Introduction: Elements of Biological Data Models

Prof. Daniel P. Miranker

Objectives:
– What is the course about?
  • Why is “data model” deserving of an entire course?
– How is the course organized?
  • What will I learn, and what is expected of me?
AFQ:
AFQ: Answers to Your First Questions

• Is this class only useful for biologists?
  – No, approaching computers from the data model is a (the) broadly accepted way of thinking about organizing computer systems. The biology applications are a means to understanding these ideas.

• How much biology do I need to know?
  – Almost none. It will be covered in class. The contemporary developments in biology that are creating the data are so new, even biology majors don’t know the story.

• Is there a lot of programming in this class?
  – Yes and no. You will be in a computer lab almost every week. You will not be writing out lines of code. You will get some visibility into this today.
  – Also, model solutions/programs are available for every homework. You are welcome to use the model code. Some team programming will be encouraged
Context of the Course

1. Genomic Revolution
2. A Discipline of Engineering
   Software is [finally] emerging

Goal: Learn about engineering large software systems through a biological application
Practical Goals:

(intended)

1. Be the non-software developer who can speak to the engineers.

(unintended)

2. If your goal is a job as a software developer, you’ll walk out of this class very employable.
What is a data model?
http://www.utexas.edu/its/windows/database/datamodeling/dm/overview.html

**Data Model**: A data model is a conceptual representation of the data structures that are required by a database application.

Key phrase: *conceptual representation*
- Think about it.
- Principles, Methods and Tools
The Revolution In Biology

• “Post-genomic era” = After the human genome was first completely sequenced, 2000.
  – Grand challenge initiated ~1990
  – (3.3 billion nucleotides, A,C,G &T)

• How was the human genome sequenced?
  – Man or machine?
Biologists discovered robots could do lab work (better).

- Not C3PO, but more like welding arms
Industrial Automation Makes it into Biology Labs.

- Mostly by the use of microfluidic pumps
  
  Keyword: “High-throughput”
Biological dogma

DNA
(coding genes)

TAC  GGA TGT TTC  GCG CTA

mRNA

AUG  CCU  ACA AAG GCG GAU

Proteins
(sequences of amino acids)

met  pro  thr  lys  ala  asp

M    P    T    K    A    D

Codon 3-nucleotides
Three Major Sources of Biological Data

1. Sequencing machines
   - Determine DNA sequences
2. DNA chips (misnomer)
   - Measures mRNA
3. Mass-spectroscopy
   - Measures proteins
Gene Expression Chips

- Each spot fluoresces if mRNA is present
- 64,000 – 4,000,000 spot per chip, record red, green
Mass-Spectrometers with Liquid Chromatography:

- Can process whole cell lysate i.e., all the proteins in a cell
  → 17,000 spectra in 12 hours., each spectra 30,000 real numbers
More coming every day (two, right here at UT)

Biology is feeling swamped by data.
  • evangelists speak to exponential growth of data.
Role of a Database? *Biology*

- Databases are assuming the role of laboratory notebooks
  - Previously, data was
    - Hard earned
    - Manually transcribed
  - Now,
    - High throughput machines
    - 1,000 - 100,000 data elements at once.

- **Archival Recording of Information**
  - Data
  - What is the data
  - How was it captured *(provenance)*
Role of a Database?

Computer Engineering

• Stores the input for functions and algorithms.
  • (starting point for doing other things.)

• How is the data used?
What is a data model?

http://www.utexas.edu/its/windows/database/datamodeling/dm/overview.html

Data Model: A data model is a conceptual representation of the data structures that are required by a database application.

Key phrase: *conceptual representation*

– *Think about it.*
– *Principles, Methods and Tools*
What goes wrong?

Example:

Hypothesis 1, temp. dependent?
Experiment 1, build a database for it:

<table>
<thead>
<tr>
<th>Input</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>temp</td>
<td>I2</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>O1</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
What goes wrong? (2)

• Scientific Method: New Hypothesis

Hypothesis 2, pressure dependent?

