Causation, probability and all that:
Data science as a novel inductive paradigm

Wolfgang Pietsch,
Munich Center for Technology in Society, TU München

Big Data in the Social Sciences, Centre for Reasoning,
University of Kent, 22.6.2017

the conceptual frontier of data science

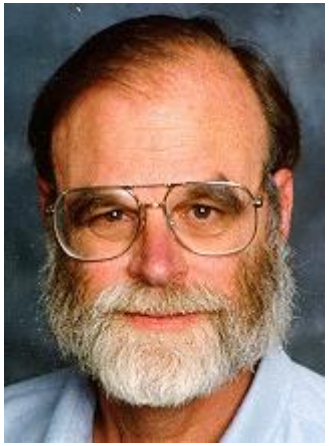
- While many data science algorithms are immensely successful, epistemological questions concerning inductivism, causation, or explanation can only be answered with a solid conceptual framework in place.
- Such a framework is still largely lacking…

Overview

- Data science – a novel inductivism?
- Two crucial insights
  - Theoretical vs. phenomenological science
  - Eliminative vs. enumerative induction
- Case study: analogical reasoning
  - Two kinds of analogy
  - Eliminative vs. enumerative approaches (or Keynes vs. Carnap)

Claims of a novel inductivism

- “The new model is for the data to be captured by instruments or generated by simulations before being processed by software and for the resulting information or knowledge to be stored in computers. Scientists only get to look at their data fairly late in this pipeline.” (Gray 2007)



- “This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory […] With enough data, the numbers speak for themselves.” (Anderson 2008)

But such inductivism is naïve!

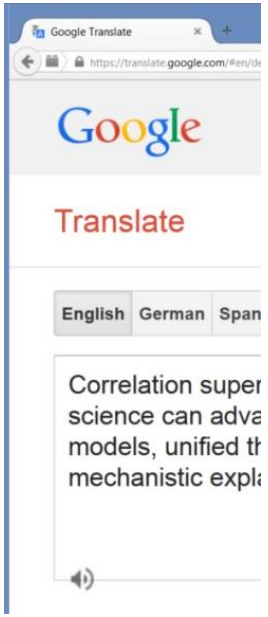
- “Inferential reasoning from data is tightly interrelated with specific theoretical commitments about the nature of the […] phenomena under investigation, as well as with experimental practices through which data are produced, tested and modelled.” (Leonelli 2012, 2)

- “With the theories and models and the scientific method in the bathwater, the baby has gone as well. Anderson’s argument is so obviously flawed that I wouldn’t have referred to it at all hadn’t it become so influential.” (Callebaut 2012, 74)

Or maybe not so naive?

There are examples of successful scientific practice, where simple models with a lot of data yield better results than complex models with comparably little data, e.g. machine translation:

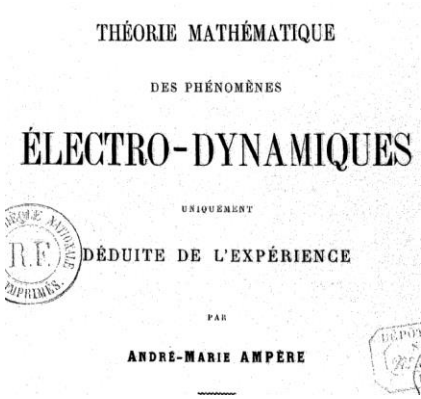
- rule-based paradigm: model both source and target languages according to grammatical structure, then try to match terms using a dictionary
- data-driven / statistical paradigm: work with probability distributions of words and sequences of words derived from large text corpora: Pr(e|f) = argmax_e Pr(e) Pr(f|e); ‚no' grammatical knowledge required



Or maybe not so naïve?

