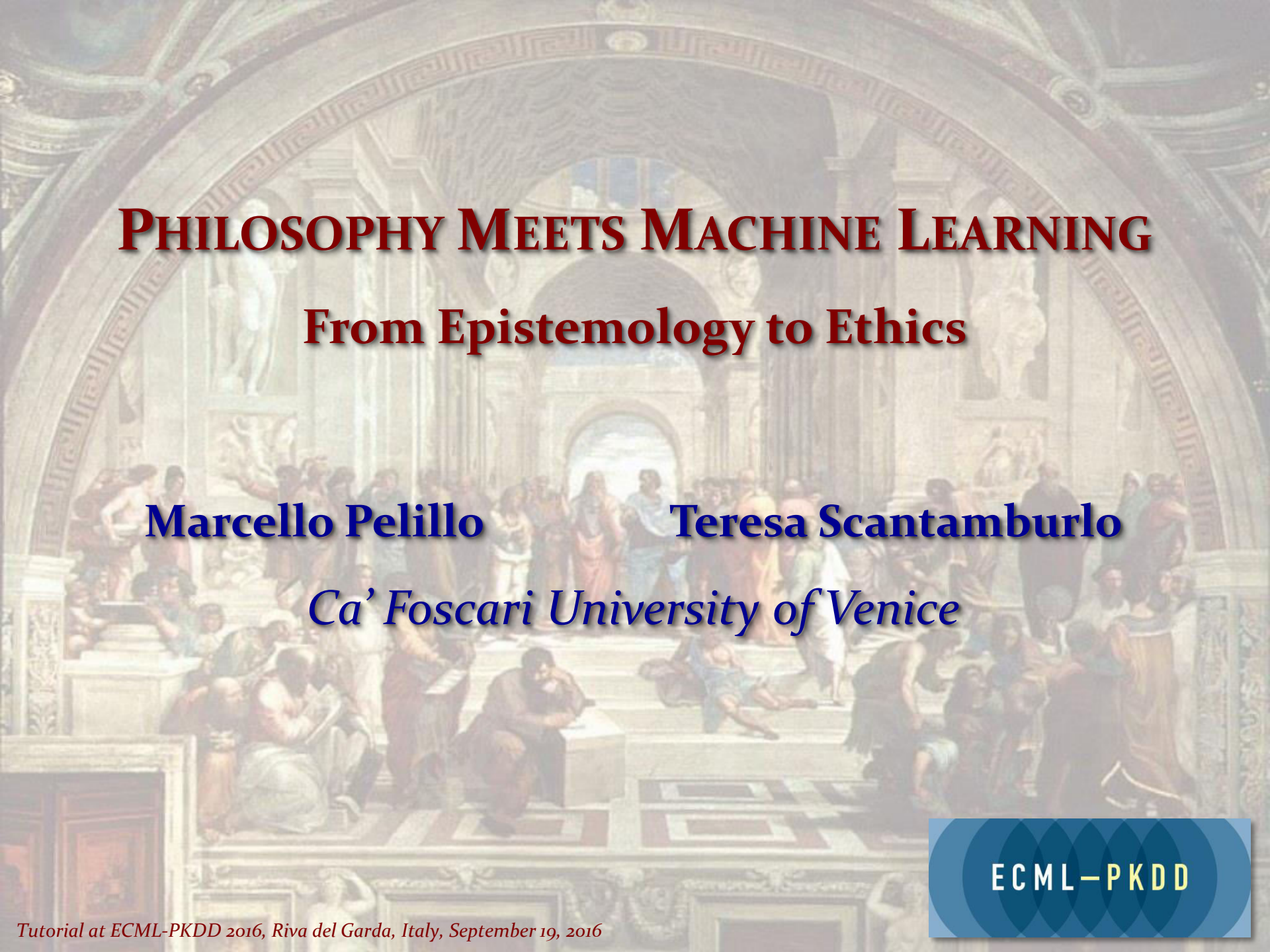


Philosophy Meets Machine Learning

From Epistemology to Ethics

Marcello Pelillo **Teresa Scantamburlo**

Ca' Foscari University of Venice

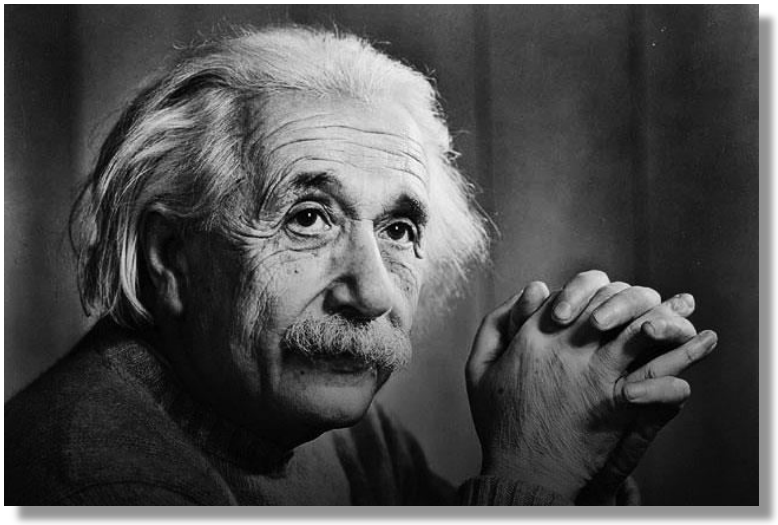
ECML–PKDD

Tutorial at ECML-PKDD 2016, Riva del Garda, Italy, September 19, 2016

Why philosophy?

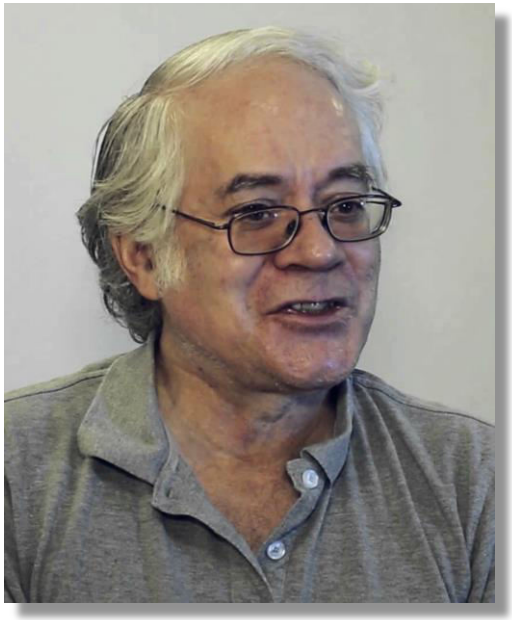
«I fully agree with you about the significance and educational value of methodology as well as history and philosophy of science. […]
A knowledge of the historic and philosophical background gives that kind of independence from prejudices of his generation from which most scientists are suffering. This independence created by philosophical insight is—in my opinion—the mark of distinction between a mere artisan or specialist and a real seeker after truth.»

Albert Einstein (1944)



Machine learning as philosophy of science

«Machine learning studies inductive strategies as they might be carried out by algorithms.
The philosophy of science studies inductive strategies as they appear in scientific practice.
[…]
the two disciplines are, in large measure, one, at least in principle.
They are distinct in their histories, research traditions, investigative methodologies; however, the knowledge which they ultimately aim at is in large part indistinguishable.»



Kevin Korb
Machine learning as philosophy of science (2004)

Tutorial outline

Part I (Pelillo): 9:00 – 10:40

- The problem of induction:
  From Bacon and Hume to Popper and Vapnik

- Scientific progress and "revolutions"

Coffee break

Part II (Scantamburlo): 11:00 – 12:40

- Introduction to ethics

- Ethics in data-driven machine learning:
  Privacy, fairness, accountability

The problem of induction

The "problem" of induction

«If we look back at the history of thinking about induction, two figures appear to stand out from the remainder.

Francis Bacon appears, as he would have wished, as the first really systematic thinker about induction;

and **David Hume** appears as perhaps the first and certainly the greatest of all inductive sceptics, as a philosopher who bequeathed to his successors a Problem of Induction.»

John R. Milton

Induction before Hume (1987)

Logical necessity?

«The bread, which I formerly eat, nourished me; [...]
but does it follow, that other bread must also nourish me at another time, and that like sensible qualities must always be attended with like secret powers?
The consequence seems nowise necessary.»

David Hume
An Enquiry Concerning Human Understanding
(1748)

Justifying induction?

«All our experimental conclusions proceed upon the supposition that the future will be conformable to the past. To endeavour, therefore, the proof of this last supposition by probable arguments, or arguments regarding existence, must be evidently **going in a circle**, and taking that for granted, which is the very point in question.»

David Hume
An Enquiry Concerning Human Understanding
(1748)

Commit it to the flames!

«If we take in our hand any volume;
of divinity or school metaphysics, for instance; let us ask,

Does it contain any abstract reasoning concerning quantity or number? No.

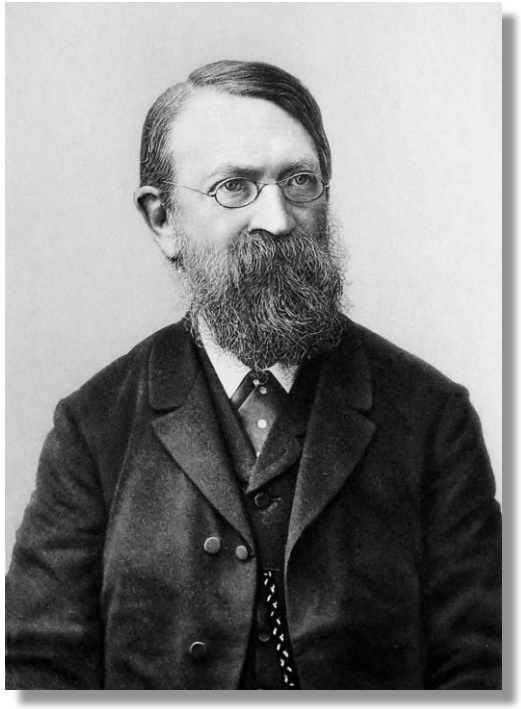
Does it contain any experimental reasoning concerning matter of fact and existence? No.

Commit it then to the flames: for it can
contain nothing but sophistry and illusion.»

David Hume
An Enquiry Concerning Human Understanding
(1748)

The elimination of metaphysics

«One and the same view underlies both my epistemological-physical writings and my present attempt to deal with the physiology of the senses — the view, namely, that all **metaphysical elements are to be eliminated** as superfluous and as destructive of the economy of science.»



Ernst Mach
The Analysis of Sensations (1897)

The Vienna circle

«After 1910 there began in Vienna a movement which regarded Mach's positivist philosophy of science as having great importance for general intellectual life

[...]

An attempt was made by a group of young men to retain the most essential points of Mach's positivism, especially his stand **against the misuse of metaphysics in science.**»

Philipp Frank

[cited by T. Uebel (2003)]

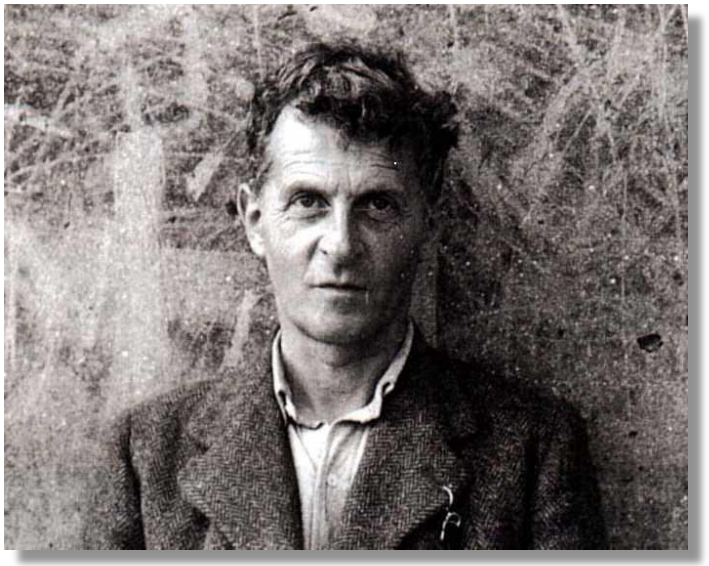
The right method of philosophy

«The right method of philosophy would be this.

