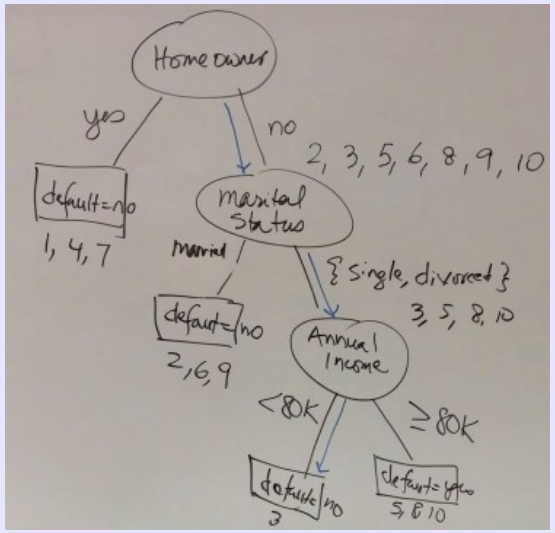


$$\begin{bmatrix} \sigma_1 & & 0 \\ & \sigma_2 & \\ 0 & & \sigma_n \end{bmatrix}$$

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$$X = \sum_{i=1}^{rank(X)} \sigma_i u_i v_i^T = U \Sigma V^T$$

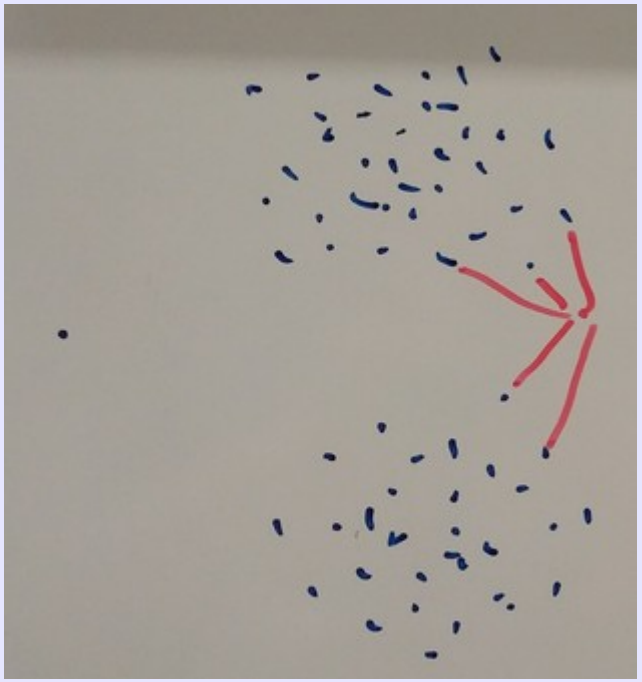
σ_i: i^{th} singular value of X

u_i: i^{th} left singular value of X (i^{th} column of U)

v_i^T: i^{th} right singular vector of X (i^{th} column of V^T)

U: Captures the patterns among attributes

V^T: Captures the patterns among the objects



CS 422-04: Data Mining
Vijay K. Gurbani, Ph.D., Illinois Institute of Technology

Lecture 1: **Introduction**

What is data mining?

