Empirical Software Engineering Research

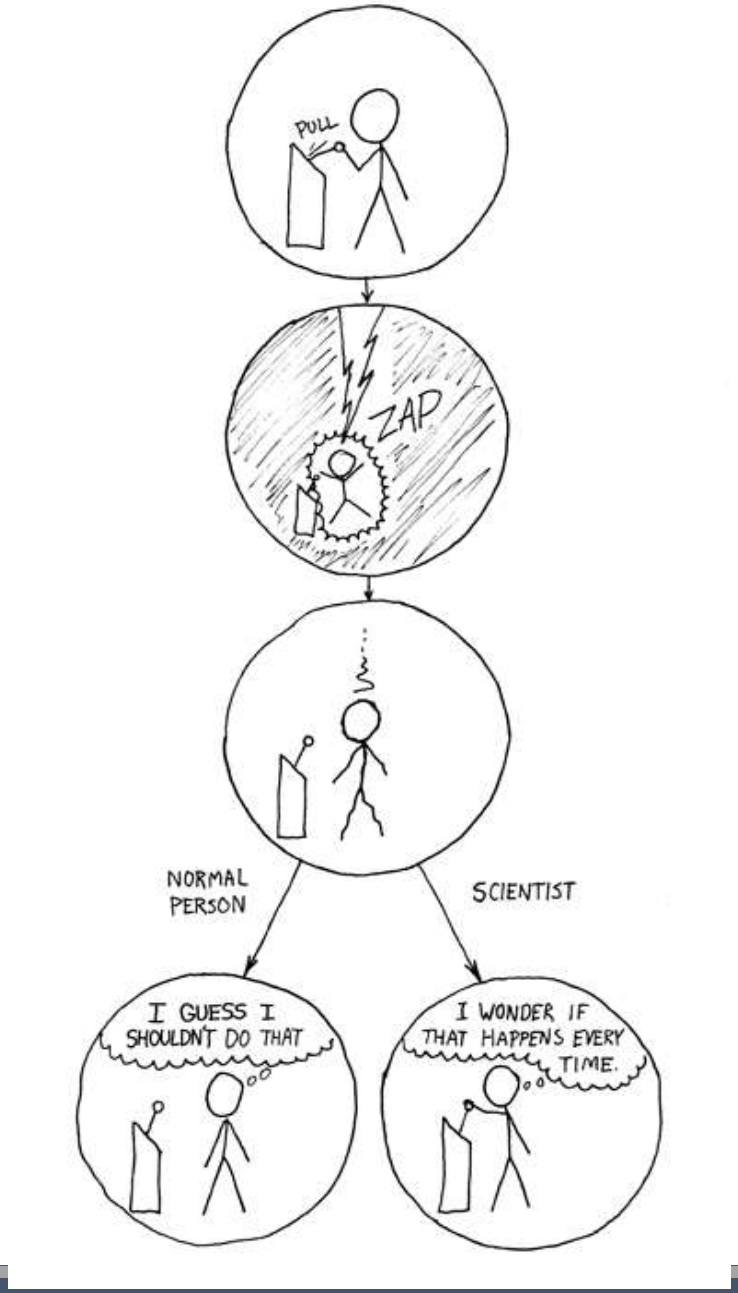
Introduction

Norman Peitek, Annabelle Bergum, Lina Lampel, Sven Apel

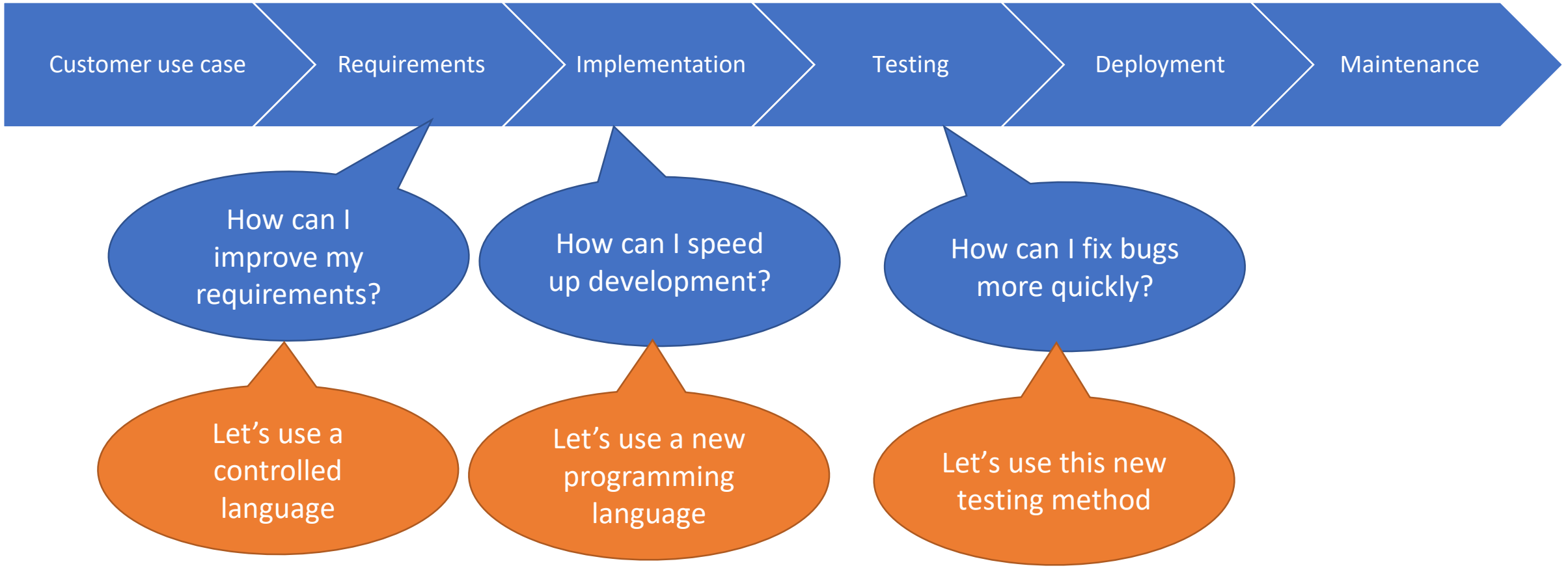
Topic & Learning Goals

- Understand empirical methods in software engineering (research)
- Why do we need empirical methods?
- Get an overview over different empirical methods in SE

Motivation


PULL
ZAP
NORMAL PERSON
SCIENTIST
I GUESS I SHOULDN'T DO THAT
I WONDER IF THAT HAPPENS EVERY TIME.

Software Engineering



All solutions fail to consider some aspect of the problem and are not proven

Software Engineering

- In software engineering, we often believe we can infinitely scale up our solutions
- The real world is complex → we cannot simply adapt from a simple example
- Software development is complex and context-dependent
- We must understand the real world to be effective

Motivation

- Many decisions in SE were (are?) opinion-based or historically driven

```
class ConferenceSpeaker {
    String SpeakerName;
    String TalkTitle;
}
```

„**By convention**, C# programs use `PascalCase`“. Microsoft C# Programming Guide

```
class ConferenceSpeaker {
    String speakerName;
    String talkTitle;
}
```

„**By convention**, … [`camelCase`] …”. Oracle Java

```
class ConferenceSpeaker {
    String speaker_name;
    String talk_title;
}
```

„[…] should be lowercase, with words separated by underscores as necessary **to improve readability**.“ Python PEP8

+20%

Empirical Research

- Empirical: based on experience and observation
- Empirical research: method to *scientifically* gain knowledge
  - Use of a (not “the”) scientific method to investigate software engineering problems
- Start from observation, formulate hypothesis, select methods, collect data, and draw conclusions
- Basic idea: if we understand how things work, we can improve them
  - Knowledge helps us to improve and solve problems in the real world