To say nothing except what can be said [...]

when someone else wished to say something metaphysical, to demonstrate to him that **he had given no meaning to certain signs in his propositions**.»

Ludwig Wittgenstein
Tractatus Logico-Philosophicus (1922)



Meaning and verification

«There is only one way of giving meaning to a sentence [...] we must describe the facts which will make the proposition "true", and we must be able to distinguish them from the facts which will make it "false". [...]

In other words: **The Meaning of a Proposition is the Method of its Verification.**

The question "What does this sentence mean?" is identical with [...] the question:"how is this proposition verified?"»

Moritz Schlick

Form and Content (1938)

A polemical target

«Here is a passage taken from the writings of a famous philosopher:

"Reason is substance, as well as infinite power, its own infinite material underlying all the natural and spiritual life; as also the infinite form which sets the material in motion. Reason is the substance from which all things derive their being."
[…]

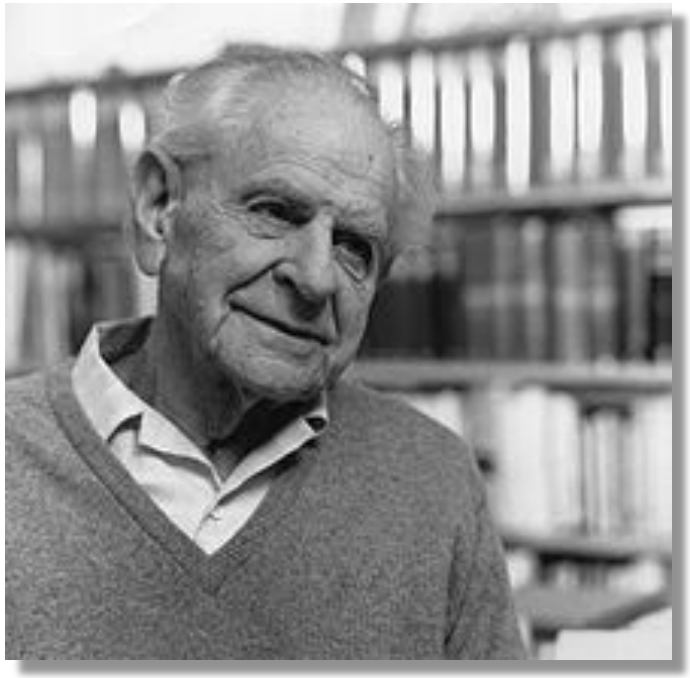
Now consider a scientist trained to use his words in such a way that every sentence has a meaning. […] What would such a man say if he read the quoted passage?»

Hans Reichenbach
The Rise of Scientific Philosophy (1951)

Against verifiability

«My criticism of the verifiability criterion has always been this: against the intention of its defenders, *it did not exclude obvious metaphysical statements; but it did exclude the most important and interesting of all scientific statements,* that is to say, the scientific theories, *the universal laws of nature.*»

Karl Popper
Conjectures and Refutations (1963)

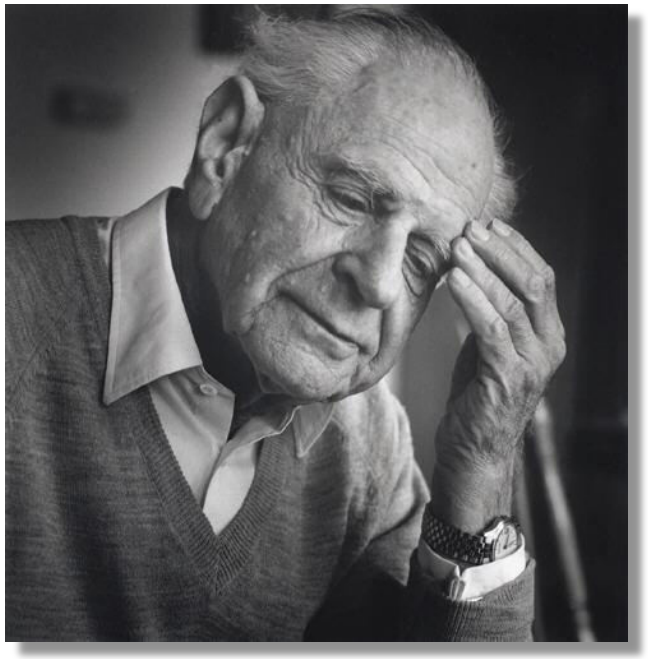


The falsifiability criterion

«My proposal is based upon an *asymmetry* between verifiability and falsifiability; an asymmetry which results from the logical form of universal statements.

For these are never derivable from singular statements, but can be contradicted by singular statements.»

Karl Popper
The Logic of Scientific Discovery (1959)



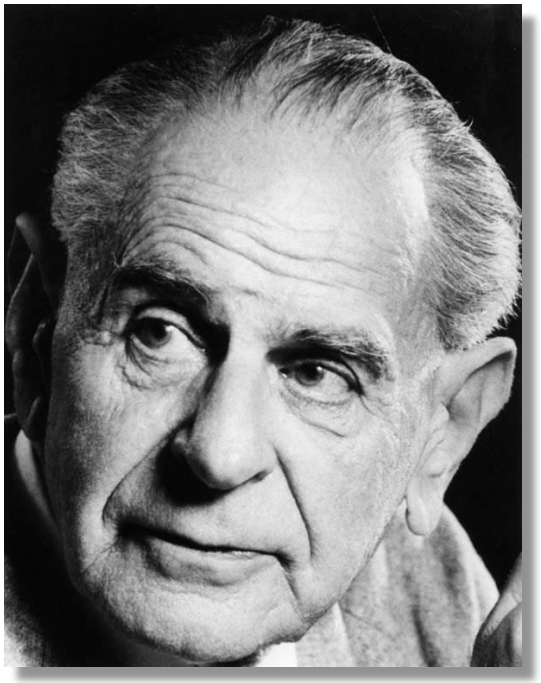
Popper on induction

«I think that I have solved a major philosophical problem:
the problem of induction.»

Karl Popper
Objective Knowledge (1972)

«Induction, i.e. inference based on many observations, is a myth.
It is neither a psychological fact, nor a fact of ordinary life, nor one of scientific procedure.»

Karl Popper
Conjectures and Refutations (1963)



Popper's scientific method

«My whole view of scientific method may be summed up by saying that it consists of these three steps:

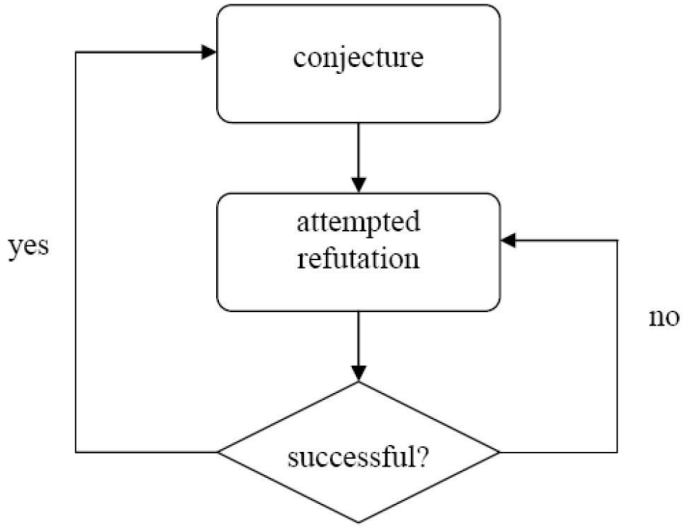
1 We stumble over some problem.

2 We try to solve it, for example by proposing some theory.

3 We learn from our mistakes, especially from those brought home to us by the critical discussion of our tentative solutions [...]

Or in three words: *problems – theories – criticism*.»

Karl Popper
The Myth of the Framework (1994)



[Wüthrich, 2010]

Feynman's version

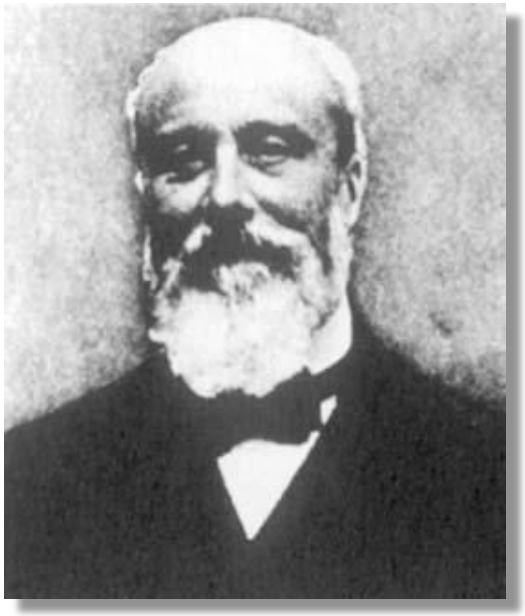
«In general we look for a new law by the following process. First we guess it. Then we compute the consequences of the guess to see what would be implied if this law that we guessed is right.
Then we compare the result of the computation to nature, with experiment or experience, compare it directly with observation, to see if it works.

If it disagrees with experiment it is wrong. In that simple statement is the key to science.»

Richard Feynman
The Character of Physical Law (1965)

The Duhem-Quine thesis

«The physicist can never subject an isolated hypothesis to experimental test, but only a whole group of hypotheses; when the experiment is in disagreement with his predictions, what he learns is that at least one of the hypotheses constituting this group is unacceptable and ought to be modified; but the experiment does not designate which one should be changed.»



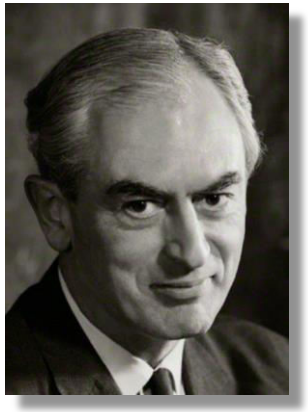
Pierre Duhem

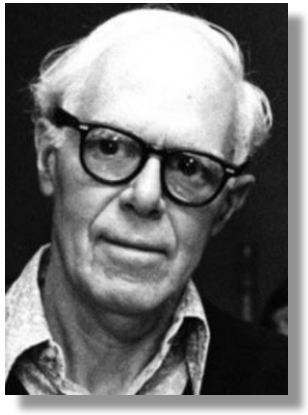
The Aim and Structure of Physical Theory (1914)

Reactions to Popper

«I think Popper is incomparably the greatest philosopher of science that has ever been.»

Peter Medawar





«Popper's great and tireless efforts to expunge the word *induction* from scientific and philosophical discourse has utterly failed.»

Martin Gardner

From verification to confirmation

«If verification is understood as a complete and definitive establishment of truth then a universal sentence, e.g a so-called law of physics or biology, can never be verified [...]

We cannot verify the law, but we can test it by testing its single instances [...]

