THE UNIVERSITY of EDINBURGH
informatics

Informatics 1, 2010
School of Informatics, University of Edinburgh

Data and Analysis

Revision Lecture: Answering the why-question

Areti Manataki

Today

- What have we learnt?
- How are the different topics related?
- Why bother?

What have we learnt?

Course Content

…the principles underlying a variety of technologies for gathering, storing, retrieving and analysing data

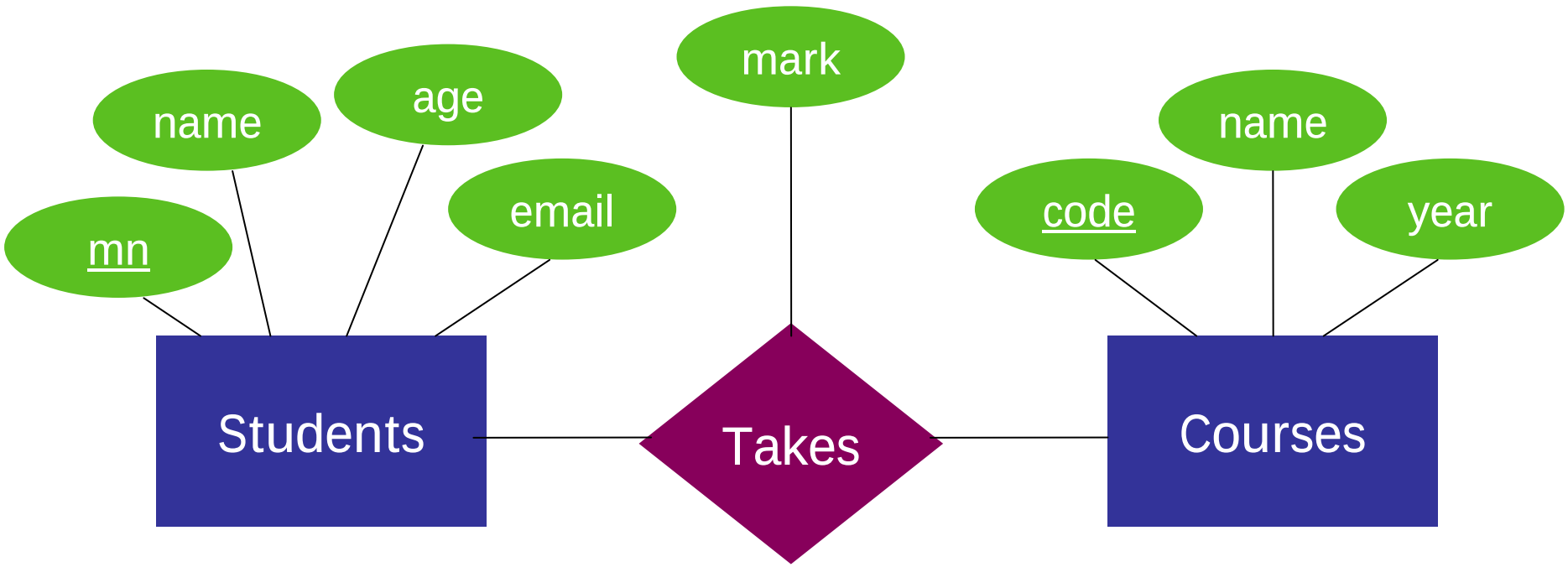
- Structured Data
- Semistructured Data and Text Corpora
- Unstructured Data and its Analysis

