MSc Applied Information and Data Science

HSLU Lucerne University of Applied Sciences and Arts

Data Analytics for Data Scientists

Design of Experiments (DoE)

Lecture 07: Paradigms

2025

Prof. Dr. Jürg Schwarz

Program: 16:15 until 17:55

16:15	**Begin of the lesson**
	Lecture: Jürg Schwarz ◦ The Fourth Paradigm ◦ The End of Theory ◦ Properties of research approaches ◦ The integrated Fourth Paradigm ◦ Preview of Lecture 08
~ 17:00	**Break**
	Lecture: Jürg Schwarz ◦ Addendum Introduction to Analysis of Variance (ANOVA)
17:55	**End of the lesson**

Detailed table of contents → Slide 24

The Fourth Paradigm

Talk 2007: Data science in the context of science paradigms

Held by James Nicholas "**Jim" Gray**, eScience Group, Microsoft Research
Ph.D. in Computer Science in 1969 at University of California, Berkeley

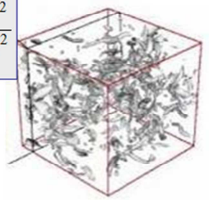
eScience -- A Transformed Scientific Method

Jim Gray,
eScience Group,
Microsoft Research
http://research.microsoft.com/~Gray
in collaboration with
Alex Szalay
Dept. Physics & Astronomy
Johns Hopkins University
http://www.sdss.jhu.edu/~szalay/

Science Paradigms

- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations
- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today:
  data exploration (eScience)
  unify theory, experiment, and simulation
  – Data captured by instruments
    Or generated by simulator
  – Processed by software
  – Information/Knowledge stored in computer
  – Scientist analyzes database / files
    using data management and statistics

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - \mathrm{K}\frac{c^2}{a^2}$$



Science paradigms according to Jim Gray and partly revised in Hey (2009)*

First **Experimental** Science** (Thousand years ago)

- Description of natural phenomena

Second **Theoretical Science** (Last few hundred years)

- Modelling and generalization: Newton's Laws, Maxwell's Equations …

Third **Computational Science** (Last few decades)

- Simulation of complex phenomena

Fourth **Data-Intensive Science** → **eScience** (Enhanced Science) (Today)

- Exploration of data: Analysis of database / files using data management and statistics
- Unification of theory, experiment, and simulation
- Synthesis of information technology and science
- Requires combination of statistics & computer science
- …
- Change of publishing data & literature (curation, access, preservation)

*Tony Hey, until 2014 Corporate Vice President, Microsoft Research / **Jim Gray uses the term "empirical" in a somewhat broader context.

How is the "Fourth Paradigm" implemented?

Example Strategic planning by the ETH Council* for the ETH sector → 2017-2020

Science has entered a data-centered Fourth Paradigm that complements theory, experiment and simulation by analyzing the most extensive data sets.

These make it possible to arrive at findings and decisions on the basis of big-data collections and by applying empirical principles.

Strategic planning 2021–2024: "The Strategic Focus Area "Data Science" […] will be continued so that their full potential can be realised."

Strategic planning 2025–2028: "It also addresses the future of the centres and platforms that have been created in the context of the Strategic Focus Areas (SFAs) 2021–2024 … Particular emphasis is placed on the future of SDSC …" → "The Swiss Data Science Center"**


Executive Summary

Contents

ETH BOARD Strategic Plan 2025–2028 of the ETH Board for the ETH Domain 3

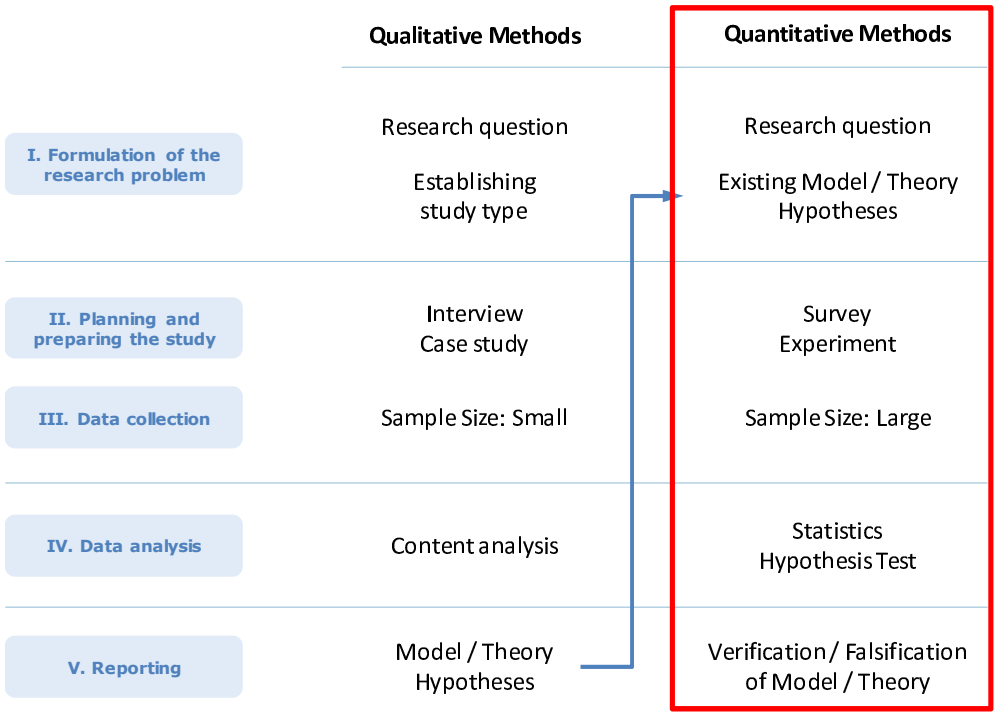
The aim of SDSC aim is to accelerate the use of data science and machine learning technologies by researchers in the ETH Domain and the Swiss academic community at large, as well as by the industrial sector.

