INF

Faculty of
Computer Science

Experiments

Janet Feigenspan

Examples

Watching stars to detect planets

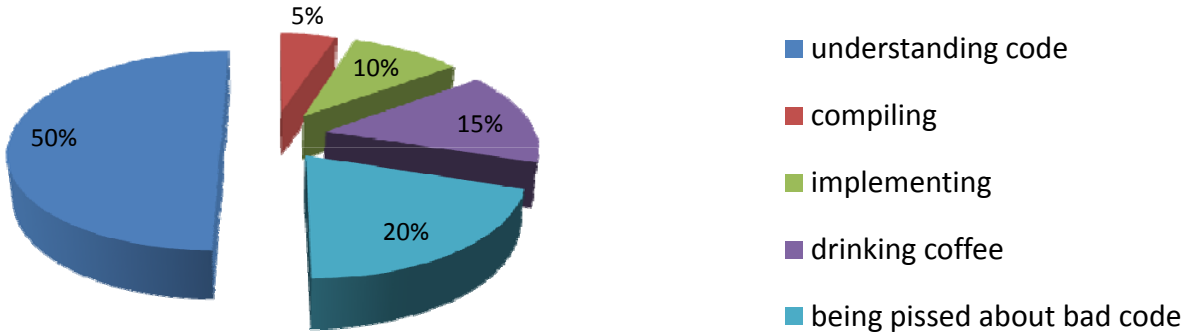
Evaluate efficiency of new warp drive

Compare two teaching methods

What are Experiments?

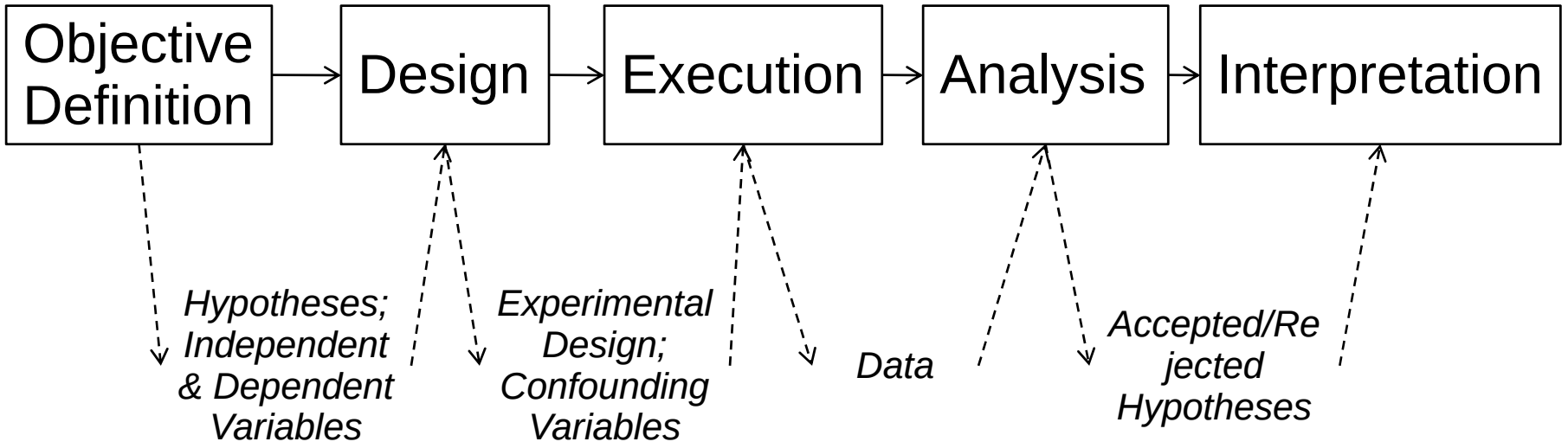
- systematic research study
- one or more factors intentionally varied
- everthing else held constant
- result of systematic variation is observed
- here: human participants
  - Wilhelm Wundt. *Grundzüge der Physiologischen Psychologie*. Engelmann, 1874.
  - Wilhelm Wundt. *Grundriß der Psychologie*. Kröner, 1914.

Why Experiments?



- In general: Human Factors

Stages of Experiments



Outline

- discuss each stage with a running example
- discuss problems and solutions

- Goal:
  - get a feeling for design of experiments
  - know what to do with data

//Comments in Source Code

- Do they make code more comprehensible?
- Do they make code more maintainable?
- Do they increase development time?
- Do they reduce maintenance costs?

