



# CSC2130: Empirical Research Methods for Software Engineering

Steve Easterbrook

sme@cs.toronto.edu

[www.cs.toronto.edu/~sme/CSC2130/](http://www.cs.toronto.edu/~sme/CSC2130/)



## Course Goals

### → Prepare students for advanced research (in SE):

- ↳ Learn how to plan, conduct and report on empirical investigations.
- ↳ Understand the key steps of a research project:
  - ↳ formulating research questions,
    - > theory building,
    - > data analysis (using both qualitative and quantitative methods),
    - > building evidence,
    - > assessing validity,
    - > publishing.

### → Motivate the need for an empirical basis for SE

### → Cover all principal empirical methods applicable to SE:

- ↳ controlled experiment, case studies, surveys, archival analysis, action research, ethnographies,...

### → Relate these methods to relevant meta-theories in the philosophy and sociology of science.





## Intended Audience

### → This is an advanced software engineering course:

- ↳ assumes a strong grasp of the key ideas of software engineering and the common methods used in software practice.

### → Focus:

- ↳ how do software developers work?
- ↳ how do new tools and techniques affect their ability to construct high quality software efficiently?
- ↳ qualitative and quantitative techniques from behavioural sciences

### → The course is aimed at students who:

- ↳ ...plan to conduct SE research that demands some empirical validation
- ↳ ...wish to establish an empirical basis for an existing SE research programme
- ↳ ...wish to apply these techniques in related fields (e.g. HCI, Cog Sci)

### → Note: we will *\*not\** cover the kinds of experimental techniques used in CS systems areas.



## Format

### → Seminars:

- ↳ 1 three-hour seminar per week
- ↳ Mix of discussion, lecture, student presentations

### → Readings

- ↳ Major component is discussion of weekly readings
- ↳ Please read the set papers before the seminar

### → Assessment:

- ↳ 10% Class Participation
- ↳ 30% Oral Presentation - *\*critique a published empirical study*
- ↳ 60% Written paper - *design an empirical study for a SE research question*

*\*As part of a mock conference program committee meeting*





## Course Outline

1. **Introduction & Orientation**
2. **What is Science?**
  - ↳ Philosophy of Science
  - ↳ Sociology of Science
  - ↳ Meta-theories
3. **What is software engineering?**
  - ↳ Engineering & Design
  - ↳ Disciplinary Analogies for SE
  - ↳ Evidence-based software engineering
4. **Basics of Doing Research**
  - ↳ Finding good research questions
  - ↳ Theory building
  - ↳ Research Design
  - ↳ Ethics
  - ↳ Evidence and Measurement
  - ↳ Peer Review Process



## Course Outline (cont)

- |   |  |
|---|--|
| <ol style="list-style-type: none"><li>5. <b>Experiments</b><ul style="list-style-type: none"><li>↳ Controlled Experiments</li><li>↳ Quasi-experiments</li><li>↳ Sampling</li><li>↳ Replication</li></ul></li><li>6. <b>Case Studies</b><ul style="list-style-type: none"><li>↳ Single and Multi-case</li><li>↳ Longitudinal Case Studies</li><li>↳ Approaches to Data Collection</li></ul></li><li>7. <b>Histories and Simulations</b><ul style="list-style-type: none"><li>↳ Artifact Analysis</li><li>↳ Archival Analysis and Post-mortems</li><li>↳ Simulation Techniques</li></ul></li><li>8. <b>Survey and Observation</b><ul style="list-style-type: none"><li>↳ Surveys</li><li>↳ Focus Groups</li><li>↳ Field Studies / Ethnographies</li></ul></li></ol> | <ol style="list-style-type: none"><li>9. <b>Interventions</b><ul style="list-style-type: none"><li>↳ Action Research</li><li>↳ Pilot Studies</li><li>↳ Benchmarking</li></ul></li><li>10. <b>Qualitative Analysis</b><ul style="list-style-type: none"><li>↳ Grounded Theory</li><li>↳ Phenomenography</li><li>↳ Mixed Methods Research</li></ul></li><li>11. <b>Quantitative Analysis</b><ul style="list-style-type: none"><li>↳ Stats, power analysis, meta-analysis</li></ul></li><li>12. <b>Publishing and Reviewing</b><ul style="list-style-type: none"><li>↳ (mock PC meeting)</li></ul></li><li>13. <b>Replication and Beyond</b><ul style="list-style-type: none"><li>↳ Internal and External Replication</li><li>↳ Biases and Influences</li><li>↳ Threats to Validity</li><li>↳ When to use empirical methods</li></ul></li></ol> |
|---|--|





