# Responsible Data Science - Dealing with Uncertainty
## DBGroup

Boris Glavic[1]

**DBGroup**
*Illinois Institute of Technology*



December 3, 2021



---

[1]bglavic@iit.edu

ILLINOIS INSTITUTE
OF TECHNOLOGY

Hi, I am **Boris**

ILLINOIS INSTITUTE
OF TECHNOLOGY

I am a **database** guy!

Hi, I am **Boris**

ILLINOIS INSTITUTE
OF TECHNOLOGY

I am a **database** guy!

Hi, I am **Boris**

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

# Outline

# What Are Databases?

## Database systems and databases

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

- **Most large software systems use databases!**
  - Business Intelligence, e.g., *IBM cognos*
  - Web-based systems
- **Desktop software**
  - Your music player
  - Your email client (most likely at least)
- **Every big company uses DBs**
  - Banks
  - Insurance
  - Government agencies
  - . . .
- **Your mobile phone**
  - Many apps use an embedded database called Sqlite

- **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

## Connections to other CS fields

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

## 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*

## 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 |

## 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 tables (closed language)

| 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 |

- *What are the names of CS students?*

| Name |
|------|
| Peter |
| Alice |

- *Who has the highest GPA?*

| Name |
|------|
| Alice |

## 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 can rely on this

## Banking Example

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

| Transfer Money | Give 10% interest | Balances | |
|---|---|---|---|
| **Action** | **Action** | **Acc. A** | **Acc. B** |
| Subtract $25 from **A** | | $25 | $50 |
| | Add 10% interest | $27.5 | $55 |
| Add $25 to **B** | | $27.5 | $80 |

We have lost interest!

## Concurrency Control

- Databases manage concurrent operations
- Prevent bad things from happening
- From the user's perspective:
  - Behaves like your program is the only one running!

## Can we loose interest?

Nope!

# The Data-Driven World

## The Data Age

- We (humanity) are generating data at an ever increasing rate
- Data has become a main driver for many businesses, governments, scientific disciplines, organizations

## Some examples

- Online encyclopedias
- Self-driving cars
- Computers beating humans in go
- Open data (e.g., https://data.cityofchicago.org/)
- IoT
- ...

ILLINOIS INSTITUTE
OF TECHNOLOGY

- **Data Science** is the process of extracting insights from data and includes ...
  - **Data Collection**: finding relevant data for the analysis task
  - **Data Preparation/Curation**: cleaning and integrating the data
  - **Data Analysis**: analyzing the data, e.g., building machine learning models
  - **Interpretation + Presentation**: creating visualizations / documents for conveying the results to a consumer

## Is this linear process realistic?

- **No!** - typically requires backtracking & iteration until the results are sufficient

## Example

- The analysis result is wrong / misleading because the dataset was too small to yield statistically significant results
- Workaround: collect more data, augment existing data with synthetically generated data, . . .
- This new data needs to cleaned be integrated with the existing data
- No we have duplicate and conflicting information
- Ok, need more cleaning
- Repeat analysis (fingers crossed)
- . . .

# Computational Notebooks

## What are computational notebooks?

- a mix of documentation, computation, and results
- consist of cells:
    - documentation cells (typically a lightweight markup language like markdown)
    - code blocks
- results of execution of code is shown inline

## Demo (Jupyter)

- `https: //www.kaggle.com/pouyaaskari/avocado-classification`

# Uncertainty in Data Science

## Almost all data is uncertain!

- missing values
- typos and manual entry errors
- logical errors (zip code with incorrect city)
- misinterpretation of semantics (*e.g., date is a contract start date instead of end data*)

## Data curation is heuristic

- typically insufficient information is available to determine how to correctly clean and integrate the data
- curation decisions are based on informed guesses (made by humans/code)

- Uncertainty is everywhere
- Traditional data cleaning methods are heuristic
- Unless a human tracks uncertainty, after cleaning we loose all information about uncertainty
- $\Rightarrow$ We do not know whether a result is based on real data or just an artifact of cleaning / errors in the data

# What can be done?

### Support for modeling uncertainty
- We need to enable humans and algorithms to model the uncertainty in their data and decisions

### Support for tracking uncertainty
- We need to track information about uncertainty through curation / analysis / visualization steps

### Support for visualizing uncertainty
- We need to present uncertainty information to the data consumer

### Demo (Vizier)
- https://vizierdb.info/

ILLINOIS INSTITUTE
OF TECHNOLOGY

## 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

## Uncertain Data Management

- How to model and track uncertainty in data?
  - Uncertainty-Annotated Databases

## Data Science

- We are data science enablers!
- Vizier - a data-centric notebook platform with uncertainty tracking, provenance, spreadsheets, reproducibility

# Questions?

- **IIT DBGroup**
  - **students**: 7 Ph.D., 2 Undergraduates
  - **research group**: http://www.cs.iit.edu/~dbgroup/
  - **personal page**:
    http://www.cs.iit.edu/~dbgroup/members/bglavic.html
  - **github**: https://github.com/IITDBGroup