Many influential scientists have held views quite similar to those of modern data scientists, e.g. André-Marie Ampère:

- “First observe the facts, while varying the conditions to the extent possible, accompany this first effort with precise measurement in order to deduce general laws based solely on experiments, and deduce therefrom, independently of all hypotheses regarding the nature of the forces which produce the phenomena, the mathematical value of these forces, that is to say, the formula which represents them, this was the path followed by Newton.
- This was the approach generally adopted by the scholars of France to whom physics owes the immense progress which has been made in recent times, and similarly it has guided me in all my research into electrodynamic phenomena. I have relied solely on experimentation to establish the laws of the phenomena and from them I have derived the formula which alone can represent the forces which are produced […]” (Ampère 1827)



Core tenets of inductivism

- scientific laws should be proven from the phenomena, i.e. from experiment and observation
- presupposing a reliable inductive method
- these laws can eventually be considered true or at least highly probable
- implying an aversion against hypotheses, which by definition are always preliminary and never proven beyond doubt
- the accumulation of evidence continuously improves the knowledge of the phenomena
- and establishes a hierarchy of laws of increasing universality

Objections against inductivism

- Ubiquity of hypotheses in scientific practice
- Problem of induction: no epistemological justification exists for inductive inferences
- Theory-ladenness of observation: there are no pure theory-independent statements of fact
- Confirmational holism: Scientific statements cannot be confirmed or falsified in isolation
- Underdetermination of theory by evidence: a variety of theories always exists that can account for any given evidence
- …

Two core insights

- Theoretical vs. phenomenological science
  - Several of the issues mentioned on the previous slide chiefly concern theoretical science, e.g. confirmational holism and underdetermination,
  - while data science mostly belongs to the realm of phenomenological science.
- Enumerative vs. eliminative induction
  - Qualms about inductive inferences are often based on enumerative induction,
  - while in scientific practice, including many data science algorithms, eliminative induction dominates.

Theoretical vs. phenomenological science

	Phenomen. science	Theoretical science
Laws	Causal, contextual	Abstract, universal
Aim	Prediction, intervention	Explanation, conceptual framework
Phenomena	„Full“ complexity	Exemplary, paradigmatic
Method	Inductive, variation of circumstances	Abstraction
Example	Engineering sciences	Physics

[Duhem 1906, Cartwright 1983]

- **Data science mostly remains on the phenomenological level**

Eliminative vs. enumerative ind.

- Enumerative induction focuses on the repetition of instances.
- Eliminative induction focuses on the variation of circumstances, e.g. in the crucial method of difference:
  - "If an instance in which the phenomenon E under investigation occurs, and an instance in which it does not occur, have every circumstance save one C in common, that one occurring only in the former; the circumstance C in which alone the two instances differ, is the […] cause, or a necessary part of the cause, of the phenomenon." (Mill 1886)

Eliminative vs. enumerative ind.

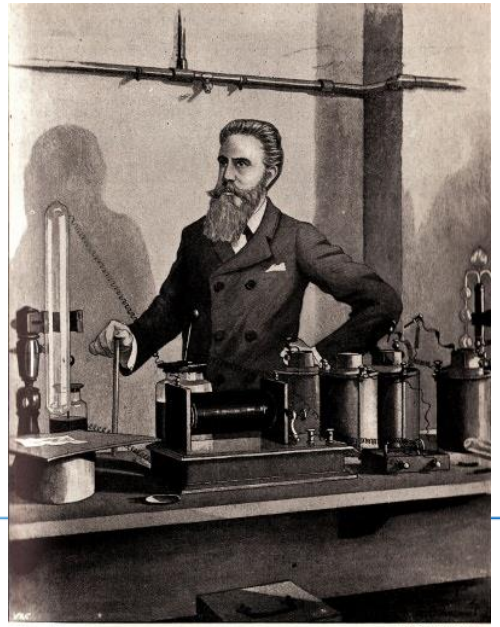
- Sophisticated inductivists are often critical of enumerative induction:
  - "[Enumerative induction] is the kind of induction which is natural to the mind when unaccustomed to scientific methods. […] It was, above all, by pointing out the insufficiency of this rude and loose conception of induction that [Francis] Bacon merited the title so generally awarded to him of Founder of the Inductive Philosophy." (Mill 1886)

