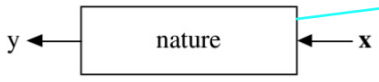
- 2 cultures in the use of statistical modeling:

① One assumes data is generated in stochastic (non-deterministic) manner (data model)

② Uses algorit. models and treats data mechanism as unknown (algo model)

Statistics starts with data. Think of the data as being generated by a black box in which a vector of input variables **x** (independent variables) go in one side, and on the other side the response variables **y** come out. Inside the black box, nature functions to associate the predictor variables with the response variables, so the picture is like this:



black box

There are two goals in analyzing the data:

Prediction. To be able to predict what the responses are going to be to future input variables;
Information. To extract some information about how nature is associating the response variables to the input variables.

There are two different approaches toward these goals:

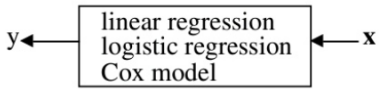
- 2 goals of data analysis
  - prediction (for future unseen data)
  - information (extract info about how nature associates outputs to inputs)
- 2 approaches towards those goals:
  ① Data modeling culture:
  (assumes stochastic data model for

inside of black box

The analysis in this culture starts with assuming a stochastic data model for the inside of the black box. For example, a common data model is that data are generated by independent draws from

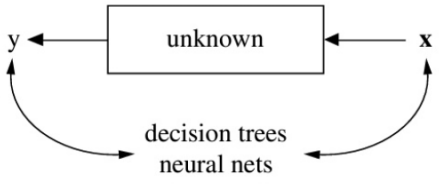
response variables = f(predictor variables, random noise, parameters)

value of params estimated from data



Model validation. Yes–no using goodness-of-fit tests and residual examination.
Estimated culture population. 98% of all statisticians.

② Algorithmic modeling
↳ considers inside of box complex and unknown



Model validation. Measured by predictive accuracy.
Estimated culture population. 2% of statisticians, many in other fields.