RESEARCH METHODS
Empirical/Experimental CS Research Methods

J. Gamper

Free University of Bozen-Bolzano
Faculty of Computer Science
IDSE

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

Course Structure and Schedule/1

- Lectures: 6 hours
  - Tuesday, March 1, 09:30–12:30, Room: E412
  - Friday, March 4, 09:30–12:30, Room: E420
- Homework: 10 hours

Course Structure and Schedule/2

- Class I
  - Initial brainstorming on the key concepts
  - Teacher presentation of research methods
  - Critical discussion
  - 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
  - Discussion and critical discussion for each presentation
  - Discussion of the material read during the homework

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
  - Critical thinking
  - Critical reading and evaluation
  - The ability to present a logical and coherent argument

Topics discussed in the lecture

- What is research
- Research methods
- Research techniques
- What is computer science
- Research paradigms in Computer Science
- Experimental computer science vs. theoretical
- Basic vs. applied computer science research
- Impact of the research

What is Research?/1

- The word re-search is a noun composed of two syllables:
  - re is a prefix, meaning again, anew or over again
  - search is verb, meaning to examine closely and carefully, to test and try, or to probe
- Together they form a noun describing a careful, systematic, patient study and investigation in some field of kenowledge, undertaken to establish facts or principles.

What is Research?/2

- Diligent and systematic inquiry or investigation in an area, with the objective of discovering or revising facts, theories, applications. The goal is to discover and disseminate new knowledge.

  [Merriam-Webster]

- Systematic investigative process employed to increase or revise current knowledge by discovering new facts. It is divided into two general categories: (1) Basic research is inquiry aimed at increasing scientific knowledge, and (2) Applied research is effort aimed at using basic research for solving problems or developing new processes, products, or techniques.

  [Business Dictionary]

- Careful study of a given subject, field, or problem, undertaken to discover facts or principles.

  [The Free Dictionary]

What is Research?/3

- 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.

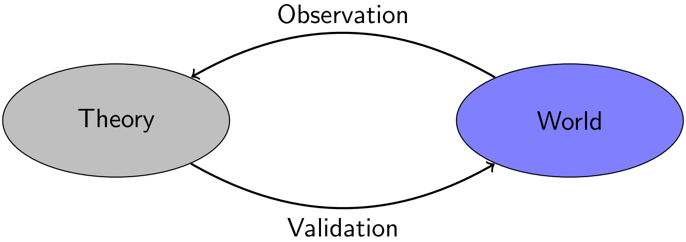
[Wikipedia]

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:



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 that doesn't need to be quite as empirically based as in other types of research.
- Nevertheless the conclusions have to be objectively argued and defined.
- This may involve evaluating the "construct" analytically against some predefined criteria or performing some benchmark tests with the prototype.

Empirical Research

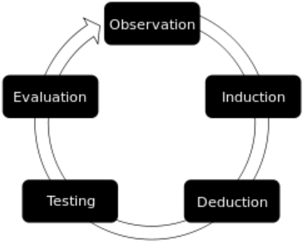
- “Empirical” comes from the Greek word for experience: ἐμπειρία (empeiría)
- Empirical research is a way of gaining knowledge by means of direct and indirect observation or experience.
- Empirical evidence/observations can be analyzed quantitatively or qualitatively.
- Through quantifying the evidence or making sense of it in qualitative form, a researcher can answer empirical questions, which should be clearly defined and answerable with the evidence collected (usually called data).
- Research design varies by field and by the question being investigated.
- A combination of qualitative and quantitative analysis is often used to better answer questions.

Empirical Research – Example

- Observation is the key: A way of gaining knowledge by direct observation or experience.
- Used to answer empirical questions, e.g., *"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.

Advantages of Empirical Research Methods

- 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

Improved Quality with Experimental 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!

Research Techniques

- Interpretivist or qualitative research techniques
- Research techniques at the scientific/interpretivist boundary
- Quantitaive and scientific research techniques
- 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.

