Characteristics of Big-Data

- Large – scaling to peta- and exa-scale.
- Noisy – with high rates of false positives and negatives.
- Multiscale – incorporating interactions at vastly different levels of abstractions.
- Dynamic – data changes over time.
- Heterogeneous – demonstrating high variability in characteristics over space and time. Significant skews in degree distribution.
- Distributed – data is typically collected and stored at distributed locations.
- Elastic – data is typically elastic.

Analysis of Large Datasets

Ad-hoc solutions do not work at scale! Analysis techniques for big-data **must**:

- Rely on suitable formulations – results are typically probabilistic. Formulations must quantify (and optimize) significance (statistical).
  - Deterministic formulations on noisy data are not meaningful. This is the norm currently.
  - Overfitting to noisy data is a major problem.
  - Distribution agnostic formulations (say, based on simple counts and frequencies) are not meaningful – once again, most work in data mining and machine learning relies on such over-simplified models (or no models at all!).
- Provide rigorously validated solutions – garbage-in, garbage-out at scale.
- Dynamic and heterogeneous datasets require significant formal basis.