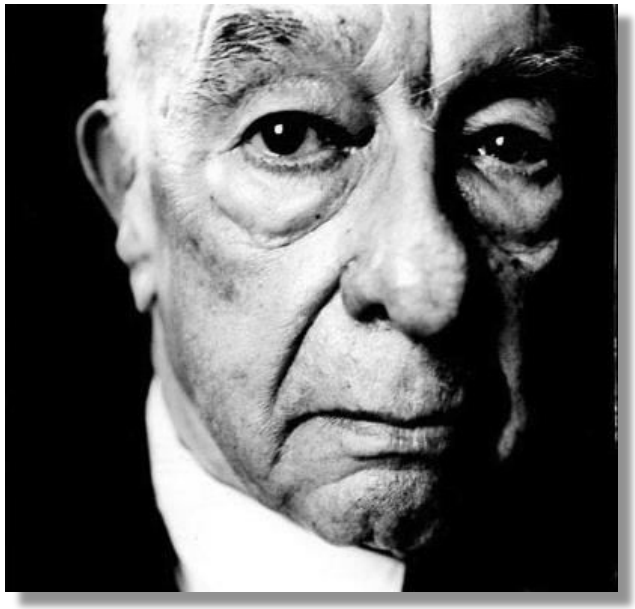
If in the continued series of such testing experiments no negative instance is found but the number of positive instances increases then our confidence in the law will grow step by step.
Thus, instead of verification, we may speak here of gradually increasing *confirmation* of the law.»

Rudolf Carnap
Testability and meaning (1936)

The paradoxes of confirmation

«What tends to confirm an induction?
This question has been aggravated on the one hand by Hempel's puzzle of the non-black non-ravens, and exacerbated on the other by Goodman's puzzle of the grue emeralds.»

Willard V. O. Quine
Natural kinds (1969)



From black ravens ...

Nicod's principle: Universal generalizations are supported or confirmed by their positive instances and falsified by their negative instances.

Example.

A black raven confirms the hypothesis "*All ravens are black*"

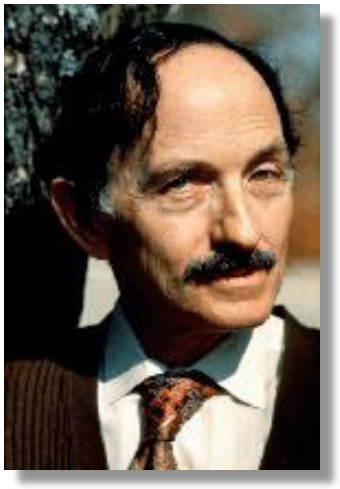
Equivalence principle: Whatever confirms a generalization confirms as well all its logical equivalents.

Example.

$\forall x$ (Ax → Bx) is logically equivalent to $\forall x$ (~Bx → ~Ax)

Hence, the hypothesis "*All ravens are black*" is logically equivalent to "*All non-black things are non-ravens*"

… to white shoes and indoor ornithology



«Hempel's paradox of confirmation can be worded thus 'A case of a hypothesis supports the hypothesis. Now the hypothesis that all crows are black is logically equivalent to the contrapositive that **all non-black things are non-crows**, and this i**s supported by the observation of a white shoe.**'»

Irving J. Good
The white shoe is a red herring (1967)

«The prospect of being able to investigate ornithological theories without going out in the rain is so attractive that we know there must be a catch in it.»

Nelson Goodman
Fact, Fiction, and Forecast (1955)

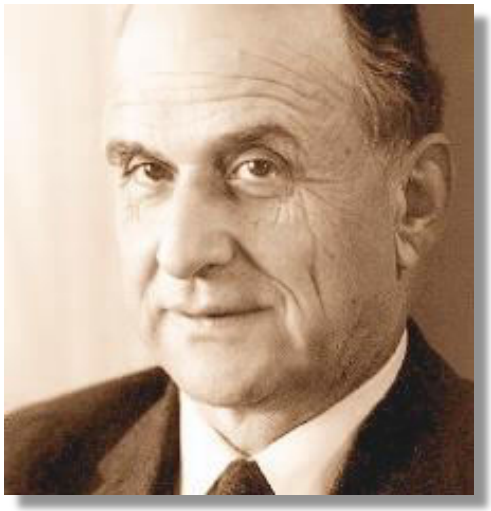
Lawlike statements?

«That a given piece of copper conducts electricity increases the credibility of statements asserting that other pieces of copper conduct electricity [...]

But the fact that a given man now in this room is a third son does not increase the credibility of statements asserting that other men now in this room are third sons [...]

Yet in both cases our hypothesis is a generalization of the evidence statement. **The difference is that in the former case the hypothesis is a lawlike statement**; while in the latter case, the hypothesis is a merely contingent or accidental generality.»

Nelson Goodman
Fact, Fiction, and Forecast (1955)



Goodman's new riddle

Argument 1:

PREMISE	All the many emeralds observed prior to 2017 AD have been green
CONCLUSION	***All emeralds are green***

Argument 2:

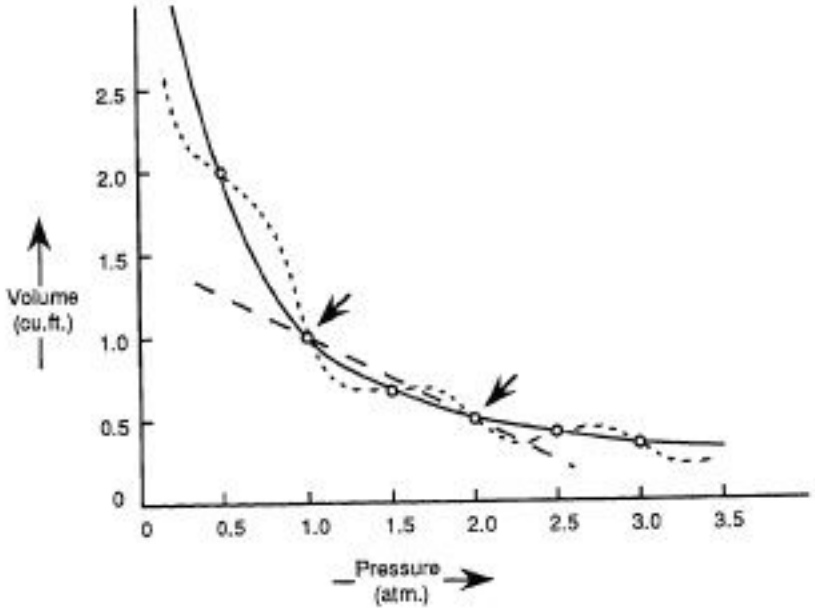
PREMISE	All the many emeralds observed prior to 2017 AD have been "grue"
CONCLUSION	***All emeralds are "grue"***

Definition: Any object is said to be *grue* if:

- ✓ it was first observed before 2017 AD and is green, or
- ✓ it was *not* first observed before 2017 AD and is blue

If all evidence is based on observations made before 2017 AD, then the second argument should be considered as good as the first …

Goodman's riddle and model selection



Boyle's Law (solid line) and alternative laws.

There's always an infinity of mutually contradictory hypotheses that fit the data, but which is best confirmed?

***Customary answer**:* choose the simplest one (Occam's razor). But … why?

The probabilistic turn

«I am convinced that it is impossible to expound the methods of induction in a sound manner, without resting them upon the theory of probability.

Perfect knowledge alone can give certainty, and in nature perfect knowledge would be infinite knowledge, which is clearly beyond our capacities. We have, therefore, to content ourselves with partial knowledge—knowledge mingled with ignorance, producing doubt.»

William S. Jevons
The Principles of Science (1874)

But … what does "probability" mean?

Classical view (*Laplace, Pascal, J. Bernoulli, Huygens, Leibniz, …*)

Probability = ratio # favorable cases / # possible cases

Frequentist view (*von Mises, Reichenbach, …*)

Probability = limit of relative frequencies

Logical view (*Keynes, Jeffreys, Carnap, …*)

Probability = logical relations between propositions ("partial implication")

Subjectivist view (*Ramsey, de Finetti, Savage*, …)

Probability = a (personal) agent's "degree of belief "

But also: Propensity (Popper), Best-system (Lewis), …

Bayesianism to the rescue?

«Through much of the twentieth century, the unsolved problem of confirmation hung over philosophy of science. What is it for an observation to provide evidence for, or confirm, a scientific theory? […]
The situation has now changed. Once again a large number of philosophers have real hope in a theory of confirmation and evidence. The new view is called *Bayesianism*.»

Peter Godfrey-Smith
Theory and Reality (2003)

The three tenets of Bayesianism

Bayesian confirmation theory (BCT) makes the following assumptions:

1. It is assumed that agents assigns *degrees of belief*, or credences, to different competing hypotheses, reflecting the agent's level of expectation that a particular hypothesis will turn out to be true

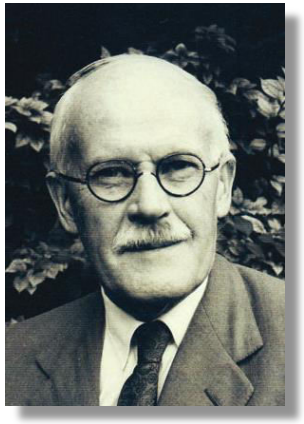
2. The degrees of belief are assumed to behave mathematically like probabilities, thus they can be called *subjective probabilities*

3. Agents are assumed to learn from the evidence by what is called the *Bayesian conditionalization rule*. The conditionalization rule directs one to update his credences in the light of new evidence in a quantitatively exact way

In BCT, evidence *e* confirms hypothesis *h* if:

$$P(h \mid e) > P(h)$$

Bayes' theorem



«[Bayes' theorem] is to the theory of probability what Pythagoras' theorem is to geometry.»

Harold Jeffreys
Scientific Inference (1931)

$$P(h \mid e) = \frac{P(e \mid h)P(h)}{P(e)} = \frac{P(e \mid h)P(h)}{P(e \mid h)P(h) + P(e \mid \quad h)P(\quad h)}$$

- ✓ $P(h)$: prior probability of hypothesis h
- ✓ $P(h \mid e)$: posterior probability of hypothesis h (in the light of evidence e)
- ✓ $P(e \mid h)$: "likelihood" of evidence e on hypothesis h

Bayesians' answer to confirmation paradoxes

The ravens: White shoes do in fact confirm the hypothesis that all ravens are black, but only to a negligible degree.

The grue emeralds: Both hypotheses ("green" and grue") are OK, but most people would assign a higher prior to the "green" hypothesis than to the "grue" one. (But… why is it so?)

The Bayesian "machine"

- ✓ determine the prior probability of h
- ✓ if e_1 is observed, calculate the likelihood $P(e_1 \mid h)$ and the probability to observe e_1 independently of h
- ✓ calculate the posterior probability $P(h \mid e_1)$ via Bayes' theorem
- ✓ consider this posterior probability as your new prior probability of h
- ✓ if e_2 is observed, calculate the likelihood $P(e_2 \mid h)$ and the probability to observe e_2 independently of h
- ✓ calculate the posterior probability $P(h \mid e_2)$ via Bayes' theorem
- ✓ consider this posterior probability as your new prior probability of h
- ✓ …

Challenges to Bayesianism

Priors. Where do they come from? Also, initial set of prior probabilities can be chosen freely ⇒ how could a strange assignment of priors be criticized, so long as it follows the axioms?

Old evidence. Existing evidence can in fact confirm a new theory, but according to Bayesian kinematics it cannot (e.g., the perihelion of Mercury and Einstein's general relativity theory).

If e is known before theory T is introduced, then we have $P(e) = 1 = P(e|T)$, which yields:

$$P_{new}(T \mid e) = \frac{P(T)P(e \mid T)}{P(e)} = P(T)$$

⇒ posterior probability of T is the same as its prior probability!