*www.ethrat.ch / **www.datascience.ch

The Fourth Paradigm in detail

The Fourth Paradigm

(Quantitative) empirical research*



Data-driven research

Reality created through data

Research question / case study

Data exploration / data mining

"Big data" (the Vs → volume, variety, ...**)

Analysis / modeling with big data analytics

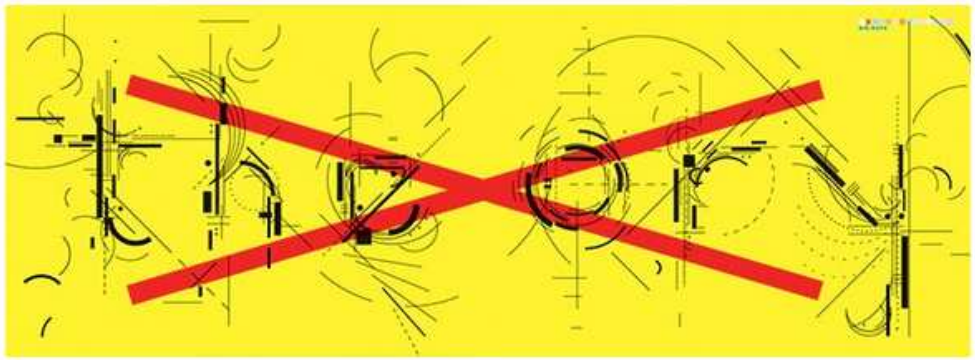
Application / description of the world

*Lecture 02 Principles / **Sivarajah et al. (2017) & (2024)

The End of Theory

Article 2008: The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

Written by Chris Anderson, Editor-in-Chief, www.wired.com
BSc in Physics 1985 at George Washington University

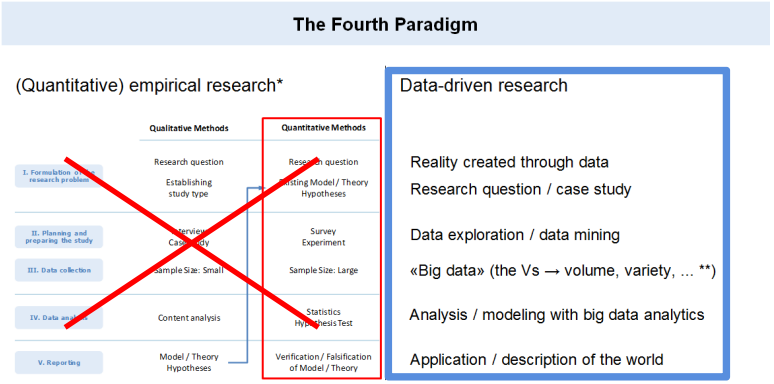


The new availability of huge amounts of data, along with the statistical tools to crunch these numbers, offers a whole new way of understanding the world. Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all.

The Article corresponds to the data-driven part of the fourth paradigm.

www.wired.com/2008/06/pb-theory

How to assess *The End of Theory*?



Arguments that speak **<u>in favor of it</u>***

- No need for *a priori* theories
- Large data quantities can cover an entire domain and provide a comprehensive solution.
- New fields like machine learning and artificial intelligence have emerged due to huge data availability, aimed at extracting knowledge directly from data and solving complex problems.
- Through the use of unbiased value-free data analysis, the data speaks for itself, free from human prejudgment and without scientific premises.

Arguments that speak **<u>against it</u>***

- Correlation is not causation!
  One of the main counterarguments is that correlation alone does not imply causality.
- Large data quantities are created in an environment influenced by many factors:
  - Aim and purpose
  - Technology and platform
  - Regulatory environment
  - …
- Large data quantities cannot be interpreted outside of their context.
- There is no such thing as unbiased value-free data analysis …

*Arguments from various sources and with the support of ChatGPT-4o (April 2025): Prompt on Slide → 24

