

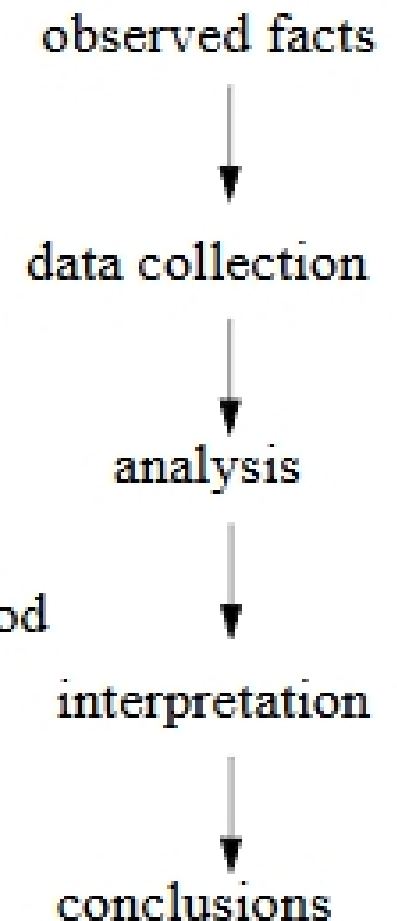
# Data Analysis

- Data in visualization:
  - digital data generated from mathematical models or computations
  - digital data generated from human or machine collection
- Purpose of data analysis:
  - all data collected are linked to a specific relationship or theory
  - relationships are detected as patterns in the data
  - note: the relationship may either be functional (good) or coincidental (bad)
  - note: data analysis and interpretation are functionally subjective
- Logical analysis:
  - applying logic to observations (the data) creates conclusions (Aristotle)
  - conclusions lead to knowledge (at this point the data become information)
- There are two fundamental approaches to generate conclusions:
  - induction
  - deduction

# Induction vs. Deduction

- Induction: make observations first, then draw conclusions

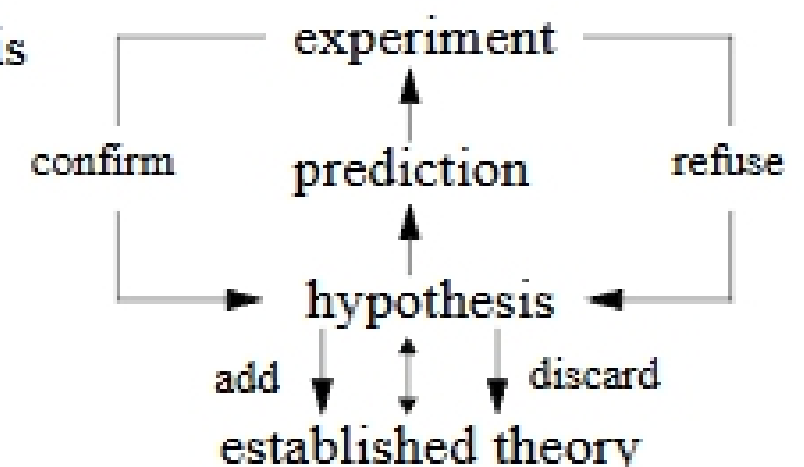
- organized data survey (structured analysis, visualization) of the raw data provide the basis for the interpretation process
- the interpretation process will produce the knowledge that is being sought
- experience of the individual scientist (the observer) is crucial
- important: selection of relevant data, collection method, and analysis method
- *data mining* is an important knowledge discovery strategy here
- ubiquitous data collection, filtering, classification, and focusing is crucial



- Deduction: formulate a hypothesis first, then test the hypothesis

via experiment and accept/reject

- data collection more targeted than in induction
- only limited data mining opportunities



# The Data

- Data origin:
  - real world data - measured from real-world objects and processes (sensors, statistics, surveys)
  - model data - computed by machines (numerical simulations, scientific computations)
  - design data - edited by humans
- Data size:
  - number of samples and data items (kB, GB, MB, TB)
- Data type:
  - scalar or multi-variate, N-dimensional: number of attributes per data item (attribute vector)
  - scalar or vector (e.g., flow direction)
- Data range and domain:
  - qualitative (non-numerical measurements) vs. quantitative (numerical measurements)
- Data value:
  - categorical (nominal): categoricies are disjunct, no intrinsic rank (e.g., {yellow, red, green})
  - ordinal data: data members of ordered sequence of categories (e.g. {tiny, small, large, huge})