TDDD89

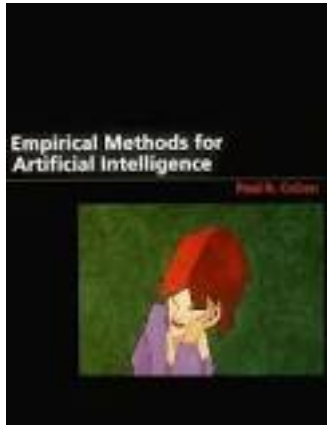
Lecture 4 - Research methods

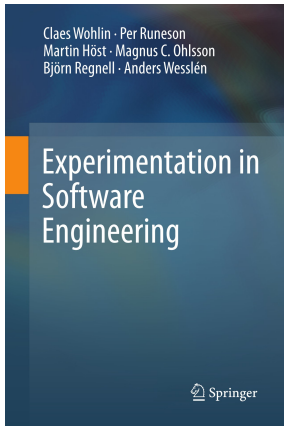
Ola Leifler

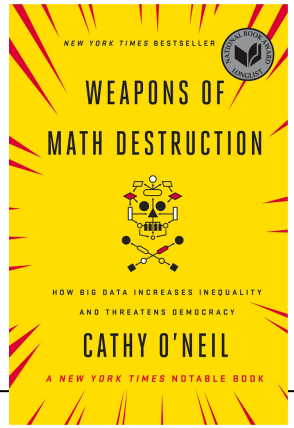
LINKÖPINGS UNIVERSITET

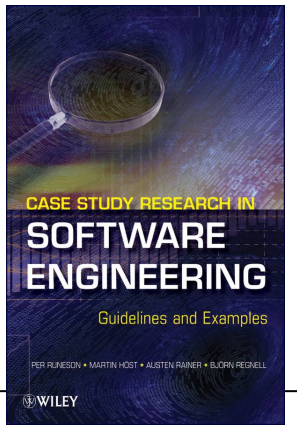
Literature

- Cohen, Paul. Empirical Methods in Artificial Intelligence
- Experimentation in Software Engineering
- Case Study Research in Software Engineering
- Weapons of Math Destruction









What is a scientific method?

- Design, implement, test?
- Acquire data, aggregate, visualise?
- …

Different types of methods

- Qualitative methods: establish concepts, describe a phenomenon, find a vocabulary, create a model
- Quantitative methods: make statistical analyses, quantify correlations, ..

Human-Centered methods

- Surveys
- Interviews
- Observations
- Think-aloud sessions
- Competitor analysis
- Usability evaluation
- …

Method choice?

- What do you want to find more about?
  - Identify the stakeholders (users, customers, and purchaser)
  - Identify their needs

Interviews

- Structured or unstructured?
- Group interviews (focus groups) or individual interviews?
- Telephone interviews

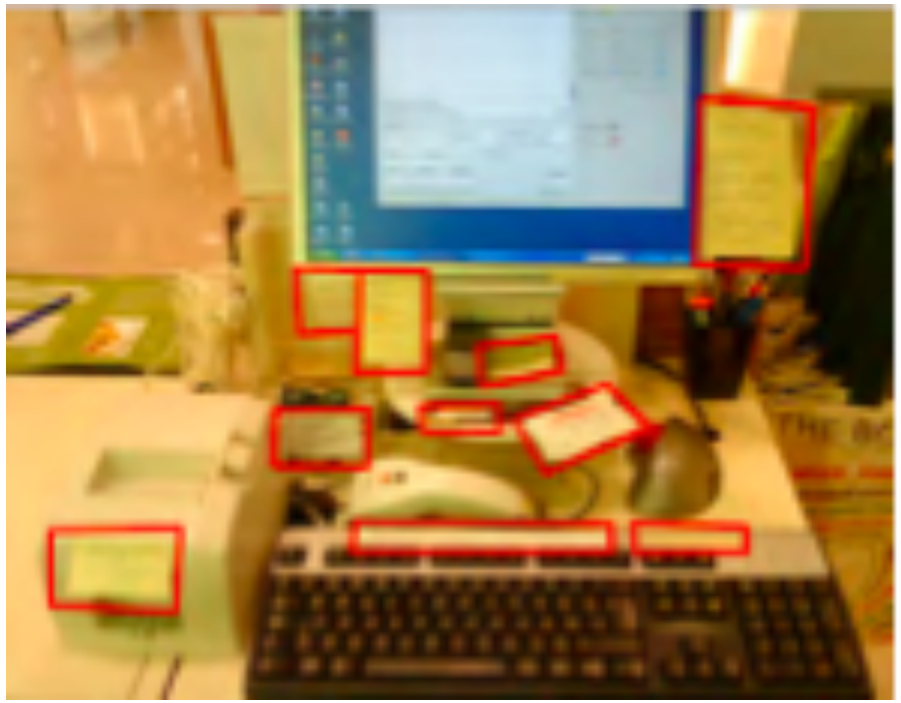
- Use open-ended questions:
  - ”Do you like your job?” vs ”What do you think about your job?"
- Active listning
- Record the interview
- Plan and schedule for that!

