Introduction to Data Science

Philosophical foundations of data analysis

23.03.2020, Data Science (SpSe 2022): T2

Prof. Dr. Claudius Gräbner-Radkowitsch
Europa-University Flensburg, Department of Pluralist Economics
www.claudius-graebner.com | @ClaudiusGraebner | claudius@claudius-graebner.com

Europa-Universität Flensburg

Europa-Universität Flensburg
International Institute of Management and Economic Education
Department of Pluralist Economics

This session has two parts

- Introduction of the philosophical background
- Solving your installation problems

Why do we need philosophy of science?

CHRIS ANDERSON SCIENCE JUN 23, 2008 12:00 PM

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

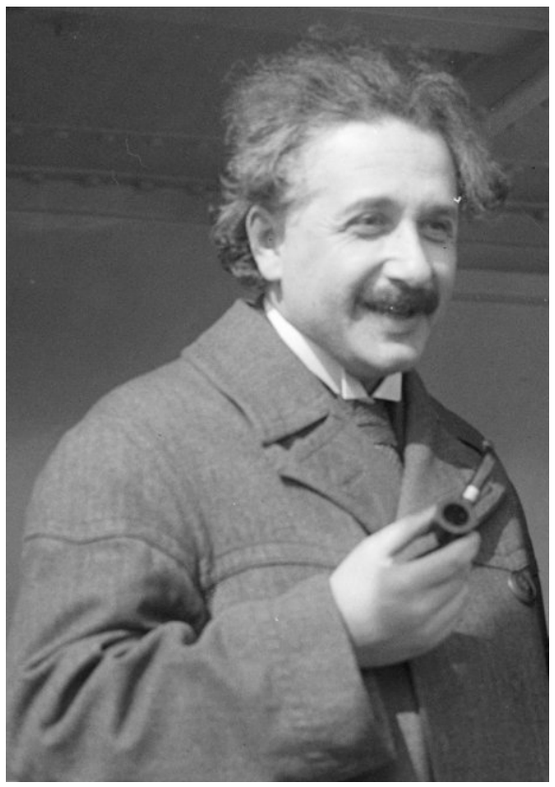
“This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. […] **With enough data, the numbers speak for themselves**.”

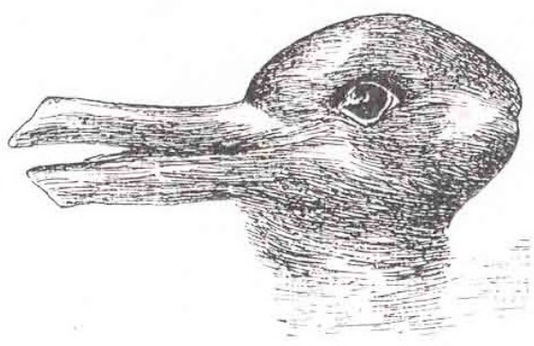
“The new availability of huge amounts of data, along with the statistical tools to crunch these numbers, offers **a whole new way of understanding the world**. […] There's no reason to cling to our old ways. It's time to ask: What can science learn from Google?”

Why do we need philosophy of science?