Solomonoff induction

«Solomonoff completed the Bayesian framework by providing a rigorous, unique, formal, and universal choice for the model class and the prior.»

Marcus Hutter
On universal prediction and Bayesian confirmation (2007)

Basic ingredients:

- ✓ Epicurus (keep all explanations consistent with the data)
- ✓ Occam (choose the simplest model consistent with the data)
- ✓ Bayes (combine evidence and priors)
- ✓ Turing (compute quantities of interest)
- ✓ Kolmogorov (measure simplicity/complexity)

Data expressed as binary sequences
Hypotheses expressed as algorithms (processes that generate data)

Bad news: Solomonoff induction is intractable …. (use approximation)

A long-lasting debate

«The dispute between the Bayesians and the anti-Bayesians has been one of the major intellectual controversies of the 20th century.»

Donald Gillies, *Was Bayes a Bayesian?* (2003)

«All that can be said about 'inductive inference' [...], essentially, reduces [...] to Bayes' theorem.»

Bruno De Finetti, *Teoria della probabilità* (1970)

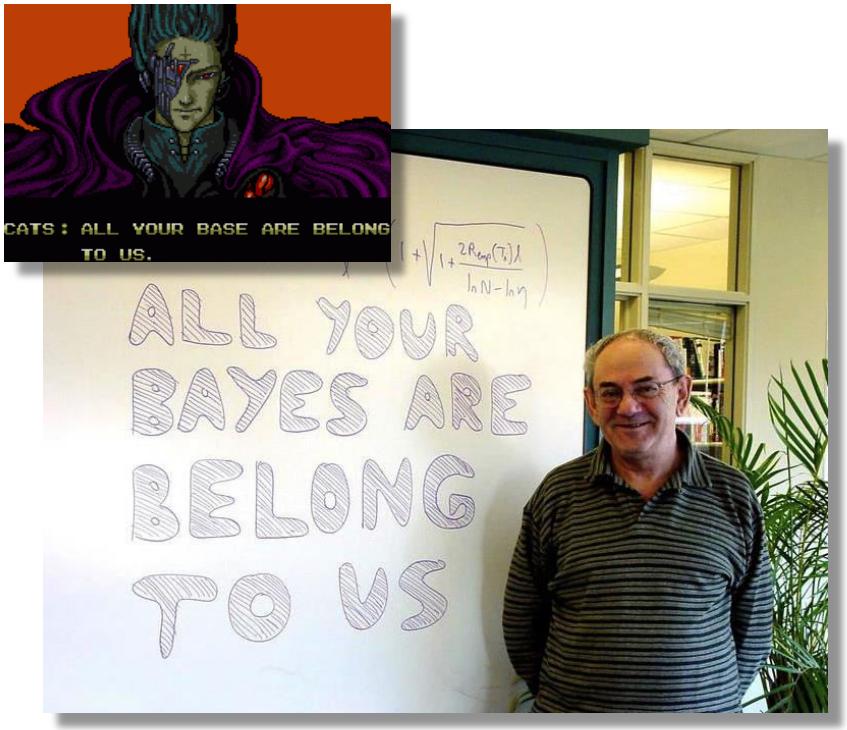
«The theory of inverse probability is founded upon an error, and must be wholly rejected.»

Ronald A. Fisher
Statistical Methods for Research Workers (1925)

Statistical learning theory

«Between 1960 and 1980 a revolution in statistics occurred: Fisher's paradigm, introduced in the 1920's and 1930's was replaced by a new one. This paradigm reflects a new answer to the fundamental question:

What must one know a priori about an unknown functional dependency in order to estimate it on the basis of observations?



In Fisher's paradigm the answer was very restrictive—one must know almost everything. [...]
The new paradigm overcame the restriction of the old one.»

Vladimir Vapnik
The Nature of Statistical Learning Theory (2000)

The formal setup of SLT

SLT deals mainly with **supervised learning** problems.

Given:

- ✓ an input (feature) space: $\mathcal{X}$
- ✓ an output (label) space: $\mathcal{Y}$ (typically $\mathcal{Y} = \{ -1, +1 \}$)

the question of learning amounts to estimating a functional relationship between the input and the output spaces:

$$f : \mathcal{X} \to \mathcal{Y}$$

Such a mapping f is called a **classifier**.

In order to do this, we have access to some (labeled) training data:

$$(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \mathcal{Y}$$

A **classification algorithm** is a procedure that takes the training data as input and outputs a classifier f.

Assumptions

In SLT one makes the following assumptions:

- ✓ there exists a joint probability distribution P on $\mathcal{X} \times \mathcal{Y}$
- ✓ the training examples (X_i, Y_i) are sampled independently from P (iid sampling).

In particular:

1. No assumptions on P
2. The distribution P is unknown at the time of learning
3. Non-deterministic labels due to label noise or overlapping classes
4. The distribution P is fixed

Losses and risks

We need to have some measure of "how good" a function f is when used as a classifier. A *loss function* measures the "cost" of classifying instance $X \in \mathcal{X}$ as $Y \in \mathcal{Y}$.

The simplest loss function in classification problems is the **0-1 loss** (or misclassication error):

$$\ell(X, Y, f(X)) = \begin{cases} 1 & \text{if } f(X) \neq Y \\ 0 & \text{otherwise.} \end{cases}$$

The *risk* of a function is the average loss over data points generated according to the underlying distribution P:

$$R(f) := E(\ell(X, Y, f(X)))$$

The *best classifier* is the one with the smallest risk $R(f)$.

Bayes classifiers

Among all possible classifiers, the "best" one is the *Bayes classifier*:

$$f_{Bayes}(x) := \begin{cases} 1 & \text{if } P(Y = 1 \mid X = x) \geq 0.5 \\ -1 & \text{otherwise.} \end{cases}$$

In practice, it is impossible to directly compute the Bayes classifier as the underlying probability distribution *P* is unknown to the learner.

Goal: Determine a function $f : \mathcal{X} \to \mathcal{Y}$ which has risk $R(f)$ as close as possible to the risk of the Bayes classifier.

Caveat. Not only is it impossible to compute the Bayes error, but also the risk of a function *f* not be computed without knowing *P*.

A desperate situation?

Empirical Risk Minimization (ERM)

Instead of looking for a function which minimizes the true risk $R(f)$, we try to find one which minimizes the *empirical risk*:

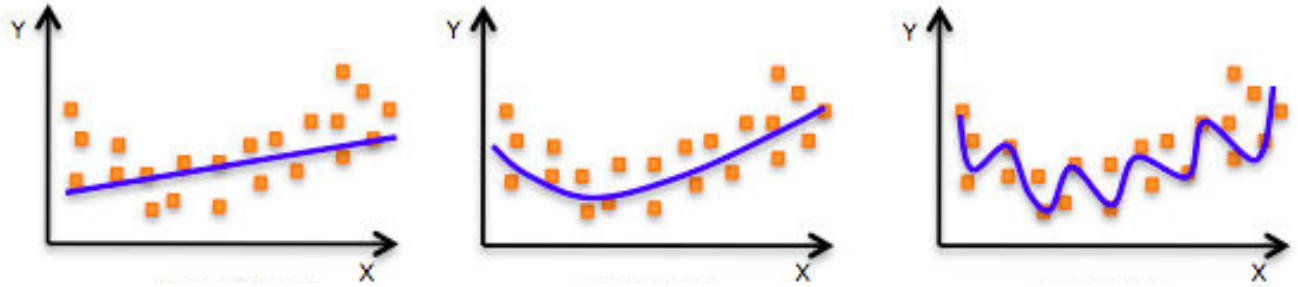
$$R_{\text{emp}}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(X_i, Y_i, f(X_i))$$

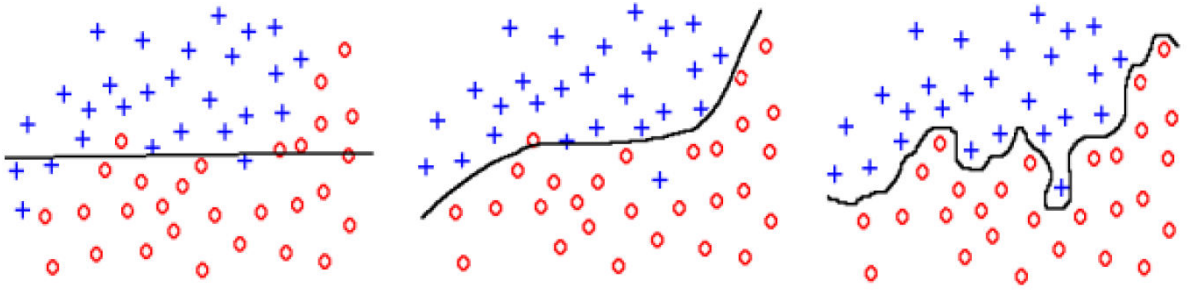
This approach is called the *empirical risk minimization* (ERM) induction principle, the motivation of which comes from the law of large numbers.

Key question: What has to be true of the function class $\mathcal{F}$ so that, no matter what the probability distribution, ERM eventually does as well as possible with respect to the rules in $\mathcal{F}$?

A fundamental result of SLT is that the set of rules in $\mathcal{F}$ cannot be too rich, where the richness of $\mathcal{F}$ is measured by its *VC dimension*.

Overfitting *vs.* underfitting





Small complexity of $\mathcal{F}$ $\Rightarrow$ underfitting / Large complexity of $\mathcal{F}$ $\Rightarrow$ overfitting

The best overall risk is achieved for "moderate" complexity

Shattering

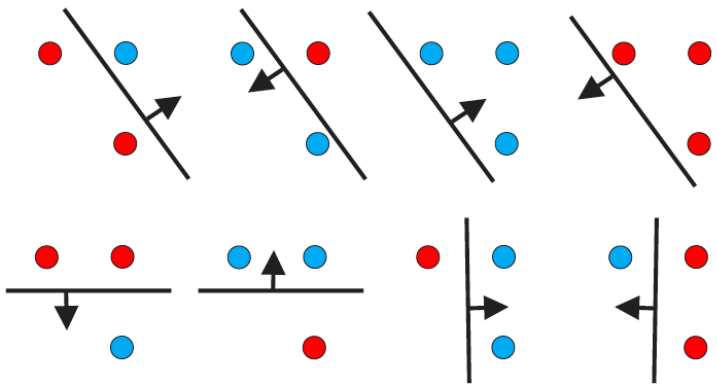
A set of n instances $X_1,\ldots, X_n$ from the input space $\mathcal{X}$ is said to be *shattered* by a function class $\mathcal{F}$ if all the 2^n labelings of them can be generated using functions from $\mathcal{F}$.

Example.

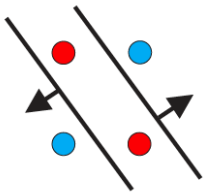
$\mathcal{F}$ = linear decision functions (straight lines) in the plane

(a) Any set of 3 non-collinear points shatters $\mathcal{F}$

(b) No set of 4 points can shatter $\mathcal{F}$



(a)



(b)

The VC dimension

The *VC dimension* of a function class $\mathcal{F}$, denoted $\mathrm{VC}(\mathcal{F})$, is the largest integer h such that *there exists* a sample of size h which is shattered by $\mathcal{F}$.

If arbitrarily large samples can be shattered, then $\mathrm{VC}(\mathcal{F}) = \infty$.

Examples.

