RESEARCH METHODS
Empirical/Experimental CS Research Methods

Anton Dignös

Free University of Bozen-Bolzano
Faculty of Computer Science
IDSE – Database Systems Group

April, 2019

Acknowledgements: I am indebted to Johann Gamper and Francesco Ricci for providing me their slides, upon which these lecture notes are based.

Course Structure and Schedule/1

- Lectures: 6 hours
  - Monday, April 8, 09:30–12:30, Room: E420
  - Friday, April 12, 09:30–12:30, Room: E331
- Homework: 10 hours

Course Structure and Schedule/2

- Class I
  - Initial brainstorming and introduction of key concepts
  - Presentation of experimental research methods in general
  - Presentation of experimental research in CS (I)
  - Paper assignment for homework
- Homework
  - Each student must read and analyze a paper about an empirical/experimental evaluation
  - Prepare a short presentation (15 mins) where you illustrate the article, focusing on the experimental evaluation
- Class II
  - Student presentations of the research paper
  - Critical discussion of each paper and presentation
  - Presentation of experimental research methods (II)

Assessment

- Critical presentation of the assigned article, showing that you have considered and evaluated all the dimensions illustrated in the lecture

Goals

- Knowledge
  - Understanding of different research methods and paradigms
  - In particular, empirical and engineering research methods
- Skills
  - Ability to set up an experimental evaluation for a reasearch topic
  - Ability of critical thinking, reading and evaluation
  - Ability to present a logical and coherent argument

What is Research?

- Research comprises creative work undertaken on a systematic basis in order to increase the stock of knowledge, including knowledge of humans, culture and society, and the use of this stock of knowledge to devise new applications. It is used to establish or confirm facts, reaffirm the results of previous work, solve new or existing problems, support theorems, or develop new theories [...] The primary purposes of basic research (as opposed to applied research) are documentation, discovery, interpretation, or the research and development (R&D) of methods and systems for the advancement of human knowledge. Approaches to research depend on epistemologies, which vary considerably both within and between humanities and sciences. There are several forms of research: scientific, humanities, artistic, economic, social, business, marketing, practitioner research, etc.

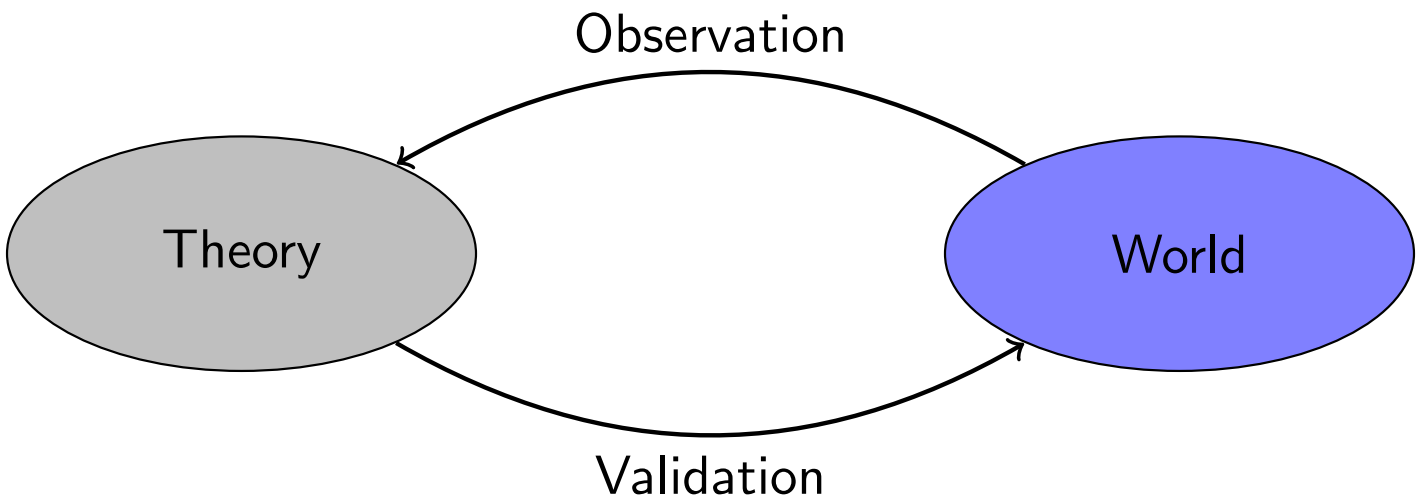
  [Wikipedia]

Research Methods, Techniques and Methodology

- Research Method: refers to the manner in which a particular research project is undertaken.
- Research Technique: refers to a specific means, approach, or tool-and-its-use, whereby data is gathered and analysed, and inferences are drawn.
- Research Methodology: refers to the study of research methods; it does not admit of a plural.

Research Methods

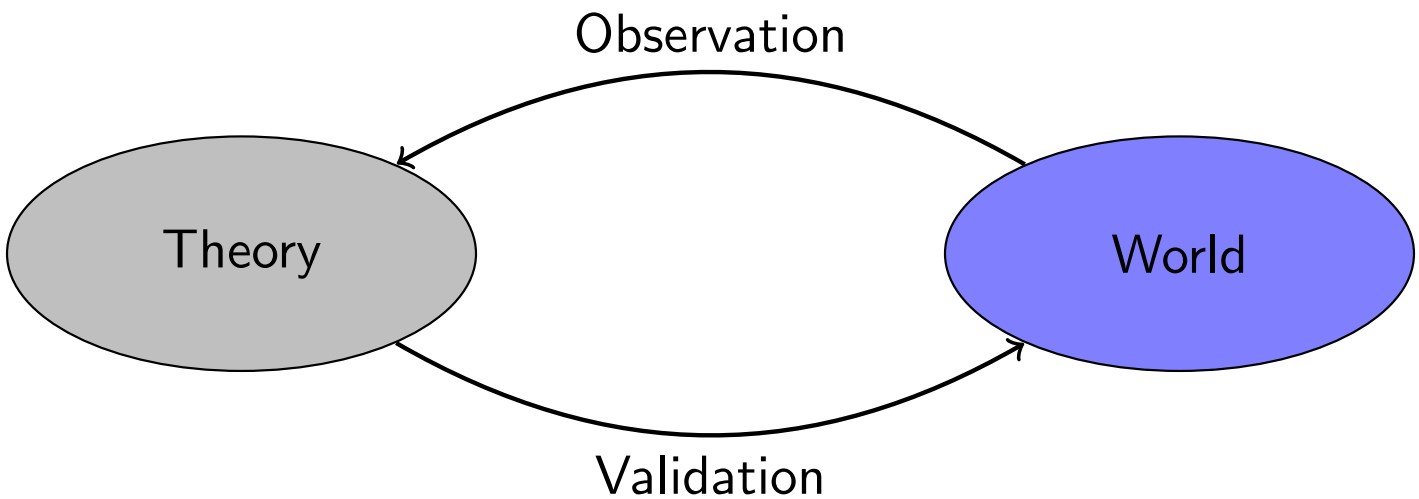
- The purpose of the research determines the method to use
- There is no single research method
- Many methods are available and have to be combined
- But somehow, scientists/researchers are supposed to do this:



How do you see your research in this cycle?

Research Methods

- The purpose of the research determines the method to use
- There is no single research method
- Many methods are available and have to be combined
- But somehow, scientists/researchers are supposed to do this:



How do you see your research in this cycle?

