Data, Politics and Society

W2 – Data II: The Bad

UCL

This week

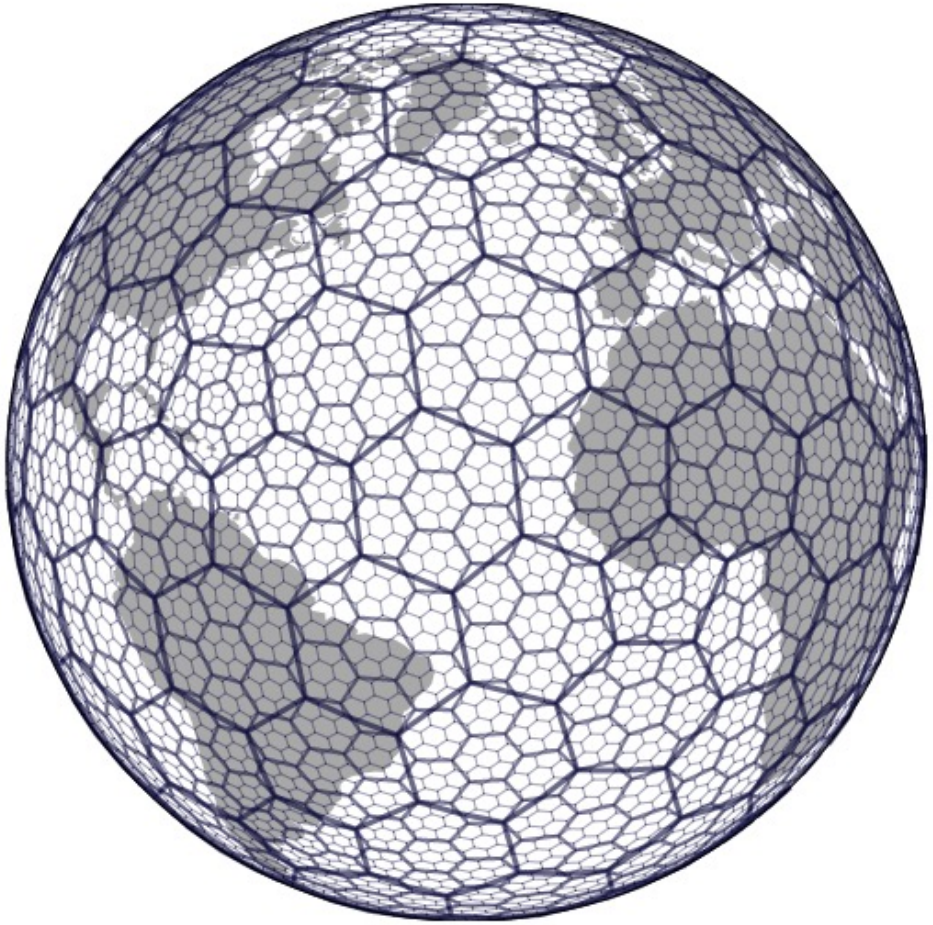
The Bad

GIS data models

Traditionally, geographic information is represented in two ways:

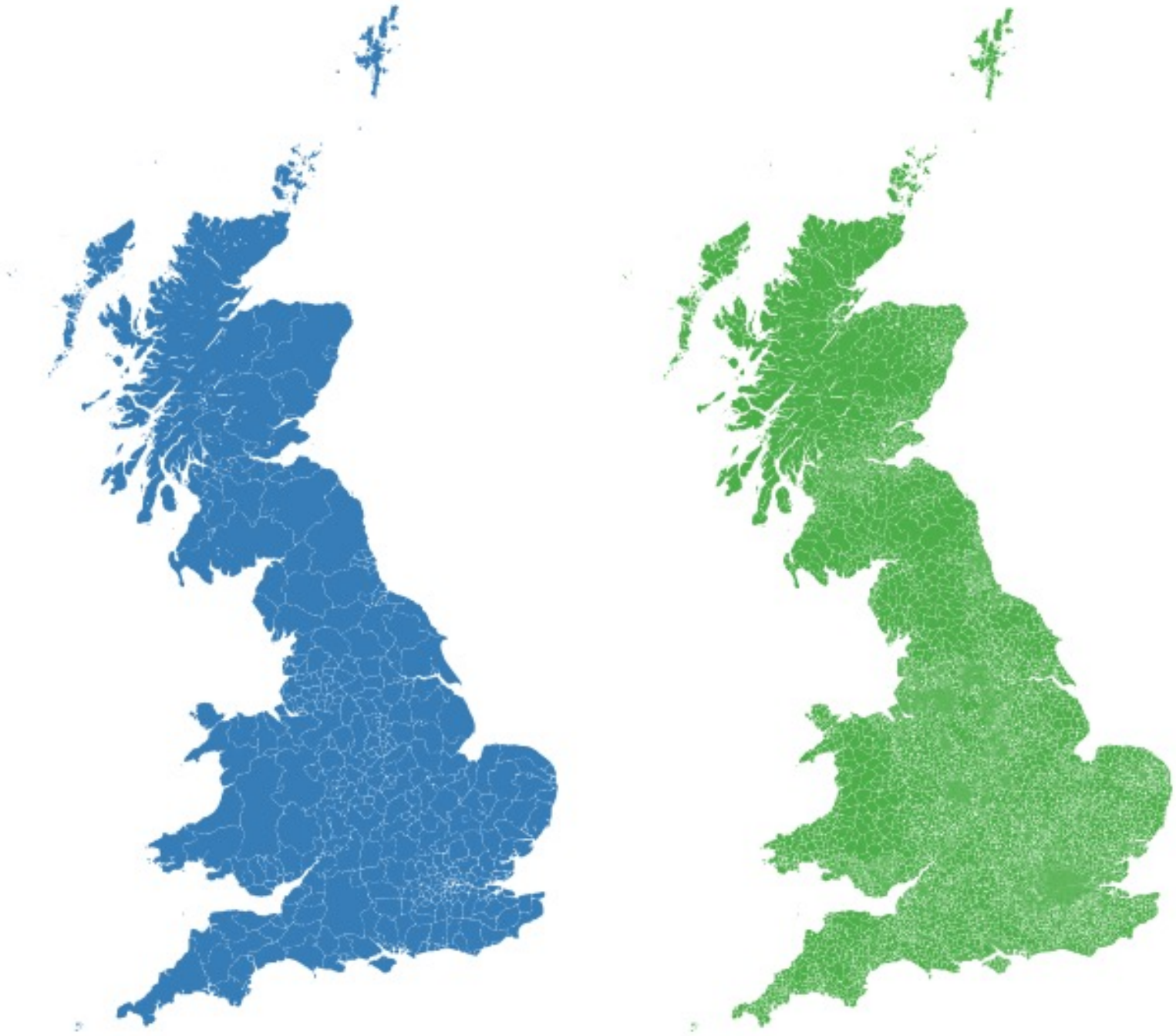
- vector: a finite set of geometric objects
- raster: images representing a surface (values, colours)

Vector I



Uber. 2018. *H3: Uber's Hexagonal Hierarchical Spatial Index*. [online] https://eng.uber.com/h3/