Quantitative empirical research

Elements of quantitative empirical research

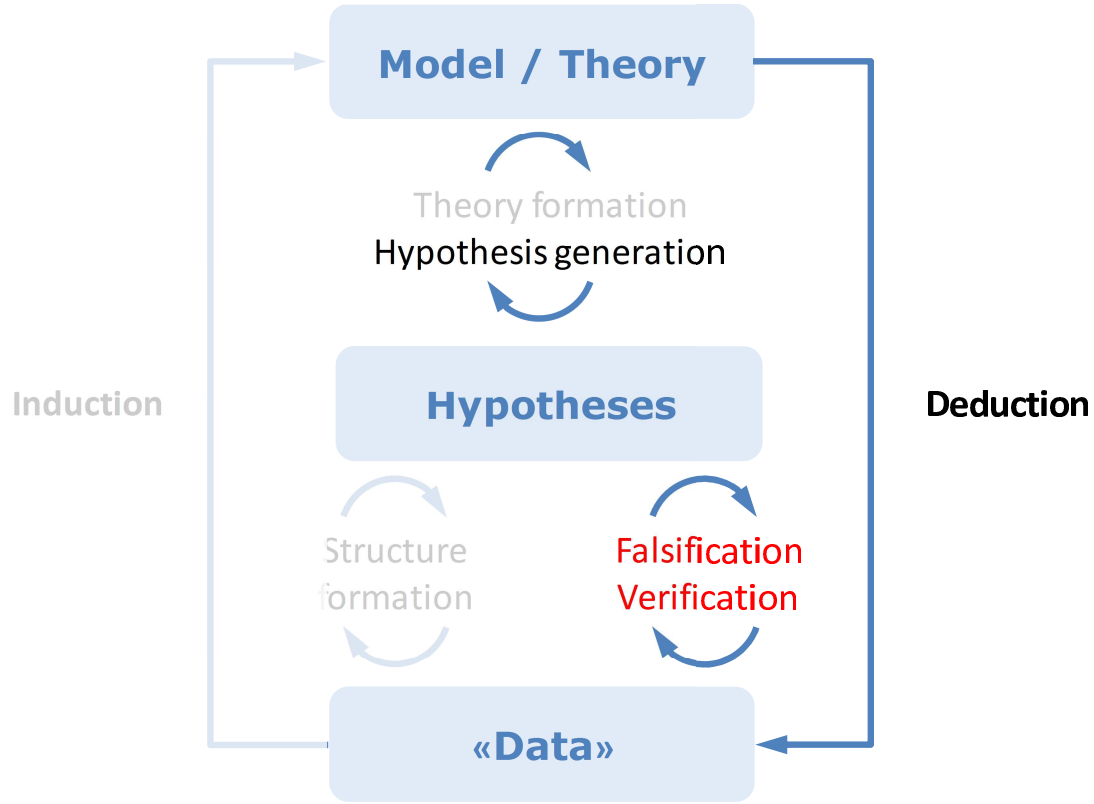
Research process*

I. Formulation of the research problem	
1 Research question / Hypothesis formation	Basic principles (→ Lecture 02)
II. Planning and preparation of the study	
2 Determination of study design	Study design etc. (→ Lectures 01 – 04)
3 Construction of survey instrument	Your curriculum
4 Definition of sampling procedures	Sampling (→ Lecture 05)
5 Pretesting	Your curriculum
III. Field phase / Data collection	
6 Application of survey instrument	Your curriculum
IV. Data Analysis	
7 Data preparation	*Module Classical and Bayesian Statistics*
8 Data Analysis / Modeling	Testing hypotheses (→ Lectures 02 & 07)
V. Reporting / Implementation	
9 Research report / Presentation	
10 Implementation of research results	

*Lecture 02: Principles

Knowledge gain in quantitative empirical research

Deductive approach → Conclusion from the general to the specific



Drawing conclusions from theory and applying them to empirical data.

The process starts with the theory from which empirically testable hypotheses are derived

- Rejection based on data → theory must be revised → Falsification
- Non-rejection based on data → theory is temporary confirmed → Verification

Typical limitations of quantitative empirical research

Research question / Forming hypotheses

- The research question cannot be implemented
- There is a poor hypothesis structure

Study design

- RCT not used / not possible
- Inadequate sampling

Instrument / Field phase

- Poor operationalization
- Poor implementation

Testing hypotheses

- Meaning of significant hypotheses vs. meaning of effect size (→ Lecture 06)
- p-hacking
  (looking for data subsets and configurations until the p-value is less than 5%)