Empirical Research

Empirical research is NOT:

- Theoretical thoughts
- Intuition
- Random selection
- Authority
- Persistence
- Empiricism

Motivation

- In 1995, the state of empirical research in CS has ben described as “unacceptable”
  - In random samples, 40-50% of published articles about new designs or models completely lacked empirical evidence
  - This situation is much inferior to other research fields, in particular natural sciences
  - “Large parts of CS may not meet standards long established in the natural and engineering sciences” [1]

- What is the situation like today?
  - Many publication vendors require empirical methods
  - Specific conferences and journals for empirical methodology (ESEM, EMSE)
  - For SE, the empirical rate has increased to 95+%

Example Question

"The programming language Python makes programmers more productive!"

→ How would you evaluate this statement? Can you (dis)prove it?

Looking for Solid Evidence

- What do you know about the topic?
- What statement/theory is linked to this topic? What have you learned in other courses?
- Which evidence do you know (e.g., from other courses)?
- Does that match with your experience?
- What kind of evidence would convince you?

Example Question

"Pair programming makes programmers more productive!"

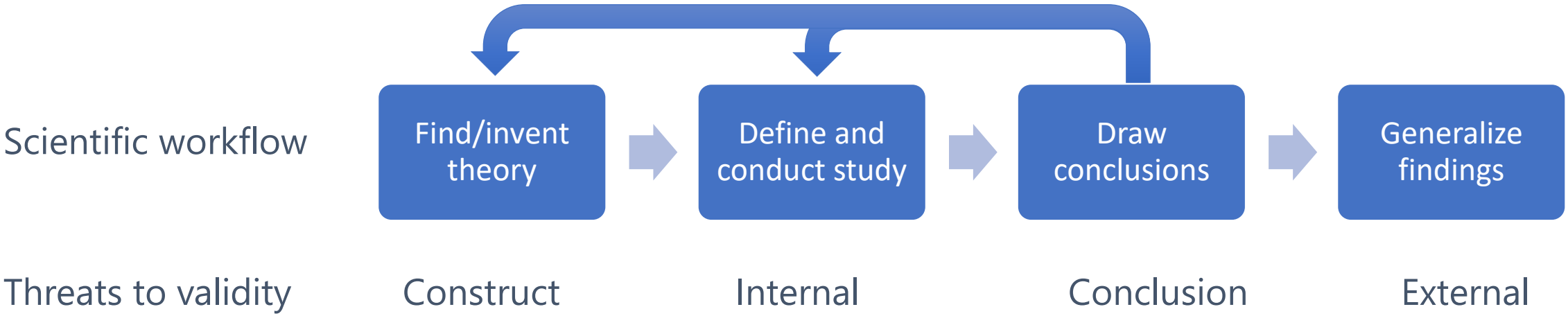
→ How would you evaluate this statement? Can you (dis)prove it?

Looking for Solid Evidence

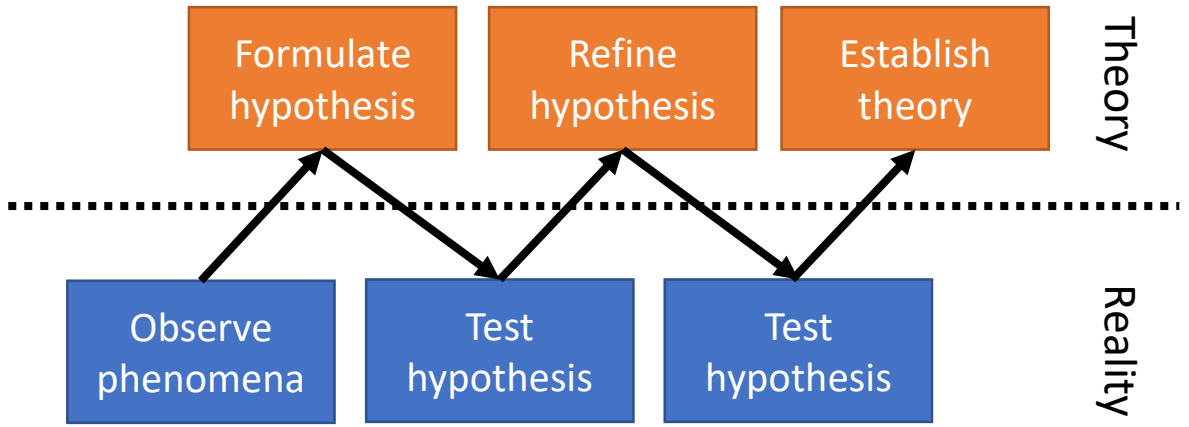
- What do you know about the topic?
- What statement/theory is linked to this topic? What have you learned in other courses?
- Which evidence do you know (e.g., from other courses)?
- Does that match with your experience?
- What kind of evidence would convince you?