Vector II



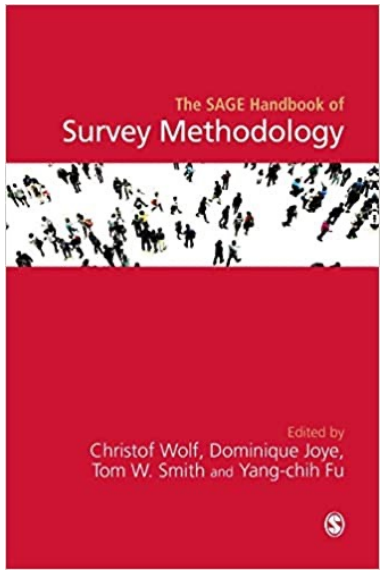
Raster

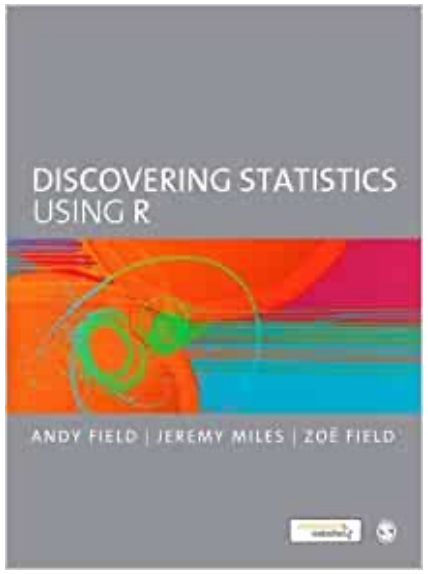
Geographic data I

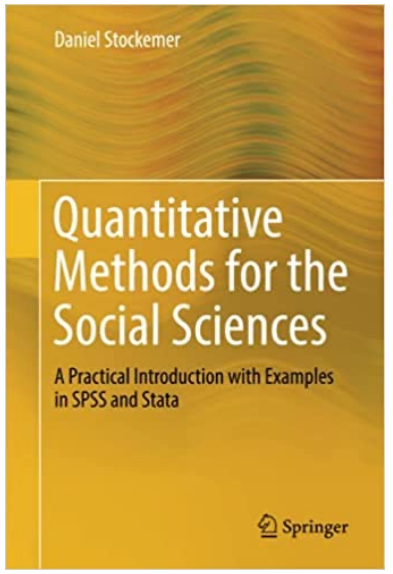
Traditionally, data sets in geography and the social sciences:

- … are collected for a specific purpose, following a careful study and design
- … contain very detailed information on a particular topic
- … are of high quality and of known provenance

Geographic data II







Geographic data III

Traditionally, data sets in geography and the social sciences:

- …very costly to administer (e.g. census, longitudinal surveys)
- …relatively poor spatial granularity (privacy preserving)
- …relatively poor temporal granularity (slow update cycles)

New geographic data I

New sources of data like the one we discussed last week tend to be:

- accidental: ‘the digital exhaust’
- diverse in quality and resolution
- arguably: higher spatial granularity?
- arguably: higher temporal granularity?

New geographic data II

Lazer and Radford 2017:

- digital life: social media (e.g. Instagram, Facebook, Twitter)
- digital traces: records of digital actions (e.g. CDR)
- digitised life: digitised records (e.g. public version of the electoral roll)

New geographic data III

Kitchen 2014:

- huge in *volume*; terrabytes of data
- high in *velocity*; being created in near real-time
- diverse in *variety*; both structured and unstructured
- *exhaustive* in scope; striving for *n = all*
- finegrained in *resolution*
- *relational* in nature; allows for conjoining different data sets
- *flexible* in terms of extensionality and scaleability

New geographic data IV

Kitchen 2014:

“Traditionally, data analysis techniques have been designed to extract insights from scarce, static, clean and poorly relational data sets, scientifically sampled and adhering to strict assumptions (such as independence, stationarity, and normality), and generated and analysed with a specific question in mind. The challenge of analysing Big Data is coping with abundance, exhaustivity and variety, timeliness and dynamism, messiness and uncertainty, high relationality, and the fact that much of what is generated has no specific question in mind or is a by-product of another activity.” (p.2)

Data bias I

Different sources of bias are contained within these ‘new’ data sources.

Data bias II

A common definition of data bias is that the available data is does not accurately represent the population or phenomenon of study. Bias can get introduced in a variety of ways.

Bias through representation I

- All data is a partial and selective representation of real-world phenomena, but normally we devise a sample scheme to collected data.
- Sampling frames are meticulously designed to reduce sample errors and sample biases.
- However, within these ‘new’ datasets: not everyone is equally represented. The primary objectives of most big datasets are not to acquire complete coverage of the population at large, thus data are prone to representing particular subsets of the population that engage with the various activities that generate data.

Bias through representation II

- Systematic distortions in demographics or other user characteristics between a population of users represented in a dataset or on a platform and some target population.
- Social media data are a good example; think Twitter, Instagram, Facebook.

Bias through quality I

- Data quality can be a source of measurement bias.
- What do we want from our data? Coverage? Completeness? Accurate? Reliable? Valid? Timely? Relevant? We want high quality.
- Data quality is a multifaceted concept with an open-ended list of desirable attributes such as completeness, correctness, and timeliness and undesirable attributes such as sparsity (lack or low amount of data) and noise (incomplete, corrupt, errors).
- Data quality is difficult to account for with ‘new’ data.

Bias through quality II

- Errors can arise in all parts of data generation and amalgamation process, from measurement to adjustment errors – which can then further contaminate other data when linked.
- Volume and velocity of ‘new’ big data prevents any efficient means of validation records.
- Simply adding more data may also increase the level of noise and reduce the quality and reliability of results.
- Sometimes quality is actually unknown.