- ✓ $\mathcal{F}$ = linear decision functions in $\mathbf{R}^2$ $\Rightarrow$ $\mathrm{VC}(\mathcal{F}) = 3$
- ✓ $\mathcal{F}$ = linear decision functions (hyperplanes) in $\mathbf{R}^n$ $\Rightarrow$ $\mathrm{VC}(\mathcal{F}) = n + 1$
- ✓ $\mathcal{F}$ = multi-layer perceptrons with W weights $\Rightarrow$ $\mathrm{VC}(\mathcal{F}) = O(W \log W)$
- ✓ $\mathcal{F}$ = nearest neighbor classifiers $\Rightarrow$ $\mathrm{VC}(\mathcal{F}) = \infty$

Note. The VC dimension is in general not related to the number of free parameters of a model (e.g., $f_\alpha(x) = \mathrm{sgn}(\sin(\alpha x))$: 1 parameter, VCdim = ∞).

Fundamental results

For all $f \in \mathcal{F}$, with probability at least $1 - \delta$, we have:

$$R(f) \quad R_{\text{emp}}(f) + \sqrt{\frac{h(\log(2n/h)+1) \quad \log(\quad /4)}{n}}$$

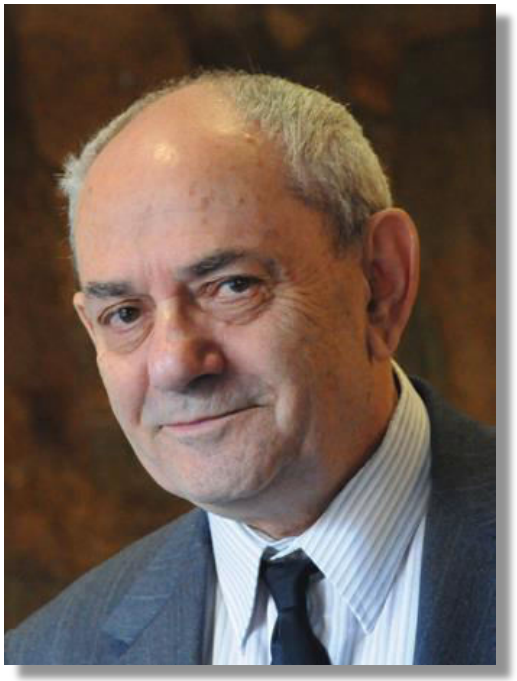
where $h = \text{VC}(\mathcal{F})$, and n is the sample size.

With probability approaching 1, no matter what the unknown probability distribution, given more and more data, the expected error for the functions that ERM endorses at each stage eventually approaches the minimum value of expected error of the functions in $\mathcal{F}$ *if and only if* $\mathcal{F}$ has finite VC dimension.

Popper as a precursor of SLT

«Let me remark how amazing Popper's idea was.
In the 1930's Popper suggested a general concept determining the generalization ability (in a very wide philosophical sense) that in the 1990's turned out to be one of the most crucial concepts for the analysis of consistency of the ERM inductive principles.»

Vladimir Vapnik
The Nature of Statistical Learning Theory (2000)



VC dimension and falsifiability

«The idea of a set of points being shattered by a class of hypotheses may bring to mind Karl Popper's notion of non-falsifiability in the following sense.
If the class of hypotheses is too rich, in the sense of having too great a capacity to discriminate, then whatever the data, a perfectly accurate classifier could be found.»

D. Corfield, B. Schölkopf, and V. Vapnik
Falsificationism and statistical learning theory (2009)

Hence, according to this interpretation:

An hypothesis class $\mathcal{F}$ is falsifiable $\Leftrightarrow$ $\mathrm{VC}(\mathcal{F}) < \infty$

Falsifiability is a matter of degree

«Theories may be more, or less, severely testable; that is to say, more, or less, easily falsifiable. The degree of their testability is of significance for the selection of theories.»

Karl Popper, *The Logic of Scientific Discovery* (1959)

Popper suggested two grounds for comparing degrees of testability.

The subclass relation. Take for instance the following two theories:

T_1 = "all planets move in circles"

T_2 = "all planets move in ellipses"

T_1 is a subclass of T_2 and hence is more easily falsified than T_2.

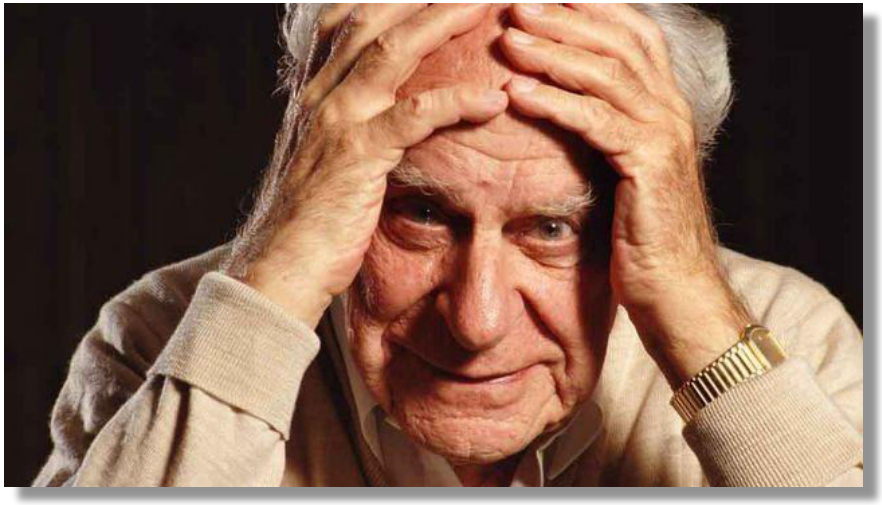
The concept of dimension. "The vague intuitive idea that a cube in some way contains more points than, say, a straight line can be clearly formulated in logically unexceptionable terms by the [...] concept of dimension."

The Popper dimension

«If there exists, for a theory *t*, a field of [...] statements such that, for some number *d*, the theory cannot be falsified by any *d*-tuple of the field, although it can be falsified by certain (*d*+1)-tuples, then we call *d* the **characteristic number** of the theory with respect to that field.»

Karl Popper

The Logic of Scientific Discovery (1959)



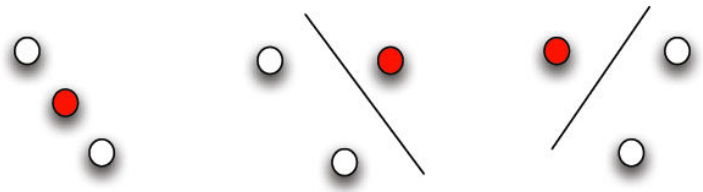
Popper *vs.* Vapnik

The Popper dimension does not quite correspond to the VC-dimension. Indeed, for a function (hypothesis) class $\mathcal{F}$:

- ✓ the VC-dimension is the largest number N such that **some** set of N points is shattered by functions in $\mathcal{F}$
- ✓ the Popper dimension is the largest number N such that **every** set of N points is shattered by functions in $\mathcal{F}$

Example. If $\mathcal{F}$ is the class of linear decision functions in the plane, we have:

- ✓ $\mathrm{VC}(\mathcal{F}) = 3$
- ✓ $\mathrm{Popper}(\mathcal{F}) = 2$



«This suggests that Popper's theory of falsifiability would be improved by adopting VC-dimension as the relevant measure in place of his own measure.»

G. Harman and S. Kulkarni, *Reliable Reasoning* (2007)

Readings

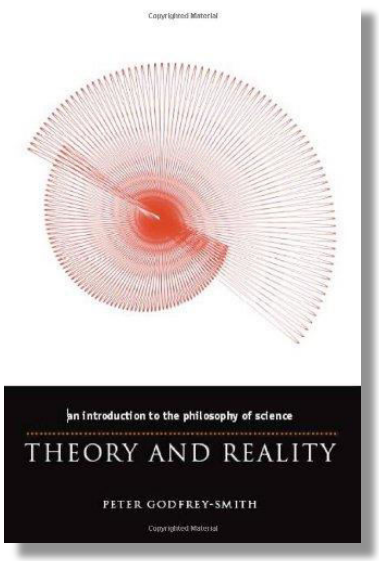
G. Harman and S. Kulkarni. *Statistical learning theory as a framework for the philosophy of induction* (2008).

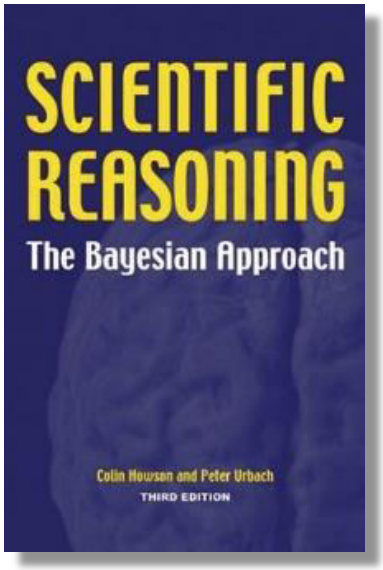
U. von Luxburg and B. Schölkopf. *Statistical learning theory: Models, concepts and results* (2008).

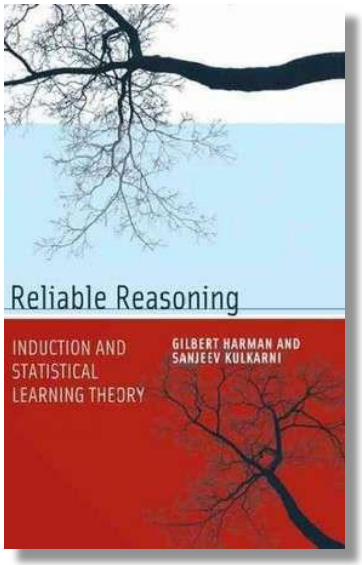
S. Kulkarni and G. Harman. *Statistical learning theory: A tutorial* (2011).

M. Hutter. *On universal prediction and Bayesian confirmation* (2007).

S. Rathmanner and M. Hutter. *A philosophical treatise of universal induction* (2011).







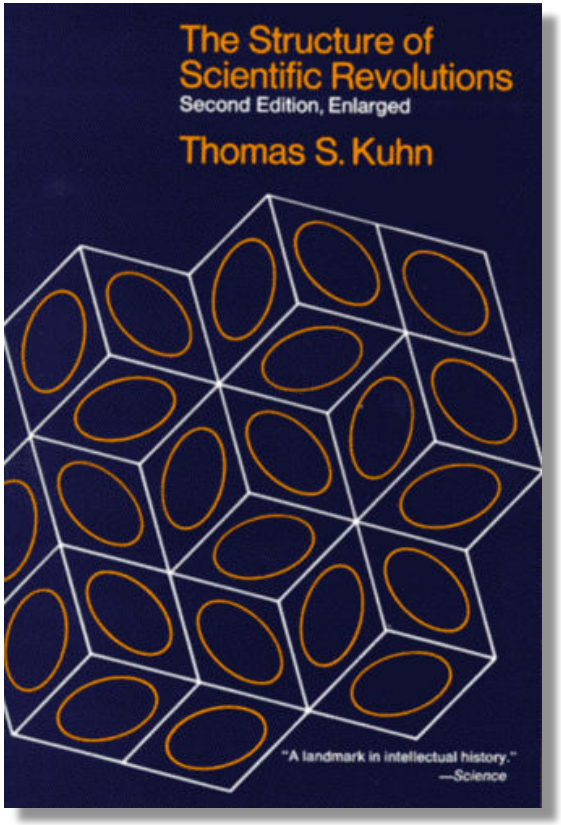
Scientific progress and "revolutions"

Scientific progress?