## Is this your research plan?

**Step 1: Build a new tool**

**Step 2: ??**

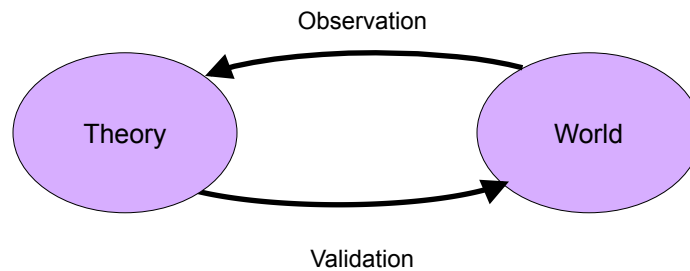
**Step 3: Profit**



## Scientific Method

→ No single “official” scientific method

→ Somehow, scientists are supposed to do this:

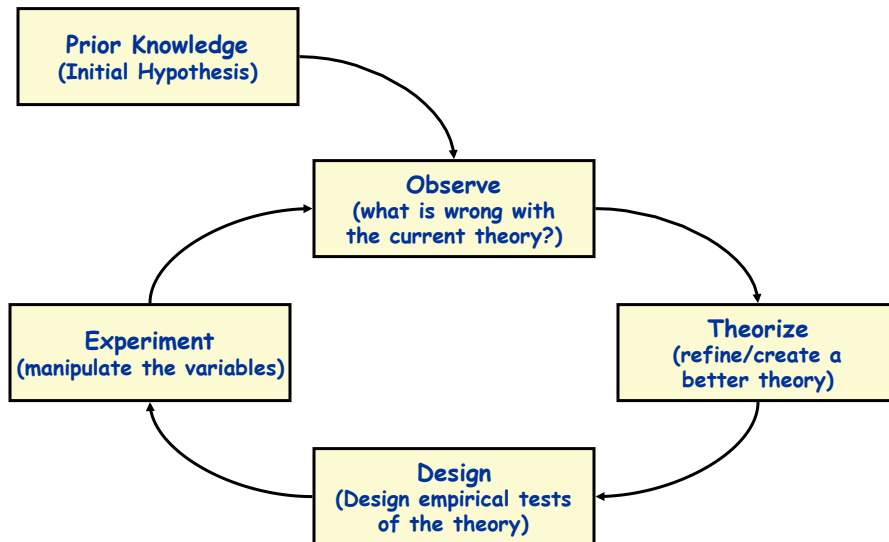




# Observe!



## Scientific Inquiry





## Some Characteristics of Science

- **Science seeks to improve our understanding of the world.**
- **Explanations are based on observations**
  - ↳ Scientific truths must stand up to empirical scrutiny
  - ↳ Sometimes “scientific truth” must be thrown out in the face of new findings
- **Theory and observation affect one another:**
  - ↳ Our perceptions of the world affect how we understand it
  - ↳ Our understanding of the world affects how we perceive it
- **Creativity is important**
  - ↳ Theories, hypotheses, experimental designs
  - ↳ Search for elegance, simplicity



## All Methods are flawed

- **E.g. Laboratory Experiments**
  - ↳ Cannot study large scale software development in the lab!
  - ↳ Too many variables to control them all!
- **E.g. Case Studies**
  - ↳ How do we know what's true in one project generalizes to others?
  - ↳ Researcher chose what questions to ask, hence biased the study
- **E.g. Surveys**
  - ↳ Self-selection of respondents biases the study
  - ↳ Respondents tell you what they think they ought to do, not what they actually do
- ...etc...





# Strategies to overcome weaknesses

## → Theory-building

- ↳ Testing a hypothesis is pointless (single flawed study!)
- ↳ ...unless it builds evidence for a clearly stated theory

## → Empirical Induction

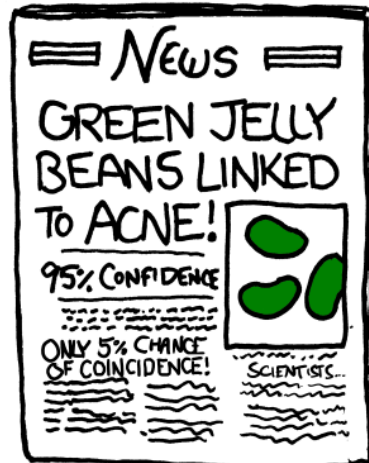
- ↳ Series of studies over time...
- ↳ Each designed to probe more aspects of the theory
- ↳ ...together build evidence for a clearly stated theory