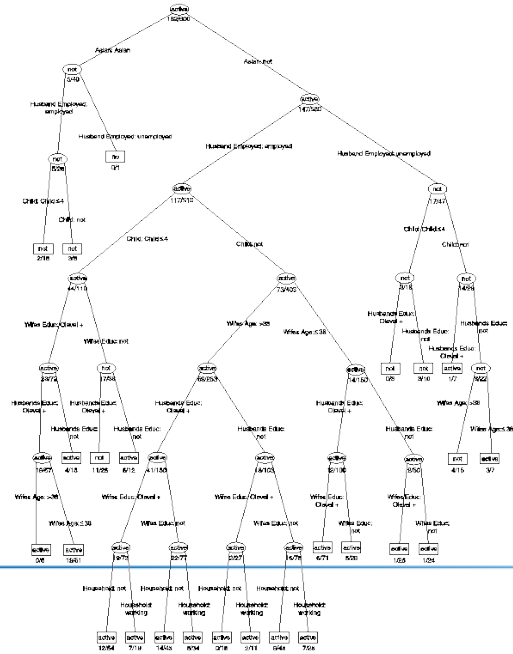
Eliminative vs. enumerative ind.

- In modern literature, a similar argument is due to Federica Russo, who advocates a rationale of variation in contrast to the prevailing rationale of regularity:
  - “The main outcome of this work is what I call the rationale of variation: quantitative causal analysis establishes causal relations by measuring variations, not by establishing regular sequences of events. I have worked hard to build empirical, philosophical, and methodological arguments to support this view.” (Russo 2009, vii)

Looking at scientific practice

- Scientific practice at the phenomenological level generally relies on eliminative induction, e.g. exploratory experimentation and data science.
- Enumerative induction plays only a marginal or subordinate role.





Taking stock

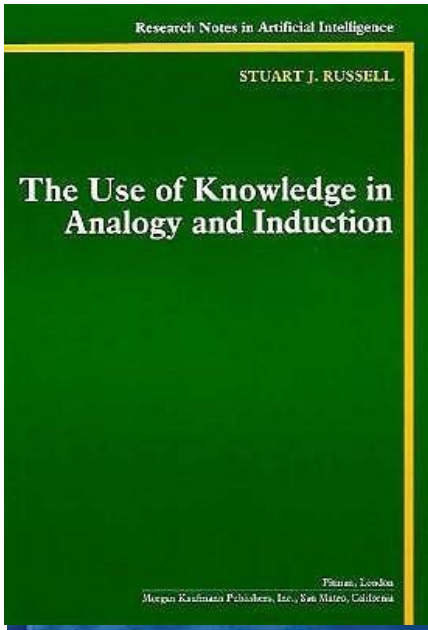
- If the mentioned two distinctions are taken into account, the prospects for an inductivism with respect to data science do not look as grim as the list of objections may have suggested:
  - Many epistemological issues such as confirmational holism or underdetermination regard chiefly theoretical science rather than phenomenological science.
  - Most arguments criticizing induction, including the hugely influential discussion by David Hume, rely on enumerative induction, while eliminative induction is primarily used in scientific practice, including in data science.

Case study: analogical reasoning

- Two types of analogy: Conceptual vs. predictive
  - The distinction between phenomenological and theoretical science suggests a distinction between two types of analogy, one aiming at concept development, the other at prediction and intervention.
- Two formal approaches
  - The distinction between enumerative and eliminative induction suggests that both types of induction could be employed to address analogical reasoning.
  - Only the latter yields a reasonable framework.

What is analogy?

- analogical inferences are inferences based on similarity:
  - If two phenomena, source A and target A*, are similar and A has a property C, under what circumstances can we assume that A* has C as well?



