

Research Methods and Strategies in Software Engineering

Part 2, Week 3 (Sept 25, 2020)

Joseph. E. McGrath.

Methodology matters: Doing research in the behavioral and social sciences. 1972

S Easterbrook, J Singer, MA Storey, D Damian.

Selecting empirical methods for software engineering research. 2008.

Actors:

human systems,
individuals, groups,
organizations,
communities

Behavior:

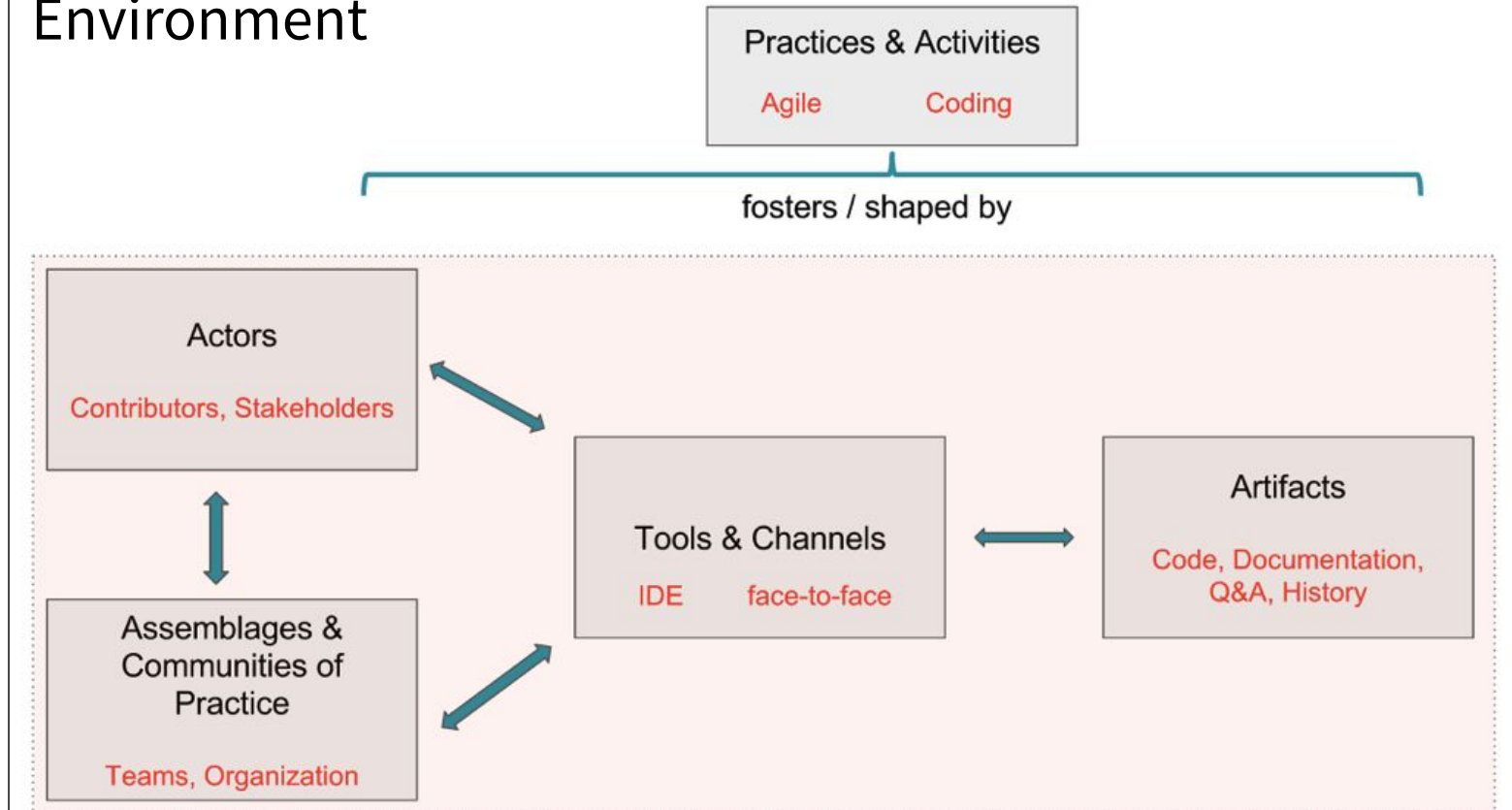
all aspects of the
states and actions of
those human
systems

Context:

temporal, locational
and situational
features in which the
human system is
embedded

2 | Who, what, where?

