

Where is (and should) Empirical Software Engineering (be) going? (or Empirical CS/AI/ML!?)

JISBD/SISTEDES 2022, Santiago de Compostela, Spain, September 5th 2022

Robert Feldt

Division of Software Engineering , Dept of Computer Science and Engineering
Chalmers University of Technology

robert.feldt@chalmers.se



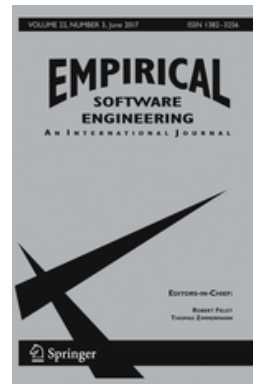
@drfeldt





CHALMERS
UNIVERSITY OF TECHNOLOGY

About me



**Founded in 1996,
recently celebrated 25th anniversary!**

Most SE papers nowadays are empirical!

But how has ESE evolved?

- Over the last ~25 years (1996-2021):
 - 1. Have the **topics** changed?
 - 2. Have the **research methods** changed?
 - 3. Have the **empirical basis** changed?
 - i.e. has the amount of and quality of empirical evidence provided changed/improved?

Relevant also for CS and ML/AI!



Please note our measures
concerning Coronavirus / Covid 19

SCHLOSS DAGSTUHL
Leibniz-Zentrum für Informatik

[About Dagstuhl](#)[Program](#)[Publications](#)

You are here: [Program](#) » [Seminar Calendar](#) » Seminar Homepage

<https://www.dagstuhl.de/22442>

October 30 – November 4 , 2022, Dagstuhl Seminar 22442

Toward Scientific Evidence Standards in Empirical Computer Science

Organizers

Brett A. Becker (University College Dublin, IE)
Christopher D. Hundhausen (Oregon State University – Corvallis,
Ciera Jaspan (Google – Mountain View, US)
Andreas Stefik (University of Nevada – Las Vegas, US)
Thomas Zimmermann (Microsoft Corporation – Redmond, US)

Seminar Calendar
All Events
Dagstuhl Seminars
Dagstuhl Perspectives
GI-Dagstuhl Seminars
Summer Schools
Events
Research Guests
Expenses
Planning your visit

nature

[Explore content](#) [About the journal](#) [Publish with us](#) [Subscribe](#)

[nature](#) > [news](#) > article

NEWS | 26 July 2022

Could machine learning fuel a reproducibility crisis in science?

‘Data leakage’ threatens the reliability of machine-learning use across disciplines, researchers warn.

[Elizabeth Gibney](#)

MIT
Technology
Review

[Featured](#) [Topics](#) [Newsletters](#) [Events](#) [Podcasts](#)

[Sign in](#)

ARTIFICIAL INTELLIGENCE

AI is wrestling with a replication crisis

Tech giants dominate research but the line between real breakthrough and product showcase can be fuzzy. Some scientists have had enough.

Methodology

- **Main idea:**
 - Use novel DNN Language models for text embedding to study abstract similarity
 - Cluster papers together if close in embedded space => topics/groups
 - Match keywords and abstract for common research methods
 - Sample some old and some new papers per research method & check “empirical basis”
- Extracted meta-data of EMSE (journal) paper data for 1996-2021
- Included every second year of ESEM (conference) up to 2019
 - IEEE Explore is not complete & ACM DL harder to extract data from
- Filtered out short papers (≤ 6 pages) as well as non-research papers (editorials etc)
- Made some effort to **exclude systematic literature** reviews but this was inexact
 - Focus is on primary research

Disclaimers

- **Springer Link data is not perfect (especially for older papers)**
 - Not uncommon that there are **merged words and other data problems**
 - I don't think this should have major impact (embedding methods are on low-level, a few characters, so very few n-grams around problem are incorrect) but I haven't verified it

[Published: December 2000](#)

An Instrument for Measuring the Key Factors of Success in Software Process Improvement

[Tore Dyba](#)

[Empirical Software Engineering](#) 5, 357–390 (2000) | [Cite this article](#)

846 Accesses | 105 Citations | [Metrics](#)

Abstract

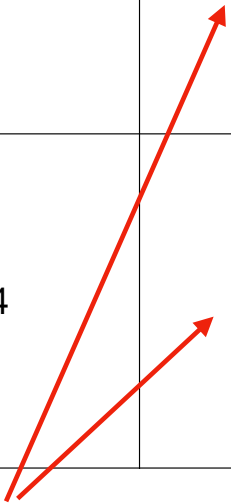
Understanding how to implement SPI successfully is arguably the most challenging issue facing the SPI field today. The SPI literature contains many case studies of successful companies and descriptions of their SPI programs. However, there has been no systematic attempt to synthesize and organize the prescriptions offered. The research efforts to date are limited and inconclusive and without adequate theoretical and psychometric justification.

More disclaimers

- **Springer Links data is not perfect (especially for older papers)**
 - It happens that **later paragraphs in the abstract has not been properly added** to meta-data
 - This sometimes happened for older papers (EMSEJ was not a Springer journal when started and they have partly imported data at some point) but very rarely for newer ones
- Overall, this analysis is in now way “perfect” and **can only be used as a rough indication** of the main topics; beware! In particular, the analysis of research methods and empirical basis is a very initial/early one.