1. Structured Data

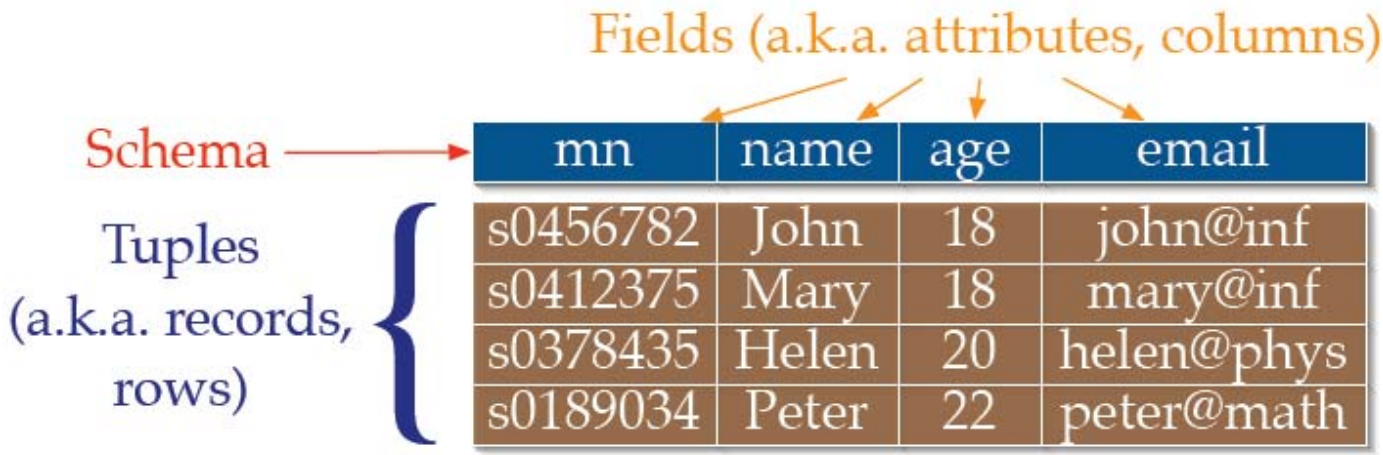
…data inherently structured: Inf1 students' data, songs' data

- Data Representation
  - ER data model
  - Relational model
- Data Manipulation
  - Relational algebra
  - Tuple relational calculus
  - SQL

Structured Data – Representation (1/3)

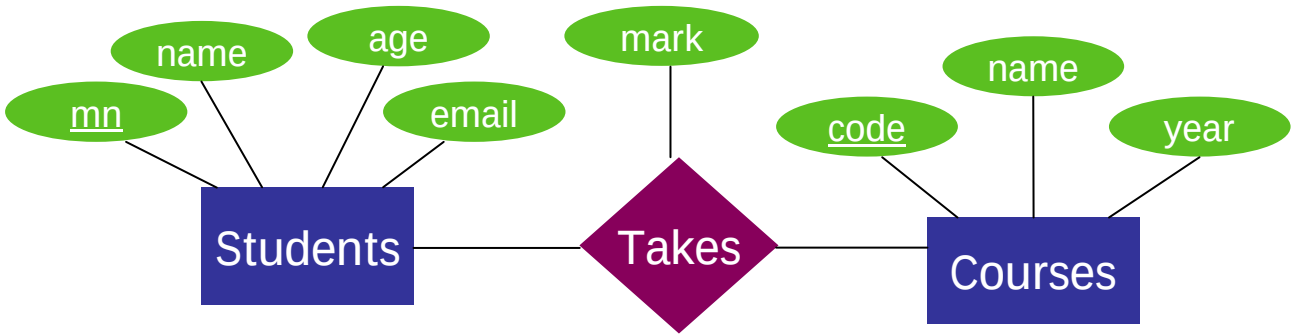


Structured Data – Representation (2/3)




```
create table Students (
  mn char(8),
  name char(20),
  age integer,
  email char(15),
  primary key (mn)
  )
```

Structured Data – Representation (3/3)



mn	name	age	email
s0456782	John	18	john@inf
s0412375	Mary	18	mary@inf
s0378435	Helen	20	helen@phys
s0189034	Peter	22	peter@math

mn	code	mark
s0456782	inf1	80
s0412375	geo1	78
s0412375	inf1	56
s0189034	math1	62

Referencing relation

code	name	year
inf1	Informatics 1	1
math1	Mathematics 1	1
geo1	Geology 1	1
dbs	Database Systems	3
adbs	Advanced Databases	4

