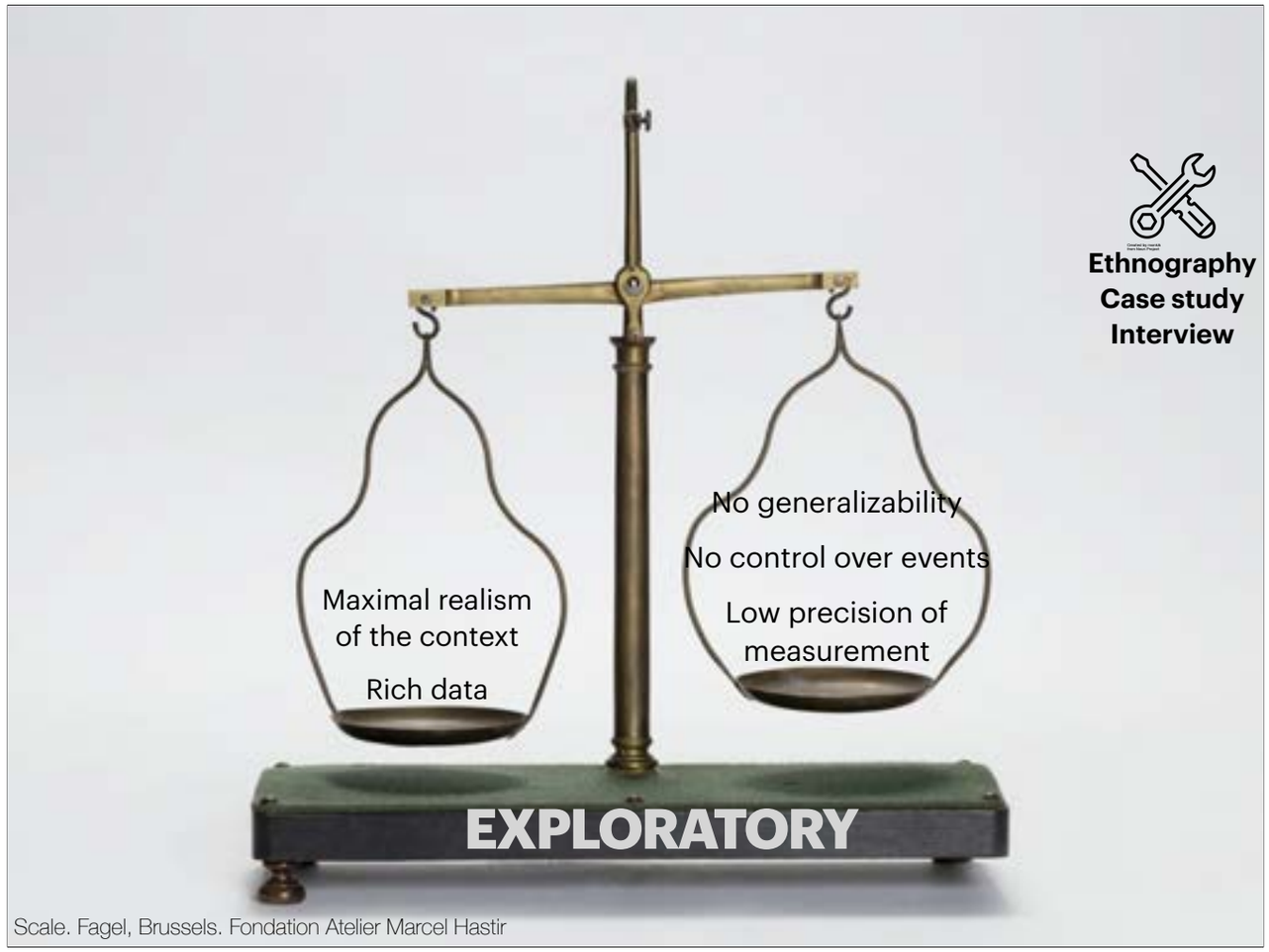


However, today we will discuss only five strategies; four of them are common in empirical software engineering research, and most likely the papers you have been working on fit one or several of these strategies. Why several? Because a single paper can cover multiple studies. Field experiments are not common because they are expensive but they are important to keep in mind.

Field study is a **jungle**. Natural setting that exists before the researcher enters it. Minimal intrusion of the setting so as not to disturb realism, only to facilitate data collection. Facilitates study of phenomena and actors and their behaviour in natural contexts. Exploratory, to understand what's going on, how things work, or to generate hypotheses.

Case study, ethnography, observational study; qualitative data incl. interviews, field notes, archival documents, may include quantitative data.

- No statistical generalizability
- No control over events
- Low precision of measurement



Scale. Fagel, Brussels. Fondation Atelier Marcel Hastir

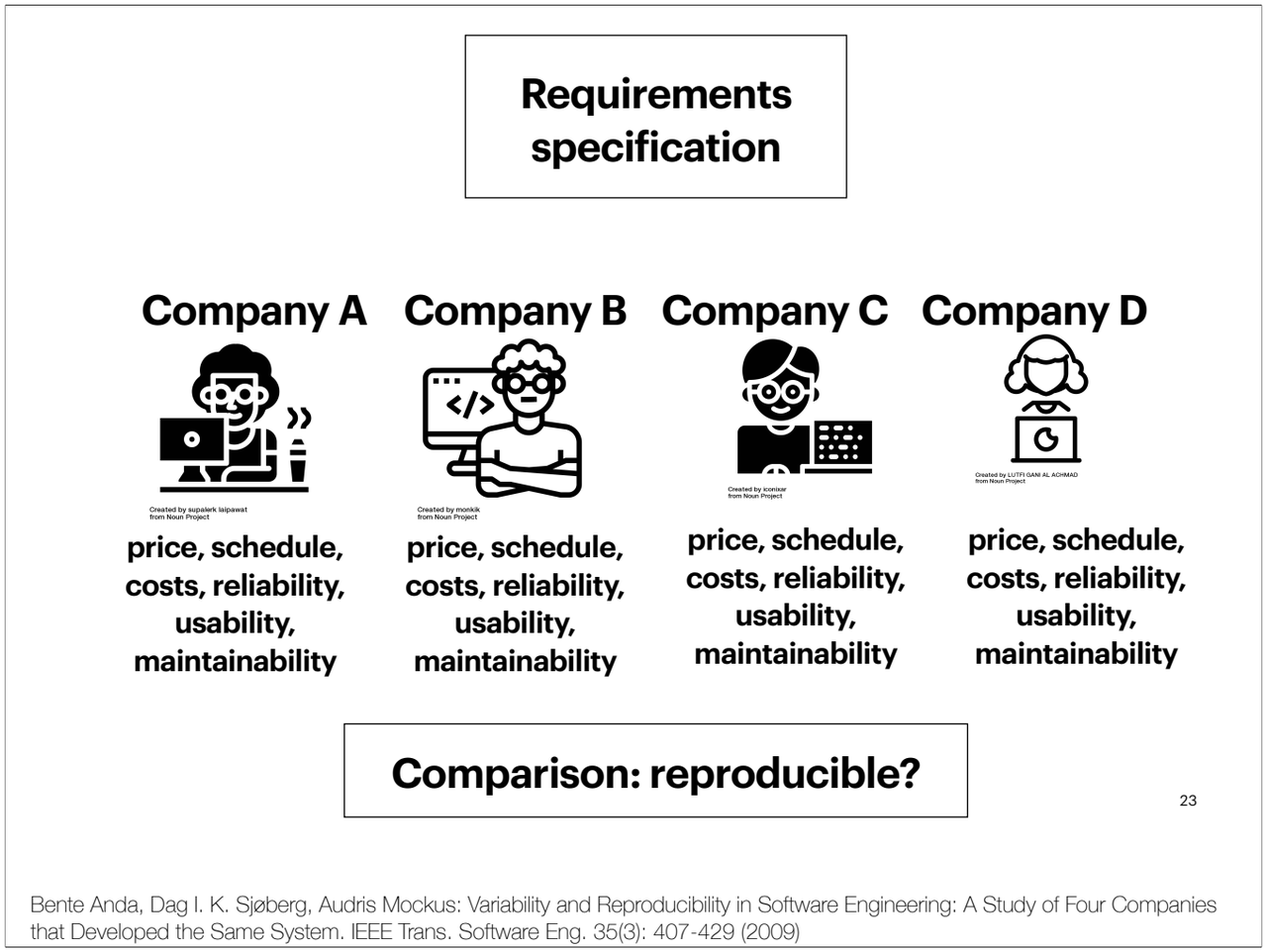
Field study is usually an exploratory strategy, to understand what's going on, how things work, or to generate hypotheses. These hypotheses should be then verified in a separate study

Klaas-Jan Stol, Brian Fitzgerald: The ABC of Software Engineering Research. TOSEM 2018, https://doi.org/10.1145/3211958
https://urbangreennewcastle.org/media/1138/havannah-three-hills-woodland-path.jpg

Nature reserve: Natural, pre-existing setting (in vivo), but some level of intrusion due to the deliberate manipulation of aspects of the setting; study affected by confounding factors. In a nature reserve, flora and fauna can still thrive as normal, but the reserve facilitates the conduct of research, for example, by placing fences so as to separate the wildlife into different treatment groups and evaluate the effects of those treatments.

To investigate, evaluate, or compare techniques, practices, processes, or approaches within a real-world and pre-existing setting.

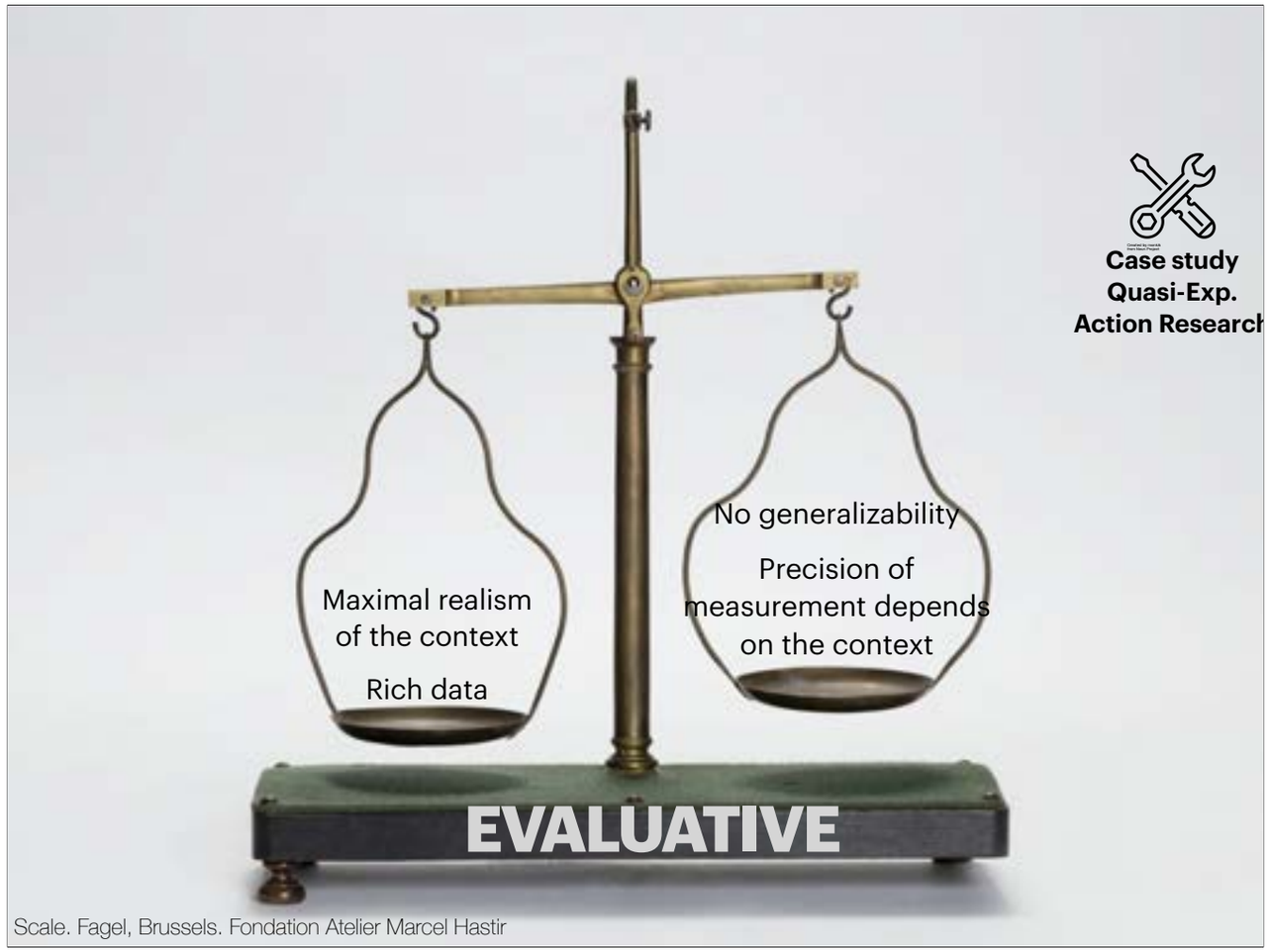
Evaluative case study, quasi-experiment, Action Research; studies may use either quantitative data or qualitative data.
• No statistical generalizability
• Precision of measurement affected by confounding contextual factors



Example of a field experiment.

A longitudinal multiple-case study of variations and reproducibility in software development, from bidding to deployment, on the basis of the same requirement specification. In a call for tender to 81 companies, 35 responded. **Four** of them developed the system independently.

You can also see why field experiments tend to be expensive.



Scale. Fagel, Brussels. Fondation Atelier Marcel Hastir

Field study is usually an exploratory strategy, to understand what's going on, how things work, or to generate hypotheses. These hypotheses should be then verified in a separate study

A **quasi-experiment** is an empirical interventional study used to estimate the causal impact of an intervention on target population without random assignment. Quasi-experimental research shares similarities with the traditional experimental design or randomized controlled trial, but it specifically lacks the element of random assignment to treatment or control.

