Real World Data Mining…

By Renée Laury

OUTLINE

- REAL WORLD DATA
- NOISY DATA
- INCONSISTANT DATA
- MISSING DATA
- DATA CLEANSING
- DATA INTEGRATION
- DATA REDUCTION
- REVIEW

REAL WORLD DATA MINING

- WHAT DOES DATA COLLECTIONS LOOK LIKE IN THE REAL WORLD?
- HOW DOES IT COMPARE WITH CLASSROOM PROJECTS?
- MOST IMPORTANTLY, HOW CAN WE FIX THESE PROBLEMS?

NOISY DATA

INCLUDES

- ERRORS
- INCONSISTANCING
- OUTLIER VALUES WHICH DEVIATE FROM THE NORM OR EXPECTED VALUES

INCONSISTANT DATA

- DATA COLLECTED WITH INCONSISTED CODES OR FROMS
- EXAMPLE
  - WILLIAM JAMES SMITH
  - W. J. SMITH
  - BILL SMITH
  - B. SMITH

DO THESE NAMES REPRESENT THE SAME PERSON?????

MISSING DATA

- IMPORTANT DATA NOT INCLUDED IN DATA BASE
- WHY???
  - EQUIPMENT MALFUNCTION
  - NOT ENTER EITHER BECAUSE OF MISUNDERSTANDING OR NOT REQUESTED AT TIME OF ENTRY
  - OVERLOOKED
  - OUT OF DATE

DATA CLEANSING

- CLEANSING DATA MAKES THE DATA BASE EASIER TO WORK WITH
- FILLS IN MISSING VALUES
- SMOOTHES NOISE
  - IDENTIFY AND REMOVES OUTLIERS
  - BINNING
- RESOLVES INCONSITANCIES

DATA INTEGRATION

- REMEMBER THE QUESTION OF NAME???
  - WILLIAM SMITH
  - W. SMITH
  - BILL SMITH
- REDUNDANT DATA
- WHAT DO YOU DO???
  - CLUSTER
  - CLEAN
  - INTEGRATE

DATA INTEGRATION con't

ANNUAL SALARY

NEED TO NORMALIZE

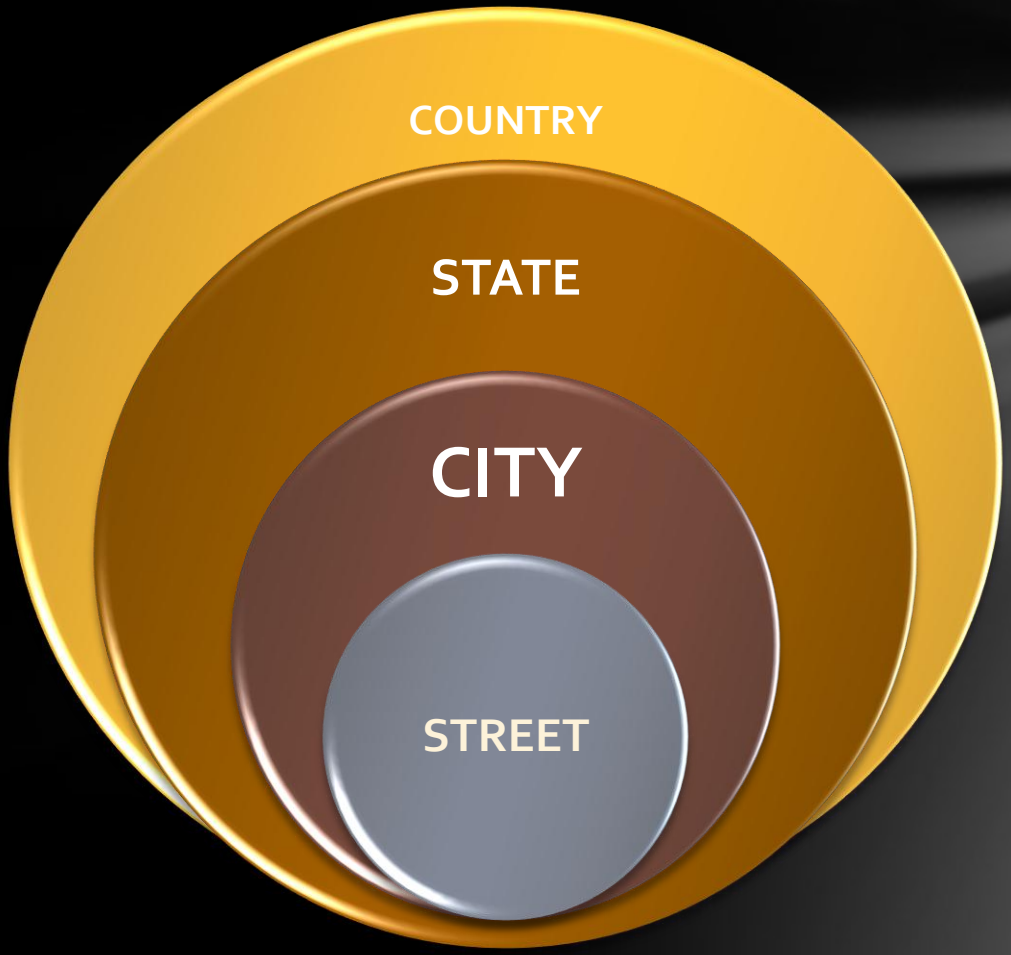
$$\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$$

DATA REDUCTION

- GENERALIZATION
  - LOCATION
- DATA CUBES
- REMOVE IRRELEVANT DATA
  - DOES GENDER MATTER TO YOUR PROJECT??

DATA REDUCTION CON'T

EXAMPLE:



REVIEW

- REAL WORLD DATA
- NOISY DATA
- INCONSISTANT DATA
- MISSING DATA
- DATA CLEANSING
- DATA INTEGRATION
- DATA REDUCTION

REFRENCES

Han, J., & Kamber, M. (2001). *Data Mining Concepts and Techniques.* San Francisco: Academic Press.