Interview analysis

- Transcribe or not?
  - Categorize what has been said (encode)

Observations

- Understand the context
- Write down what you see, hear, and feel
- Take pictures
- Combine with interview
- Ask users to use systems if availabe



Usability evaluation

- System usability scale (SUS)
- Post-Study System Usability Questionnaire (PSSUQ)
- Heuristic evaluations
- Eye tracking
  - First click Testing
  - …

System usability scale (SUS)

Note the differences

		Strongly Disagree				Strongly Agree
1.	I think that I would like to use this website frequently.	☐	☐	☐	☐	☐
2.	I found this website unnecessarily complex.	☐	☐	☐	☐	☐
3.	I thought this website was easy to use.	☐	☐	☐	☐	☐
4.	I think that I would need assistance to be able to use this website.	☐	☐	☐	☐	☐
5.	I found the various functions in this website were well integrated.	☐	☐	☐	☐	☐
6.	I thought there was too much inconsistency in this website.	☐	☐	☐	☐	☐
7.	I would imagine that most people would learn to use this website very quickly.	☐	☐	☐	☐	☐
8.	I found this website very cumbersome/awkward to use.	☐	☐	☐	☐	☐
9.	I felt very confident using this website.	☐	☐	☐	☐	☐
10.	I needed to learn a lot of things before I could get going with this website.	☐	☐	☐	☐	☐

Usability performance measurement

- Task success
- Time (time/task)
- Effectiveness (errors/task)
  - Efficiency (operations/task)
  - Learnability (performance change)

Describing a method

- ”To implement a Flux controller, I first needed to learn about Flux”

Don’t write a diary!

Write that which convinces someone you have done a good job

”The Flux controller was evaluated using the Flux controller evaluation protocol [1]”

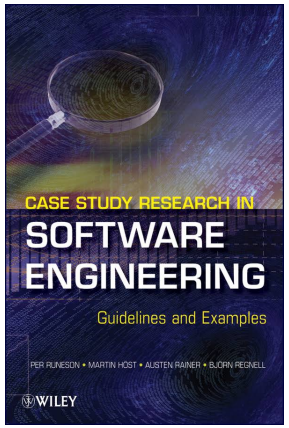
Engineering method vs scientific method

Method questions	Engineering aspect	Scientific aspect
Can I trust your work?	Have you properly tested your solution?	Have you verified that you obtain the same data in different settings/scenarios?
Can I build on your work?	Can I run/create the same system somewhere else?	Can I replicate the results of the study?

Case Study

- Investigates a phenomenon in a context,
- with multiple sources of information,
- where the boundary between context and phenomenon may be unclear
  - Uses predominantly qualitative methods to study a phenomenon

P. Runeson and M. Höst, “Guidelines for conducting and reporting case study research in software engineering,” Empirical Softw. Engg., vol. 14, pp. 131–164, Apr. 2009.



Experimental study design

Experiment idea

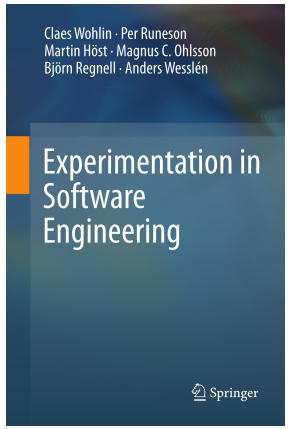
Experiment goal

Experiment planning

Hypothesis

Experiment operation

Experiment analysis



C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, Experimentation in Software Engineering. Springer Berlin Heidelberg, 2012.