«The acquisition and systematization of positive knowledge are the only human activities which are truly cumulative and progressive
[…]
progress has no definite and unquestionable meaning in other fields than the field of science»

George Sarton
The Study of the History of Science (1936)

Kuhn's *Structure*



«[*The Structure of Scientific Revolutions* is] the most famous book about science written during the twentieth century.
[…]
Kuhn's book was first published in 1962, and its impact was enormous. Just about everything written about science by philosophers, historians, and sociologists since then has been influenced by it.»

Peter Godfrey-Smith
Theory and Reality (2003)

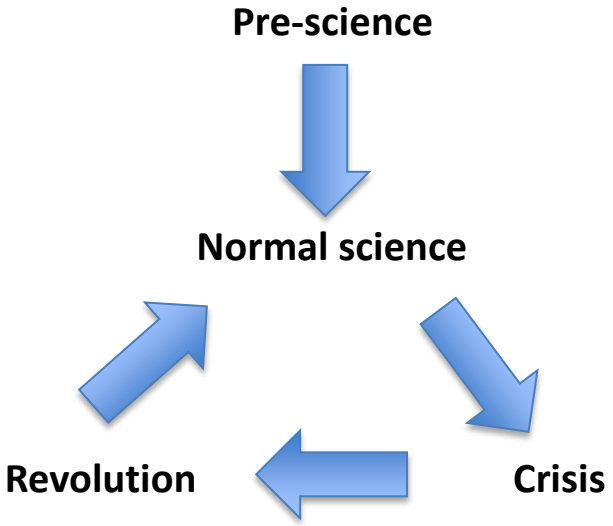
«Great books are rare. This is one. Read it and you will see.»

Ian Hacking (2012)

Kuhn's view of science

«That is the structure of scientific revolutions: normal science with a paradigm and a dedication to solving puzzles; followed by serious anomalies, which lead to a crisis; and finally resolution of the crisis by a new paradigm.»

Ian Hacking
Introduction to the 50th anniversary edition of Structure (2012)



Paradigm shifts

Dick Fosbury (Mexico, 1968)

Paradigms and maturity

«[Paradigms] I take to be universally recognized scientific achievements that for a time provide model problems and solutions to a community of practitioners.
[...]

Acquisition of a paradigm and of the more esoteric type of research it permits is a sign of **maturity** in the development of any given scientific field.»

Thomas Kuhn
The Structure of Scientific Revolutions (1962)

Paradigms:
Broad and narrow

«In much of the book the term ‘paradigm’ is used in two different senses.

On the one hand, it stands for the entire constellation of beliefs, values, techniques, and so on shared by the members of a given community.

On the other, it denotes one sort of element in that constellation, the concrete puzzle-solutions which, employed as models or examples, can replace explicit rules as a basis for the solution of the remaining puzzles of normal science.»

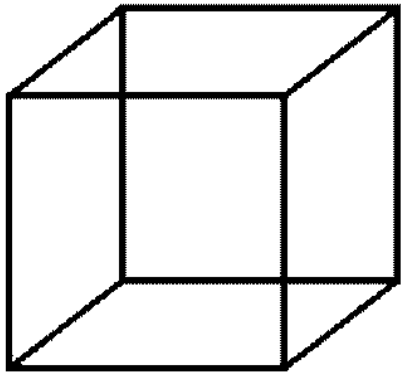
Thomas Kuhn
The Structure of Scientific Revolutions (1969)

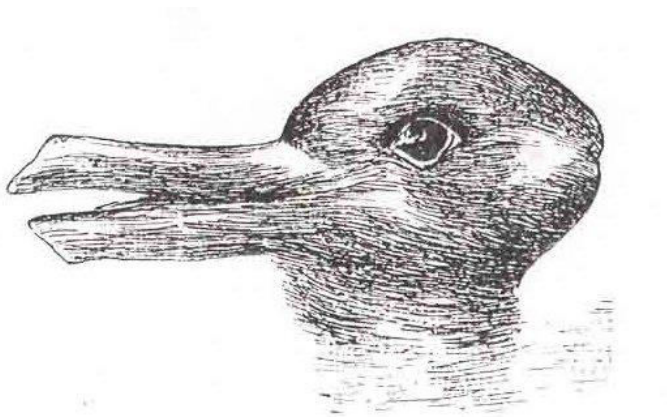
Tycho and Kepler on the hill

«Let us consider Johannes Kepler: imagine him on a hill watching the dawn. With him is Tycho Brahe. Kepler regarded the sun as fixed: it was the earth that moved. But Tycho followed Ptolemy and Aristotle in this much at least: the earth was fixed and all other celestial bodies moved around it.
Do Kepler and Tycho see the same thing in the east at dawn?»

Norwood R. Hanson
Patterns of Discovery (1958)

Paradigm shifts as Gestalt switches

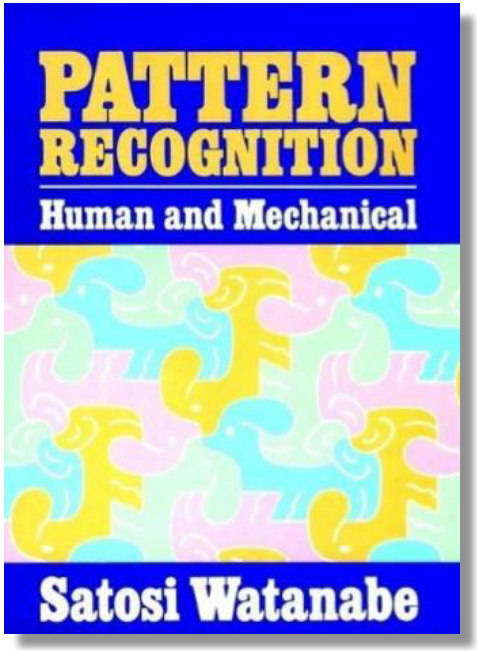




Our essentialist assumption

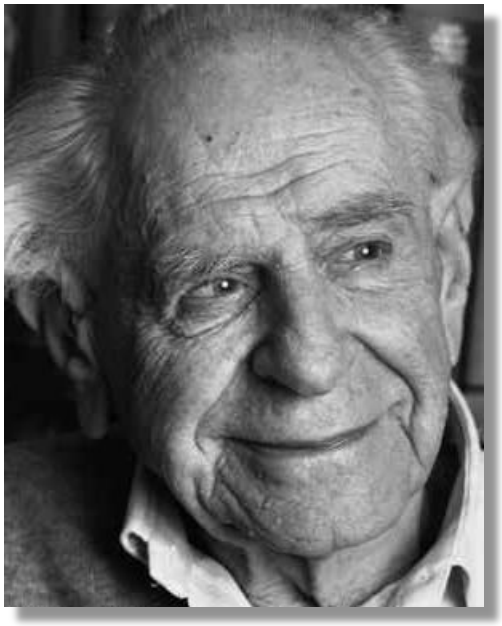
«Whether we like it or not, under all works of pattern recognition lies tacitly the Aristotelian view that the world consists of a discrete number of self-identical objects provided with, other than fleeting accidental properties, a number of fixed or very slowly changing attributes. Some of these attributes, which may be called "features," determine the class to which the object belongs.»

Satosi Watanabe
Pattern Recognition: Human and Mechanical (1985)



Essentialism and its discontents

«The development of thought since Aristotle could be summed up by saying that every discipline, as long as it used the Aristotelian method of definition, has remained arrested in a state of empty verbiage and barren scholasticism, and that **the degree to which the various sciences have been able to make any progress depended on the degree to which they have been able to get rid of this essentialist method.**»



Karl Popper
The Open Society and Its Enemies (1945)

Essentialism under attack

During the XIX and the XX centuries, the *essentialist* position was subject to a massive assault from several quarters and it became increasingly regarded as an impediment to scientific progress.

Strikingly enough, this conclusion was arrived at independently in various different disciplines:

- ✓ Physics
- ✓ Biology
- ✓ Psychology
- ✓ Mathematics

not to mention Philosophy …

Definitions in physics

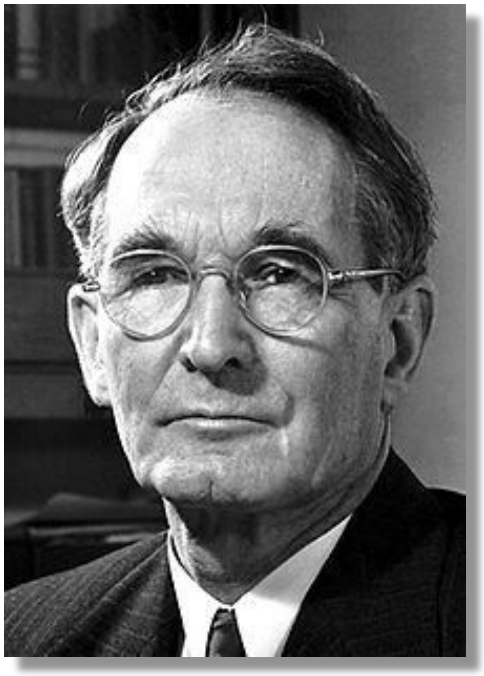
«**What do we mean by the length of an object?**

[…]

To find the length of an object, we have to perform certain physical operations. The concept of length is therefore fixed when the operations by which length is measured are fixed

[…]

In general, we mean by any concept nothing more than a set of operations; **the concept is synonymous with the corresponding set of operations.**»

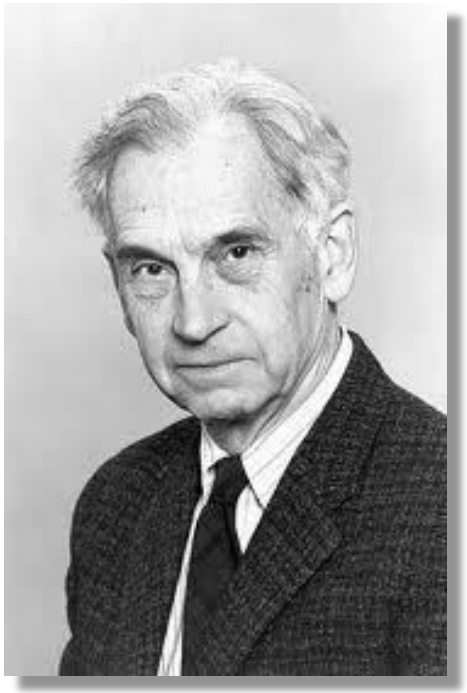


Percy W. Bridgman
The Logic of Modern Physics (1927)

Can we be essentialist after Darwin?

«Essentialism [...] dominated the thinking of the western world to a degree that is still not yet fully appreciated by the historians of ideas. [...]

It took more than two thousand years for biology, under the influence of Darwin, to escape the paralyzing grip of essentialism.»



Ernst Mayr
The Growth of Biological Thought (1982)

Against "classical" categories

«Categorization is a central issue. The traditional view is tied to the classical theory that categories are defined in terms of common properties of their members.

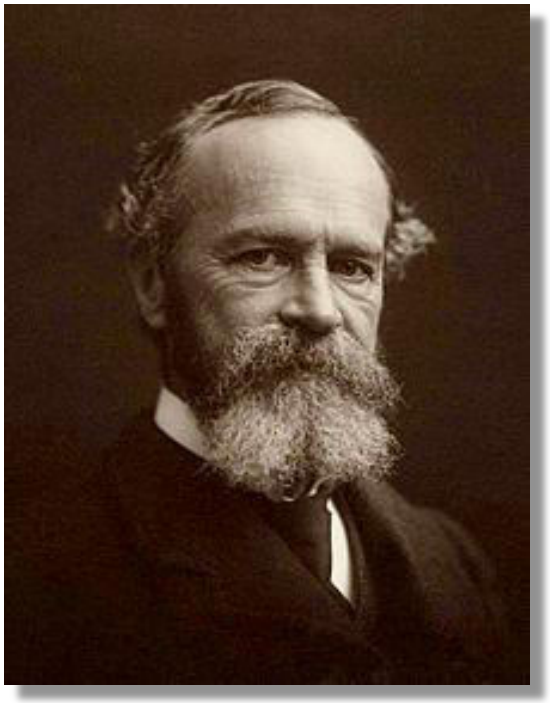
But a wealth of new data on categorization appears to contradict the traditional view of categories. In its place there is a new view of categories, what Eleanor Rosch has termed *the theory of prototypes and basic-level categories*.»