The Scientific Method

- Traditional experiment-based scientific research follows a relatively strict conceptional workflow
  - The goal is to achieve the most reliable results (= high degree of validity)
  - The main concern is reliability and reproducibility



The Scientific Method



1. Observe phenomena
2. Formulate hypothesis
3. Test hypothesis via rigorous experiment
   - If necessary, modify hypothesis and go back to Step 2
4. Establish theory based on repeated validation of results

→ Research is slow! Discovery of Higgs-Boson took decades (and billions of $), we expect SE to be much faster (and cheaper)

THE SCIENTIFIC METHOD... FOR TEN-YEAR OLDS

* THIS IS SURPRISINGLY CLOSE TO HOW REAL SCIENTISTS ACT AT CONFERENCES.

Engineering versus Science

The scientist builds in order to study;

the engineer studies in order to build.

F. Brooks. *The Computer Scientist as Toolsmith II*. Communications of the ACM, 39:3, 1996.

- Scientist
  - Understanding as goal (facts, relationships)
  - Construction as far as necessary to fulfill goal
- Engineer
  - Construction of something useful as goal
  - Understanding as a way to better construction

Computer Science

- Rooted in mathematics (theory)
- Electrical engineering
- Today: huge engineering part in many areas (e.g., when constructing UIs)
- Used by people (psychology, politics)
- Empirical research is becoming more and more important

Mathematical Proof vs. Empirical Research

- Proof of a closed system
- Formalization of statement and research topic
- E.g., mathematical induction

- Unchallengeable

- Cannot always be formalized
- E.g., interaction with people
- Result is observable, but not provable

- No final result
- Collect evidence
- Falsification

Statements

Statements cannot be proven, but observed

- Example: Copy & Paste causes errors
- Behavior of users (errors) cannot be proven, because there is no formal model of a user
- But behavior of users can be observed (e.g., during or after development, we can examine whether errors are related to copy & paste)

Human Factor

Humans use a software tool or develop software

- Human behavior is typically non-deterministic (mood, daily state of mind)
- Intra-individual differences are difficult to determine

Most likely large difference between individuals

- Skills, education, personal preferences

Many (possibly causal) relationships are currently unknown

- When does a user/programmer make an error?
- When is a UI/source code less usable/comprehensible?

Towards Empirical Research

- Can one single observation be used as evidence for a statement?
- Example: Write a program „Hello World“ in Java. Let your colleague write „Hello World“ Python. Who needs more time?
  - Would this support Java or Python?

- Example: Write „Hello World“ in Java. The next day, write „Hello World in Python. The day after that, write „Hello World“ in Java.
  - Development time on Day 3 will be different than on Day 1 and 2. Can you draw conclusions based on that?

Towards Empirical Research

- Can the personal opinion/perception be used to confirm a statement?
- Example: Assume that you just love the new material design UI on Android. Assume that your neighbor also loves the new UI.
  - Does that mean that the new UI is good?
- Example: Assume that the results of a survey show that most users love the new UI.
  - Does that mean that the new UI is good?

