

# Scientific methods and data ethics

732A60 - Seminar 7

**Louis Ohl**

Division of Statistics and Machine Learning (STIMA)  
Department of Computer and Information Science (IDA)  
Linköping University

- 1 Introduction
- 2 What is science? What is knowledge?
- 3 Hypothetico-deductive method
- 4 Models
- 5 Scientific revolutions
- 6 Ethics in science and statistics
- 7 Summary

- 1 Introduction
- 2 What is science? What is knowledge?
- 3 Hypothetico-deductive method
- 4 Models
- 5 Scientific revolutions
- 6 Ethics in science and statistics
- 7 Summary

# Intro

In this lecture:

- science = “all science”, not just “natural science”
- A *smörgåsbord* of different topics

But before I tell some stories...

Take 5 minutes to discuss with your friend(s) / neighbour(s): **What defines science?**

- Type in your answer in Mentimeter!

If you struggle to have ideas to discuss:

If you have no idea, consider the following cases:

- Is it scientific to know that the Earth is round?
- My grandmother uses eggshells to prevent slugs in her garden, is that scientific?
- By giving candies to mice, I can test and compare if they will trigger a cancer among humans
- Does using master-level maths make you a scientist?
- Is it scientific to believe in science?

## Example

Consider the observed dataset

$$x = [ 0 \quad 1 \quad 2 \quad 3 ] \quad y = [ 2 \quad 3 \quad 4 \quad 5 ]$$

**Problem** We want to understand the relation between  $x$  and  $y$ . The assumption  $y = f(x)$  is given. Which function  $f$  should be used? Why?

$\implies$  Science to the rescue!

- 1 Introduction
- 2 What is science? What is knowledge?
- 3 Hypothetico-deductive method
- 4 Models
- 5 Scientific revolutions
- 6 Ethics in science and statistics
- 7 Summary



# What is science?

- “Large amount of relatively secured knowledge”
- Results - description of facts or explanation of dependencies
  - Results are often published in scientific journals, conferences or books
  - Peer review
- A process - the methods and activities that lead to the results
  - *e.g.* experiments, computations, theories

# What is science?

- Methods are established within the scientific community
- Scientific methods have clear and explicit rules and procedures
- Replication is important: details matter
  - Methods that are arbitrary and cannot be repeated are not scientific
- Science should be as objective as possible
  - Researcher bias should be reduced
- Scientific knowledge is created within the scientific community
- All new results should be related to existing knowledge in the field

# Categories

- Nomothetic (general) studies
  - General laws
- Idiographic (specific) studies
  - Describes specific objects and processes
  - What happened at the Battle of Hastings?
- Formal science
  - The study of constructed and formal systems
  - Logic, mathematics, statistics
- Empirical science
  - Empirical evidence gained through experimentation or collection of data
  - The objects and processes are studied in the “real world”
  - Medicine, history, economics

## Classic definition of knowledge

- Classic definition of knowledge (Plato): **Justified true belief**
- $A$  knows that a proposition,  $P$ , holds if and only if the three following conditions are met
  - $P$  is true
  - $A$  believes that  $P$  is the case
  - $A$  is justified in believing that  $P$  is the case
- Epistemology
  - Rationalism: knowledge through logic and deductive reasoning
  - Empiricism: knowledge through empirical evidence

# Knowledge

- Correspondance theory of truth
  - Truth or falsity of a statement is determined only by how it relates to the world
- Coherence theory of truth
  - A statement is true if it is coherent with other statements in the system under study
- Pragmatism
  - A statement is true if the predicted consequences are correct

# Scientific explanations

- Deductive explanations
  - Based on a number of premises
  - With help of the premises, conclusions are deduced through logic
  - If the premises are true, then the conclusions are also true
- Probabilistic explanations
  - No general law. Premises that have a (high) probability to be true (or to happen)
  - The “probabilistic conclusions” are not true in a formal way, but may be probable
  - Probability theory and statistical inference are used to formalize the probabilistic explanations

# Philosophy of statistics

- Probability
  - Relative frequencies in series of events
  - Degree of belief, epistemic
- Statistical inference
  - Frequentist statistics
  - Bayesian statistics
- Philosophy of statistics:  
<https://plato.stanford.edu/entries/statistics/>

# Induction

- Induction
  - From the observed (specific) to the general
  - Based on *verification* of theories
- Can we be absolutely sure that...
  - The sun will rise tomorrow
  - All swans are white?
- The problem of inductions
  - We cannot logically verify cause and effect based on induction
- Circular reasoning
  - The future will resemble the past, because we know that
  - The future will resemble the past, because we know that
  - ...