Environment



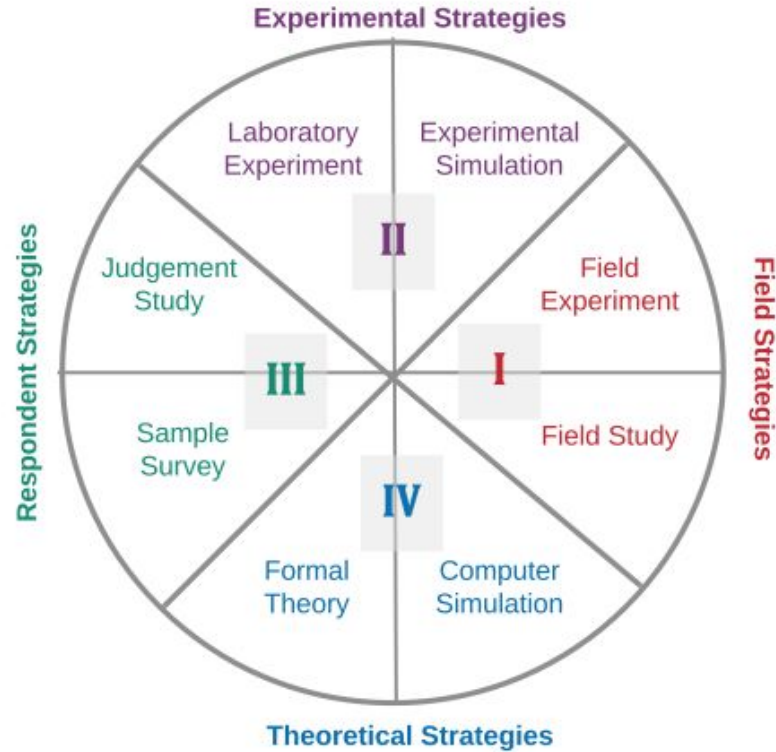
3 | Who, what, where?

Some **content** that is of interest

Some **ideas** that give meaning to that content

Some **techniques** or procedures for studying the content and ideas

4 | Research process



5 | McGrath's Circumflex

Joseph. E. McGrath.
Methodology matters: Doing research in the behavioral and social sciences. 1972

Generalizability

of the evidence over
the populations of
actors

Precision (control)

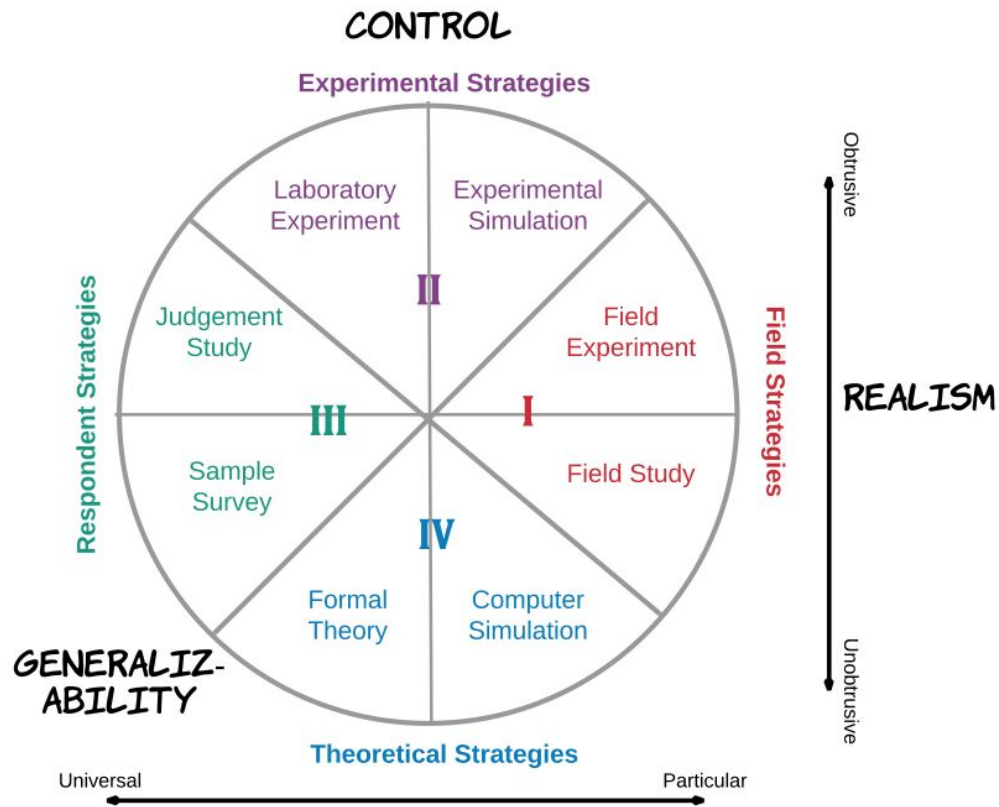
of measurement of
the human
behaviours being
studied

