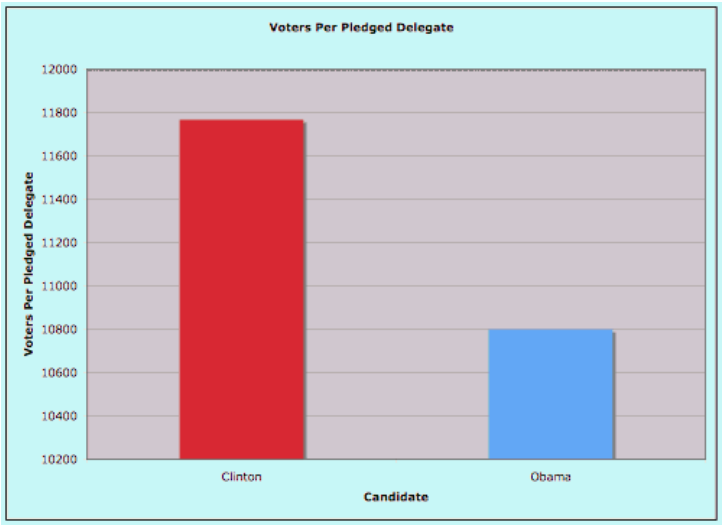
Exam #2 Review

Zuyin (Alvin) Zheng

Data Visualization

Basic principles of Data Visualization

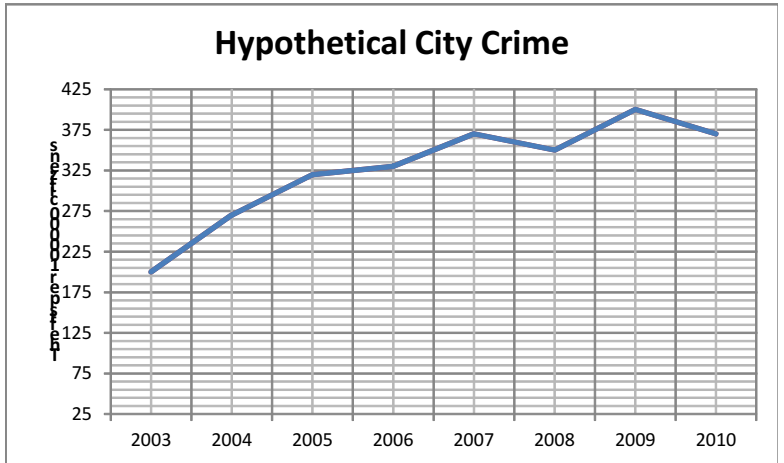
- Principle 1: The chart should tell a story
  - ✓Graphics should be clear on their own
  - ✓The depictions should enable meaningful comparison
  - ✓The chart should yield insight beyond the text
  - ✓"If the statistics are boring, then you've got the wrong numbers." (Tufte 2009)
- Principle 2: The chart should have graphical integrity
- Tufte's "Lie Factor":
  - ✓$Lie\ Factor = \frac{size\ of\ effect\ shown\ in\ graphic}{size\ of\ effect\ in\ data}$
  - ✓>1 Overstate the effect,
  - ✓<1 Understate the effect