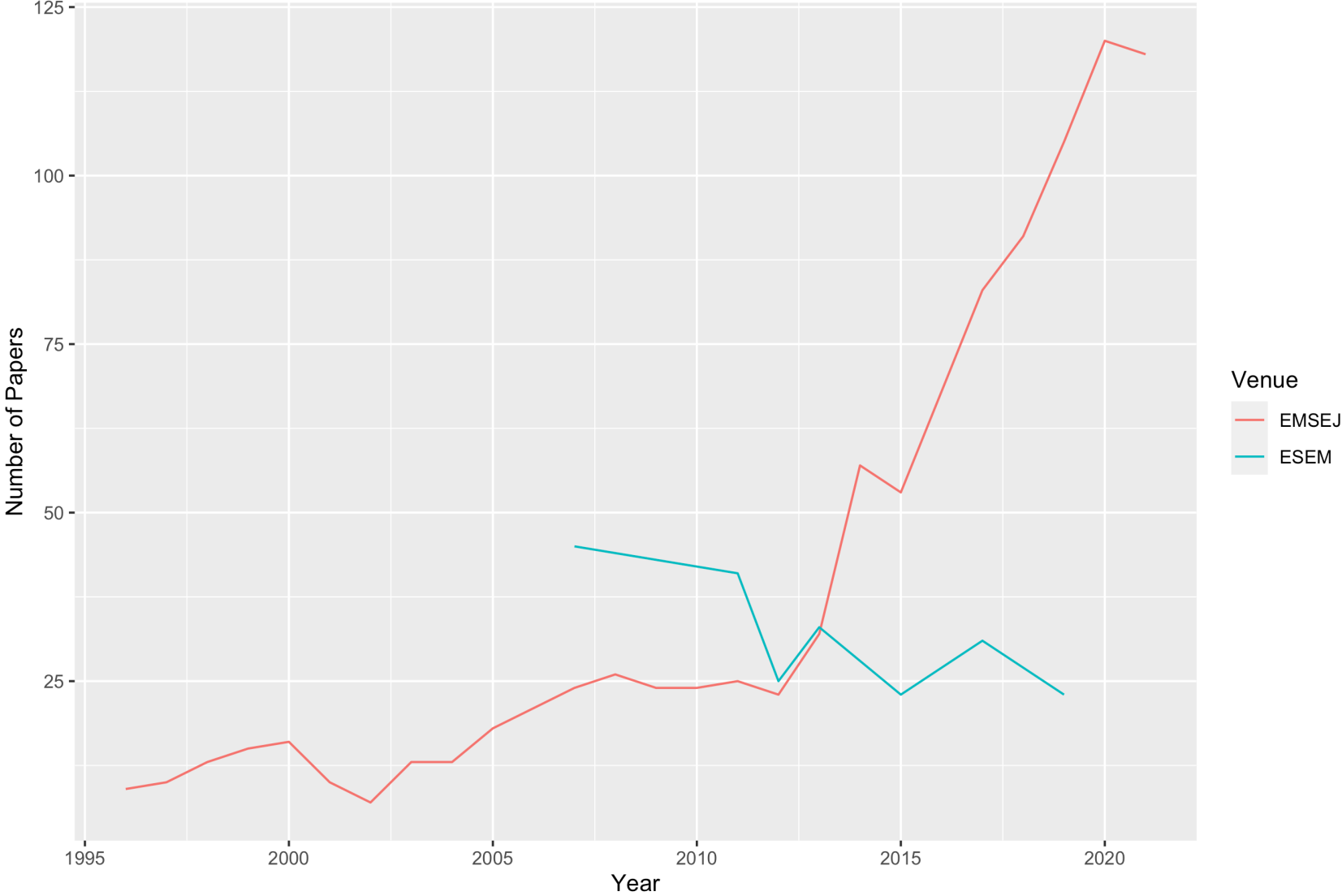
Overview of included papers

Venue	Years	Num papers	Median # pages	Mean # pages	InterQuartileRange # pages
EMSE Journal	1996-2021	1017	35	36	29-42
ESEM Conference	2007, 2009, 2011, 2012, 2013, 2015, 2017, 2019	264	10	10	10-10



BUT please note: Cannot directly compare number of pages, since formats differ a lot. In number of words, journal papers tend to be ~40-110% bigger than conference papers IMHO.

Number of papers (included) per year



Topic modeling

- BERTopic library
 - BERT
 - all-mpnet-base-v2 LLM
 - 768 floats in embedding vectors
 - UMAP for plotting in 2D
 - HDBScan for clustering (no need to select number of clusters)
 - Downside is a rather large “REST” topic with the documents without a cluster
- When applied to our sample:
 - Finds 35 clusters (topics) and 1 “REST” cluster with 286 (22%) papers with no topic assigned

BERTopic

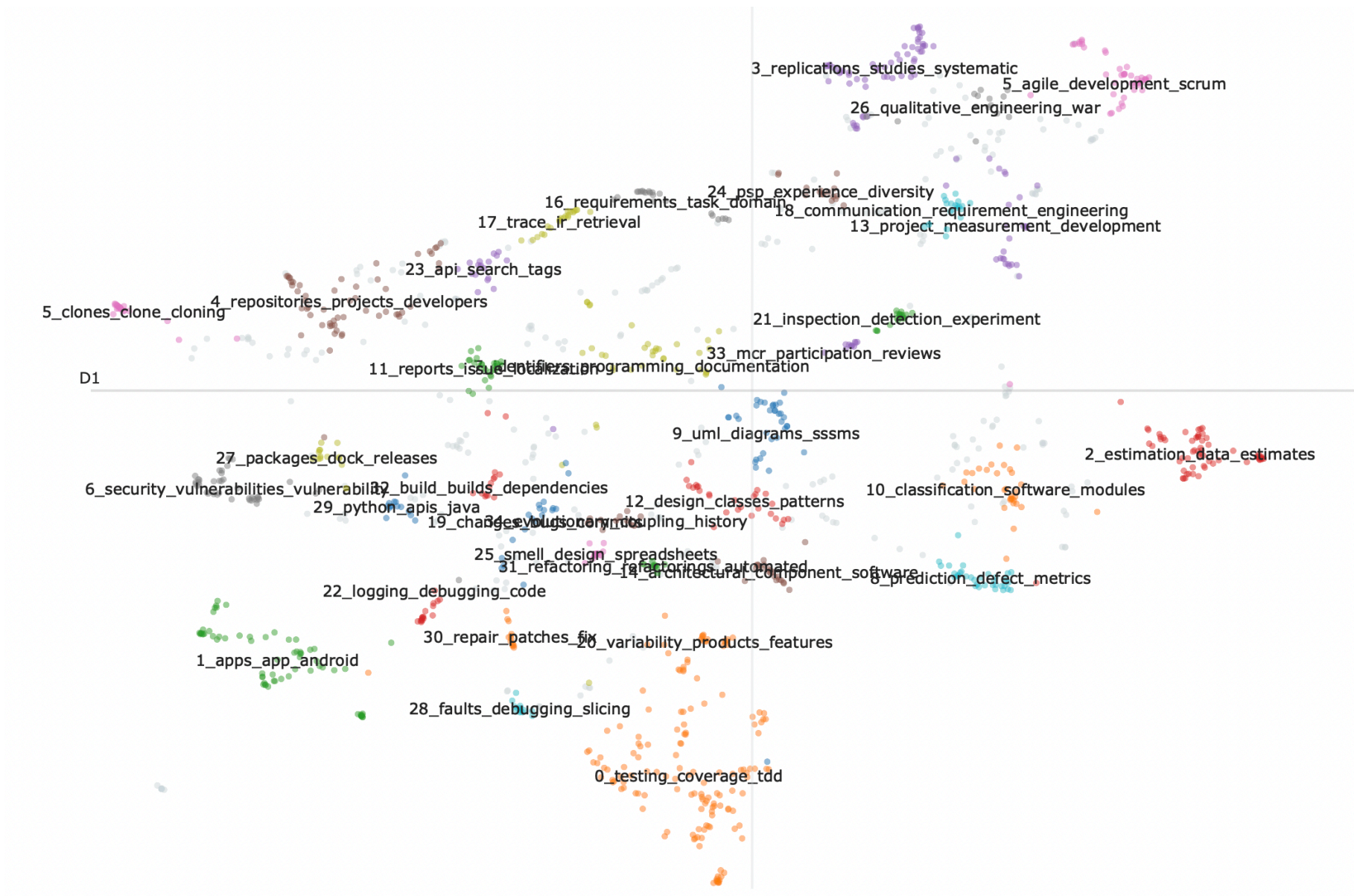
BERTopic is a topic modeling technique that leverages 🧠 transformers and c-TF-IDF to create dense clusters allowing for easily interpretable topics whilst keeping important words in the topic descriptions.

BERTopic supports **guided**, (semi-) **supervised**, **hierarchical**, and **dynamic** topic modeling. It even supports visualizations similar to LDAvis!

Corresponding medium posts can be found [here](#) and [here](#). For a more detailed overview, you can read the [paper](#).