Interpretivist/Qualitative Research Techniques/1

- Interpretivists work out people's interpretations of the world by putting themselves in their shoes, hence are subjective and biased.
- Assumption that people make own choices that are not connected to laws of science or nature.
- Research tends to be done in greater detail and looks at culture and how people live their lives.
- Science can explain people's actions but interpretivists don't just want descriptions, they want reasons why.
- Results will be personal and in depth, therefore cannot be necessarily generalised.
- Tends to undermine realiability and representativeness
- Interpretivists tend to involve emotions and bias in their views but, this may not always be beneficial as they may get in the way of what is really happening.

Interpretivist/Qualitative Research Techniques/2

- Descriptive/interpretive research: empirical observation is subjected to limited formal rigor. Controls over the researcher's intuition include self-examination of the researcher's own pre-suppositions and biases, cycles of additional data collection and analysis, and peer review;
- Focus group research: gathering of a group of people, commonly members of the public affected by a technology or application, to discuss a topic. Its purpose is to surface aspects, impacts and implications that are of concern.
- Action research: the researcher plays an active role in the object of study, e.g. by acting as a change-agent in relation to the process being researched.
- Ethnographic research: applies insights from social and cultural anthropology to the direct observation of behaviour.

Quantitative Methods

- Origin in the natural sciences – scientific method
- Systematic empirical investigation of quantitative properties and phenomena and their relationships.
- The goal is 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.

Computer Science

Shifting Definition of Computer Science/1

- *Computer Science is the study of phenomena related to computers*
  [Newell, Perlis and Simon, 1967]
- *Computer Science is the study and management of complexity*
  [Dijkstra, 1976]
- *The discipline of Computing is the systematic study of algorithmic processes that describe and transform information: their theory, analysis, design, efficiency, implementation, and application*
  [Denning, 1985]
- *Computer Science is the mechanization of abstraction*
  [Aho and Ullman, 1992]
- *Computer Science is a field of study that is concerned with theoretical and applied disciplines in the development and use of computers for information storage and processing, mathematics, logic, science, and many other areas*
  [Mahoney, 2001]

Shifting Definition of Computer Science/1

- *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]

Top Level of The ACM Computing Classification System (1998)

- A. General Literature
- B. Hardware
- C. Computer Systems Organization
- D. Software
- E. Data
- F. Theory of Computation
- G. Mathematics of Computing
- H. Information Systems
- I. Computing Methodologies
- J. Computer Applications
- K. Computing Milieux

(ACM = Association for Computing Machinery)

Top Level of The ACM Computing Classification System (2012)

- General and reference
- Hardware
- Computer systems organization
- Networks
- Software and its engineering
- Theory of computation
- Mathematics of computing
- Information systems
- Security and privacy
- Human-centered computing
- Computing methodologies
- Applied computing
- Social and professional topics

Changes in CS: the Role of Technology

- Much of the change that affects computer science comes from advances in technology:
  - The World Wide Web and its applications
  - Networking technologies and distributed systems
  - Graphics and multimedia
  - Embedded systems
  - Ubiquitous computing
  - New types of databases
  - Interoperability and data integration
  - Object-oriented programming
  - Human-computer interaction (new interfaces)
  - Software safety
  - Security and cryptography
  - Application domains

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).

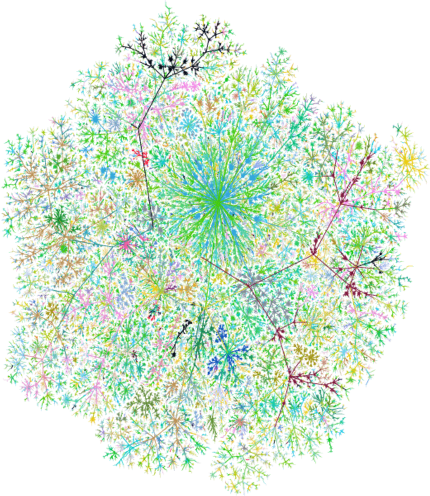
[Wegner, 1976]

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

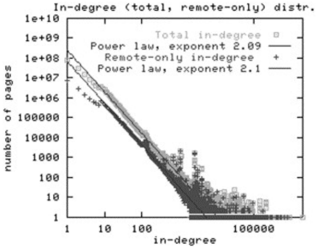
- Web does not have an engineered architecture: hundreds of billions of pages created by billions of users.
- Web contains a large strongly connected core (each page can reach every other).
- The shortest path from one page in the core to another involves 16–20 links (a small world).
- Analysis of web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