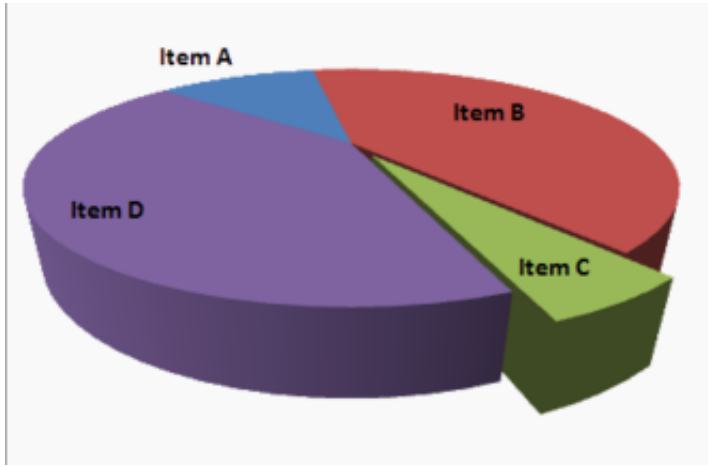


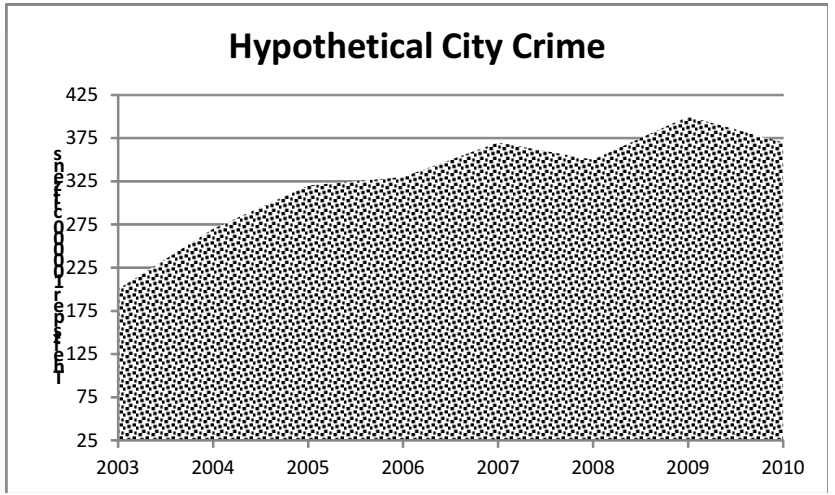
Basic principles of Data Visualization

- Principle 3: The chart should minimize graphical complexity
  - ✓Generally, the simpler the better
  - ✓A chart (table) is not necessarily better than a table (chart)
  - ✓Data Ink:
    - $Data - ink\ ratio = \frac{data-ink}{total\ ink\ used\ in\ graphic}$
    - = 1 implies all ink devoted to data
    - < 1 = more non-data related ink in graphic

Bad Examples







Extract,Transform,Load

Why ETL: Data consistency

The same data element stored in different **formats**

- Social Security number (123-45-6789 versus 123456789)
- Date (10/9/2015 versus 9/10/2015)

Redundant data across the organization

- Customer record maintained by accounts receivable and marketing

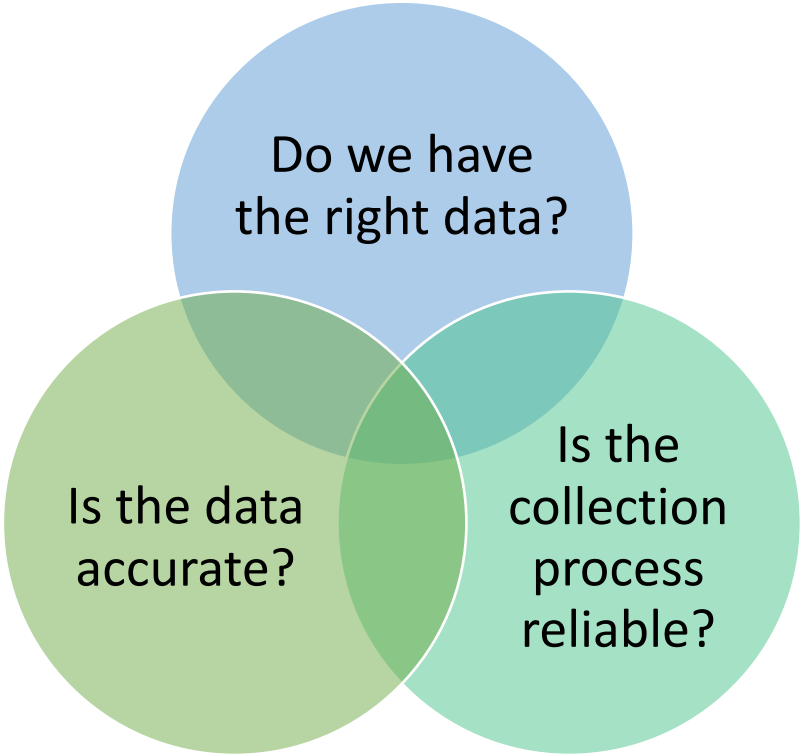
Different **naming** conventions

- “Management Information Systems” versus “MIS” verus “Man. Info. Sys.”

Different **unique identifiers** used

- AccessNet account versus Temple ID

Why ETL: Data Quality



How to do ETL in Excel?

- =CONCATENATE(value1, value2…)
  - ✓ Combines two or more string values or data in cells
  - ✓ Example…
    - CONCATENATE(A2, “, HELLO”) will append the string “, HELLO” to the end of whatever is in cell A2. Like this:
- VLOOKUP(lookup_value, table_array, column_index, range_lookup)
  - ✓ lookup_value = value that you’re looking for
  - ✓ table_array = the table where you’re going to do your search (e.g., A2:E5)
  - ✓ column_index = column number to return from matched record
  - ✓ range_lookup = TRUE for approximate matches and FALSE for exact matches

Example of VLOOKUP

	A	B	C
4	Exercise: fomulas for VLOOKUP		
5	Employee Name	Hours Worked	Hourly Rate
6	Benson, Paul	37.5	32
7	Cooper, David	40	28
8	Dawson, Helen	39	32