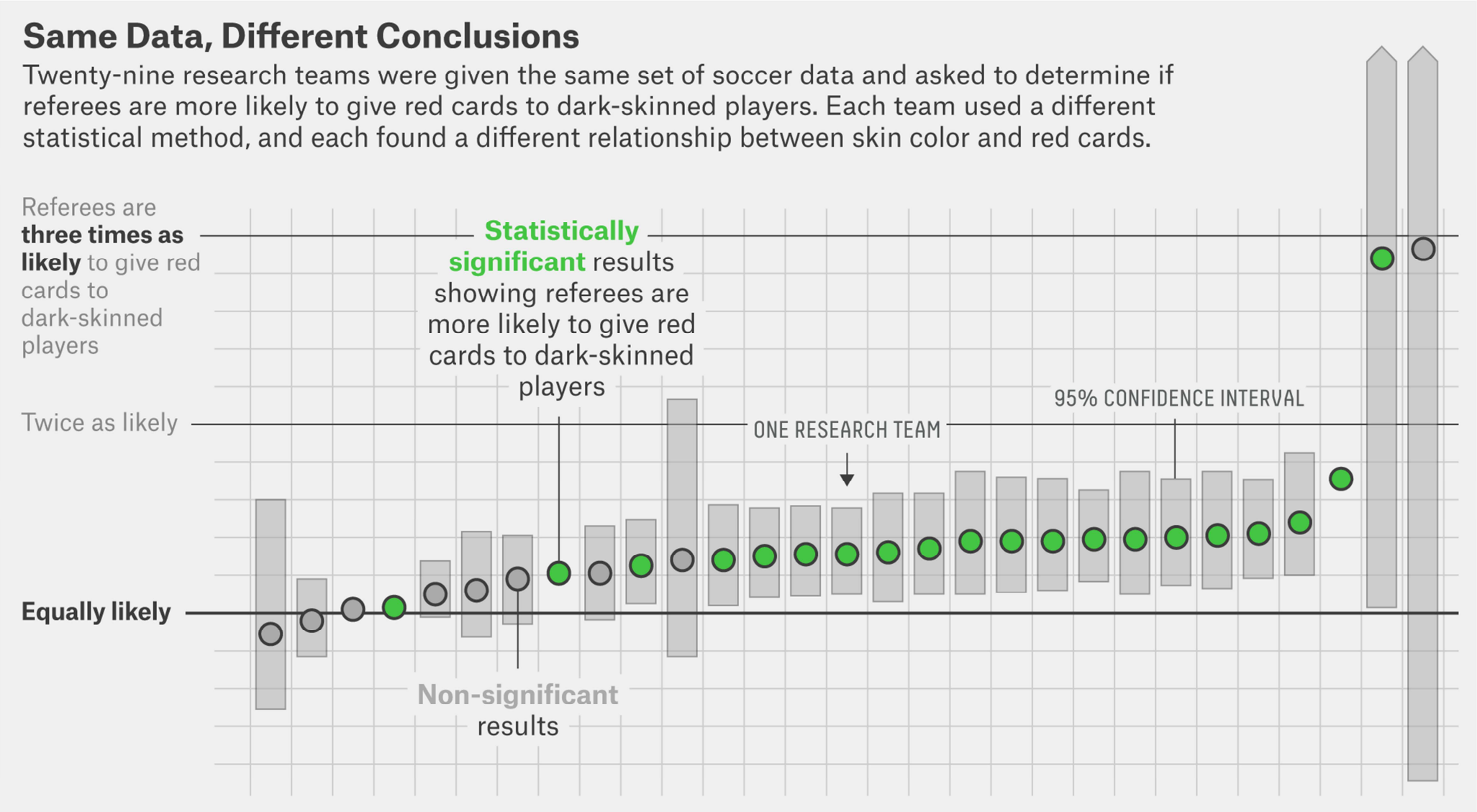
Prerequisites

- Assumptions about the distribution of variables are violated (normal distribution, etc.)
- Assumption of homogeneity of variance is violated

Et cetera …

Example – p-hacking

Soccer: Do referees give dark-skinned players more red cards than light-skinned players? Identical data from 29 teams were analyzed with different statistical methods.



fivethirtyeight.com/features/science-isnt-broken

Data-driven research

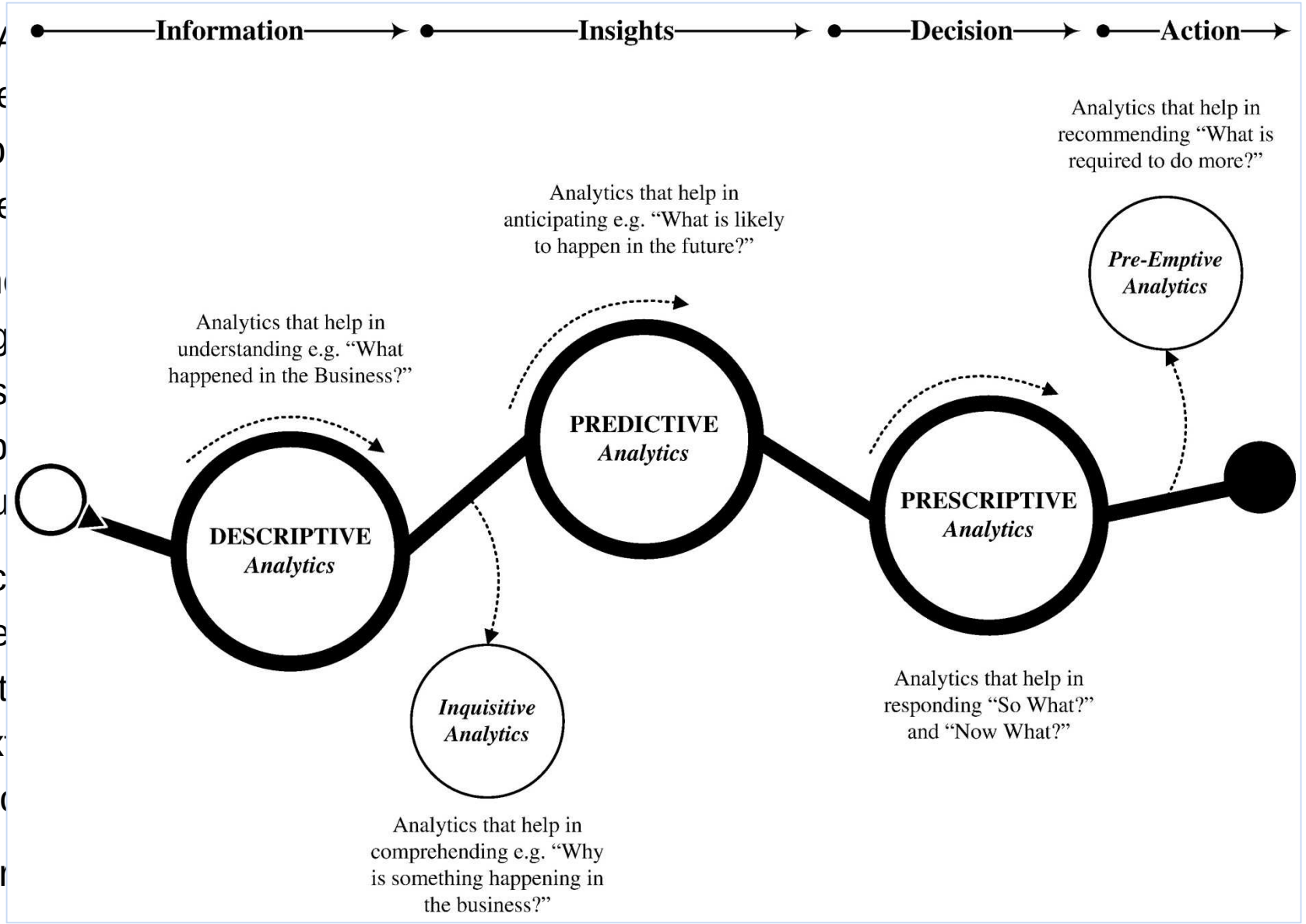
Elements of data-driven research – Appendix on Slide 23