- 1 Introduction
- 2 What is science? What is knowledge?
- 3 Hypothetico-deductive method
- 4 Models
- 5 Scientific revolutions
- 6 Ethics in science and statistics
- 7 Summary

## Hypothetico-deductive method

- Axiomatic method: summarizes and justifies a large number of statements by proving that they can be deduced from a small number of axioms
- Problems
  - How to justify the axioms?
  - How to relate to the “real world”?
- *Karl Poppers’* suggestion: the hypothetico-deductive method
  - Focus on hypotheses
  - Falsification
  - Popper: psychoanalysis cannot be considered a science
    - It is not falsifiable

# Hypothetico-deductive method

- Hypothesis: a proposed explanation for a phenomenon
  - Scientific hypothesis: should be testable (falsifiable)
- Hypothesis: a statement that fulfills the following requirements
  - We are not sure if it is true
  - We can deduce logical consequences from it, with the purpose of
    - Testing it
    - Making predictions
    - Explaining observed facts

# Hypothetico-deductive method

- Infinite number of observations cannot verify an hypothesis
- Degree of belief
  - Hypotheses are *confirmed* to different degrees, when we fail to falsify them
  - The best option with current knowledge, but we do not know if they are true

## Auxiliary hypotheses

- Theoretically we only need one incorrect observation to falsify an hypothesis
- Auxiliary hypothesis: necessary hypothesis for the deduction of a prediction, but which is not tested in the study
- Reality is complex: occasional incorrect observations are often explained as measurement error or are assumed to depend on some of the auxiliary hypotheses

## Hypothetico-deductive system

- Axiomatic system: consequences are absolutely based on the premises
- Hypothetico-deductive system: a deductive system with hypotheses as starting point
  - All principles and statements that are not entirely secure are called hypotheses
  - Justification is based on the fact that the consequences deduced from the hypotheses conforms with our experience
  - We are never entirely sure about the truth value of our hypotheses

## Creation of hypotheses: example

Let's play a game!

- I have a rule in mind that binds a True/False statement to a pair of numbers
- Can you guess it?
- You may give me 10 pair of numbers max
- You may give as many hypotheses as you want on the actual rule

Here are the starting tuples

Pair of numbers	Answer
(0,2)	TRUE
(1,5)	FALSE
(-2, 0)	TRUE

## Creation of hypotheses

- Hypotheses are created as “guesses”
  - Deduce the consequences of hypotheses with logic



## Choices of hypotheses: simplicity

- Several (infinite number) hypotheses can be used to explain the same observations
  - We must choose!
  - Ex: regression, what is a good fit to data?
- Criteria
  - Security: should we be aggressive or defensive when we create hypotheses?
  - Strength: how much can the hypothesis explain?
  - **Simplicity**: Most important criteria

## Choices of hypotheses: simplicity

- In practice, researchers often choose the simplest hypothesis
  - Aesthetic
  - Makes understanding easier
- Occam's razor: if  $H_1$  and  $H_2$  do the same prediction
  - Choose the one that has the fewest assumptions
- Poppers' recipe
  - Boldness in the guesses/hypotheses and stringency in rejection
  - You should declare precisely under which circumstances you are willing to abandon your hypothesis

- 1 Introduction
- 2 What is science? What is knowledge?
- 3 Hypothetico-deductive method
- 4 Models
- 5 Scientific revolutions
- 6 Ethics in science and statistics
- 7 Summary

# Models

- A model can be compared to a map
- What kind of map is important for the following persons?
  - A taxi driver
  - An orienteerer
  - An epidemiologist

# Models

- “All models are wrong. Some are useful”. - George E.P. Box
- Science often speaks about models
- Model: a representation of a process or a system
  - Important features are a part of it (often emphasized)
  - Other features are not included
- Historically: mechanical models were important
- Nowadays: theoretical or mathematical models are much more important

# Models

- Purpose of scientific models
  - Understand, define, quantify, visualize or simulate a process or a system
  - Common approach when working with complex problem
  - Calculations and predictions become easier
- A theory is always a model: makes a phenomenon understandable
  - All models are not theories, *e.g.* mechanical model
- Causal models
  - Describes the causal mechanisms of a system
  - Causal diagram is a directed graph that displays causal relationships
  - Confounding factor: important in statistics
    - **Correlation does not imply causation**

