Freie Universität Berlin

Course "Empirical Evaluation in Informatics"

The Scientific Method

Prof. Dr. Lutz Prechelt
Freie Universität Berlin, Institut für Informatik
http://www.inf.fu-berlin.de/inst/ag-se/

- Science and insight
- Informatics on the landscape of sciences
- The scientific method
- Variables, hypotheses, control
- Internal and external validity
- Validity, credibility, and relevance

Freie Universität Berlin

"Empirische Bewertung in der Informatik"

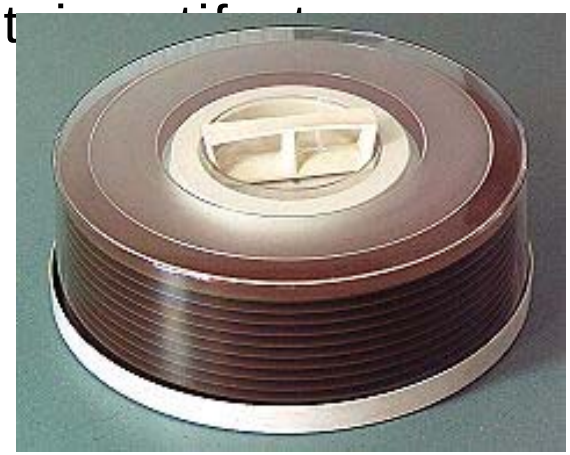
Die wissenschaftliche Methode

Prof. Dr. Lutz Prechelt
Freie Universität Berlin, Institut für Informatik
http://www.inf.fu-berlin.de/inst/ag-se/

- Wissenschaft und Erkenntnismethoden
- Einordnung der Informatik
- Die wissenschaftliche Methode
- Variablen, Hypothesen, Kontrolle
- Interne und externe Gültigkeit
- Gültigkeit, Glaubwürdigkeit und Relevanz

Our goal

- In empirical evaluation, we have given a certain artifact in a situation, e.g.
  - - a new (or old) design method or
  - - a new kind of hard disk, etc.



- and want to obtain an understanding of it
  - often with respect to specific attributes, e.g.
  - - the effort for accomodating later requirements changes
  - - or the bandwith and latency of data transfer to/from the disk

Obtaining understanding

Freie Universität Berlin

- There are different ways how people obtain understanding
  - by intuition (direct insight)
  - from some authority (tradition, teacher, book etc.)
  - by rational thought (reasoning, deduction)
  - by direct observation combined with induction
  - via the scientific method

- Each method can produce valid understanding

- No method can make totally sure that the understanding is valid
  - but the scientific method comes closest
  - and, just as importantly, has the best chance of convincing other people to accept the same understanding

The landscape of knowledge and science

Freie Universität Berlin

- The arts
  - "Geisteswissenschaften"
  - Special case: Mathematics
    - pure logic: principles of deduction are fixed, anything else is arbitrary
- The (natural) sciences
  - "Naturwissenschaften"
  - examines characteristics and behavior of the real world
- Special case: the social sciences
  - "Sozialwissenschaften"
  - examines human behavior
- Engineering
  - "Ingenieurwissenschaften"
  - solves practical problems; interested in usefulness and cost

The landscape and T, C, E

Freie Universität Berlin

T, C, E: Theory, Construction, Empiricism

- Mathematics
  - Mostly theory
  - Auxiliary C and E have entered recently (computational math.)
- The (natural) sciences
  - Theory and empiricism fertilize each other
  - Construction is purely auxiliary
- The social sciences
  - Empiricism drives Theory
  - Construction is purely auxiliary (at least mostly, at least today)
- Engineering
  - Theory, construction, and empiricism fertilize each other
  - Much theory is borrowed from the natural sciences
  - Construction is the goal

