SQL Database for Research 101

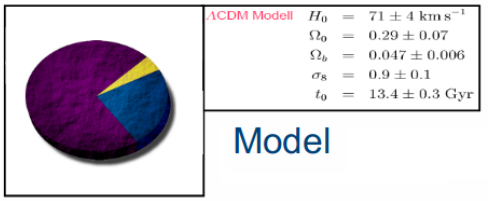
RSE Seminar @ University of Strathclyde

Dr. Anton Buyskikh

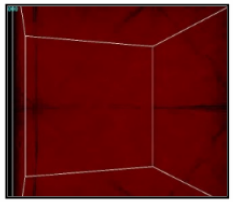
https://github.com/anton-buyskikh/python_sqlite3_workflow_for_rse

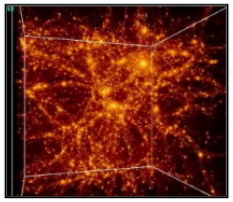
26 June 2019

Motivation: workflow is clear, right?



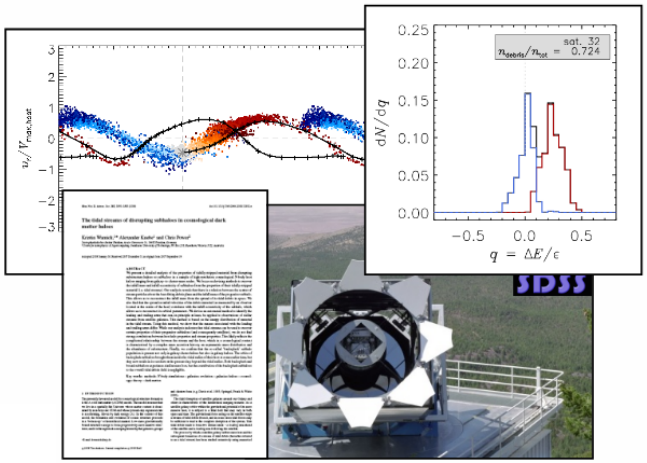
Model

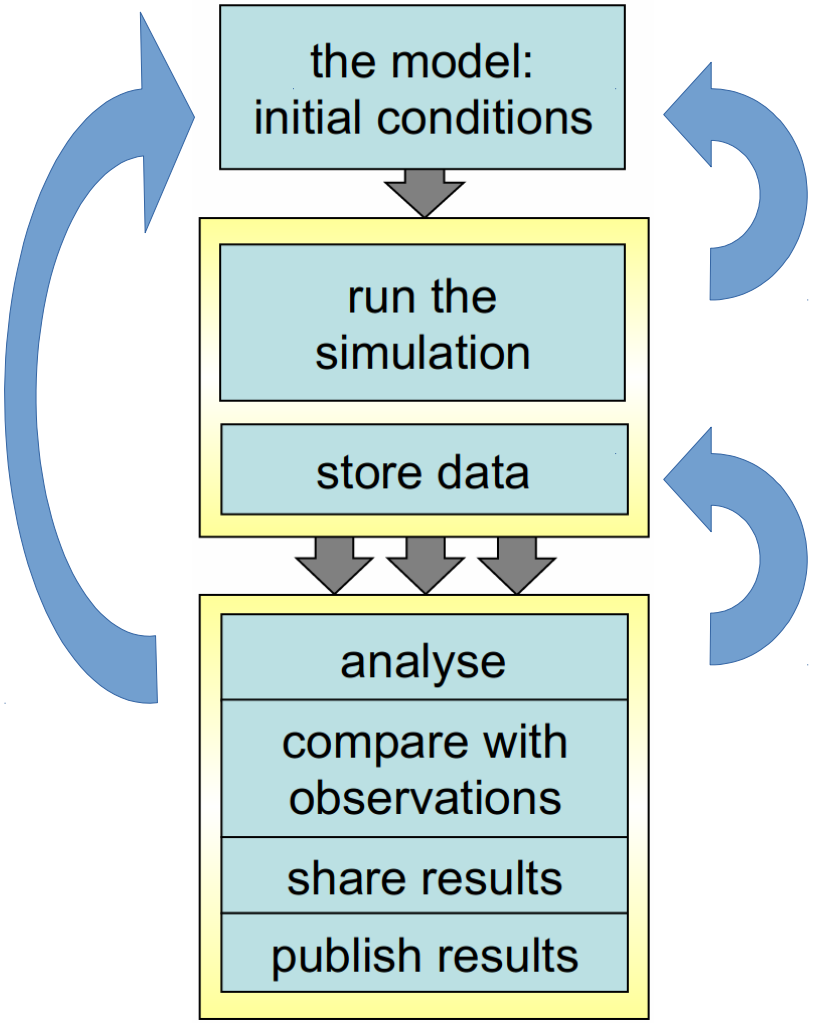




simulation snapshots

computer cluster



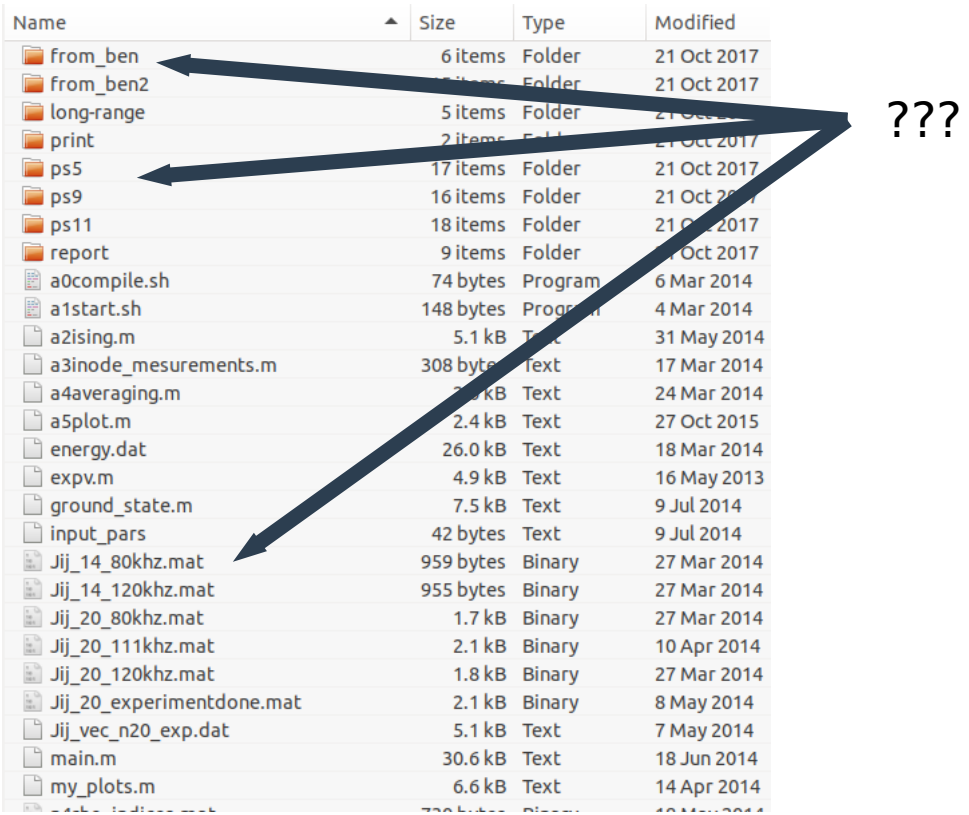


Motivation: organization is important

Project beginning

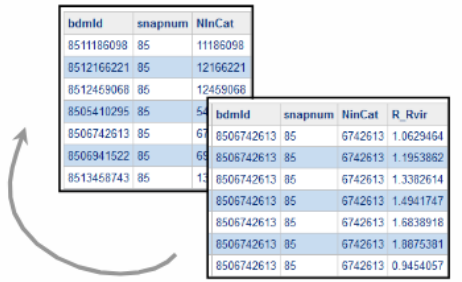
Name	Size	Type	Modified
main.py	10.7 kB	Text	28 Apr

Project end



Simulation Databases can help

- store results of simulations in database, as tables and links between them
- Why?
  - simulations produce TB of data => hard to handle and share
  - post-processing results have variety of formats, individual software for reading
  - visibility of data?
  - reproducability of data?



Just get the subset you need, do (basic) calculations directly on the database server