- 1 Introduction
- 2 What is science? What is knowledge?
- 3 Hypothetico-deductive method
- 4 Models
- 5 Scientific revolutions
- 6 Ethics in science and statistics
- 7 Summary

# Scientific revolutions

- Research is often a cumulative process
  - Continuous revision of old knowledge
- Sometimes there are revolutions: old theories are rejected and replaced with new ones
  - Chemical revolution: Lavoisier
  - Scientific theory of evolution: Darwin
  - Theory of relativity: Einstein
  - Quantum mechanics
  - Convolutional neural network and deep learning within image classification (Krizhevsky, Sutskever and Hinton, 2012)



# Scientific revolutions

- *Thomas Kuhn* "The Structure of Scientific Revolutions", 1962
  - Normal science  $\implies$  revolution and crisis  $\implies$  normal science
- Paradigm = central hypotheses
  - Researchers are laying a puzzle
  - After a while: many pieces do not fit
  - A revolution happens when the central hypotheses are rejected: a new puzzle starts
- It can be hard in practice to define and observe scientific revolutions
  - When does a hypothesis become a paradigm?
- Paradigms can be subjects of "religious" belief

- 1 Introduction
- 2 What is science? What is knowledge?
- 3 Hypothetico-deductive method
- 4 Models
- 5 Scientific revolutions
- 6 Ethics in science and statistics
- 7 Summary

# Ethics in science

- Ethics or moral philosophy
  - Deals with what is right and wrong
  - How to act?
- Research ethics: how to handle moral issues that arise during or as a result of research activities?

# Ethics in science

Bad examples from history:

- *Nazi human experimentation*
  - A large number of prisoners were forced to participate, the experiments typically led to death, trauma, permanent disability, etc
  - Led to the *Nuremberg Code* after the Nuremberg trials
- *Tuskegee syphilis experiment* (US, 1932-1972)
  - African-American men were used to see the effect of untreated syphilis infection, without consent of the participants
- *Vipeholm experiments* (Sweden, 1945-1955)
  - Intellectually disabled patients were fed with sweets in order to provoke dental caries
  - The aim was to determine the role of carbohydrates

## Ethics in science

- A researcher's work is regulated by rules and regulations
- Researcher's own ethical responsibility that
  - Research has good quality
  - Is morally acceptable
- Professional ethics
  - Research activity is driven by a number of implicit and explicit norms that decide what good science is, e.g. Helsinki Declaration about ethical principles regarding human experimentation
  - Follow national and local rules: issues like discrimination, harassment and humiliation, gifts to the researcher
  - Field specific codes of ethics: ethical code

# Ethics in statistics

- American Statistical Association (USA): "Ethical Guidelines for Statistical Practice"
- Royal Statistical Society (UK): "Code of conduct"
- International Statistical Institute: "Declaration of Professional Ethics"
- Swedish Statistical Society: "Svenska statistikfrämjandets etiska kod för statistiker och statistisk verksamhet"

## Ethics in statistics

“Declaration of Professional Ethics”:

- Pursuing objectivity
- Clarifying obligations and roles
- Assessing alternatives impartially
- Conflicting interests
- Avoiding preempted outcomes
- Guarding privileged information
- Exhibiting professional competence
- Maintaining confidence in statistics

# Ethics in statistics

## “Ethical Guidelines for Statistical Practice”

- Professional integrity and accountability
- Integrity of data and methods
- Responsibilities to science/public/funder/client
- Responsibilities to research subjects
- Responsibilities to research team/colleagues
- Responsibilities to other statisticians and other practitioners
- Responsibilities regarding allegations of misconduct
- Responsibilities of employers, including organizations, individuals, attorneys, or other clients, employing statistical practitioners.



# Ethics in statistics

Handling of data has a special role in statistics and data science.

- Special care must be taken when data is collected, stored and used for statistics and machine learning
- General Data Protection Regulation (GDPR)

# GDPR

Look [here](#) and [here](#)

- Scope: "The General Data Protection Regulation exists to protect individuals' fundamental rights and freedoms, in particular their right to protection of their personal data".
- GDPR
  - EU law on data protection
  - Regulate the use of personal data
  - Called "Dataskyddsförordningen" in Sweden

## Ethics in big data

“In today’s most common digital business model, consumers pay for ‘free’ products with their personal data.”

