Databases and Big Data

Course Introduction

What Do You Expect To Achieve?

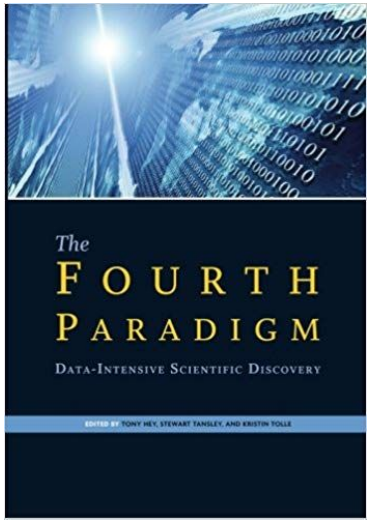
Data Owner: how do I store my data?

Database user: how do I use the data?

Database designer: how do I build a database? (DON’T DO IT!)

Database

- The world is drowning in data
- Changes the way we:
  - Make scientific discoveries
  - Live our lives (for better or for worse)



amazon

NETFLIX

Database

What is a database?

Database is an **organized collection** of **data**

What is a database management system (DBMS):

- System that **manages** the organized collection of data
  - Create, delete, store, query, analyze, etc.

Database & DBMS

Warning!

- I use the 2 terms interchangeably
- That’s me being sloppy
- If not clear from context what I meant, ASK.

Database

In practice, DBMSs rely on filesystems to persist data on disk.

Database is an **organized collection** of **data**

Database example:

- Bank accounts
- Facebook
- Amazon's products
- Experiment data

DBMS is a system that manages the database

NO

Is FILE a database?

Is FILE SYSTEM a DBMS?

Is Python LIST a database?

Is Python a DBMS?

Database

- Let’s store everything on **flat** files
  - Flat = all values are equal. E.g., Comma Separated Values (CSV) format
- Bike share dataset

Problems With File System

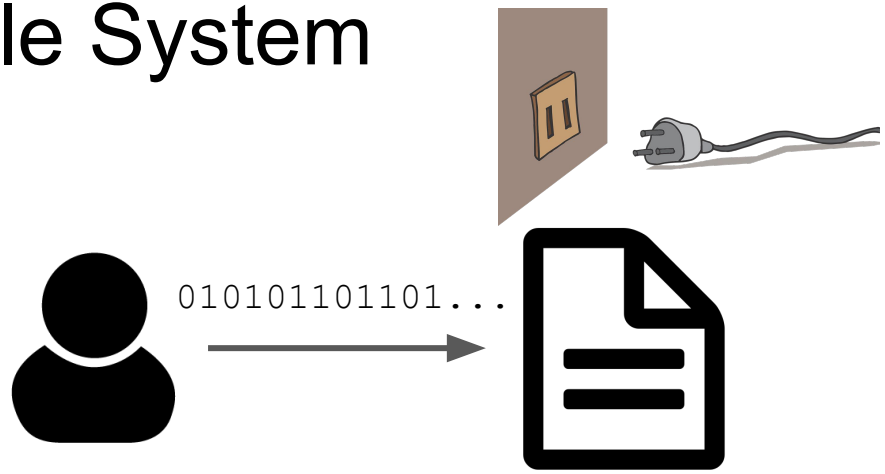
Performance

```
dinhtta@homer:~/Research/istd50043_demo$ time python3 bikeshare.py
742280971

real    0m0.499s
user    0m0.463s
sys     0m0.036s
```

```
mysql> select sum(duration) from trip;
+---------------+
| sum(duration) |
+---------------+
|     742280971 |
+---------------+
1 row in set (0,11 sec)

mysql>
```

