Unification and Evidence

FEW , May 2005

Malcolm R. Forster

Department of Philosophy

University of Wisconsin-Madison

http://philosophy.wisc.edu/forster

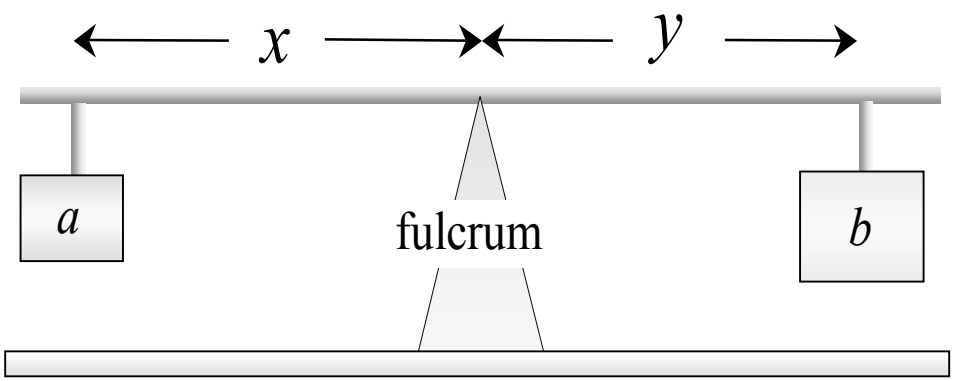
Introduction

There is a huge audience interested in theory and evidence.

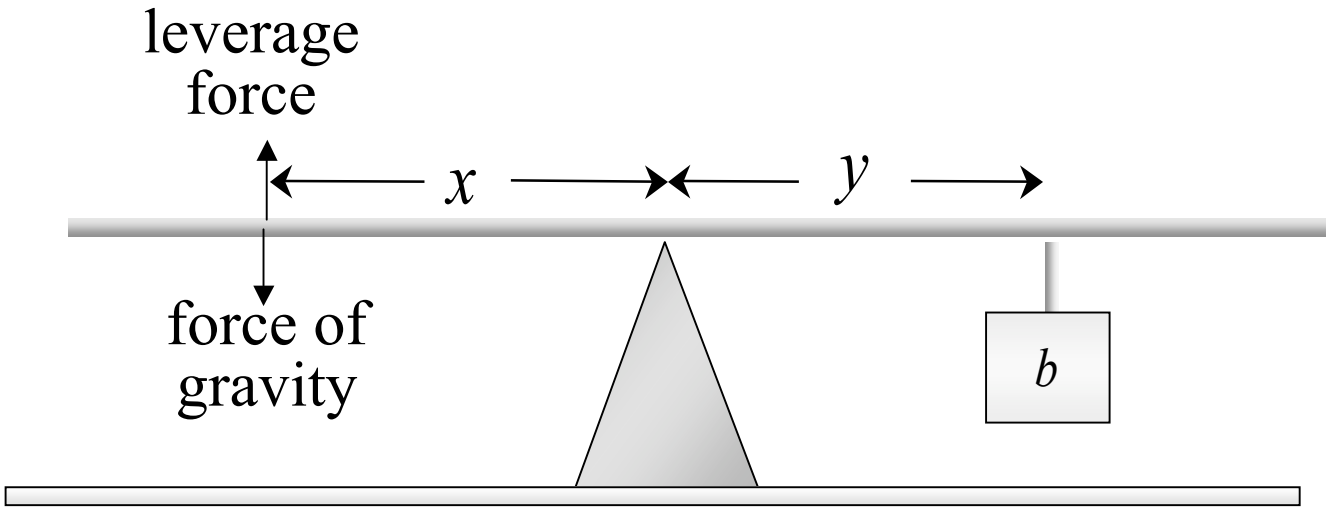
Besides historians of science and philosophers, there are scientists, statisticians, and college students. And it's relevant to machine learning, AI, and automated reasoning.

To communicate with this audience one needs good illustrative examples. We need examples that are simple and expandable, but not too simple!