## → Mixed Methods Research

- ↳ Use multiple methods to investigate the same research question
- ↳ Each method compensates for the flaws of the others
- ↳ ...together build evidence for a clearly stated theory



WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE ( $P > 0.05$ )	WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE ( $P > 0.05$ )	WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE ( $P > 0.05$ )	WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE ( $P > 0.05$ )	WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE ( $P > 0.05$ )
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE ( $P > 0.05$ )	WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE ( $P > 0.05$ )	WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE ( $P > 0.05$ )	WE FOUND NO LINK BETWEEN PINKISH JELLY BEANS AND ACNE ( $P > 0.05$ )	WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE ( $P > 0.05$ )
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE ( $P > 0.05$ )	WE FOUND NO LINK BETWEEN TANK JELLY BEANS AND ACNE ( $P > 0.05$ )	WE FOUND NO LINK BETWEEN CRANK JELLY BEANS AND ACNE ( $P > 0.05$ )	WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE ( $P < 0.05$ )	WE FOUND NO LINK BETWEEN MARINE JELLY BEANS AND ACNE ( $P > 0.05$ )
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE ( $P > 0.05$ )	WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE ( $P > 0.05$ )	WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE ( $P > 0.05$ )	WE FOUND NO LINK BETWEEN PENCH JELLY BEANS AND ACNE ( $P > 0.05$ )	WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE ( $P > 0.05$ )



Source: <http://xkcd.com/882/>





## What is a research contribution?

- A better understanding of how software engineers work?
- Identification of problems with the current state-of-the-art?
- A characterization of the properties of new tools/techniques?
- Evidence that approach A is better than approach B?

**How will you validate your claims?**



## Meet Stuart Dent

- **Name:**
  - ↪ Stuart Dent (a.k.a. "Stu")
- **Advisor:**
  - ↪ Prof. Helen Back
- **Topic:**
  - ↪ Merging Stakeholder views in Model Driven Development
- **Status:**
  - ↪ 2 years into his PhD
  - ↪ Has built a tool
  - ↪ Needs an evaluation plan





## Stu's Evaluation Plan

### → Formal Experiment

- ↳ Independent Variable: *Stu-Merge vs. Rational Architect*
- ↳ Dependent Variables: *Correctness, Speed, Subjective Assessment*
- ↳ Task: *Merging Class Diagrams from two different stakeholders' models*
- ↳ Subjects: *Grad Students in SE*
- ↳ H<sub>1</sub>: *"Stu-Merge produces correct merges more often than RA"*
- ↳ H<sub>2</sub>: *"Subjects produce merges faster with Stu-Merge than with RA"*
- ↳ H<sub>3</sub>: *"Subjects prefer using Stu-Merge to RA"*

### → Results

- ↳ H<sub>1</sub> *accepted (strong evidence)*
- ↳ H<sub>2</sub> & H<sub>3</sub> *rejected*
- ↳ *Subjects found the tool unintuitive*



## Threats to Validity

### → Construct Validity

- ↳ *What do we mean by a merge? What is correctness?*
- ↳ *5-point scale for subjective assessment - insufficient discriminatory power*
  - > *(both tools scored very low)*

### → Internal Validity

- ↳ *Confounding variables: Time taken to learn the tool; familiarity*
  - > *Subjects were all familiar with RA, not with Stu-merge*

### → External Validity

- ↳ *Task representativeness*
  - > *class models were of a toy problem*
- ↳ *Subject representativeness*
  - > *Grad students as sample of what population?*

### → Theoretical Reliability

- ↳ *Researcher bias*
  - > *subjects knew Stu-merge was Stu's own tool*





## What went wrong?

- What was the research question?
  - ↳ “Is tool A better than tool B?”
- What would count as an answer?
- What use would the answer be?
  - ↳ How is it a “contribution to knowledge”?
- How does this evaluation relate to the existing literature?



## Experiments as Clinical Trials

Why would we expect it to be better?

Why do we need to know?

What will we do with the answer?

Is drug A better than drug B?

Better at doing what?

Better in what way?

Better in what situations?





Why would we expect it to be better?

You gotta have a theory!