Informatics on the landscape

Freie Universität Berlin

- Informatics has its roots in
  - Mathematics: logic, formal languages
  - (Electrical) Engineering: constructing computers
- Today, the larger part is clearly engineering
  - (In this course, we look at this part only)
- However, the engineering is not purely technical:
  - The artifacts have to be used by people
  - Brings psychology, sociology, and politics into play
- Hence, Informatics needs a lot of empiricism

Mathematics vs. natural science

Freie Universität Berlin

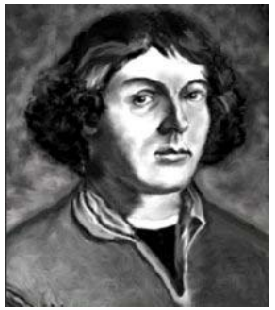
- Historically, all of science was philosophy
  - at least in the western culture
  - Greek philosophers
- and much of that was mathematics

- The notion that nature could be understood by pure thought (rationalism) was prevalent in the middle ages

- The idea that observation and experimentation was necessary to understand the world began to get accepted during the renaissance

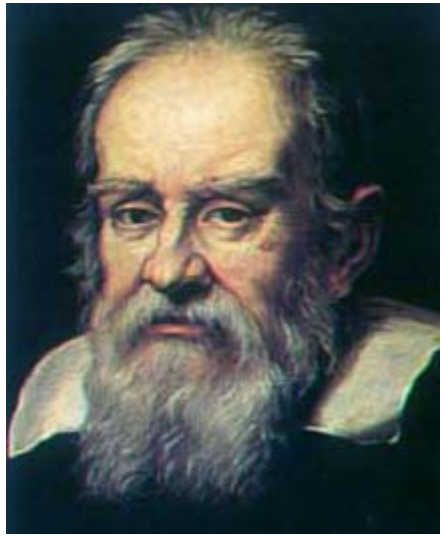
Early empiricists

- Some of the earliest modern empiricists were the astronomers Kopernikus, Brahe, and Galilei
  - around 1500–1600

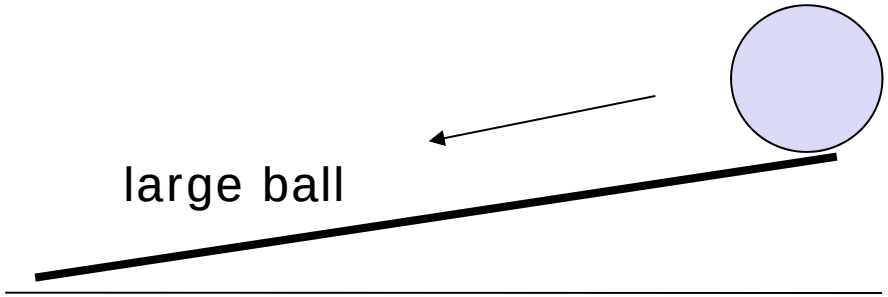

Kopernikus


Brahe

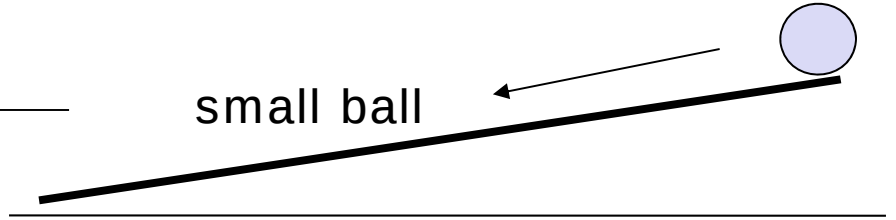
- One of the first modern experimental scientists was Galileo Galilei
  - At the time, it was generally accepted that heavy objects fell down faster than lighter ones
    - as claimed by Aristotle (384–322 BC)
  - Galilei did not believe this and experimented with brass spheres, inclined planes, and water clocks (1589–1604)
    - He systematically varied the weight of the ball and the steepness of the plane and found weight-independent acceleration
    - These were controlled experiments