Basic Beam Balance Model



Possible Data	
x	y
1 cm	1 cm
2 cm	2 cm
3 cm	3 cm



leverage force = force of gravity

$$m(b)g\frac{y}{x} = m(a)g$$

$$y = \frac{m(a)}{m(b)}x$$

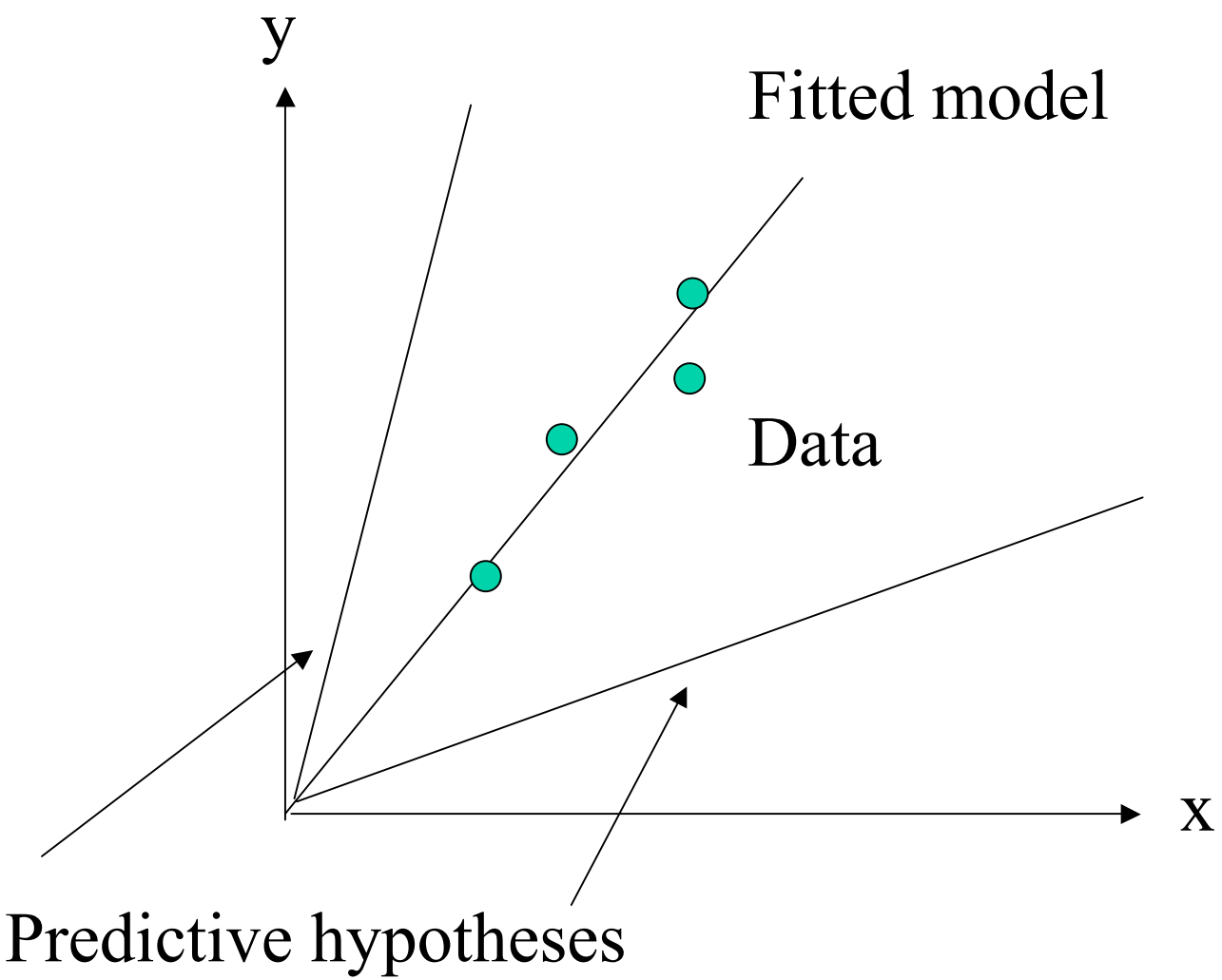
Function Fitting and Prediction

A *predictive hypothesis* is a function y = f(x). It can be represented by a curve in the x-y plane.

Predictive hypotheses make point predictions of y given x.

A *model* is an equation, or set of equations, with at least one adjustable parameter. A model is represented by a *family* of x-y curves.

Curve Fitting



Parameter Estimation = Theoretical measurement

A *fitted model* is obtained from a model by estimating the value of theoretical parameters from a set of seen data.

A fitted model is a predictive hypothesis. A model does not make *point* predictions.

The method of least squares, or the method of maximum likelihood, are common statistical methods of parameter estimation.

The Problem of Many Models

Even when models are derived from an entrenched background theory, there are many rival models.