Experiment goal

Analyze <Object>
for the purpose of <Purpose>
with respect to their <Quality>
from the point of view of the <Perspective>
in the context of <Context>

	Example
Object	Product, process, resource, model, metric, …
Purpose	evaluate choice of technique, describe process, predict cost, …
Quality	effectiveness, cost, …
Perspective	developer, customer, manager
Context	Subjects (personell) and objects (artifacts under study)

Experiment analysis

H0 hypothesis: there are no underlying differences between two sets of data

Type I error: Reject H0 even though H0 is true

Type II error: Accept H0 even though it is false

Example

H0 hypothesis: ”Data-corrupting faults are as common as non-corrupting faults”

There are 11 non-corrupting faults and 4 corrupting faults

What is the probability of up to four corruptive faults?

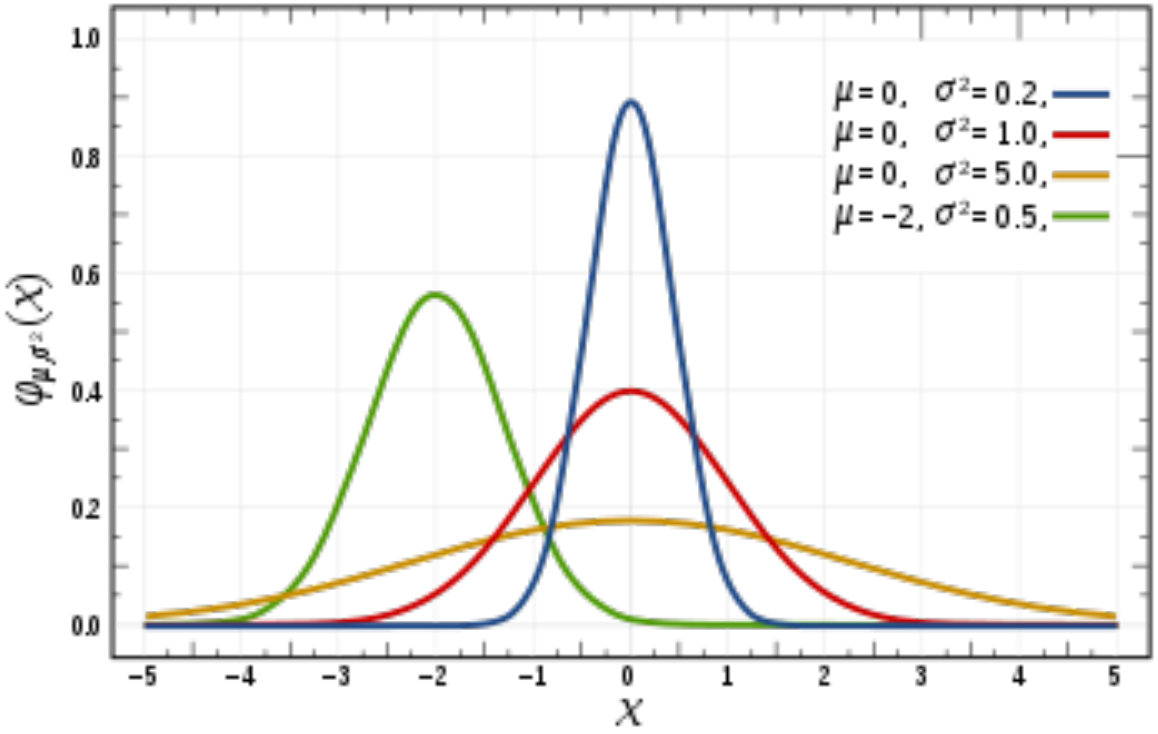
$$\sum_{i=0}^{4} \binom{15}{i} \left(\frac{1}{2}\right)^{i} \left(\frac{1}{2}\right)^{15-i}$$

What is the risk of a type I error,
given the probability ’a’ (!= 1/2) of the outcome?

$$\sum_{i=0}^{4} \binom{15}{i} a^{i} (1-a)^{15-i}$$

Parametric vs nonparametric tests

Can your data be described by an underlying (normal) probability distribution?



https://en.wikipedia.org/wiki/Normal_distribution#/media/File:Normal_Distribution_PDF.svg

Parametric distribution?

Non-parametric distribution?

One factor?

Chi-2, Binomial test

One treatment/sample?