Visual A
- (Inte
- Exp
- (Inte

Machin
- Reg
- Res
- Sup
- Neu

Advanc
- Dee
- Patt
- Tex
- (Soc

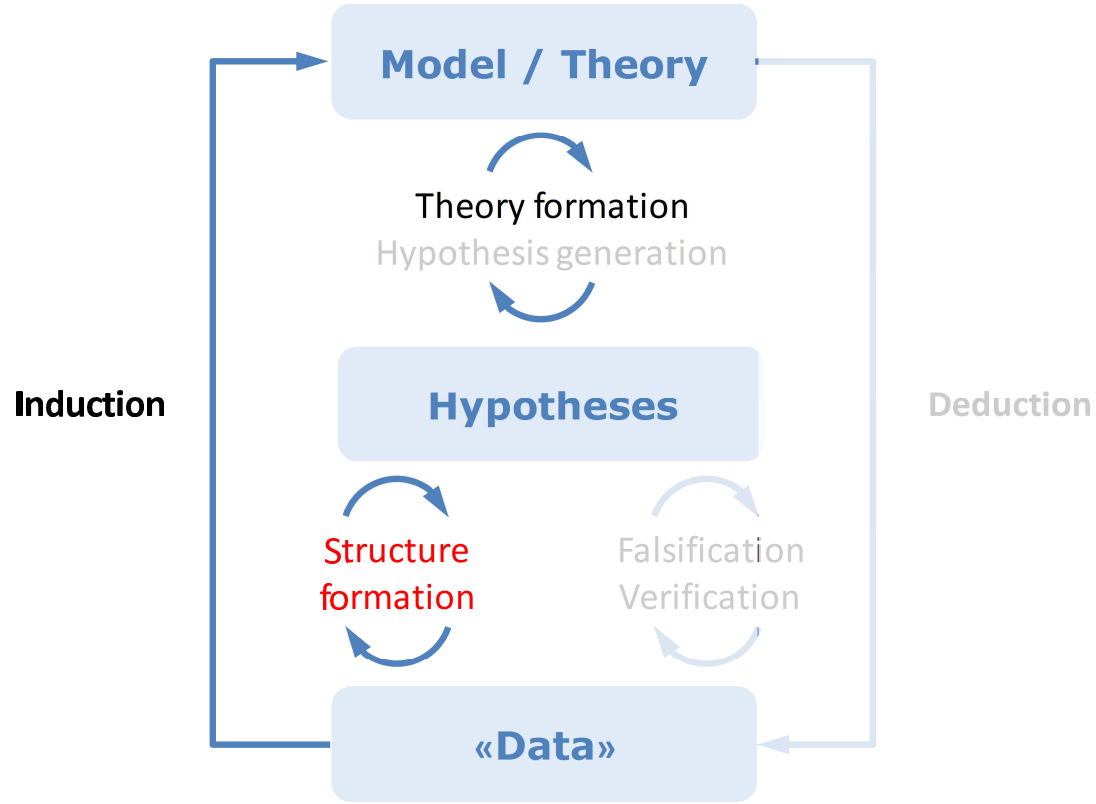
Et ceter



Sivarajah et al. (2018) & (2024)

Knowledge gain in data-driven research

Inductive approach → Conclusion from the specific to the general



Drawing conclusions from empirical data on scientific theory to a higher level.