Realism

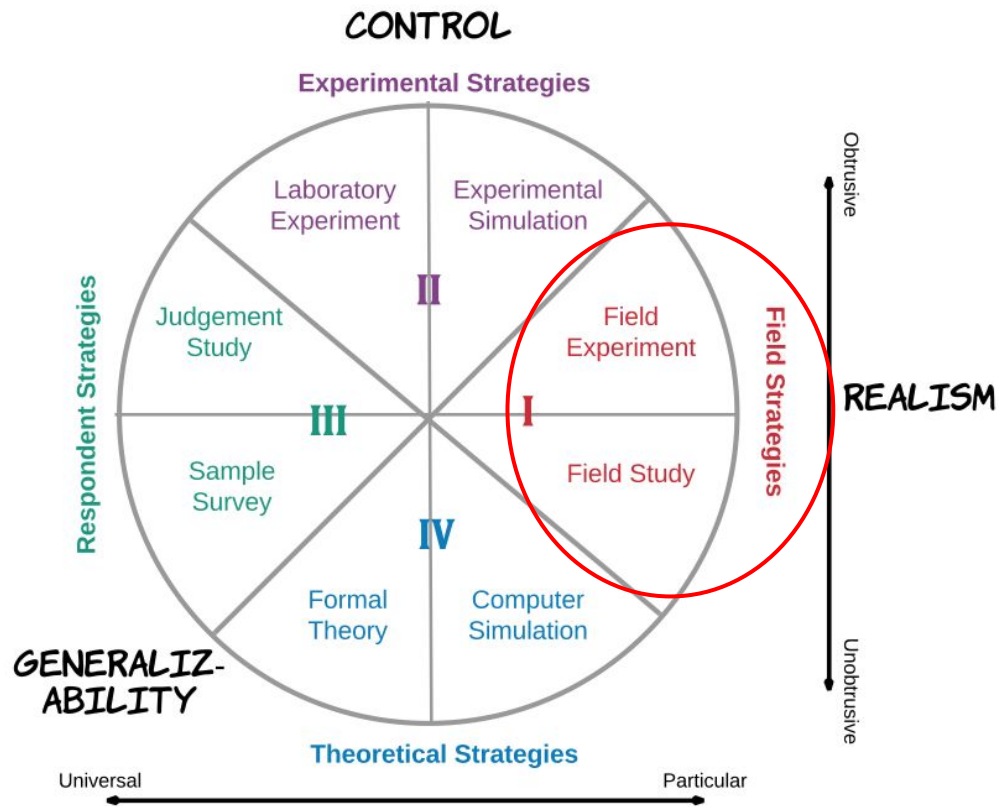
of the situation or
context where the
evidence is gathered



6 | Desirable features of research evidence



7 | Circumflex



8 | Circumflex

Field studies:

Case studies

Ethnography

Action research

Field experiments:

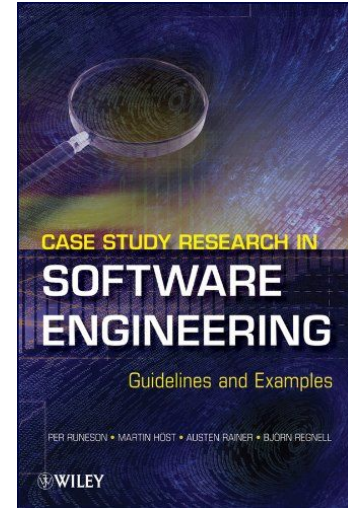
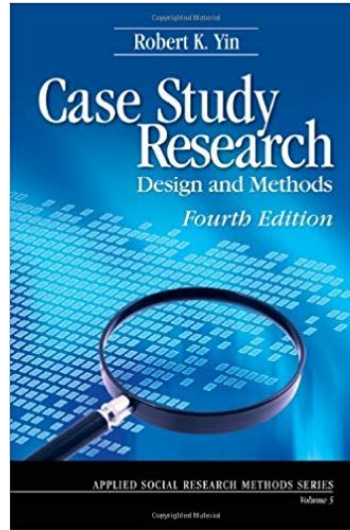
Hypothesis driven

Naturalness is given up for
increased measurement precision

Some variable may be manipulated
(e.g. tool used or process)

Highly obtrusive

*“an empirical inquiry that investigates a contemporary phenomenon within its real-life context, especially when the **boundaries between phenomenon and context are not clearly evident**” Yin*



See also:

http://www.cs.toronto.edu/~sme/case-studies/case_study_tutorial_slides.pdf

Suitable for:

How and why questions

To understand cause and effect

To test a theory

Requires:

A study *proposition* to guide the selection of cases and types of data to collect

Must know your *unit of analysis*

Exploratory case studies derive new theories

Confirmatory case studies test existing theories


But... findings are hard to generalize...