Referenced relations

Primary key

Foreign key

Structured Data - Manipulation

RA

$$\pi_{\texttt{Students.name}}(\texttt{Students} \bowtie_{\texttt{Students.mn=Takes.mn}} (\texttt{Takes} \bowtie_{\texttt{Takes.code=Courses.code}} (\sigma_{\texttt{name='Informatics 1'}}(\texttt{Courses}))))$$

TRC

$$\{P \mid \exists S \in \texttt{Students}\ \exists T \in \texttt{Takes}\ \exists C \in \texttt{Courses}\ (C.\texttt{name} = \text{'Informatics 1'} \wedge C.\texttt{code} = T.\texttt{code} \wedge S.\texttt{mn} = T.\texttt{mn} \wedge P.\texttt{name} = S.\texttt{name})\}$$

SQL

```
SELECT S.name
FROM Students S, Takes T, Courses C
WHERE S.mn = T.mn AND T.code = C.code
      AND C.name = 'Informatics 1'
```

Structured Data – Tricky points

- Key & participation constraints @ ER model
- Mapping relationship sets to relational schemas, foreign key constraints
- Queries returning a new table (with fields of interest)

2. Semistructured Data

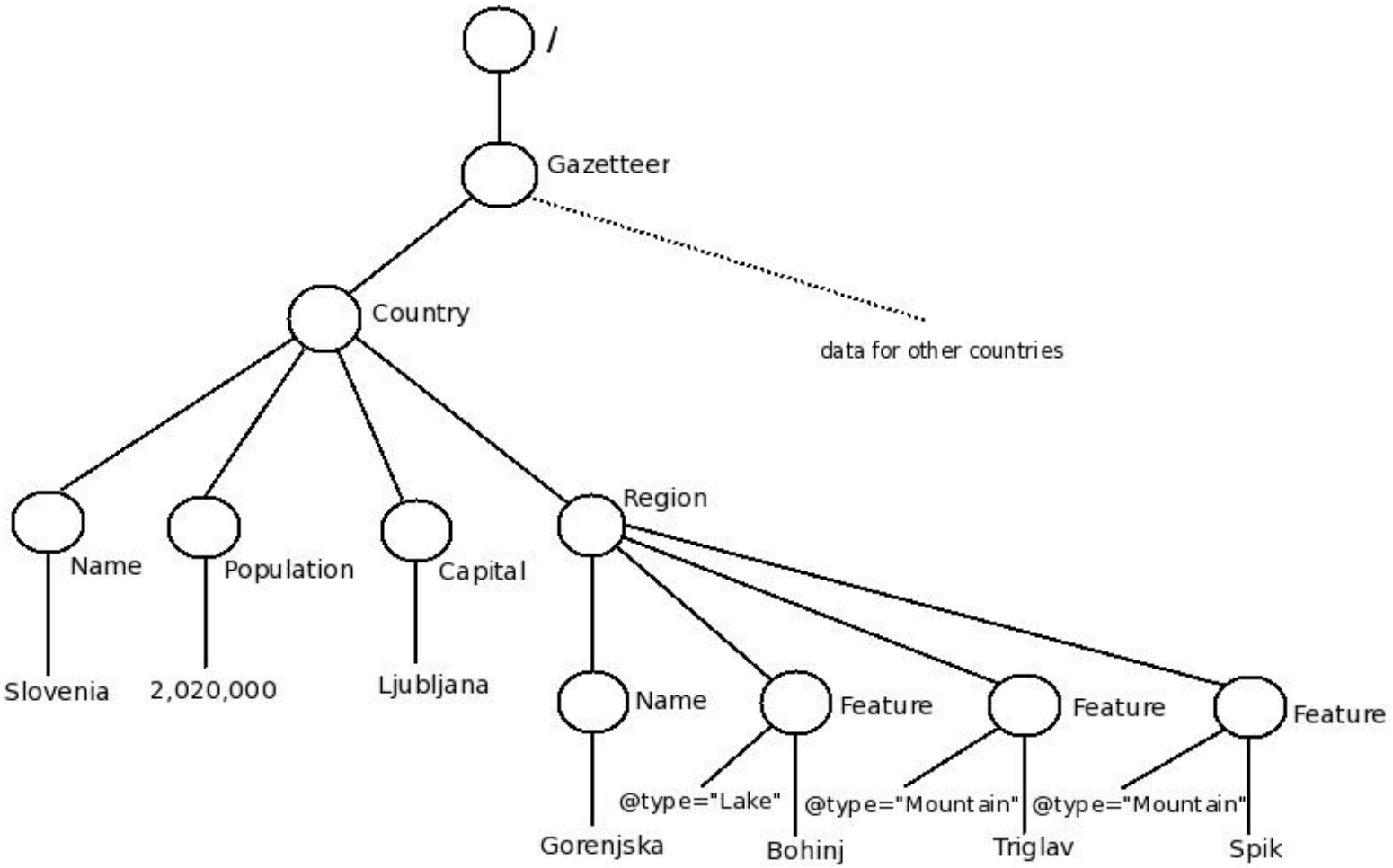
…data loosely structured: restaurants in Edinburgh, Dickens corpus