Objective Definition

Independent Variable

- factor, predictor (variable)
- intentionally varied
- influences dependent variable

- comments

Operationalization

- levels, alternatives

- presence/absence of comments
- good comments
- bad comments
- useless comments

Dependent variable

- response variable
- outcome of experiment

- Program comprehension
- Maintainability
- Development time

Operationalization

- Specify a measure

- Program comprehension:
  - subjective rating
  - solutions to tasks (correctness? response time?)
  - think aloud

Hypotheses

- Expectations about outcome
- Based on theory or practice -> expectations must have reason

Hypotheses - Example

- Bad comments bad for program comprehension
- good comments good for program comprehension

Good/Bad Hypotheses

- what are good/bad comments?
- what does good/bad for program comprehension mean? -> slower, more errors? by how much?

- Hypothesis must be falsifiable

  - Karl Popper. *The Logic of Scientific Discovery*. Routledge, 1959.

Better Hypotheses

- comments describing each statement of source code have no effect on the response time of understanding source code
- comments containing wrong information about statements slow down comprehension
- comments describing the purpose of statements speed up comprehension

Why Hypotheses?

- why not just measure and see what the result is?
  - influences experimental design
  - fishing for results

Summary

- Independent Variable
  - Comment
  - two levels: presence/absence
- Dependent Variable
  - Program Comprehension
  - Response time for tasks
- Hypotheses
  - comments describing the purpose of statements speed up comprehension