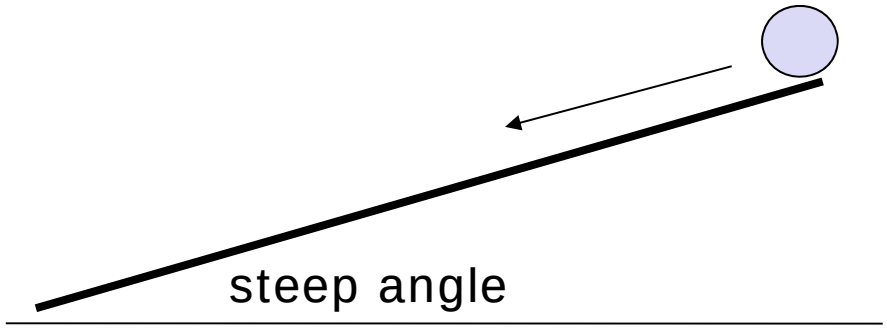

Galilei

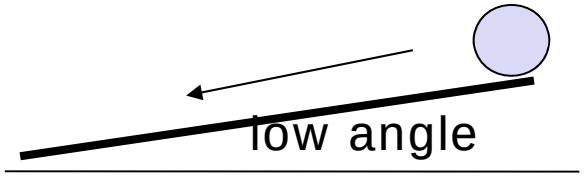
Galilei's experiments



- Weight of the sphere is <u>not</u> relevant







The scientific method

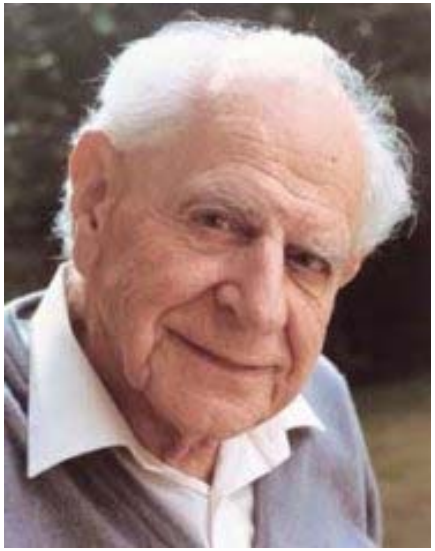
- Since Galilei, physics and other sciences work according to this model:
  - Formulate a theory T about how (some aspect of) the real world behaves
  - Design and conduct experiments E for testing this theory
- Is accepted in all subjects where experimentation is possible
  - Natural sciences: Physics, chemistry, biology, medicine etc.
  - Engineering
  - Parts of many social sciences (such as economics, sociology, etc.)
- Is problematic where experiments cannot be performed
  - for technical or ethical reasons

The scientific method (2)

Freie Universität Berlin

- Note the following:
  - T is called a scientific theory only if it predicts something specifically and hence can be tested
  - Even if T is wrong, it may happen that the results of E are as expected
  - But if E contradicts predictions of T, then T must be false

- This view of science was suggested by Karl Popper (1904–1994)
  - It is the prevalent scientific paradigm today
  - In this view, theories cannot be directly confirmed, only refuted
  - If a theory cannot be refuted for a long time, it will gradually be accepted as confirmed
    - example: special theory of relativity