CHRIS ANDERSON SCIENCE JUN 23, 2008 12:00 PM

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

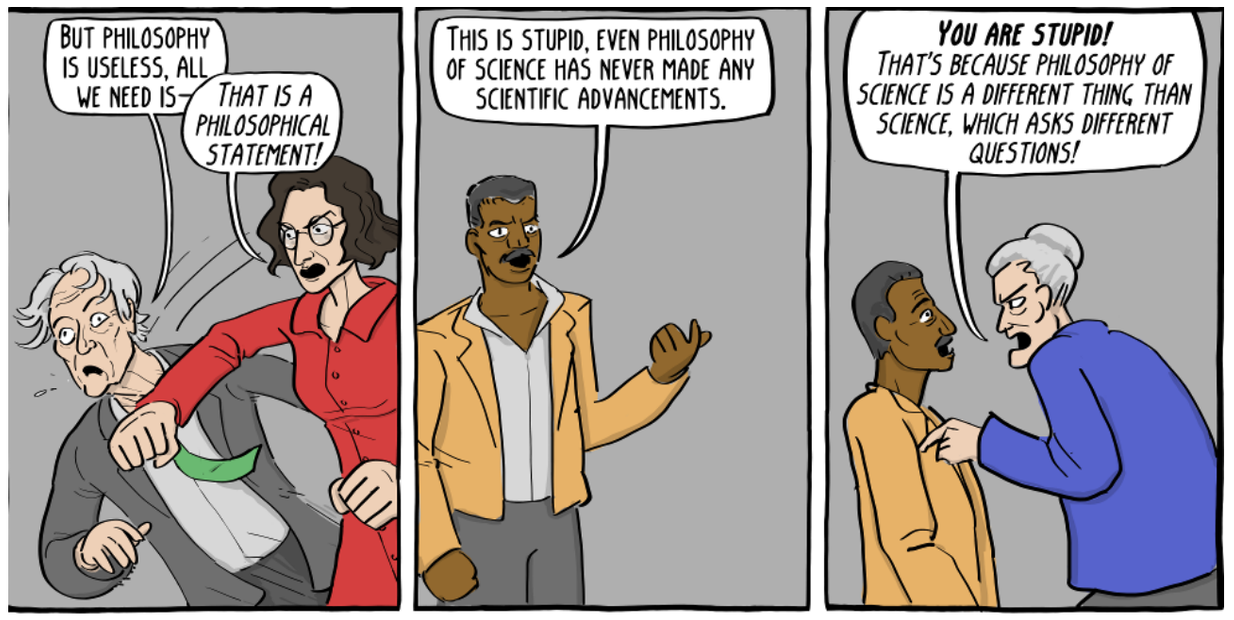


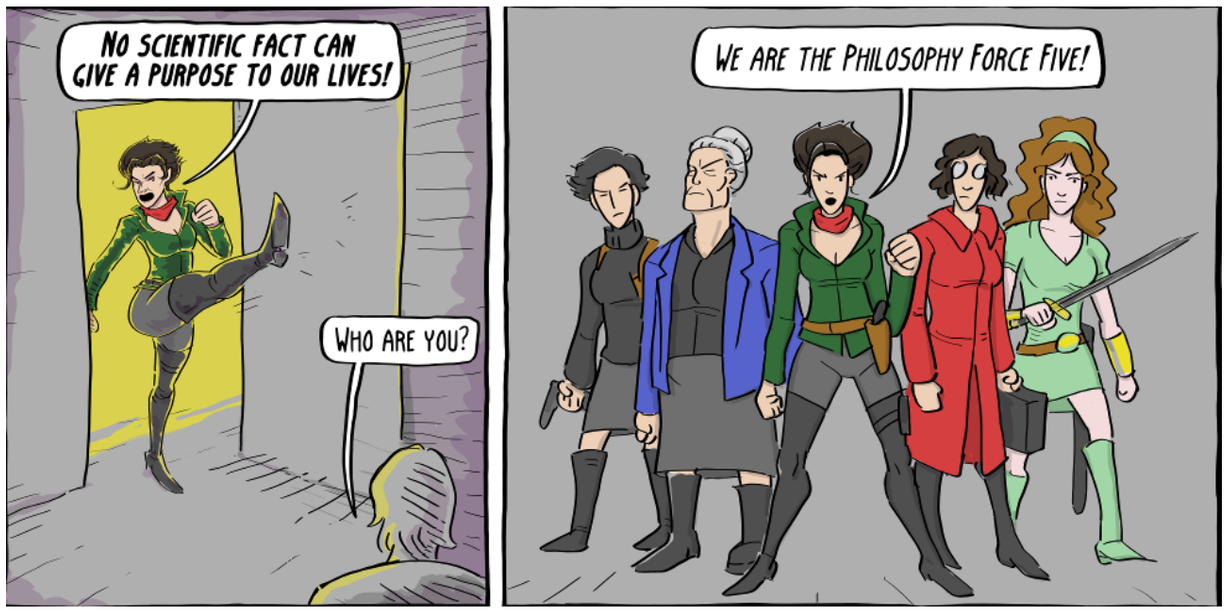


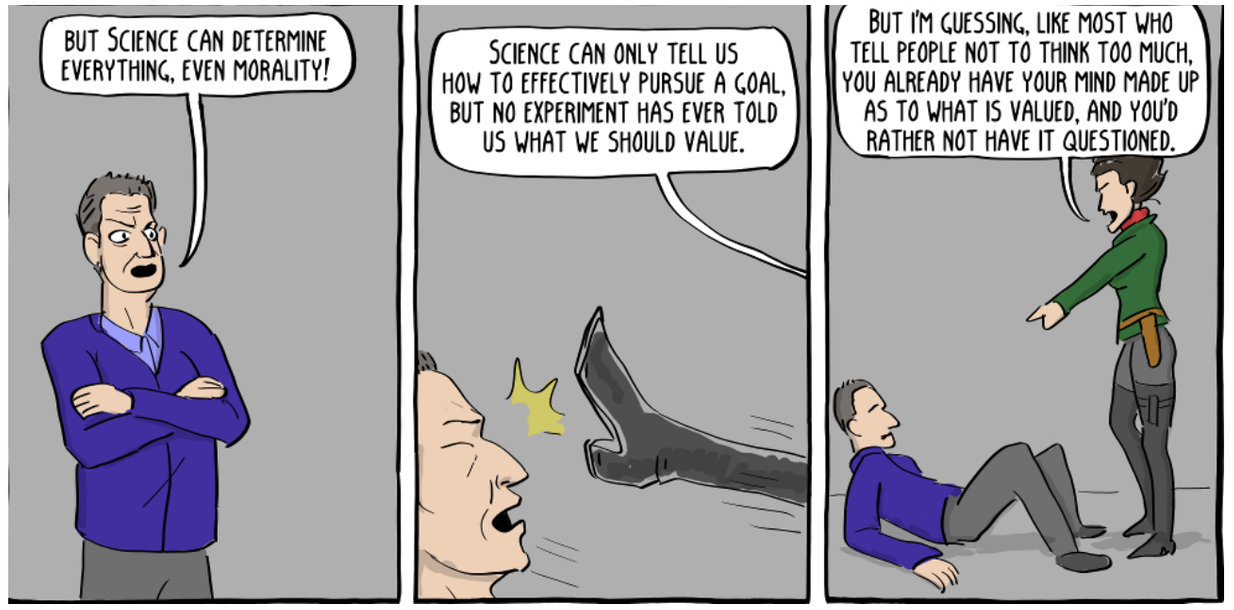
> “It is absolutely wrong to build a theory only on observations. Because it is only the prior theory that decides what we can actually observe.
>
> Albert Einstein

Why do we need philosophy of science?

We had this debate long ago…







Why do we need philosophy of science?

CHRIS ANDERSON SCIENCE JUN 23, 2008 12:00 PM

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

Philosophy of science helps you…

- …not to ridicule yourself in public!
- …to use modern tools in a reflected and responsible way!
- …to communicate with other scientists and practitioners!
- …to become a better data **scientist**!

Learning goals of this lecture

- You will learn about how to express your thought in a logical language…
  - …and how this can guide you through your data analysis