QUESTION

Caneill et al. present a case study of Debian, which investigates its evolution, rate of change, use and popularity of programming languages, and use and evolution of licenses within Debian. No attempt at generalisation is made.

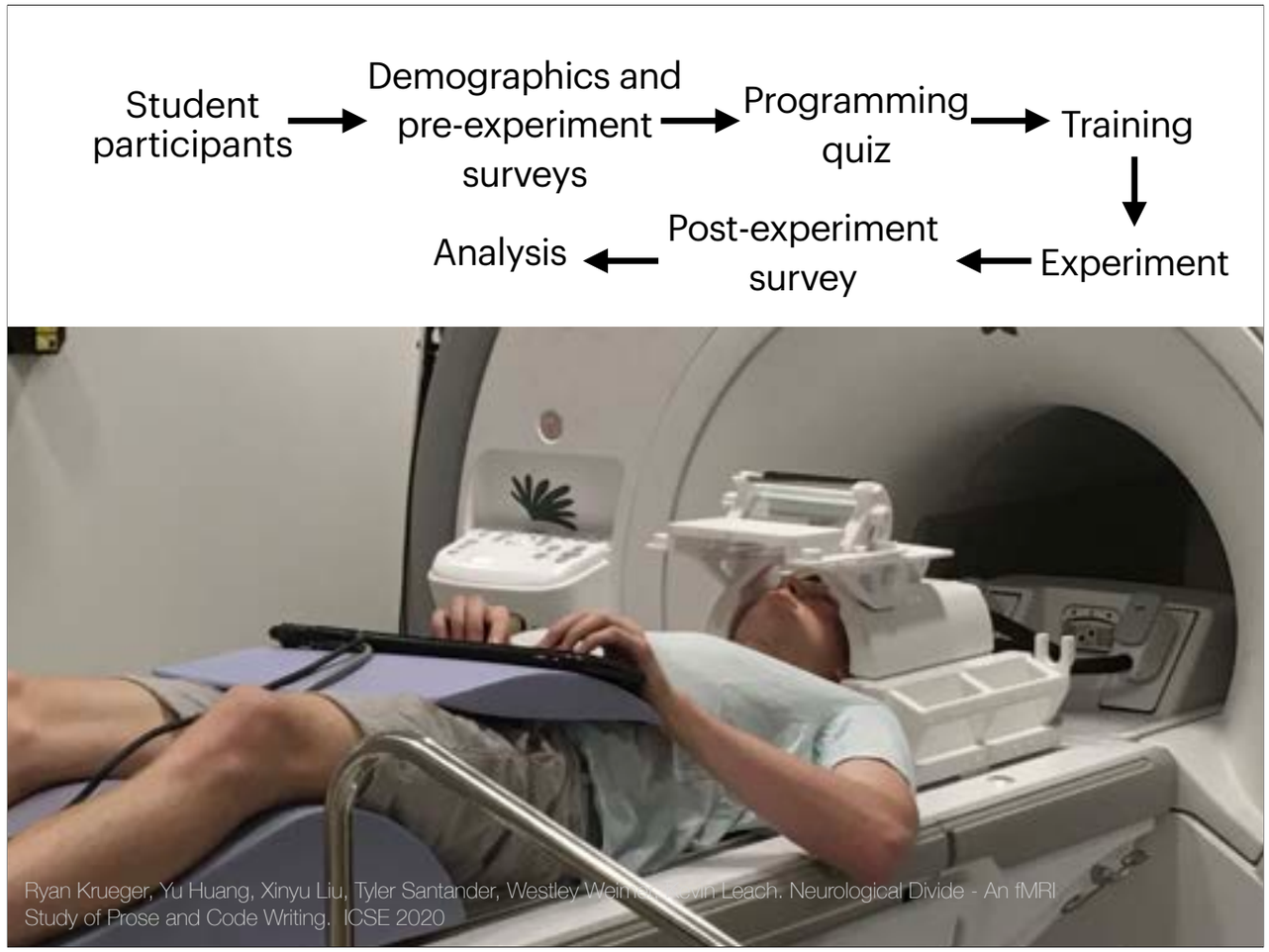
(A) FIELD STUDY **(B) FIELD EXPERIMENT**

There is no discussion of experimentation at all, so this is a field study

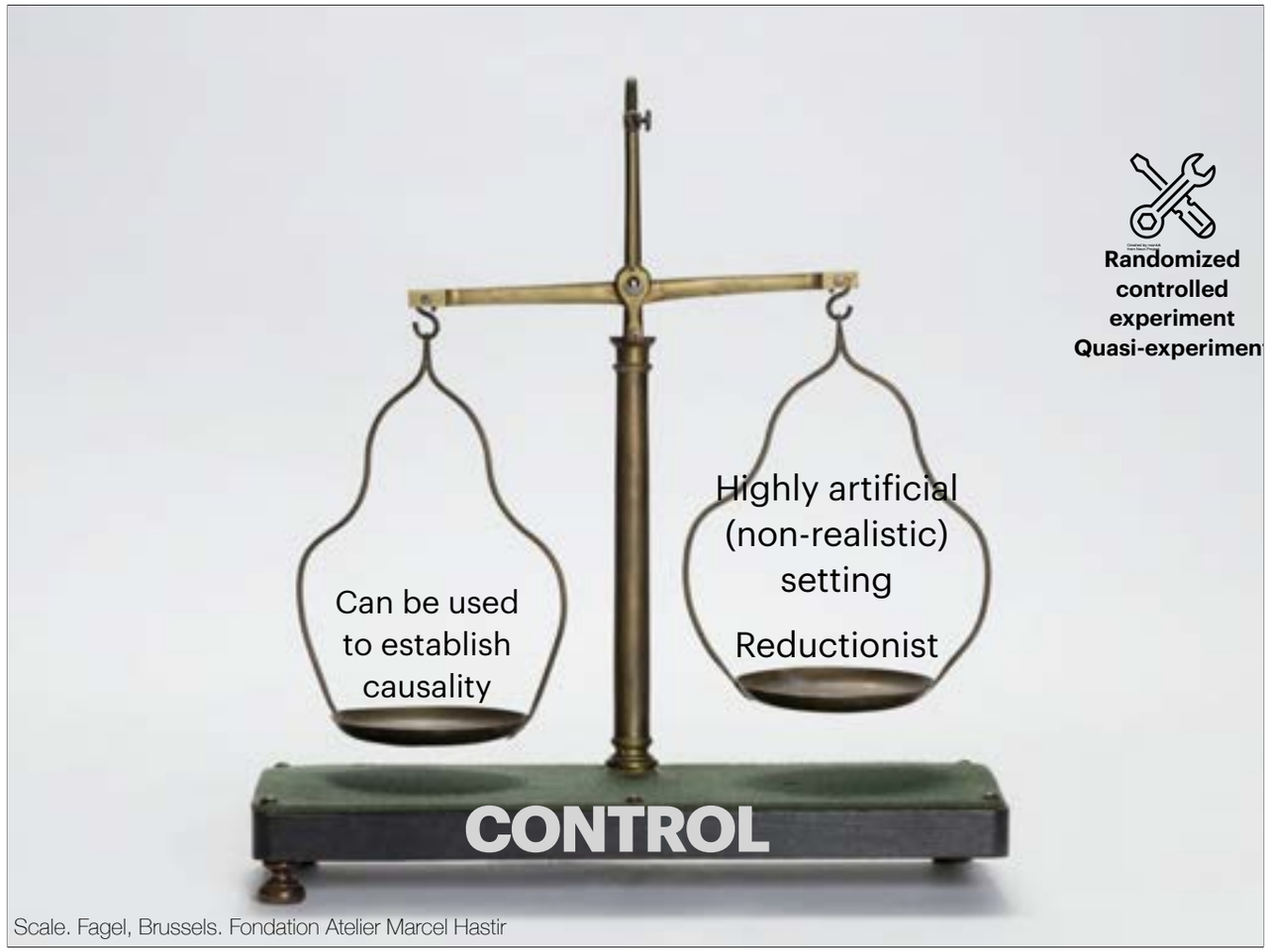
Klaas-Jan Stol, Brian Fitzgerald: The ABC of Software Engineering Research. TOSEM 27(3): 11:1-11:51 (2018)
https://o-ring.info/o-ring.info-images-new-design/eriks_cleanroom-belgium-1535031592.jpg

Cleanroom, test tube: Contrived setting (in vitro) created specifically for a study, with high degree of control of all measured variables. To study with a high degree of precision relationships between variables, or comparisons between techniques; may allow establishment of causality between variables.
Randomized controlled experiments and quasi experiments, comparative evaluations with benchmark studies; usually quantitative data exclusively.

- Abstract or unrealistic context due to highly artificial setting
- Typically scope of problem reduced to study the “essence,” optimizing internal validity at cost of external validity



Here the researchers wanted to understand whether writing source code involves the same brain areas as writing prose. The study was conducted using fMRI. Stress control: excluding left-handedness, ensuring comparable knowledge of C/C++ (quiz), randomzing the order of experimental tasks



Scale. Fagel, Brussels. Fondation Atelier Marcel Hastir

Field study is usually an exploratory strategy, to understand what's going on, how things work, or to generate hypotheses. These hypotheses should be then verified in a separate study

QUESTION

Jongeling et al. investigate to what extent results from software engineering studies using sentiment analysis depend on the choice of sentiment analysis tool. Through experimental comparison, finds that tools do not compare with manual labelling, nor do different tools agree with each other.

(A) FIELD STUDY

(B) FIELD EXPERIMENT

(C) LAB EXPERIMENT

This might appear odd but this is a lab experiment: participants are systems (sentiment analysis tools) rather than people!

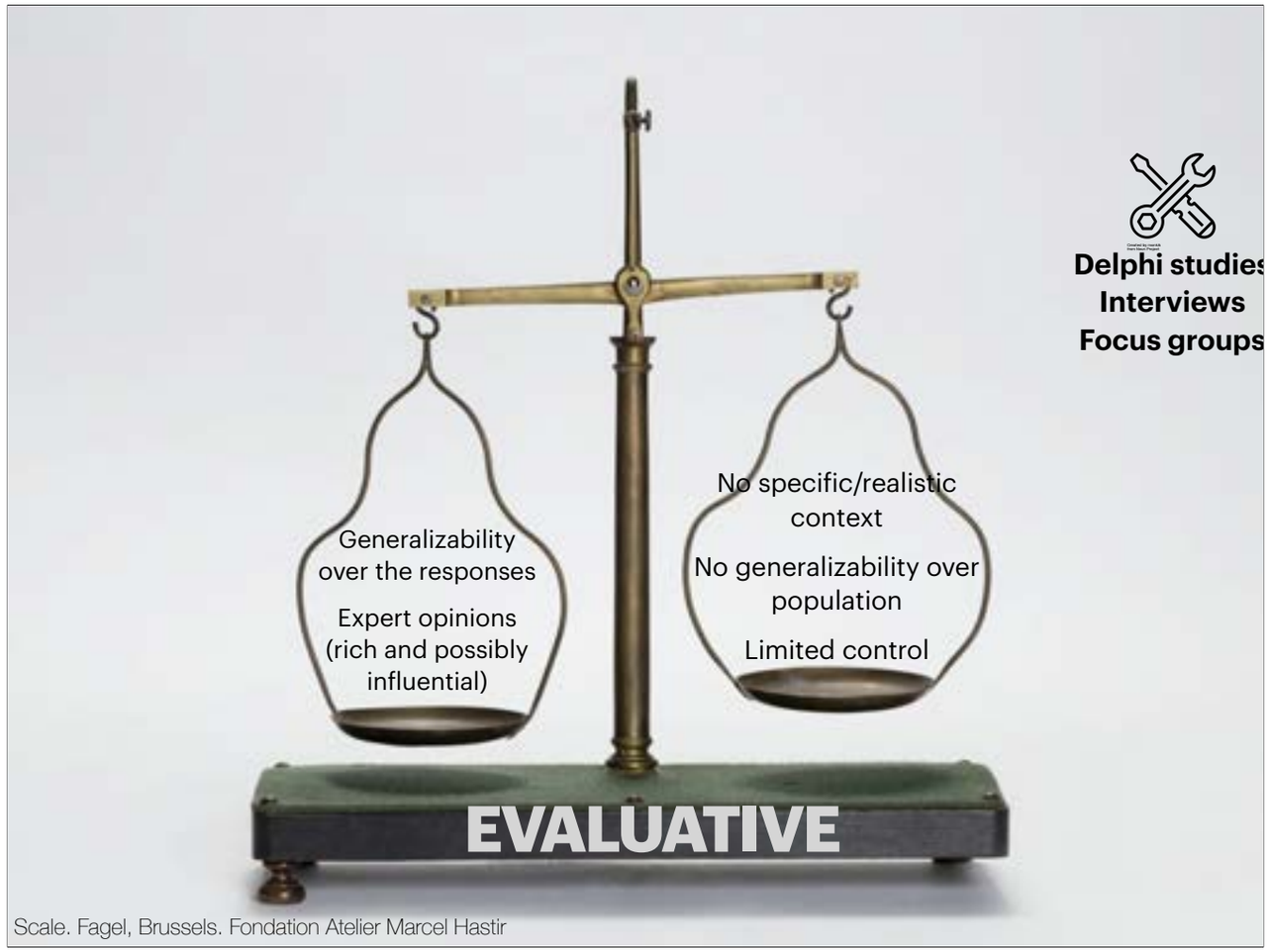
Courtroom: Neutral setting; may be actively designed to nullify the context, so that “responses” are in relation to some stimulus (question or instructions), independent of setting.

To elicit information from subjects for purposes of evaluation or study of an object. A judgment study involves gathering empirical data from a group of participants who are asked to judge or rate behaviors, to respond to a request or “stimulus” offered by a researcher, or to discuss a given topic of interest. Judgment studies rely on systematic sampling rather than representative sampling and should involve experts, appropriately informed to respond to a certain question or stimulus (**jury**). The goal of a judgment study is to seek generalizability over the responses, rather than generalizability to a population of actors.

