Where do data come from?

CS100 - Guest Lecture - Databases and Provenance

Boris Glavic[1]

DBGroup

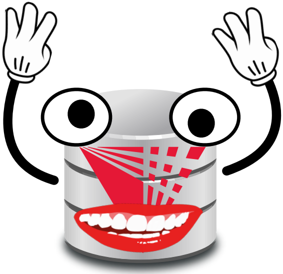
Illinois Institute of Technology

October 25, 2019

[1] bglavic@iit.edu

Outline

1. Who I am
2. What are Databases?
3. Data Provenance
4. Questions



Who I am

Hi, I am **Boris**

Who I am

Hi, I am **Boris**

I am a **database** guy!

Who I am

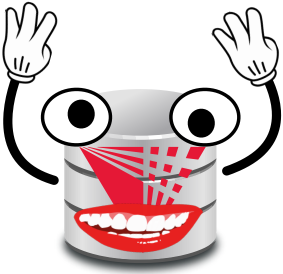
Hi, I am **Boris**

I am a **database** guy!

I will tell you:
1) Why DBs are important
2) Why DBs are interesting
3) My Research

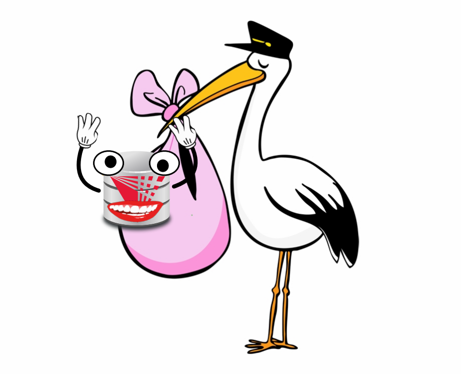
Outline

1. Who I am
2. What are Databases?
3. Data Provenance
4. Questions



Where do data come from?

You might have heard ...



What Are Databases?

Database systems and databases

- **database systems** manage databases
- a **database** is a structured collection of data

What do database systems do?

1. Provide **persistent** storage of data
2. Efficient **declarative** access to data: querying
3. **Protection from data loss** under failures
4. **Safe concurrent access** to data

Who Uses Databases?

- **Most large software systems use databases!**
  - Business Intelligence, e.g., *IBM cognos*
  - Web-based systems
- **Desktop software**
  - You music player
  - You email client (most likely at least)
- **Every big company uses DBs**
  - banks
  - insurance
  - government agencies
  - . . .

Who Creates Databases?

- **Relational databases is big business**
  - IBM DB2
  - Oracle
  - Microsoft SQLServer
  - Teradata
  - Open Source Systems: PostgreSQL, MySQL
- **Distributed systems**
  - Cloud storage and Key-value stores
    - Amazon S3, Google Big Table, Cassandra
  - Big Data Analytics
    - MapReduce, Spark, Flink

Why Are Database Interesting?

Combination of systems and theoretical research

- **Interesting systems problems**
  - Hacking complex and large systems
  - Low-level optimizations
    - exploit modern hardware
- **Interesting theoretical foundations**
  - Complexity of answering queries
  - Expressiveness of query languages
  - Strong connections to logic

Why Are Databases Interesting (cont.)

Connections to other CS fields

- Distributed systems
  - getting more and more important
- Compilers
- Modeling
- AI and machine learning
  - Data mining
- Operating and File Systems

The Relational Datamodel

Relations aka Tables

- a table consists of **columns** and **rows**
- tables store one type of entity
  - *e.g., students, bank accounts, loans, . . .*
- each row is one entity
  - *e.g., one student*
- columns store a particular type of information about an entity
  - *e.g., name of a student*

The Relational Datamodel (cont.)

Example Tables

Students table

CWID	Name	Major	GPA	Phone
A1333331	Peter	CS	3.5	312 555 8888
A5552341	Alice‘	CS	4.0	312 555 7777
A1325324	Elisa	Bio	3.2	312 555 5555

Grades table

CWID	Course	Grade
A1333331	CS100	A
A5552341	CS425	C
A1325324	CS525	A
A1325324	CS566	B

Queries

What do I do with the data in my database?