Different Research (Methods) Exist

- Exploratory research structures and identifies new problems.
- Constructive research develops solutions to a specific persisting problem.
- Empirical research tests the feasibility of a solution using empirical evidence.

Exploratory Research

- This is done to improve the basic knowledge on the concept and walk in to the unknown realms of the subject.
- It is a type of research conducted for a problem that has not been clearly defined.
- It should draw definitive conclusions only with extreme caution.
- Given its fundamental nature, exploratory research often concludes that a perceived problem does not actually exist.

Constructive Research

- This is done by technical professionals to find a new solution to a specific persisting problem.
- It is very commonly used in computer science research.
- The term "construct" is often used in this context to refer to the new contribution being developed, such as a new theory, algorithm, model, software, or a framework.
- This approach demands a form of validation
- This may involve evaluating the "construct" analytically against predefined criteria or performing some benchmark tests with the prototype.

Empirical Research

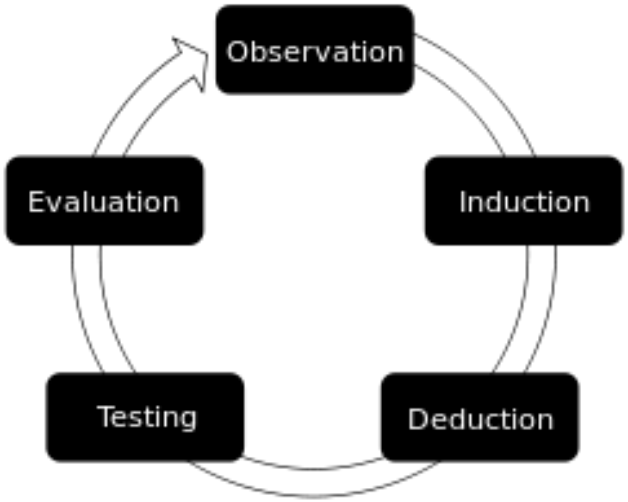
- "Empirical" comes from the Greek word for experience: ἐμπειρία (empeiría)
- Observation is the key: Empirical research is a way of gaining knowledge by means of direct and indirect observation or experience.
- Empirical evidence/observations can be analyzed quantitatively or qualitatively.
- Through quantifying the evidence or making sense of it in qualitative form, a researcher can answer empirical questions
- A combination of qualitative and quantitative analysis is often used to better answer questions.

Empirical Research – Example

- Empirical question: *"Does listening to music during learning have an effect on later memory?"*
- Based on existing theories about the topic, some hypotheses will be proposed, e.g., *"Listening to music has a negative effect on learning."*
- This prediction can then be tested with a suitable experiment.
- Depending on the outcomes of the experiment, the theory on which the hypotheses and predictions were based will be supported to a certain degree of confidence or not
- e.g., *"People who study while listening to music will remember less on a later test than people who study in silence."*

A.D. de Groot's Empirical Cycle

- A.D. de Groot's empirical cycle:
  1. Observation: The collecting and organization of empirical facts.
  2. Induction: Formulating hypothesis.
  3. Deduction: Deducting consequences of hypothesis as testable predictions.
  4. Testing: Testing the hypothesis with new empirical material.
  5. Evaluation: Evaluating the outcome of testing



- Adrianus Dingeman de Groot (1914–2006) was a Dutch chess master and psychologist
- Conducted some of the most famous chess experiments of all time.
- In 1946 he wrote his thesis *"Het denken van den schaker"* (*Thought and choice in chess*).
- Played in the Chess Olympiads 1937 and 1939.

Research Techniques

- Qualitative research techniques
- Quantitative and scientific research techniques
- Research techniques at the scientific/interpretivist boundary
- Non-empirical techniques
- Engineering research techniques

Qualitative Techniques

- Have their roots in the social sciences
- Primarily concerned with inreasing and in-depth understanding of an area
- Investigate why and how of decision making, not just what, where, when.
- Often associated with fieldwork, face-to-face interviews, focus groups, site visits
- Focus on the analysis of a limited number of samples/settings
- Produce information only on the particular cases studied
  - Any more general conclusions are only hypotheses (informative guesses).
  - Quantitative methods can be used to verify such hypotheses.
- As humans and organisational conditions change over time, the pre-condition for the study and the analysis of the problem change
  $\Longrightarrow$ repeatability of experiments may not be possible.

Quantitative Techniques

- Origin in the natural sciences
- Systematic empirical investigation of quantitative properties and phenomena and their relationships.
- The goal is to develop models, theories, and hypotheses pertaining to natural phenomena (how it works)
- The research is generally driven by hypotheses, which are formulated and tested rigorously.
- Measurement is fundamental since it gives the connection between observation and the formalization of the model, theory and hypothesis
- Repeatability of the experiments and testing of hypotheses are vital to the reliability of the results, since they offer multiple opportunities for scrutinising the findings.

Scientific Research Techniques

- Forecasting: involves the application of regression and time-series techniques, in order to extrapolate trends from past data.
- Field experimentation and quasi-experimental designs: opportunities are sought in the real-world which enable many factors, which would otherwise confound the results, to be isolated, or controlled for.
- Laboratory experimentation: this involves the creation of an artificial environment, in order to isolate and control for potentially confounding variables.

Research Techniques at the Scientific/Interpretivist Boundary

- Field study: the object of study is subjected to direct observation by the researcher.
- Questionnaire-based survey: involves the collection of written data from interviewees, or the collection of verbal responses to relatively structured questions.
- Case study: this involves the collection of considerable detail, from multiple sources, about a particular, contemporary phenomenon within its real-world setting.
- Secondary research: this technique analyses the contents of existing documents. Commonly, this is data gathered by one or more prior researchers, and it is reexamined in the light of a different theoretical framework from that previously used.

Non Empirical Techniques

- Conceptual research: opinion and speculation, comprising philosophical or 'armchair' analysis and argumentative/dialectic analysis.
- Theorem proof: applies formal methods to mathematical abstractions in order to demonstrate that, within a tightly defined model, a specific relationship exists among elements of that model.
- Futures research, scenario-building, and game- or role-playing: individuals interact in order to generate new ideas, gather new insights into relationships among variables, and postulating possible, probable, and preferable futures.
- Review of existing literature, or 'meta-analysis': the opinions and speculations of theorists, the research methods adopted by empirical researchers, the reports of the outcomes of empirical research, and materials prepared for purposes other than research.

Engineering Research Techniques

- Construction: involves the conception, design and creation (or 'prototyping') of an artifact and/or technique.
  - The new technology is designed to intervene in some setting, or to enable some function to be performed.
  - The design is usually based upon a body of theory
  - Artifact/technology is usually subjected to some form of testing, in order to establish the extent to which it achieves its aims.
- Destruction: new information is generated concerning the characteristics of an existing class of technologies.
  - Typically achieved through testing the technology, or applying it in new ways.
  - The design is usually based upon a body of theory.

Empirical Research Techniques in Computer Science

- Two important classes of empirical research techniques in CS
  - Run experiments to measure parameters of software artefacts
  - User studies using questionaires, etc.

