MINI WORLDS

Database systems are designed to capture well-defined subsets of the real world, the so-called mini worlds.

**Mini World**
A mini world contains the relevant objects (or: entities, things) of a real-world subset. Only the significant attributes (or: characteristics) of these objects are preserved. Objects may relate to each other. Specific constraints (or: rules of the world) are captured as well.
MINI WORLDS

Mini worlds may represent subsets of our (true, physical) environment as well as any of the many *virtual worlds* that we create.

**Mini World Example: US Geological Survey Earthquake Maps**

**Real world**
- The Earth

**Subset covered**
- Global real-time earthquake information

**Relevant objects**
- Quakes, locations, date/time

**Significant attributes**
- Magnitude, latitude, longitude, depth, day, hour, min, sec, ...
MINI WORLDS

Available at http://earthquake.usgs.gov/earthquakes/map/
MINI WORLDS

Mini World Example: Enterprise Data (→ TPC-H[1])

Real world
  Company/Corporation

Subset covered
  Ordering and Fulfillment, Client Relationships, Supply Chain

Relevant objects
  Orders, Lineitems, Products, Suppliers, Customers, Shipments, ...

Significant attributes
  Product IDs, order/shipment dates, ordered quantities, prices, names, ...

Constraints
  “The price of an order must be the sum of the prices of its individual lineitems”
MINI WORLDS

Mini World Example: Web Sites (Amazon, Wikipedia, YouTube)

Real world
  The World Wide Web

Subset covered
  Web site (shop, encyclopedia, social networking)

Relevant objects
  Store inventory, shopping baskets, payment data, wiki page contents, video stream data

Constraints
  “When stock of item is below 10, that item has an order immediately notice”
Mini World Example: Movie Script

Real world
  Cinematography, movies, films

Subset covered
  Movie scripts (story, setting, roles, scenes)

Relevant objects
  Chapters, scenes, actors, characters, locations, character (co-)occurrence, dialogue, ...

Relationships
  Character is played by actor, scene is part of chapter, character occurs in scene, ...

Constraints
  “If an actor impersonates more than one character, these characters may not meet”
MINI WORLDS

Mini World Example: LEGO™ Sets, Bricks, Mini Figures

Real world
   LEGO toys

Subset covered
   Catalog of available LEGO sets (or: models) and their contents

Relevant objects
   Categories ("space", "city", ...), sets, individual bricks, mini figures, colors

Relationships
   Set contains bricks, brick is available in color, brick<sub>1</sub> is equivalent to brick<sub>2</sub>

Significant attributes
   Names, product IDs, quantity, 3D size (measured in studs), weight, image, ...

Constraints
   "If a set contains a piece, details for the piece must be available either in the brick or mini figure listing", "No two pieces share the same product ID"
LEGO SET 5610

- A sample object of this particular mini world:

LEGO Set 5610 “Builder”, Category Town (City, Construction)
Web site *BrickLink* hosts a database of the LEGO sets mini world:

Inventory of Set 5610–1

http://www.bricklink.com/catalogItemInv.asp?S=5610-1
DATA MODELS AND DATA LANGUAGES

**Data Model**
A data model defines a limited toolbox of constructs (or **types**) that can be used to represent the objects, attributes, and relationships of a mini world.

**Data Language**
Once the types are fixed, this also largely prescribes the **operations** we can perform with data of these types.

- Database systems are designed to effectively and efficiently support a **single data model and language** (we will see that support for “foreign” data models often feels awkward)
DATA MODEL: TEXT

Types

Text (strings of characters) in a particular encoding (e.g., Unicode/UTF–8). Typically, newline characters '\n' are used to break the text into lines.

Besides the line-breaking convention, the text data model imposes no further structure on the data (→ unstructured data model).