- You will learn about the difference between deduction and induction…
  - …and how both areas influence different areas of data science

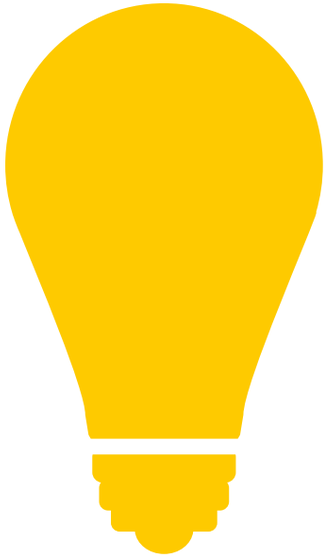
Note: we will come back to the philosophy a couple of times during the course and explore topics such as the relationship between prediction and explanation, the ways theories affects our perception of data, how we can effectively triangulate different data scientific methods, and much more…

What is philosophy of science anyways?

- **Science** seeks to generate reliable knowledge about a certain topical area by applying particular methods
  - Under which circumstances does a minimum wage cause unemployment?
  - What are the determinants of economics growth?
- **Philosophy of science** studies questions about this process of generating reliable knowledge
  - What distinguishes scientific from non-scientific knowledge?
  - Can models based on false assumptions yield true results?
  - Can empirical research lead to true statements about the world, despite our senses being imperfect?