Design

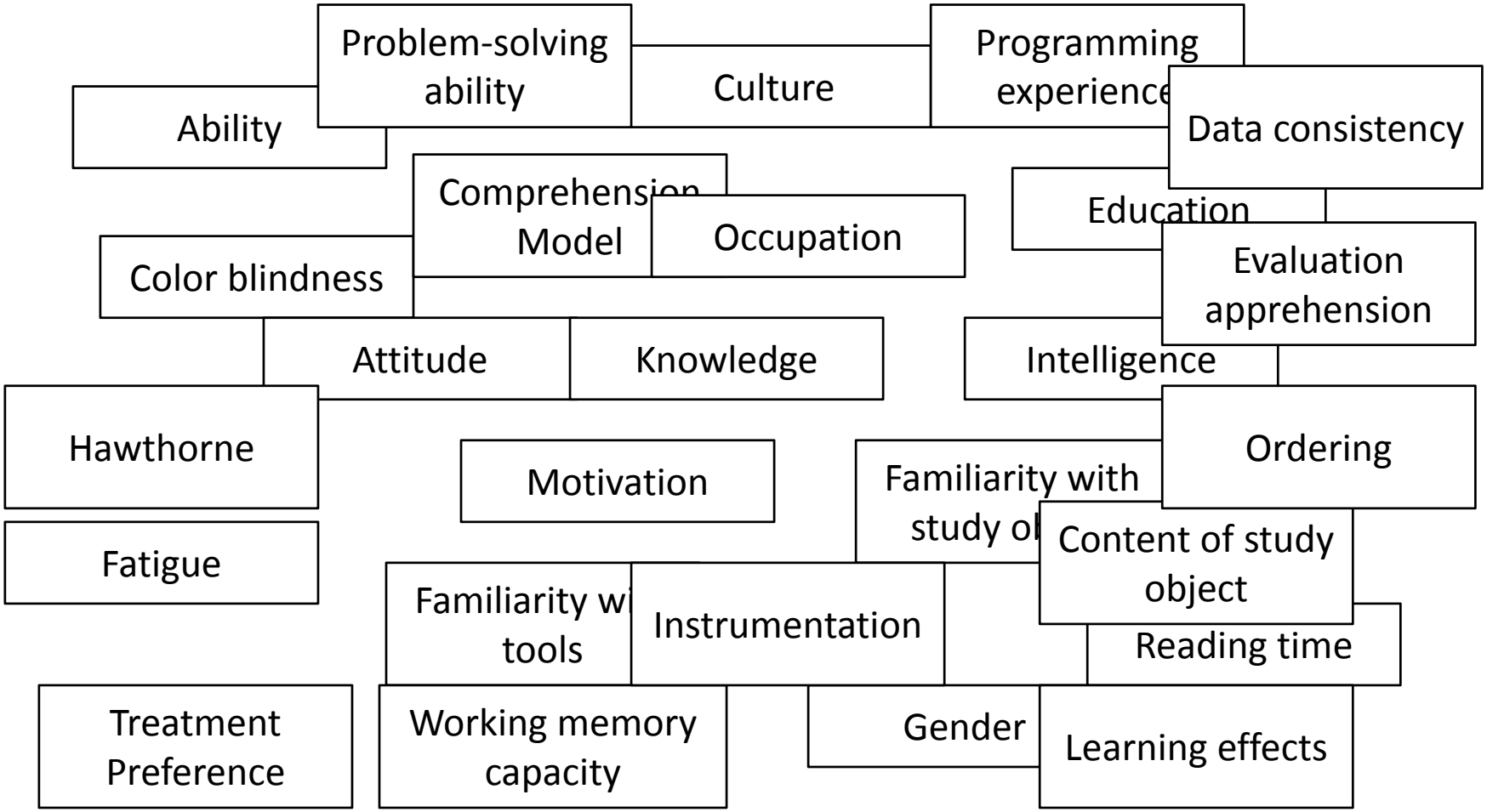
Validity

- Do we measure what we want to measure?
- Internal:
  - degree to which the value of the dependent variable can be assigned to the manipulation of the independent variable
- External:
  - degree to which the results gained in one experiment can be generalized to other subjects and settings

Confounding Parameters

- influence depending variable besides variations of independent variable

Confounding Parameters



Controlling Confounding Variables

- Randomization
- Matching
- Keep confounding parameter constant
- Use confounding parameter as independent variable
- Analyze influence of confounding parameter on result

Randomization

- use random number generator
- roll a dice
- toss a coin
- ...

Matching

Proband	Wert
P5	65
P9	56
P3	42
P4	34
P10	24
P6	23
P7	21
P8	16
P2	12
P1	5

Group A	Group B
65	56
34	42
24	23
16	21
12	6

odd-even-even-odd

Matching

- We have to measure a confounding parameter
- Programming experience
  - questionnaire, pretest, number of years, size of projects,...
- Intelligence
  - test (exhausting, more time), university grades as indicator,...

Keep Parameter Constant

- Programming experience
  - recruit students as participants (bachelor, master, PhD)
  - recruit programming experts
- Intelligence
  - only participants with certain grades

Use parameter as Independent Variable

- Reminder: 2 level of independent variable (comment/no comment)
- Example: 2 levels of programming experience
  - Comment/low experience
  - Comment/high experience
  - No comment/low experience
  - No comment/high experience

Analyze Influence of Parameter on Result

- when we cannot assign participants to groups, for example when comparing two companies
- when something happened during the experiment, e.g., power failure in one session, but not in an other session

Which control technique is the best?

Randomization	Large sample
Matching	Measurement
Keep Constant	Measurement
Use as factor	Measurement, large sample
Analyze afterwards	Measurement

Validity

- Internal and external validity need different things:
  - internal: controlling everything
  - external: broad setting so that we can generalize

- first maximize internal validity
- step by step increase external validity

Experimental Designs

- One-factorial designs

Group	Levels
A	Comment
B	No comment

- comparable groups

One Group	Session 1	Session 2
	Comment	No Comment

- ordering effects

Group	Session 1	Session 2
A	Comment	No Comment
B	No comment	Comment

- learning effects
- mortality

Experimental Designs

- Two-factorial designs

Group	Session 1	Session 2	Session 3	Session 4
Comment/ Low Experience	Group A	Group D	Group C	Group B
Comment/ High Experience	Group B	Group A	Group D	Group C
No comment/Low Experience	Group C	Group B	Group A	Group D
No comment/High Experience	Group D	Group C	Group B	Group A

Execution

What can go wrong?

Everything!

