Basically Everyone Except My Bank

~~Biologists~~

~~Physicists~~

Why ~~Computer Scientists~~ Don't Use Databases

Sam Madden

The Answer

- Benefit(DBMS) < Suckiness(DBMS)
- Disadvantages of DBMSs are growing
  - Impoverished data manipulation language
  - Lack of modern cleaning and modeling tools
- Advantages of a DBMSs are shrinking
  - Large data sets? Transactions? High-level languages?
- DBMS setup & *boundary crossings* painful
  - Especially if you have to do it multiple times!

MATLAB
Python
awk

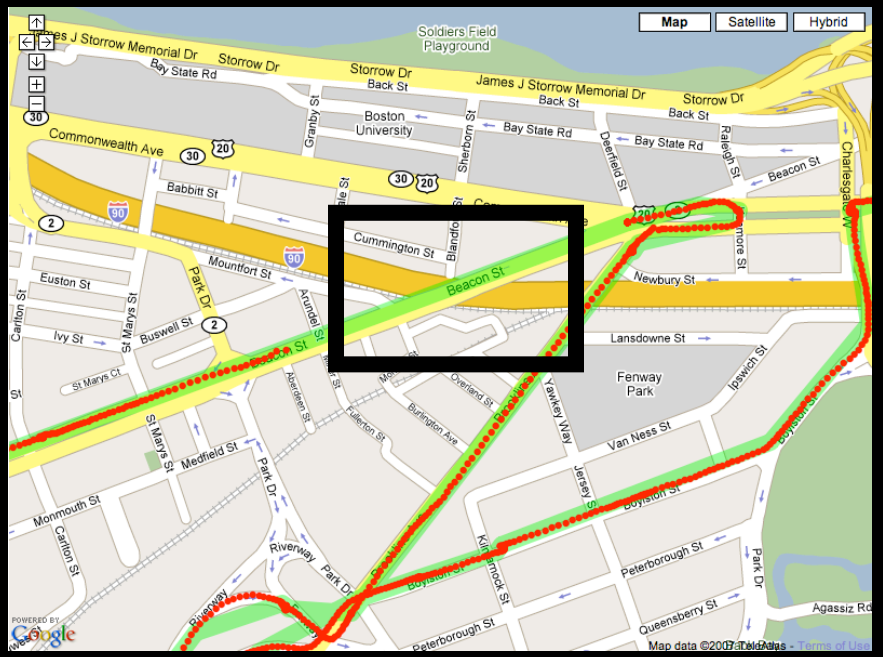
Regression
Graphical models
Interpolation

Example

- CarTel - ~1M GPS points per day from a fleet of 40 cabs on Boston streets

- Pipeline

IMPORT	Raw data in DBMS
EXPORT	Trajectories with Matlab
IMPORT	Queries with SQL
EXPORT	Route Planning with C++
EXPORT	Visualization on Google Maps