Delphi studies, interview studies, focus group, evaluation studies; use of qualitative and/or quantitative data.

• Responses not related to any specific or realistic context
• Less generalizability than sample studies due to lack of representative sampling
• Less control and precision of measurement than a lab. exp.

Three rounds of expert interviews (so called Delphi method).



Scale. Fagel, Brussels. Fondation Atelier Marcel Hastir

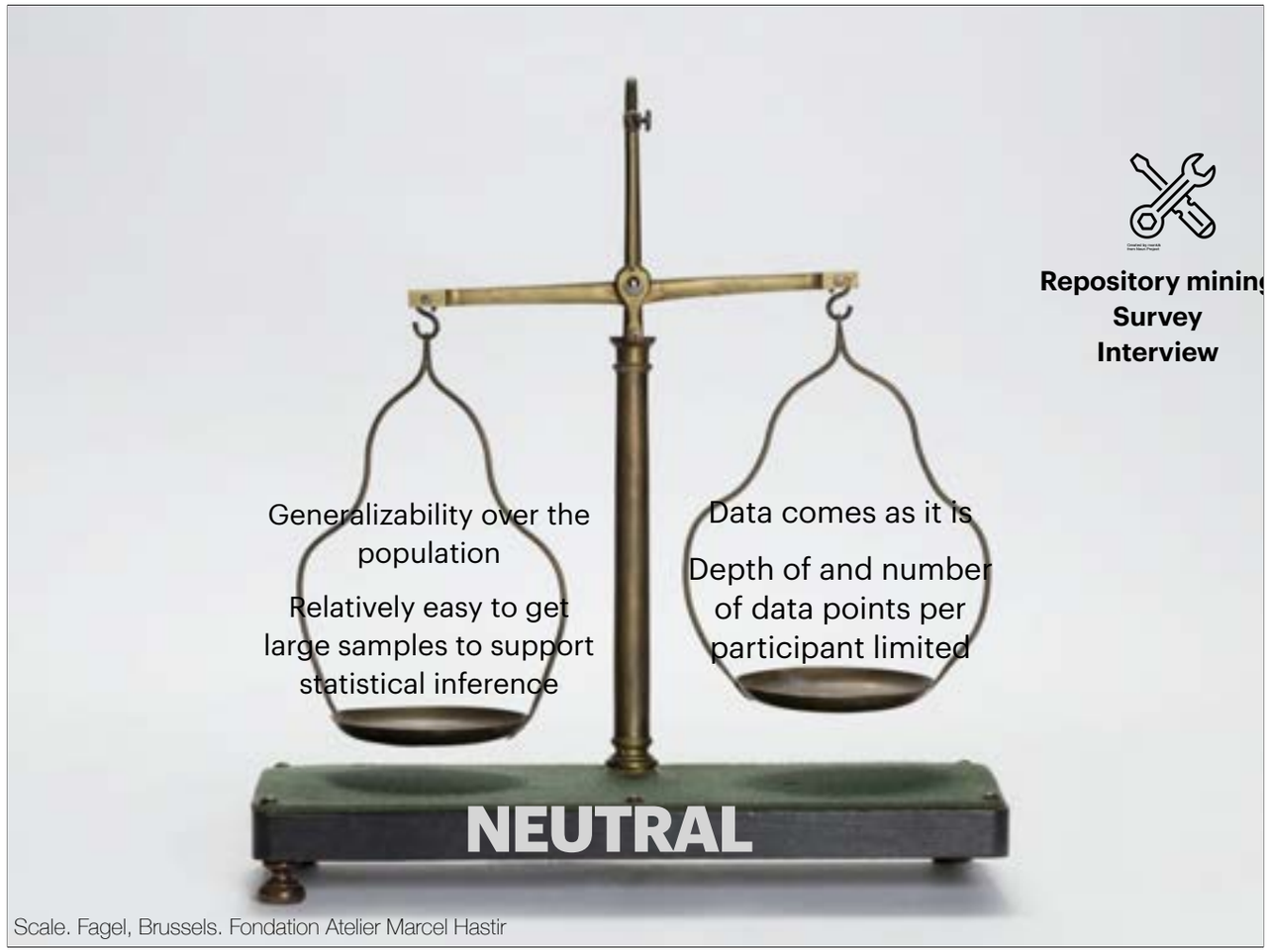
Field study is usually an exploratory strategy, to understand what's going on, how things work, or to generate hypotheses. These hypotheses should be then verified in a separate study

Referendum: Neutral setting. Limited level of precision of measurement; no variables are manipulated. The researcher must deal with whatever data is collected.

To study the distribution of a particular characteristic in a population (of people or systems), or the correlation between two or more characteristics in a population. Information is sought of the subjects.

Software repository mining, questionnaires, interviews; analysis includes correlational methods, e.g., regression. Typically, quantitative data (e.g., Likert scales) but can include qualitative data.

- Reductionist—depth of and number of data points per participant limited
- Data collection not “interactive”: no option to clarify questions; repository data comes as is, no opportunity to manipulate variables, only to correlate them



Scale. Fagel, Brussels. Fondation Atelier Marcel Hastir

Field study is usually an exploratory strategy, to understand what's going on, how things work, or to generate hypotheses. These hypotheses should be then verified in a separate study

QUESTION

Joblin et al. investigate the evolution of developer coordination in open-source projects, through an analysis of 18 large open-source software projects, with findings that aim to generalize.

(A) FIELD STUDY

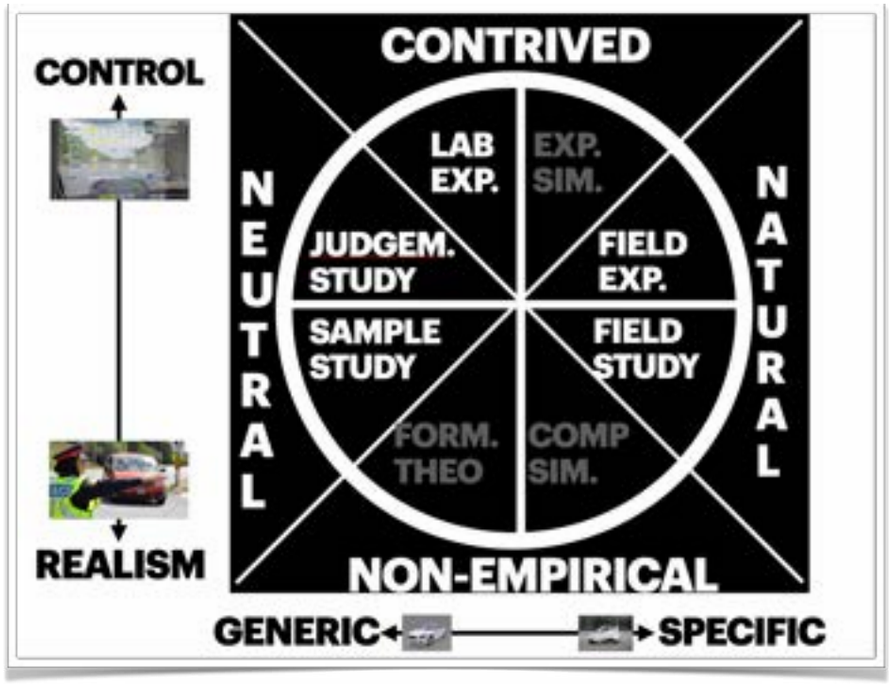
(B) SAMPLE STUDY

(C) JUDGEMENT STUDY

(D) LAB EXPERIMENT

This is a sample study, no discussion of specially selected experts

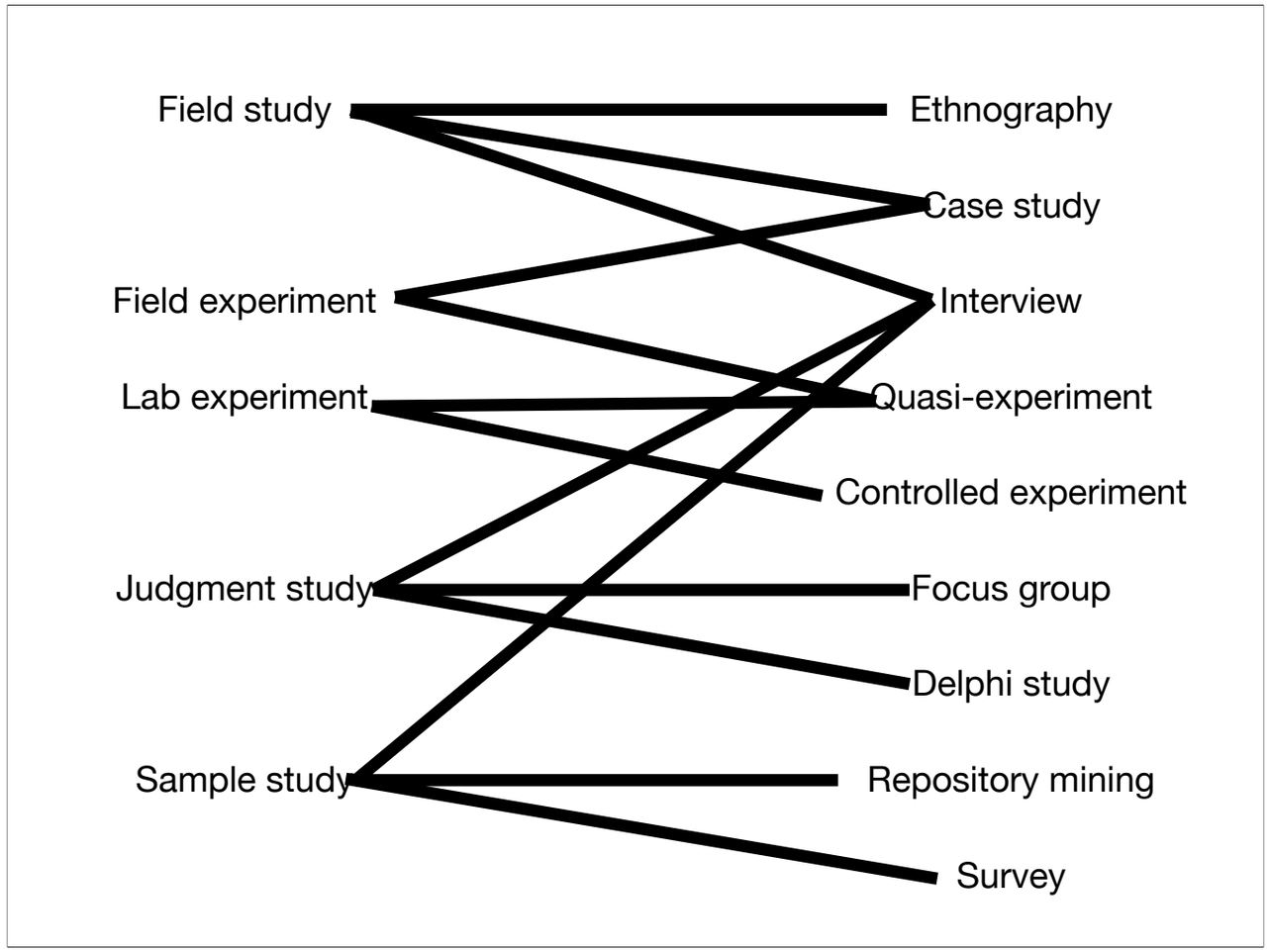
What **age- and gender-specific experiences** have veteran software developers of marginalized genders had in their careers?



Klaas-Jan Stol, Brian Fitzgerald: The ABC of Software Engineering Research. TOSEM 27(3): 11:1-11:51 (2018)

Sterre van Breukelen, Ann Barcomb, Sebastian Baltes, Alexander Serebrenik: "STILL AROUND": Experiences and Survival Strategies of Veteran Women Software Developers. ICSE 2023: 1148-1160

Question to you: which research strategy would you choose to conduct such a study and why? Reminder: we have chosen the neutral quadrant but you can choose something else.

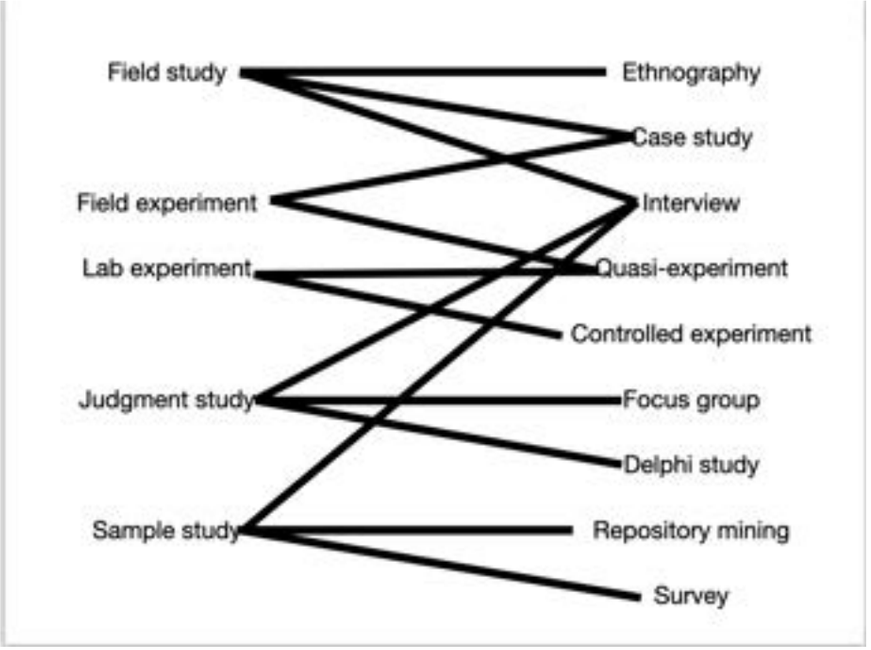


Research strategies on the left, research methods on the right

A **quasi-experiment** is an empirical interventional study used to estimate the causal impact of an intervention on target population without random assignment. Quasi-experimental research shares similarities with the traditional experimental design or randomized controlled trial, but it specifically lacks the element of random assignment to treatment or control.