Advantages of Empirical Research Methods and Techniques

- Go beyond simply reporting observations or proving theorems
- Prove relevancy of theory by working in a real world environment (context)
- Help integrating research and practice
- Understand and respond more appropriately to dynamics of situations
- Provide respect to contextual differences
- Provide opportunity to meet standards of professional research

Experimental Research Improves Quality of Research

- Empirical research is often the final step in research with the aim to "prove" theories in real life
- Research is an iterative and contiuous process from ideas to final verification of the ideas in the real-world
  1. Initial ideas, concepts, intuition, ... in your head
  2. Write down and explain your thoughts
  3. Prove theorems, lemmas, propositions, ...
  4. Implementation of research prototype
  5. Empirical/experimental evaluation against synthetic and real-world data
- Each step
  - reveals weaknesses, errors, ...
  - refines the theory
- At the end, empirical research pushes research to another level of quality!

Computer Science

Shifting Definition of Computer Science

- *Computer science is in part a scientific discipline concerned with the empirical study of a class of phenomena, in part a mathematical discipline concerned with the formal properties of certain classes of abstract structures, and in part a technological discipline concerned with the cost-effective design and construction of commercially and socially valuable products* [Wegner, 1971]

- *Since its beginnings in the late 1930s, computer science has been a unique combination of math, engineering, and science. It is not one, but all three.* [Denning]

Research Paradigms in CS

- Empirical: Computer science is concerned with the study of a class of phenomena
- Mathematical: Computer Science is concerned with the study of algorithms and properties of information structures (abstraction from real objects)
- Engineering: managing the cost-effective design and construction of complex software-hardware systems (commercially and socially valuable).

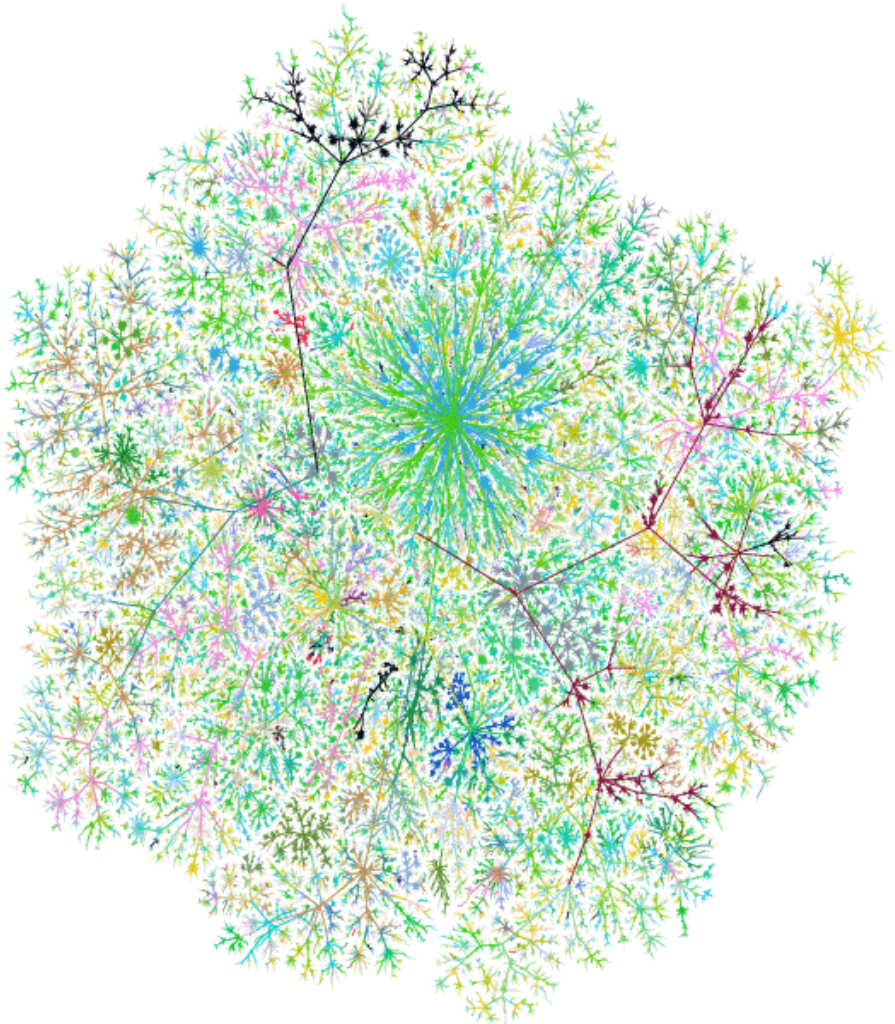
[Wegner, 1976]

Programming Languages – the Diachronic Perspective

- 1950–1960 the age of empirical discovery:
  discovery of basic techniques such as look-up techniques or the stack algorithm for evaluating arithmetic expressions. Prog. Lang. were considered as tools for facilitating the specification of programs.
- 1961–1969 the age of elaboration and abstraction: theoretical work in formal languages and automata theory with application to parsing and compiling.
- 1970–? the age of technology: decreasing HW costs & increasing complex SW projects created a "complexity barrier". Development of tools and methodologies for controlling the complexity, cost and reliability of large programs.

[Wegner, 1976]

Empirical



The Structure of the Web

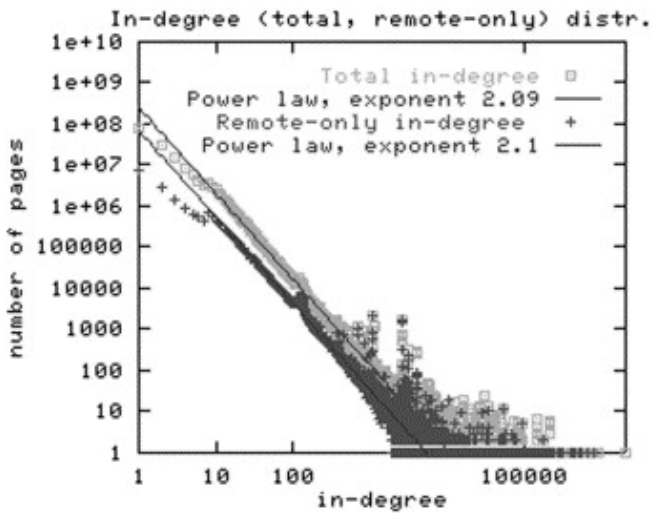
- Web does not have an engineered architecture: hundreds of billions of pages created by billions of users.
- Web contains a large strongly connected core (each page can reach every other).
- The shortest path from one page in the core to another involves 16–20 links (a small world).
- Analysis of web structure led to better search engines (e.g., Google PageRank method) or content filtering tools.

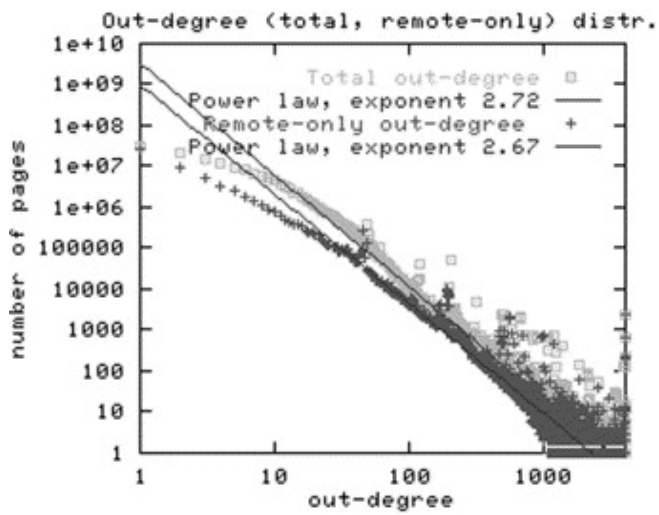
[Broder et al., 2000]

Distribution Links

- The number of links to and from individual pages is distributed according to a power law: e.g., the fraction of pages with n in-links is roughly $n^{-2.1}$

[Broder et al., 2000]





Mathematical

- Study of algorithms (Knuth)
  - Design and analysis of (optimal) algorithms for particular problems
  - Computational complexity
- Study of representation, transformation and interpretation of information structures
  - Models for characterizing general-purpose tools
  - Mechanisms and notations for computing all computable functions.