George Lakoff
Women, Fire, and Dangerous Things (1987)

“Signal” *vs.* “noise”

«*There is no property* ABSOLUTELY *essential to any one thing.* The same property which figures as the essence of a thing on one occasion becomes a very inessential feature upon another.»

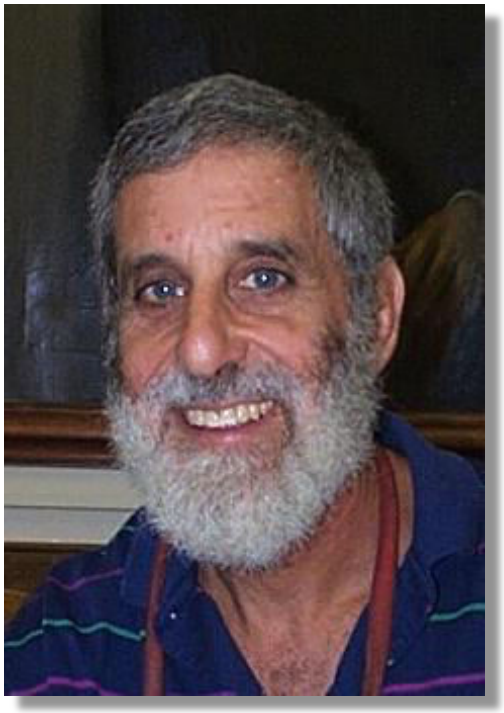
William James
The Principles of Psychology (1890)



What is the subject-matter of math?

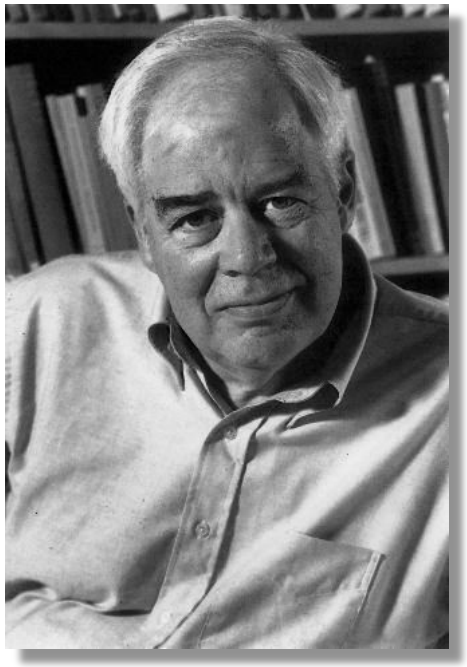
«In mathematics the primary subject-matter is not the individual mathematical objects but rather the structures in which they are arranged.»

Michael D. Resnik
Mathematics as a Science of Patterns (1997)



Epistemic anti-essentialism

«We antiessentialists would like to convince you that it [...] does not pay to be essentialist about tables, stars, electrons, human beings, academic disciplines, social institutions, or anything else. We suggest that you think of all such objects as resembling numbers in the following respect: **there is nothing to be known about them except an initially large, and forever expandable, web of relations to other objects**.



There are, so to speak, relations all the way down, all the way up, and all the way out in every direction: you never reach something which is not just one more nexus of relations.»

Richard Rorty
A World Without Substances or Essences (1994)

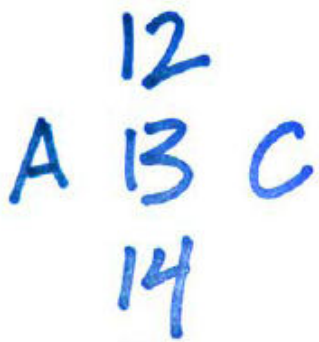
Two consequences of the essentialist assumption in ML

Our essentialist attitude has had two major consequences which greatly contributed to shape the ML/PR fields in the past few decades.

- ✓ it has led the community to focus mainly on **feature-vector representations**, where, each object is described in terms of a vector of numerical attributes and is therefore mapped to a point in a Euclidean (geometric) vector space

- ✓ it has led researchers to maintain a **reductionist position**, whereby objects are seen in isolation and which therefore tends to overlook the role of contextual, or relational, information

Context helps ...



c → cat
→ circus

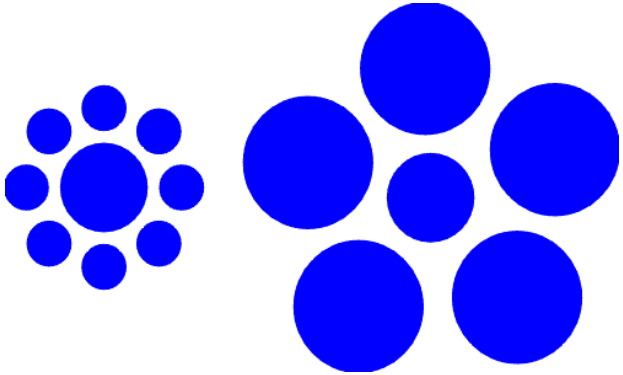
i → sin
→ fine

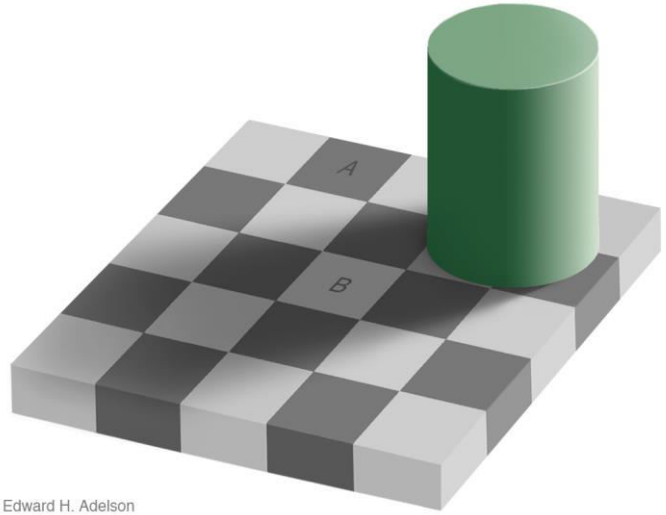
e → red
→ read

festival

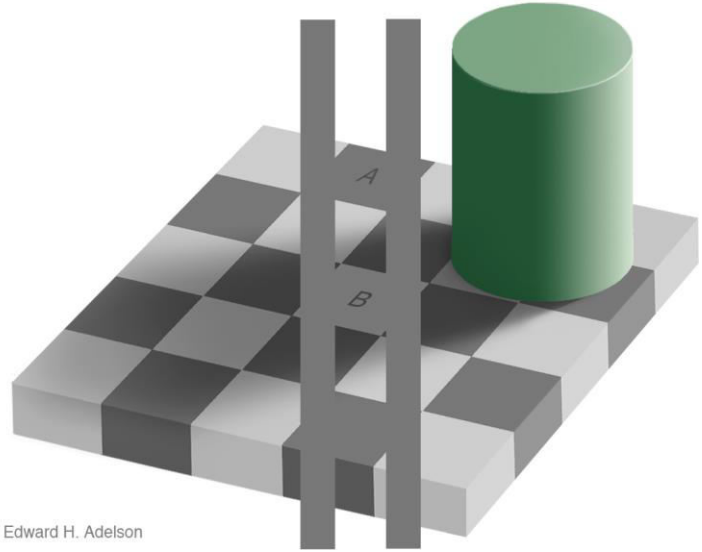
graphics

… but can also deceive



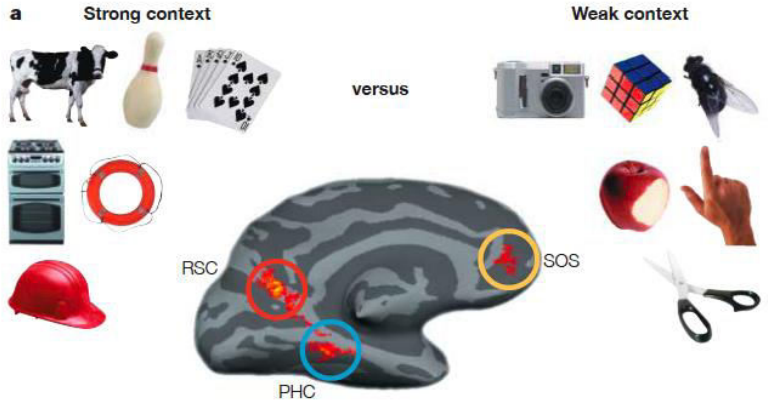


Edward H. Adelson



Edward H. Adelson

Context and the brain



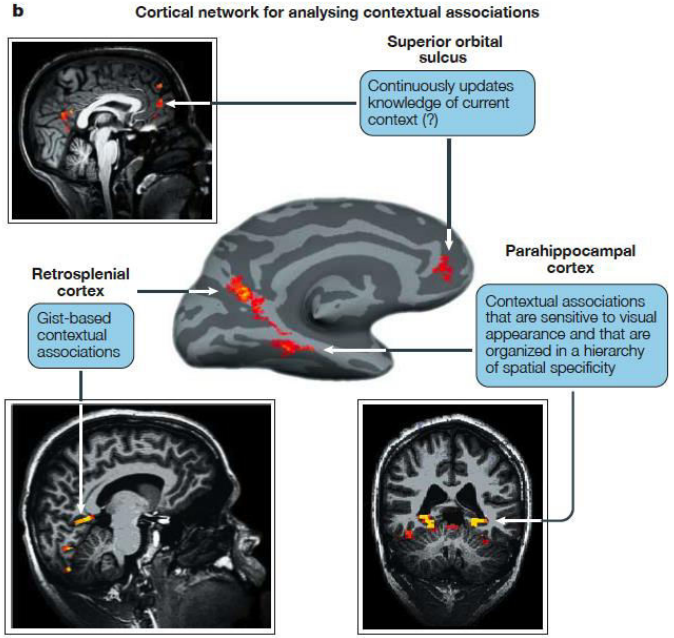


Figure 3 | **Cortical areas involved in processing context. a** | A functional magnetic resonance imaging (fMRI) statistical activation map representing the difference between perceiving objects that are strongly associated with a specific context and perceiving objects that are not associated with a unique context. This is a medial view of the left hemisphere, shown using a precise computer reconstruction where the sulci have been exposed by 'inflation'. The parahippocampal cortex (PHC) is circled in blue; the retrosplenial cortex (RSC) is circled in red; the superior orbital sulcus (SOS) is circled in yellow. Note that in all experimental conditions, subjects viewed similar looking colour photographs of meaningful, everyday common objects that were equally recognizable. Consequently, activation due to low-level processes was presumably subtracted out, and the differential activation map shown here represents only processes that are related to the level of contextual association. **b** | The cortical network for contextual associations among visual objects, suggested on the basis of existing evidence. Other types of context might involve additional regions (for example, hippocampus for navigation[125] and Broca's area for language-related context). Modified, with permission, from REF. 12 © (2003) Elsevier Science.