The links merely indicate *typical* methods. Software development in case of simulation means development of simulation environments. Of course, repository mining and controlled experiments often require building software tools: repository mining needs analysis tools and controlled experiments need experimental environments. In the following lectures we will consider many of these research methods and will discuss in context of what research strategies they are useful.

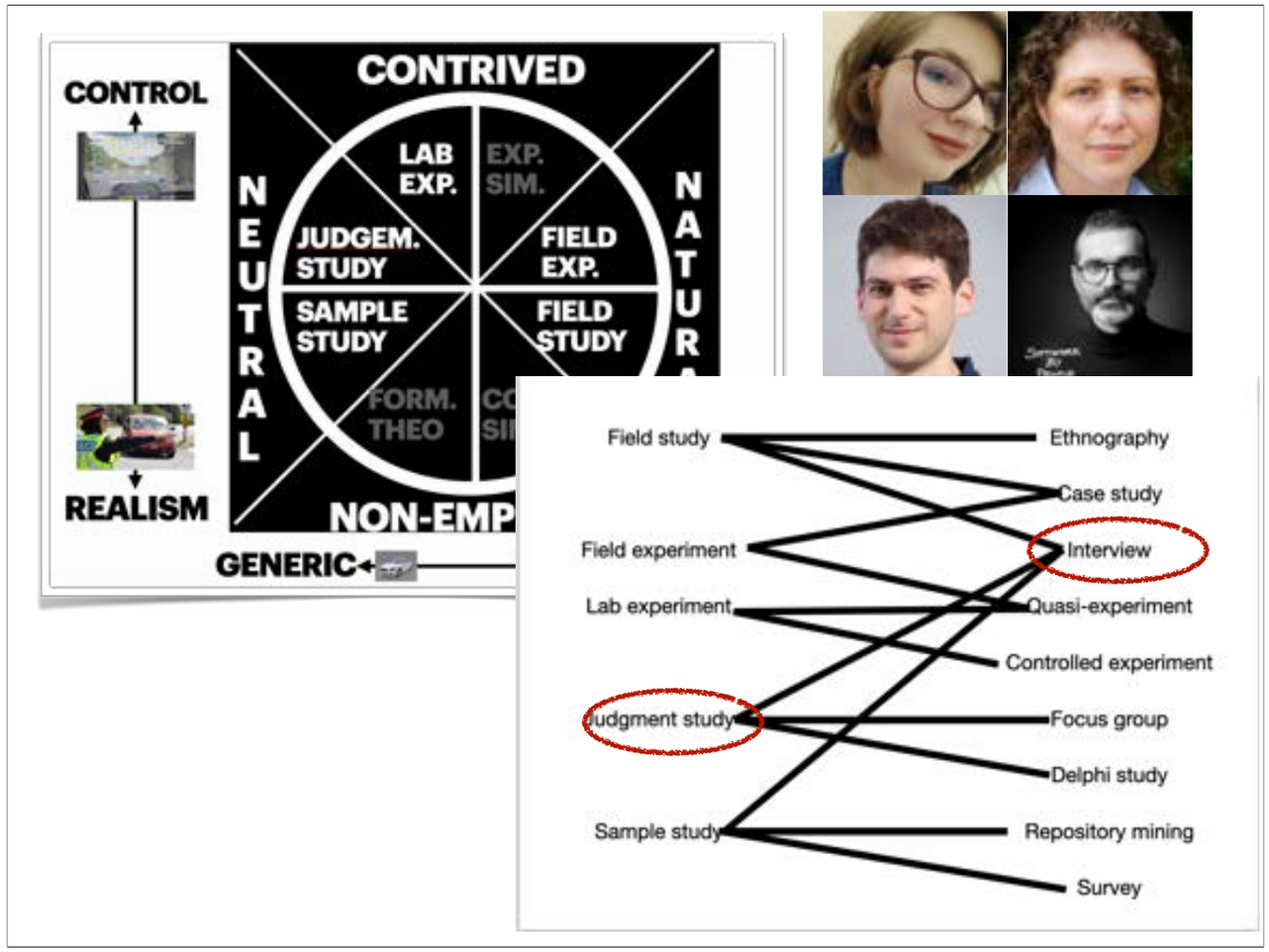
What **age- and gender-specific experiences** have veteran software developers of marginalized genders had in their careers?



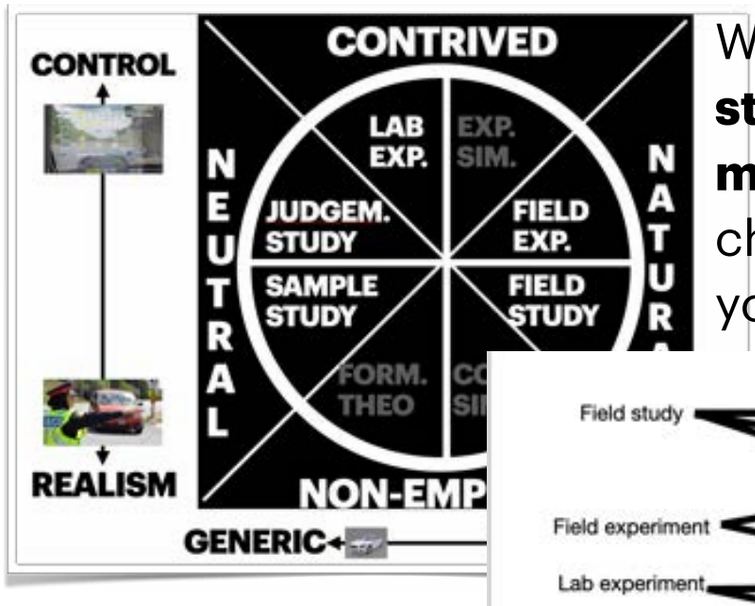
Klaas-Jan Stol, Brian Fitzgerald: The ABC of Software Engineering Research. TOSEM 27(3): 11:1-11:51 (2018)

Sterre van Breukelen, Ann Barcomb, Sebastian Baltes, Alexander Serebrenik: "STILL AROUND": Experiences and Survival Strategies of Veteran Women Software Developers. ICSE 2023: 1148-1160

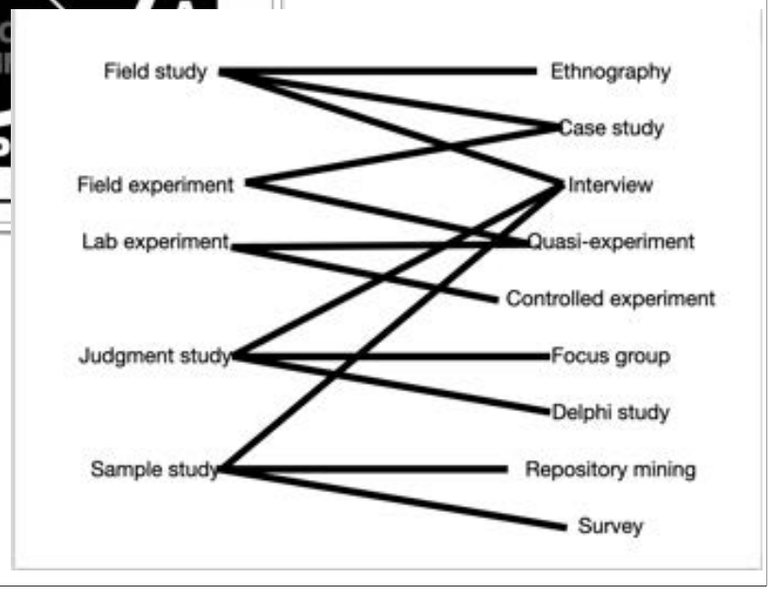
Question to you: which research methods would you choose to conduct a study of gender- and age-related experiences and why?



This is what we have chosen



What **research strategy and method(s)** were chosen in the paper you brought today?



=

Since we are interested in obtaining profound insights in the individuals' experience, we have decided to conduct interviews. From this point on we discuss the decisions in small.

What age- and gender-specific experiences have **veteran software developers of marginalized genders** had in their careers?

Sterre van Breukelen, Ann Barcomb, Sebastian Baltes, Alexander Serebrenik: "STILL AROUND": Experiences and Survival Strategies of Veteran Women Software Developers. ICSE 2023: 1148-1160

Answering this RQ calls for understanding of two difficult constructs, namely “being old” and “being of minoritized gender”.

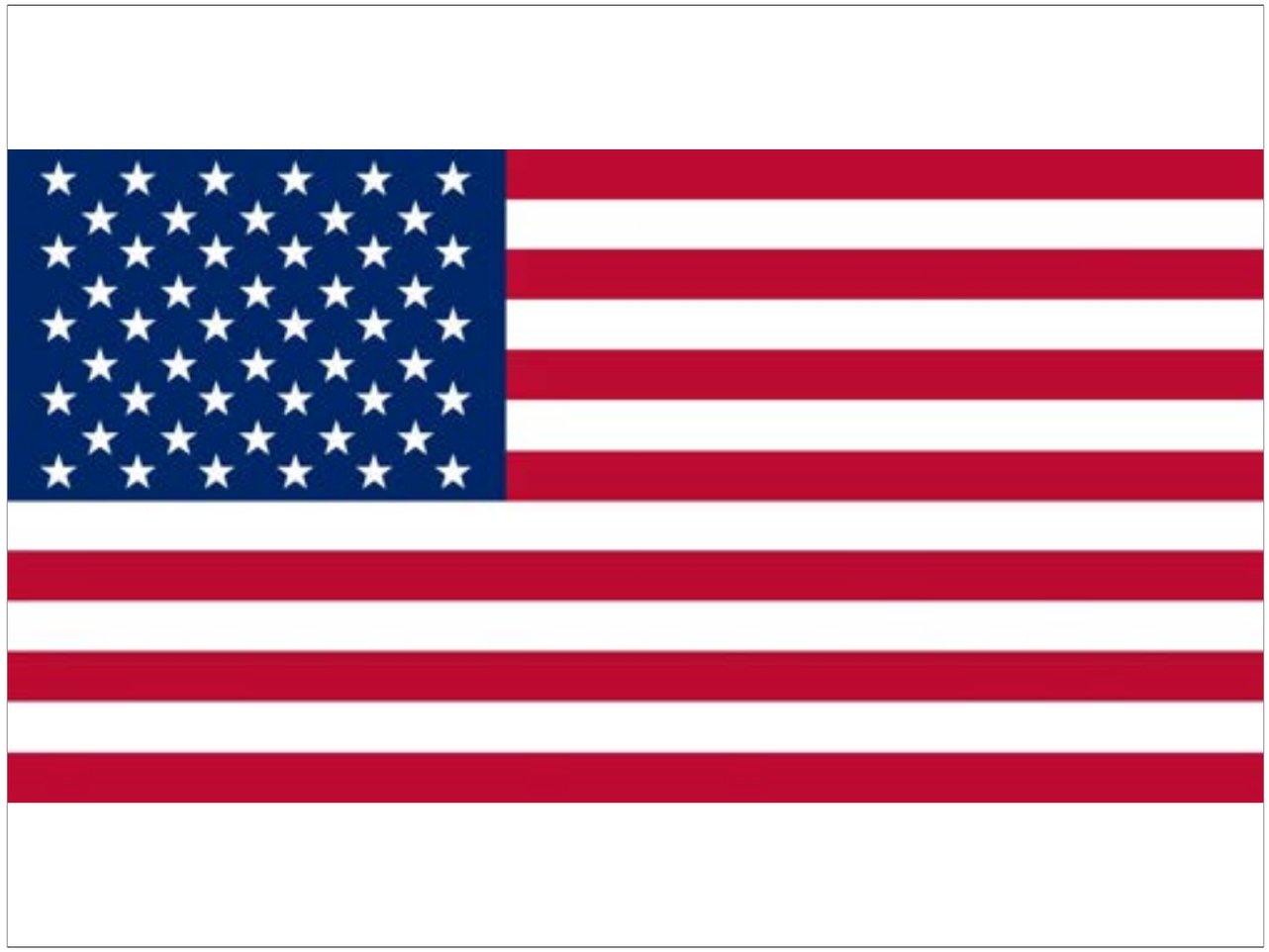
HOW OLD IS "OLD"?

30 or older	4
35 or older	3
35-40 or older	2
40 or older	**7**
45 or older	2
50 or older	1
Not stated explicitly	5

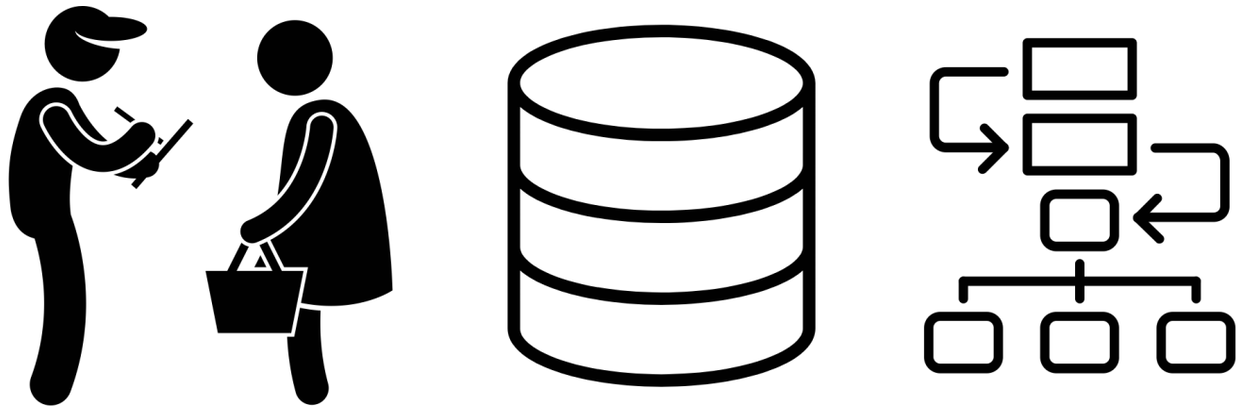
Sebastian Baltes, George Park, Alexander Serebrenik. Is 40 the new 60? How popular media portrays the employability of older software developers, IEEE Software, 37(6):26-31, 2020

The problem here is that not much research has been conducted on the perception of age among software developers, so we used our own previous work.