Mathematical – Example

- How to deal with the problem of empty result set for Boolean queries, i.e., queries that contain a set of key-words and fail to return any item
- e.g., $q = \{prolog, language, comparison, survey, rating\}$ fails to retrieve any record (web page)

q'	url1	prolog		comparison	survey	
	url2		language	comparison	survey	rating
	url3	prolog	language		survey	
	url4		language	comparison	survey	
q''	url5	prolog	language	comparison		rating
q'	url6	prolog		comparison	survey	
	url7		language	comparison		
	...	...	...	...	...	...

- but there are results for $q' = \{prolog, comparison, survey\}$ or $q'' = \{prolog, language, comparison, rating\}$.

Relaxation of Boolean Queries

- Godfrey [1997] studied extensively the problem of empty result set for Boolean queries, i.e., queries that contain a set of keywords and fail to return any item
- Solution: Find a maximal succeeding subquery
  - one of these succeeding subqueries can be found in $O(|q|)$
  - two in $O(|q|^2)$
  - all makes the problem intractable

Engineering

- Building a robot for the new mission to Mars

- . . . showing that it works (better than the previous model)

My system is better ...

- This does not work!

My Creation is Better

- Discovering a fact about nature or about the math world, it is a contribution *per se*, no matter how small
- But in the engineering field anyone can create some new thing
- One must show that the creation is better
  - Solves a problem in less time
  - Solves a larger class of problems
  - Is more efficient of resources
  - Is more expressive by some criterion
  - Is more visually appealing
  - Presents a totally new capability
  - ...
- The "better" property is not simply an observation, but is much more complex and demanding!

Empirical-Mathematical-Engineering

- **Example:** Dealing with failing queries
  - Analyse the failing queries that users generate
  - Define a tractable problem, e.g., find all the maximal succeeding subqueries of q of length $|q-1|$
  - Design an algorithm that can run in linear time and solve the above problem
  - Design and implement a middleware that get such a query, call a standard SQL-based DBMS and returns the found subqueries
  - Empirically test the middleware on a set of real queries (user input) and characterize when such an algorithm is useful (enough powerful to solve the majority of real queries).

[Mirzadeh et al., 2004]

. . . and apply it to Tourism

. . . and show that is better

- IQM is the intelligent query management component that suggests query relaxation (and tightening)
- 40 users tried to plan their vacation in Trentino using NutKing
- Half of them used a system version with IQM: NutKing+
- The other half used a system version that did not support query relaxation: NutKing-

Objective Measures	**NutKing-**	**NutKing+**
Queries submitted by a user	20 ± 19.2	13.4 ±9.3 *
# of constraints in a query	4.7 ±1.2	4.4 ± 1.1
Avg query result size	42.0 ± 61.2	9.8 ±14.3**
# of times relaxation suggested	n.a.	6.3 ± 3.6
# of times the user accepted a suggested relaxation	n.a.	2.8 ± 2.1

Basic vs. Applied Research

- Basic research (aka fundamental or pure)
  - Driven by a scientist's curiosity or interest in a scientific question
  - Main motivation is to expand man's knowledge, not to create or invent something
  - There is no obvious commercial value to the discoveries that result from basic research.
  - e.g., How did the universe begin?
    What are protons, neutrons, and electrons composed of?
- Applied research
  - Designed to solve practical problems of the modern world, rather than to acquire knowledge for knowledge's sake
  - One might say that the goal of the applied scientist is to improve the human condition.
    - e.g., improve agricultural crop production
    - treat or cure a specific disease
    - help consumer to find best deals

Experimental vs. Theoretical CS

- Experimental Computer Science
  - Experimental computer science (ECS) refers to the building of and/or experimentation with nontrivial HW or SW systems
  - ECS does not depend on a formalized theoretical foundation in the same way that experimental physics can draw on theoretical physics
    - According to theory XXX we must observe this – then experimentally we look for it (if it is not observed the theory is falsified, see K. Popper)
  - Good experimentalists do create models (theories) and test (reject or accept) hypotheses
  - Experiments are most often conducted to validate some informal thesis derived from a computational model (but not rigorously specified by theory) that may have been developed for the experiment
  - Due to the complexity of the systems built in ECS, experimental implementation/evaluation is necessary to evaluate the ideas and the models or theories behind them.
- Theoretical Computer Science
  - "Theory" in CS is very close to mathematics – theoreticians prove theorems

Technique- and Problem-Driven Research

- Technique-Driven Research
  - Primarily interested in a technique (e.g. neural networks)
  - Look for applications of it
  - Much computer science is here
  - Tend to "abuse" and push unnecessary techniques not justified by the problem at hand
- Problem-Driven Research
  - Primarily interested in a goal (e.g., support autistic children)
  - Use whatever methods are appropriate
  - Tend to be considered as "naive" and not enough "formal"
- Technique people "learn" about many applications
- Problem-driven people "learn" about many techniques.

Experimental Evaluation in Computer Science/1

- Tichy et al.: *Experimental evaluation in computer science: a quantiative study*. Journal of Systems and Software, 1995.
  - A survey of over 400 recent research articles suggests that computer scientists publish relatively few papers with experimentally validated results.
  - The survey includes complete volumes of several refereed CS journals, a conference, and 50 titles drawn at random from all articles published by ACM in 1993. The journals Optical Engineering (OE) and Neural Computation (NC) were used for comparison.
  - Of the papers in the random sample that would require experimental validation, 40% have none at all. In journals related to software engineering, this fraction is over 50%.
  - In comparison, the fraction of papers lacking quantitative evaluation in OE and NC is only 15% and 12%, respectively.
  - Conversely, the fraction of papers that devote one fifth or more of their space to experimental validation is almost 70% for OE and NC, while it is a mere 30% for the CS random sample and 20% for software engineering.
  - The low ratio of validated results appears to be a serious weakness in CS research. This weakness should be rectified for the long-term health of CS.

Experimental Evaluation in Computer Science/2

- Of course, there are top journals and conference with a strong emphasis on experimental evaluation
- Selected examples include:
  - e.g., SIGMOD, VLDB, VLDB journal, ICDE (databases), KDD (data mining), IR (information retrieval)
  - Experimental evaluation papers in VLDB since a few years
  - Bioinformatics Journal:
    - Provides a strict structure on the paper: Background, Methods, Results

Different Experimental Evaluation Techniques

- Depending on the objective, various evaluation techniques shall be used
- Quantitative testing/experiments of algorithms/programs/databases/...
- Usability tests with users
- Questionnaires
- Surveys
- Case studies
- ...

Parameters to be Evaluated

- Runtime
- Preprocessing time
- Disk space (overhead)
- Memory
- Correctness of results
- Accuracy of approximation algorithms
- User satisfaction
- Usability
- ...
- . . .
- Dive into the details! You will discover/explore new features of the problem!

