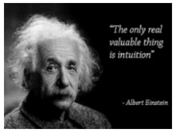
Week 4:

Data Pre-Processing

Overview



Data Pre-Processing: Intuition

- About **80%** to **90%** of the work in analytics is getting the **data ready** for analysis.
- Data is often pre-processed to correct problems with the data
- But pre-processing is sometimes done to improve model performance – **"feature engineering"**
- For example, a popular technique to detect **spam mail** is to look for the proportion of upper case letters in the subject. This requires pre-processing the subject headers.
- Often the pre-processing needs are based on the functional domain expertise of the modeler rather than mathematical reasons

Why Pre-Process Data?

- Typical **problems** with the data include: missing values; inconsistent data; outliers; skewed or bi-modal distributions.
- Some times the data is simply not in the **appropriate format** for the type of analysis to be done.
- Some times the data needed for the analysis needs to be gathered from multiple data sources and joined
- Or the data needs to be **scaled**, **centered** or **normalized**
- Or it needs to be aggregated, summarized, etc.
- Especially when data is not at the **"unit of analysis"** level
- For example, you may have data on prices and features for individual houses, but you need to analyze housing by counties in the US – i.e., the unit of analysis is the **county**, not the individual homes → the data needs to be **aggregated**

Kogod School
of
Business