The process starts with given data

- From the data, patterns are gradually worked out → Structure formation
- By means of induction, hypotheses about theories are formed and theories are formulated

Typical limitations of data-driven research

Research question / Forming hypotheses

- There is no research question
- There is no hypothesis structure

*Big Data Hubris** (*hubris*: excessive pride or self-confidence) → see also Slide 7

- Big Data is used as a replacement for traditional methods and not as a supplement
- Correlation is understood as causality

Data basis

- Population is unknown
- *Sparse data not recognized / not considered*
  (*Sparse data: Although the data basis is "big," it contains little information*)

Instrument / Field phase

- Unknown operationalization
  (*Example Social Media: It is not clear what information is generated and why*)
- Underlying conditions that change over time are not recognized / not taken into account

Data analysis / Modeling

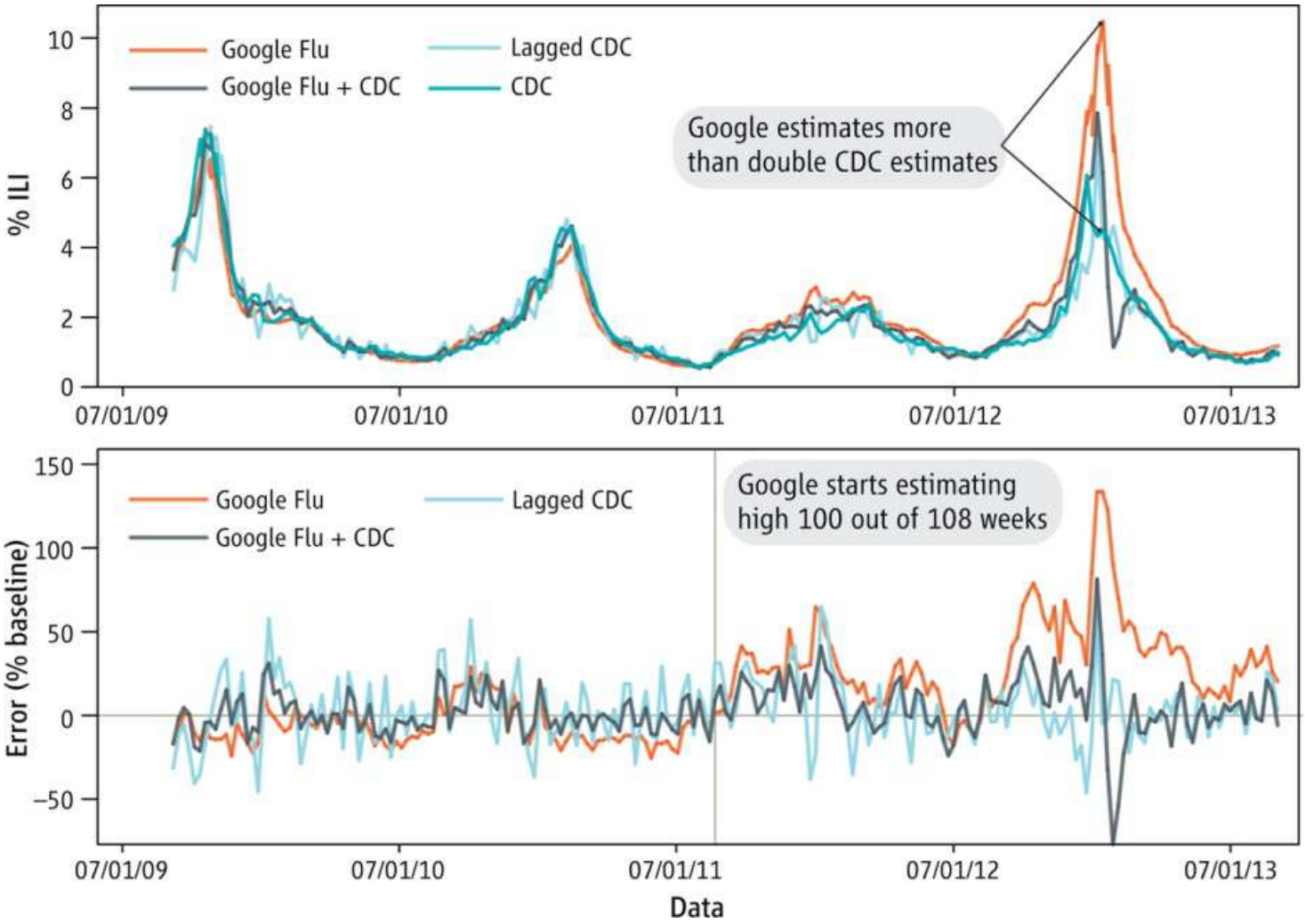
- A whole range of methodical errors

Et cetera …

*The term "Big Data Hubris" was first mentioned in Lazer et al. (2014) The Parable of Google Flu

Example – The Parable of Google Flu

GFT overestimated the prevalence of flu in the 2012–2013 season and overshot the actual level in 2011–2012 by more than 50%. From 21 August 2011 to 1 September 2013, GFT reported overly high flu prevalence 100 out of 108 weeks.



GFT
Google Flu Trends

ILI
Influenza-like Illness

CDC
Centers for Disease Control and Prevention

Lazer et al. (2014) / "Google's Big Data Flu Flop" on science.sciencemag.org

Comparison – Examples of research papers I

Quantitative empirical research	Data-driven research
Study on life satisfaction	**Sentiment analysis in Facebook**
Research question The explanation of happiness in Munich	Development question Method for sentiment analysis
Study design Analytical study Survey over several waves	Setting Big data analysis Data-driven techniques
Operationalization Questionnaire	Development Application *SentBuk*
Data ("Small Data") Population → Population of Munich (1'450'000) Random sample → 3,000 Munich households	Data ("Big Data") Data basis → Status messages from Facebook Random selection from three message classes → 140,568 "... messages written in Spanish."