Operations

1. **Iterate** over the lines of a given text (e.g. contained in a file)
2. Inside a line, use **pattern matching** to **extract** individual/groups of characters
GenBank entry for *Saccharomyces cerevisiae* (Baker’s Yeast)

<table>
<thead>
<tr>
<th>LOCUS</th>
<th>SCU49845</th>
<th>5028 bp</th>
<th>DNA</th>
<th>PLN</th>
<th>21-JUN-1999</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEFINITION</td>
<td>Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Ax12p (AXL2) and Rev7p (REV7) genes, complete cds.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACCESSION</td>
<td>U49845</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VERSION</td>
<td>U49845.1</td>
<td>GI:1293613</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KEYWORDS</td>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SOURCE</td>
<td>Saccharomyces cerevisiae (baker's yeast)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ORGANISM</td>
<td>Saccharomyces cerevisiae</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes; Saccharomycetales; Saccharomycetaceae; Saccharomyces.
### DATA MODEL: TEXT

- GenBank entry for *Saccharomyces cerevisiae* (Baker’s Yeast) [cont’d]

<table>
<thead>
<tr>
<th>FEATURES</th>
<th>Location/Qualifiers</th>
</tr>
</thead>
<tbody>
<tr>
<td>source</td>
<td>1..5028</td>
</tr>
<tr>
<td></td>
<td>/organism=&quot;Saccharomyces cerevisiae&quot;</td>
</tr>
<tr>
<td></td>
<td>/db_xref=&quot;taxon:4932&quot;</td>
</tr>
<tr>
<td></td>
<td>/chromosome=&quot;IX&quot;</td>
</tr>
<tr>
<td></td>
<td>/map=&quot;9&quot;</td>
</tr>
<tr>
<td>CDS</td>
<td>&lt;1..206</td>
</tr>
<tr>
<td></td>
<td>/codon_start=3</td>
</tr>
<tr>
<td></td>
<td>/product=&quot;TCP1-beta&quot;</td>
</tr>
<tr>
<td></td>
<td>/protein_id=&quot;AAA98665.1&quot;</td>
</tr>
<tr>
<td></td>
<td>/db_xref=&quot;GI:1293614&quot;</td>
</tr>
<tr>
<td></td>
<td>/translation=&quot;SSIYNGISTSGDLNNGTIADMRQLGIVESYKLKRAVVSASEA AEVLLRVDNIRARPRTANRQHM&quot;</td>
</tr>
<tr>
<td>gene</td>
<td>687..3158</td>
</tr>
<tr>
<td></td>
<td>/gene=&quot;AXL2&quot;</td>
</tr>
</tbody>
</table>

 [...]
DATA MODEL: TEXT

- GenBank entry for *Saccharomyces cerevisiae* (Baker’s Yeast) [cont’d]

```
[...]
ORIGIN
  1 gatcctccat atacaacggt atctccacct caggttttga tctcaacaac ggaaccattg
  61 ccgacatgag acagtttagg atcgctcggaga gttacaagct aaaacgagca gtagtcagct
 121 ctgcatctga agccgctgaa gttctactaa gggtggataa catcatccgt gcaagaccaa
181 gaaccgcctaa tagacaacat atgtaacata tttaggatat acctcgaaaa taataacccg
241 ccacactgtc attattataa ttagaaacag aacgcaaaaa ttatccacta tataattca
[...]
//
```

- Aims for readability by humans *and* machines. Formatting conventions are obeyed to facilitate the construction of parsers for GenBank entries:

```
//⟨key⟩=⟨value⟩
```
Example: LEGO Set 5610 (BrickLink)

- Represent catalog information about LEGO Set 5610 ("Builder") along with a detailed listing of the set contents (bricks, minifigures).
- This text file format primarily aims for **human readability**. The listing of the contents follows line-based formatting conventions that provide **hooks for parsing**.

**LEGO™ Set "Builder" (set no 5610-1)**

Category: Town (City, Construction)

Contains 20 pieces: 19 bricks, 1 minifigure

5610-1 Builder is a City impulse set released in 2008. It contains a construction worker with a rolling cement mixer, along with 3 dark grey studs that resemble mortar or concrete. When the mixer is pushed, the drum turns. The drum can also tilt side-to-side, but not enough to dump the studs.