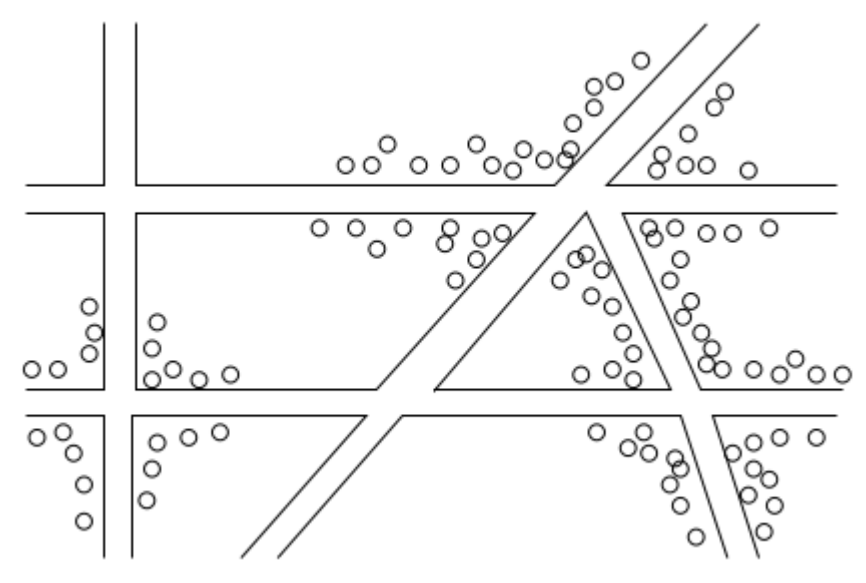
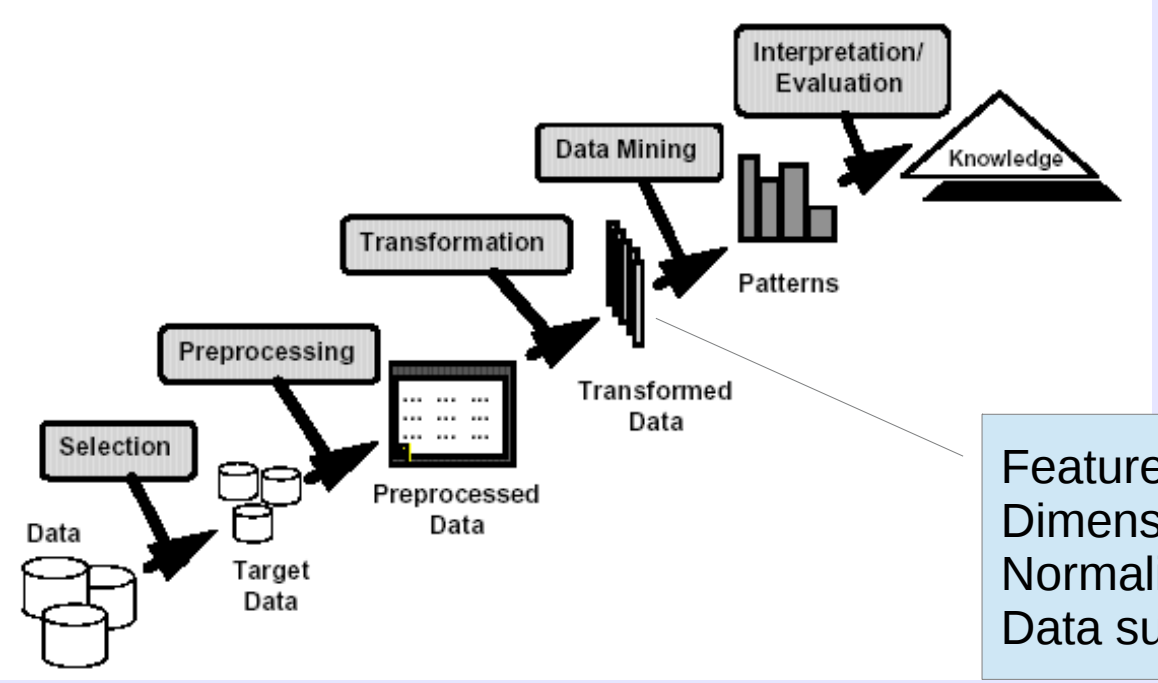


Figure source: Mining of Massive Datasets, Leskovec et al., 2014.

In 1854, Dr. John Snow noticed a pattern with respect to cholera outbreaks in London.