Steinbeißer & Bader (2014) / Ortigosa et al. (2013)

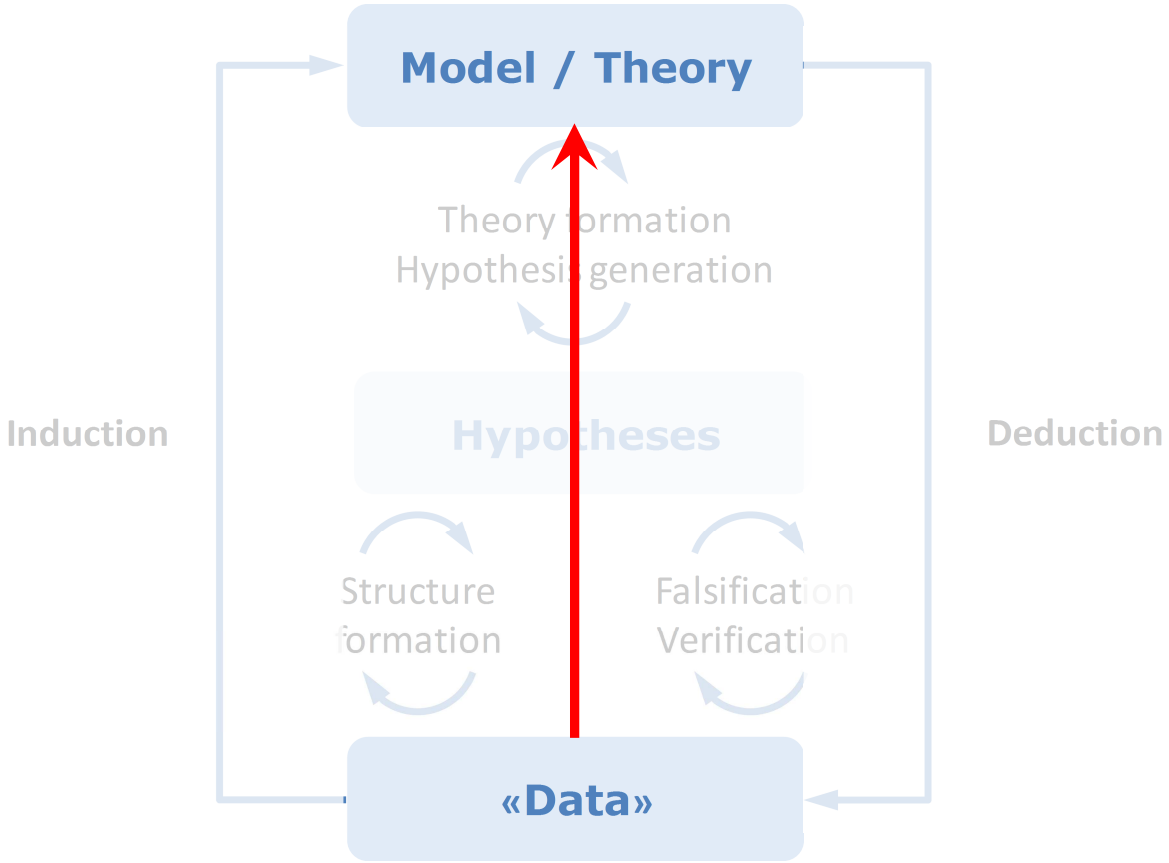
Comparison – Examples of research papers II

Quantitative empirical research	Data-driven research
Study on life satisfaction	**Sentiment analysis in Facebook**
Field phase Questionnaire distributed to households	Field phase Collecting the messages with *SentBuk*
Data analysis Descriptive statistics Inferential statistics Regression models Life satisfaction ~ age, income	Data analysis Basic analytics Decision tree Machine learning methods J48, Naive-Bayes, SVM Kernel
Result Life satisfaction among the citizens of Munich is almost the same as between 2010 and 2014 and averages 6.9 points on an eleven-point scale. The explanatory power of the models used is low …	Conclusions We have presented a new method for sentiment analysis in Facebook. The work described in this paper demonstrates that it is feasible to extract information about the student's sentiments from the messages they write in Facebook with high accuracy.

The integrated Fourth Paradigm

Integration of abduction in the Fourth Paradigm

Abductive approach



The theory is formed directly from the empirical data.

Abductive approach and combination of approaches

Neither deduction nor induction can produce an entirely new theory.
Deductive and inductive scientific research is subject to certain rules.

In the case of abduction, the process of gaining knowledge starts with data.
In contrast to induction, the patterns in the data are not systematically worked out step by step, but a new explanatory theory is formed by a sudden mental leap.

Abduction is a creative process of generating new theories from data, whereby the context and also the attitude of the researchers are particularly influential.

Induction → Search and generation of theories that fit the research context.
*Induction shows that something actually is operative.**

Deduction → Verification or falsification of existing theories.
Deduction proves that something must be.

