ECE 356 • Crime Statistics Database Project

```
SELECT
    Student,
    Course,
    Term,
    Institution
FROM WaterlooStudentsView
ORDER BY Student ASC;
```

Student	*Course*	*Term*	*Institution*
Kyle Pinto	ECE 356 - Database Systems	Fall 2021	University of Waterloo
Puranjoy Saha	ECE 356 - Database Systems	Fall 2021	University of Waterloo
Zahin Zaman	ECE 356 - Database Systems	Fall 2021	University of Waterloo

Table of Contents

Introduction
- Overview
- Datasets
  - London Police Records
  - NYPD Complaint Data Historic
  - Crimes in Chicago
  - LA Crimes

Design
- Analysis of Datasets
- Design Options
  - UK & US Data Separation Option
  - Crimes, Complaints & Stop-and-Searches Option
  - Crimes Only Option
  - Decision Matrix
- Entity-Relationship Model
  - Location
  - Code
  - Person
  - Incident
  - Complaint
  - Crime

- Search
- Relational Schema
- **Application**
  - Installation
  - Database Setup
  - Operations
- **Data Mining**

Introduction

Overview

This project involves the collection of crime records datasets from law enforcement departments in UK and US, and the process of developing an optimally designed database and a client interface for the definition, manipulation and storage of this data.

Datasets

Datasets used in this project have been collected from **Kaggle**. Each dataset provides crime data from a different city and state in UK or US.

Name	Location	Link
London Police Records	London, England, UK	[↗]
NYPD Complaint Data Historic	New York City, New York, US	[↗]
Crimes in Chicago	Chicago, Illinois, US	[↗]
LA Crime Data	Los Angeles, California, US	[↗]

London Police Records

This dataset includes crime data from London from late 2014 to mid 2017, held in the following three CSV files:

- `london-outcomes.csv`
- `london-street.csv`
- `london-stop-and-search.csv`

`london-outcomes.csv` and `london-street.csv` hold data on instances of crime committed in London and their relevant information. `london-stop-and-search.csv` holds data on "stop-and-searches" conducted by London police and their relevant information.

The data references the location of each incident by the Lower Layer Super Output Area (LSOA) code of the neighborhood, which can be mapped to specific area names using the **Lookup Table of UK Local Government Areas** dataset.

NYPD Complaint Data Historic

This dataset includes records of complaints of incidents reported to the New York City Police Department (NYPD) from 2006 to the end of 2017, in CSV file `NYPD_Complaint_Data_Historic.csv`. The data contained in this file also includes the NYPD crime code corresponding to each complaint, which is unique to a specific type of crime.

Crimes in Chicago

This dataset includes crimes reported and committed from the records of the Chicago Police Department (CPD) between 2001 and 2017, divided into four CSV files:

- `Chicago_Crimes_2001_to_2004.csv`
- `Chicago_Crimes_2005_to_2007.csv`
- `Chicago_Crimes_2008_to_2011.csv`
- `Chicago_Crimes_2012_to_2017.csv`

The columns of this dataset includes the Illinois Uniform Crime Reporting (IUCR) code corresponding to the committed crime, which can be extracted from the **IUCR** dataset hosted on the City of Chicago website.

LA Crimes

This dataset includes crimes reported and committed from the records of the Los Angeles Police Department (LAPD) between 2010 and mid-2021, divided into two CSV files:

- `Crime_Data_from_2010_to_2019.csv`
- `Crime_Data_from_2020_to_Present.csv`

The columns of this dataset includes the LAPD crime code corresponding to the committed crime, which is based on the FBI Uniform Crime Reporting codes. These codes can be compiled from the **FBI UCR Handbook**.

Design

Analysis of Datasets

The first step of the design process is a thorough investigation of the given datasets and their attributes in order to prepare practical options for merging the datasets. By examining each dataset, we discover the following:

- Every dataset entry has a location associated with it. Usually that location is specified by latitude and longitude coordinates, along with a few other address parameters that depend on the area, such as ward, precinct, LSOA code, borough, city, state, country, etc. These address parameters, however, vary significantly between UK and US datasets.
- Every dataset entry also has a few attributes that are common between all or most of the datasets. These attributes include date of occurrence, type of crime, description of crime, and other similar attributes that describe a general incident.
- Every US dataset entry has a unique crime code defined that describes the category of the crime. The uniqueness of these codes also depend on the organization that reports this crime data. For instance, NYPD and IUCR crime codes are not identical.
- Some datasets include victim information, and some do not. Because information related to individual people may lead to privacy issues, the datasets omit personal information such as names, contact

numbers etc. and only store their ages (or age ranges), genders and ethnicities.
- The `london-stop-and-searches.csv` file from the London Police Records dataset contains information that is slightly different from the rest of the datasets. It includes information about stop-and-searches conducted by London police, which may or may not have resulted in the discovery of criminal activity. While there are common attributes between this and other datasets, such as location and date, the context of this information is different.