Bias through availability

- Using whatever data is available because it is available, rather than because you believe it will truly answer your research question.
- Easier to use passive, large-scale data sets as a proxy than to set up an extensive study?

Bias through temporal factors

- Data set being limited to the time in which it is created due to systematic distortions across user populations or behaviours over time.
- Data collected at different points in time may differ along diverse criteria, including who is using the system, how the system is used.
- Some human phenomena vary seasonally – or change completely, e.g. altered movement patterns due to COVID-19.
- Not per se exclusively related to ‘new’ data (e.g. Census).

Bias through spatial factors

- Modifiable Areal Unit Problem (MAUP) – the idea that outcomes change when data or processes are summarised in some way over different spatial units.
- Digital divide – which areas are the data coming from, data are collected ‘somewhere’.

Bias through measurement

- Systematic or non-random error that occurs in the collection of data (also known as detection bias).
- For any measured variable, the difference between the true score and the observed score results from measurement error. This error is common, but it can be controlled through systematic measure development in small controlled studies.
- For ‘new’ data: users may be unaware of how study variables were measured and low reliability of some measures should be expected.

And some other possible sources of bias

Olteanu *et al*. 2019

- Linkage bias: behavioural biases that are expressed as differences in the attributes of networks obtained from user connections, interactions or activity.
- Redundancy bias: single data items that appear in the data in multiple copies, which can be identical (duplicates), or almost identical (near duplicates).
- Non-individual accounts: Interactions on social platforms that are produced by organizations or automated agents.
- But also: further biases introduced when data processing and analysing.

The problem I

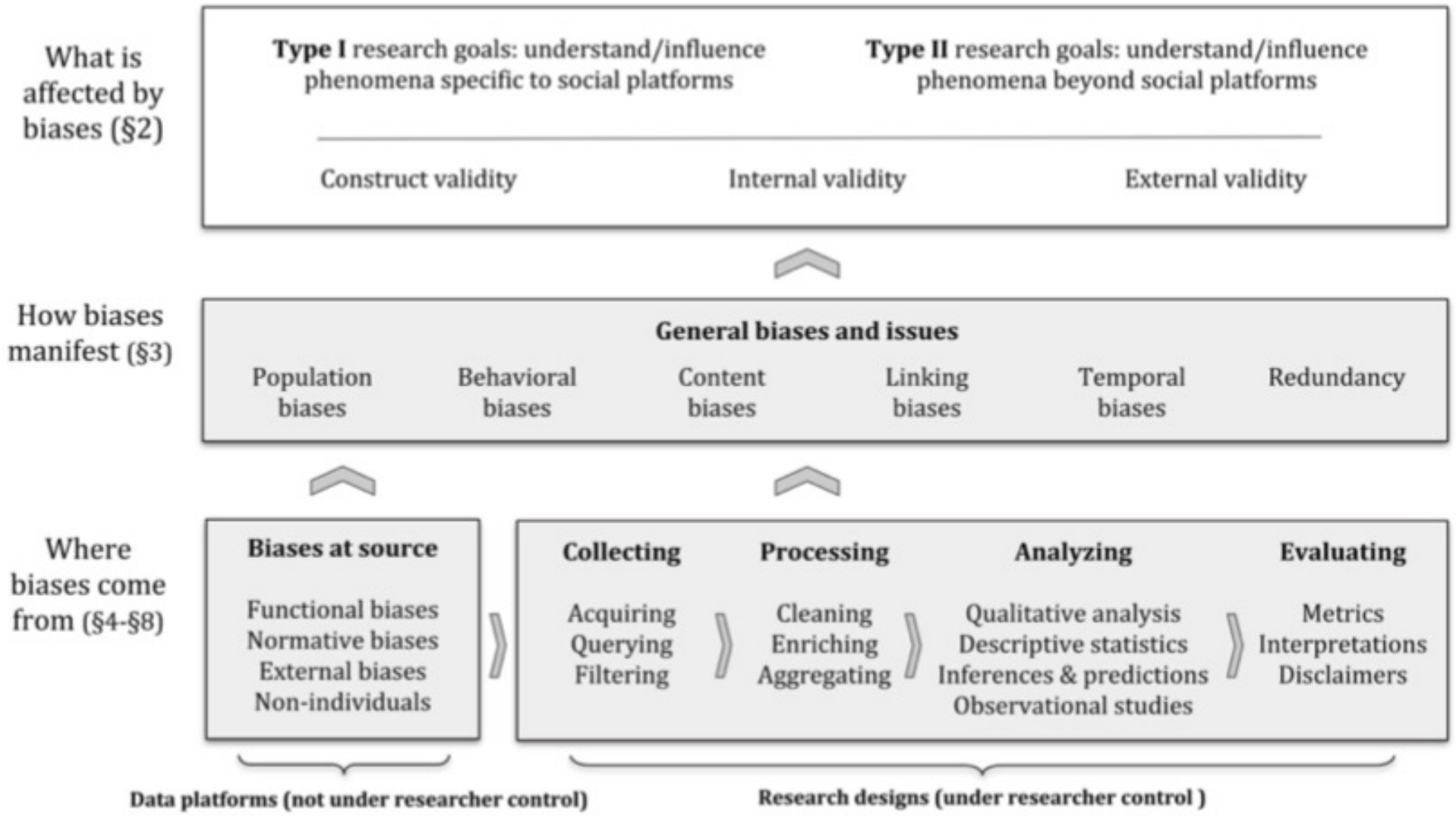
We want:

- True positives
- True negatives

Through biases we are more likely to get:

- False positives
- False negatives

The problem II



The problem III

- construct validity: do our measurements of our data measure what we think they measure?
- internal validity: does our analysis correctly lead from the measurements to the conclusions of the study?
- external validity: to what extent can research findings be generalized to other situations?

The problem IV

- epistemological challenges
- ‘what do we know and how can we know it’
- ‘what tools do we use to study the world’
- “Big Data analytics enables an entirely new epistemological approach for making sense of the world; rather than testing a theory by analysing relevant data, new data analytics seek to gain insights ‘born from the data’.” (Kitchen 2014, p.2)

Epistemology

- What is knowledge in the first place?
- What is a good basis for judgements about ‘truth’?

… but this is a controversial and contested question!

Skepticism

- Academic skeptics would argue that sensory impressions, often taken to be the foundational knowledge of the world, don't enable you to know anything.
- Zhuang Zhou: butterfly or man?
- Modern day: The Matrix?
- How can present experience prove you are not dreaming or trapped in the Matrix?
- Extreme position of knowledge being impossible.

Verification and falsification

- Logical positivism / empirical positivism
- Only statements verifiable through direct observation or logical proof are meaningful in terms of conveying truth value. Basic principle: verification / induction
- Contested by Karl Popper ('theories are never fully verified'. Basic principle: falsification/deduction
- Further contested by Duhem-Quine thesis: hypothesis do not get tested in isolation, so if a hypothesis gets refuted an entire theory should get refuted.

Some further interesting thoughts I

What does constitute as evidence for a statement?

Some further interesting thoughts II

I will prove the hypothesis that all ravens are black.

I will use my grey laptop as evidence.

Some further interesting thoughts II

(1) Hypothesis: *All ravens are black.*

(2) This can be expressed as: *If something is a raven, then it is black.*

(3) This is equivalent to: *If something is not black, then it is not a raven.*

(4) This means that for all situations where (2) is true, (1) is also true—and likewise, in all circumstances where (2) is false (1) is also false.

(5) If you own a black pet raven you could say: *My pet raven is black.*

This is then evidence supporting the hypothesis that all ravens are black. Agreed?

Some further interesting thoughts II

Now look at my grey computer. I can now say:

This grey computer is not black, and not a raven.

This supports the statement that:

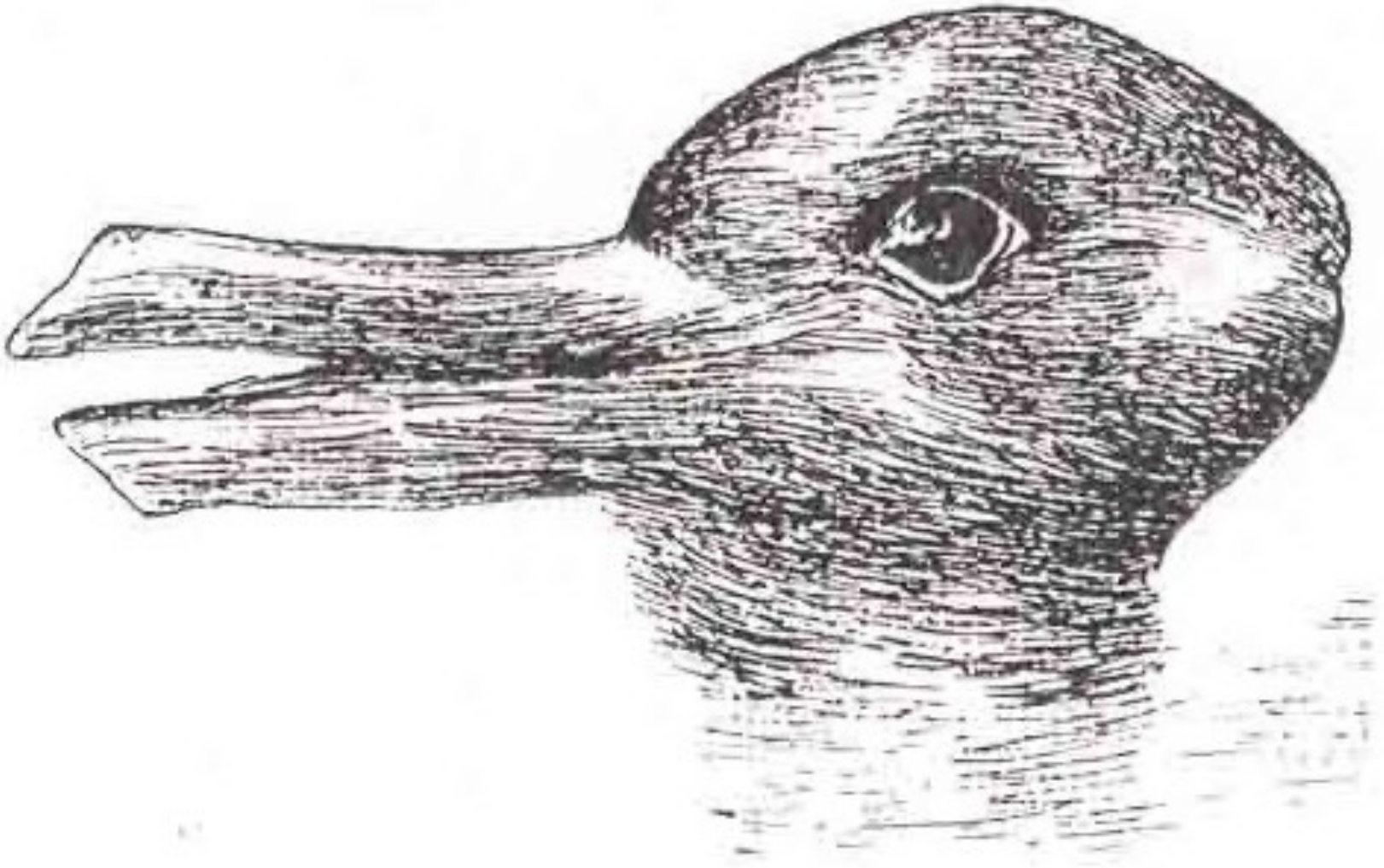
If something is not black then this is not a raven.

However, remember that this statement was equivalent to:

All ravens are black.

Hempel's paradox: what constitutes as evidence?

Some further interesting thoughts III



Thomas Kuhn

- More concerned with the ‘workings’ of science.
- Periods of ‘normal science’ with incremental advantages with occasional paradigm shifts with a sudden shift to a new explanatory framework (e.g. heliocentrism, evolution).
- A paradigm constitutes an accepted way of interrogating the world and synthesizing knowledge common to a substantial proportion of researchers in a discipline at any one moment in time
- Big Data as new research paradigm?

Kitchen I

Table 1. Four paradigms of science.