- **Previous research has shown that 40 is the threshold when developers are starting to be seen as old**, so we have reused the same threshold in the current study.
- We further **operationalized 'older' as having at least 18 years of experience** in the software industry, based on a typical career including school and university.

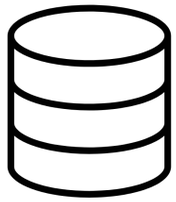


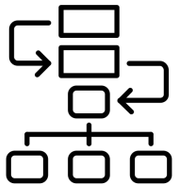
However, this study has been conducted on the US public discourse, which means that the subsequent study should be US-based as well.



Alexander Serebrenik. How to Ask about Gender Identity of Software Engineers and “Guess” it from the Archival Data, Equity, Diversity, and Inclusion in Software Engineering: Best Practices and Insights (Damian et al. eds.) 2024.

When it comes to gender, there are roughly three ways: ask people about their gender, select the dataset that explicitly records gender-related information or use an algorithmic tool to guess it based on such information as names or profile images.





Best answers possible

Good but possibly outdated answers

As many data points as needed

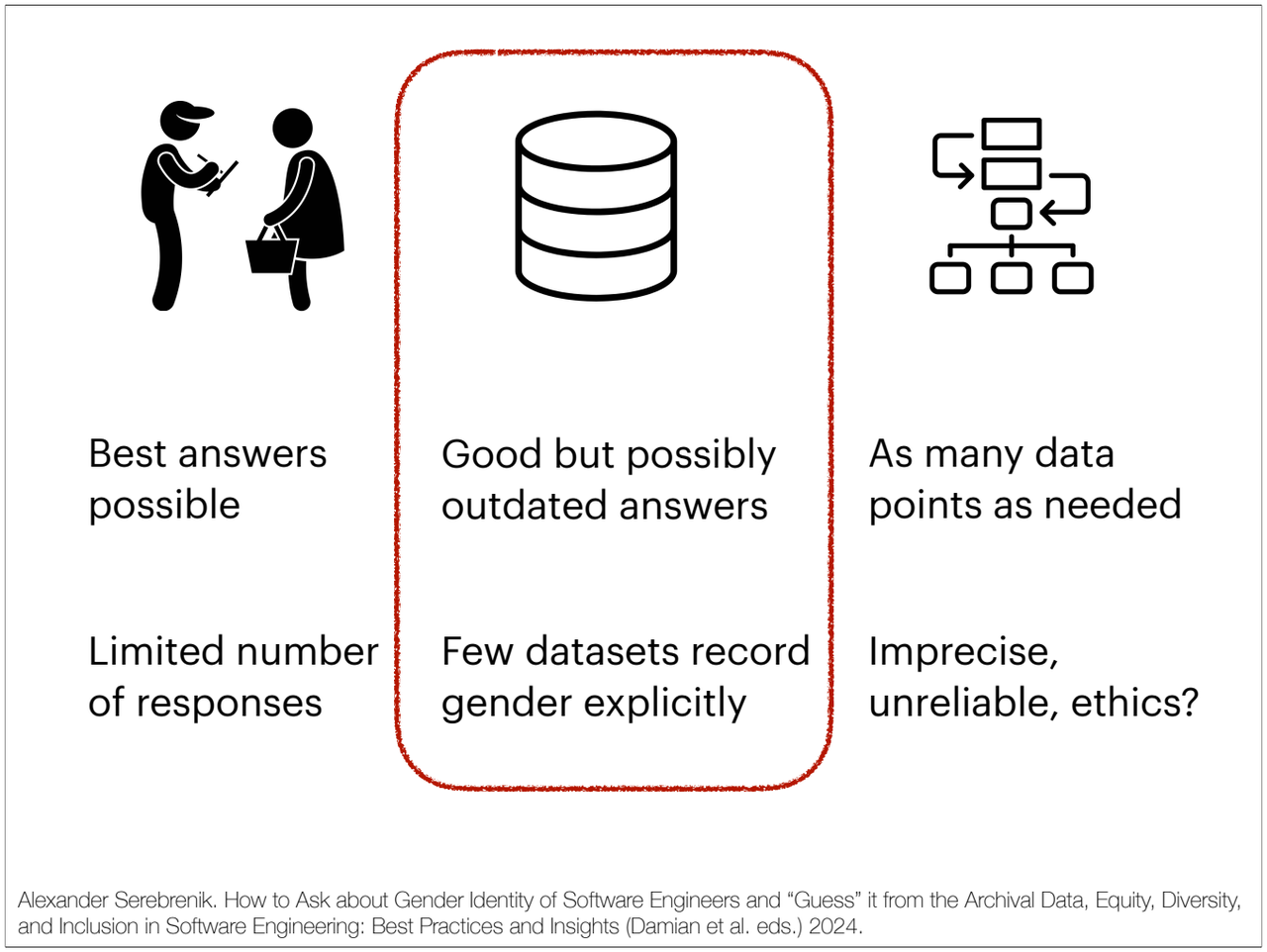
Limited number of responses

Few datasets record gender explicitly

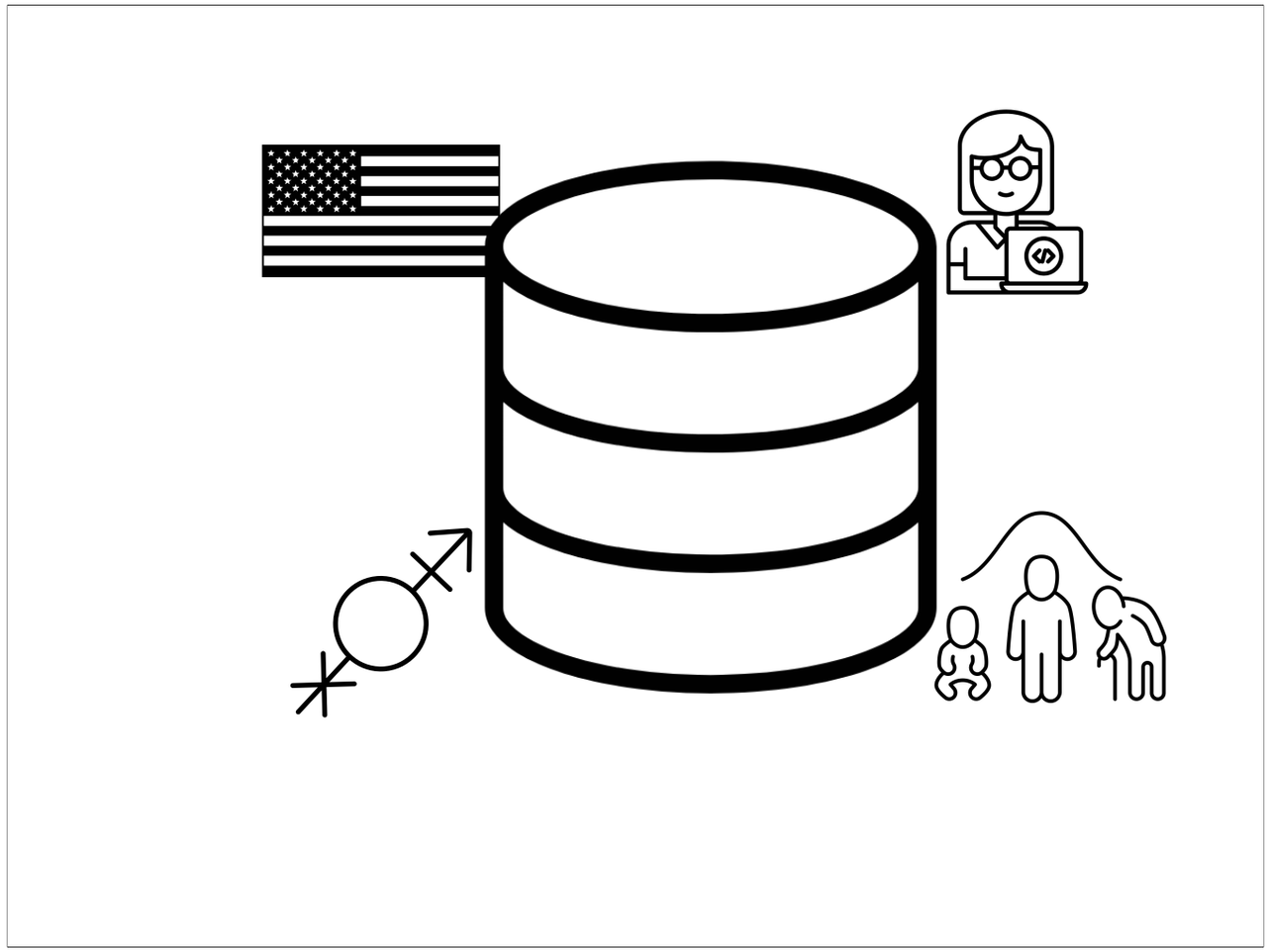
Imprecise, unreliable, ethics?

Alexander Serebrenik. How to Ask about Gender Identity of Software Engineers and “Guess” it from the Archival Data, Equity, Diversity, and Inclusion in Software Engineering: Best Practices and Insights (Damian et al. eds.) 2024.

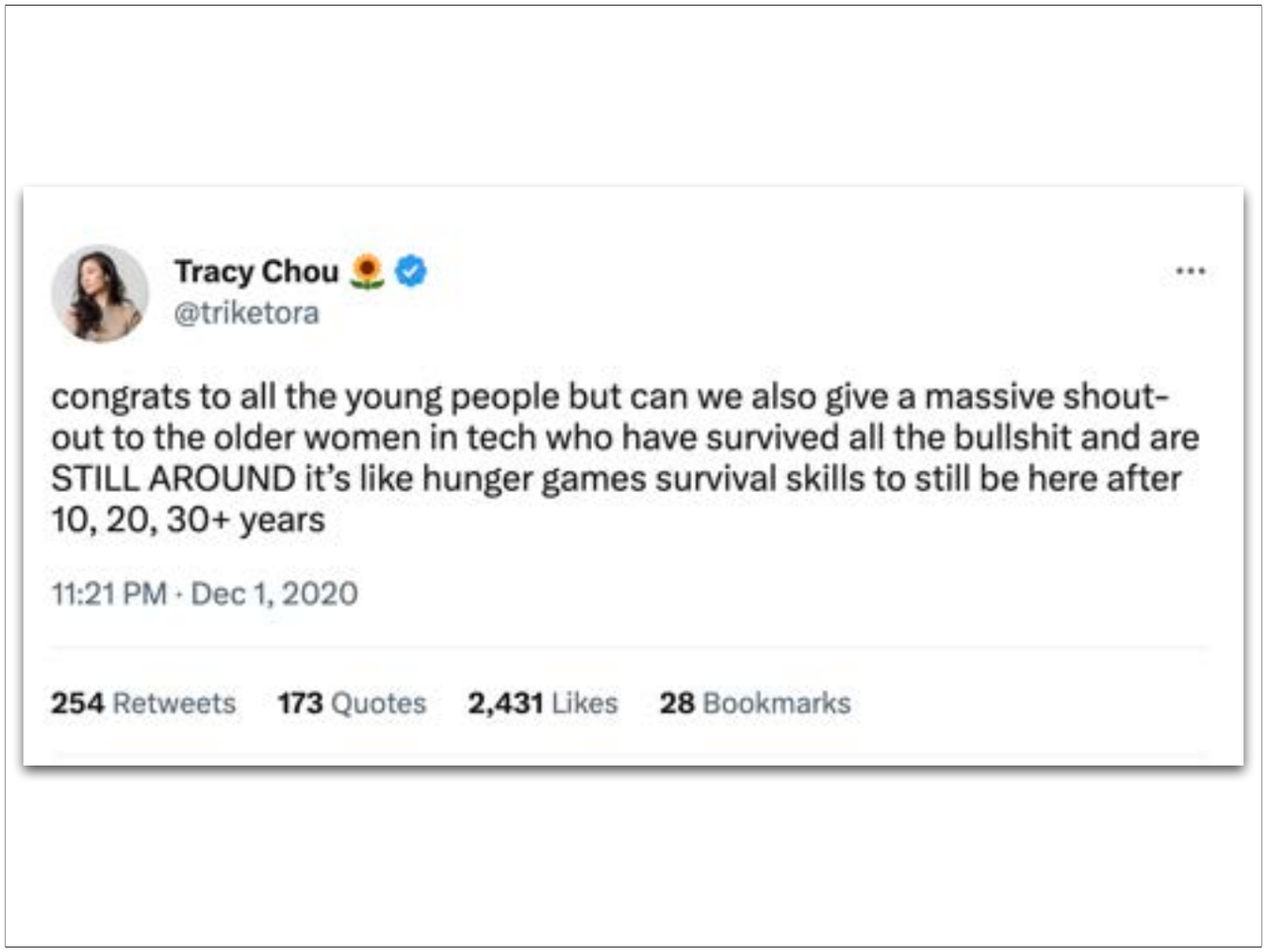
Advantages and disadvantages of each one of the solutions



