The good traveller is flexible and has a sense of humor.

What is a

Vague concept -
   Make precise in various ways

Datum - undefined concept

Data - \{datum\}
   Some things became data only recently

Data analysis -
   Ancient

Confirmatory data analysis -
   Deciding seems established
   The model is sacred, clear question
   Careful planning

E.g. cloud seeding

Exploratory data analysis -
   What seems to be going on
   The data are sacred, generate questions
   Human interaction basic

idea \rightarrow\ question/design \rightarrow\ collection \rightarrow\ analysis \rightarrow\ answer
Kepler-Newton-Lagrange-Gauss

Relation of EDA and CDA
Need both
Scientific method, cyclic, Popper

idea→question/design→collection→analysis→answer→idea

Data mining –
Large data sets, (perhaps collected for other purposes), retrospective
Search for patterns
Brings diverse fields together, e.g. computing
Often profane, opportunistic

Model – Suppes

References.

J. W. Tukey (1986). “We need both exploratory and confirmatory”.


"Theories of data analysis: from magical thinking through classical statistics" – P. Diaconis

**Magical thinking** – a term from anthropology and psychiatry

- assuming can wish for things and get them
- reading too much into patterns

There are patterns in noise!

**EDA can come close to magical thinking**

Classical mathematical statistics

pick models and hypotheses in advance

Scientific thinking

repetition of experiments – cold fusion

“uncomfortable science” – replication is not feasible – astronomy, economics

**INTUITIVE STATISTICS**

Scatterplots
most subjects judged a small plot more associated than big plot of same points

Anchoring/experimenter bias

Representativeness

Examples
  clinical trials
  legal cases
  ESP

Multiplicity
  preliminary data screening
  many comparisons
  transformation

Remedies
  Publish without p-values
    Success stories
      air pollution
      economics
      medicine
      psychology

Theories for data analysis
  Probability-free, GHA, Finch, Mallows
  Ad hoc inference with non-experimental data

Mathematics can help
Ozone study

22 sites in New Jersey

Highest readings at rural Ancora

Error???

There was some theory suggesting OK Philadelphia was 23 miles away

Scatter plot of ozone vs. direction of wind at Philadelphia

When curves added clearly some association

Later other support for the hypothesis of “transport”
Crucial elements

(i) willingness to collect and study data

(ii) use of diagnostic techniques to show unexpected

(iii) an ability to recognize striking patterns – QQ plot – high at Ancorra

(iv) enough understanding to enable patterns to be recognized as potentially meaningful

(v) avoidance of precipitate commitment to models of clearly inadequate complexity; use of robust summarised and graphical displays

(vi) energetic following-up of clues