- Database isn't the most valuable tool in this picture

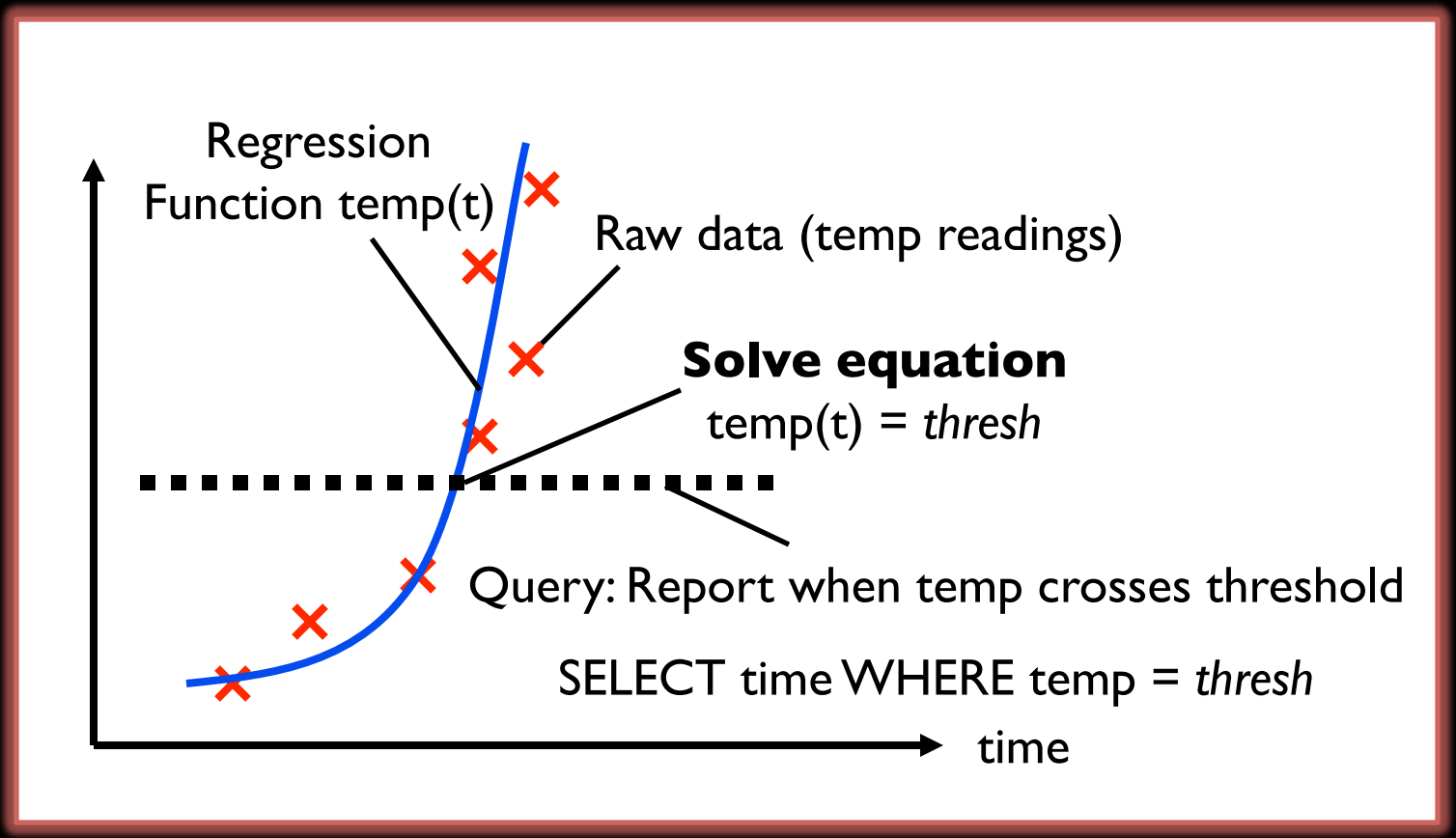
- Import/Export is non-trivial
  - Database is least flexible tool, requires most maintenance

Two Solutions

- Decrease Frequency of Boundary Crossings
  - Cram more stuff into the DBMS
    - FunctionDB
    - Probabilistic databases
    - ArrayDBs
    - XML
    - …

FunctionDB

- DBMS that can fit *continuous functions* to raw data, query data represented by these functions using SQL



Two Solutions

- Decrease Frequency of Boundary Crossings
  - Cram more stuff into the DBMS
    - FunctionDB
    - Probabilistic databases
    - ArrayDBs
    - XML
    - …

```
> checkpoint data.txt
> awk -f process.awk data.txt > data.txt
> history data.txt
   awk process.awk
> rollback data.txt
```

- Decrease Pain of Boundary Crossings
  - Don't insist on DBMS-specified storage representation
    - Text and fileystem based tools to manage, edit, manipulate data
  - Don't insist on SQL
  - Don't insist on structured data
    - Add transactions, rollback, lineage to Unix toolchain and filesystems

If you can't beat them, join them