Understand work
practices in context

Insights on culture and
meanings

Actionable insights

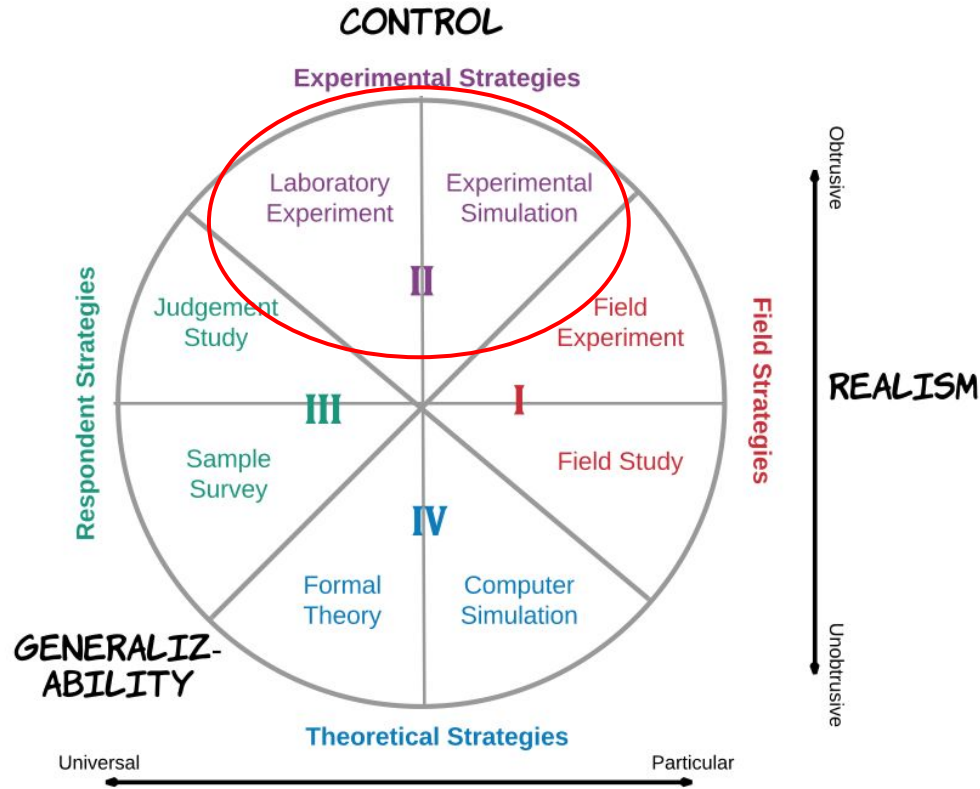
Time consuming analysis,
hard to build theories



"If you want to understand what motivates a guy to pick up skateboarding, you could bring him into a sterile laboratory and interrogate him... or you could spend a week in a skatepark observing him interacting with his friends, practicing new skills and having fun."

Ethnography is observing people's behavior in their own environments so you can get a holistic understanding of their world—one that you can intuit on a deeply personal level."

—LiAnne Yu, cultural anthropologist



13 | Circumflex

Laboratory experiment:

Hypothesis driven

Fabricated setting,

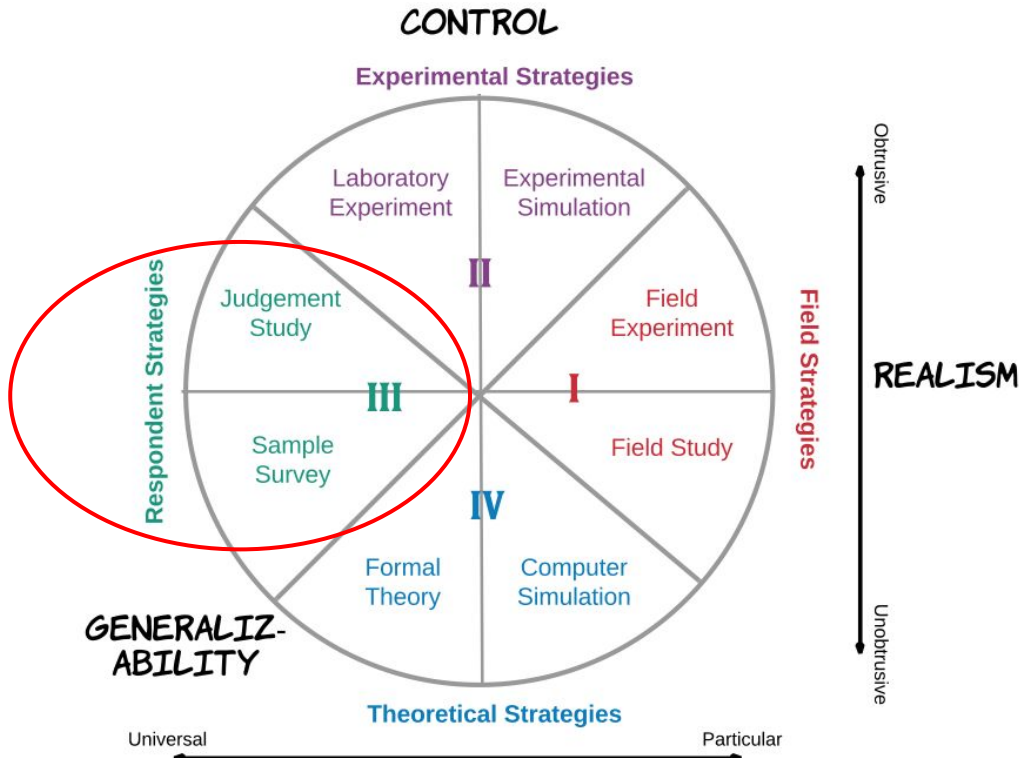
defined rules for its operation,
induces human actors to participate