- conduct pilot tests
  - test material
  - tools
  - data storage
- tell participants **exactly** what they have to do
- observe that participants do what they are instructed to do
- make backups of the data

Analysis

Experiment

```
public static void main(String[] args) {
   String word = "Hello";
   String result = new String();
   for (int j = word.length() - 1; j >= 0; j--)
     result = result + word.charAt(j);
   System.out.println(result);
}

public static void main(String[] args) {
   String word = "Hello";
   String result = new String();
   //reverse character order
   for (int j = word.length() - 1; j >= 0; j--)
     result = result + word.charAt(j);
   System.out.println(result);
}
```

Group	Time [s]
A (no comment)	42
A	60
A	30
A	77
A	58
A	49
A	38
B (comment)	48
B	48
B	26
B	30
B	50
B	34

Descriptive Statistics

- what do we do with these data?
- look at the data
  - mean/average (=arithmetic mean)
  - median
  - standard deviation
  - boxplots

Median

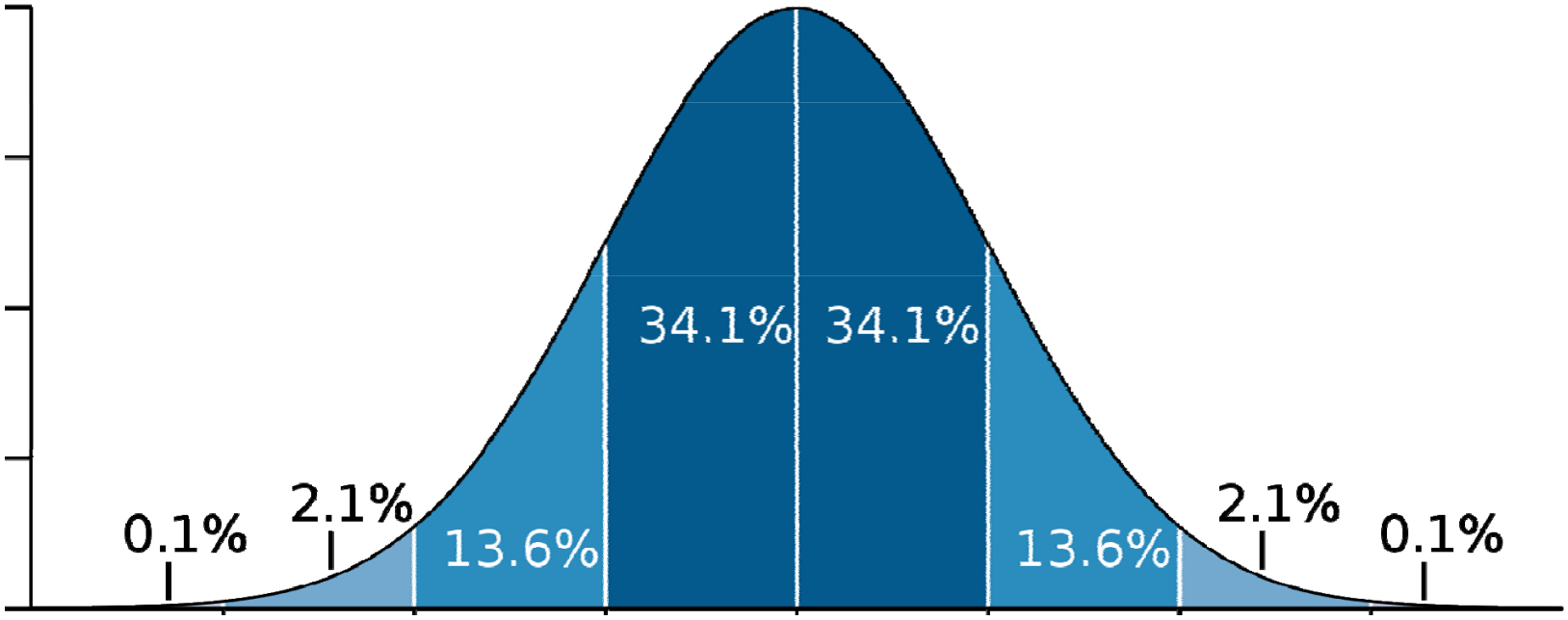
Group	Time [s]	Time [s]
A	42	30
A	60	38
A	30	42
A	77	49
A	58	58
A	49	60
A	38	77

