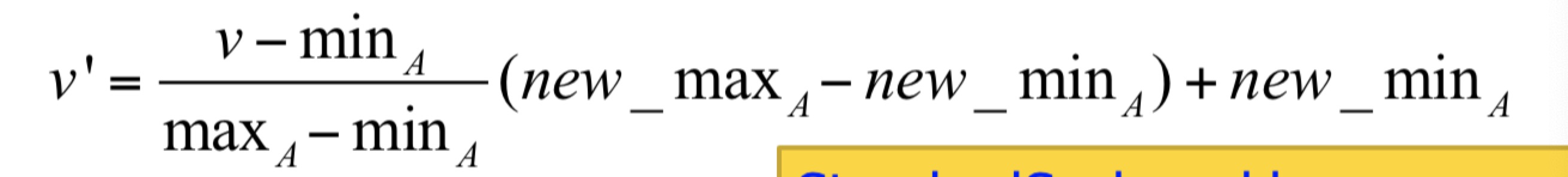
# 9318 week3-4

## Major task in Data Pre-processing

1. Data cleaning
2. Data Integration
3. Data transformation: Normalization
4. Data reduction: PCA and so
5. Data discretization and Data Type Conversion

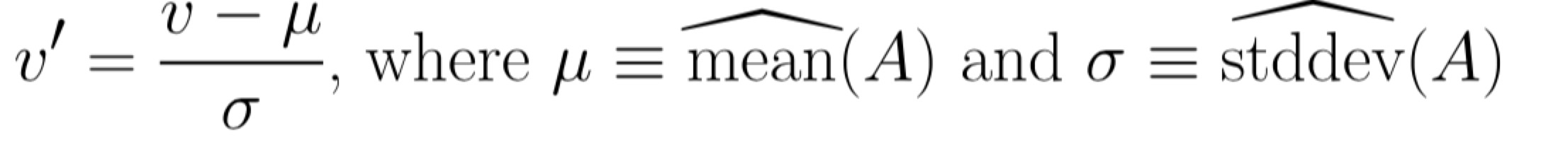
Normalization:

1. MinMax



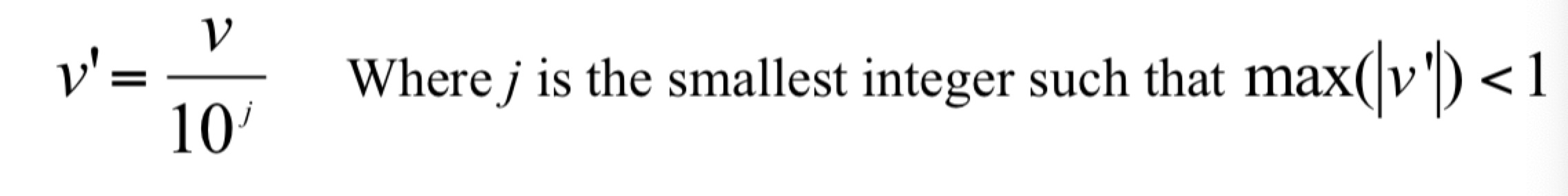
规定new max and new min, normalize data in between these new max and new min.

1. z-score



先算平均和标准差，然后对每一项用上面的公式

1. decimal scaling



找出数据中最大的位数n，然后对每一项/10^n

Some data reduction methods:

1. Dimensionality reduction
2. Data Compression
3. Numerosity reduction – fit data into models
4. Discretization and concept hierarchy generation
5. Dimensionality reduction

High dimensional data is always sparse which make them hard to learn

Goal: reduce dimensionality of the data and still maintain the meaning fulness of the data

Methods:

**Feature selection**: choose a subset of the features.

**Feature extraction**: create new features by combining new ones.