Design Options

From the analysis of the datasets, the most obvious design options we can draw are the definitions of separate entities for crime codes and for individual people.

Each crime code entry should be unique depending on the code and the organization that reports that crime data. Each code should also have a category definition and a description of the crime.

Individual people can also be considered a separate entity. A person could be described as the victim of a crime, the perpetuator of a crime, or the suspect of a stop-and-search. An issue that can be identified here is that the datasets have omitted personal information to avoid privacy issues, which includes information that is typically present in a police department database. Thus, for the sake of completeness, it may make sense to generate fake names and phone numbers to go with the rest of the information about individual people.

The rest of the information from the datasets mostly consist of partially overlapping attributes, which necessitates the exploration of the differences between the datasets and the design choices to accommodate these differences.

UK & US Data Separation Option

There is a significant difference between location-based attributes of the UK and US datasets. UK addresses use attributes such as LSOA code and borough for location which usually is not relevant for US addresses, and US addresses use attributes such as precinct, ward, distrinct and state for location, which is irrelevant for UK addresses.

Additionally, UK datasets do not report any crime codes that correspond to the crimes. This is only relevant to crimes in US.

Considering these points, we can outline the separation of UK and US datasets into individual UK and US based addresses and crimes as a practical design option.

Crimes, Complaints & Stop-and-Searches Option

Another practical design option is to divide the datasets into three separate entities: complaints, which would include information from the NYPD Complaints Data Historic dataset, stop-and-searches, which would include information from the `london-stop-and-searches.csv` file of the London Police Records dataset, and crimes, which would include crime data from all other datasets.

This design option considers the different entities that the datasets provide and try to minimize the number of irrelevant attributes, while also providing a generalized division between the dataset attributes.

Crimes Only Option

In order to avoid making the entities of the database too specific, we can also consider further generalizing all the information from all the datasets into a single entity. The advantage of this option is avoidance of over-specialization, but has a possible disadvantage of too many irrelevant attributes caused by over-generalization.

Decision Matrix

We can consider our outlined options and construct a weighted decision matrix to decide between our design choices. We will weigh our alternatives in terms of the following criteria:

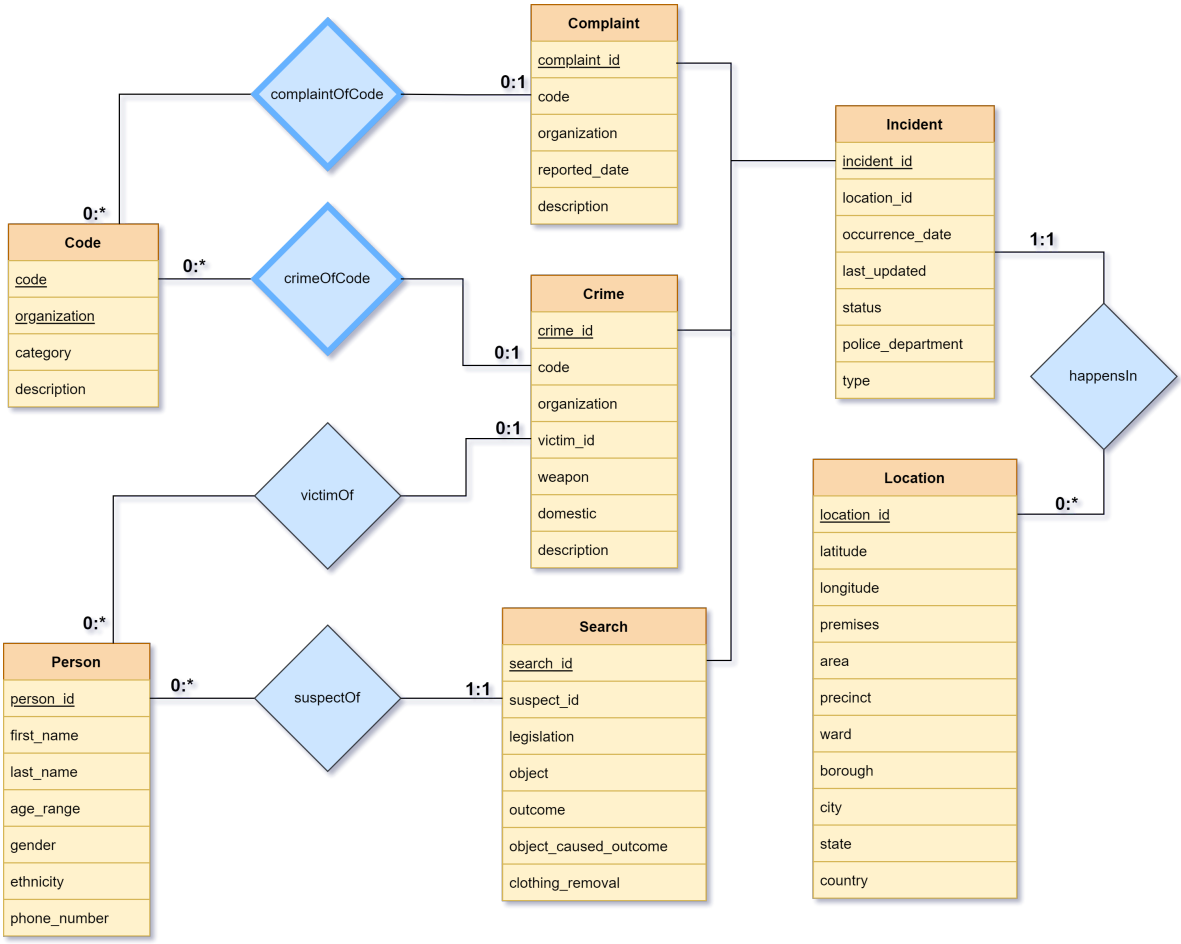
- Generalization of information
- Simplicity of database relationships
- Reduced attribute redundancy
- Reduced possibility of error
- Accommodation of user needs (assuming users are police station employees)