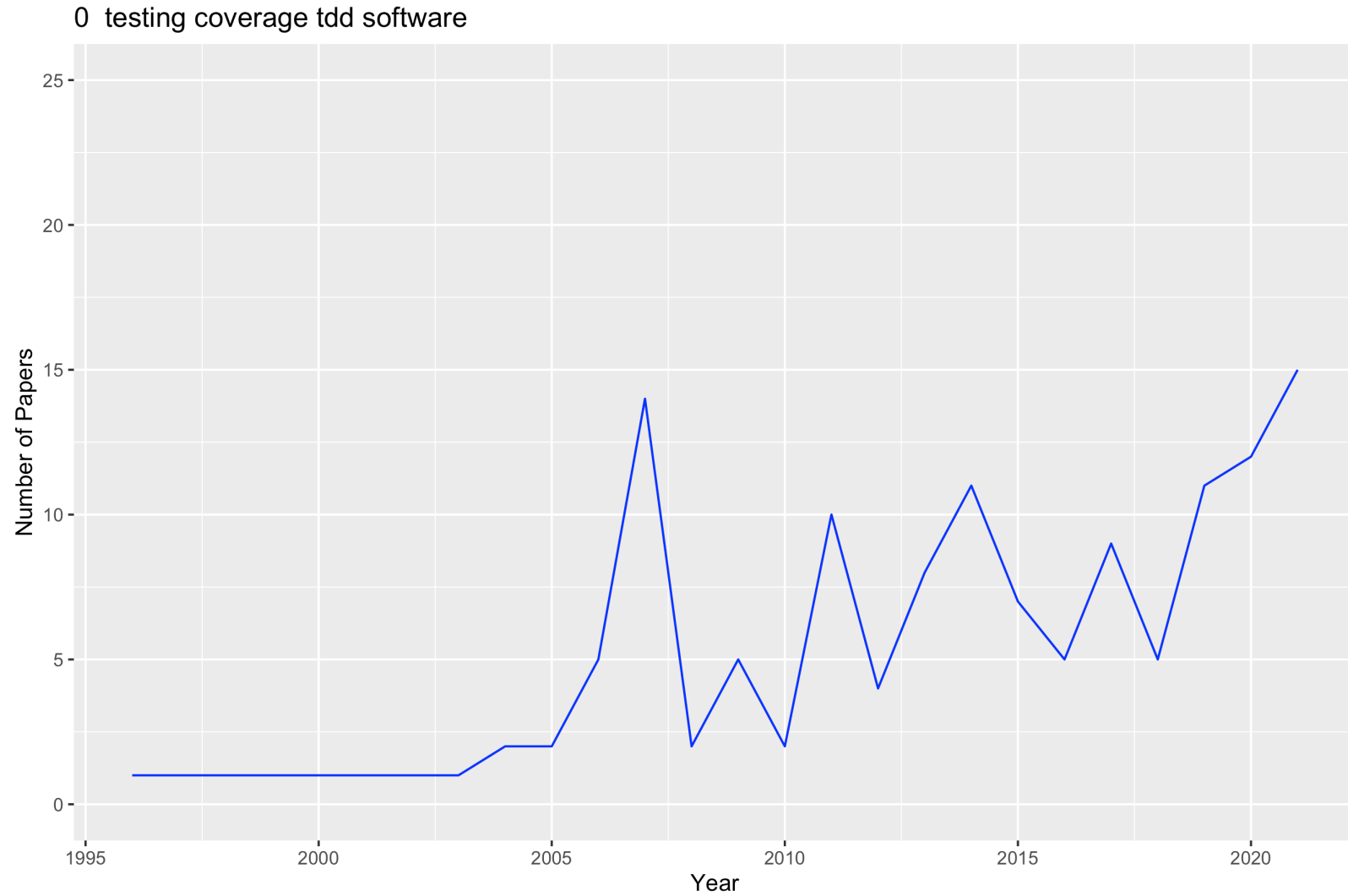
BERTopic



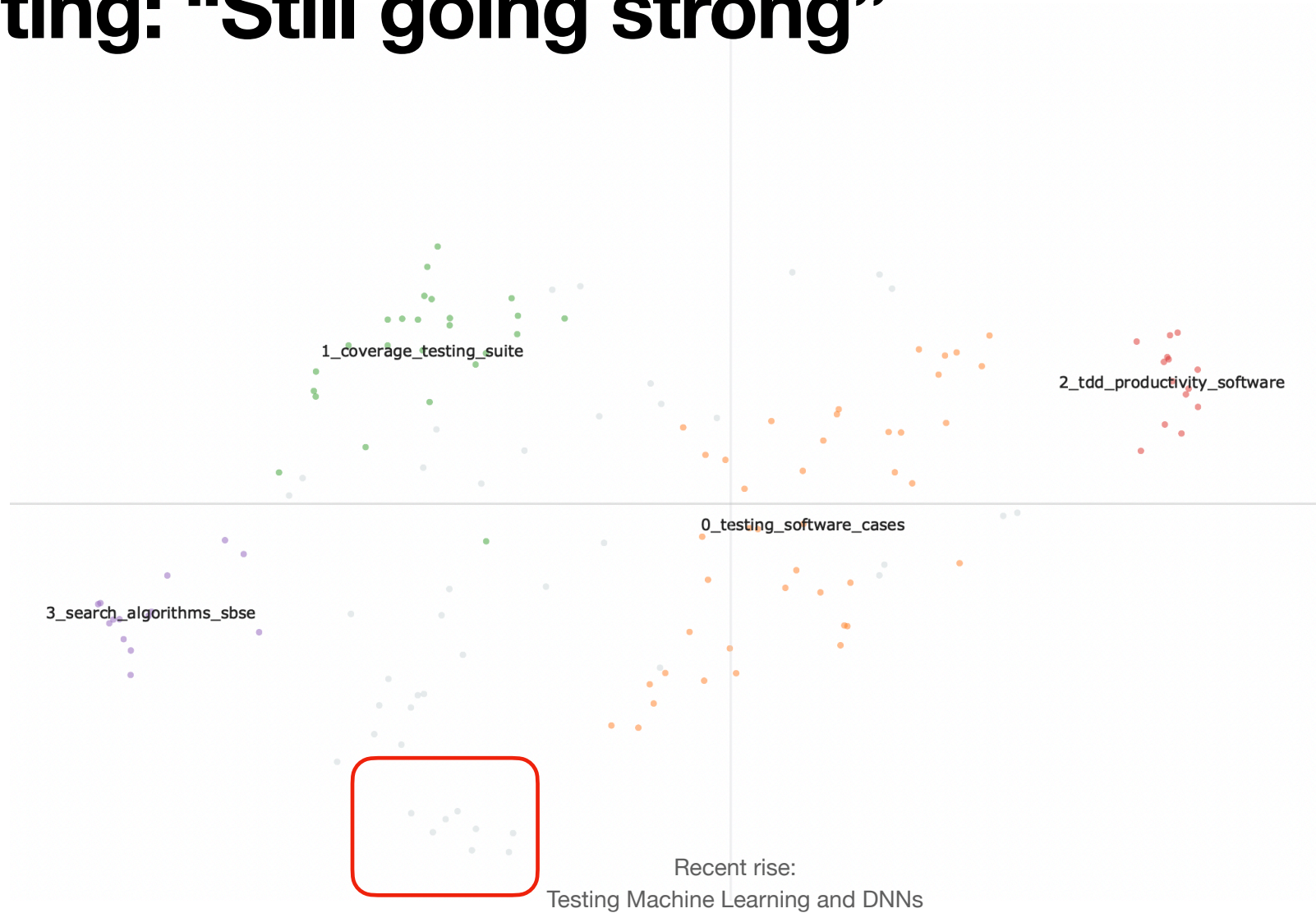
Top-20 Topics, 1996-2021

Topic String	NumPapers Int64	FirstYr Int64	MeanYr Float64	NumLast5 Int64	PctLast5 Float64
testing coverage tdd software	132	1996	2013.9	52	39.4
apps app android reviews	65	2012	2018.6	57	87.7
estimation data estimates cost	62	1996	2008.8	7	11.3
replications studies systematic...	60	1999	2012.4	19	31.7
repositories projects developer...	54	2009	2017.3	36	66.7
agile development scrum teams	44	2004	2013.4	12	27.3
security vulnerabilities vulner...	39	2007	2015.3	19	48.7
identifiers programming documen...	35	2007	2015.6	15	42.9
prediction defect metrics cpdp	34	2007	2015.6	18	52.9
uml diagrams sssms design	34	1997	2010.9	8	23.5
classification software modules...	30	1996	2007.5	3	10.0
reports issue localization rele...	26	2013	2018.0	17	65.4
design classes patterns metrics	26	1996	2008.0	2	7.7
project measurement development...	25	1998	2009.8	3	12.0
architectural component softwar...	22	1999	2011.8	7	31.8
clones clone cloning plagiarism	22	2008	2015.5	12	54.5
requirements task domain method...	20	1997	2013.1	10	50.0
trace ir retrieval eye	20	2008	2014.6	7	35.0
communication requirement engin...	19	1996	2014.3	9	47.4
changes bugs commits files	18	2004	2014.9	8	44.4