From: M. Bar, "Visual objects in context", Nature Reviews Neuroscience, August 2004.

The importance of similarities

«Surely there is nothing more basic to thought and language than our sense of similarity.

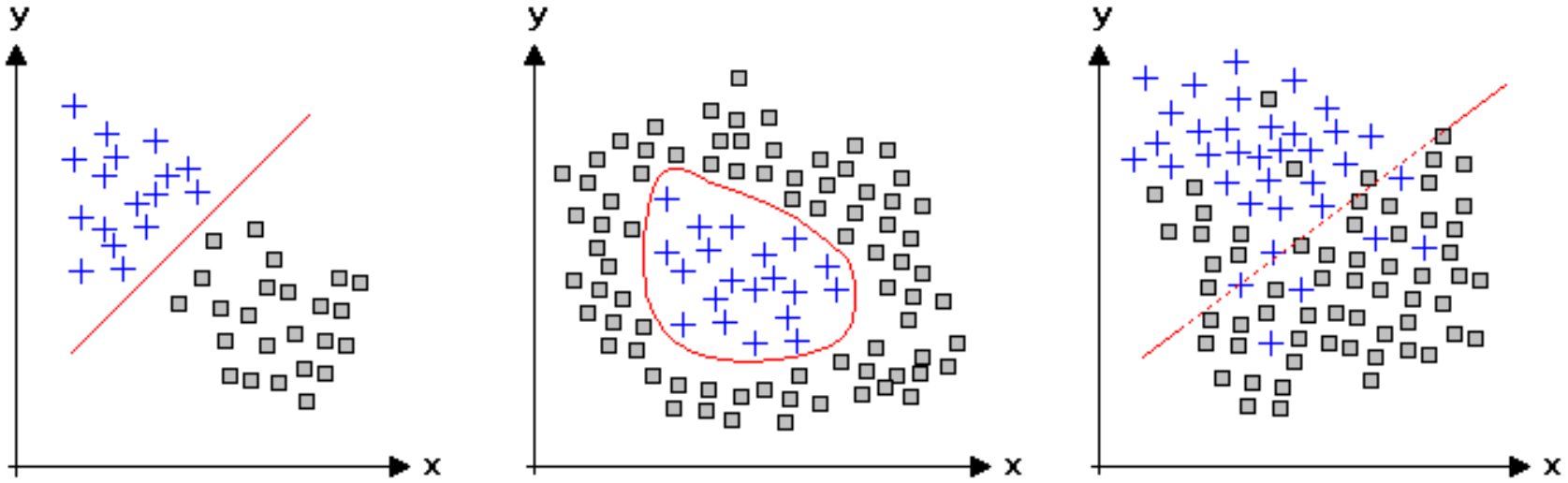
[…]

And every reasonable expectation depends on resemblance of circumstances, together with our tendency to expect similar causes to have similar effects.»

Willard V. O. Quine
Natural Kinds (1969)

Today's view: Similarity as a by-product

Traditional machine learning and pattern recognition techniques are centered around the notion of **feature-vector,** and derive object similarities from vector representations.



Limitations of feature-vector representations

There are situations where either it is not possible to find satisfactory feature vectors or they are inefficient for learning purposes.

This is typically the case, e.g.,

- ✓ when data are high dimensional (e.g., images)
- ✓ when features consist of both numerical and categorical variables
- ✓ in the presence of missing or inhomogeneous data
- ✓ when objects are described in terms of structural properties, such as parts and relations between parts, as is the case in shape recognition
- ✓ in the presence of purely relational data (graphs, hypergraphs, etc.)
- ✓ …

Application domains: Computational biology, adversarial contexts, social signal processing, medical image analysis, social network analysis, document analysis, network medicine, etc.

Signs of a transition?

The field is showing an increasing propensity towards anti-essentialist/relational approaches, e.g.,

- ✓ Kernel methods
- ✓ Pairwise clustering (e.g., spectral methods, game-theoretic methods)
- ✓ Graph transduction
- ✓ Dissimilarity representations (Duin et al.)
- ✓ Theory of similarity functions (Blum, Balcan, …)
- ✓ Relational / collective classification
- ✓ Graph mining
- ✓ Adversarial learning
- ✓ Contextual object recognition
- ✓ …

See also “link analysis” and the parallel development of “network science” …

Taking stock

In summary: is machine learning a “mature” science (according to Kuhn)?

The answer depends on the scope of the notion of a paradigm …

Narrow? ⇒ a whole series of paradigms: neural networks, SVM’s, kernel methods, random forests, deep learning, …

Broad? ⇒ the field has been dominated by an **essentialist outlook**

But … there are signs of a transition from an essentialist to an anti-essentialist/relational paradigm.

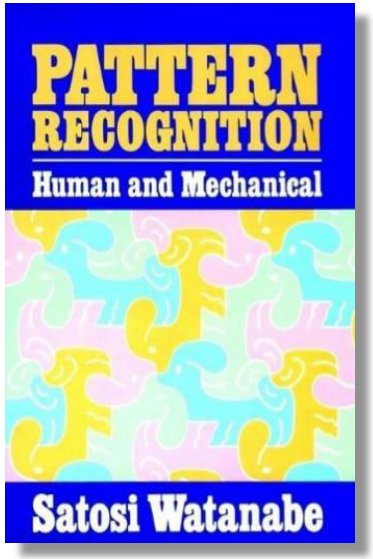
Other fields have made substantial progress by abandoning a purely essentialist position.

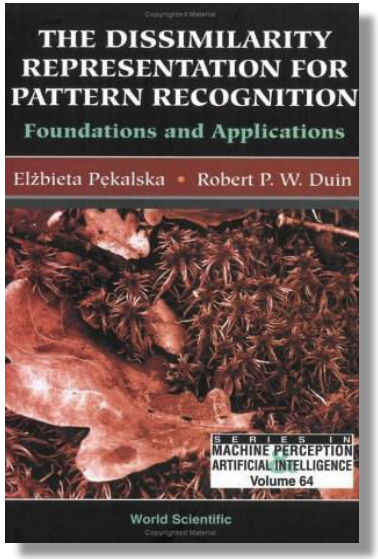
Readings

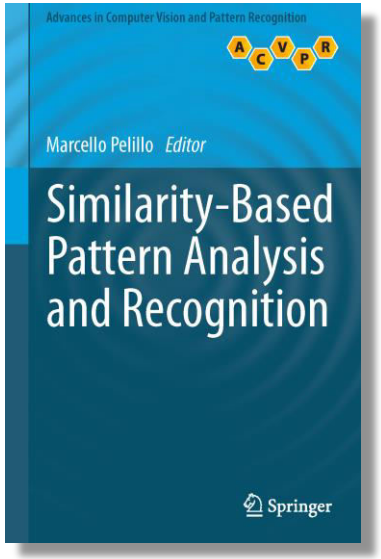
M. Pelillo and T. Scantamburlo. How mature is the field of machine learning? In: *Proc. AI*IA* (2013).

N. Cristianini. On the current paradigm in artificial intelligence. *AI Communication* (2014).

R. P. W. Duin and E. Pekalska. The science of pattern recognition. Achievements and perspectives. *Studies in Computational Intelligence* (2007).







Epistemology and machine learning

Other epistemological topics of interest to the machine learning community (not treated today):

- ✓ Causality (Pearl, Spirtes, Glymour, Schölkopf, …)
- ✓ Complexity and information (Kolmogorov, Solomonoff, Hutter, …)
- ✓ Model selection
- ✓ Emergentism
- ✓ Scientific method
- ✓ Abstraction and categorization
- ✓ Decision theory
- ✓ Philosophy of technology

and many more …

Philosophy and machine learning

http://www.dsi.unive.it/PhiMaLe2011/

Philosophy and Machine Learning - Workshop @ NIPS 2011

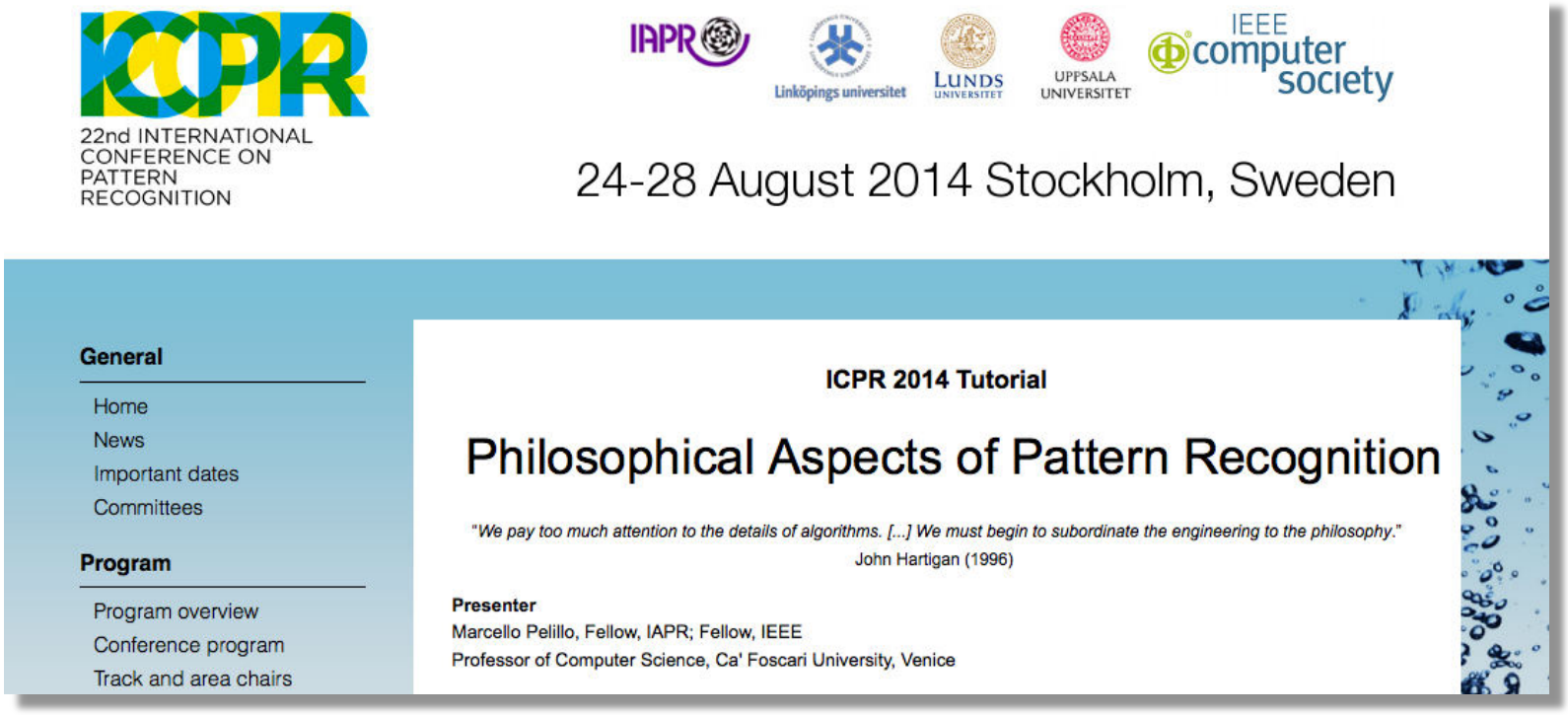
Sierra Nevada, Spain - 17 December 2011

Special issue on
"Philosophical aspects of pattern recognition"

Vol. 64, October 2015

Guest editor: M. Pelillo

Philosophy and machine learning



http://www.icpr2014.org/tutorialpages/philosophicalaspects