	A	B
1	Employee Name	Hourly Rate
2	Atkins, James	35.50
3	Benn, Carol	25.00
4	Benson, Paul	32.00
5	Cooper, David	28.00
6	Daley, Ann	41.00
7	Dawson, Helen	32.00

4	Exercise: fomulas for VLOOKUP		
5	Employee Name	Hours Worked	Hourly Rate
6	Benson, Paul	37.5	32
7	Cooper, David	40	=VLOOKUP(A7,'Hourly Pay'!A:B,2,FALSE)
8	Dawson, Helen	39	

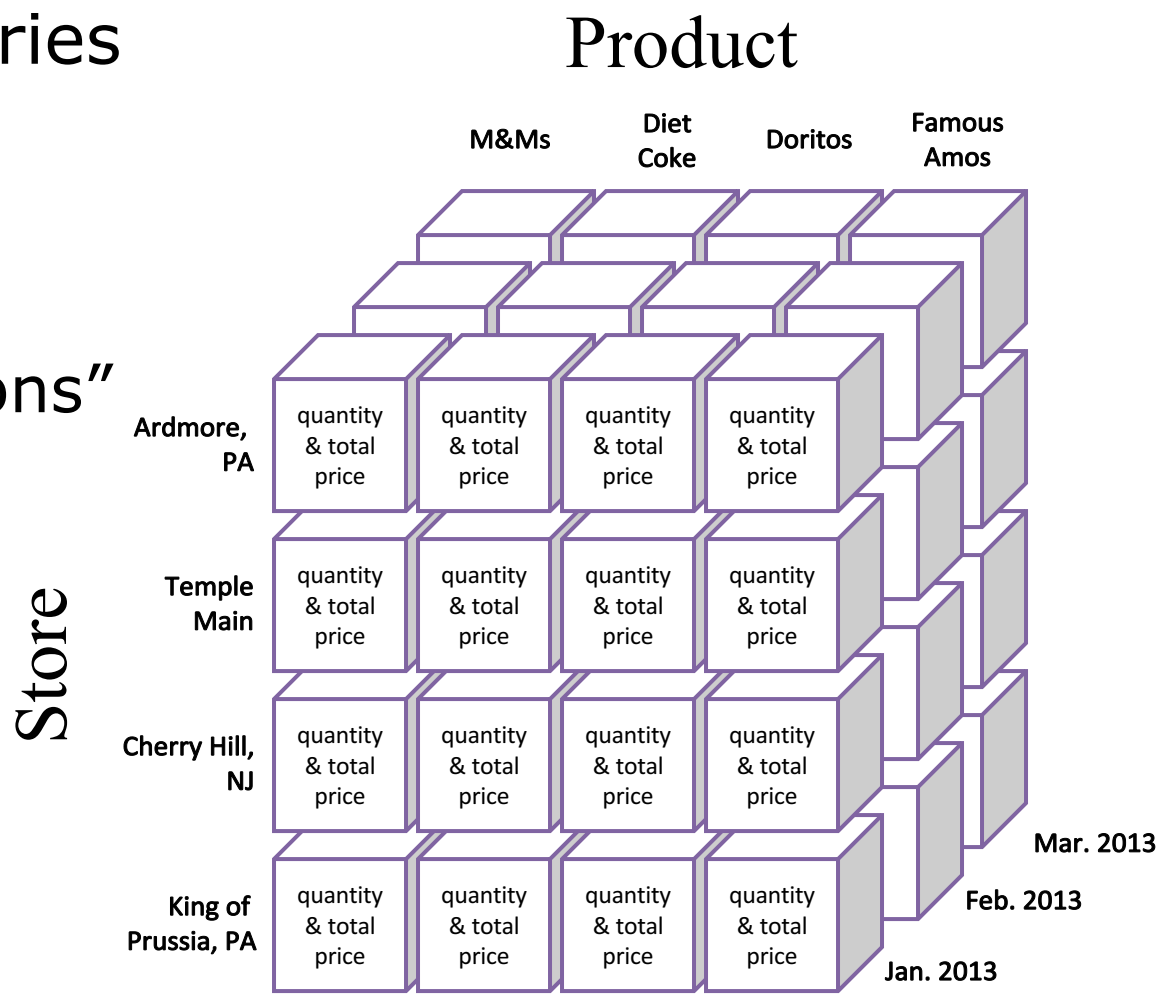
VLOOKUP(lookup_value, table_array, col_index_num, [range_lookup])

Dimensional Modeling
Pivot Tables

Data Cube

- A data cube is a set of summaries from different angles (or dimensions).