Mann-Whitney

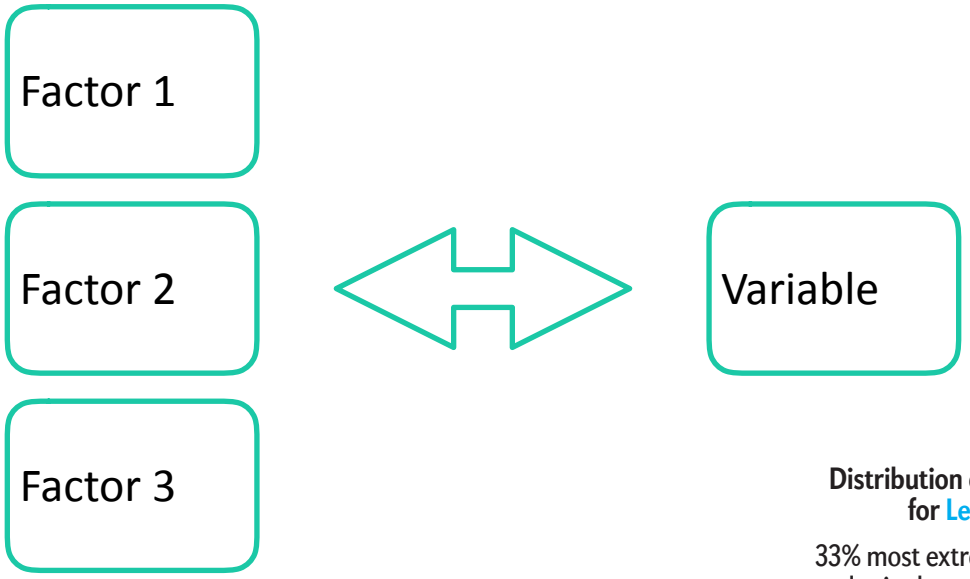
Paired comparison/ randomized design?