Median: 49

Group	Time [s]	Time [s]
B	48	26
B	48	30
B	26	34
B	30	48
B	50	48
B	34	50

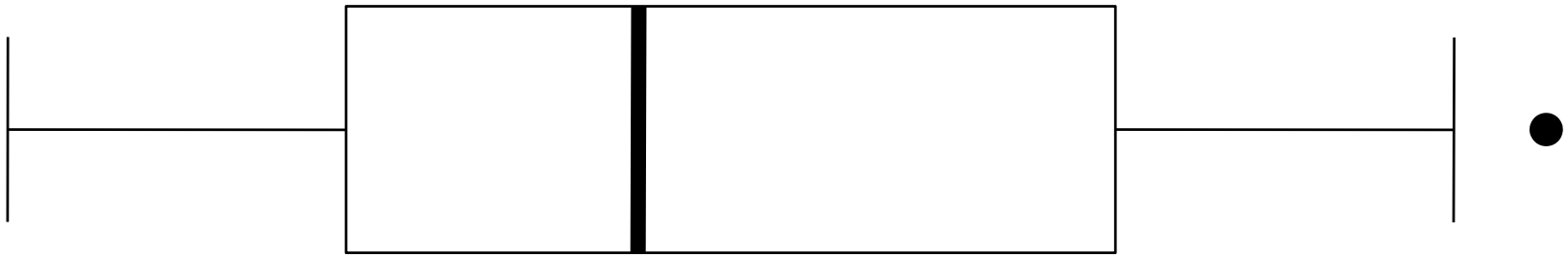
Median: (34 + 48)/2 = 41

Standard Deviation



Group A:	19	35	51	67	83
Group B:	17	28	39	50	61

Boxplot



- box: 50% of all values
- line: median
- whiskers: upper and lower 25% of data
- dot:
  - outlier (=values that deviate too much from mean/median)
  - what is too much?
  - 1.5/2 standard deviations

Statistical Tests

When is a difference real, not coincidental?

-Assumption: both values are the same (= null hypothesis; H_0)

-Conditional probability: probability of observed result under assumption that values should be the same

-if probability is low, then assumption must be wrong

- typical: 5%
- possible: 10%

T Test

- interesting values:
  - p value: smaller/larger than 0.05?
  - (t value/degrees of freedom-df): when you report the test

- p value > 0.05? -> no significant difference
- p value <= 0.05? -> significant difference

Interpretation of T Test

- we reject hypothesis, that comments speed up comprehension

- in case p value is <= 0.05
  - we **did not confirm** hypothesis
  - we just did not find any evidence against it
  - hence: we do not say that we confirmed a hypothesis, but that we can accept it
  - (or even more correct: we can reject the null hypothesis)

Preconditions t test

- normally distributed data
- metric data

- scale types
  - metric data (e.g., response times)
  - ordinal data (e.g., rankings, grades)
  - nominal data (e.g., gender, party members)

Further tests

- Mann-Whitney-U test (non-parametric test)
  - ordinal scale type
  - metric scale type, but not normally distributed

- χ^2-Test
  - nominal scale type

Interpretation

Are comments any good?

- look at the size of the difference
- did anything noteworthy happened during execution?
- comments of participants?
- what does that result mean for practice?

Ethics

- Be nice to your participants, they voluntarily invest their time for you
- Assure anonymity
- Assure that benefit for science is worth the effort for participants

- Urban Wiesing. *Die Ethik-Kommissionen – Neuere Entwicklungen und Richtlinien*. Deutscher Ärzte-Verlag, 2003.

And now?

- You might feel frustrated
- You might think that there is no way to create an absolutely waterproof experiment design
- That is correct, there is no perfect design
- Accept that every experiment has flaws, it is unavoidable
- Do not look for a perfect design, look for a good, sufficient design to evaluate your hypotheses

Literature

- Historical
  - Wilhelm Wundt. *Grundzüge der Physiologischen Psychologie*. Engelmann, 1874.
  - Wilhelm Wundt. *Grundriß der Psychologie*. Kröner, 1914.
  - Karl Popper. *The Logic of Scientific Discovery*. Routledge, 1959
- Experiment Design
  - Jutta Markgraf, Hans-Peter Musahl, Friedrich Wilkening, Karin Wilkening, and Viktor Sarris. *Studieneinheit Versuchsplanung*, 2001. FIM-Psychologie Modellversuch, Universität Erlangen-Nürnberg.

Literature

- Experimentation in general:
  - C. James Goodwin. *Research in Psychology: Methods and Design*. Wiley Publishing, Inc., 1999.
  - Claes Wohlin. *Experimentation in Software Engineering: An Introduction*. Kluwer Academic Publishers, 2000.

Literature

- Analysis
  - Theodore W. Anderson and Jeremy D. Finn. *The New Statistical Analysis Analysis of Data*. Springer, 1996
  - Jürgen Bortz. *Statistik für Human- und Sozialwissenschaftler*. Springer, 2005.