- *Clusters* around water pumps

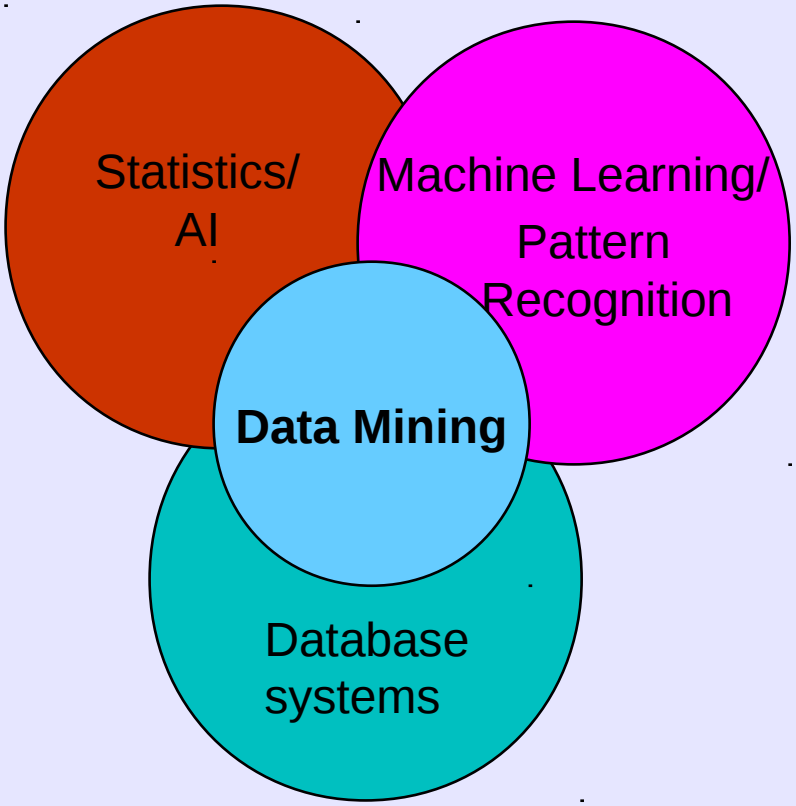
Introduction to Data Mining



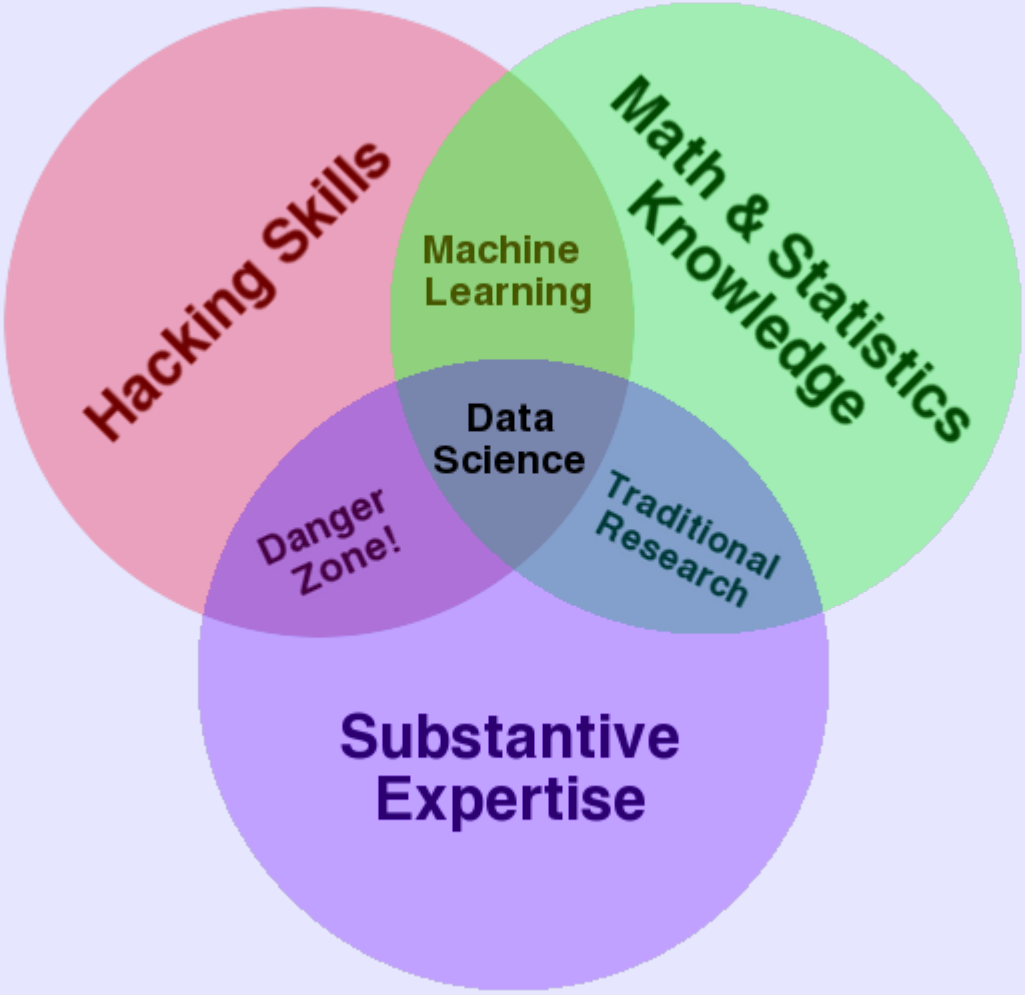
Feature selection
Dimensionality reduction
Normalization
Data subsetting

Introduction to Data Mining: Origins

- **Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems**
- **Traditional techniques may be unsuitable due to**
  - **Enormity of data**
  - **High dimensionality of data**
  - **Heterogeneous, distributed nature of data**



Introduction to Data Mining: Origins



Resources

- R
  - https://www.r-project.org/
- RStudio
  - https://www.rstudio.com/