Analogies in theory-development

Social physics, social atoms, social forces…

- Analogies with physics have a long history in the social sciences, reaching back to Quetelet and Comte among others
- In such analogies (parts of) the structure of a well-developed, successful physical theory are transferred to the social sciences

=> this kind of reasoning is generally thought to be reserved for ‘human science’, since it involves creativity and intuition

Analogies for prediction

- The other type of analogy is not used for concept development, but rather for prediction
  - Predict which books someone might buy based on similarities with other customers
  - Predict for which candidate someone will vote based on the intentions of similar people
  - Etc.

- => most big data applications in the social sciences have a strong analogical component

Predictive vs. conceptual analogies

	Predictive analogy	Conceptual analogy
Level	Phenomenological science	Theoretical science
Aim	Reliable prediction; effective intervention	Development of a conceptual framework
Vertical relationships	Causal	Conventional, definitional
Evaluation	in terms of truth and probability	Pragmatic; *not* in terms of truth and probability
Framework	Carnap‘s continuum; eliminative induction (e.g. Keynes)	Gentner‘s structure-mapping theory
Data science	Most inferences involve predictive analogy	Conceptual analogies play a minor role

Carnap: the inadequacy of enumerative approaches

- Rudolf Carnap has developed one of the most extensive inductive frameworks in the 20th century, which explicitly aimed to include considerations of analogy.
  - It is based on the so-called straight rule of induction.
  - The confidence in a hypothesis h based on evidence e is spelled out in terms of the confirmation function c(h|e).
  - Carnap defines analogical inferences as follows: “The evidence known to us is the fact that individuals b and c agree in certain properties and, in addition, that b has a further property; thereupon we consider the hypothesis that c too has this property.” (1945, 87)
- Regarding analogy, it is largely a failure.

The continuum of inductive methods

- Based on the "straight rule" of induction, according to which the degree of confirmation is the relative frequency s_j/s of a property P_j in the first s individuals.
- Extended by Carnap to the λ-γ system:
  - $c_j(s_j, \dots, s_k) = \frac{s_j + \lambda\gamma_j}{s+\lambda}$
  - corresponding to s real and λ virtual individuals; among the latter $\lambda\gamma_j$ have the property P_j
- The confirmation function can be rewritten in terms of an empirical and a logical part:
  - $c_j(s_j, \dots, s_k) = \frac{s}{s+\lambda}\frac{s_j}{s} + \frac{\lambda}{s+\lambda}\gamma_j$
  - For large s, the empirical part dominates; for small s, the logical part (essentially representing prior considerations).

The continuum of inductive methods

- Analogy is treated in terms of the mentioned γ corresponding to the width (or weight) of properties and an additional η corr. to the distance between properties.
  - instances with properties that are closer to the predicted property confirm better than those with more distant properties
  - the greater the individual weight of a property the larger its influence
- In general, the analogy influence
  - is small
  - belongs to the logical / a priori part of the confirmation function
  - vanishes with a large number of instances

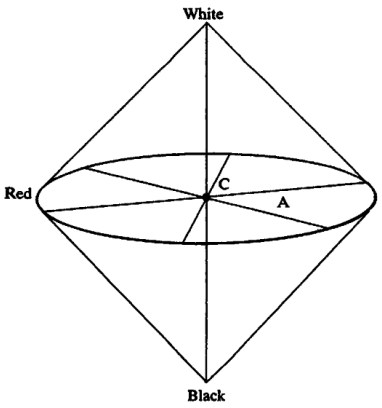


Fig. 14-1. The Color Space.