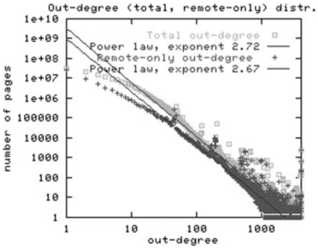
[Broder et al., 2000]

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

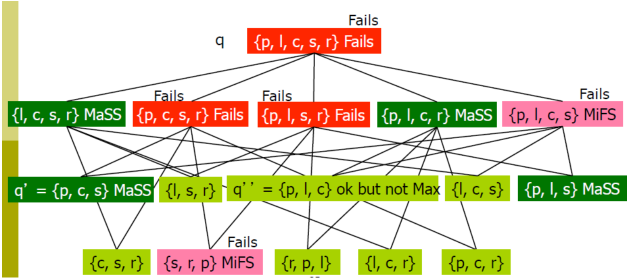
- How to deal with the problem of empty result set for Boolean queries, i.e., queries that contain a set of key-words and fail to return any item
- e.g., $q = \{prolog, language, comparison, survey, rating\}$ fails to retrieve any record (web page)

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 $q'' = \{prolog, language, comparison, rating\}$.

Formal Definition of the Problem

- Let q be a query with empty result size.
- Maximal succeeding subquery problem q': returns some results, and there is no other succeeding subquery q'' that contains q'
- Minimal failing subquery q^*: is a failing subquery of qbut any of its subqueries are succeeding



How to use them

- Minimal failing subqueries: give a "compact" reason of why the query is failing
  - The user can generate one succeeding subquery removing one constraint from each minimal failing subquery
  - e.g., $\{l, c, s, r\}$ (maximal succeeding subquery) is obtained by removing p from $\{p, l, c, s\}$ and $\{s, r, p\}$ (the two minimal failing queries)
  - Still difficult to find the query that relaxes less constraints as possible (search for the smallest set of conditions that make satisfiable all the minimal failing subqueries)
- Maximal succeeding subqueries: provide full description of all best relaxations.

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
- Maximal succeeding subquery problem
  - one of these succeeding subqueries can be found in $O(|q|)$
  - two in $O(|q|^2)$
  - all makes the problem intractable
- Minimal failing subquery problem
  - similar results as above

Engineering

- Building a robot for the new mission to Mars

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

My system is better . . .

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.

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

Example: all minimal failing subqueries

ShowMe: Please enter your query:

User: price ≤ 1000, month = august, region = ireland, persons = 2, duration = 14, type = skiing, accom = flat, transport = plane

ShowMe: There are no matches for the following combinations of constraints in your query:

region — price ≤ 1000, region = ireland
type — month = august, type = skiing
region — region = ireland, persons = 2
region = ireland, type = skiing
region = ireland, transport = plane
type = skiing, transport = plane
transport — accom = flat, transport = plane
price ≤ 1000, month = august, transport = plane
price ≤ 1000, persons = 2, transport = plane
price ≤ 1000, duration = 14, transport = plane
type — price ≤ 1000, duration = 14, type = skiing, accom = flat

To solve this problem, you need to relax one of the constraints in each of the unmatched combinations

By relaxing *transport* you can eliminate 6 of the unmatched combinations

[McSherry, 2004]

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 (ECS) refers to the building of, or the experimentation with or on, 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
- "Theory" in CS is very close to mathematics – theoreticians prove theorems
- 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
- The complexity of the systems built in ECS and of the underlying models and theories means that experimental implementation is **necessary** to evaluate the ideas and the models or theories behind them.

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.

R. Clarke's Quality Characteristics

- Research should reflect the state of knowledge of the domain, at the time the project is commenced (in order to advance knowledge).
- Research should reflect the state of knowledge of research methodology, at the time the project is commenced (in order to advance knowledge).
- Research should combine research techniques in such a manner that the weaknesses of each are complemented by the strengths of the others (in order to contribute to rigour).
- Research should produce data that reflect the phenomena under study. For scientific research, these need to be subjected to validation testing, and to be submitted to powerful statistical techniques in order to tease out the relationships among the variables (in order to contribute to rigour).
- Research should be practicable (in order to avoid wastage of resources).
- Research should produce results relevant to the world (in order to address the interests of organisations which use the data and provide the funding).
- Research should be likely to be publishable (in order to satisfy the interests of the researcher and their sponsor);
- Research should be ambitious (in order to drive theory and practice forward).