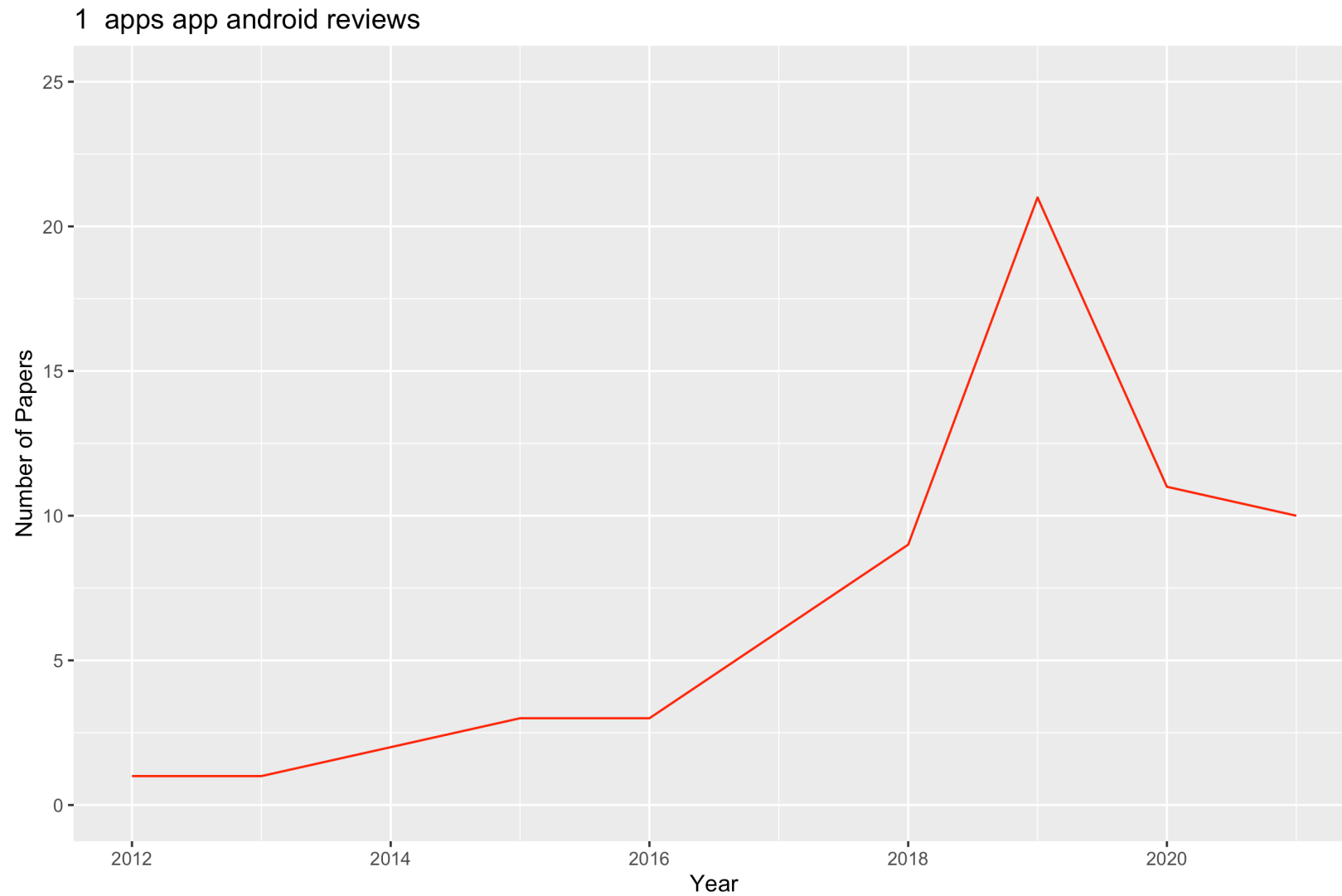
Testing: “Still going strong”



Testing: “Still going strong”

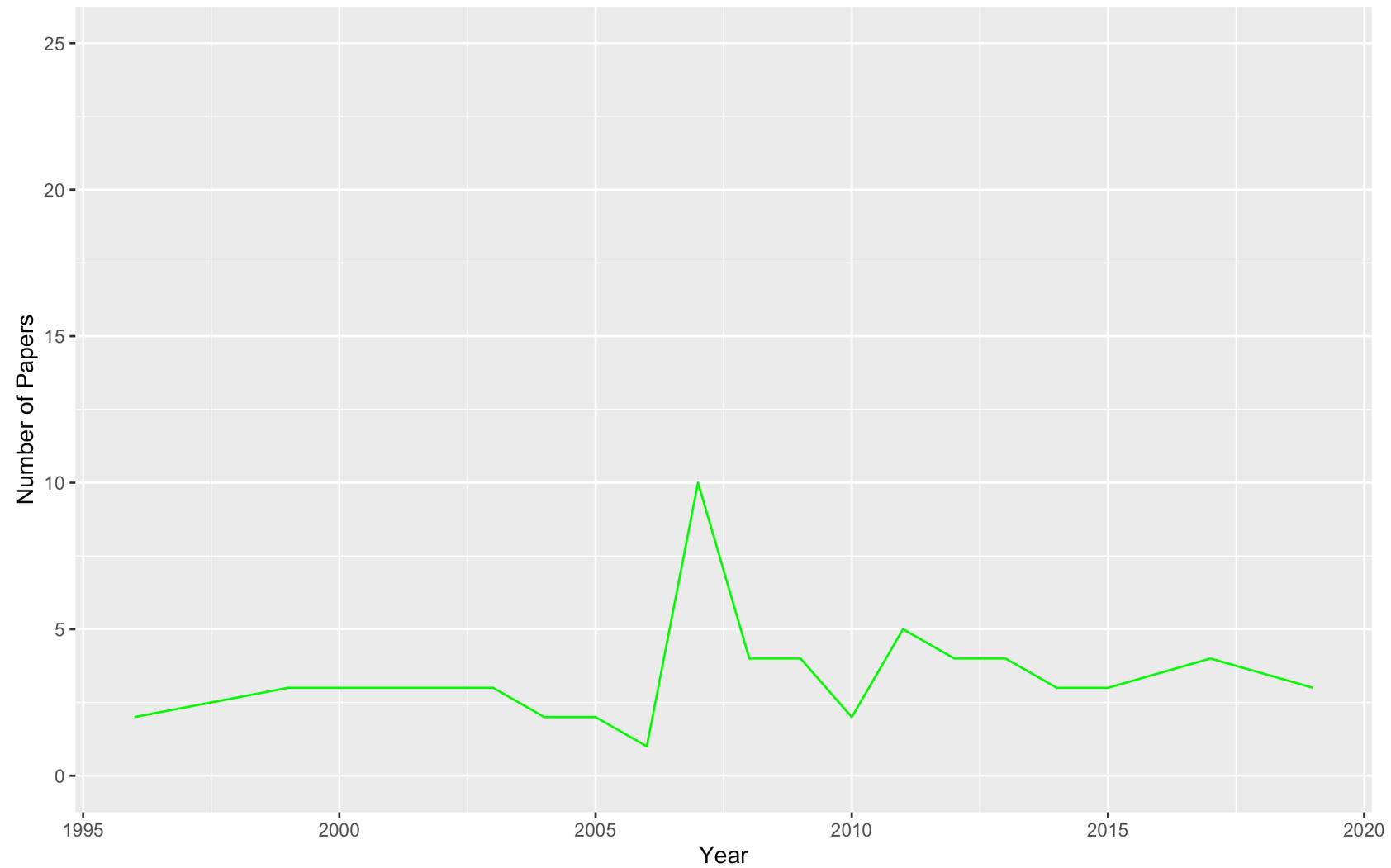


App analysis: “Explosive growth, but now what?”