Increased precision of measurement,
control over human behaviour

Increased **obtrusiveness**, unrealistic
setting and reduced generalizability

Experimental simulation:

Experimenter has **control** over the
setting and conditions – but feels
more like the real setting
e.g. flight simulators



15 | Circumflex

Sample survey:

Collects evidence across a distribution of some variables or relationships among them, within a specific human actor population

Careful sampling must be done to maximize generalizability

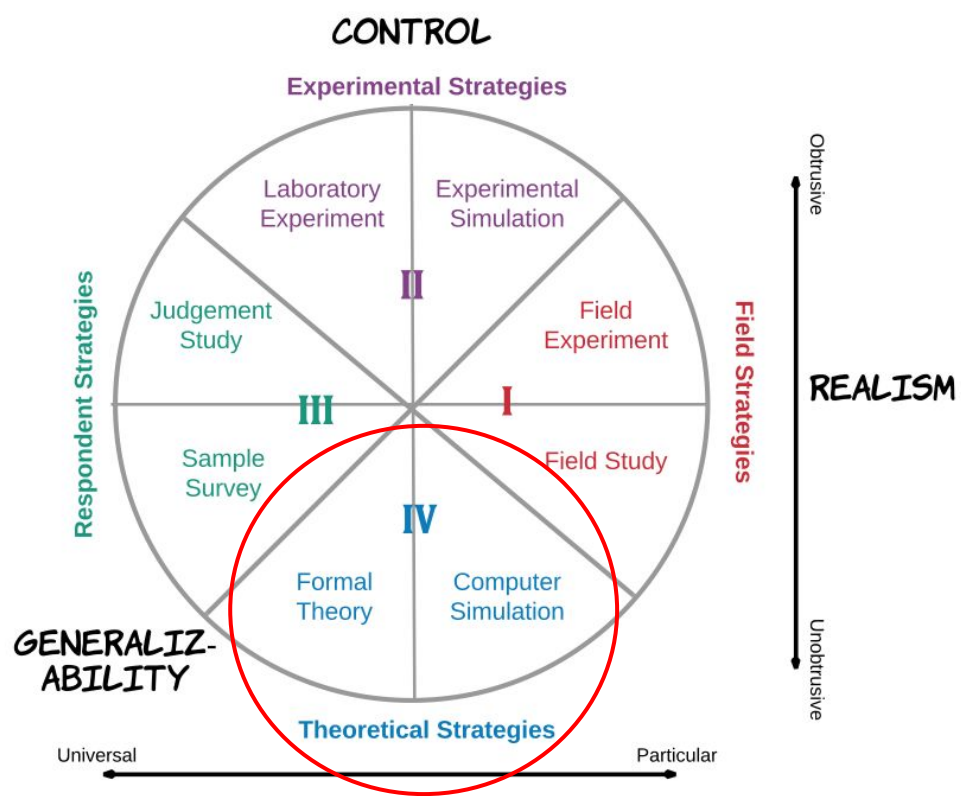
Often imprecise measurements, biases may be present

Judgment study:

Obtaining information about a set of stimulus materials

Usually done with actors of convenience

More precise measurements, but low generalizability



17 | Circumflex

Formal theory:

Theories based on previous empirical evidence or theories

No gathering of empirical observations – relations among variables of interest are formulated

Relations (hypotheses, propositions) should hold over a broad range of populations

Computer simulation:

Contrived setting

A closed system that models the operation of the concrete system but without participants

Behaviour must also be modeled, so all behavioural parameters must be known in advance – based on previous empirical evidence

Actors?

Are the participants aware the record has been made?

Investigators?

When is the record made?

Third parties?

| **Who** makes the record?

Trace data:

“Outcroppings of human behaviour”

Unobtrusive (unaware)

But incomplete, ethical?

Archival records:

Made by third party

Ethical concerns

| Types of data

Self reports:

Versatile, low cost

But may be inaccurate, reactive

Researcher generated data:

Can be precise, controlled

But respondent and researcher bias, reactivity, ambiguity in instruments and collected data

Data studies (based on trace data, archival records)
in software engineering

Program data:

runtime traces,
program logs, system
events, failure logs,
performance logs,
continuous
deployment,...

User data:

usage logs, user
surveys, user forums,
A/B testing, Twitter,
blogs, ...

Development data:

source code versions,
bug data, check-in
information, test
cases and results,
communication
between developers,
social media





“Measurement is the empirical, objective assignment of numbers, according to a rule derived from a model or theory, to attributes of objects or events with the intent of describing them.”