Science and the creation of new knowledge

- Science, and data science should help in learning something new
  - *Wissenschaft* → create new knowledge
- But how to do this? What does it mean?



Asked spontaneously: what does it mean for ***you*** to learn something new about an economic phenomenon? How could you do it?

Science and the creation of new knowledge

- Science, and data science should help in learning something new
  - *Wissenschaft* → create new knowledge
- But how to do this? What does it mean?
- So far, humans came up with two fundamental ideas about how this works…

Deduction ← → **Induction**

None of this is generally superior to the other, so we somehow need to do both and come up with some clever middle ground…

Deduction

- Deduction means to derive new statements (‘**conclusions**’) from certain assumptions (‘**premises**’ / ‘assumptions’)
- This means…
  - …the truth of the premises guarantees the truth of the conclusions
  - …the informational content of the conclusion never exceeds that of the premises
  - …that deduction helps us to derive implications from true statements
  - …that deduction does not help find new true statements as such
- Lets look at some examples…

Deduction - some examples

Either R or Julia will be used in this course.
R will be used in this course.

Julia will not be used in this course.

- Logical validity and empirical correctness are fundamentally different

If two times three equals four, Flensburg is a metropolis
Two times three equals four

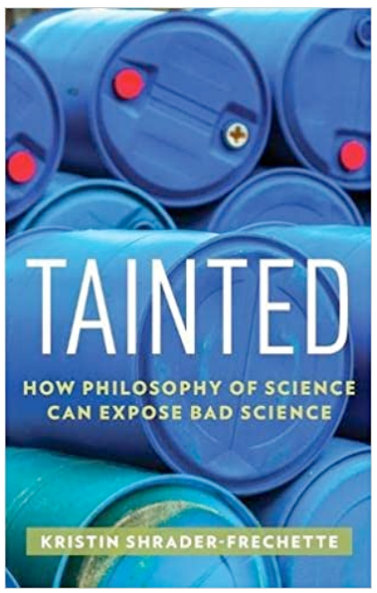
Flensburg is a metropolis

- But expressing thoughts through logic and exploiting logical rules is extremely helpful for communication and for preparing computational analysis

Note: what deduction can(not) show us

- Logical validity and empirical correctness are fundamentally different
- In fact, the rules are as follows:
  - Premises are true → Conclusion is also true
  - Premises are false → Conclusion may be true or false
  - Conclusion is false → At least one premise is false
  - Conclusion is true → Premises may be true of false
- Keeping this in mind helps in organising your thoughts!
- It is surprising how many fundamental policy flaws are due to the ignorance towards these simple rules

  → check examples in the further readings



Rules of logic

- Logical rules help us to connect premises to reach true conclusions
- Among the most widely used rules is the *modus ponens*:

p	Premise 1	*If it rains, the street will be wet*
$p \rightarrow q$	Premise 2	*It rains*
q	Conclusion	*The street is wet*

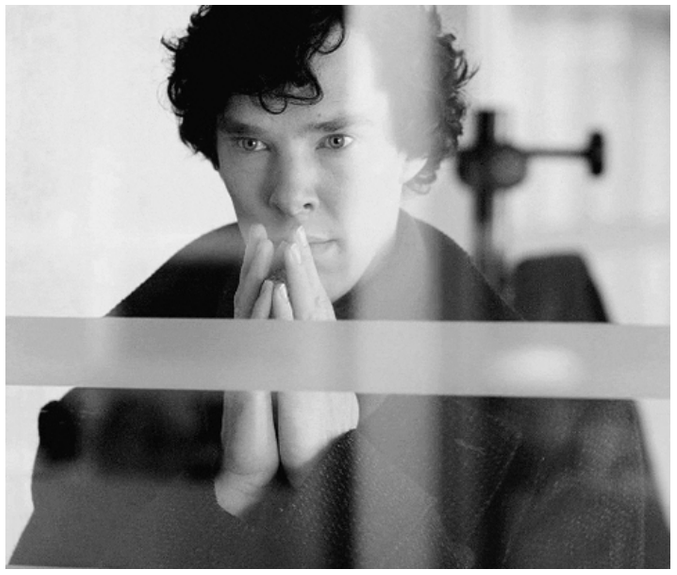
- Also frequently applicable is *modus tollens*:

$p \rightarrow q$	Premise 1	*If it rains, the street will be wet*
$\neg p$	Premise 2	*The street is not wet*
¬p	Conclusion	*It does not rain*

Reminder: logical symbols

- $\neg p$: non-p (negation)
- $p \wedge q$: p and q (conjunction)
- $p \vee q$: p or q (disjunction / inclusive or)
- $p \veebar q$: either p or q (exclusive disjunction / exclusive or)
- $p \rightarrow q$: if p then q (implication)
- $p \leftrightarrow q$: p means the same as q (if and only iff)

Logical fallacies

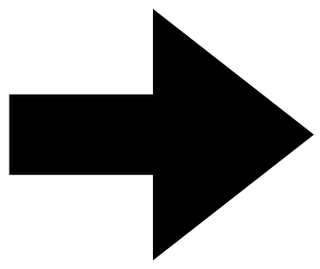


- Are these deductions correct?

If it rains, the street will be wet $p \rightarrow q$

The street is wet q

It rains p

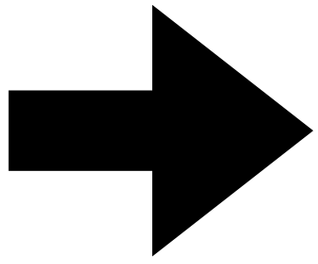


“Affirming the Consequent”

If it rains, the street will be wet $p \rightarrow q$

It does not rain $\neg p$

The street is not wet $\neg q$



“Argumentum ad ignorantiam”

→ Ignores ‘equifinality’ → There are many ways for a street to get wet!

→ Concluding p swaps $q \rightarrow p$ for $p \rightarrow q$ without reason!

→ Also ignores ‘equifinality’ → Aus $(p \rightarrow q)$ folgt nicht $(\neg p \rightarrow \neg q)$

Deduction - Summary

- Deduction is about how we can derive true conclusions from given premises
  - Allows us to derive **implications from things we know**
- Translating our thoughts into logical language…
  - …helps in communicating them
  - …is a great preparation before doing programming work
  - …helps structuring our research questions and to derive hypotheses
- Every data scientist must know basics of deduction!
- But deduction **says nothing about the empirical validity** of claims
- Deduction also does **not** help us to **come up with new premises**

Some logical quests

- What is the logical problem in the following deduction? Explicate the problem by translating it into formal language!

Industrialisation improves standard of living

Luxembourg has a high standard of living

Luxembourg is highly industrialised

- Translate the following reasoning into logical symbols and discuss it. It paraphrases a report about whether Yucca Mountain, Nevada, is a viable spot for long-term nuclear waste storage:

> The presence of tectonic activity would render a place unsuitable as a long-term nuclear-waste storage. The Department of Energy did not find any evidence for the presence of dangerous tectonic activity, which is why Yucca Mountain should be considered a viable location for a long-term nuclear waste storage.

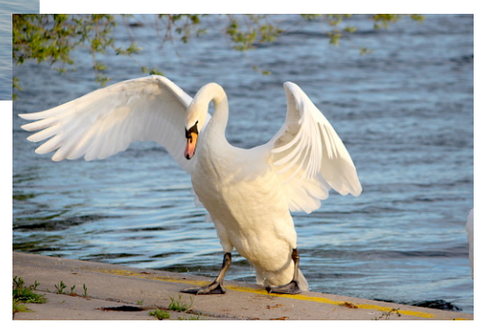
Induction

- Induction is about the generalisation of particular observations
- Inductivism: science advances by (1) observation and (2) generalisation

There is a white swan! $Sx_1 \wedge Wx_1$

The second swan there is white! $Sx_2 \wedge Wx_2$

The third swan there is white! $Sx_3 \wedge Wx_3$



$\vdots$

The *n*th swan there is white! $Sx_n \wedge Wx_n$

All swans are white! $(x) : (Sx \rightarrow Wx)$

The problem of induction

- David Hume: fundamental empirical and logical problems with induction
  - The logical problem: you generalise beyond known cases without justification:

> “[induction assumes a] principle that instances, of which we have had no experience, must resemble those, of which we have had experience, and that the course of nature continues always uniformly the same.

