# Introduction to Big Data

Big Data analytics mainly deals with the management of a large amount of data and how to extract meaningful information from them.

In the big data style of analytics, algorithms are not important, the data is, meaning, the domain knowledge of the data itself may not be required to perform analytics on the data.

A spurious correlation between data occurs when two or more events or variables are associated but are not causally related, usually by coincidence or by a third party factor. Another pitfall that occurs in big data analytics is gaps in the data, where it is assumed that negative results are more likely to go missing, introducing a bias to certain types of data.

Error handling is done by using domain knowledge to check the model for validity, and estimation of errors is done by comparing the analysed data with the ground truth theory about the same data.

Why is weather forecasting successful?

1. It is a chaotic, dynamic and a nonlinear system
2. Data is adjusted by humans
3. More sensitivity towards errors in predicting no rain than rain.

The collapse of the housing market in 2008

1. The problem with risk calculations for combined mortgages was done assuming independence
2. This is a huge problem, and the shorthand is to never predict a situation when you are out of sample, or never seen a situation before.

Error estimation is usually done by the Gaussian Central Limit Theorem which says if the variation is due to multiple factors, the factors are considered to be independent, each with the same importance, and the error always follows a Gaussian distribution.

Big Data error estimation is purely empirical, by dividing the data into a training set and a test set, and training on the training set and verifying the model with the test set.

The old style of data had a fixed schema and format, was clean, consistent, predictable and had elaborate procedures for managing and analysing it. The new style of data is characterised by

1. Volume - Exponential growth rate
2. Velocity - Data inflow rate
3. Veracity - Trustworthiness of data, data cleaning or scrubbing required
4. Variety - Types of data are different in different formats

Taking the example of Google's PageRank algorithm

1. First implementation
    a. Web designed as a graph
    b. User starts at a random page and follows a link
    c. The rank of a page is proportional to the number of users on that page and the number of pages pointing to that page
    d. The main assumption of pagerank was that, the more popular the page, the better its quality.
2. Storage and Retrieval
    a. A URL server sends URLs to a crawler
    b. The crawler sends pages to a Store server
    c. A store server compresses these pages and stores it in a repository
    d. An indexer parses the pages, generates index records containing word, URL, position, font size, capitalisation and the link database
    e. The sorter generates the inverted index by sorting the words.
3. Optimizations
    a. URLs and words are converted to integers to reduce space and speedup lookup
    b. Usually stored as a hashtable