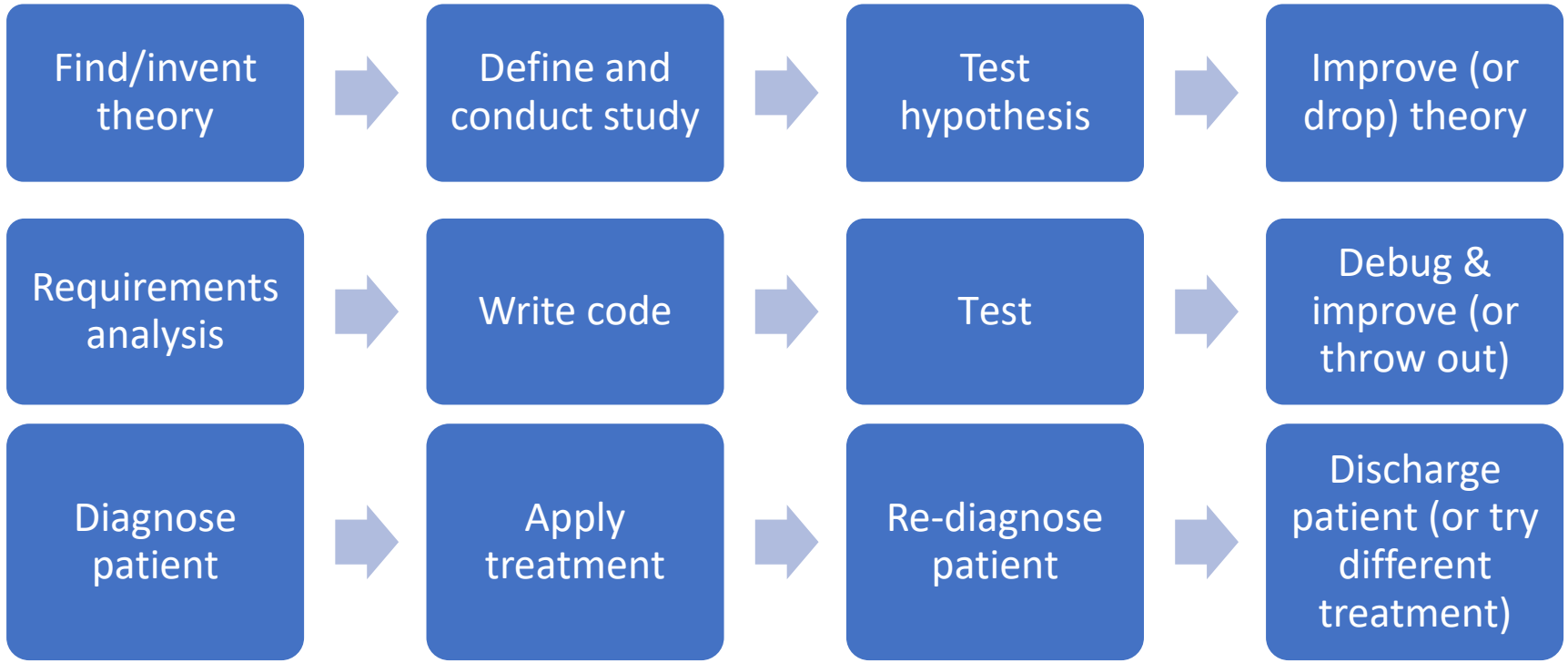
Towards Empirical Research

- How can we use observations as scientific method?

- Empirical methods:
  - Data collection: What kind of data can we observe where?
  - Qualitative vs. quantitative observations: Which kind of information can we collect?
  - Logic of empirical research: How can we conclude statements or contradictions from data?
  - Experiment, field studies, case study, etc.: Under which conditions can we conclude what kind of statements/contradictions?