The first reckless passing manoeuvre was successful

The second reckless passing manoeuvre was successful

⋮

The last reckless passing manoeuvre was successful

Reckless passing manoeuvres are always successful!

} Logically as good or bad as any nonsense…

The problem of induction

- David Hume: fundamental empirical and logical problems with induction
  - The logical problem: you generalise beyond known cases without justification:

> “[induction assumes a] principle that instances, of which we have had no experience, must resemble those, of which we have had experience, and that the course of nature continues always uniformly the same.

- The empirical problem: a circular justification

The first application of induction was successful

The second application of induction was successful

⋮

The last application of induction was successful

Induction is always successful!

} Induction can justify itself only by means of itself

The problem of induction

- David Hume: fundamental empirical and logical problems with induction
  - The logical problem: you generalise beyond known cases without justification:

> “[induction assumes a] principle that instances, of which we have had no experience, must resemble those, of which we have had experience, and that the course of nature continues always uniformly the same.

  - The empirical problem: induction can justify itself only by means of itself → circularity

- The problem of induction has inspired numerous philosophical and practical innovations, several of which we will learn about later in the course!

Induction and deduction in data science

- Two fundamental ideas of **how to create knowledge**: deduction and induction
  - **Deduction** helps us to ensure logical consistency and to derive implications of known truths, but does not create new true statements
  - **Induction** helps us to process observations and conduct exploratory research, but it does not help us to prove the truth of general statements

Both deductive and inductive methods play a role in data science

Visualisation techniques and unsupervised machine learning tools

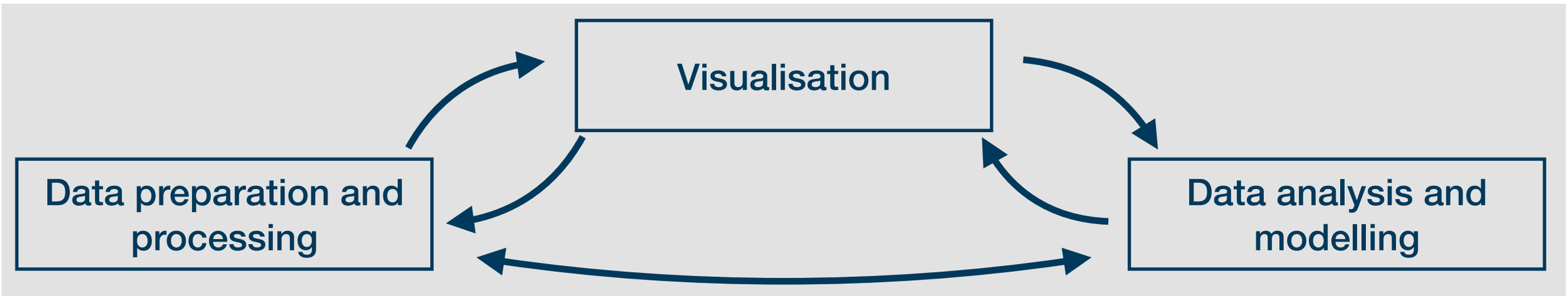
exploratory analysis → inductive spirit

Tools from supervised machine learning, regressions and hypothesis testing

explanatory analysis → deductive spirit

Summary

- Philosophy of science analyses how new knowledge can be created
  - Without philosophy of science, our thinking often remains fuzzy and full of implicit assumptions
  - Using basic tools of philosophy of science helps us structuring our thoughts and our analysis, and prepares us for the application of computational methods
- We learned about the two fundamental approaches of creating new knowledge: induction and deduction
  - Both of them are practically relevant and both have their place in data science



Questions for self study

1. What are the two central issues that David Hume discussed as the *problem of induction*?

2. Translate the following claim into logical language. How could that help your structure your further investigation of its validity?

   Our current economic system ist non sustainable because it rests upon an imperative to ever-lasting economic growth.

3. You and a friend are playing a cards game. The back of the cards is either red or blue. Your friend claims: “Did you notice that all aces have red backs?” If you want to examine the validity of this statement, should you (a) check all red cards or (b) check all blue cards? What logical rule is underlying the correct answer?