Data Sets

- Real-world data
  - Always good to have – show that system works in practice
  - Sometimes difficult to obtain
  - Do not allow to test all aspects of an algorithm/system
- Synthetic data
  - Allow to test specific aspects of the algorithm
  - Often (very) difficult to generate
- If possible, try to use the same data as your competitors
- It is easy to show that your approach is better if only very particular data is used
- Describe the most important aspects of the data

Benchmarks

- In some areas, well known benchmarks are available
  - TPC benchmarks for databases
  - DIMACS benchmark for road networks
  - UCR Time Series Classification Archive
  - ...
- Use existing benchmarks as much as possible
- Facilitates the comparison of different solutions

Organizing Experiments

- Running experiments is time-consuming and requires care
- Important to have a good handling on it
- Do a fair comparison with state of the art competitors
  - Might require a lot of implementation of other methods!
- Keep repeatability in mind: you will have to run the experiments again and again, before the submission, during the preparation of the final version, ...
- Hence, all steps of running experiments must be automatic as much as possible
- Bash scripts are a useful tool
  - 1 script for each experiment
  - 1 meta-script that runs all experiments, e.g., over night
- Consider how to import the results into a gnuplot or tikz to draw plots
  - Must be simple and automatic, otherwise you will do mistakes
- Other (scripting) languages might be used as well: perl, awk, pyton, etc.
- Parameter settings for evaluated solutions is critical and need "good choices" and explanations!

Example Bash Script for an Experiment

```
#!/bin/bash

. ../env.sh

if [ "$#" -gt 0 ]; then
  repeat=$1
else
  repeat="1"
fi

OUTPUT="fig15a.dat"
PARSE="java -cp .. ParseRuntime"

echo -ne "" > $OUTPUT

for INPUTK in 1 3 5 10 50 100
do
  $PSQLC -f init${INPUTK}k.sql
  $PSQLC -f index.sql
  $PSQLC -f analyze.sql
  echo -ne ${INPUTK} >> $OUTPUT
  echo -ne " " >> $OUTPUT

  rm -f tmp.out
  for (( i=0; $i < $repeat; i=$i+1)) do
    $PSQLC -f ljoin-align-true.sql >> tmp.out
  done
  $PARSE tmp.out >> $OUTPUT
  echo -ne " " >> $OUTPUT

  if [ $INPUTK -gt 5 ]; then #ignore sql for larger 5k
    echo -ne "nan" >> $OUTPUT
  else
    rm -f tmp.out
    for (( i=0; $i < $repeat; i=$i+1)) do
      $PSQLC -f ljoin-sql-true.sql >> tmp.out
    done
    $PARSE tmp.out >> $OUTPUT
  fi
```

Example Bash Script to Run all Experiments

```
#!/bin/bash

login=xxx/yyy
logdir="log$HOSTNAME"
date="100504"
input="lineitem50M"

# echo -n "EXP. PARTITIONING (Large GT) "; date
# Epart $login $input 5000000 > $logdir/Epartl.$date.log
# echo "DONE"; date

echo -n "EXP. INDEXING (1K GT): "; date
Eindex $login $input 1000 > $logdir/Eindex1k.$date.log
echo "DONE"; date

echo -n "EXP. INDEXING (2K GT): "; date
Eindex $login $input 2000 > $logdir/Eindex2k.$date.log
echo "DONE"; date

# echo -n "EXP. 1 DETAIL TABLE (Large GT): "; date
# E1 $login 5000000 > $logdir/E1l.$date.log
# echo "DONE"; date

echo -n "EXP. 2 GROUP TABLE (5M GT): "; date
E2 $login $input 5000000 > $logdir/E2l.$date.log
echo "DONE"; date

echo -n "EXP. INDEXING (10K GT): "; date
Eindex $login $input 10000 > $logdir/Eindex10k.$date.log
echo "DONE"; date

...
```

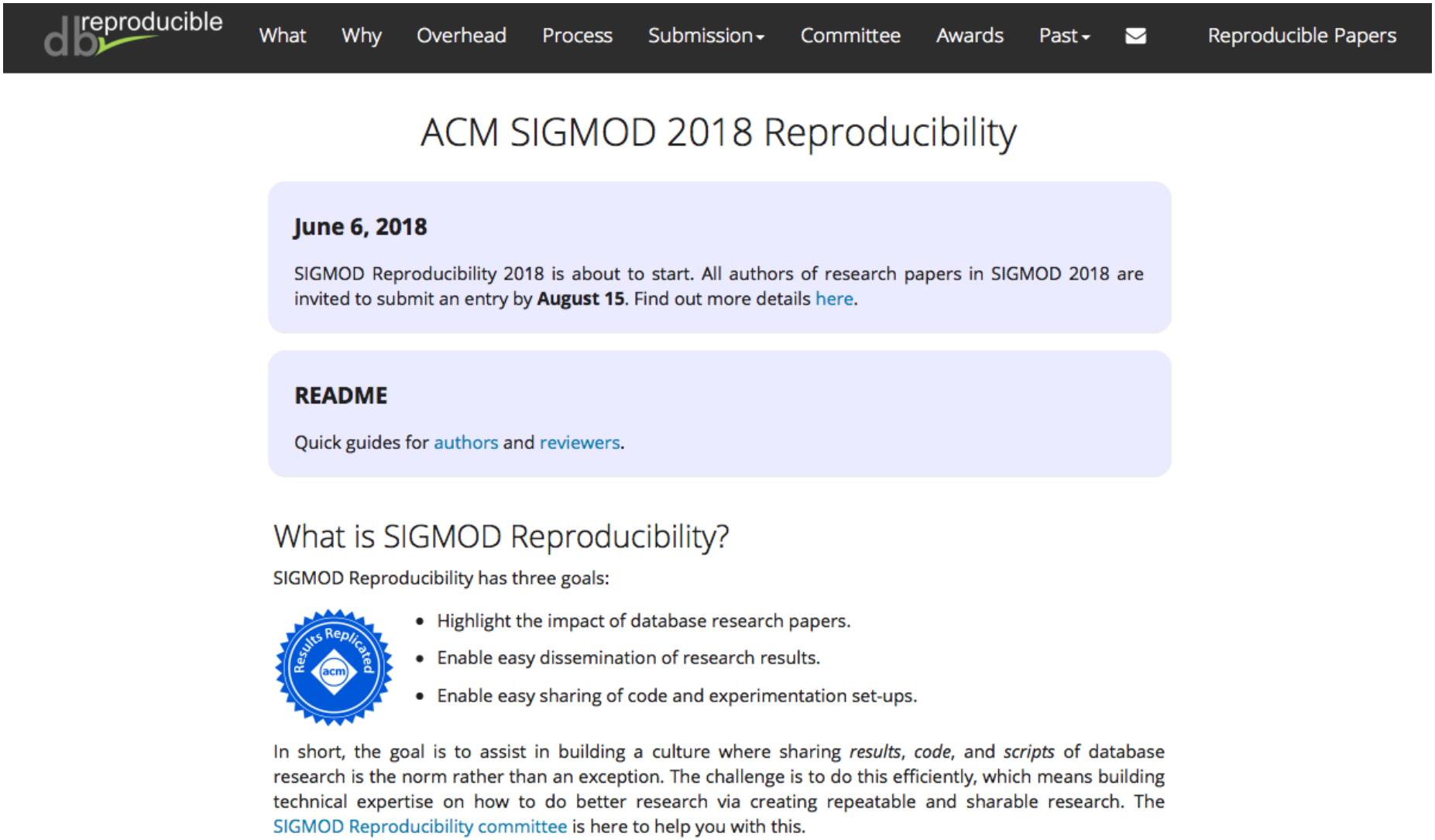
Reproducibility/1

- Initiated by SIGMOD 2012

  The goal of establishing reproducibility is to ensure your SIGMOD research paper stands as reliable work that can be referenced by future research. The premise is that experimental papers will be most useful when their results have been tested and generalized by objective third parties.