Pre-theoretical empiricism

Freie Universität Berlin

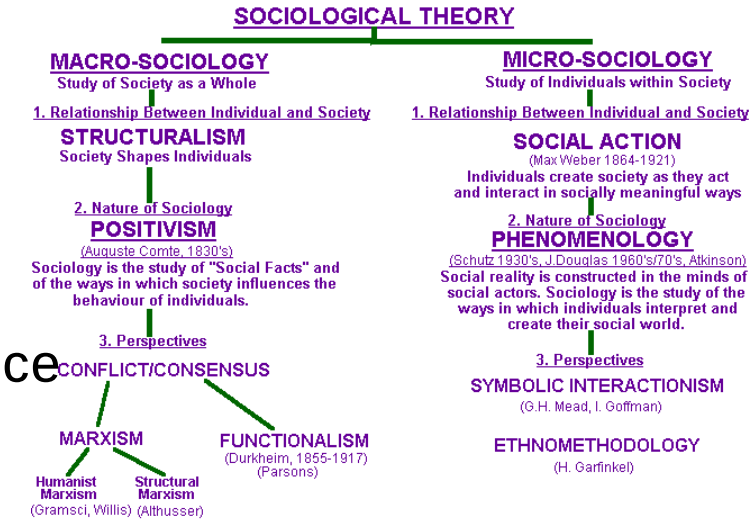
- In many areas, too little is known for formulating a plausible, testable theory
  - Often true where people are involved and the situation is complex
    - such as in software engineering
- Even then empiricism is useful:
  - Observe things that lead to hypotheses from which one could build theories
  - Often these observations have to be qualitative rather than quantitative in order to be useful
    - Qualitative research is a large and interesting branch of research methodology
    - but not the topic of this course (half-exception: Case Studies)

Hard science vs. soft science

- Many people claim that a subject is a science only if it produces theories that are precise and reliable
  - "hard science", such as physics formulas

- and hence claim that subjects involving human behavior are not scientific ("physics envy")
  - "soft science"

- This is not true: The scientifc principle can be applied
  - but the theories will be more complex and make weaker (e.g. probabilistic) predictions
- Hard science is simpler than soft science
  - That is why it is farther advanced



Terminology of Empiricism

Freie Universität Berlin

- When we empirically investigate something
  - we characterize the situation by a set of ***input variables***
    - usually quantitative or categorial
    - e.g. "team size = 4" or "design method used = A"
  - and the observations by a set of ***output variables***
  - If we choose the value of at least one input variable, the study is called an ***experiment***

- The act of consciously manipulating the values of input variables is called ***control***

- Every empirical study assumes that there is some systematic relationship between inputs and outputs
  - If we have a certain expectation about this relationship, this is called a ***hypothesis***
  - Any additional factors influencing the outputs are called ***extraneous variables***

Example: Case study

Freie Universität Berlin

- Assume we want to evalute a design method A
- We pick a representative team of people
  - a capable, but not unrealistically clever team
- We pick a task of interest
  - a "normal" one: not unusually small or large or difficult or ...
- We have them do the design using method A
  - (hopefully they receive some training beforehands...)
- We see what happens (using many sources of observations):
  - What goes well?
  - What goes not so well?
  - How good is the resulting design?

Control in the case study

Freie Universität Berlin

- This case study has little control
  - We have controlled the task to be done and the method to be used
    - (and even this is unusual for a case study)
  - but not the capabilities of the people
    - Precisely how intelligent, knowledgable, interested etc. are they?
  - Worse, we cannot judge the results without comparing them to other results
- Hence, it is not so clear what the results mean

Example: Controlled experiment

Freie Universität Berlin

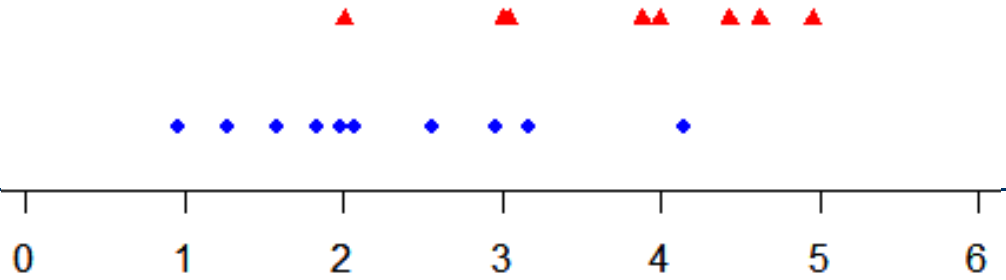
- This time, we compare design methods A and B
- Again, we pick a task T and a set of people P
  - but this time a large set of people
  - we train all of them equally well in both methods
- But now we use separate teams working with A or with B
- and have 20 different teams solve T with each method
  - People are assigned to the teams at random
- We compare the average result obtained by the method A teams and method B teams