	Generalization of Information	Simplicity of Relationships	Reduced Redundancy	Reduced Error Possibility	Accommodation of needs	Score
Weights	2	1	3	4	5	
UK & US Data Separation	1	1	1	4	3	37
Crimes, Complaints & Stop-and-Searches	3	4	3	5	5	**64**
Crimes Only	5	5	2	1	1	30

By analyzing our decision options using a decision matrix, we can conclude that separating crimes, complaints and stop-and-searches while merging the UK and US datasets is the most optimal solution given the context of our project.

Entity-Relationship Model

We can now build an entity-relationship model based on our selected design options.



Location

The *Location* entity is a combination of all the location-based attributes from all datasets. This includes attributes that are common between UK and US datasets, such as `latitude`, `longitude`, `premises`, `city` and `country`, as well other attributes that are only unique to either UK or US, such as `area`, `precinct`, `ward`, `borough`, and `state`. Note that LSOA code was omitted from this entity since it was uniquely functionally dependent on `borough`, and not particularly relevant if we already have the name of the borough.

The primary key for this entity is an artificial primary key, `location_id`.

Code

The *Code* entity contains all the different US crime codes as reported by the NYPD, IUCR and the LAPD, along with additional information regarding the codes, including `category` and `description`.

Since the uniqueness of the entries of this entity is dependent on both the crime code and the reporting organization, the primary key for this entity is a composite key made up of attributes `code` and `organization`.

Person

The *Person* entity holds information about individual people that are relevant to the database (including information that may have to be auto-generated, such as `first_name`, `last_name` and `phone_number`).

The primary key for this entity is an artificial primary key, `person_id`.

Incident

The *Incident* entity represents an incident reported in any of the datasets and is meant to hold general information that is relevant to all specific incidents. This includes attributes `occurrence_date`, `type`, `status`, `police_department` and `last_updated`.

This entity also contains a `location_id` attribute which is related to the *Location* entity through relation *happensIn*. Ideally, this is a many-to-one relation between *Incident* and *Location* (i.e. multiple incidents can happen in the same location).

The primary key for this entity is an artificial primary key, `incident_id`.

Complaint

The *Complaint* entity is a specialization of the *Incident* entity and provides additional information about incidents that are complaints about crimes, through attributes such as `reported_date` and `description`.

The `code` and `organization` attributes of the *Complaint* entity are related to the *Code* entity that describes the specific crime code referenced by the complaint, through the relation *complaintOfCode*. Note that this is a weak entity set because a complaint about a crime cannot exist if the criminal law (a.k.a crime code) that prohibits that crime does not exist.

The primary key for this entity is an artificial primary key, `complaint_id`.

Crime

The *Crime* entity is a specialization of the *Incident* entity and provides additional information about incidents that are reported crimes, through atttributes such as `weapon`, `domestic` and `description`.

Like those of the *Complaint* entity, the `code` and `organization` attributes of the *Crime* entity are related to the *Code* entity that describes specific crime code of the crime that's reportedly committed, through the relation *crimeOfCode*. Note that, once again, this is a weak entity set.

The `victim_id` attribute is related to the *Person* entity and describes the information related to the victim of the crime, through relation *victimOf*.

The primary key for this entity is an artificial primary key, `crime_id`.

Search

The *Search* entity is a specialization of the *Incident* entity and provides additional information about incidents that are stop-and-searches conducted by the police, through attributes such as `legislation`, `object`, `outcome`, `object_caused_outcome` and `clothing_removal`.

The `suspect_id` attribute of the *Search* entity is related to the *Person* entity and describes the information related to the suspect of the stop-and-search, through relation *suspectOf*.

The primary key for this entity is an artificial primary key, `search_id`.

Relational Schema

Application

Installation

Database Setup

Operations

Data Mining