Experiment 2, build a database for it:

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>pres</td>
<td>O1</td>
</tr>
<tr>
<td>I2</td>
<td>O2</td>
</tr>
<tr>
<td>I3</td>
<td>O3</td>
</tr>
</tbody>
</table>
This goes wrong:

- Some time later

Hypothesis, both temp & pressure dependent?

Experiment 3 - NOT, just analyze the previous experiments together

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
</table>

The schema don’t match
Goals/Content of Course

1. Mini-course in Data/Software Engineering
   – Process & methods for organizing data/programs

2. Tools to support this
   – A picture says a thousand words…

3. Walk through developing an application
Data Modeling In the Context of Database Design

1. planning and analysis
2. conceptual design  // logic without the details
3. logical design
4. physical design
5. implementation
## Inventor - Invention as DB Tables

<table>
<thead>
<tr>
<th>Invention</th>
<th>Inventor</th>
</tr>
</thead>
<tbody>
<tr>
<td>iid</td>
<td>name</td>
</tr>
<tr>
<td>1</td>
<td>structure</td>
</tr>
<tr>
<td>2</td>
<td>sequencing_machine</td>
</tr>
<tr>
<td>3</td>
<td>expression_chips</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>iid</th>
<th>first</th>
<th>last</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Francis</td>
<td>Crick</td>
</tr>
<tr>
<td>1</td>
<td>James</td>
<td>Watson</td>
</tr>
<tr>
<td>1</td>
<td>Rosalyn</td>
<td>Franklin</td>
</tr>
<tr>
<td>1</td>
<td>Maurice</td>
<td>Wilkens</td>
</tr>
<tr>
<td>2</td>
<td>Lee</td>
<td>Hood</td>
</tr>
<tr>
<td>3</td>
<td>David</td>
<td>Botstein</td>
</tr>
</tbody>
</table>
Inventor-Invention, Object Model

A list of inventions, each with their list of inventors

<table>
<thead>
<tr>
<th>Invention</th>
<th>Inventor</th>
</tr>
</thead>
<tbody>
<tr>
<td>name()</td>
<td>first()</td>
</tr>
<tr>
<td>inventors()</td>
<td>last()</td>
</tr>
</tbody>
</table>

1 *
Computer Aided Software Engineering (CASE)

• Computer’s help Civil Engineer’s and Architects (CAD)

• Why not, have computer’s help write software?

• The can & do:
  – We will learn to use Rational Rose
Just to show you a pretty picture (1)

Figure 2.27a UML Representation of a 1:1 Relationship

(a)
CREATE TABLE T_Invention (  
    iname VARCHAR (255) NOT NULL,  
    T_Invention_ID INTEGER NOT NULL,  
    CONSTRAINT PK_T_Invention0 PRIMARY KEY (T_Invention_ID)  
);  
CREATE TABLE T_Inventor (  
    Firnname VARCHAR (255) NOT NULL,  
    LastName VARCHAR (255) NOT NULL,  
    name SMALLINT NOT NULL,  
    T_Inventor_ID INTEGER NOT NULL,  
    T_Invention_ID INTEGER NOT NULL,  
    CONSTRAINT PK_T_Inventor1 PRIMARY KEY (T_Inventor_ID)  
);  
CREATE INDEX TC_T_Inventor1 ON T_Inventor (T_Invention_ID );  
ALTER TABLE T_Inventor ADD CONSTRAINT FK_T_Inventor0  
    FOREIGN KEY (T_Invention_ID) REFERENCES T_Invention  
    (T_Invention_ID)  
    ON DELETE NO ACTION ON UPDATE NO ACTION;
A commercial database has an average of ______ attributes per table
Application Example:

• Rosetta Sequence Analysis to
  – determine gene/protein function
Monomeric proteins that are found fused in another organism are likely to be functionally related and physically interacting.

Introduce self & Administrivia
Student’s turn

• Name, dept., year

• Why did you register for this course
  – (especially if you are not a biology major)

• What are you hoping to get out of this course?