## Some Definitions

- A **model** is an abstract representation of a phenomenon or set of related phenomena
  - ↳ Some details included, others excluded
- A **theory** is a set of statements that explain a set of phenomena
  - ↳ Serves to explain and predict
  - ↳ Precisely defined terminology
  - ↳ Concepts, relationships, causal inferences
  - ↳ (operational definitions for theoretical terms)
- A **hypothesis** is a testable statement derived from a theory
  - ↳ A hypothesis is not a theory!
- In SE, we have mostly *folk theories*





## A simpler definition

A **Theory** is the best explanation of all the available evidence



## The Role of Theory Building

- **Theories lie at the heart of what it means to do science.**
  - ↳ Production of generalizable knowledge
- **Theory provides orientation for data collection**
  - ↳ Cannot observe the world without a theoretical perspective
- **Theories allow us to compare similar work**
  - ↳ Theories include precise definition for the key terms
  - ↳ Theories provide a rationale for which phenomena to measure
- **Theories support analytical generalization**
  - ↳ Provide a deeper understanding of our empirical results
  - ↳ ...and hence how they apply more generally
  - ↳ Much more powerful than statistical generalization





# Stu's Theory

## → Background Assumptions

- ↳ Large team projects, models contributed by many actors
- ↳ Models are fragmentary, capture partial views
- ↳ Partial views are inconsistent and incomplete most of the time

## → Basic Theory

- ↳ (Brief summary:)
- ↳ Model merging is an exploratory process, in which the aim is to discover intended relationships between views. 'Goodness' of a merge is a subjective judgment. If an attempted merge doesn't seem 'good', many need to change either the models, or the way in which they were mapped together.
- ↳ [Still needs some work]

## → Derived Hypotheses

- ↳ Useful merge tools need to represent relationships explicitly
- ↳ Useful merge tools need to be complete (work for any models, even if inconsistent)



# What type of question are you asking?

## → Existence:

- ↳ Does X exist?

## → Description & Classification

- ↳ What is X like?
- ↳ What are its properties?
- ↳ How can it be categorized?
- ↳ How can we measure it?
- ↳ What are its components?

## → Descriptive-Comparative

- ↳ How does X differ from Y?

## → Frequency and Distribution

- ↳ How often does X occur?
- ↳ What is an average amount of X?

## → Descriptive-Process

- ↳ How does X normally work?
- ↳ By what process does X happen?
- ↳ What are the steps as X evolves?

## → Relationship

- ↳ Are X and Y related?
- ↳ Do occurrences of X correlate with occurrences of Y?

## → Causality

- ↳ Does X cause Y?
- ↳ Does X prevent Y?
- ↳ What causes X?
- ↳ What effect does X have on Y?

## → Causality-Comparative

- ↳ Does X cause more Y than does Z?
- ↳ Is X better at preventing Y than is Z?
- ↳ Does X cause more Y than does Z under one condition but not others?

## → Design

- ↳ What is an effective way to achieve X?
- ↳ How can we improve X?





# What type of question are you asking?

## → Existence:

↳ Does X exist?

## → Description & Classification

- ↳ What is X like?
- ↳ What are its properties?
- ↳ How can it be categorized?
- ↳ How can we measure it?
- ↳ What are its components?

## → Descriptive-Comparative

↳ How does X differ from Y?

## → Frequency and Distribution

- ↳ How often does X occur?
- ↳ What is an average amount of X?

## → Descriptive Process

- ↳ How does X normally work?
- ↳ By what process does X happen?
- ↳ What are the steps as X evolves?

## → Relationship

- ↳ Are X and Y related?
- ↳ Do occurrences of X relate with occurrences of Y?

## → Causality

- ↳ Does X cause Y?
- ↳ Does X prevent Y?
- ↳ What causes X?
- ↳ What effect does Y have on Y?

## → Causality-Comparative

- ↳ Does X cause more Y than does Z?
- ↳ Is X better at preventing Y than is Z?
- ↳ Does X cause more Y than does Z under one condition but not other?

## → Design

- ↳ What is an effective way to achieve X?
- ↳ How can we improve X?



# Stu's Research Question(s)

## → Existence

↳ Does model merging ever happen in practice?

## → Description/Classification

↳ What are the different types of model merging that occur in practice on large scale systems?

## → Descriptive-Comparative

↳ How does model merging with explicit representation of relationships differ from model merging without such representation?

## → Causality

↳ Does an explicit representation of the relationship between models cause developers to explore different ways of merging models?

