



Data Challenges

Gregory Piatetsky-Shapiro
KDnuggets



Need Big Data Repository

- UCI KDD Repository : only 6 medium size datasets
- KDD Cup data – limited
- NCDM, Chicago (Bob Grossman) has facilities for TB data storage



Big Data Repository

- For commercial data, anonymize
 - Avoid AOL Search Data fiasco
- Encourage data submission for DM Conferences submissions



Need Common Meta-data format

Many data formats

- Open source:
 - .csv, .txt, .arff
- Proprietary: SAS, Excel
- In SQL databases
- “non-standard” data: text, images, ...

Can have a common meta-data format



Reuse PMML subset for data

Excerpt from Visits.xml file for Visits database

```
<DataDictionary numberOfFields="2">  
  <DataField name="id" optype="continuous" dataType="integer">  
    <Extension name="storageType" value="numeric"/>  
  </DataField>  
  <DataField name="category" optype="categorical"  
    dataType="string">  
    <Extension name="storageType" value="string"/>  
    <Value value="bbs" property="valid"/>  
    <Value value="business" property="valid"/>  
  </DataField>  
</DataDictionary>
```



Meta-data Language

- Standard types
 - Numeric
 - String
 - Date
 - Categorical
 - ...

But also



Add Semantic Types

- E.g. IP address
 - Country code, Country name, organization, ...
- Telephone:
 - Country code, area code, ...
- Many micro-types



Data Mining Metadata Language

- DMML – should be simple, easy to use
 - Subset of PMML
- Converters from common types
- Common ontology for semantic types can be developed
 - (similar to GO: Gene Ontologies)
- US and EU and Asian organizations can fund development