- XML
  - Structuring XML
  - Navigating XML
- Corpora
  - Building a corpus
  - Querying a corpus

Semistructured Data – Structuring XML (1/3)

```
<Gazetteer>
  <Country>
      <Name>Slovenia</Name>
      <Population>2,020,000</Population>
      <Capital>Ljubljana</Capital>
      <Region>
        <Name>Gorenjska</Name>
        <Feature type="Lake">Bohinj</Feature>
        <Feature type="Mountain">Triglav</Feature>
        <Feature type="Mountain">ˇSpik</Feature>
      </Region>
  </Country>
<!-- data for other countries here -->
</Gazetteer>
```

Semistructured Data – Structuring XML (2/3)



Semistructured Data – Structuring XML (3/3)

```
<!ELEMENT Gazetteer (Country+)>
<!ELEMENT Country (Name,Population,Capital,Region*)>
<!ELEMENT Name (#PCDATA)>
<!ELEMENT Population (#PCDATA)>
<!ELEMENT Capital (#PCDATA)>
<!ELEMENT Region (Name,Feature*)>
<!ELEMENT Feature (#PCDATA)>
<!ATTLIST Feature type CDATA #REQUIRED>
```

Semistructured Data – Navigating XML

```
//Region
/descendant::Region

//Feature/@type
/descendant::Feature/attribute::type

//Feature[@type='Mountain']/text()
/descendant::Feature[attribute::type='Mountain']
  /child::text()
```

Semistructured Data – Building a corpus

- Criteria for a corpus:
  - sampling and representativeness
  - finite size
  - machine-readable form
  - a standard reference
- Tasks for building a corpus
  - Collect data (balancing & sampling)
  - Preprocessing & Annotation