Problems With File System

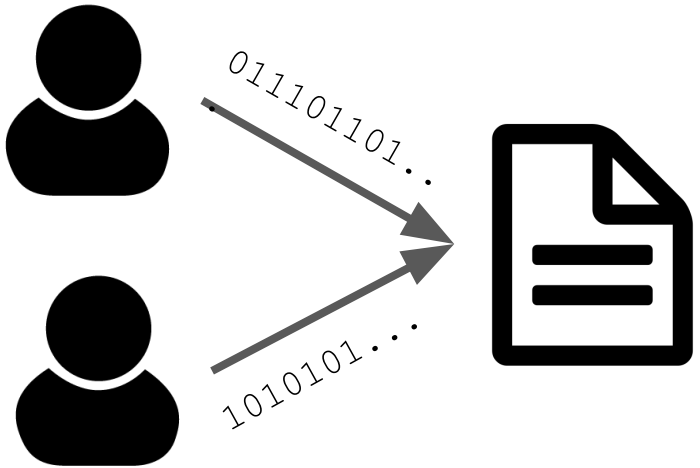
Crash




```
82,Broadway St at Battery St,37.7985,-122.401,15,San Francisco,2014-01-22,94107
83,Mezes Park,37.4913,-122.236,15,Redwood City,2014-02-20,94063
84,Ryland Park,37.3427,-121.896,15,San Jose,2014-04-09,95113
85,Time Squa
```

Problems With File System

Concurrent access




```
84,Ryland Park,37.3427,-121.896,15,San Jose,2014-04-09,95113
85,Ryland Green,37.5427,85,Google Head Quater,35.427,-110.6,17,San Francisco,2014-05-19,95113
 -120.79,15,San Jose,2014-05-011,95113
```

Problems With File System

Easy to use?

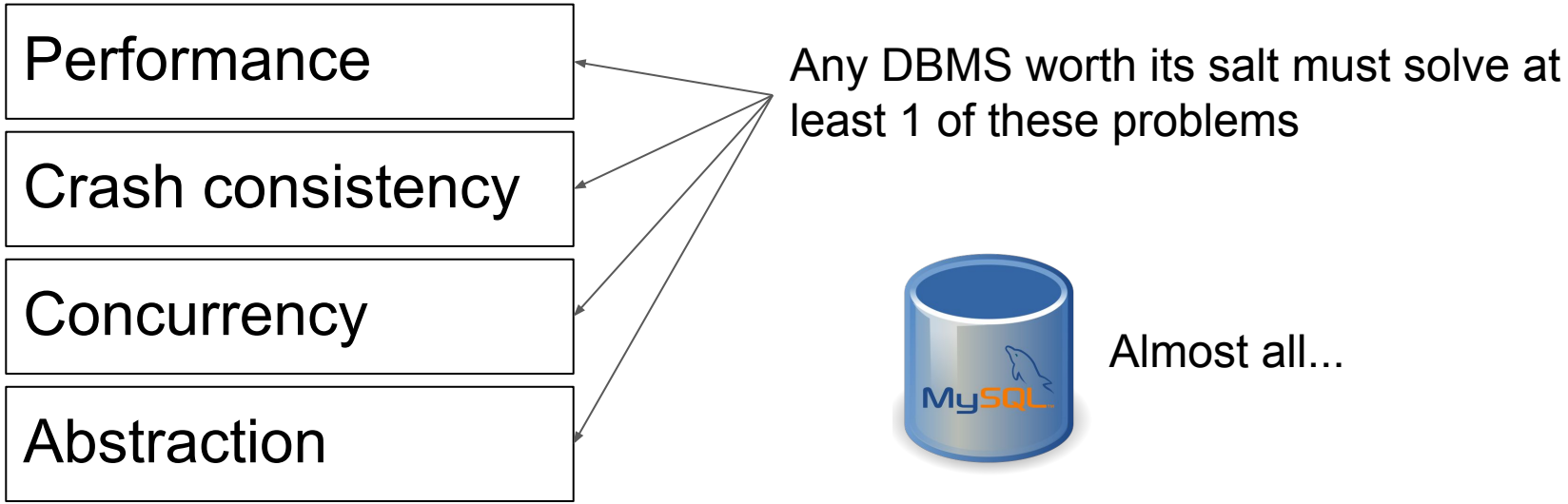
```
def most_popular_bike(lines):
        count = {}
        for l in lines:
                ls = l.strip().split(',')
                if not (ls[8] in count):
                        count[ls[8]] = 0
                count[ls[8]] = count[ls[8]]+1
        m = 0
        bid = 0
        for x in count:
                if count[x] > m:
                        m = count[x]
                        bid = x
        print(bid, m)

def popular_per_day(lines, DATE):
        count = {}
        for l in lines:
                if l.find(DATE) !=-1:
                        ls = l.strip().split(',')
                        if not (ls[8] in count):
                                count[ls[8]] = 0
                        count[ls[8]] = count[ls[8]]+1
        m = 0
        id = 0
        for x in count:
                if count[x] > m:
                        m = count[x]
                        id = x
        print(id, m)

sum_duration(lines)
#most_popular_bike(lines)
popular_per_day(lines, "2013-08-29")
```