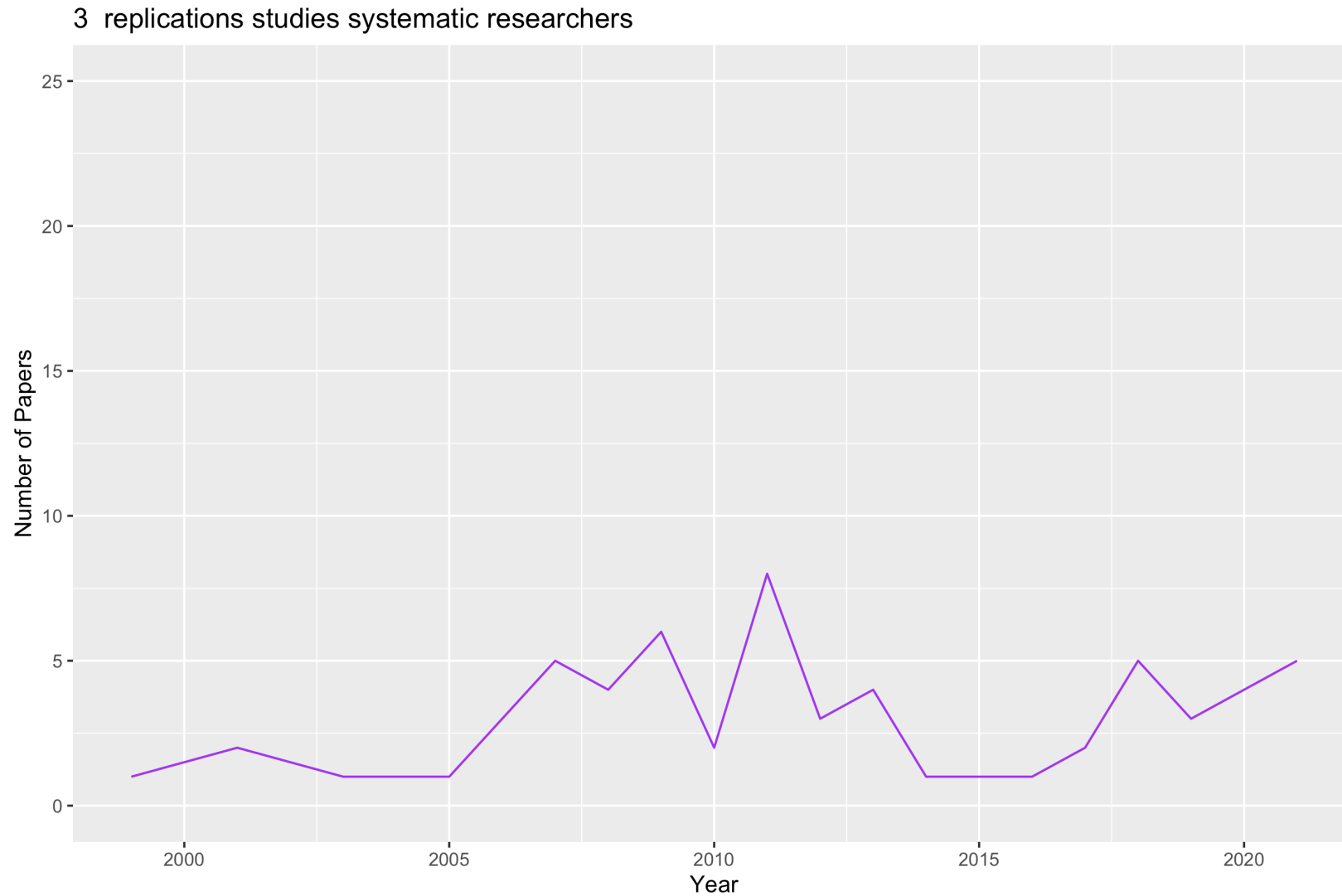


Cost estimation: “Decline or Slow but steady!?”

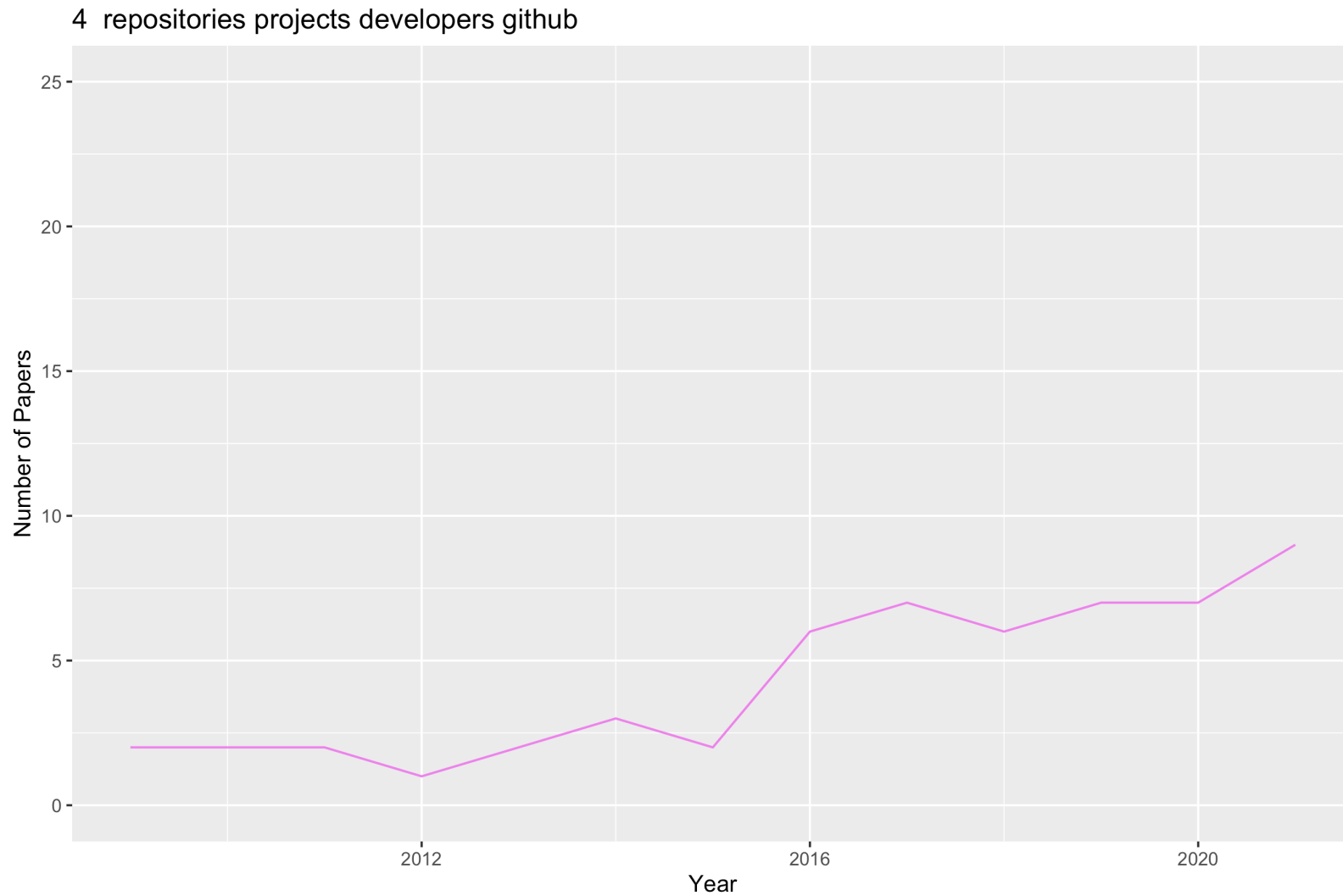
2 estimation data estimates cost



Replications & secondary: “Rising again?”