Semistructured Data – Querying a corpus

- Interested in:
  - concordances
  - frequencies
  - bigrams & collocations

- CQP [*not examinable*]
  - `"great" "deal";`
  - `"kiss.*";`
  - `[(pos = "N.*") & (word = "lock")];`
  - ```
    q4 = [pos = "JJ.*"] [pos = "N.*"];
    count q4 by word;
    ```
    ```

Semistructured Data – Tricky points

- XPath data model: attribute nodes, root node/root element, order in tree/XML document
- XPath queries: predicates

3. Unstructured Data

…data with no identifiable structure: bitmaps for pictures, digitalised sound, experimental results

- Information Retrieval
- Statistical Analysis

Unstructured Data – Information Retrieval

- Specification issues
  - Evaluation: precision VS recall
  - Query type
  - Retrieval model: vector space model

	Term_1	Term_2	Term_3	...	Term_n
Doc_1	14	6	1	...	0
Doc_2	0	1	3	...	1
Doc_3	0	1	0	...	2
...	...	...	...	...	...
Doc_N	4	7	0	...	5

$$\cos(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}||\vec{y}|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

Unstructured Data – Statistical Analysis

- Data scales
- Summary statistics: mean, median, mode, variance, standard deviation
- Hypothesis testing & correlations
- χ^2 test

$$\rho_{x,y} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N\sigma_x\sigma_y}$$

O_{ij}	sub	¬sub	
att	O_{11}	O_{12}	$R_1 = O_{11} + O_{12}$
¬att	O_{21}	O_{22}	$R_2 = O_{21} + O_{22}$
	$B_1 = O_{11} + O_{21}$	$B_2 = O_{12} + O_{22}$	N

E_{ij}	sub	¬sub	
att	$E_{11} = B_1R_1/N$	$E_{12} = B_2R_1/N$	$R_1 = E_{11} + E_{12}$
¬att	$E_{21} = B_1R_2/N$	$E_{22} = B_2R_2/N$	$R_2 = E_{21} + E_{22}$
	$B_1 = E_{11} + E_{21}$	$B_2 = E_{12} + E_{22}$	N

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Unstructured Data – Tricky points

- Different formulas for variance, standard deviation and correlation for actual population / sample-based estimation

$$\sigma = \sqrt{\mathrm{Var}} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}} \qquad s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n-1}}$$

- p is not the probability that the null hypothesis is true, but represents the probability that we would obtain a result similar to R if the null hypothesis were true

p	0.1	0.05	0.01	0.001
χ^2	2.706	3.841	6.635	10.828

Brief Overview

Data Type	Representation	Analysis
Structured Data	• ER model • Relational model	• Relational algebra • Tuple relational calculus • SQL
Semistructured Data	• XML • Corpora	• XPath • CQP
Unstructured Data		• Information retrieval • Statistics

How are the different topics related?

ER model VS. Relational model

- ER model
  - conceptual modelling
  - to visualise data and their dependencies
- Relational model
  - logical design, implementation
  - data can be queried

Relational model VS. XML (1/4)

- Relational model
  - rigid structure on data
  - relational nature
- XML
  - less rigid structure
  - hierarchical nature
  - publicly available in a standard and easily readable data format
  - to markup unstructured data with additional information

Relational model VS. XML (2/4)

```
<!ELEMENT Students (Student*)>
<!ELEMENT Student (mn,name,age,email)>
<!ELEMENT mn (#PCDATA)>
<!ELEMENT name (#PCDATA)>
<!ELEMENT age (#PCDATA)>
<!ELEMENT email (#PCDATA)>
```