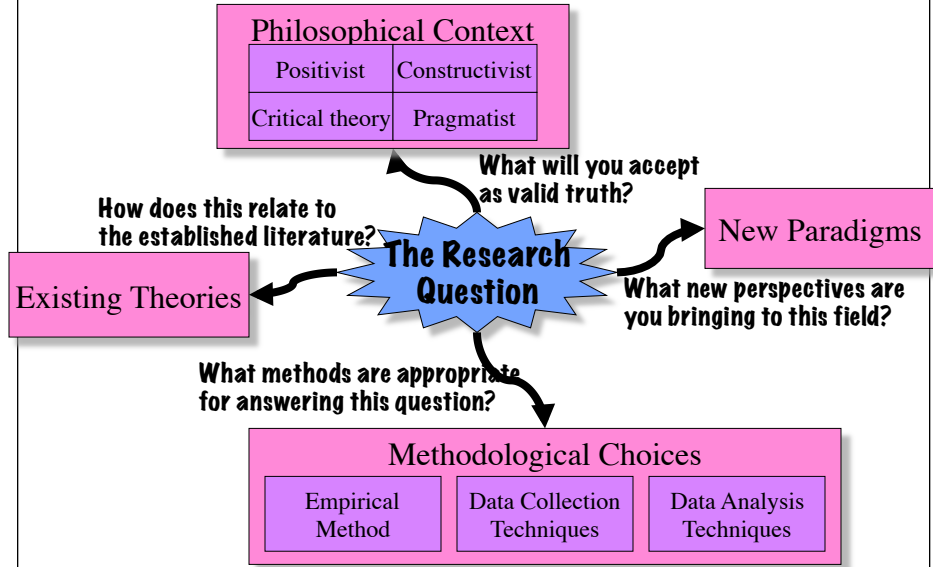
## → Causality-Comparative

↳ Does the algebraic representation of relationships in Stu's tool lead developers to explore more than do pointcuts in AOM?





# Putting the Question in Context

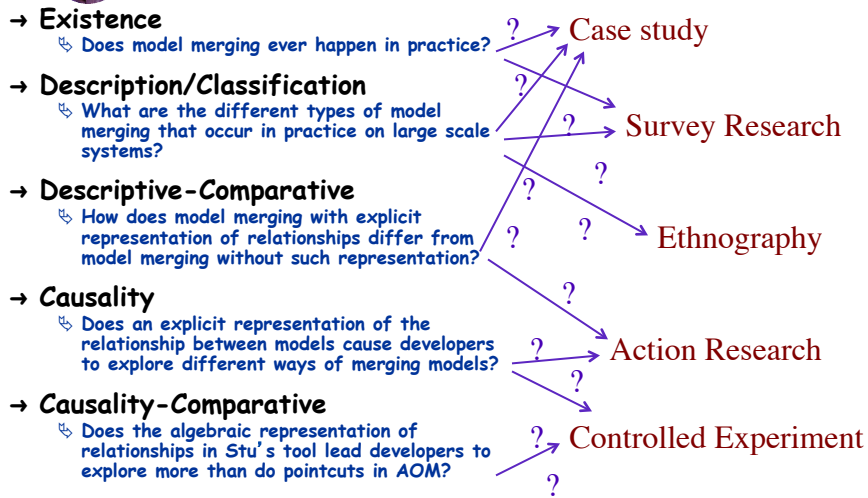


# Many available methods...

- | Common "in the lab" Methods  | Common "in the wild" Methods   |
|--|--|
| <ul style="list-style-type: none"> <li>● Controlled Experiments</li> <li>→ Rational Reconstructions</li> <li>→ Exemplars</li> <li>→ Benchmarks</li> <li>→ Simulations</li> </ul> | <ul style="list-style-type: none"> <li>● Quasi-Experiments</li> <li>● Case Studies</li> <li>● Survey Research</li> <li>● Ethnographies</li> <li>● Action Research</li> </ul> |
- Artifact/Archive Analysis ("mining"!) (located below the table)



# Stu's Method(s) Selection...



## Warning

**No method is perfect**

**Don't get hung up on methodological purity**

**Pick something and get on with it**

**Some knowledge is better than none**



Okay, but...



## Why Build a Tool?

- **Build a Tool to Test a Theory**
  - ↳ Tool is part of the experimental materials needed to conduct your study
- **Build a Tool to Develop a Theory**
  - ↳ Theory emerges as you explore the tool
- **Build a Tool to Explain your Theory**
  - ↳ Theory as a concrete instantiation of (some aspect of) the theory