- Joined by PVLDB in 2018

Reproducibility/2



Reproducibility/3

pVLDB Reproducibility Submission What Why Overhead Papers Process Committee

pVLDB Reproducibility 2018

Starting with pVLDB 2018, pVLDB joins SIGMOD in encouraging the database community to develop a culture of sharing and cross-validation. pVLDB's reproducibility effort is being developed in coordination with SIGMOD's.

News

Reproducibility submissions are now open through CMT.

What is pVLDB Reproducibility?

pVLDB Reproducibility has three goals:

- Increase the impact of database research papers.
- Enable easy dissemination of research results.
- Enable easy sharing of code and experimentation set-ups.

In short, the goal is to assist in building a culture where sharing results, code, and scripts of database research is the norm rather than an exception. The challenge is to do this efficiently, which means building technical expertise on how to do better research via creating repeatable and sharable research. The pVLDB Reproducibility committee is here to help you with this.

Reproducibility/4

Why should I be part of this?

You will be making it easy for other researchers to compare with your work, to adopt and extend your research. This instantly means more recognition for your work and higher impact.

Taking part in the SIGMOD Reproducibility process enables your paper to take the **ACM Results Replicated** label. This is embedded in the PDF of your paper in the ACM digital library.

There is an option to also host your data, scripts and code in the ACM digital library as well to make them available to a broad audience, which will award the **ACM Artifacts Available** label.

ACM Results Replicated label

The experimental results of the paper were replicated by the committee and were found to support the central results reported in the paper.

ACM Artifacts Available label

The experiments (data,code,scripts) are made available to the community.

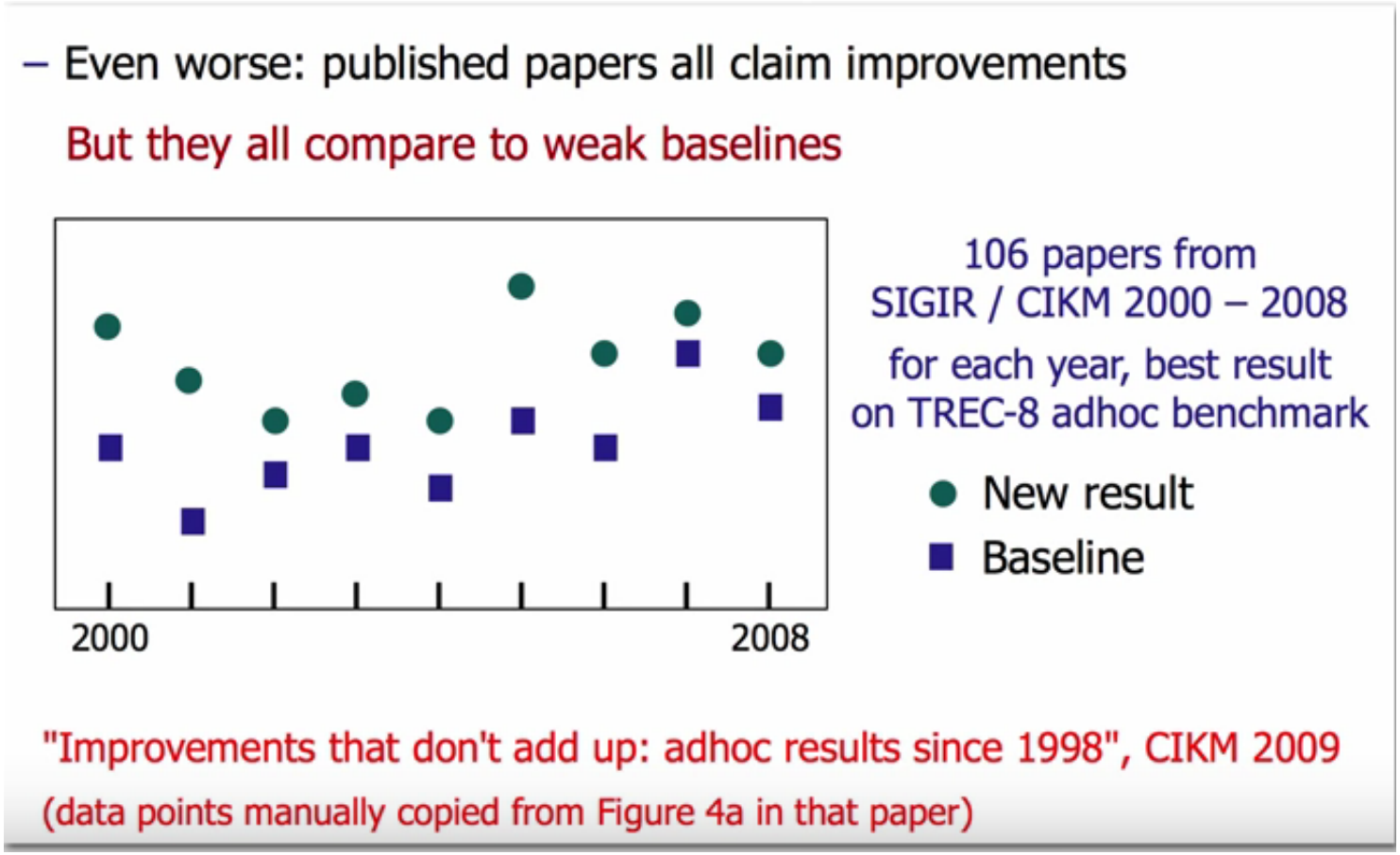
> Both the "ACM Results Replicated label" and the "ACM Artifacts Available label" are visible in the **ACM digital library**.

How are Experiments Done - A Small Case Study/1

- Armstrong et al.: *Improvements That Don't Add Up: Ad-Hoc Retrieval Results Since 1998*, CIKM 2009
  - The existence and use of standard test collections in information retrieval experimentation allows results to be compared between research groups and over time. However, such comparisons are rarely made. Most researchers only report results from their own experiments, a practice that allows lack of overall improvement to go unnoticed.
  - The critical experimental failing, in our view, is that the great majority of papers only report on experiments that the researchers have carried out themselves, without reference to past result.
  - Our longitudinal analysis of published IR results in SIGIR and CIKM proceedings from 1998-2008 has uncovered the fact that ad-hoc retrieval is not measurably improving.
  - A central repository of effectiveness results presents a solution to this problem: best known results could be quickly found by authors, and readers and reviewers could more effectively assess claims made in papers.