Paradigm	Nature	Form	When
First	Experimental science	Empiricism; describing natural phenomena	pre-Renaissance
Second	Theoretical science	Modelling and generalization	pre-computers
Third	Computational science	Simulation of complex phenomena	pre-Big Data
Fourth	Exploratory science	Data-intensive; statistical exploration and data mining	Now

Kitchen 2014, p.3

Kitchen II

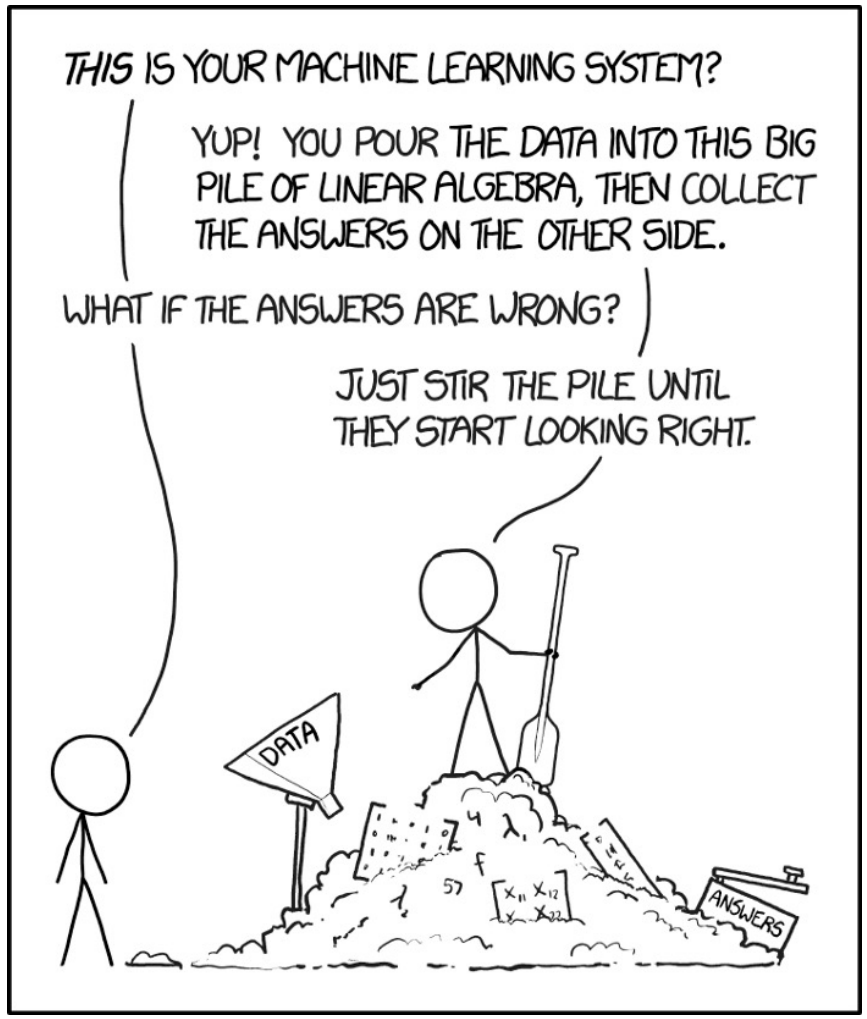
Kitchen 2014 discusses:

- Anderson: ‘the death of theory’
- ‘Return’ of empiricism and induction? But what about interpretation, how data are generated?

Kitchen III

- Data-driven science as model?
- “Whilst Big Data analytics might provide some insights, it needs to be recognized that they are limited in scope, produce particular kinds of knowledge, and still need contextualization with respect to other information, whether that be existing theory, policy documents, small data studies, or historical records, that can help to make sense of the patterns evident.” (Kitchen 2014, p.9)

Machine learning and AI?



Conclusion

- We explored some of the ‘bad’ side of human-generated datasets.
- Lots of different sources of representation and bias. Bad!
- Bias can get introduced at many different stages; think collecting, processing, analysing, evaluating.
- Clearly problematic to think of ‘new’ data as resulting in the ‘death of theory’.
- A need for a new data-driven epistemology: an approach that, grounded in scientific theories, extends their traditional approaches, adopting data and computation as an additional tool not only to test existing theories but also to develop new ones.

Seminar preparation

In preparation for the next seminar, identify and carefully read a published article that makes use of ‘new’ geographic data. *This cannot be an article that is currently on the reading list!* For this article:

- Write a 100-words summary of what you think is the article's main contribution.
- Identify at which points the dataset used may be biased, how the authors have tried to mitigate these biases, and what the authors could have done more to account for the biases you identified.
- Add it to the Google Spreadsheet linked on the module page!

Questions

Justin van Dijk

j.t.vandijk@ucl.ac.uk