Data dependence:
rewrite if format changed

Poor abstraction:

- write ad-hoc code for every query → need to know queries in advance
- cannot enforce integrity constraints:
  - data of right type
  - data that shouldn't have existed (trip with unregistered bikes)

Problems With File Systems



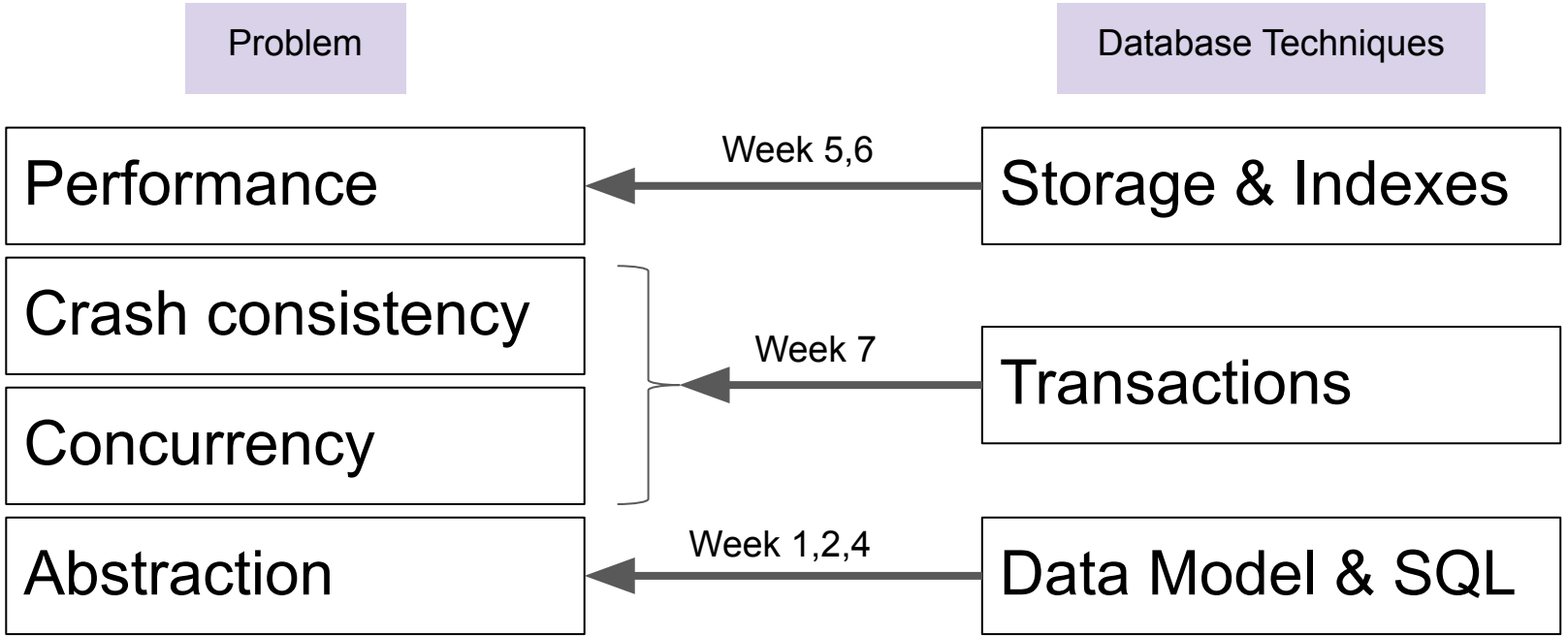