Statistical power

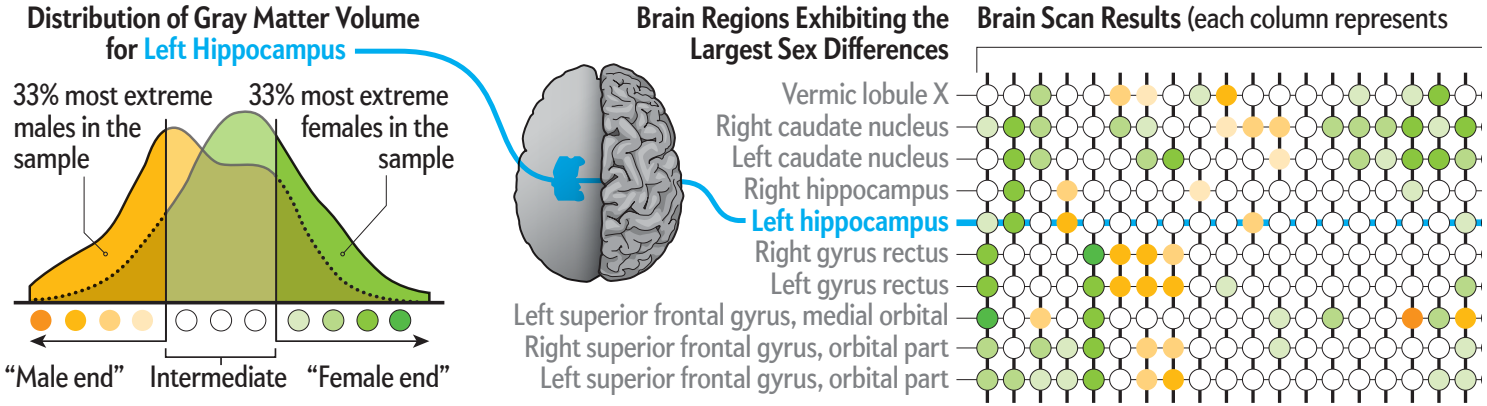
- P = 1 - risk of type II error

Classification problems

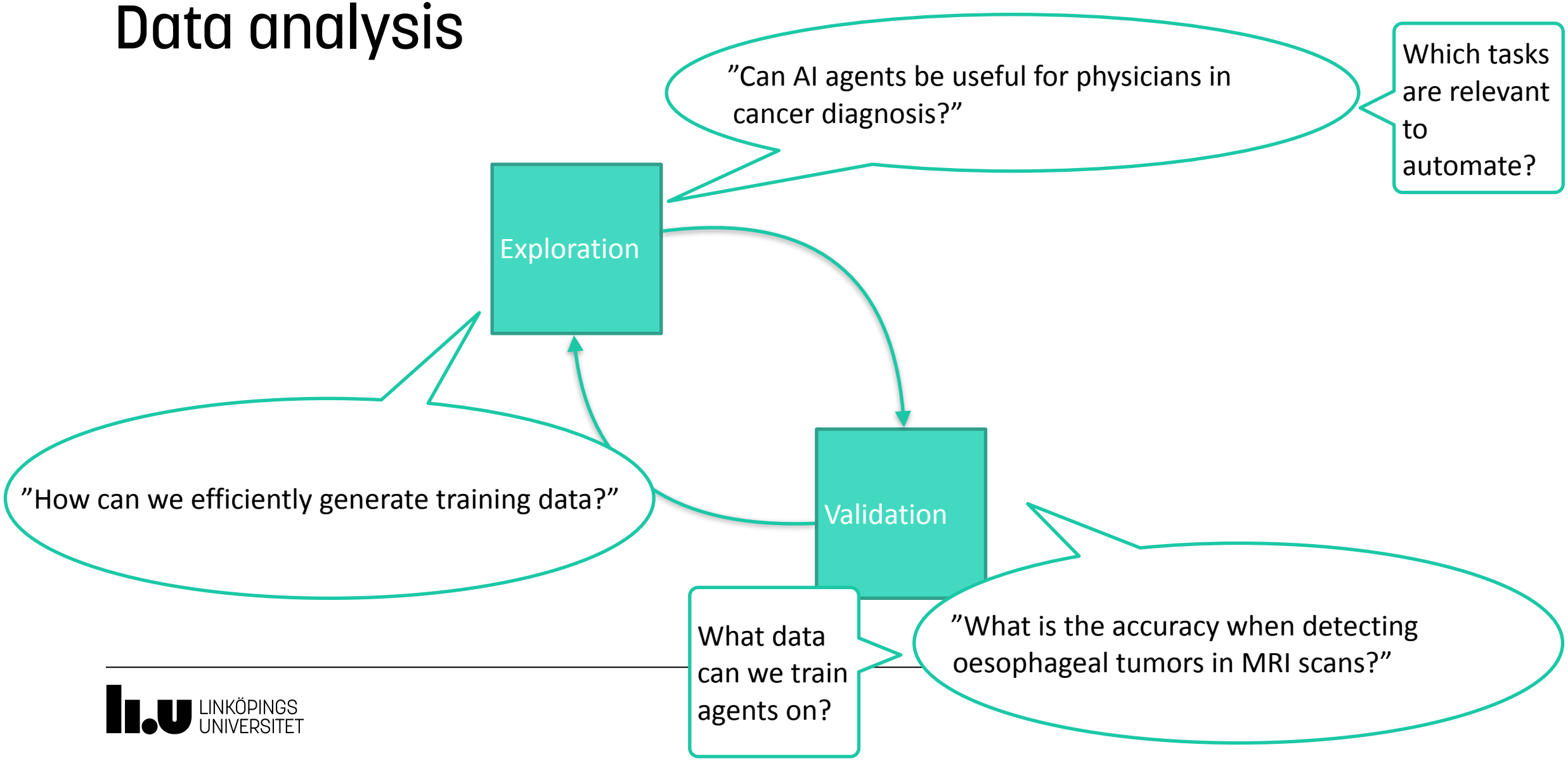
”Given luminosity, hue and saturation regional values, determine whether the picture contains a face”



”Given that an image contains a face, determine luminosity, hue and saturation regional values”



Data analysis



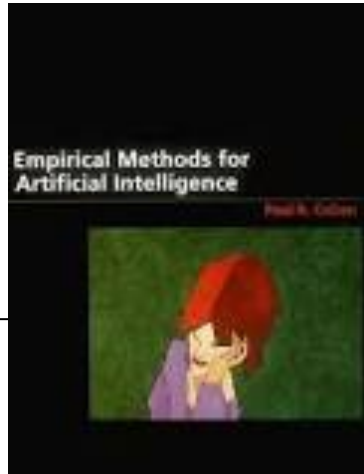
Data analysis, exploration

Trial	Wind speed	RTK	First Plan	Num plans	Fireline built	Area burned	Finish time	Outcome
1	high	5	model	1	27056	23.81	27.8	Success
2	high	1.67	shell	1	14537	9.6	20.82	Success
3	high	1	mbia	3	0	42.21	150	Failure
4	high	0.71	model	1	27055	40.21	44.12	Success
5	high	0.56	shell	8	0	141.05	150	Failure
6	high	0.45	model	3	0	82.48	150	Failure
7	high	5	model	1	25056	25.82	29.41	Success
8	high	1.67	model	1	27054	27.74	31.19	Success
9	medium	0.71	model	1	0	63.86	150	Failure
10	medium	0.56	mbia	7	0	68.39	150	Failure
11	medium	0.45	mbia	5	0	55.12	150	Failure
12	medium	0.71	model	1	0	13.48	150	Failure
13	medium	0.56	shell	4	42286	10.9	75.62	Success
14	low	0.71	model	1	11129	5.34	20.69	Success