– Kaner, 2004

Product Metrics:

KLOC, Complexity measures (cyclomatic complexity, function points), OO metrics, #defects

Field metrics:

User engagement, user sentiment

Process metrics:

Testing, code review, deployment, agile practices (e.g., #sprints, burndown rate)

Productivity:

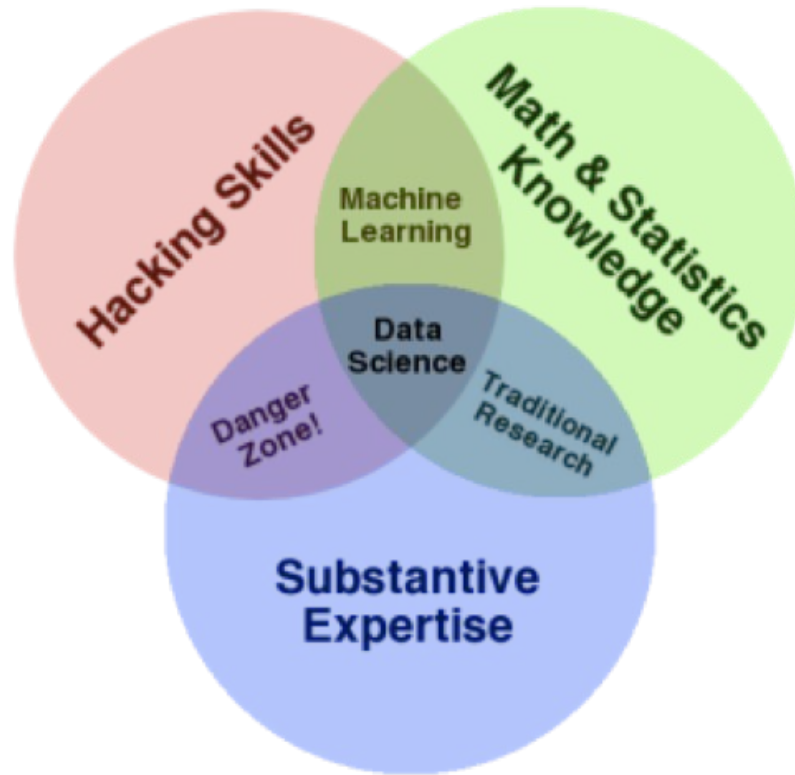
KLOC, Mean time to repair, #commits, team sprint velocity

Developer metrics:

Skills, followers, biometrics

Estimation:

cost metrics and models



Association rules and frequent patterns

Classification

Clustering

Text mining/Natural Language processing

Searching and mining

Qualitative analysis

| Techniques



Ownership
Churn
Tangled code changes



Poor replication
Poor actionability
“Secret life of bugs”

Data

Data may have low
construct validity

Data assumes humans
are “rational animals”

Data does not tell you
why

A repository is not necessarily a (development) project

Most projects are inactive or have few commits

Most projects are for personal use only

Only 10% of projects use pull requests

History can be rewritten on GitHub

A lot happens outside of GitHub

[The Promises and Perils of Mining GitHub](#), Eirini Kalliamvakou et al, MSR 2013.

Data

Data may have low construct validity

Data assumes humans are “rational animals”

Data does not tell you why

Analysis

Correlations are not cause and effect

Big data and small effects

Researcher bias

Martin Shepherd's meta analysis of 24 studies defect prediction:

Technique used had **a small effect**

but

Research group that did the study had a **bigger effect**

Data

Data may have low construct validity

Data assumes humans are “rational animals”

Data does not tell you why

Analysis

Correlations are not cause and effect

Big data and small effects

Researcher bias

Aftermath

Low actionability

Ethics of using data

Unexpected consequences

Biases in algorithms (feedback loops)

“Defect prediction approaches are evaluated on the past history of a system’s bugs, where that history is treated as the future. **A real prediction perturbs the space-time continuum.** Without real world adoption, you simply can’t measure the predictor’s effect. A real prediction perturbs the space-time continuum.”

– [Lanza et al., 2016](#)





34 | Biases in algorithms and diversity

<https://www.sciencemag.org/news/2017/04/even-artificial-intelligence-can-acquire-biases-against-race-and-gender>

We can identify and **predict bugs** in mobile apps

We can identify **insecure code**

We can predict slow **build times**

We can identify **useful features** (and which ones are not)