Empirical Methods to Non-Scientists (Programmers)

- Scientists:
- Programmers:
- Medicine:



Overview of Empirical Methods in SE

- No universal taxonomy of research methods
- We try to stay consistent
  - But the methods may be named differently in the literature

Glass et al. [63]	Zannier et al. [230]	Sjøberg et al. [190]	Höfer and Tichy [75]	Easterbrook et al. [48]
Action research	Controlled experiment	Controlled experiment	Case study	Experimentation
Conceptual analysis	Quasi experiment	Surveys	Correlational study	Case study
Concept implementation	Case study	Case studies	Ethnography	Survey
Case study	Exploratory case study	Action research	Ex post facto study	Ethnography
Data analysis	Experience report		Experiment	Action research
Discourse analysis	Meta-analysis		Meta-analysis	
Ethnography	Example application		Phenomenology	
Field experiment	Survey		Survey	
Field study	Discussion			
Grounded theory				
Hermeneutics				
Instrument development				
Laboratory experiment (human/software)				
Literature review				
Meta-analysis				
Mathematical proof				
Protocol analysis				
Phenomenology				
Simulation				
Descriptive/expl. survey				

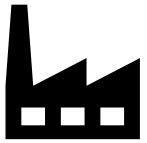
Overview of Empirical Methods in SE

Controlled experiments

Survey

Interviews

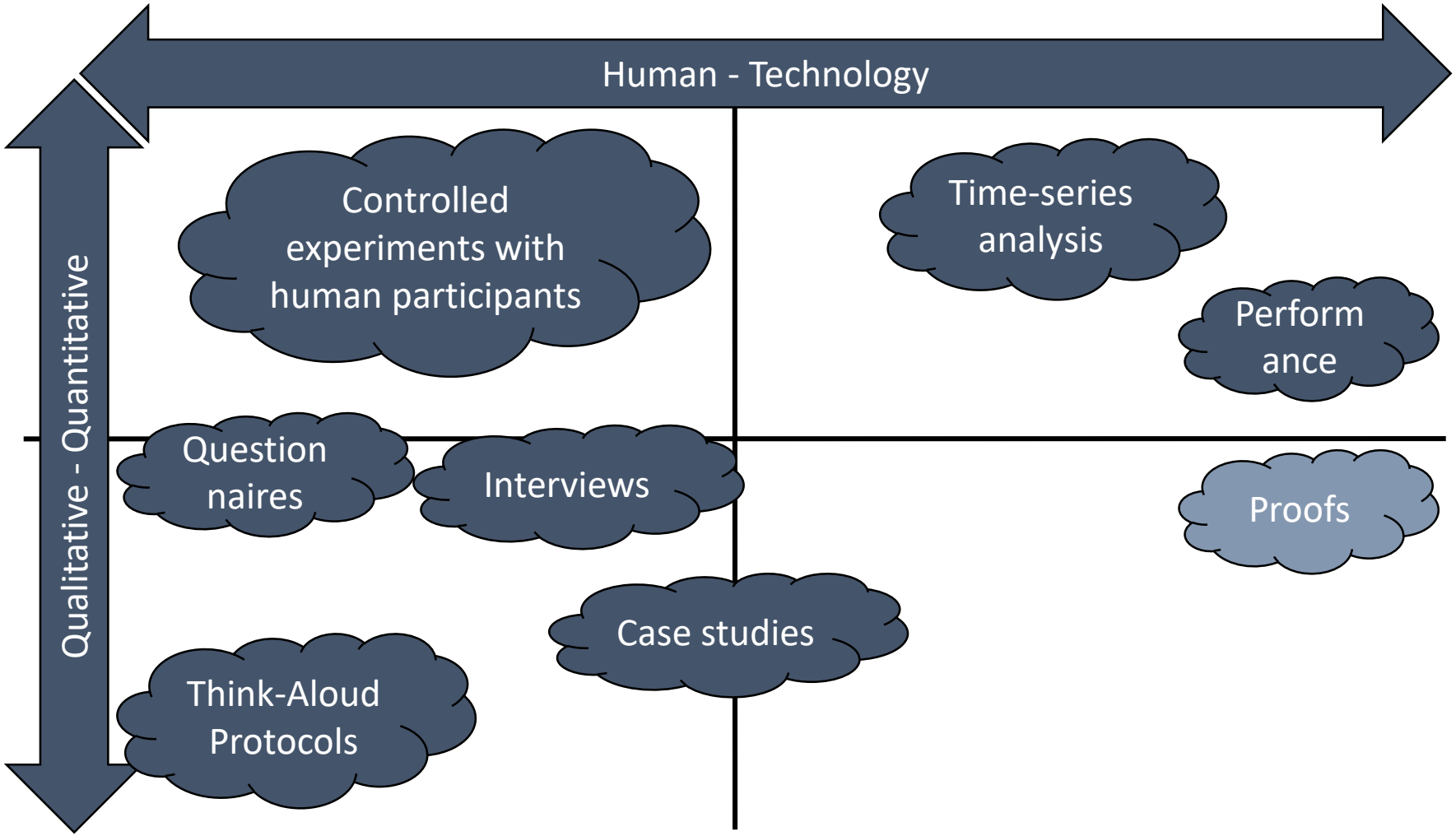
Case studies



Field studies

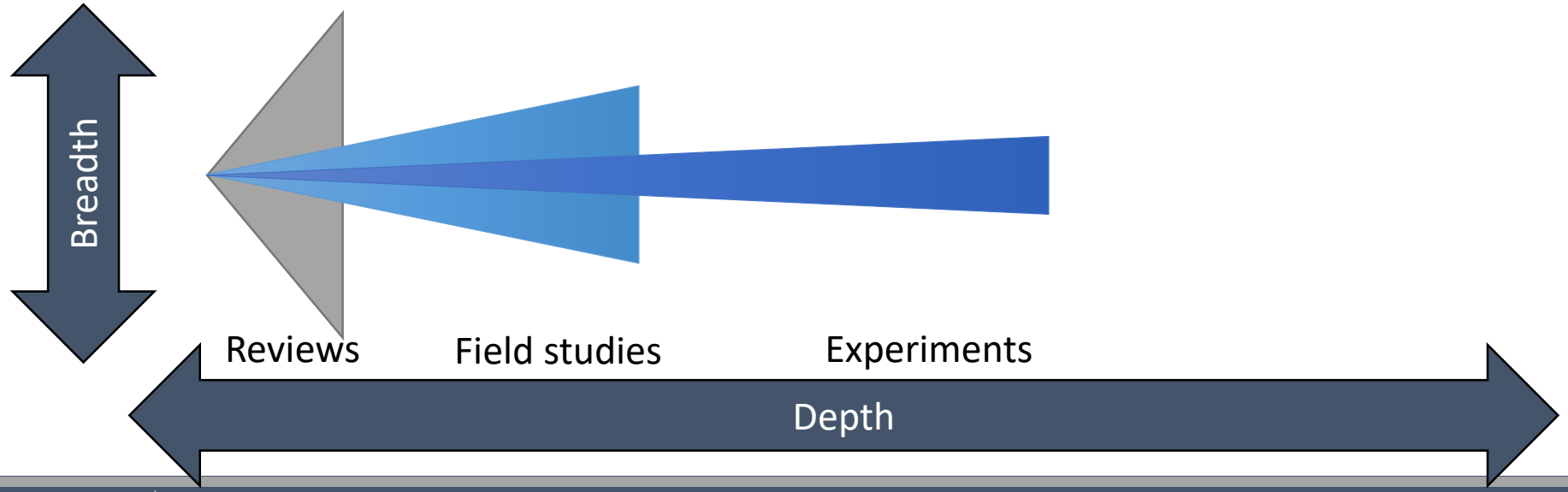
Literature review

Overview of Primary Empirical Methods in SE



Overview of Empirical Methods in SE

- On a high level, there are three classes of empirical methods:
  - Controlled experiments
  - Field studies (interviews, case studies, …)
  - Secondary studies (literature reviews, meta analysis, …)
- They all have their respective strengths and weaknesses