Problems With File Systems

- Performance
- Crash consistency
- Concurrency
- Abstraction

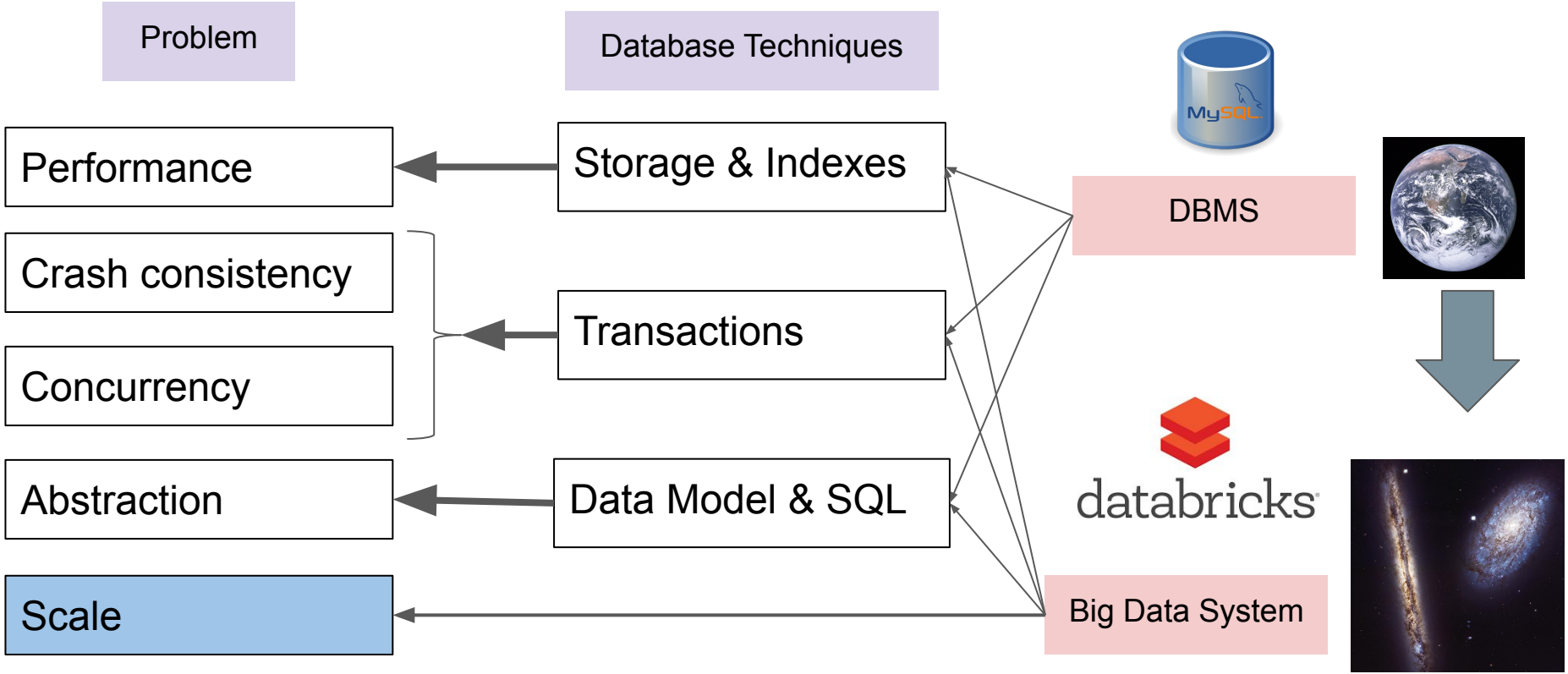
Any DBMS worth its salt must solve at least 1 of these problems

But Prof, why people use file system AT ALL?

How DBMS Solve These Problems



How DBMS Solve These Problems



How Do Users See It?