Control in the controlled experiment

Freie Universität Berlin

- This time we have controlled all variables:
  - task and method as before
  - the comparison to method B allows for interpreting the results
  - replication turns all kinds of individual differences into a noise signal
    - we will get different results for different teams although they are using the same condition
    - but given enough teams, the differences cancel out
  - random group assignment avoids systematic accumulations of individual differences
    - e.g. if more capable people favor working with method A

- Hence, we can decide whether A works better than B
  - at least for this kind of people, in this setting, and for this task



Internal and external validity

Freie Universität Berlin

- Internal validity
  - the degree to which the observed results were caused by only the intended input variables
  - rather than extraneous variables
- External validity
  - the degree to which the results can be generalized to other circumstances
    - in our example: other people, settings, and tasks
- Improving external validity tends to reduce internal validity
  - because it will strengthen the influence of extraneous variables

Threats to internal validity

Freie Universität Berlin

- Have all plausible extraneous variables been controlled completely?
- Has the act of observing influenced the observations?
- Are the results that are compared really comparable?

A related concept is *construct validity*:

- Do my measurements really represent the characteristic that I want to observe?
  - e.g. does the number of pages of a design document really represent the size of a design task?

Threats to external validity

Freie Universität Berlin

- The results rely on specific characteristics of the task
  - and these are uncommon
  - e.g. task is unusually well suited for method A, but not for B
- The results rely on specific characteristics of the people
  - and these are uncommon
  - e.g. they have an unrealistically good understanding of the ideas of method A, because they were thoroughly taught by its inventor

- The results rely on specific characteristics of the experimental setting
  - and these are uncommon
  - e.g. the subjects were enthusiastic about A, but not B.

Credibility, relevance, validity

Freie Universität Berlin

- Credibility is achieved when
  - there is high internal validity
  - there is a reasonable amount of external validity
    - in particular: no bias of the task
  - there is no doubt that both is the case

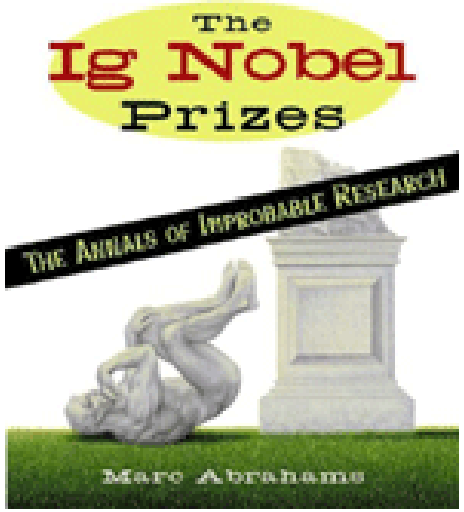
- Relevance is achieved when
  - the question investigated is of general interest and
  - there is high external validity

Judging empirical results

- Some fraction of the empirical results in scientific publications is dubious or even plain wrong
- Outside of science, this is even much worse

- How can we discriminate valid results from dubious ones?
- The following questions help:
  - How do they know this?
    - in particular: Are the conclusions warranted by the facts?
  - What has not been said (but should have)?
  - Is this information really relevant?

(More about this in the next lecture)



Summary

Freie Universität Berlin

- Our goal is insight into objective facts and relationships
- The most powerful method for this is the scientific method:
  - Formulate a theory, derive hypotheses
  - Test them by experiments
    - Can only refute the theory, not prove it!
- It is accepted wherever experiments are possible
  - and can be approximated in many further settings
  - In Informatics, control in the experiments is often incomplete
- The goal is high internal and external validity
  - because they are key to good credibility and relevance
- Results should be judged by these criteria

Thank you!