Paul R. Cohen, Empirical Methods in Artificial Intelligence. The MIT Press, 1995

Data types

- Categorical data (Outcome) => Count frequency
- Ordinal values (Wind speed) => Correlation coefficients
- Interval or ratio scales (time to finish/best time to finish) => linear correlation coefficients



Distributions of data

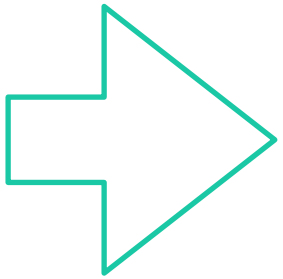
- Parametric distributions (assuming a probability distribution)

Sample/Value frequency	1	2	3
A	1/2	1/3	1/4
B	1/3	4	1/3
C	4	5	6

Transformations of data

1 4 5 7 45

2 5 4 8 35



1 1 -1 1 -10 or 1 1 -1 1 -1

Quantitative studies

- Uses statistical analyses of some empirical data
  - Randomization of subjects
  - Blocking (grouping) subjects based on confounding *factors*

Factors

- That which may correlate with (and possibly cause) an effect
  - ”How does *SCRUM* affect product quality as measured by the number of bugs?”
  - ”How is code quality affected by the choice of *programming language*?”
  - ”How understandable is a design document when creating procedural and OO design, based on *good/bad requirements*?”

Analysis

- There must be a *null hypothesis* which we can test our data against
- One factor, two treatments: t-test, Mann-Whitney
- One factor, several treatments: ANOVA
- Two factors: ANOVA

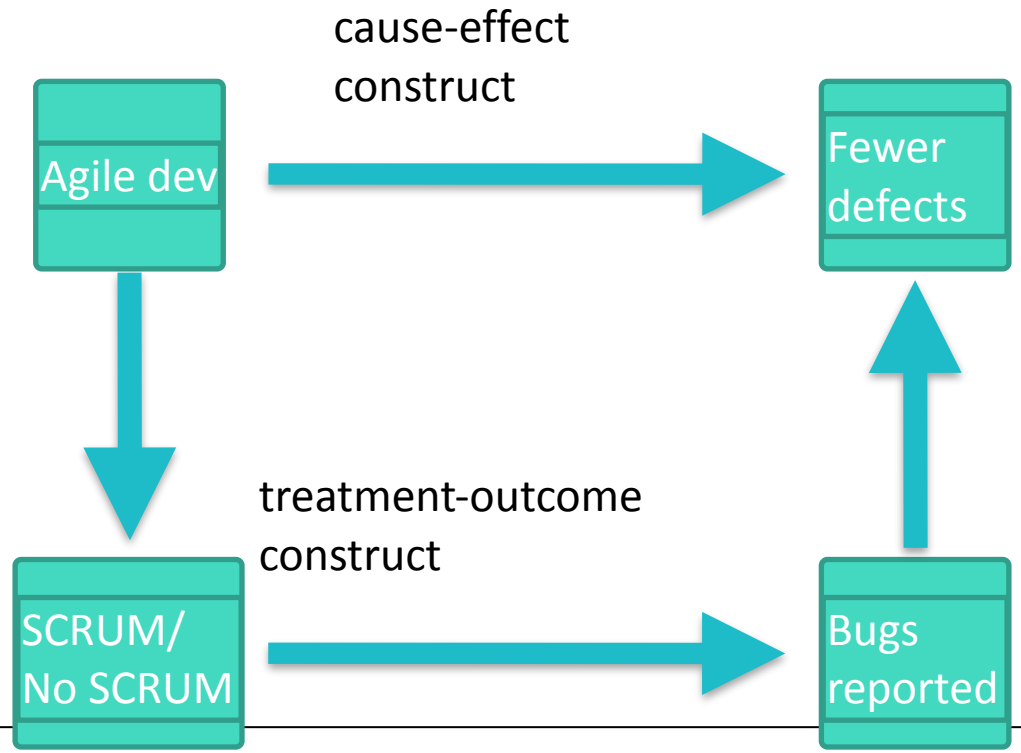
Statistics

- There are separate statistics courses, but..
  - Separate correlation and causality
  - Unless >= 95% confidence, there is no correlation
  - Confidence only part of statistical *power* (confidence + effect size + sample size)

Discussion, example

Does agile development lead to higher quality code?