How are Experiments Done - A Small Case Study/2

- Jens Dittrich: The Case for Small Data Managment
  `https://youtu.be/O7Qgo6RSzmE?t=19m`

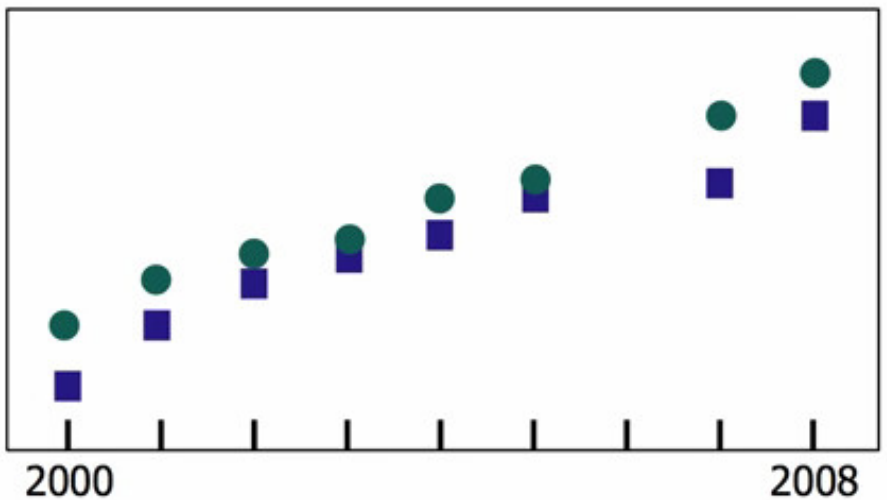


[Jens Dittrich, 2015]

How are Experiments Done - A Small Case Study/3

– Even worse: published papers all claim improvements

But they all compare to weak baselines



The picture should rather look like this

- New result
- Baseline

"Improvements that don't add up: adhoc results since 1998", CIKM 2009

(data points manually copied from Figure 4a in that paper)

[Jens Dittrich, 2015]

Towards Executable Papers

- Dittrich and Bender: *Janiform Intra-Document Analytics for Reproducible Research*. PVLDB 2015.

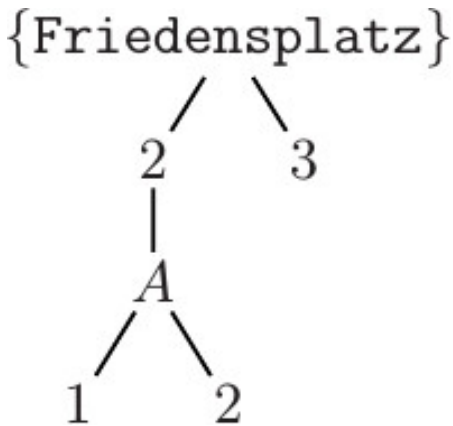
  `https://bigdata.uni-saarland.de/publications/p1972-dittrich.html`

1. Save it
2. Use it
3. Change extension to .pdf
4. Future work: allow also .ova (virtual box) extension

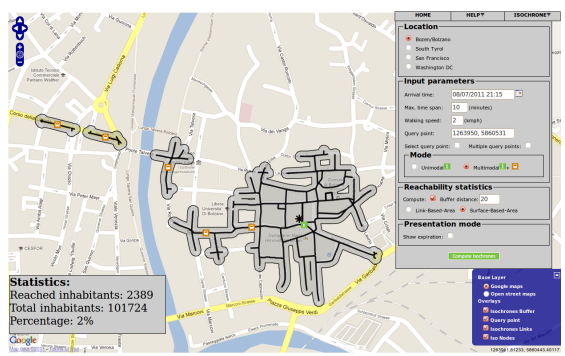
Working with Real Data is Rewarding – eBZ Project

- Working with real-world data not only helps for the evaluation, but reveals interesting insights and helps to identify particularities of problems, which often leads to new research.

- Example 1: Synchronization of residential addresses in databases of the Municipality of Bozen-Bolzano
- One big sub-problem was the matching of street names
- Solution: Represent a street as an address tree
- $\Longrightarrow$ PhD of Nikolaus Augsten



- Example 2: Reachability analysis in Bozen-Bolzano
- Solution: compute isochrones
- $\Longrightarrow$ PhD of Markus Innerebner



VLDB Experimental Evaluation Papers/1

Experiments & Analysis Papers

Green turtle, Chelonia mydas, Hawaii, Photo by Brocken Inaglory, GNU Free Documentation/Creative Commons Attribution license, via Wikimedia

Experiments and Analysis Papers focus on the experimental evaluation of existing algorithms, data structures, and systems. Papers proposing new techniques should continue to be submitted to the regular research track. The primary contribution of Experiments and Analysis papers is performance evaluation through analytical modelling, simulation, and/or experiments. Suitable papers can fit in different categories:

1. **Experimental Surveys**: papers that compare a wide spectrum of approaches to a problem and, through extensive experiments, provide a comprehensive perspective on the results available and how they compare to each other.
2. **Result Verification**: papers that verify or refute results published in the past and that, through a renewed performance evaluation, help to advance the state of the art.
3. **Problem Analysis**: papers that focus on relevant problems or phenomena and through analysis and/or experimentation provide insights on the nature or characteristics of these phenomena.

VLDB Experimental Evaluation Papers/2

- Ding et al.: *Querying and mining of time series data: experimental comparison of representations and distance measures*, VLDB 2008
- Zhang et al.: *Crowdsourced top-k algorithms: an experimental evaluation*, VLDB 2016
- Lu et al.: *Large-scale distributed graph computing systems: an experimental evaluation*, VLDB 2016
- Papenbrock et al.: *Functional dependency discovery: an experimental evaluation of seven algorithms*, VLDB 2016
- Wu et al.: *Shortest path and distance queryies on road networks: an experimental evaluation*, VLDB 2012
- Li et al.: *An experimental study on hub labeling based shortest path algorithms*, VLDB 2018.
- Jiang et al.: *String similarity joins: an experimental evaluation*, VLDB 2014
- Chen et al.: *Spatial keyword query processing: an experimental evaluation*, VLDB 2013
- Han et al.: *An experimental comparison of Pregel-like graph processing systems*, VLDB 2014
- Lu et al.: *Large-scale distributed graph computing systems*, VLDB 2014
- Weber et al.: *A quantiative analysis and performance study for similarity-search methods in high-dimensional spaces*, VLDB 1998
- Huang et al.: *Experimental evaluation of real-time optimistic concurrency control schemes*, VLDB 1991
- Zhang et al.: *An experimental evaluation of simrank-based similarity search algorithms*, VLDB 2017.
- Memarzia et al.: *A Six-dimensional Analysis of In-memory Aggregation*, EDBT 2019
- Mann et al.: *An Empirical Evaluation of Set Similarity Join Techniques*, VLDB 2016

Homework

- Choose one of the papers on the previous slide (or propose another experimental evaluation paper)
- Read the paper carefully, in particular with respect to the experimental evaluation part
  - Are the experiments clear and carefully done?
  - Are all relevant parameters evaluated?
  - Are the measures used meaningful?
  - Are the datasets large enough, real-world, realistic, ...?
  - What did you learn?
  - etc.
- Prepare a 15 minutes presentation of the paper for Friday

Benchmarking

- The following slides are taken from:
  Susan Elliott Sim, Steve Easterbrook, and Richard C. Holt. Using Benchmarking to Advance Research: A Challenge to Software Engineering, Proceedings of the Twenty-fifth International Conference on Software Engineering, Portland, Oregon, pp. 74-83, 3-10 May, 2003.

What is a Benchmark?