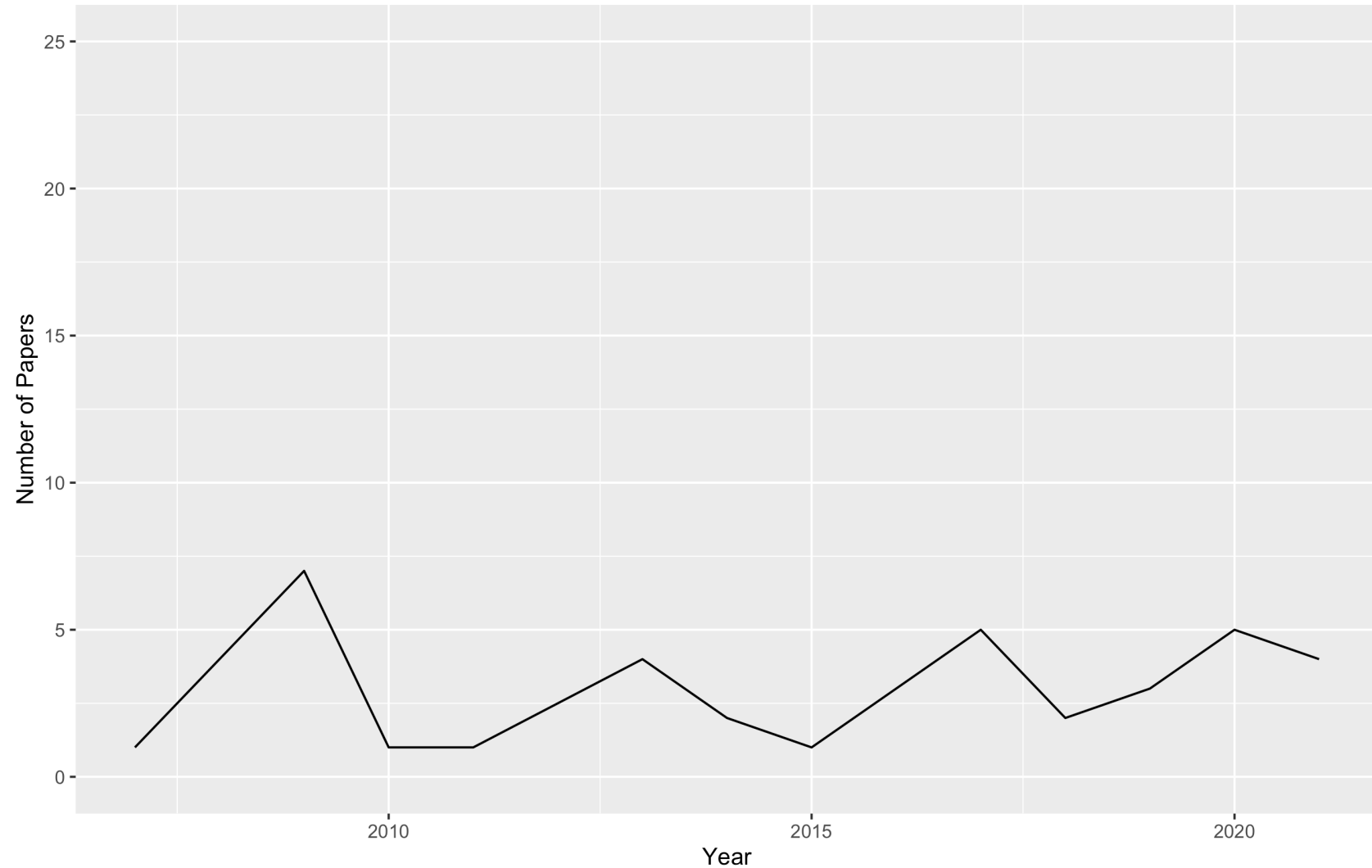


Repository mining: “Clearly rising”



Security / vulnerabilities: “Never really taking off?”

6 security vulnerabilities vulnerability privacy



Can we see some Empirical SE topic trends?

- Topic modeling can help find interesting patterns!
- But many topics => hard to analyse trends over time, some patterns:
 - **Strong showing for long & no signs of slowing down:**
 - Testing, Replications & secondary studies
 - **Recent rising “stars”:**
 - App analysis, Repository mining, Issue/report localization
 - **Classic but possibly slowing down:**
 - Cost estimation, Design/classes and metrics, Project measurement
- Many papers in ESE are in-between topics or hard to classify

Which research methods are used the most?

Method String	NumStudies Int64	FirstYr Int64	MeanYr Float64	NumLast5 Int64	PctLast5 Float64
Case study	220	1996	2013.0	78	35.5
Controlled experiment	101	1996	2010.8	22	21.8
Interview study	89	2005	2015.9	51	57.3
Review	68	2000	2014.4	24	35.3
Questionnaire	33	1998	2012.2	12	36.4
Action research	9	2002	2013.8	3	33.3
User Study	13	2013	2017.2	8	61.5

- Overall hard to judge from abstract; need more advanced method (NLP!?) for the future!
 - In particular for older papers (more variation, less well defined/named)
 - In particular for conference papers (often say just “empirical study”)
- Interview studies and User Studies on the rise
- **Main problem** is that for most papers **we couldn’t judge the method used** (with our (admittedly too simple) analysis). But was often hard to do even if done manually!

Which methods are used together?

Method1	Method2	NumStudies
Case study	Interview study	26
Case study	Review	10
Interview study	Questionnaire	9
Controlled experiment	Questionnaire	5
Case study	Questionnaire	4
Case study	Action research	4
Controlled experiment	Review	4
Controlled experiment	User Study	3
Interview study	Review	3
Controlled experiment	Interview study	2
Interview study	User Study	2
Review	Questionnaire	2
Case study	Controlled experiment	1

**Overall, few multi-method papers
(or not clear from abstract)!**

Differences in “empirical basis”?

- Empirical basis = “amount” & “quality”/relevance of empirical evidence/data (in absolute terms)
- “Value” of evidence is typically relative to novelty/maturity of topic:
 - Relative value of (same) absolute amount of information decreases over time
 - “Yet another study showing that code size is a good proxy for X”
 - But first time this is shown the scientific value can be very high!
 - Can be contrasted with:
 - “Some (new) information is better than no (new) information!”
 - Novelty not a key aspect if research is sound

Lets look at “empirical basis” of interview studies

- Study 1 from **2005**:
 - 4 companies for theory/method formation/development:
 - One interviewee at each company
 - Discourse analysis of interviews and ethnographic data
 - 1 validation company:
 - Observed and interviewed developers for a week
 - “interviewed numerous people”
 - No clear discussion of analysis method, quotes from interviews, interleaved in an interpretation / narrative of the authors
- **Summary:**
 - Interviewees not clarified in a table, nor described in common form
 - Even their number is not clear, 4+4?

“empirical basis” of interview studies #2

- 2013 study
 - Multiple data sources: tool evaluations, interviews, a survey
 - 12 interviews in 4 companies, mapping to companies or roles unclear
 - 6 survey responses from 2 companies, rest answered “in groups”
- **Summary:**
 - Interviewees not clarified in a table, just described overall
 - Roles and mapping to companies unclear

“empirical basis” of interview studies #3

- 2021 study
 - Multiple data sources: experiment, survey, semi-struct interviews
 - 40 participants from 12 companies
 - Full replication package with (quantitative) data + analysis scripts
 - Study design clarified in diagram (pointers to detailed tables)
 - Thematic analysis for qualitative data, Bayesian analysis for quant.
- **Summary:**
 - However, not clarified if all participants also interviewed or only some
 - Length of interviews unclear (post-task so presumably short)
 - Roles and mapping to companies unclear

2 proposals & 2 warnings

- Increase methodological clarity!
 - Both in thinking, designing and reporting of research
 - Pre-registration can help, fosters early clarity!
 - JSON format for methodological description & empirical basis!?
- Causal graphs / modeling
 - Clearly suited to experiments & confirmatory research
 - But also for qualitative studies (see IS) and tools/tech
- Risk of uniformity of study designs
 - Reduced creativity, new methodologies, new combinations
- Focusing only on quantity of empirical basis
 - High quantity of relevant, high-quality data is key