- you can interrogate the database system to extract information about your data
- this is done using a programming language called SQL
- SQL is a **declarative** language
    - say what data you want not how to compute it
- Queries return table (closed language)

Queries (cont.)

CWID	Name	Major	GPA	Phone
A1333331	Peter	CS	3.5	312 555 8888
A5552341	Alice‘	CS	4.0	312 555 7777
A1325324	Elisa	Bio	3.2	312 555 5555

- *How many students are in my database?*

#Students
3

- *Who has the highest GPA?*

Name
Alice

- *What are the names of CS students?*

Name
Peter
Alice

Persistence and Recovery

What if you shutdown your computer?

- will you loose your precious data?

What happens when your computer crashes?

- will you loose your precious data?

Persistence and Recovery

What if you shutdown your computer?

- will you loose your precious data?

What happens when your computer crashes?

- will you loose your precious data?

No!

- the database system stores your data on stable storage (disk)
- database systems know how recover from failures
- when the database system signals to you that a change you made was applied then you will never loose it

Concurrent Access

Banking Example

- **Account A**: $50
- **Account B**: $50
- Transfer $25 from A to B
- Bank gives all accounts 10% interest

Transfer Money

Action
Subtract $25 from **A**
Add $25 to **B**

Give 10% interest

Action
Add %5 interest

Balances

Account A	**Account B**
$25	$50
$27.5	$55
$27.5	$80

We have lost interest!

Concurrent Access (cont.)

Concurrency Control

- databases manage concurrent operations
- prevent bad things from happening
- from user perspective:
  - behaves like your program is the only one running!

Can we loose interest?

Nope!

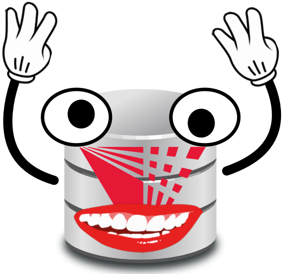
Outline

1. Who I am

2. What are Databases?

3. **Data Provenance**

4. Questions



What is Provenance?

Provenance in Art

- record of ownership of a piece of art

Arnolfini Portrait

The provenance of the painting begins in 1434 when it was dated by van Eyck and presumably owned by the sitter(s). At some point before 1516 it came into the possession of Don Diego de Guevara (d. Brussels 1520), a Spanish career courtier of the Habsburgs . . .
By 1516 he had given the portrait to Margaret of Austria, . . .

What is Provenance?

Provenance in Databases

- Records how data was produced
  - which other data was used in the creation process
  - which operations were involved in its creation
- For sake of this lecture, only provenance of queries

Provenance of a query result

- Select one row from the result of a query
- Which input rows were used to compute it?
- Maybe also: how were these rows combined

Example

Compute average salary of employees per department

```
SELECT dept, avg(salary) AS avgsal
FROM emp
GROUP BY dept
```

name	salary	dept
Peter	10	HR
Bob	20	HR
Alice	5	IT

Example

Compute average salary of employees per department

dept	avgsal
HR	15
IT	5

```
SELECT dept, avg(salary) AS avgsal
FROM emp
GROUP BY dept
```

name	salary	dept
Peter	10	HR
Bob	20	HR
Alice	5	IT

Example (cont.)

The first result row depends on the first two input rows

dept	avgsal
HR	15
IT	5

```
SELECT dept, avg(salary) AS avgsal
FROM emp
GROUP BY dept
```

name	salary	dept
Peter	10	HR
Bob	20	HR
Alice	5	IT

Some Observations

Provenance maps output rows of queries to input rows

- here track this per row
- could also track attribute values (higher fidelity)
- could also track tables (lower fidelity)

Why should I care?