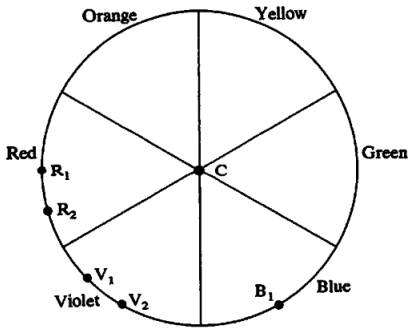
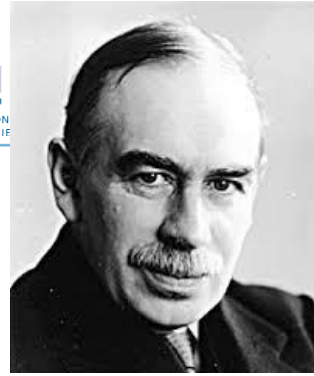


Fig. 14-2. Area A of the Color Space.



Keynes: The ubiquity of analogy

- While Carnap‘s approach, based on the straight rule, stands in the tradition of enumerative induction, John Maynard Keynes‘ framework, focusing on the variation of circumstances, broadly belongs to eliminative induction.
- According to Keynes, all inductive inferences are analogical inferences:
  - „In an inductive argument, therefore, we start with a number of instances similar in some respects AB, dissimilar in others C.
  - We pick out one or more respects A in which the instances are similar, and argue that some of the other respects B in which they are also similar are likely to be associated with the characteristics A in other unexamined cases.
  - The more comprehensive the essential characteristics A, the greater the variety amongst the non-essential characteristics C, and the less comprehensive the characteristics B which we seek to associate with A, the stronger is the likelihood or probability of the generalisation we seek to establish.” (Keynes 1920, 219-220)

Critique of enumerative approaches

- For Keynes, the variety of circumstances is important, not the number of instances
  - „In the case, however, of most scientific arguments, which would commonly be called inductive, the probability that we are right, when we make predictions on the basis of past experience, depends not so much on the number of past experiences upon which we rely, as on the degree in which the circumstances of these experiences resemble the known circumstances in which the prediction is to take effect.“ (Keynes 1920, 241)

- Enumerative induction and relatedly relative frequencies in statistics are of dubious value. Keynes explicitly rejects the straight rule:
  - „I do not myself believe that there is any direct and simple method by which we can make the transition from an observed numerical frequency to a numerical measure of probability.” (Keynes 1920, 367)

Enumerative (Carnap) vs. eliminative (Keynes)

- Carnap's approach implements
  - a clear distinction between enumerative induction and analogy
  - confines analogical influence to a priori considerations whose importance vanishes with increasing evidence
  - according to the 'principle of instantial relevance' any positive instance strictly increases confirmation
  - natural measure of confirmation exists in terms of relative frequencies of events.

- By contrast, Keynes argues that
  - all induction relies on analogy
  - seeming 'enumerative induction' controls for unaccounted circumstances
  - identical instances do not confirm at all
  - no obvious measure of confirmation exists as this would have to rely on counting properties

Taking stock

- Analogical reasoning in theoretical science is mostly heuristic, while analogical inferences in phenomenological science aim at truth and probability.
- Enumerative approaches to predictive analogical inferences have largely failed, while eliminative approaches are conceptually much more sensible.

$\Rightarrow$ Data science works with predictive analogical inferences.

$\Rightarrow$ The role of analogical reasoning in data science fits much better with a variational approach a la Keynes than with a regularity approach a la Carnap.

Conclusions

- It was argued that data science stands in an old and venerable tradition of inductivism in science.
- Classical objections against inductivism can be mitigated or weakened when taking into account two crucial distinctions, which are core tenets for an epistemology of data science: one between theoretical and phenomenological science, the other between enumerative and eliminative induction.
- The same distinctions provided useful guidelines for discussing analogical reasoning, which is crucial for many inferences in data science.
- The distinction between predictive and conceptual analogies delineates where analogical reasoning is merely heuristic and where it aims at truth or prob.
- Comparing an enumerative approach to analogy (Carnap) with an eliminative (Keynes), only the latter proved somewhat promising.