Experiments: Correlation and Causation

- A controlled experiment is the ideal type for all quantitative empirical methods
  - Strictly speaking, results of an experiment are always correlations, though many people do not understand the difference
- Demanding tight control over experimental variables is usually only possible in lab environments
  - In the “field”, there are usually too many disturbing variables
  - But, if the context is important, a lab experiment is useless

Field Studies: Understanding

- Experiments take place in a lab to control environmental conditions
  - This is problematic when the environment is important (for example, studying hunting behavior of a lion in the wild versus in a zoo)
  - In these cases, field studies is required with typically qualitative methods (interviews, case studies, …)
- The advantage is the ability to study a subject in context (without restrictions of an experiment)
- The drawback is that we are typically not able to derive correlations (or causality), no control over events
- We will derive understanding and insights

Research Goals

- There are five basic goals any empirical study can pursue
- Exploration
  - Try to understand a problem in greater detail, particularly with a view towards making the problem known in a wider context, and informing further research
- Explanation
  - Try to develop a theory for a domain, based on previous explorations and/or disproved theories
- Demonstration
  - Prove that something can be done in a specific way, in particular one that is derived from or linked to a theory
- Codification
  - Try to establish causal relationships to codify a theory and make predictions about phenomena

Research Goals

- Big questions need big studies (= a research program rather than single study)
  - Each research instrument can only contribute to some of these goals
  - By combining different research instruments, we can achieve both greater scope and greater depth of insight

- Triangulation is a method to increase validity of research by studying an object from several points of view
  - Originally from social sciences, but applies to SE as well

Truth versus Results

- We do not accept a scientific result as being the truth unless adequate evidence is presented in an argument that convinces us
- Frequently, we trust in a result for the wrong reasons
  - People tend to believe a result if previous work or some authority says so
  - Own circumstantial experience is often valued higher than that of others
  - Intuition can be deceiving, things that make sense are not necessarily true
- Trust is really only justified by applying a rigorous procedure
  - Scientists apply a scientific method (and they did not cheat)
  - Fellow scientists have peer-reviewed the procedure (and are fair, attentive, and unbiased)
  - Other fellow scientists have replicated the results
  - Of course, this process does not guarantee perfection: mistakes happen, but should be found eventually

Scientific Proof

- A scientific proof in SE earns trust-worthiness by the applied scientific procedure
- Scientists prove do not the correctness of a given hypothesis, but that the opposite hypothesis (null hypothesis) is wrong (“Falsification”)
  - “Proving wrong” means that the plausible alternative hypothesis have extremely low likelihood
  - “Proving right” means to compile sufficient supportive evidence so that the likelihood of being wrong converges to zero
- “I have seen white swans. I have not seen a black one. All swans must be white.”
  - Clearly, any such hypothesis can be falsified at any time, but never proven right (as there could be an infinite number of swans)

Scientific Truth

- A scientific truth requires evidence
  - In the end, scientific truth is about believing in a given result because of the strength of the argument and the evidence presented
- Scientific truth is provisional
  - Newton's laws of motion from 1687 hold up for centuries before limitations were found

Non-Results & Negative Results

- A non-result is where no evidence was found (typically: p-values too high)
  - There a three possible reasons
    - The study design was poor or unsuitable
    - The study execution was poor
    - The expectations are genuinely wrong, the theory is incomplete/wrong
- A negative result is where the evidence was convincing, but did not support the tested hypothesis
  - This means the wrong question was asked → implies the correct question to ask
  - Or, the theory does not fit the observations → leading to an improved theory
- In empirical work, negative results are almost as good as positive ones
  - But sometimes hard to publish (or only under specific tracks/journals)

Further Reading

- [1] Tichy, Walter, et al. "Experimental evaluation in computer science: A quantitative study." *Journal of Systems and Software* 28.1 (1995): 9-18.
- [2] Tichy, Walter "Should computer scientists experiment more?." *Computer* 31.5 (1998): 32-40.