```
<Students>
   <Student> <mn>s0456782</mn> <name>John</name>
     <age>18</age> <email>john@inf</email> </Student>
   <Student> <mn>s0412375</mn> <name>Mary</name>
     <age>18</age> <email>mary@inf</email> </Student>
   <Student> <mn>s0378435</mn> <name>Helen</name>
     <age>20</age> <email>helen@phys</email> </Student>
   <Student> <mn>s0189034</mn> <name>Peter</name>
     <age>22</age> <email>peter@math</email> </Student>
</Students>
```

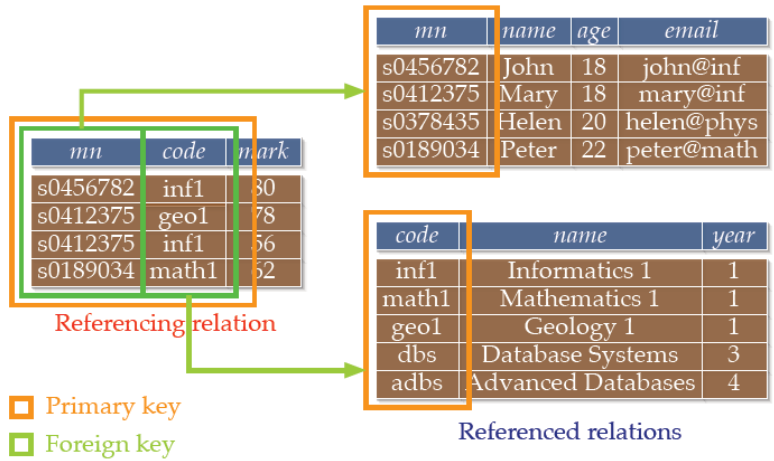
```
create table Students (
   mn char(8),
   name char(20),
   age integer,
   email char(15),
   primary key (mn)
   )
```

mn	name	age	email
s0456782	John	18	john@inf
s0412375	Mary	18	mary@inf
s0378435	Helen	20	helen@phys
s0189034	Peter	22	peter@math

Students

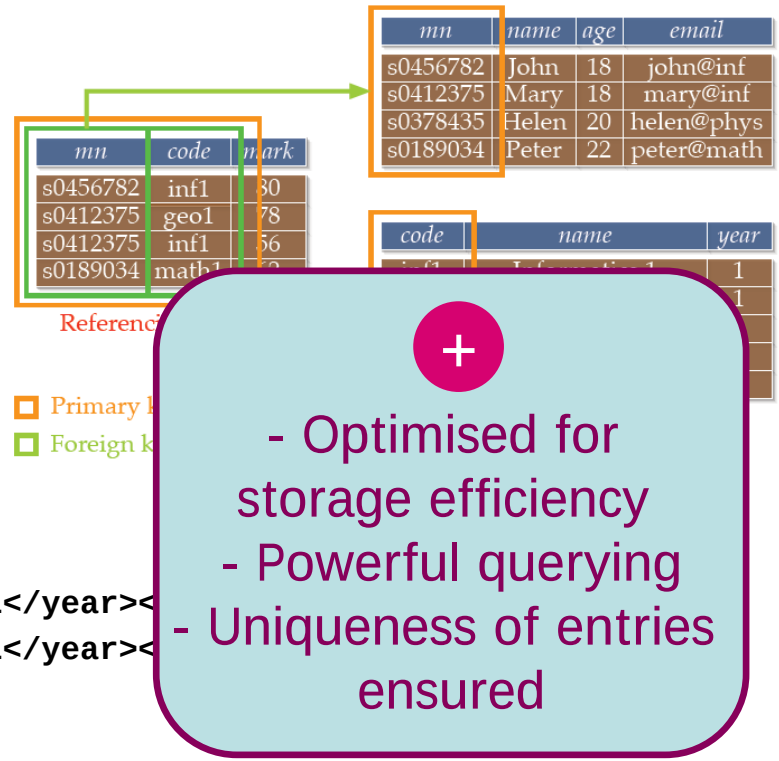
Relational model VS. XML (3/4)

```
<UniversityData>
   <Students>
     <Student> <mn>s0456782</mn> <name>John</name>
       <age>18</age> <email>john@inf</email> </Student>
     <Student> <mn>s0412375</mn> <name>Mary</name>
       <age>18</age> <email>mary@inf</email> </Student>
     <Student> <mn>s0378435</mn> <name>Helen</name>
       <age>20</age> <email>helen@phys</email> </Student>
     <Student> <mn>s0189034</mn> <name>Peter</name>
       <age>22</age> <email>peter@math</email> </Student>
   </Students>
   <Courses>
     <C><code>inf1</code><name>Informatics 1</name><year>1</year></C>
    <C><code>math1</code><name>Mathematics 1</name><year>1</year></C>
   </Courses>
   <Takes>
     <T><mn>s0412375</mn><code>inf1</code><mark>80</mark></T>
     <T><mn>s0378435</mn><code>math1</code><mark>70</mark></T>
   </Takes>
</UniversityData>
```



Relational model VS. XML (4/4)

```
<UniversityData>
   <Students>
     <Student> <mn>s0456782</mn> <name>John</name>
       <age>18</age> <email>john@inf</email> </Student>
     <Student> <mn>s0412375</mn> <name>Mary</name>
       <ag                      /email> </Student>
     <Stud                      e>Helen</name>
       <ag                      s</email> </Student>
     <Stud                      e>Peter</name>
       <ag                      h</email> </Student>
   </Stude
   <Course
     <C><c                      atics 1</name><year>1</year><
    <C><cod                     matics 1</name><year>1</year><
   </Courses>
   <Takes>
     <T><mn>s0412375</mn><code>inf1</code><mark>80</mark></T>
     <T><mn>s0378435</mn><code>math1</code><mark>70</mark></T>
   </Takes>
</UniversityData>
```



\+

- Easily readable & transferable across platforms
- Ordered

\+

- Optimised for storage efficiency
- Powerful querying
- Uniqueness of entries ensured

XML VS. Corpora

…XML is the most widely used markup language for corpora