Use cases

- Debugging queries and data
- Auditing
- Explainability
- Optimizing DB operations
- Determining trust in data

Queries are Functions

Functional View of Querying

- A query takes as **input** a database and **outputs** a table
- We can think about queries as functions from databases to result tables!

What then is provenance?

- Select one of the outputs of the query
- Which inputs were used to compute it?

Recap Functions

What are functions in math?

- you already know functions from high school math!

Recap Functions

What are functions in math?

- you already know functions from high school math!

Examples

- $f(x) = x$
    - $f(1) = 1$

Recap Functions

ILLINOIS INSTITUTE OF TECHNOLOGY

What are functions in math?

- you already know functions from high school math!

Examples

- $f(x) = x$
  - $f(1) = 1$
  - $f(2) = 2$

Recap Functions

What are functions in math?

- you already know functions from high school math!

Examples

- $f(x) = x$
  - $f(1) = 1$
  - $f(2) = 2$
  - $\ldots$

Recap Functions

What are functions in math?

- you already know functions from high school math!

Examples

- $f(x) = x$
  - $f(1) = 1$
  - $f(2) = 2$
  - $\ldots$
- $f(x) = x^2$

Recap Functions

ILLINOIS INSTITUTE OF TECHNOLOGY

What are functions in math?

- you already know functions from high school math!

Examples

- $f(x) = x$
  - $f(1) = 1$
  - $f(2) = 2$
  - $\dots$
- $f(x) = x^2$
  - $f(1) = 1$

Recap Functions

What are functions in math?

- you already know functions from high school math!

Examples

- $f(x) = x$
  - $f(1) = 1$
  - $f(2) = 2$
  - $\ldots$
- $f(x) = x^2$
  - $f(1) = 1$
  - $f(2) = 4$

Recap Functions

What are functions in math?

- you already know functions from high school math!

Examples

- $f(x) = x$
  - $f(1) = 1$
  - $f(2) = 2$
  - $\ldots$
- $f(x) = x^2$
  - $f(1) = 1$
  - $f(2) = 4$
  - $\ldots$

Recap Functions

What are functions in math?

- you already know functions from high school math!

Examples

- $f(x) = x$
  - $f(1) = 1$
  - $f(2) = 2$
  - $\ldots$
- $f(x) = x^2$
  - $f(1) = 1$
  - $f(2) = 4$
  - $\ldots$
- $f(x, y) = x + y$

Recap Functions

What are functions in math?

- you already know functions from high school math!

Examples

- $f(x) = x$
  - $f(1) = 1$
  - $f(2) = 2$
  - $\ldots$
- $f(x) = x^2$
  - $f(1) = 1$
  - $f(2) = 4$
  - $\ldots$
- $f(x, y) = x + y$
  - $f(1, 2) = 1 + 2 = 3$

Recap Functions

What are functions in math?

- you already know functions from high school math!

Examples

- $f(x) = x$
  - $f(1) = 1$
  - $f(2) = 2$
  - $\dots$
- $f(x) = x^2$
  - $f(1) = 1$
  - $f(2) = 4$
  - $\dots$
- $f(x, y) = x + y$
  - $f(1, 2) = 1 + 2 = 3$
  - $\dots$

Recap Functions (cont.)

What makes a function a function?

- Does it have to return numbers?
- Does have to take numbers as input?

Counterexamples

- f takes names as an input and returns the name converted to lower case
    - $f(Peter) = peter$
    - $f(Bob) = bob$
- f takes text as input and returns the numbers of characters in the text
    - $f(Bob) = 3$
    - $f(Alice) = 5$

Recap Functions (cont.)

Definition (Function)

- **Input domain:** A set of values $\mathcal{I}$
- **Output domain:** A set of value $\mathcal{O}$
- **Mapping**: Associate each value from $\mathcal{I}$ with **one** value from $\mathcal{O}$

Queries as Functions

- **Input domain:** databases
- **Output domains**: tables

What is the Provenance of a Function?

Provenance = Inversion?