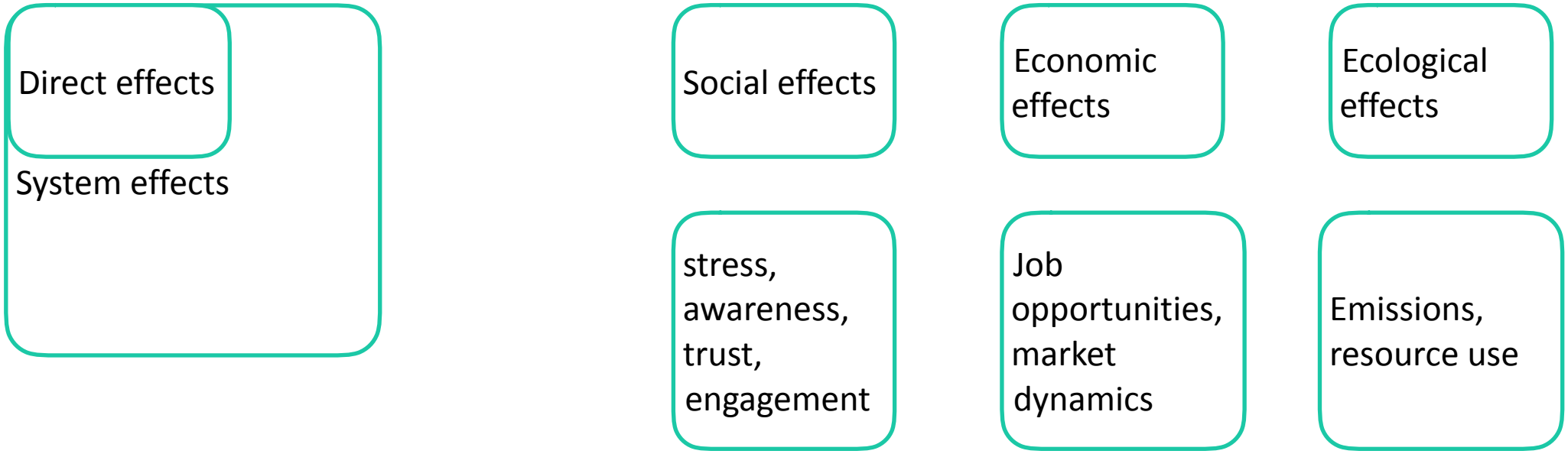
Hypothesis



Your work in a wider context

Why do we as humans have to solve this problem?

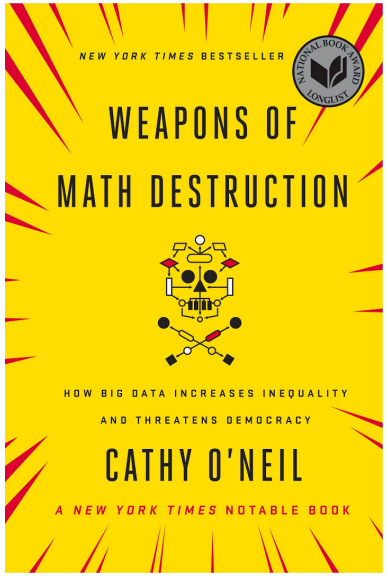
Your work in a wider context



C. Becker, R. Chitchyan, L. Duboc, S. Easterbrook, B. Penzenstadler, N. Seyff, and C. C. Venters, “Sustainability design and software: the Karlskrona manifesto,” in IEEE International Conference on Software Engineering (ICSE), vol. 2, pp. 467–476, IEEE, 2015.

The effects of Big Data

- A level 1 non-linear, chaotic dynamic system: the climate system, turbulence, population dynamics
- A level 2 chaotic system: Human activities such as stock markets



Example

- ”Automating the classification of fMRI images for oncologists”
- ”Directed media content through topic modeling”