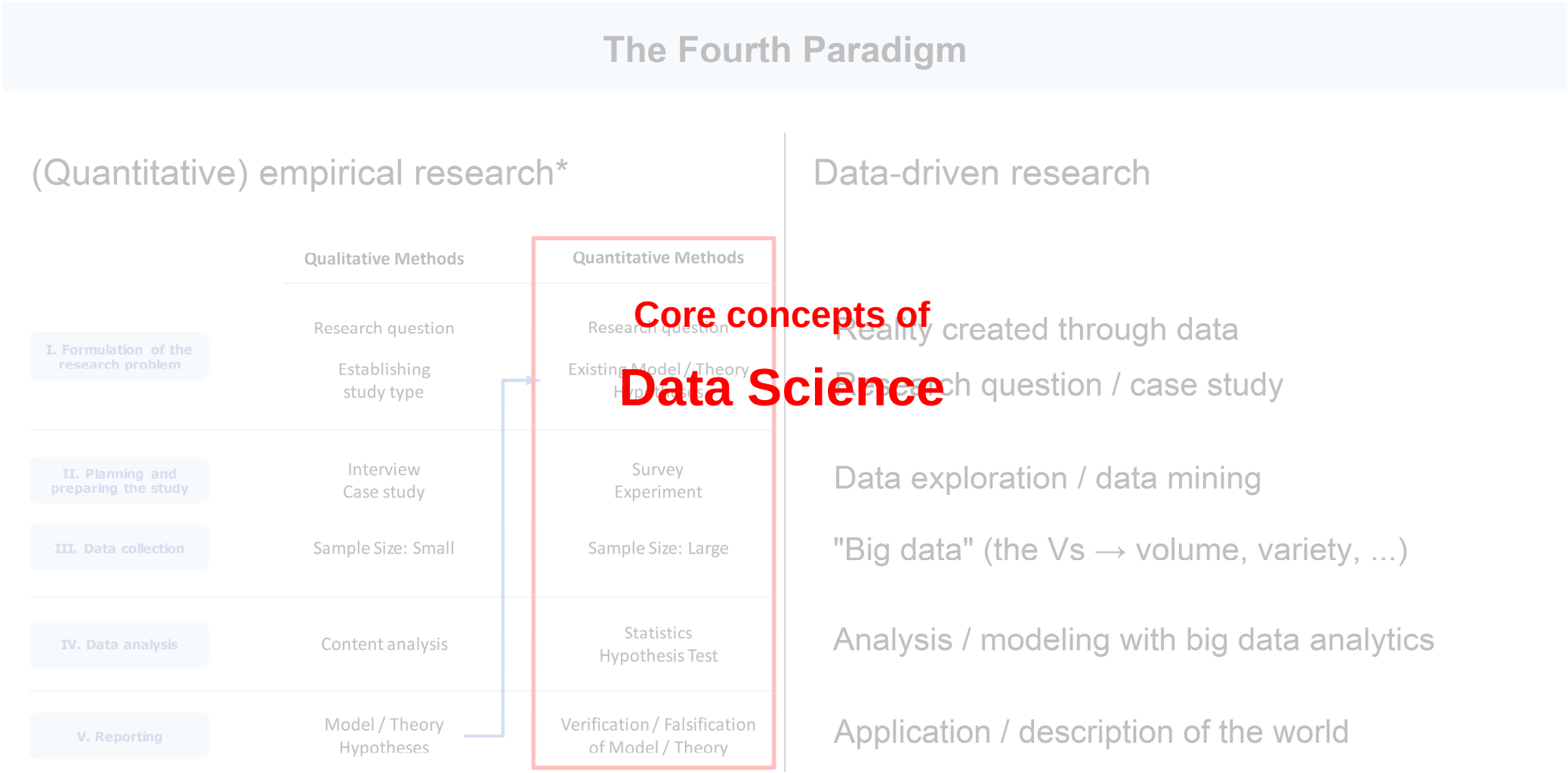
Abduction → Search and generation of **new**, **also speculative** theories.
Abduction merely suggests that something may be.

Ideally, all three methods are used cyclically

→ Triad (three steps) of scientific research

*Peirce et al. (1978) Collected papers of Charles Sanders Peirce

The integrated Fourth Paradigm



Combination of induction, deduction and abduction

Preview of Lecture 08

What has happened so far

Paradigms, paradigms …

The two main ideas are known: "The Fourth Paradigm" and "The End of Theory"

They are integrated into the research paradigms.

The current paradigm corresponds to a combination of induction, deduction and abduction.

What follows in Lecture 08

What is **A/B-Testing**?

A/B testing is used to optimize websites and applications,
especially in terms of their management and content.

It will turn out that the study design of A/B Testing is a variant of RCT.

However, there are some special features that need to be looked at in detail.

Appendix

Elements of data-driven research

Visual Analytics

- (Interactive) Data Visualization
- Exploratory Data Analysis
- (Interactive) Descriptive Statistics

Machine Learning and Predictive Modelling

- Regression / Classification / Decision Trees
- Resampling, model selection and regularization
- Support vector machines
- Neural Networks & Deep Learning

Advanced Analytics for Unstructured Data

- Deep Learning in Vision
- Pattern Recognition in Audio-Signals
- Text Mining and Analytics
- (Social) Network Analysis

Et cetera …

Reflection on "The End of Theory"? – Prompt for ChatGPT-4o

In 2008, Chris Anderson wrote the article "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete" in the online magazine www.wired.com (source: www.wired.com/2008/06/pb-theory).
From today's perspective, which arguments speak in favor of the statements in the article, and which arguments speak against the statements in the article?

Table of contents