- We want to understand which input was used to generate an output
- In math this is called function inversion
- The **inverse** f^{-1} of a function f takes an output of f and returns the corresponding input
  - When $f(x) = y$ then $f^{-1}(y) = x$

Examples

What is the Provenance of a Function?

Provenance = Inversion?

- We want to understand which input was used to generate an output
- In math this is called function inversion
- The **inverse** f^{-1} of a function f takes an output of f and returns the corresponding input
    - When $f(x) = y$ then $f^{-1}(y) = x$

Examples

- if $f(x) = 2x$ then $f^{-1}(x) = 0.5x$

What is the Provenance of a Function?

Provenance = Inversion?

- We want to understand which input was used to generate an output
- In math this is called function inversion
- The **inverse** f^{-1} of a function f takes an output of f and returns the corresponding input
  - When $f(x) = y$ then $f^{-1}(y) = x$

Examples

- if $f(x) = 2x$ then $f^{-1}(x) = 0.5x$

What is the Provenance of a Function?

Provenance = Inversion?

- We want to understand which input was used to generate an output
- In math this is called function inversion
- The **inverse** f^{-1} of a function f takes an output of f and returns the corresponding input
    - When $f(x) = y$ then $f^{-1}(y) = x$

Examples

- if $f(x) = 2x$ then $f^{-1}(x) = 0.5x$
- if $f(x) = x^3$ then $f^{-1}(x) = \sqrt[3]{x}$

What is the Provenance of a Function?

Provenance = Inversion?

- We want to understand which input was used to generate an output
- In math this is called function inversion
- The **inverse** f^{-1} of a function f takes an output of f and returns the corresponding input
  - When $f(x) = y$ then $f^{-1}(y) = x$

Examples

- if $f(x) = 2x$ then $f^{-1}(x) = 0.5x$
- if $f(x) = x^3$ then $f^{-1}(x) = \sqrt[3]{x}$

What is the Provenance of a Function?

Provenance = Inversion?

- We want to understand which input was used to generate an output
- In math this is called function inversion
- The **inverse** f^{-1} of a function f takes an output of f and returns the corresponding input
  - When $f(x) = y$ then $f^{-1}(y) = x$

Examples

- if $f(x) = 2x$ then $f^{-1}(x) = 0.5x$
- if $f(x) = x^3$ then $f^{-1}(x) = \sqrt[3]{x}$
- if $f(x) = x^2$ then $f^{-1}(x) = \sqrt[2]{x}$?

What is the Provenance of a Function?

Provenance = Inversion?

- We want to understand which input was used to generate an output
- In math this is called function inversion
- The **inverse** f^{-1} of a function f takes an output of f and returns the corresponding input
  - When $f(x) = y$ then $f^{-1}(y) = x$

Examples

- if $f(x) = 2x$ then $f^{-1}(x) = 0.5x$
- if $f(x) = x^3$ then $f^{-1}(x) = \sqrt[3]{x}$
- if $f(x) = x^2$ then $f^{-1}(x) = \sqrt[2]{x}$?

What is the Provenance of a Function?

Provenance = Inversion?

- We want to understand which input was used to generate an output
- In math this is called function inversion
- The **inverse** f^{-1} of a function f takes an output of f and returns the corresponding input
  - When $f(x) = y$ then $f^{-1}(y) = x$

Examples

- if $f(x) = 2x$ then $f^{-1}(x) = 0.5x$
- if $f(x) = x^3$ then $f^{-1}(x) = \sqrt[3]{x}$
- if $f(x) = x^2$ then $f^{-1}(x) = \sqrt[2]{x}$?
  - this does not work (two possible solutions)

From Invertible Functions to Queries

ILLINOIS INSTITUTE OF TECHNOLOGY

Queries are typically not invertible!

- Return the number of students in CS100
- Let's say the result is 3 students
- Inverse function would have to magically guess who these 3 students are!

Queries operate on tables

- We want more fine-granular information:
    - Which rows from the input affected which rows from the output!