(Pre-)Registered Reports helps clarify methods, data, analysis

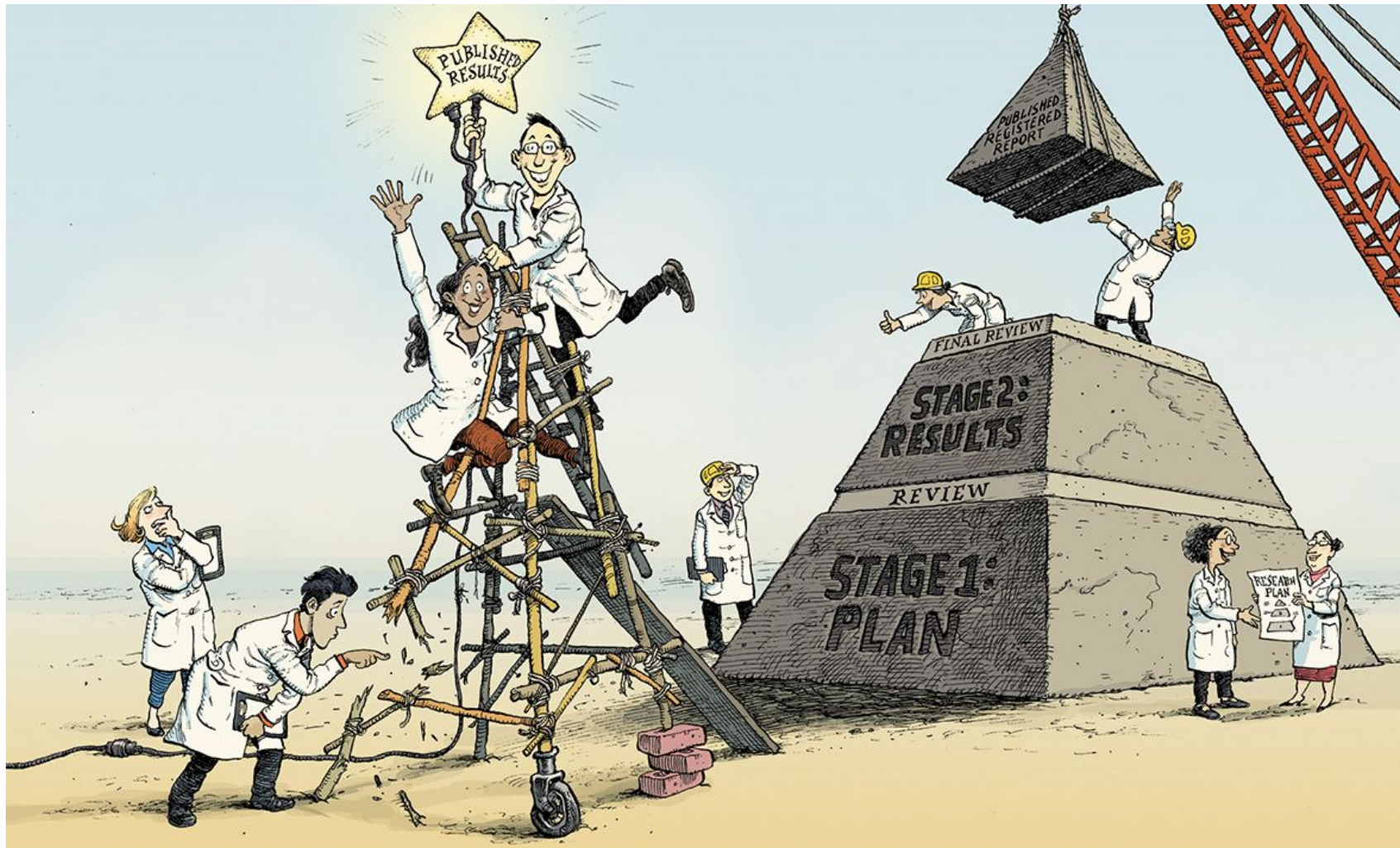
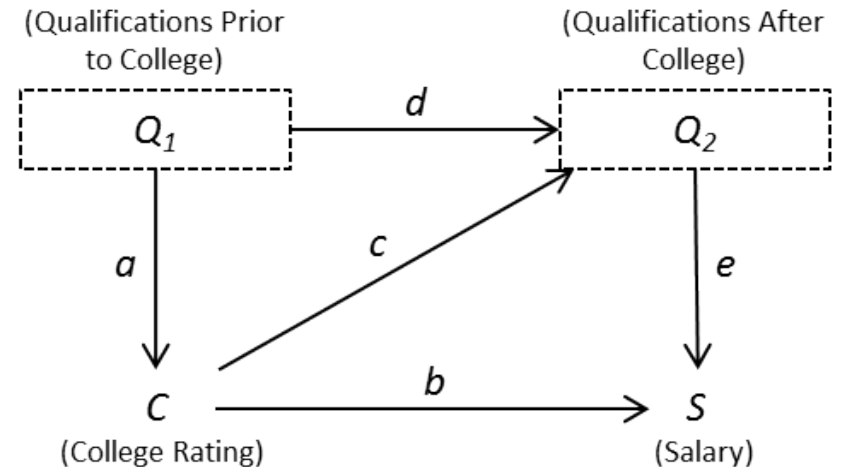
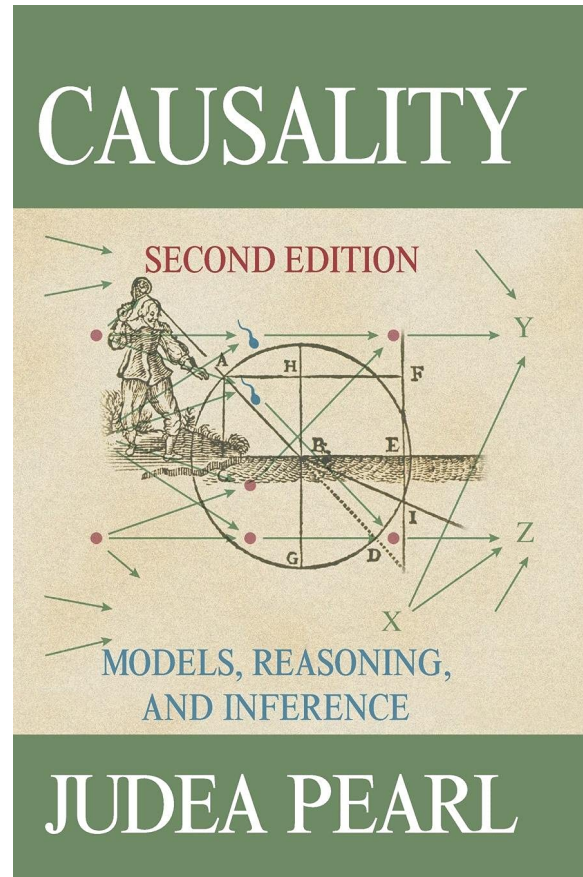
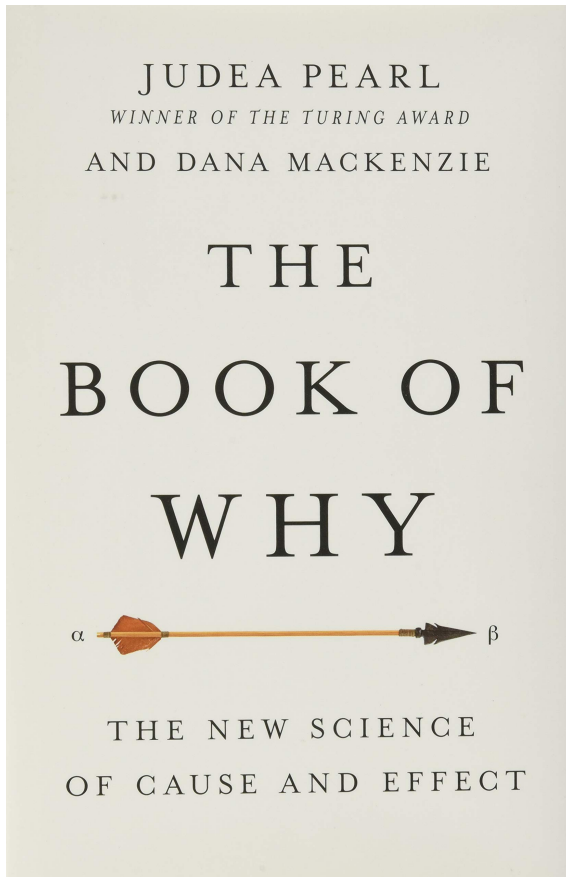


Illustration by David Parkins in Nature, September 2019

research_method.json - meta-data to clarify methods & empirical basis

```
1  {
2    "Title": "Measuring affective states from technical debt",
3    "Authors": ["Jesper Olsson", "Erik Risfelt", "Terese Besker",
4               "Antonio Martini", "Richard Torkar"],
5    "Estimand": {
6      "Overall": "Affective states",
7      "Operationalisations": ["positive/negative emotions"],
8    },
9    "Methods": [
10     {
11       "Method": "Experiment, repeated-measures",
12       "N": 40,
13       "Repetitions": 5},
14     {
15       "Method": "Survey",
16       "Instrument": "Self-Assessment Manikin (SAM)",
17       "N": 40},
18     {
19       "Method": "Interviews",
20       "Sub-method": "semi-structured",
21       "Interview length": "30-40 minutes, directly after experiment session",
22       "N": "possibly 40, unclear",
23       "Data": "transcripts missing",
24       "Analysis": "unspecified"},
25   ]
26 }
```

Causal Modeling with graphs (à la Pearl)



Causal Graphs also for Qualitative & Exploratory?!