We can identify which parts of the code are not **“green”**

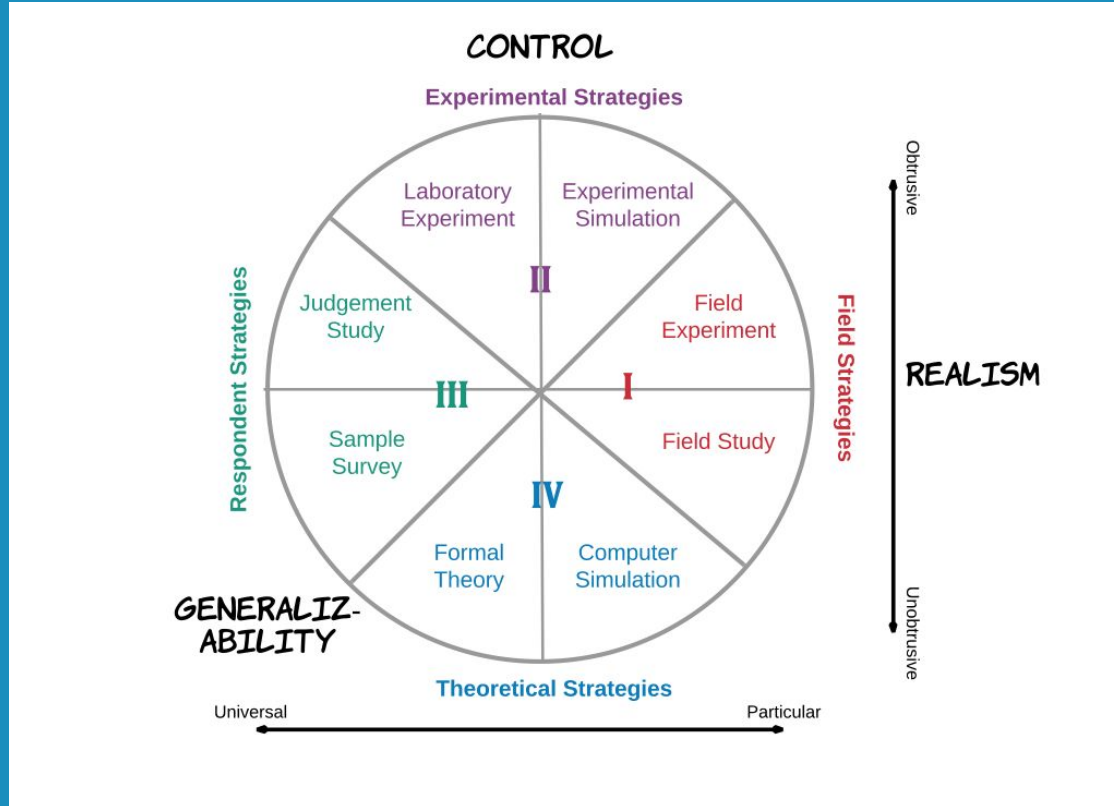
What else? See

<https://www.researchgate.net/publication/264799710> The Road Ahead for Mining Software Repositories

| But we can learn a lot from data mining studies

Data studies in software engineering

How do these fit in McGrath's model?



Triangulation and mixed methods

Investigator triangulation: to reduce bias

Data triangulation: same method, different sources of data

Methodological triangulation: using different strategies and/or methods,

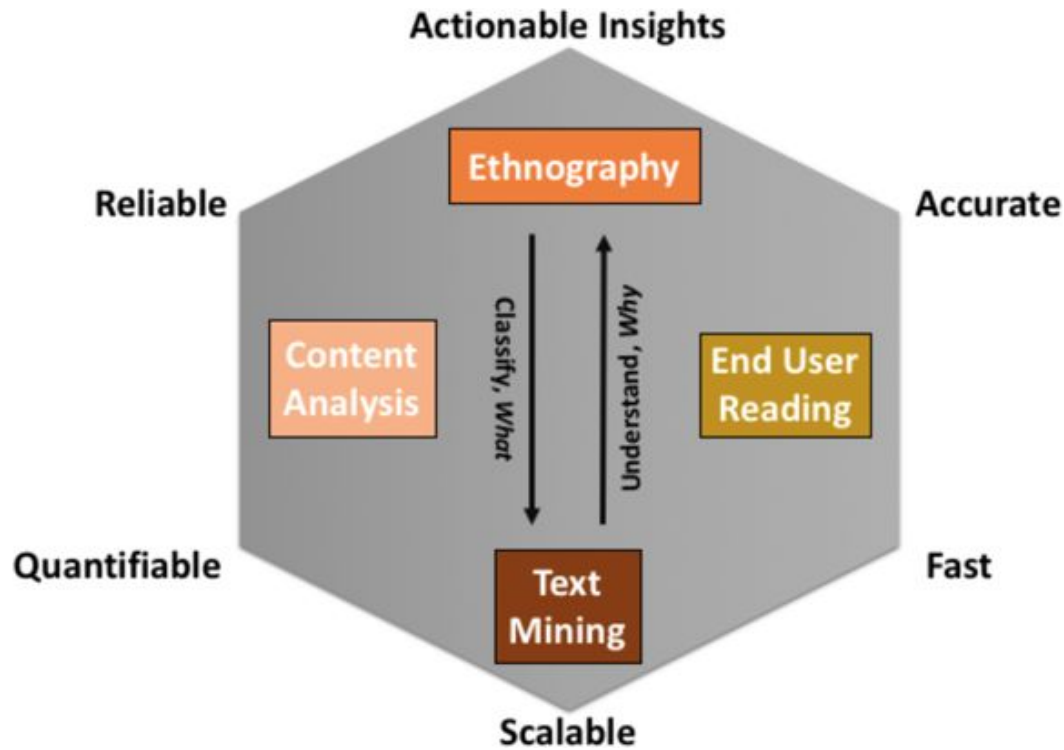
Triangulation of research strategy is how researchers can improve the balance of desirable research quality criteria of generalizability, precision/control and realism

| Triangulation

Sequential explanatory strategy: e.g., quantitative analysis of trace data followed by qualitative analysis of interview data (latter helps explain the former)