The problem of comparing or evaluating models using seen data is called the problem of *model selection*.

SIMPLE $$y = \frac{m(a)}{m(b)}x$$

COMPLEX $$y = \delta + \frac{m(a)}{m(b)}x$$

Naïve Empiricism Does not Work!

Naïve empiricism says: Favor the model that best fits the seen data.

Problem: For nested models, the simpler model is *never* favored.

The argument is that the best fitting curve in SIMPLE is also in COMLEX, therefore COMPLEX can always do as well as SIMPLE, and almost always better.

It doesn't matter how goodness-of fit is defined.

Sophisticated Empiricism Does Work!

Select the model with the highest CV score

1. Select one data point as the test datum. Fit the model to the remaining $N - 1$ data pointsnd record how well the test datum fits the curve.
2. Repeat N times, for each of N data points.
3. By averaging the N scores, we extract full information from the data.

What Went Wrong with Goodness of-Fit?

By using *all* the data to "construct" the fitted model, and then the *same data* as the test data, one introduces a BIAS towards models that are good at ACCOMMODATING the data.

We want to select models that are good at PREDICTION. The CV score measures a model's ability to predict "data of the same kind".

Maximizing goodness-of-fit is works for *parameter estimation*, but NOT for model selection.

Model Selection: the rest of statistics

Neyman-Pearson hypothesis testing.

Akaike's criterion (A1C).

Bayes factors (BIC).

Countless others.

BUT they all have similar properties: They tend to select models that are good at predicting data of the same kind! What about predicting data of a different kind? Is that even possible?

Instrumentalist and Realist Goals

Instrumentalist Goals: Prediction and control of Nature. (Predictive accuracy, not empirical adequacy.)

Realist Goals: The representation of reality behind the observed phenomena. (Approximate truth.)

What I call *monolithic* philosophies of science assume that one truth-related goal subsumes all others. (I favor pluralistic approaches)

An version of the beam balance example will show that pluralism is right.

The Distinction between Methods and Goals

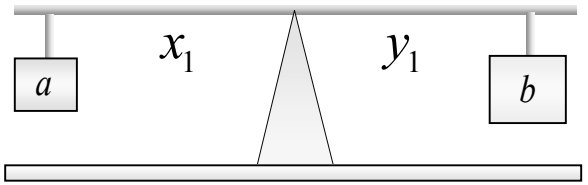
Methods or criteria of model selection are *means* for achieving goals.

Different truth-related goals sometimes optimized by different methods of model selection.

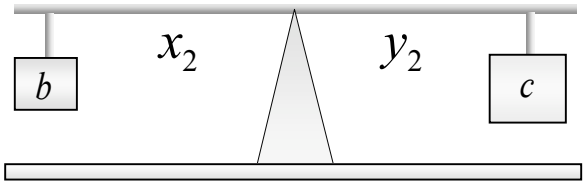
Which truth-related goal are optimized by which methods, modes of inference, or criteria?

Standard model selection methods primarily optimize a model's predictive accuracy (of data of the same kind).