# Ethics in big data

A few starting principles:

1. Ownership: individuals own their own data
2. Transaction transparency: the use of the data should be transparent
3. Consent: informed and explicitly expressed consent is needed to use the data
4. Privacy
5. Currency: individuals should be aware of financial transactions resulting from the use of their personal data and the scale of these transactions
6. Openness: aggregated data sets should be freely available

# Ethics in big data I

Five principles for big data ethics from “[Medium](#)”:

1. **Private customer data and identity should remain private:** Privacy does not mean secrecy, as personal data might need to be audited based on legal requirements, but that private data obtained from a person with their consent should not be exposed for use by other businesses or individuals with any traces to their identity.
2. **Shared private information should be treated confidentially:** Third-party companies share sensitive data — medical, financial or locational — and need restrictions on whether and how that information can be shared further.
3. **Customers should have a transparent view** of how our data is being used or sold and the ability to manage the flow of their private information across massive, third-party analytical systems.

## Ethics in big data II

4. **Big Data should not interfere with human will:** Big data analytics can moderate and even determine who we are before we make up our minds. Companies need to consider the kind of predictions and inferences that should be allowed and those that should not.
5. **Big data should not institutionalize unfair biases like racism or sexism.** Machine learning algorithms can absorb unconscious biases in a population and amplify them via training samples.

## Facebook-Cambridge Analytica data scandal I

- [Facebook-Cambridge Analytica data scandal](#): big political scandal in early 2018
- The company Cambridge Analytica had collected personal data from millions of peoples' Facebook profiles
  - without consent
  - and used it for political advertising purposes
- Whistle-blower: [Christopher Wylie](#)
- Aleksandr Kogen researcher at Cambridge University created an app
  - "This Is Your Digital Life"
- Several hundred thousands of Facebook users gave consent to be part of the survey only for academic use
- Facebook's design allowed data to be collected from the social network of the participants
  - This allowed Cambridge Analytica to collect data from up to 30/50/87 million users

## Facebook-Cambridge Analytica data scandal II

Cambridge Analytica used the data to:

- Create psychographic profiles of Facebook users
- Profiles used to choose advertisement that most effectively persuade specific groups of persons
- Used in political campaigns with the aim to affect elections, for example:
  - 2016 US presidential election
  - 2016 UK referendum on the EU membership



## Discussion

- What is ethical to do with user data on social media platforms?
- “Personal data is gold”: companies using data as main source of profit, what is good ethics in such businesses? Do the users understand what their data are used for?
- What responsibilities does a machine learner or data scientist working for a social media company have?
  - What to do if your boss asks you to do something that maybe feels wrong? For instance, collect and analyze personal data when it is unclear if consent is given
- Is it right to produce a machine learning system (in a democratic country) and then sell the system to totalitarian regimes who want to use the system for controlling its citizens?

- 1 Introduction
- 2 What is science? What is knowledge?
- 3 Hypothetico-deductive method
- 4 Models
- 5 Scientific revolutions
- 6 Ethics in science and statistics
- 7 Summary

# Summary

Take-home messages:

- Be critical! Question it!
  - Data, plots, graphs, tables
- Science is hard
- Correlation does not imply causation
- Who is behind the results? What are the incentives?
  - Money, power, prestige, reputation
- Don't do bad things with data: ethics matter
  - "With big data comes big responsibility"

# References I

## Books:

- Ladyman, James, Understanding Philosophy of Science, Routledge, London, 2002
- Dagfinn Føllesdal, Lars Wallöe, Jon Elster, Argumentationsteori, språk och vetenskapsfilosofi, Thales, Stockholm, 2001
- Data Ethics - The New Competitive Advantage,  
<https://dataethics.eu/book/>

## References II

Links:

- Stanford Encyclopedia of Philosophy [[link](#)]
  - Philosophy of Statistics [[link](#)]
  - Scientific Method [[link](#)]
  - Science and Pseudo-Science [[link](#)]
  - Scientific Progress [[link](#)]
  - Scientific Revolutions [[link](#)]
  - The Problem of Induction [[link](#)]
  - Bayes' Theorem [[link](#)]
- Probability and Induction [[link](#)]
- DataEthics [[link](#)]

Thank you for your attention!

[www.liu.se](http://www.liu.se)