Sequential exploratory strategy: e.g., analysis of qualitative data from surveys followed by analysis of quantitative trace data (for testing emerging theory, explain early exploratory findings)

Concurrent triangulation strategy: different methods used concurrently, improve validity



| Another view

<https://www.epicpeople.org/hospital-satisfaction-data/>

Threats to Validity

In all cases, we need to think what are the threats to validity...

What alternative **hypotheses** could explain the results?

Mono-method bias?

Did you **measure** what you thought you measured? Did your participants understand the vocabulary terms the way you did?

Interaction effects?

Is an experiment better described as a pre or quasi experiment – stating the limitations

Threats to validity

Threats to Validity: Positivist stance

Construct validity

Internal validity

External validity

Reliability: would the study yield the same results if done by different researchers?

Internal Validity

What can we **conclude** from the study?

Could it have been due to **chance** (statistical conclusion validity)?

Some **other variables** may have been covarying with X (e.g. age and money) that we did not measure/control

Have you considered all plausible **rival hypotheses**?

Construct validity

How well defined are the **theoretical ideas** in your study?

Do the **methods** you select match the problem?

Are you really **measuring** what you are trying to measure?

External validity

Will the findings hold under replication, that is how **generalizable** are they? What are the limits of how they hold?

External validity (typically) can not be determined from one study – need follow-up/multiple studies

“Validity”: Constructive stance

Triangulation

Member checking

Rich, thick descriptions

Clarify bias (report researcher bias)

Report discrepant information

Prolonged contact with participants

Peer debriefing (plan ahead for this!)

External auditor (also need to plan)

Validity and Qualitative Research: An Oxymoron?

<https://link.springer.com/article/10.1007/s11135-006-9000-3>

Selecting a method

Choice of method depends on the **research question** being asked (exploratory, confirmatory, relationship) as well as the **researcher's philosophical perspective**

S Easterbrook, J Singer, MA Storey, D Damian.
Selecting empirical methods for software engineering research.

Asking vague questions!

Jane: “Is a fisheye view file navigator more efficient than the traditional view for file navigation?”

Joe: “How widely used are UML diagrams used as collaborative shared artifacts during design?”

Your thoughts on these questions?

What kind of research question are you asking?

Exploratory questions:

- Existence question

- Description and classification question

- Descriptive-comparative questions

Base-rate questions:

- Frequency and distribution questions

- Descriptive-process questions

Relationship questions:

- Correlation questions

- Causality questions

- Causality-comparative questions

- Causality-comparative interaction questions

Design questions

1. Why do engineers ignore security warnings in their code?
2. Does test driven development improve code quality?
3. Which code review tool reveals more bugs?
4. Do the topics discussed in online technical forums deter the involvement of female students? Has this changed since online learning?
5. How often does this software fail and in what ways?

Activity: In breakouts, choose best method for answering the above questions (if time discuss limitations/threats to validity)

Comparison techniques

Critical to every empirical study, comparisons are at the heart of the research – depend on elements, relations and context

Three basic forms of comparison techniques:

- Base rates
- Correlational questions
- Difference (or comparison) questions

Baselines

- How **often** does Y occur? (at what rate, what proportion of the time)
- Often done as a precursor to more complex questions...
- Need to know what the **rate** of something is in the general case

Correlational questions

- Is there a **systematic or covariation** in the values of 2 or more properties (or features) of some system?
 - Positive correlation: As X increases (decreases), so does Y
 - Negative correlation: As X increases (decreases), then Y decreases (increases)
 - Zero or low correlation: No observable connection between X and Y
 - Non-linear correlations: e.g. X and Y may covary, then flatten, then covary again... need more powerful statistical tools to study non-linear correlations
- May look at more than two variables
- Note correlations do not necessarily indicate causal relationships

Difference (comparison) questions

- Is Y **present** (**absent**) when X is present or high (absent or low)?
 - E.g. Do software engineers collaborate more effectively when they have had face-to-face meetings?
- Need to look for **interaction effects** of other variables – not always easy to hold other factors constant

Dealing with other factors!

- **Randomization** – 2 aspects
 - Sampling: how we select actors from a given population
 - How we allocate cases to conditions
- Note: you do not select a random sample, you select a sample using a *random procedure*!
- *Sample size* is critical – the larger the sample, the more likely it is you have a random sample (but be careful with this too!)
- Even doing all of the above won't lead to logical conclusions – just increases the likelihood or **probability** that X causes Y (could be other factors that were not evenly distributed)
- Need to reduce the scope to improve the power of the randomization -- realism is removed as we selected the participants, created the tasks and created the conditions