MIS
Quarterly

RESEARCH ARTICLE

ATTAINING INDIVIDUAL CREATIVITY AND PERFORMANCE IN MULTIDISCIPLINARY AND GEOGRAPHICALLY DISTRIBUTED IT PROJECT TEAMS: THE ROLE OF TRANSACTIVE MEMORY SYSTEMS¹

Wei He

Area of Information Systems and Quantitative Sciences, Rawls College of Business
Texas Tech University, Lubbock, TX, U.S.A. {wei.he@ttu.edu}

J. J. Po-An Hsieh

Department of Computer Information Systems, Robinson College of Business
Georgia State University, Atlanta, GA, U.S.A. {jjhsieh@gsu.edu}

Andreas Schroeder

Operations & Information Management Group, Aston Business School
Aston University, Birmingham, U.K. {a.schroeder@aston.ac.uk}

Causal Graphs also for Qualitative!

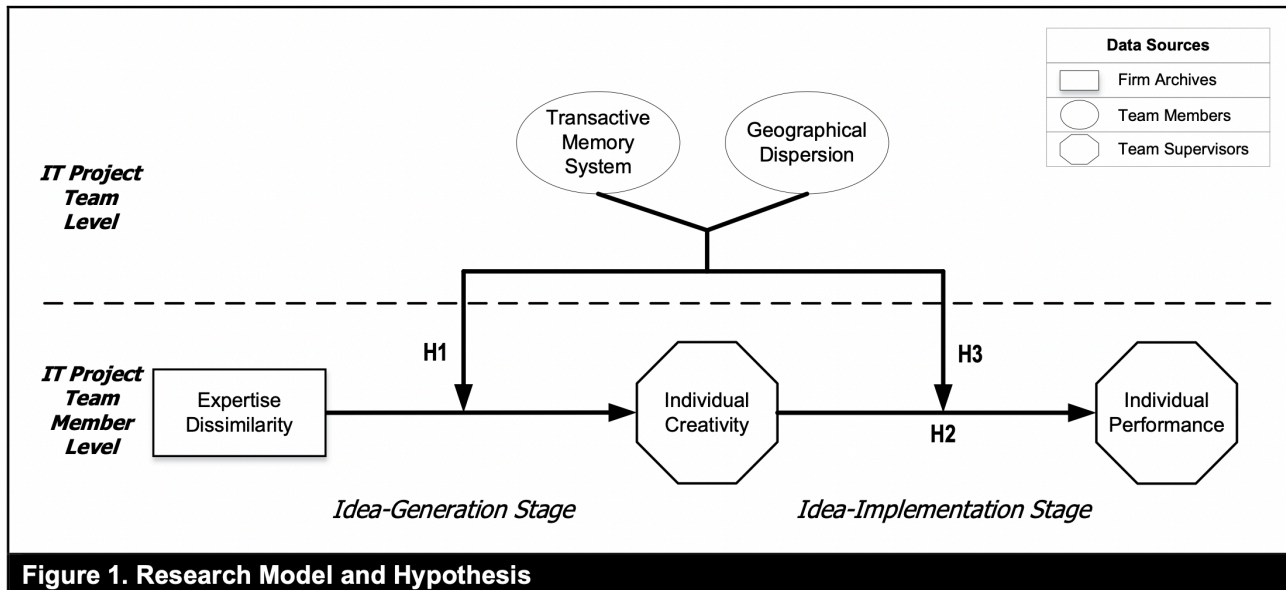
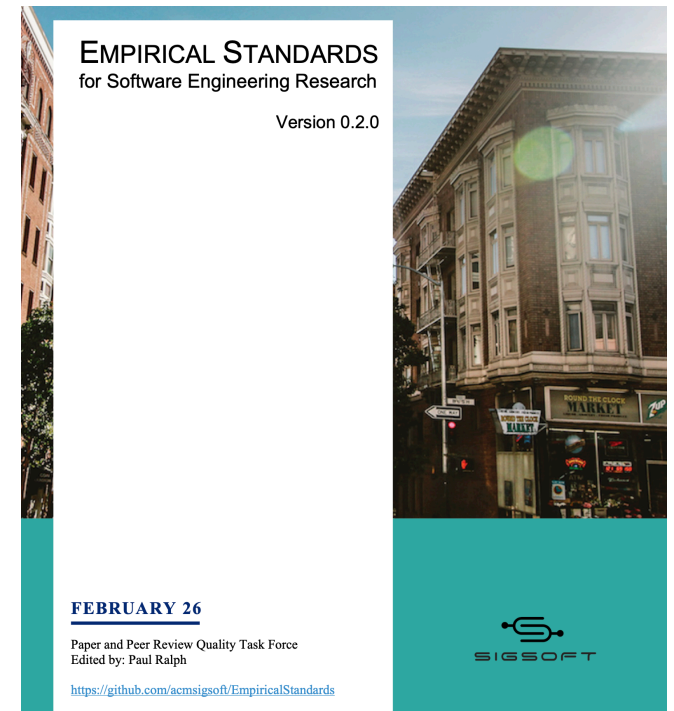


Table 2: Construct Definitions

Construct name	Level	Definition	Supporting references
Expertise dissimilarity (ExpDiss)	Individual	The difference in expertise between a focal team member and his or her fellow team members.	Harrison & Klein, 2007; Huang et al. 2014; van der Vegt et al., 2003
Individual creativity	Individual	The generation of new and useful ideas by an individual team member.	Zhou & George, 2001
Individual performance	Individual	The actions specified and required by an individual team member's job description.	Borman & Motowidlo, 1997; Janssen & Van Yperen, 2004
Transactive memory system (TMS)	Team	A team's cooperative division of cognitive effort for storing, retrieving, and communicating team knowledge.	Hollingshead, 2001; Lewis, 2003, 2004; Liao et al., 2012; Wegner, 1987
Geographical dispersion (GeoDisp)	Team	The extent to which a team is geographically dispersed.	Ganesh & Gupta, 2010; Gilson et al., 2015

But beware!

- Risk of uniformity of study designs
 - Can reduce creativity,
 - Can reduce exploration of alt methodologies
 - Can reduce multi-method studies
- Focusing only on quantity of empirical basis
 - Repository studies that just increase
 - # of projects/files/classes/methods
 - But are all projects relevant?
 - Bad example: Classify projects based on majority of programming language used in its files - interaction, differences, nuance discarded





Emeritus Prof



PI



PhD

Recommendations! (1)

Dare think early about your causal model

Clarify factors and estimand(s). Postulate causal model in line with prior research and common sense.

Explicitly state research model

Use Pearl DAG or other path model/diagram to clarify constructs & hypotheses. Define constructs & operationalisations clearly. Also for qualitative studies.

Pre-Register your research

Several conference tracks & journals (EMSE, TOSEM) now support this.

Share data and analysis scripts

Replication package with data as well as scripts. Even for qualitative data (when allowed).



Emeritus Prof



PI



PhD

Recommendations! (2)

Clearly report on model, data, and analysis

Also in abstract.

Consider empirical standards to help guide reporting & clarity.

Explore multiple- & alternative methodologies

Avoid using only the “standard”, normal, or currently accepted/trendy methodologies. Consider which data your context give access to and adapt to it.

Increase quantity & quality of empirical basis

Fewer but stronger studies is often better for science longer-term.

But consider quality and relevance of data for your hypotheses.

Quantity is rarely a quality in itself.

Manifesto for Empirical Software Engineering 2.0

Empirical evidence over theoretical & formal arguments

Systematic & explicit methods over one-off, unique studies

Practical context & impact over clean but simplified lab studies

Truth over novelty, relevance and importance

Plurality & nuance over simple, dichotomous claims

Human factors over algorithms & technology

Explanations & theories over descriptions of data at hand