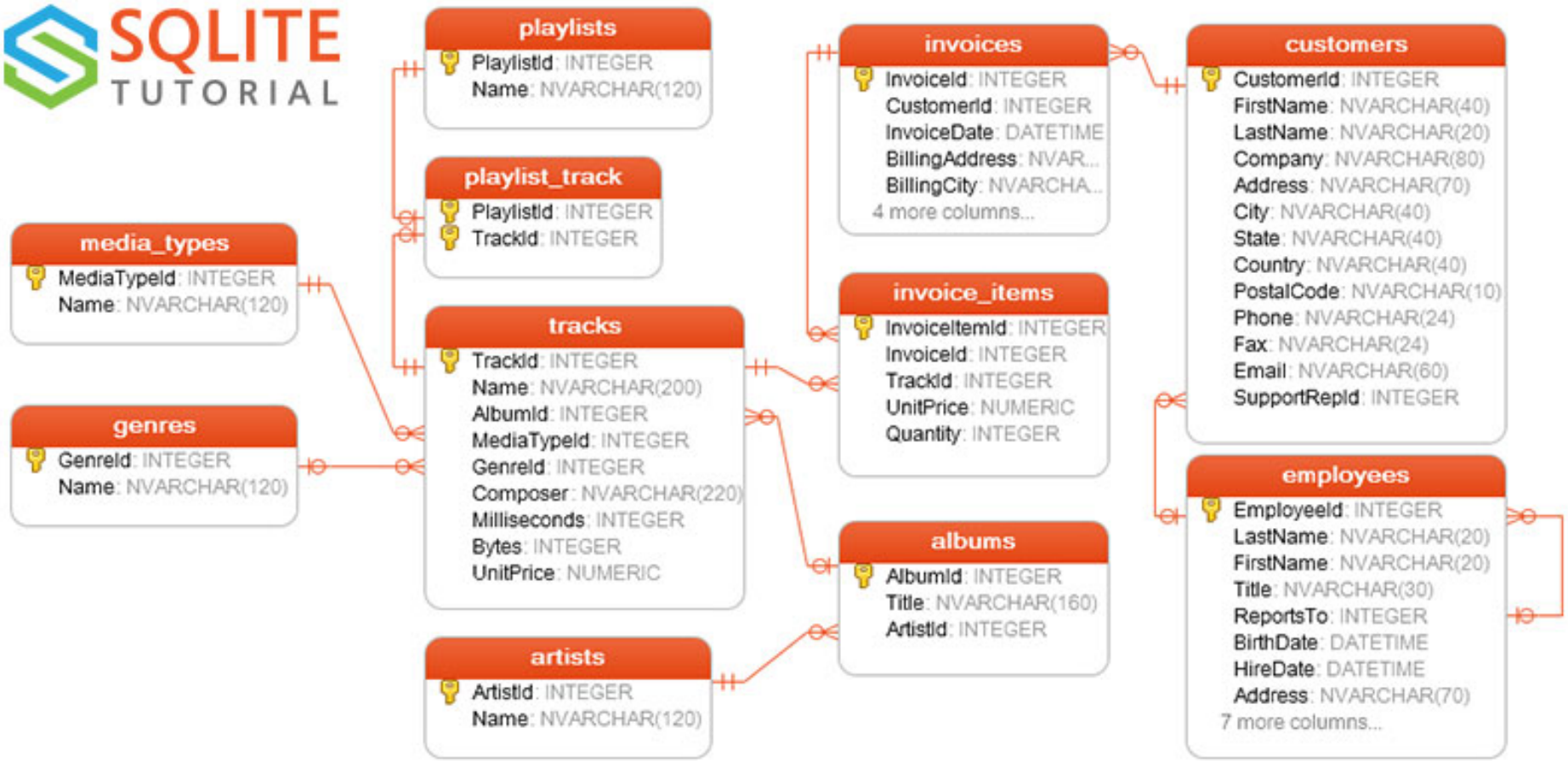
Uniform data format, SQL as standard

```
select top 20 * from MDR1..FOF
where snapnum=85
order by mass desc
```

extracts 20 most massive FOF groups at z=0

< 1 s

Example Database



SQL Language

- **SQL: Structured Query Language**
- **Originally called SEQUEL**
  - **S**tructured **E**nglish **Que**ry **L**anguage
  - Designed/Implemented at IBM Research for an experimental DBMS called **System R**.
- **SQL is now a standard**
  - American National Standards Institute (ANSI)
  - International Standards Organization (ISO)

There are different flavours of SQL, but their syntax is very similar

SQL Language

- **Declarative Language**

A user *only* specifies *what* the result is to be...

The database figures out *how* to retrieve the result!

This allows for greater flexibility in the language
and more opportunity for an SQL compiler
to optimize queries to achieve increased performance!

Create Table

```
CREATE TABLE COMPANY (
 Fname VARCHAR(15) NOT NULL,
 Lname VARCHAR(15) NOT NULL,
 Ssn   CHAR(9) NOT NULL,
 Bdate DATE,
 Dno   INT NOT NULL,
 PRIMARY KEY (Ssn),
 FOREIGN KEY (Dno) REFERENCES DEPT(no)
);
```

Data Types

- **Numeric**
- **Character string**
- **Bit string (BLOB)**
- **Boolean**
- **Time**

- integer
- float
- double
- quad
- ...

Non-standard formats can be saved in binary:

- images
- numpy arrays
- objects
- ...

Retrival Queries

SELECT-FROM-WHERE Structure

```
SELECT   <attribute list>
FROM     <table list>
WHERE    <condition>;
```

```
SELECT   Pnumber, Dnum,
         Lname, Address,
         Bdate
FROM     PROJECT, DEPARTMENT,
         EMPLOYEE
WHERE    Dnum=Dnumber AND
         Mgr_ssn=Ssn AND
         Plocation='Stafford';
```

Research application

- **All** modern languages have **A**pplication **P**rogramming **I**nterface for SQL databases
- Python solutions:
  - sqlite3 – interface SQLite
  - SQLAlchemy – Python SQL toolkit
  - pandas – data analysis toolkit
  - more...
- Let’s look at the practical...