Annotation Propagation to the Rescue

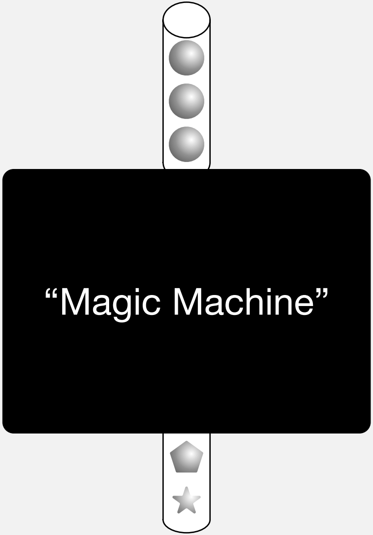
ILLINOIS INSTITUTE OF TECHNOLOGY

Quadratic function

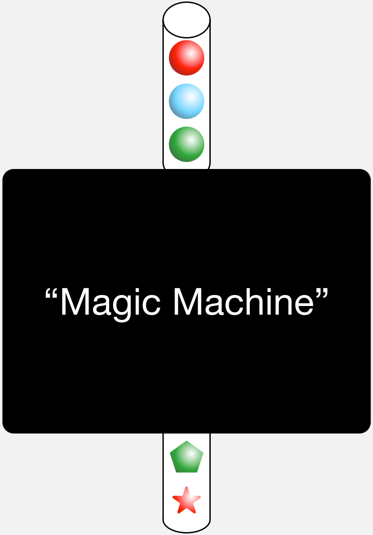
Input	Output
-2	4
-1	1
0	0
1	1
2	4

- Cannot invert this
- The output is not enough to compute provenance!
- How can we deal with that?

Magic Machine Riddle



Magic Machine Solution



Generalizing the solution

Approach

- annotate input data with unique identifiers (colors)
- outputs annotated with the color of the input they are derived from

Quadratic function

Input	Output
(-2,⬛)	(4,⬛)
(-1,🟥)	(1,🟥)
(0,🟪)	(0,🟪)
(1,🟩)	(1,🟩)
(2,🟦)	(4,🟦)

Caveats

- We assumed that the function happily accepts inputs that are pairs of numbers and colors
- If inputs and outputs are tables then we need to understand the internals of the function to know how they are related

Provenance for Queries (how it is actually done)

Encode annotations by extending tables

- each row is extended with extra attributes
- these attributes are used to store provenance

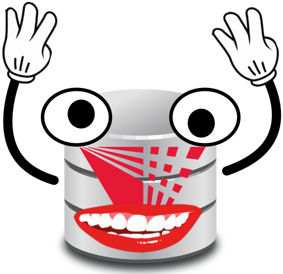
Query instrumentation

- We rewrite queries input queries into queries that
  1. create annotations for each input
  2. propagate these annotations to produced annotated outputs

Outline

ILLINOIS INSTITUTE OF TECHNOLOGY

1. Who I am
2. What are Databases?
3. Data Provenance
4. Questions



DBGroup - What do we do?

Distributed and High-performance Databases

- HRDBMS - a scalable database with high per-node performance
- **HCDF** - operating system - database co-design

Data Integration and Cleaning

- how to systematically evaluate cleaning and integration systems
  - Bart
  - iBench

Data Provenance

- GProM - a generic provenance middleware
- Relevance-based Data Management - optimizing data operations based on what data is relevant

DBGroup - What do we do? (cont.)

Data Science

- We are data science enablers!
- Vizier - a data-centric notebook platform with uncertainty tracking

Questions?

- IIT DBGroup
  - **students**: 7 Ph.D., 1 Master, 1 Undergraduate
  - **research group**: `http://www.cs.iit.edu/~dbgroup/`
  - **personal page**: `http://www.cs.iit.edu/~dbgroup/members/bglavic.html`
  - **github**: `https://github.com/IITDBGroup`