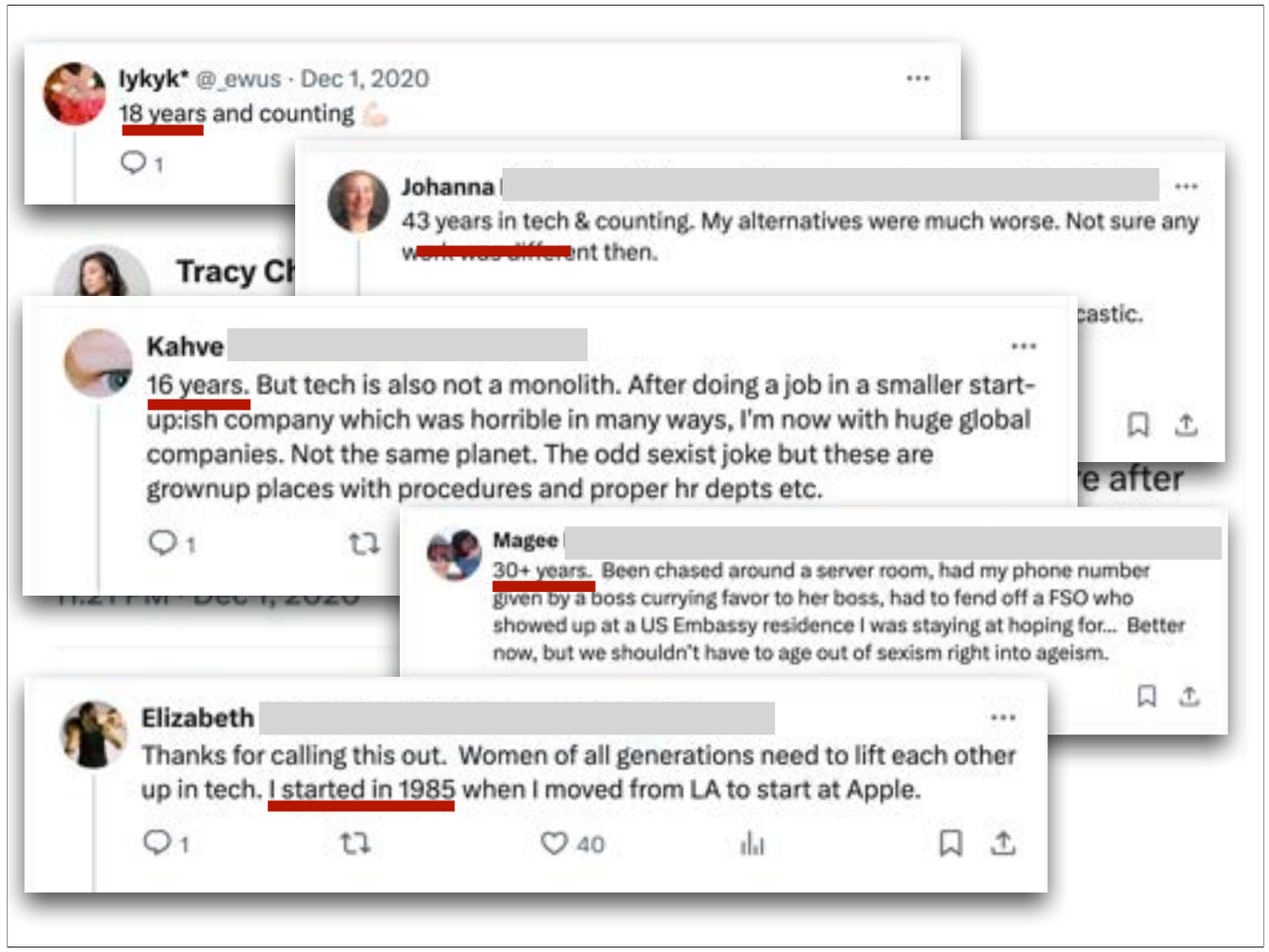
In our work we have focused on the second way but we will see that we have made use of the two other techniques as well.



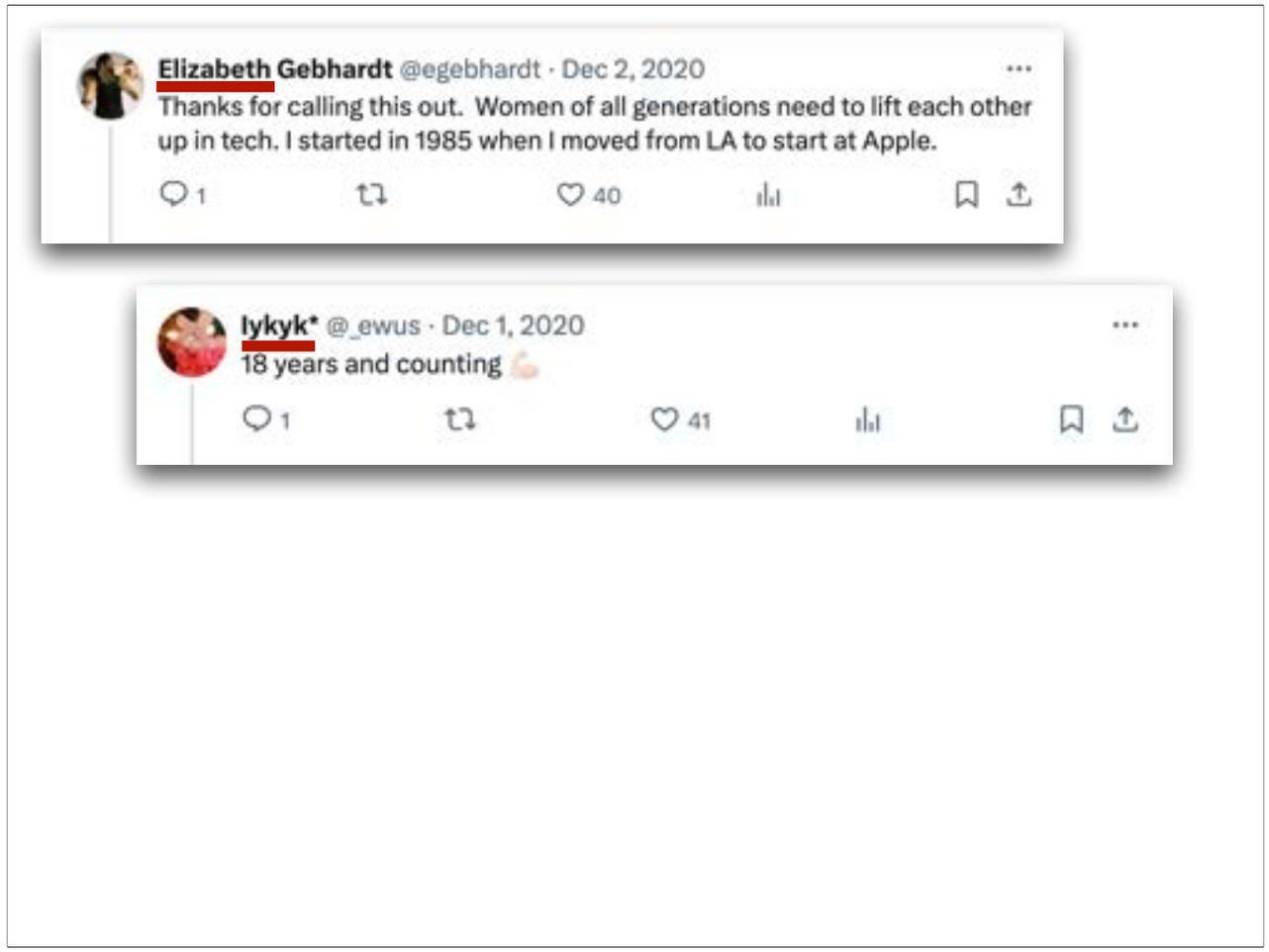
So we need a dataset with the US-based data about software developers that explicitly records gender and age.



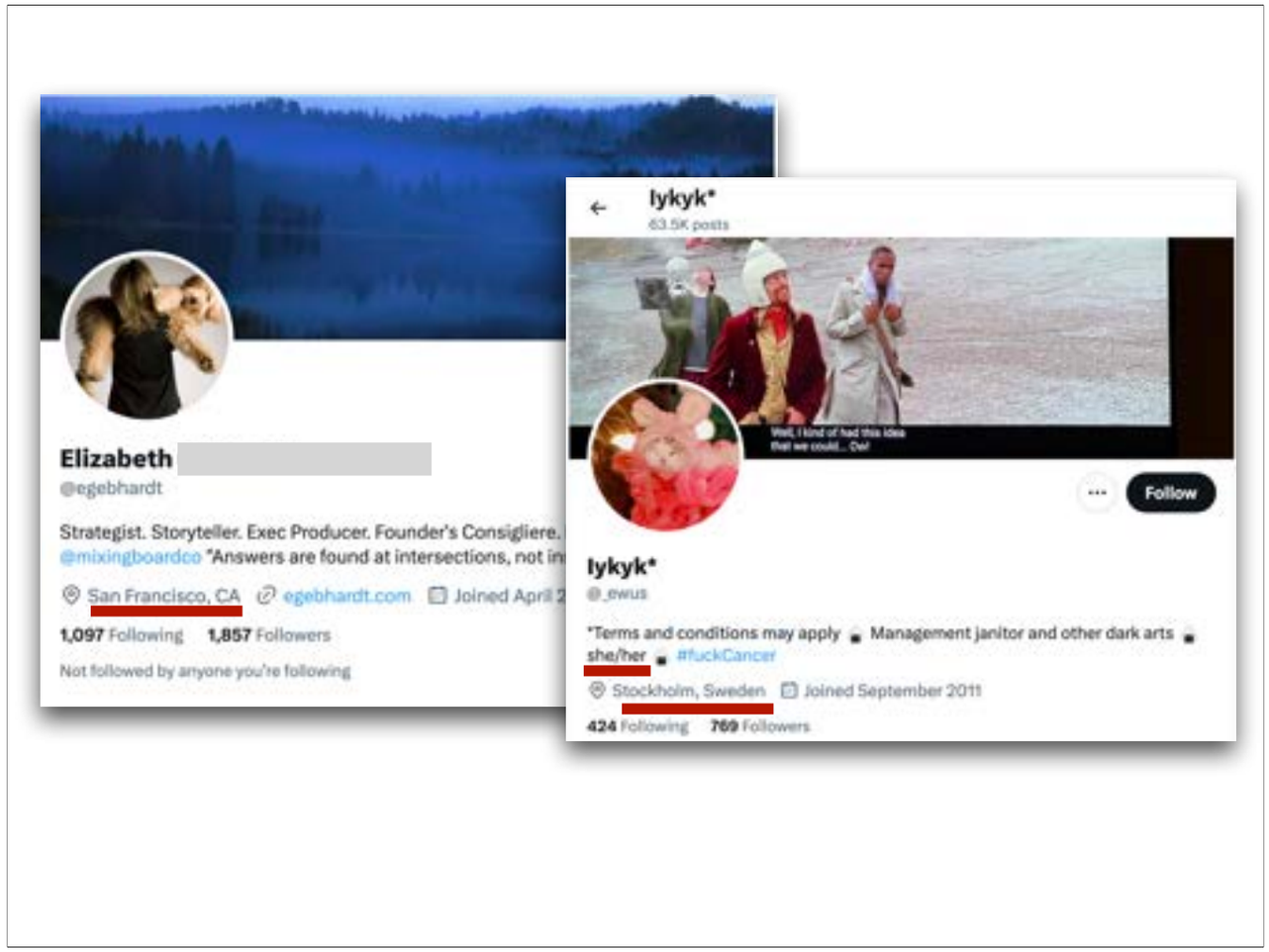
So we have found a tweet by **Tracy Chou.** She is one of **Time's 12 Women of the Year** (2022), a **software engineer** and **advocate for diversity in technology-related fields**. Her **tweet alludes to the negative experiences and rarity of older women** who are still active in the field of software development.



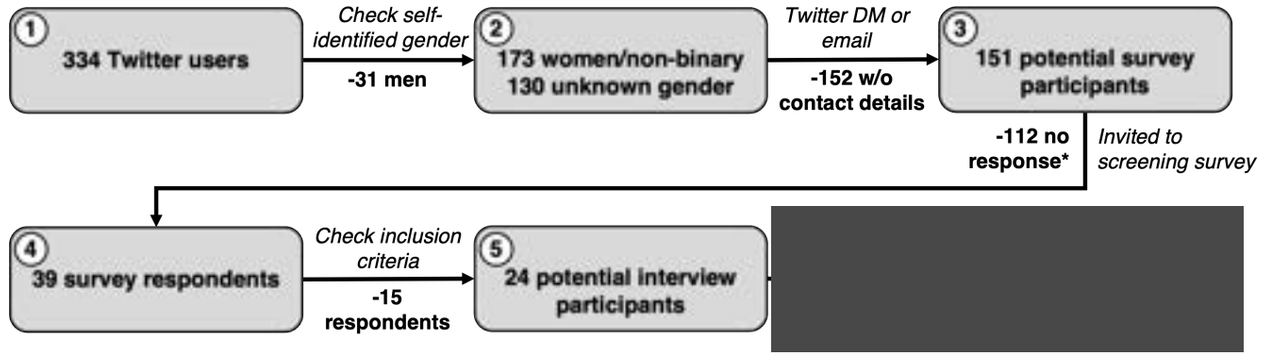
This tweet has started an entire thread and we have found several similar threads. We see that individuals responding to the message by Tracey Chou voluntarily divulge their number of years of experience in tech and subsequently their age.



So what about their gender? As I mentioned before, one of the options would be to use a guessing tool that would try to infer gender based on the first name. While this is oftentimes the only option, this solution is suboptimal for many reasons including that we can only infer how others perceive the person, not how they perceive themselves. Moreover, this would not work for the second example.