[Clarke, 2000]

How to do?

The Three Golden Rules

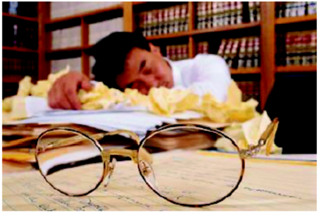
- Raise your quality standards as high as you can live with and always try to work as closely possible at the boundary of your abilities.
- If you can find a topic that is socially relevant and scientifically sound you are lucky: if the two targets are in conflict let the requirement of scientific soundness prevail.
- Never tackle a problem of which you can be pretty sure that it will be tackled by others who are, in relation to that problem, at least as competent and well-equipped as you.

[Dijkstra, 1982]

Impact – the Criterion of Success

- The fundamental basis for academic achievement is the impact of one's ideas and scholarship on the field
- Dimension of impact:
  - Who is affected by a result
  - The form of the impact
  - The magnitude of the impact
  - The significance of the impact

Who is affected



Other researchers

Engineers and Practitioners

Users

The form of the impact

- The contribution may be used directly or be the foundation for some other artifact
- It may help other to understand better a topic or a question
- It may change how others conduct their research
- It may affect the questions they ask or the topics they choose to study
- It may even indicate the impossibility of certain goals and kill off lines of research.

Magnitude and Significance

- Assessing the magnitude and significance of the impact is done observing "indicators"
- The number of quotations
- The quality of the journal/conferences that published the result (acceptance rate – impact factor)
- The role taken by the researcher in the scientific community (e.g. conference program chair)
- The patents
- The amount of money collected by the result (projects, consultancy, products)
- The quality and quantity of the scientific connections (collaborations)

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
```

Example Bash Script to Run all Experiments

```
#!/bin/bash

login=gamper/gamper
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

...
```

How are Experiments Done - A Small Case Study/1

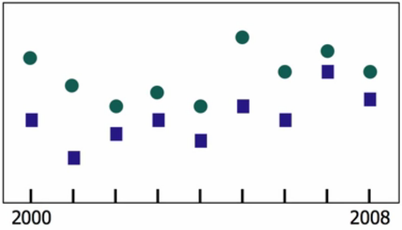
- Armstrong et al.: *Improvements That Don't Add Up: Ad-Hoc Retrieval Results Since 1998*, CIKM 2009
  - The existence and use of standard test collections in information retrieval experimentation allows results to be compared between research groups and over time. However, such comparisons are rarely 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/07Qgo6RSzmE?t=19m`

– Even worse: published papers all claim improvements

But they all compare to weak baselines



106 papers from
SIGIR / CIKM 2000 – 2008
for each year, best result
on TREC-8 adhoc benchmark

● New result
■ Baseline

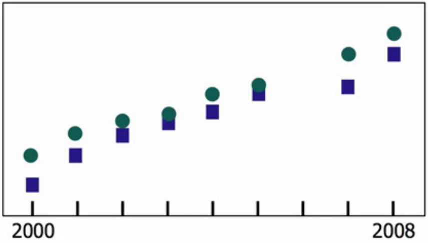
"Improvements that don't add up: adhoc results since 1998", CIKM 2009
(data points manually copied from Figure 4a in that paper)

[Jens Dittrich, 2015]

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

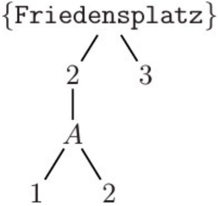
(data points manually copied from Figure 4a in that paper)

[Jens Dittrich, 2015]

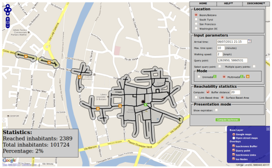
Working with Real Data is Rewarding – eBZ Project

- 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

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

- Wu et al.: *Shortest path and distance queryies on road networks: an experimental evaluation*, VLDB 2012
- 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

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.