```
<wtext type="FICTION">
  <div level="1">
    <head> <s n="1">
      <w c5="NN1" hw="chapter" pos="SUBST">CHAPTER </w>
      <w c5="CRD" hw="1" pos="ADJ">1</w>
    </s> </head>
    <p> <s n="2">
      <c c5="PUQ"> </c>
      <w c5="CJC" hw="but" pos="CONJ">But</w>
     ....
</wtext>
```

…an alternative to using a dedicated concordance program (e.g. CQP) is to use XML query technology (XPath and XQuery) to search any corpus implemented in XML

Unstructured data VS. (Semi)structured

- (Semi)structured data
  - clear, imposed structure
- Unstructured data
  - no (imposed) structure
  - need for uncovering implicit structure
- Information Retrieval VS. Corpora
- Statistical Analysis VS. (Semi)structured data querying

What representation and analysis method should I use and when?

- The data will tell you
  - structure?
  - hierarchical? relational?
- The application will tell you
  - type of queries?
  - transferability?
  - efficiency?
  - frequent update?

Scenario: Relational model or XML?

- Scenario#1: You are hired by an online music store, Shalalala.com, to represent and analyse data of interest. The store has several registered *customers* that buy *music albums*. Relevant customer information includes account information (i.e. username, password), personal information (i.e. name, title) and contact information (i.e. email and address, consisting of street name, street number, city, zip code and country). Shalalala.com is interested in answering questions like *"who buys what"*.
- Scenario#2: Shalalala.com decides to sell its customer data to *another* company.

Scenario: SQL queries or statistical analysis?

- Scenario#3: Shalalala.com wants to employ targeted marketing methods. Therefore, it wants to find out whether there is a *relation* between the city where customers live and the genre of music they listen to. There are also suspicions (*hypothesis*) that Portishead albums are more popular during the winter, and Shalalala.com wants to find out whether this is actually true.

Relation to INF1-Computation&Logics?

Relation to INF1-Computation&Logics

- TRC
- DTD
- CQP

Why why why?

…why not just SQL?

…the focus of this course is not directly on individual technologies, but on the principles underlying these technologies

- SQL is influenced by RA and TRC
- Real-world database management systems use query optimisation techniques based on RA to find more efficient strategies for evaluating queries

Significant?

Course Content	Science	Industry	Everyday Life
Structured Data • ER / Relational model • RA, TRC, SQL			
Semistructured Data • XML & Xpath • Corpora & CQP			
Unstructured Data • Information retrieval • Statistics			

Significance

Course Content	Science	Industry	Everyday Life
Structured Data • ER / Relational model • RA, TRC, SQL	•Storing and searching chemical structures	•Sales records	•Mobile contacts
Semistructured Data • XML & Xpath • Corpora & CQP	•Markup for gene expression •Compare theories in linguistics	•Organising & publishing real estate listings •Development of translation tools	•Data published on the web •Basis for better machine translation
Unstructured Data • Information retrieval • Statistics	•Searching scientific databases •Analyse experimental results	•Searching for managerial reports •Analyse marketing/ financial data	•Searching the web •Estimating the duration of the Edinburgh-London journey

Relevant to you

…Informatics studies the representation, processing, and communication of information in natural and engineered systems

- Relevant Inf courses: database systems, machine learning & pattern recognition, multi-agent semantic web systems, natural language generation, querying and storing xml, applied databases, probabilistic modelling & reasoning, advanced databases, data mining & exploration…
- Relevant jobs: data analyst, database administrator, business intelligence architect, computational linguist, web developer…

…and to others

thanks and good luck!