[...]
---

**DATA MODEL: TEXT**

Catalog information for LEGO Set 5610 [cont’d, here: listing of set contents]

<table>
<thead>
<tr>
<th>Brick#</th>
<th>Color/Weight</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1x 6157</td>
<td>Black/1.12g</td>
<td>Plate, Modified 2 x 2 with Wheels Holder Wide</td>
</tr>
<tr>
<td>2x 3139</td>
<td>Black/0.4g</td>
<td>Tire 14mm D. x 4mm Smooth Small Single</td>
</tr>
<tr>
<td>1x 3839b</td>
<td>Black/0.61g</td>
<td>Plate, Modified 1 x 2 with Handles - Flat Ends, [...]</td>
</tr>
<tr>
<td>1x 30663</td>
<td>Black/0.4g</td>
<td>Vehicle, Steering Wheel Small, 2 Studs Diameter</td>
</tr>
<tr>
<td>1x 6222</td>
<td>Dark Bluish Gray/3.57g</td>
<td>Brick, Round 4 x 4 with Holes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Minifig#</th>
<th>Weight</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1x cty052</td>
<td>3.27g</td>
<td>Construction Worker - Orange Zipper, Safety Stripes, Orange</td>
</tr>
</tbody>
</table>
- Sample problem (or query):
  What is the overall weight (in grams) of LEGO Set 5610?

- Possible plan of attack:
  1. **Iterate** over the lines of the catalog entry
  2. Use **pattern matching** to identify lines of the form (\_ = space)

    \[
    \langle \text{quantity} \rangle \times \text{...} / \langle \text{weight} \rangle \text{g ...} \\
    \langle \text{quantity} \rangle \times \text{...} \_ \langle \text{weight} \rangle \text{g ...}
    \]

  3. **Extract** quantity and weight in each such line
  4. **Multiply** quantity and weight and aggregate (i.e. sum up) as needed
The unstructured text data model provides poor support for queries even of this simple kind. One option: rely on UNIX’ text processing tools like grep, sed, and awk to implement the plan.

**sed** (stream editor)

Operates over '\n'-separated lines of text, can filter lines based on regular expressions, can modify and then print selected lines. Example:

```
    sed -E -e 's/⟨regular expression⟩/⟨modification⟩/p; ⋯'
```

Good match for the text data model: 
**sed** implicitly **iterates** over the lines (of its standard input), **pattern matches** and can **extract** select portions of matched lines.
**DATA MODEL: TEXT**

**awk** (Aho, Kernighan, Weinberger)

Iterate over '\n'-separated lines of text, apply rules of the following form to each line

\[
\langle \text{pattern} \rangle \{ \langle \text{action} \rangle \}
\]

- If \langle \text{pattern} \rangle matches, \langle \text{action} \rangle can extract the fields $1$, $2$, ... of the matched line and perform (simple) computation.
  Specific patterns: //\langle \text{regular expression} \rangle//, //, BEGIN, END.
- What constitutes a field is determined by field separator string FS
DATA MODEL: TEXT

- sed and awk script to compute the weight of LEGO Set 5610:

```bash
#! /bin/sh
# Compute the overall weight of all pieces in LEGO set 5610-1.

# Notes:
# - assumes one piece per line and input of the form:
#       <quantity>x ... <weight>g ...
#   (everything else is considered noise and skipped over)
# - sed command `p': print pattern space, then process next line
#   `d': delete pattern space, next line

sed -E -e 's/^([0-9]+)x.+/\1 \2/p; d' |

awk '
    BEGIN { sum = 0 }
    //   { sum += $1 * $2 }
    END   { print sum }
'`
- Sample problem (or query):
  Extract a subsequence, specified by a \langle from \rangle-\langle to \rangle location, from the DNA origin sequence of a GenBank entry

- Possible plan of attack:
  1. **Iterate** over the lines of the GenBank entry
  2. Use **pattern matching** to identify the start and end of the DNA sequence (**ORIGIN** and **/**, respectively)
  3. Only when inside a sequence, use **pattern matching** to identify lines of the form
     \[
     \langle offset \rangle \cdots \langle amino acids \rangle \langle amino acids \rangle \langle amino acids \rangle \cdots
     \]
  4. **Extract** \langle amino acids \rangle fields, **aggregate** (here: concatenate) the extracted fields
  5. *(Cut the requested subsequence from the concatenated result)*
**awk** script to extract a DNA subsequence from a GenBank entry:

```bash
#!/bin/sh

FROM=$1
TO=$2

awk 'BEGIN { ORS = """; dna = 0; seq = """ } /ORIGIN/ { dna = 1 } dna && /^ *[0-9]+/ { for (i = 2; i <= NF; i++)
    seq = seq $i
 }
 /\//\// { dna = 0 } END { print seq }
|'
cut -c $FROM-$TO
```
DATA MODEL: NESTED ARRAYS AND DICTIONARIES

Types
- **Container types**: arrays and key/value pair dictionaries (or: hashes, association lists) and **basic atomic types** (e.g. numbers, strings, Booleans). Containers may contain atomic values as well as other containers (**nesting**).

Nested containers provide a multitude of data structuring options. Data models of this kind are commonly referred as being **semi-structured**.

Operations

1. Index- or key-based **lookup** into containers
2. (Nested) **iteration** over and **filtering** of container contents
3. **Construction** of new containers
4. **Computation** over basic types (comparison, arithmetics, ...), **aggregation**
Recent and now widespread instance of this data model: **JSON (JavaScript Object Notation)**, excerpt of the JavaScript language definition (notation for literal JavaScript objects).

Find a complete and compact one-page(!) JSON definition on [http://json.org](http://json.org). JSON has become popular as an inter-application data exchange format.

A recent proposal for a data language for this data model: **JSONiq** ([http://jsoniq.org](http://jsoniq.org)), derived from and interoperable with XQuery (the language for the XML data model).
DATA MODEL: NESTED ARRAYS AND DICTIONARIES

- JSON syntax (⟨string⟩ and ⟨number⟩ follow usual syntactic conventions):

\[
\begin{align*}
\langle \text{value} \rangle & \::= \langle \text{string} \rangle \\
& \quad | \langle \text{number} \rangle \\
& \quad | \text{true} | \text{false} \\
& \quad | \text{null} \\
& \quad | \langle \text{dict} \rangle \\
& \quad | \langle \text{array} \rangle \\
\langle \text{dict} \rangle & \::= \{ \} \\
& \quad | \{ \langle \text{members} \rangle \} \\
\langle \text{members} \rangle & \::= \langle \text{pair} \rangle \\
& \quad | \langle \text{pair} \rangle , \langle \text{members} \rangle \\
\langle \text{pair} \rangle & \::= \langle \text{string} \rangle : \langle \text{value} \rangle \\
\langle \text{array} \rangle & \::= [ ] \\
& \quad | [ \langle \text{elements} \rangle ] \\
\langle \text{elements} \rangle & \::= \langle \text{value} \rangle \\
& \quad | \langle \text{value} \rangle , \langle \text{elements} \rangle
\end{align*}
\]
DATA MODEL: NESTED ARRAYS AND DICTIONARIES

- Sample JSON value `construction` (JSONiq):

```json
let $set5610 := { "set": "5610-1",          
                 "pieces": [ { "brick": "6157", "quantity": 1 },       
                               { "brick": "3139", "quantity": 2 },       
                               { "brick": "3839b", "quantity": 1 } ] } 
return ⋯
```

- Key-based `lookup` into dictionary `d` (via `d.k`, "dot notation"), index-based `lookup` (via `a[[n]]`) into array `a` (read "→" as "evaluates to"):

```json
$set5610."set"  →  "5610-1"
$set5610."pieces"[[2]]  →  { "brick": "3139", "quantity": 2 }
$set5610."pieces"[[2]]."brick"  →  "3139"
```
DATA MODEL: NESTED ARRAYS AND DICTIONARIES

let $xs := [ 1, 2, 3 ]
let $ys := { "one": "eins", "two": "zwei", "three": "drei" }
return ⋯

- Iterating over container contents. Need to convert contents into a sequence of values first.
  Iteration yields a sequence again:

  ```
  for $x in members($xs)
  return $x         ↦ (1, 2, 3)

  for $y in keys($ys)
  return $y         ↦ ("one", "two", "three")

  for $y in keys($ys)
  return $ys.$y     ↦ ("eins", "zwei", "drei")
  ```
DATA MODEL: NESTED ARRAYS AND DICTIONARIES

let $xs := [ 1, 2, 3, 4, 5, 6 ]
return ⋯

- **Grouping** and **aggregation** of containers (+ computation, construction):

  for $x in members($xs)                         ( { "even": true, 
    group by $even := $x mod 2 = 0             ⧵->    "nums": [2, 4, 6] },
  return { "even": $even, "nums": $x }      { "even": false, 
                                                 "nums": [1, 3, 5] } )

  for $x in members($xs)                         ( { "even": true, 
    group by $even := $x mod 2 = 0             ⧵->    "sum": 12 },
  return { "even": $even, "sum": sum($x) }    { "even": false, 
                                                 "sum": 9 } )

- Note: In the **group by** clause, variable $x$ is bound to individual members of $xs$. In the **return** clause, $x$ is bound to a **sequence** of members (i.e., a group).
DATA MODEL: NESTED ARRAYS AND DICTIONARIES

Example: LEGO Set 5610 (BrickLink data modelled as JSON)

```json
{
  "set": "5610-1",
  "name": "Builder",
  "category": "town",
  "year": 2008,
  "pieces": [
    {
      "brick": "6157",
      "quantity": 1,
      "extra": false,
      "color": "Black",
      "weight": 1.12,
      "name": "Plate, Modified 2 x 2 with Wheels Holder Wide"
    },
    ...
    {
      "minifig": "cty052",
      "quantity": 1,
      "extra": false,
      "weight": 3.27,
      "name": "Construction Worker - Orange Zipper, ...
    }
  ]
}
```
Sample problem (or query):
What is the overall weight (in grams) of LEGO Set 5610?

Possible plan of attack (JSONiq):
1. **Access** the JSON representation of LEGO Set 5610
2. **Iterate** over the set’s **pieces** array:
   1. Inside each piece, **lookup** the values for the **quantity** and **weight** keys
   2. **Multiply** quantity and weight
3. **Aggregate** (**sum**) the multiplied weights
Example: USGS Real-Time Earthquake Data

```json
{
  "type":"FeatureCollection",
  "metadata":{
    "generated":1381237557000,
    "url":"http://earthquake.usgs.gov/earthquakes/feed/v1.0/summary/2.5_day.geojson",
    "title":"USGS Magnitude 2.5+ Earthquakes, Past Day",
    "status":200,
    "api":"1.0.11",
    "count":31
  },
  "features":[
    { "type":"Feature",
      "properties":{
        "mag":2.9,
        "place":"38km E of King Salmon, Alaska",
        "time":1381234105000,
        :}
      } ]
}
```
DATA MODEL: NESTED ARRAYS AND DICTIONARIES

- Sample query:
  What was the magnitude of the worst earthquake on the northern hemisphere?

- Possible plan of attack (JSONiq):
  1. **Access** the JSON representation of USGS earthquake data
  2. **Iterate** over the data’s **features** array of quakes:
     - **Filter** quakes to retain only those that affected the northern hemisphere (lookup geometry to check whether latitude > 0)
  3. **Iterate** over the qualifying quakes:
     - **Lookup** **mag** (magnitude) among the quake’s **properties**
  4. **Aggregate** (**max**) the collected magnitudes
- A *slight* variation of the original sample *query*: 
  What was the magnitude and place of the worst earthquake on the northern hemisphere?

- In a nutshell: we need `argmax()` not `max()`

- We require one of many possible different plans of attack here:
  1. Once we computed the maximum magnitude $\text{mag}$, iterate over all quakes again to find those with magnitude $\text{mag}$.
  2. Remember that for $m \in S$ we have
     $$\max(S) = m \iff \{ y \mid y \in S, y > m \} = \emptyset$$
     for any set $S$ of comparable elements.
  3. Order all quakes by descending magnitude, then pick the first in that order.
DATA MODEL: NESTED ARRAYS AND DICTIONARIES

- These plans of attack represent typical query formulation techniques:
  - The use of **quantifiers** (here: \(\forall / \text{empty}()\), but \(\exists\) is as important)
  - The use of **nested iteration** and **correlation**
  - The use of **ordered** containers and **positional lookup**
- We will revisit all of these in this course.

**Query Equivalence?**

Are all of these queries equivalent? Will they return the same earthquake regardless of the current earthquake data?
Replace the role of arrays and dictionaries by ordered trees and obtain XML, another semi-structured data model in wide use today.

- XML defines a textual representation for data trees whose leaves contain strings:

```
<a>
  <b><c>foo</c></b>
  <d>bar</d>
</a>
```

- Data languages for XML (XPath, XQuery) navigate trees (descend to child nodes, collect all nodes in subtree, collect all nodes on path to the root, ...).
DATA MODEL: TABULAR

Types

Container types: tables of rows (or: records, tuples), each row having the same number of fields. Fields contain values of basic atomic types (e.g. numbers, strings, Booleans) only. Inside a row, entries are identified either by name or position.

We essentially obtain a flat, tabular data model comprised of strictly rectangular data grids.

Operations

1. (Nested) iteration over the rows of a table
2. Filter the rows based on given criteria (or: predicates)
3. Access one or more fields of a row
4. Computation over field values (comparison, arithmetics, ...)
DATA MODEL: TABULAR

Thanks to its restrictiveness, the tabular data model can have particularly simple textual representations. Quite common: **CSV (comma-separated values)**.

**CSV** (here: tab-separated values)

1. Table \equiv file, row \equiv line. Rows are separated by \n. Fields in a row are separated by \t (tab character)
2. First line of file contains field names, lines 2, 3,… contain data rows
3. Field value syntax uses usual conventions, strings are enclosed in "\ldots"
4. A missing field value is represented as \N
**DATA MODEL: TABULAR**

Example: USGS Earthquake Data in CSV Format

<table>
<thead>
<tr>
<th>time</th>
<th>latitude</th>
<th>longitude</th>
<th>depth</th>
<th>mag</th>
<th>place</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013-10-08T12:08:25</td>
<td>58.6193</td>
<td>-156.005</td>
<td>85.2</td>
<td>2.9</td>
<td>&quot;38km E of King Salmon, ...&quot;</td>
</tr>
<tr>
<td>2013-10-08T11:45:19</td>
<td>31.6814</td>
<td>131.7314</td>
<td>21.34</td>
<td>4.8</td>
<td>&quot;35km ENE of Nichinan, Japan&quot;</td>
</tr>
<tr>
<td>2013-10-08T11:31:04</td>
<td>12.2779</td>
<td>-88.4332</td>
<td>48.22</td>
<td>4.3</td>
<td>&quot;109km SSW of Chirilagua, ...&quot;</td>
</tr>
<tr>
<td>2013-10-08T10:48:22</td>
<td>36.5863</td>
<td>-98.0361</td>
<td>\N</td>
<td>3.2</td>
<td>&quot;21km ENE of Helena, Oklahoma&quot;</td>
</tr>
</tbody>
</table>

(excerpt only, beautified)

- Note the regular row-wise/column-wise organization of data.
- This particular variant of CSV format is also used by PostgreSQL when relational data is to be exported/imported.
There is no agreed-upon (let alone standardized) data language for CSV. To touch CSV data, in this course we will build our own data language based on Python.

To name the beast, let’s call it Python QL or PyQL for short (pickle /ˈpɪkəl/: messy situation).

- Needed: a Python 2.7 (or Python 3) installation
- A supporting Python module DB1 is available on the course homepage
- All other PyQL operations and constructs are in fact regular Python operations
- Sample PyQL query: access all rows of table *earthquake.csv*. Iterate over table, bind each row to *row variable eq*, print each row:

```python
from DB1 import Table

earthquakes = Table('earthquakes.csv')
for eq in earthquakes:
    print(eq)  # A indentation indicates block structure
```

- Function *Table(⟨f⟩)* reads CSV file *⟨f⟩* and returns a Python iterator. Iteration (e.g. via *for ... in ...*) yields each row in the file in the form of Python dictionary.
In PyQL, use standard Python constructs to access a field in a row or to construct new rows:

```python
from DB1 import Table

earthquakes = Table('earthquakes.csv')

for eq in earthquakes:
    print({ 'place': eq['place'], 'mag': float(eq['mag']) })
```

- Compare to JSONiq’s key-based dictionary lookup and dictionary construction syntax
- Conversion from string (to numeric) is explicit (`float()`, `int()`)

**DATA MODEL: TABULAR**
DATA MODEL: TABULAR

- PyQL syntax:

\[
\langle e \rangle ::= \text{Table(} \langle \text{csv} \rangle \text{)}
\]

\[
\text{for } \langle v \rangle \text{ in } \langle e \rangle : \langle e \rangle
\]

\[
\text{[ } \langle e \rangle \text{ for } \langle v \rangle \text{ in } \langle e \rangle \text{ ... ]}
\]

\[
\text{if } \langle e \rangle : \langle e \rangle \text{ else: } \langle e \rangle
\]

\[
\text{print(} \langle e \rangle, \langle e \rangle, ...)\]

\[
\langle e \rangle[\langle f \rangle]
\]

\[
\{ \langle f \rangle: \langle e \rangle, ..., \langle f \rangle: \langle e \rangle \}
\]

\[
\text{float(} \langle e \rangle \text{)} \mid \text{int(} \langle e \rangle \text{)} \mid ...
\]

\[
\langle e \rangle + \langle e \rangle \mid \langle e \rangle == \langle e \rangle \mid ...
\]

\[
\langle \text{float} \rangle \mid \langle \text{int} \rangle \mid \langle \text{string} \rangle \mid ...
\]

\[
\langle v \rangle
\]

\[
\langle v \rangle = \langle e \rangle
\]

\[
\langle e \rangle \cdots \langle e \rangle
\]

\[
\langle \text{csv} \rangle ::= \text{CSV file name}
\]

\[
\langle f \rangle ::= \text{field name}
\]

\[
\langle v \rangle ::= \text{variable name}
\]
PyQL’s `comprehensions` provide elegant and compact notation for iteration and filtering:

```python
# Magnitude of worst earthquake on the northern hemisphere
from DB1 import Table

earthquakes = Table('earthquakes.csv')

magnitudes = [ float(eq['mag']) for eq in earthquakes if float(eq['latitude']) >= 0.0 ]

print(max(magnitudes))
```

- Compare to set comprehensions as they are common in mathematics:

```latex
\{ f(x) \mid x \in S, p(x) \}
```
Above we have used a **list-oriented style** of query formulation (list comprehensions, \texttt{max()} list aggregate).

- Same query in **imperative style**. Now, the focus is on updating the state of \texttt{float} variable \texttt{mag}:

```python
earthquakes = Table('earthquakes.csv')

mag = 0.0  # ▲ variable initialization

for eq in earthquakes:
    if float(eq['latitude']) >= 0.0:
        if float(eq['mag']) > mag:
            mag = float(eq['mag'])  # ▲ variable update

print(mag)
```
Consider the slight variation of the sample query again:

*What was the magnitude and place of the worst earthquake on the northern hemisphere?*

---

The imperative query style can be convenient. It does, however, effectively allow writing programs that are arbitrarily complex to evaluate (or never terminate at all).

For this (and more good) reason, by design **data languages are considerably more restricted than general programming languages**.
The tabular data model is **flat**: fields contain **atomic** values.

In the absence of nesting, how to represent complex structured information? Recall:

```
{ "set": "5610-1",
  "pieces": [ { "brick": "6157", ..., "quantity": 1 },  # "pieces": non-flat
    { "brick": "3139", ..., "quantity": 2 },
    { "minifig": "cty052", ..., "quantity": 1 } ] }
```

One option: **flatten out** the nested data, attach set identifier **5610-1** to each row:

```
[ { "set": "5610-1", "brick": "6157", ..., "quantity": 1 }  
{ "set": "5610-1", "brick": "3139", ..., "quantity": 2 }  
{ "set": "5610-1", "minifig": "cty052", ..., "quantity": 1 } ]
```
- In this flat model, where to keep the brick/minifigure (or: piece) details?
- Consider the full LEGO set mini-world. Brick 6157 occurs in many sets:

```json
...
{ "set": "5610-1", "brick": "6157", ¶, "quantity": 1 }
...
{ "set": "10048-1", "brick": "6157", ¶, "quantity": 4 }
...
{ "set": "1029-1 ", "brick": "6157", ¶, "quantity": 2 }
...
```

- Keeping details for brick 6157 here (¶) would replicate data — wastes space and comes with the risk that the copies go out of sync over time.
- Such redundancy is almost always to be avoided!
DATA MODEL: TABULAR

- Option adopted by the tabular data model:
  1. Keep brick information in a separate CSV file (i.e., in a separate bricks table)
  2. Use brick identifiers (e.g., 6157) to locate brick details in this new table:

```json
{ "piece": "6157", "type": "B", "name": "Plate, Modified …", "weight": 1.12 }
```

- There is no redundancy. The different LEGO sets share a single copy of the brick details.
- If brick details change, a single row in the bricks table is affected.

**Note:** Brick identifier 6157 must indeed be present as well as unique in the bricks table. This is a typical mini-world rule (or constraint).
DATA MODEL: TABULAR

- Data of the LEGO mini-world is now spread over three tables:

  **contains**
  
<table>
<thead>
<tr>
<th>set</th>
<th>piece</th>
<th>color</th>
<th>extra</th>
<th>quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>c</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

  **bricks**
  
<table>
<thead>
<tr>
<th>piece</th>
<th>type</th>
<th>name</th>
<th>cat</th>
<th>weight</th>
<th>img</th>
<th>x</th>
<th>y</th>
<th>z</th>
</tr>
</thead>
<tbody>
<tr>
<td>b</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

  **minifigs**
  
<table>
<thead>
<tr>
<th>piece</th>
<th>type</th>
<th>name</th>
<th>cat</th>
<th>weight</th>
<th>img</th>
</tr>
</thead>
<tbody>
<tr>
<td>m</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- The two predicates \( c = b \) and \( c = m \) identify related rows in the three tables.
DATA MODEL: TABULAR

- Sample query:
  *What is the overall weight (in grams) of LEGO Set 5610?*

- Plan of attack (recall: a set’s piece is either a brick or a minifig):
  1. **Iterate** over *contains*, **filter** rows for pieces *c* in LEGO Set 5610
     - **Iterate** over *bricks*, **filter** rows for the brick *b* that corresponds to *c*
     - **Multiply** quantity and weight of *b*, **aggregate** (sum up)
  2. **Iterate** over *minifigs*, **filter** rows for the minifig *m* that corresponds to *c*
     - **Multiply** quantity and weight of *m*, **aggregate** (sum up)
  2. Return aggregate (sum) of steps 1.1 and 1.2
# Compute the overall weight of all pieces in LEGO set 5610-1 ("Builder")

```python
from DB1 import Table

contains = Table('contains.csv')
bricks = Table('bricks.csv')
minifigs = Table('minifigs.csv')

weight = 0

for c in contains:
    if c['set'] == '5610-1':
        for b in bricks:
            if c['piece'] == b['piece']:
                weight = weight + int(c['quantity']) * float(b['weight'])
        for m in minifigs:
            if c['piece'] == m['piece']:
                weight = weight + int(c['quantity']) * float(m['weight'])

print(weight)
```

2. Here and in the following, \( \langle \text{var} \rangle \) denotes a *meta-syntactic variable* of name \text{var} (i.e. a placeholder).