- A benchmark is a standard test or set of tests used to compare alternatives. It consists of the following components:
  1. Motivating comparison
     - Motivation for research area and benchmark
  2. Task sample
     - Representative sample of problems from a problem domain
     - Most controversial part of benchmark design
  3. Performance Measures
     - Can be qualitative or quantitative, measured by human, machine, or both
- Becomes a standard through acceptance by a community
- Though benchmarks exist in many scientific disciplines, we are primarily concerned with technical benchmarks in computer science research communities

Benchmarking as an Empirical Method

Characteristics from experiments	Characteristics from case studies
Features • Use of control factors • Replication • Direct comparison of results	**Features** • Little control over the evaluation setting, (e.g., choice of technology and user subjects) • No tests of statistical significance • Some open-ended questions possible
Advantages • Direct comparison of results	**Advantages** • Method is flexible and robust
Disadvantages • Not suitable for building explanatory theories	**Disadvantages** • Limited control reduces generalizability of results

Benefits of Benchmarking

- *"... benchmarks cause an area to blossom suddenly because they make it easy to identify promising approaches and to discard poor ones."*
  – Walter Tichy
- Stronger consensus on the community's research goals
- Greater collaboration between laboratories
- More rigorous validation of research results
- Rapid dissemination of promising approaches
- Faster technical progress
- Benefits derive from process, rather than end product

Successful Benchmarks in Computer Science

- TREC Ad Hoc Task
- TPC-A^{TM} (http://www.tpc.org/)
- UCR Time Series Classification Archive (http://www.cs.ucr.edu/~eamonn/time_series_data/)
- SPEC CPU2000
- Calgary Corpus and Canterbury Corpus
- Penn treebank
- xfig benchmark for program comprehension tools
- C++ Extractor Test Suite (CppETS)

Surveys and Questionnaires: Some Definitions[1]

- A type of research to collect data and facts about some certain situation or issue from a target population which is relevant for the study
- Survey research is the research strategy to study the relationships and characteristics
- Surveys are based on the desire to collect information about a well defined issue or situation/hypothesis from a well defined population
- Surveys are now used in all areas of life (e.g., business, politics, agriculture, industry, education, media, etc.), but also in computer science

- The main technique to collect data in surveys is through oral or written questionnaires

[1] Most of the slides in this section are adapted from Naveed Iqbal Ch.

Why Surveys?

- To describe/explain situations (analytical surveys)
- To identify or solve problems
- To measure the change, acceptance of products, etc.
- To study attitudes, behavior and habits
- To examine cause-effect relationships
- To study the characteristics
- To test or formulate a hypothesis
- etc.

Types of Surveys

- Descriptive survey
  - A descriptive survey attempts to picture or document current conditions or attitudes, i.e., to describe what exists at the moment
  - Examples:
    - Audience survey to determine the program taste
    - Study the need for a certain program
- Analytical survey
  - An analytical survey attempts to describe and explain why certain situations exist. Here we examine two, or more variable to test our research hypothesis
  - Examples:
    - Impact of software to learning behaviour
    - Impact of shopping app to consumer behavior
- Opinion surveys
  - Respondents expresses their viewpoint
- etc.

How to Conduct a Survey Study

1. Prepare questionnaire
   - Develop hypotheses
   - Decide on type of survey (mail, interview, telephone)
   - Write survey questions, decide on response categories and design layout
2. Plan how to record data and test the survey instrument
3. Decide on target population
   - Fix sampling size and select sample
4. Locate respondents, conduct interviews, record data
5. Enter data into computer and validate data, perform statistical analysis
6. Describe methods and findings in research report
   - Present findings to others for critique and evaluation

Methods of Surveys

- Mailed questionnaire
- Personal interview
- Telephone interview

(Mailed) Questionnaires

- One of the most important data collection survey methods
- Involves sending a questionnaire to a specific person (by mail, email or web)
- Advantages
  - Low cost
  - Reduction in biasing errors
  - Greater anonymity
  - Less time and trained staff required
- Disadvantages
  - Requires simple questions
  - No probing opportunity
  - No control over WHO fills?
  - Low response rate

Personal Interviews

- Together with the questionnaire, interviews make up the survey method, which is one of the most popular technique of data collection.
- Advantages
  - Flexibility in questioning
  - Control over the interview situation
  - High response rate
  - Collection of supplement data
- Disadvantages
  - Higher cost
  - Interviewer bias
  - Respondent's hesitation on sensitive topics
  - Greater staff requirement

Telephone Interview

- Telephone interview demonstrates the same structural characteristics as standard interviewing technique, except that it is conducted by telephone.
- Advantages
  - Moderate cost
  - Less time consumption
  - Higher response rate
  - Quality (supervision , recording)
- Disadvantages
  - Hesitation to discuss sensitive topics
  - "Broken-Off" interviews

Sampling is Critical and Difficult

- The process of choosing some representative members from the target population
- Probabilistic sampling methods
  - Simple random sampling
  - Systematic sampling
  - Cluster sampling
  - Stratified random sampling
  - Multi phase sampling
  - Spatial sampling
  - etc.

Structure of Questionnaire

- Cover letter
  - Main objectives and research team
  - Reasons why the respondent should complete the questionnaire
  - Assurance of anonymity and confidentiality
  - Requirements for completion such as maximum time, conditions, etc
- Instructions
  - How to fill the questionnaire
- Main body
  - Includes the questions
  - Be careful about content, structure, format, wording, etc.

Types of Questions and Responses[2]

- Type of questions and in particular the response sets is crucial for the subsequent analysis and interpretation of the data

- Open questions
  - \+ participants can answer exactly what they want; more information
  - − time-consuming and difficult to interpret
- Closed questions
  - \+ time-efficient, easy to interpret
  - − participants are required to choose a response that might not exactly reflect their opinion
- Contingency questions
  - Need to be answered only when the respondent provides a particular response to a previous question (aka filter question)
    - \+ Avoids asking people questions that are not applicable to them
- etc.

[2] Adapted from https://infoactive.co/data-design/

Closed Question Types

- Dichotomous questions: yes/no
- Multiple choice questions: choose from a set
  - Single answer vs. multiple answers
  - Answer choices should be mutually exclusive and exhaustive, i.e., not overlaping and cover all possible options
- Scaled questions
  - Likert scale: answers consist of ordered categories
    - e.g., poor, fair, good, very good, excellent
    - usually not more than 5 categories
  - Slider scale: answer is a mark anywhere along a numerical scale
    - e.g., [1,2,3,4,5,10], where 1 = strongly disagree and 5 = strongly agree
    - data are measured at an interval/metric level

Question Wording

- It is extremely important to think about how you word your questions
- Each question should be specific and have a defined focus
  - Bad example: Do you use Thunderbird because you don't like Microsoft?
- Questions should be relatively short (except where additional wording is absolutely necessary)
- Try to formulate neutral questions; biased or leading questions can easily skew the answers

References

- ACM Computing Classification System, http://www.acm.org/about/class/class/2012
- M. Berndtsson, J. Hansson, B. Olsson, B. Lundell, Thesis Projects: A Guide for Students in Computer Science and InformationSystems, Springer 2008
- R. Clarke, Appropriate Research Methods for Electronic Commerce http://www.anu.edu.au/people/Roger.Clarke/EC/ResMeth.html
- E. W. Dijkstra, Selected Writings on Computing: A Personal Perspective, Springer-Verlag, 1982.
- Dodig-Crnkovic G., Scientific methods in computer science, Conference for the Promotion of Research in IT at New Universities and at University Colleges in Sweden, Skövde. 2002
- National Research Council, Academic Careers for ExperimentalComputer Scientists and Engineers, National Academy Press, Washington, D.C., 1994.
- Wegner P., Research paradigms in computer science, Proceedings of the 2nd international conference on Software engineering, SanFrancisco, California, United States, Pages: 322 – 330, 1976.
- Chiasson T. and Gregory D.: Data + Design: A Simple Introduction to Preparing and Visualizing Information (https://infoactive.co/data-design)