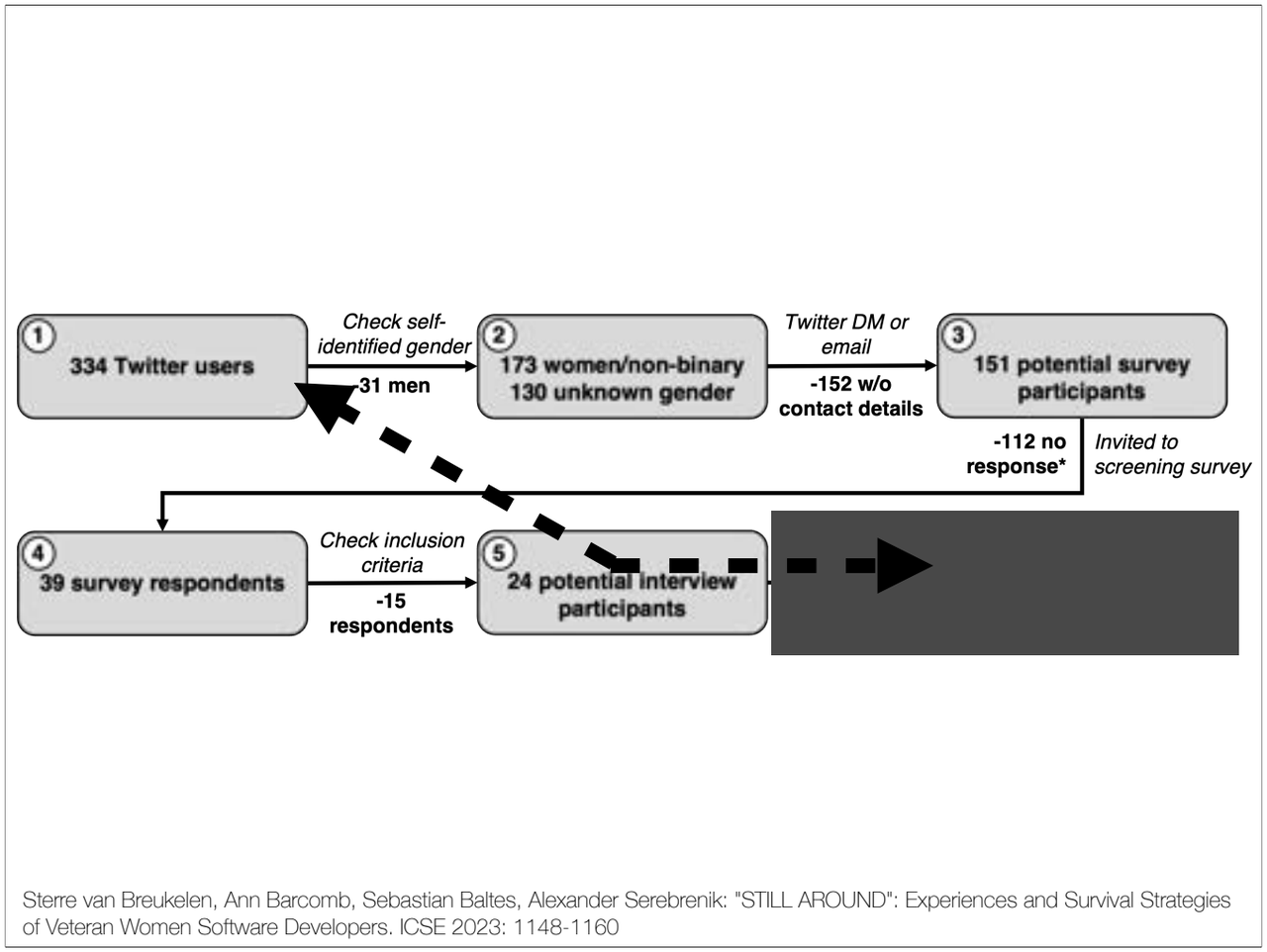


This is why we have focused on individuals that have either indicated pronouns typically associated with women and non-binary people (she/they), used gender-specific terms (mother, wife) or at least do not include pronouns and terms that are typically associated with men (he, husband). Of course, this is a rather coarse grained way of identifying people and this still needs to be verified.



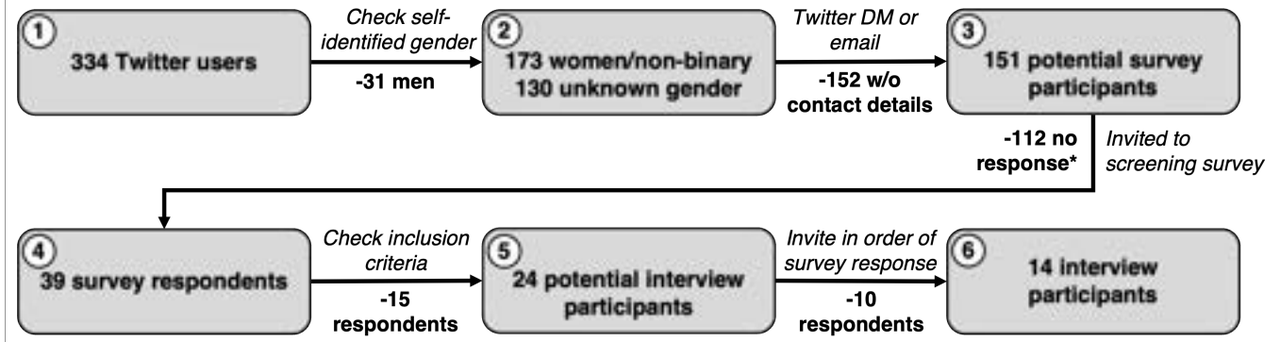
Sterre van Breukelen, Ann Barcomb, Sebastian Baltes, Alexander Serebrenik: "STILL AROUND": Experiences and Survival Strategies of Veteran Women Software Developers. ICSE 2023: 1148-1160

This is why to ensure correctness of our dataset, we used a prescreening survey to verify that the respondents actually belong to our demographics.



Sterre van Breukelen, Ann Barcomb, Sebastian Baltes, Alexander Serebrenik: "STILL AROUND": Experiences and Survival Strategies of Veteran Women Software Developers. ICSE 2023: 1148-1160

Now we have two options: either to proceed with the interviews or go back to our data collection phase and consider additional threads.



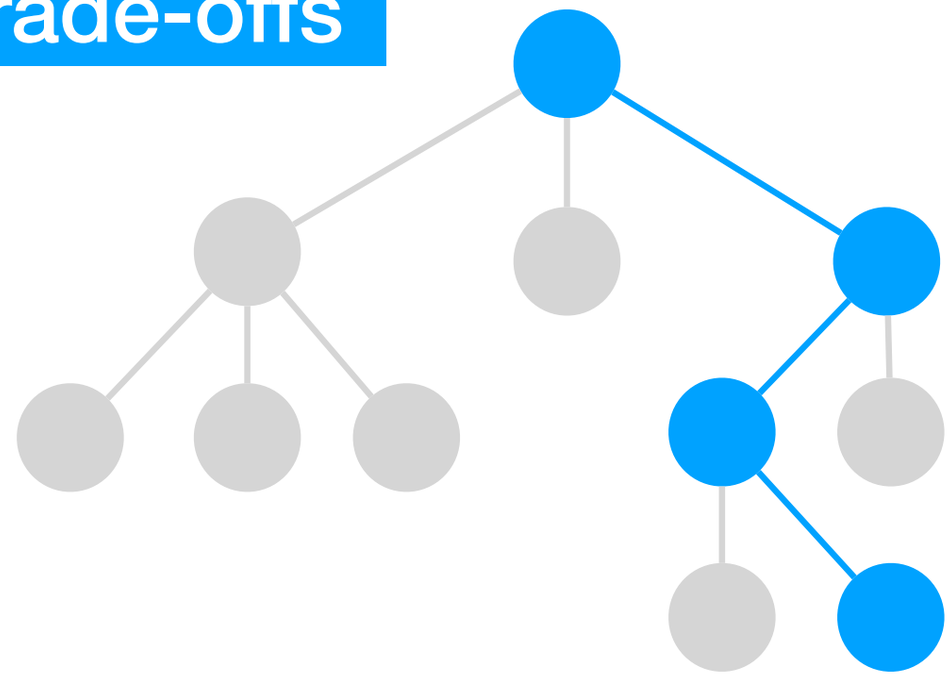
Sterre van Breukelen, Ann Barcomb, Sebastian Baltes, Alexander Serebrenik: "STILL AROUND": Experiences and Survival Strategies of Veteran Women Software Developers. ICSE 2023: 1148-1160

We have decided to continue with the group we had and interviewed the participants. We have stopped after 14 interviews since we have reached theoretical saturation. The notion of theoretical saturation comes from the grounded theory

Greta Hoffman from Pexels

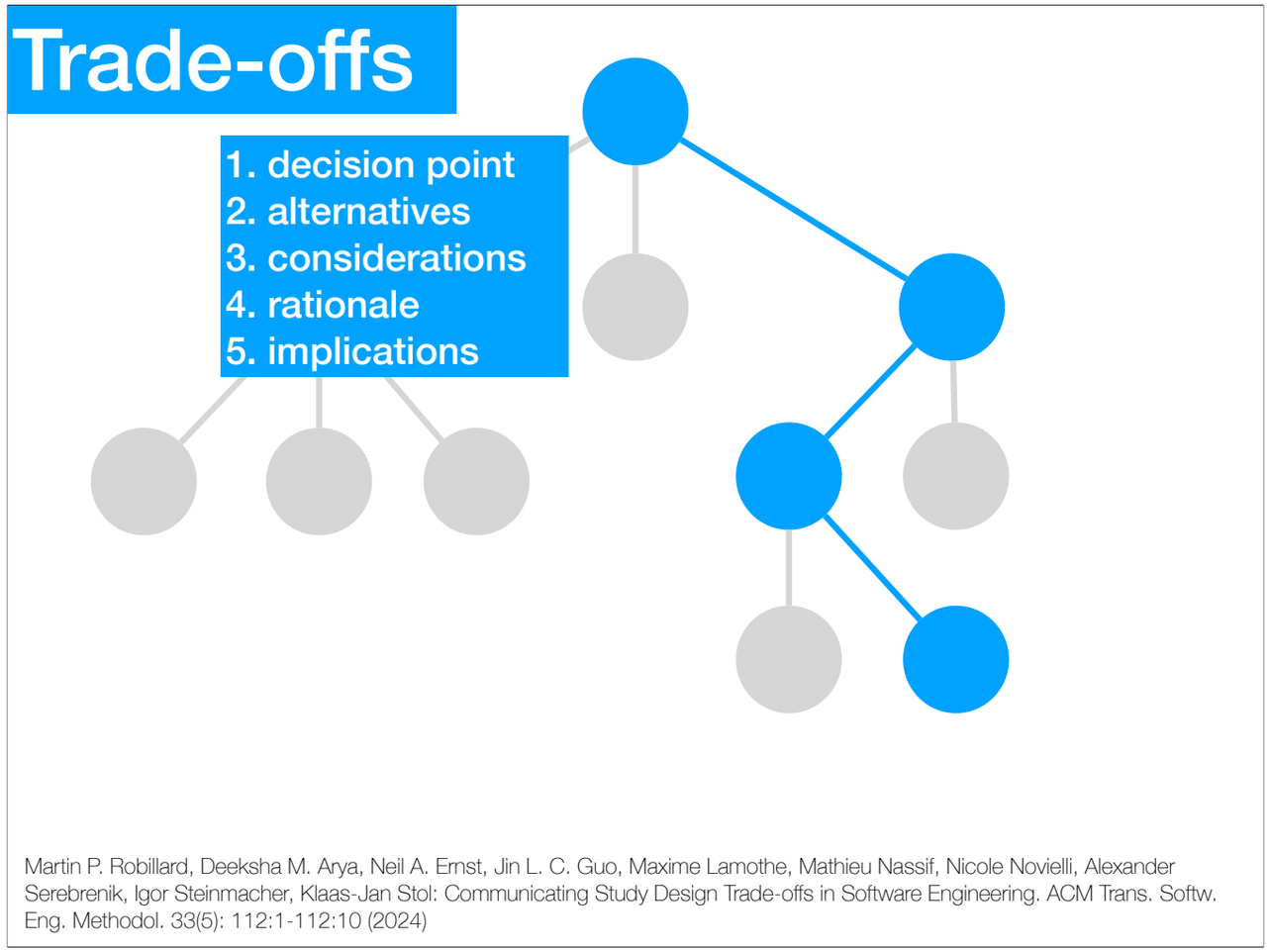
- To **evaluate the stability of our findings**, we also included three participants that, while not strictly belonging to the target demographics, might share experiences and strategies with other interviewees.
- One participant slightly **younger**, one who **left the industry**, and one identified as **non-binary**.
- The answers of those participants **were in line with the other feedback** we got from our interviews.
- In total, we have conducted **14 interviews** until reaching **saturation**.

Trade-offs



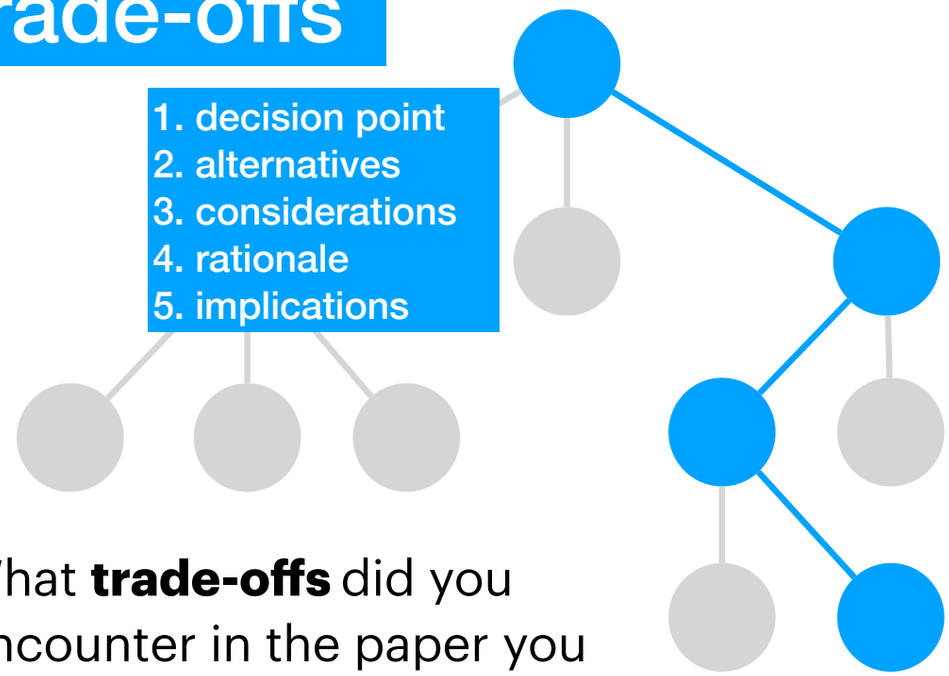
Martin P. Robillard, Deeksha M. Arya, Neil A. Ernst, Jin L. C. Guo, Maxime Lamothe, Mathieu Nassif, Nicole Novielli, Alexander Serebrenik, Igor Steinmacher, Klaas-Jan Stol: Communicating Study Design Trade-offs in Software Engineering. ACM Trans. Softw. Eng. Methodol. 33(5): 112:1-112:10 (2024)

Essentially whether we are working on a large scale or on a small scale, what we have is a series of decisions and trade-offs: what research method do we use (interviews or surveys), how do we define “being old”, how do we guess gender?



