Data Quality

Data Uncertainty may occur during :

Data collection: data collection devices (including sensor, RFID readers)are sometimes imperfect; data extraction techniques(某些其他的信息获取技术、算法，而不是设备） are often inaccurate

Data transmission: may occur in sensor networks(packet losses, transmission errors), GPS, (Reflection or refraction of the satellite signal)

Data processing : Privacy preserving, lossy data compression, data integration

Data Uncertainty 的分类：

从sources of data uncertainty角度
Undesirable uncertainty
desirable uncertainty

Granularity
Tuple uncertainty
Attribute uncertainty

Correlations
Independent Uncertainty
Correlated Uncertainty
Uncertainty with local correlations

Data Quality 的五个指标：ACUTC (Acute 把e换成c，敏锐的）
Accuracy
Completeness
Uniqueness
Timeliness
Consistency

改善数据质量的方法：
Anomaly Data Identification and Data Cleansing
什么是data cleaning?
Data Cleansing (or data scrubbing) is detecting and correcting (or removing) errors and inconsistencies（错误和不匹配）in order to improve the quality of data（目的）.Used mainly in database(在数据库中的另一种描述）,the term refers to identifying incorrect, incomplete, inaccurate, irrelevant parts of the data and then replacing, modifying or deleting this dirty data.

Anomaly Data detection:怎么找到异常数据
classification
clustering
statistical(前三个是我在机器学习里遇到过的)
nearest neighbor-based (anomaly detection)
spectral

data cleaning 的四个任务：
Fill in missing values(缺少的正确的)
Resolve redundancy caused by data integration（多余的正确的）
Identify outliers and smooth out noisy data（多余的错误的）
Correct inconsistent data（多余的错误的）

如何处理Missing data (丢失数据)
Missing data may need to be inferred.
Ignore the tuple
fill in the missing value manually
fill in it automatically

如何处理Noisy Data?
binning
regression
clustering
combined computer and human inspection

重点介绍binning的过程：
1. first sort data and partition into bins
2. smooth by bin means, median, boundaries.

什么是Data integration?
combines data （from multiple sources） into a （coherent） store（存储）.

redundancy in data integration 的原因：
1.object identification: the same attribute or object may have different names in different databases.
2. Derivable data: one attribute my be a “derived” attribute in another table

如何解决redundancy in data Integration?
redundant attributes may be able to be detected by correlation analysis

有哪些data transformation的手段：
1. smoothing
2. aggregation and generalization
3. normalization
4. attribute/feature construction

什么是data reduction?
Obtain a reduced representation of the data set that is much smaller in volume but yet produce the same(or almost the same) analytical results
为什么需要data reduction?
1. a database’s storage is limited
2. complex data analysis may take a long time run on the data set

Sensor cleaning 的步骤（pipeline)
virtualize <- arbitrate <- merge <- smooth <- point
利用了temporal and spatial characteristics of sensor data （使用了时间和空间特点）
point: filter individual values（例如errant RFID tags, obvious outliers（异常值）, conversion of raw data into tuples)

smoothing: interpolates lost readings (temporal)
方法：window based queries

Merge:spatial interpolation
例如取一个空间内的均值

arbitrate : remove conflicting readings and duplication

virtualize: multi-source integration（有点类似于data integration)

什么是数据融合？Data fusion
data fusion, is generally defined as the use of techniques that combine data from multiple sources and gather that information in order to achieve inferences（推断）more efficient and more accurate. (than if they were achieved by means of a single source) （比data integration多了个目的）

三个数据融合处理架构：
data-level fusion : direct fusion of sensor data
feature-level fusion : representation of sensor data via feature vectors, with subsequent fusion of the feature vectors.
decision-level fusion : processing of each sensor to achieve high-level inferences or decisions

if the multi-sensor data are commensurate (measuring the same physical phenomena) then the raw sensor data can be directly combined ——data-level fusion

feature-level fusion involves the extraction of features from sensor data. Features are extracted from multiple sensor observations and combined into a single concatenated feature vector（不同类型的，比如说新闻舆论和股票价格concat）

Decision-level fusion:
decision-level fusion combines sensor information after each sensor has made a preliminary determination .

Database 是什么?
collection of persistent data

Data是什么?
known facts that can be recorded and have an implicit meaning.

Database management system(DBMS)是什么?
（software） system that supports creation, population and querying （of a database ）

database system = DBMS + Database

typical DBMS Functionality:DMBS的功能是什么?
1. Define a particular database in terms of its data types, structures and constraints
2. construct or load the initial database （contents） （on a secondary storage medium）
3. Manipulate the database : retrieval, modification, accessing the database through Web application
4. Share a database allows multiple users and program to access the database simultaneously

Database system的特性：self-describing
database system contains not only the database itself but also a complete definition of the database structure and constrains.

补充：The information stored in the catalog is called meta-data(data about data), and it describes the structure of the primary database.

另外，数据库的查看方式也有很多种
each user may see a different view of the database, which describes only the data of interest to that user.
没有显式存储，但是derived from the database 叫做virtual data

allowing a set of concurrent users to retrieve from and to update the database（好多人同时获取和上传数据）,concurrency control within the DBMS guarantees that each transaction is correctly executed.

数据存储的三种解决方案：
Data storage solutions:
Direct attached storage(DAS)
Storage area network(SAN)
Network attached storage(NAS)
三个图也要记住
DAS Characteristics:
(storage) devices attached (directly) to servers(only point of access)
好处:
they are relatively inexpensive, familiar to most IT departments, and are widely useable.
坏处：
single-server access is limiting and DAS networks have distance limitations.

NAS:
好处：
enables faster access to applications. Multiple servers can access the same device, and a NAS is more reliable than a DAS.
坏处：
NAS is limited by LAN bandwidth.

SAN:
好处：provide highly scalable performance.
坏处：much more expensive than other storage networking solutions.

Data Mining 是什么?
the （automatic） discovery of relationships in typically large database. And in some instances, the use of the discovery results in predicting relationships.
data mining 的功能：
data mining functionalities are used to specify the kind of patterns to be found in data mining tasks.

datamining 的task任务分类：
descriptive: characterize （general） properties of data in the database.
predictive: perform inference on data （to make predictions）

为什么要做数据挖掘：
1. 其实是前提条件：lots of data is being collected and warehoused.
2. 一样是前提条件computing has become affordable
3. competitive pressure is strong.

主要的major data mining task:
classification
association
clustering

classification的定义：
find a model for class attribute as a function of the values of other attributes.

association（关联）的定义：

transaction(交易) data analysis

given a database （of transactions）, find all association rules: the presence of one set of items implies the presence of another set of items.

clustering 定义：

given a set of data points, （each having a set of attributes, and a similarity measure（相似度的分析方法，例如范数） among them, ）find clusters such that data points in one cluster are more similar, data points in separate clusters are less similar.

Multimedia 的定义：

Multimedia is a combination of text, graphic, sound, animation and video that is delivered interactively to the user by electronic or digitally manipulated means.

text: is the most basic elements of multimedia. convey the intended message to the users.

graphic: could be produced manually or computer graphics technology

audio: produced by vibration (震动)

animation: created by the consecutive display of images （of static elements）

video: capturing, recording, processing, transmitting and reconstructing moving pictures.

什么是digital image: is a representation of 2d image as a finite set of digital values, called pixels.

digitization implies that digital image is an approximation of a real scene.