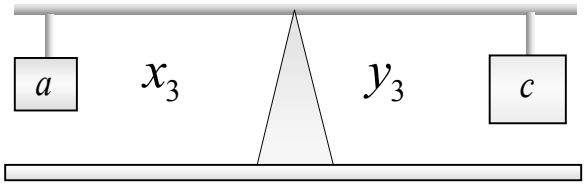
UNIFIED Beam Balance Model



$$y_1 = \frac{m(a)}{m(b)} x_1$$



$$y_2 = \frac{m(b)}{m(c)} x_2$$

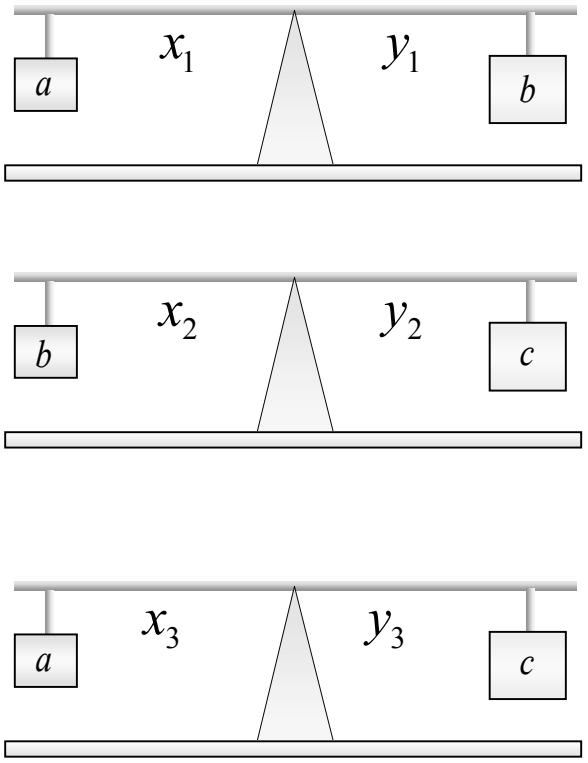


$$y_3 = \frac{m(a)}{m(c)} x_3$$

Constraint:

$$\frac{m(a)}{m(c)} = \frac{m(a)}{m(b)} \times \frac{m(b)}{m(c)}$$

Disunified Beam Balance Model (DUM)



$$y_1 = \alpha\, x_1$$

$$y_2 = \beta\, x_2$$

$$y_3 = \gamma\, x_3$$

NO CONSTRAINT: $\gamma = \alpha\,\beta$ or $\gamma \neq \alpha\,\beta$

Prediction versus Accommodation Again

It's not true that UNIFIED predicts the constraint, while DUM does not.

UNIFIED predicts $\hat{\gamma} \approx \hat{\alpha}\hat{\beta}$

DUM merely accommodates $\hat{\gamma} \approx \hat{\alpha}\hat{\beta}$

SIMPLE predicts $\hat{\delta} \approx 0$

COMPLEX merely accommodates $\hat{\delta} \approx 0$

Is Unification Special?

Given that standard model selection methods, such as CV, succeed in comparing SIMPLE and COMPLEX, will they work for UNIFIED versus DUM?

There is a very important difference between SIMPLE and UNIFIED.

δ is measured only once.

While mass parameters can be independently measured many times.

An Intuition

Suppose that 21 objects are hung pair-wise on a beam balance. There are 210 experiments, so that DUM has 210 adjustable parameters.

Choose one object as the unit mass, so that UNIFIED has 20 adjustable parameters. Each mass has roughly 10 independent measurements.

If the measurements agree for 20 different parameters, then this is strong statistical evidence in favor of UNIFIED.

The Example

Suppose that half of the massless weight hangers are replaced by lead hangers at random. Then some of the estimated mass values estimated using UNIFIED will be too high, and some too low, but these errors will tend to cancel out.

There is good evidence in favor of UNIFIED over DUM.

Yet DUM will make the best *predictions* if the lead hangers remain attached to the same objects.

UNIFIED will make more accurate predictions in different experiments, while also fulfilling realist goals.

Error correction

Moreover, UNIFIED can be used to infer where the lead hangers are, thereby arriving at a corrected model with only 20+1 adjustable parameters.

Clearly no model can "correct itself" without a fairground theory. Which shows that theory, unification, and evidence are linked together.

This understanding of unification and evidence has many applications in philosophy of science.

Historical Examples

Copernicus did not achieve better "next instance" prediction than Ptolemaic astronomy. It did provide a unified model of planetary motions.

Kepler's model preserved the unification, and improved predictive accuracy.

From a sun-centered viewpoint, Copernican sub-models were disunfied.

Newton argued that Kepler's harmonic law provided independent measurements of the sun's gravitational mass.

The Composition of Causes

Suppose that a and b are hung together on the left side of the beam, and a unit mass u is hung on the right. The model is:

$$y = \big(m(a) + m(b)\big)x$$

It is impossible to estimate the value of each mass separately from this single experiment.

But, if we hang *a* with *c*, and *b* with *c*, then there are 3 equations with 3 unknowns, and each mass is measured.

WIDENING the variety of evidence can remove non-identifiably.

Variety of Evidence

Hempel (1966) considered a 3- experiment version of Snell's law, which is structurally equivalent to the 3-experiment beam balance example.

Compare 2 possible data sets:

A: 300 data points in the {a,b} experiment and none in the others.

B: 100 data points in each experiment.

Why is the more varied evidence better?

Unification and Evidence

Unified models can predict the agreement of independent measurements of theoretical parameters, such as masses.

Disunfied models like DUM merely accommodate these relational facts.

Prediction has evidential value, while accommodation does not.

Unification has no evidential value by itself.

Accommodation "saves the phenomena," while *prediction* can provides evidence that disparate phenomena are connected by an underlying reality.