1. A trade-off is identified by its decision point, which can act as its identifier.
2. Alternatives, relative importance of these alternatives. A review of alternatives can include properties, such as whether or not the set of alternatives is closed or whether or not they are mutually exclusive.
3. The selection of one alternative over competing options is the outcome of a system of considerations that relates the costs and benefits of each alternative, as well as constraints limiting the design space.
4. The rationale for selecting an alternative can then be expressed in terms of these considerations.
5. Additional discussion of the implications of the decision supports an in-depth exploration of the consequences of the choice made, in contrast to the inevitably more general cost-benefit calculus involved in the previous point (i.e., considerations).

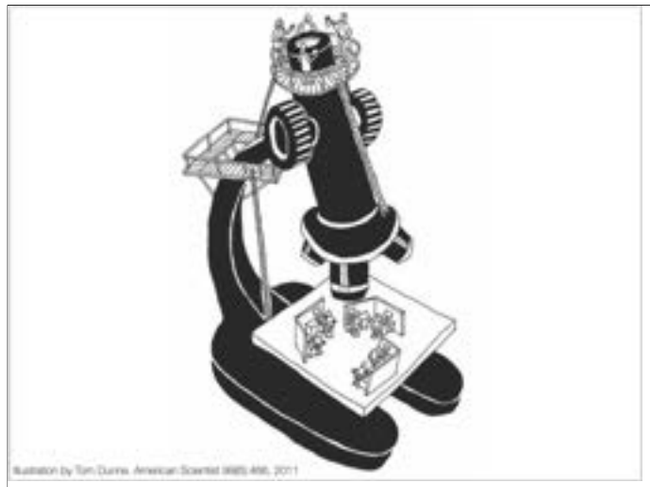
Trade-offs



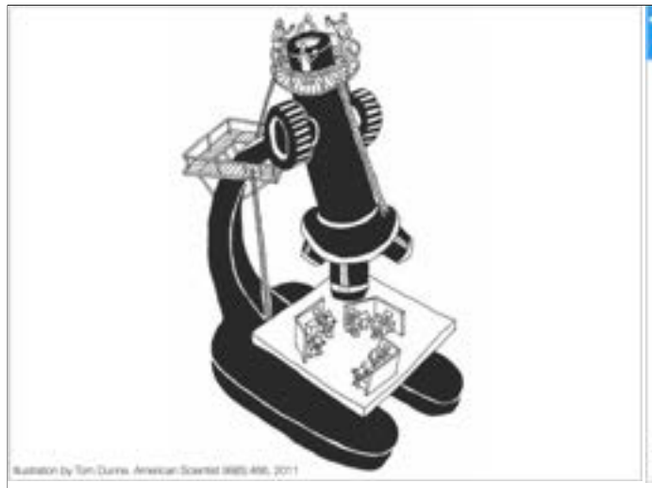
What **trade-offs** did you encounter in the paper you brought today?

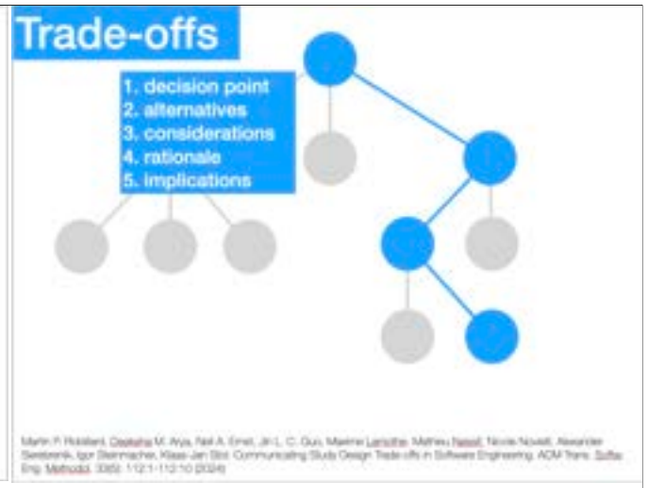
Martin P. Robillard, Deeksha M. Arya, Neil A. Ernst, Jin L. C. Guo, Maxime Lamothe, Mathieu Nassif, Nicole Novielli, Alexander Serebrenik, Igor Steinmacher, Klaas-Jan Stol: Communicating Study Design Trade-offs in Software Engineering. ACM Trans. Softw. Eng. Methodol. 33(5): 112:1-112:10 (2024)

1. A trade-off is identified by its decision point, which can act as its identifier.
2. Alternatives, relative importance of these alternatives. A review of alternatives can include properties, such as whether or not the set of alternatives is closed or whether or not they are mutually exclusive.
3. The selection of one alternative over competing options is the outcome of a system of considerations that relates the costs and benefits of each alternative, as well as constraints limiting the design space.
4. The rationale for selecting an alternative can then be expressed in terms of these considerations.
5. Additional discussion of the implications of the decision supports an in-depth exploration of the consequences of the choice made, in contrast to the inevitably more general cost-benefit calculus involved in the previous point (i.e., considerations).

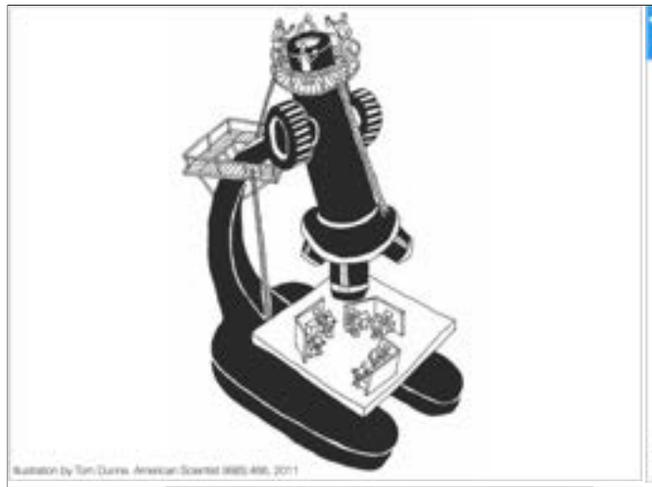


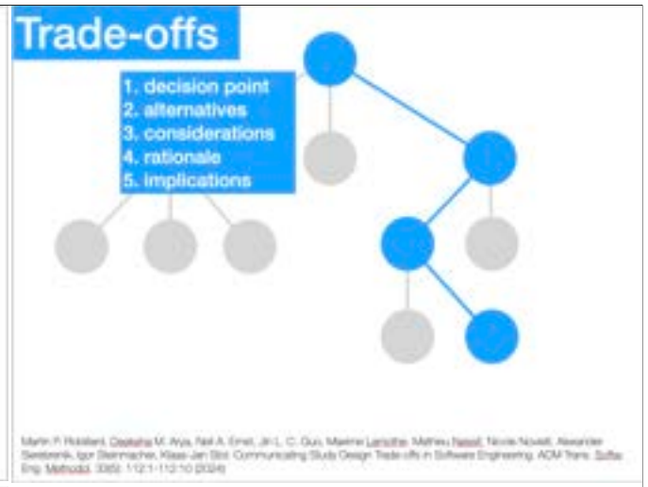
Let us summarize. In empirical studies **observation is the only source of knowledge, while reasoning or intuition can be source of belief or conjecture**. In software engineering researchers describe **how software is being developed** to suggest best practices how to develop better software (product) and how to develop software better (process).

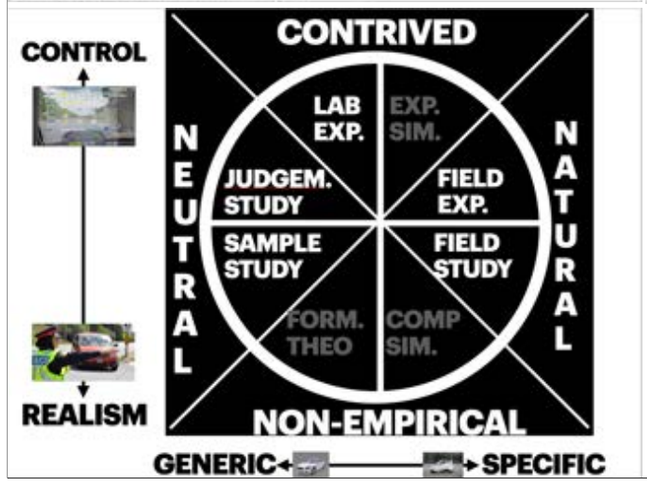




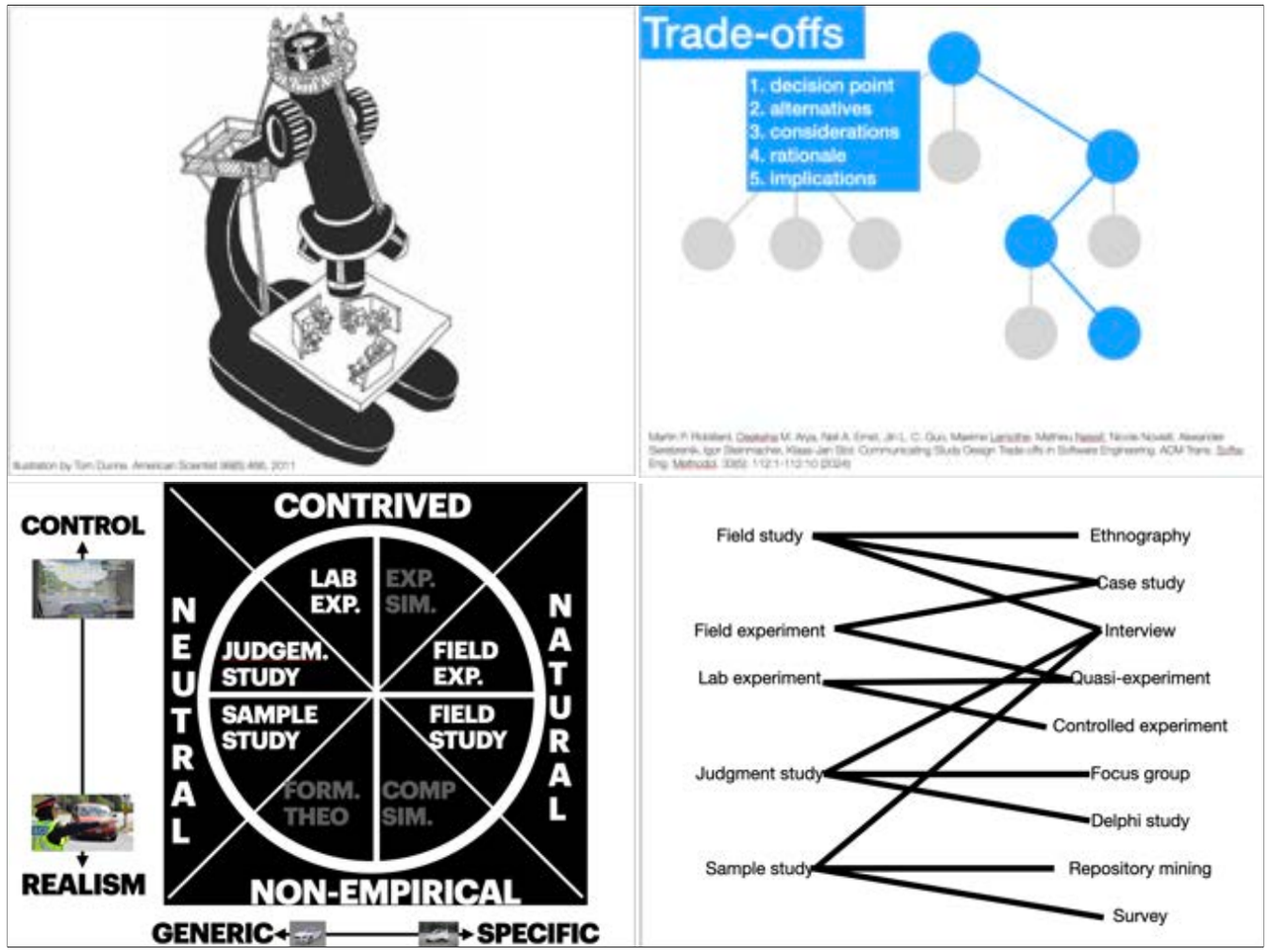
Empirical studies involve numerous decisions on a large scale or on a small scale. Whether we are working on a large scale or on a small scale, what we have is a series of decisions and trade-offs: what research method do we use (interviews or surveys), how do we define “being old”, how do we guess gender?







When it comes to decisions on the large scale, we first need to decide whether we aim at controlling the study or making it as realistic as possible, and whether we want to go broad (general) or deep (specific). None of these choices is inherently right or wrong but they influence the choice of the research strategies.



Each research strategy comes with a series of commonly used research methods. This list is not exhaustive. Furthermore, studies might involve multiple research methods making them multi-method studies or mixed-method studies. Each of the research methods comes with its own decisions: if you are conducting a survey what items will you use and why? if you are mining repositories - which repositories and why? Etc