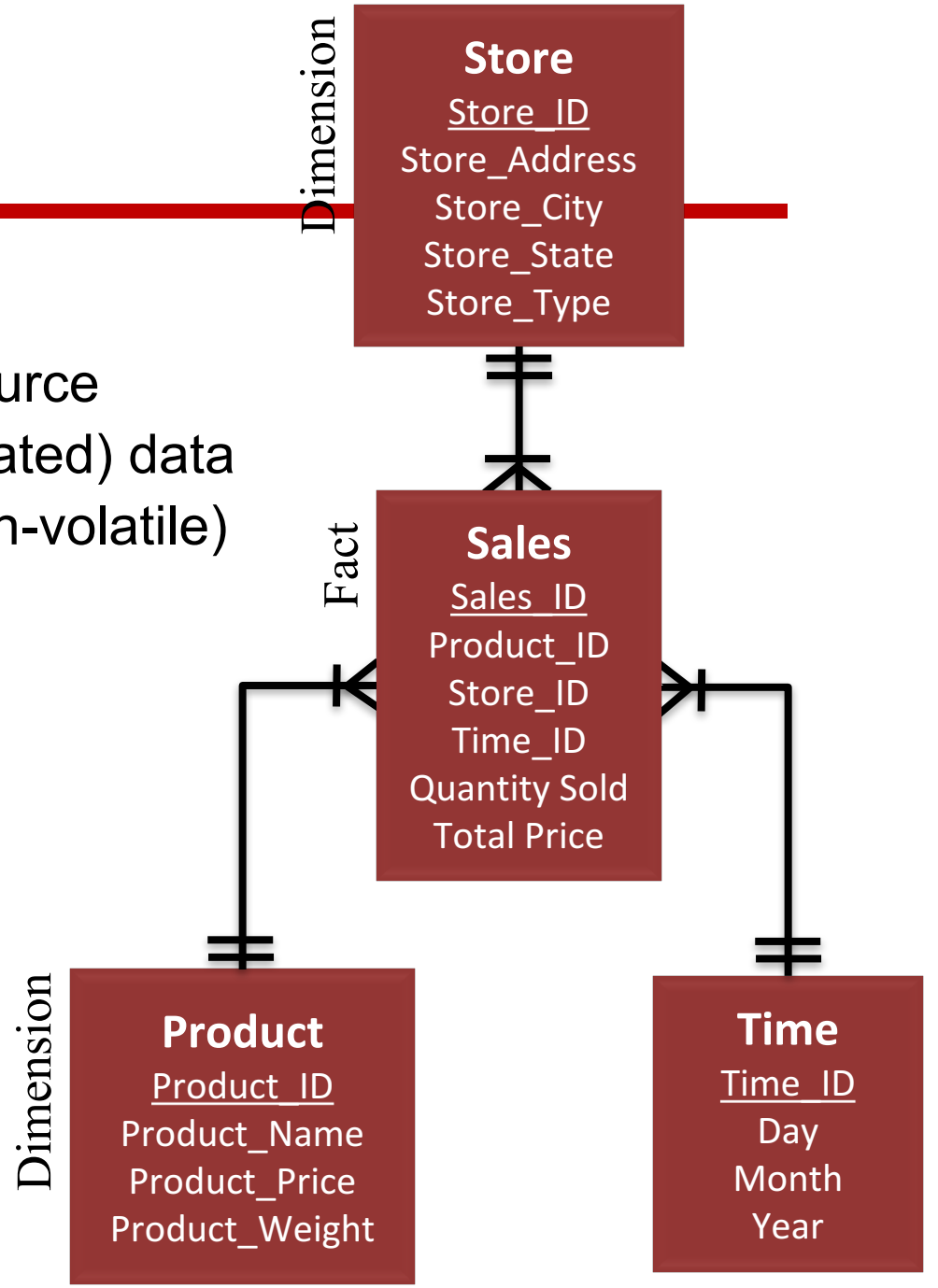
- Made up of "facts" and "dimensions"
  - ✓We can have more than 3 dimensions
- Slicking
  - ✓Dimension reduction
- Dicing
  - ✓A sub-cube
  - ✓No dimension reduction



Modeling a data cube: The Star Schema

- From Star Schema to Data Cube
  - ✓A Cube typically uses a Star Schema as its source
  - ✓and stores precomputed summarized (aggregated) data
  - ✓Much more efficient, but can't be changed (non-volatile)

- Interpret the star schema
  - ✓Cardinality
  - ✓Dimension tables
  - ✓Fact tables
  - ✓Primary key/Foreign keys



Design the Star Schema

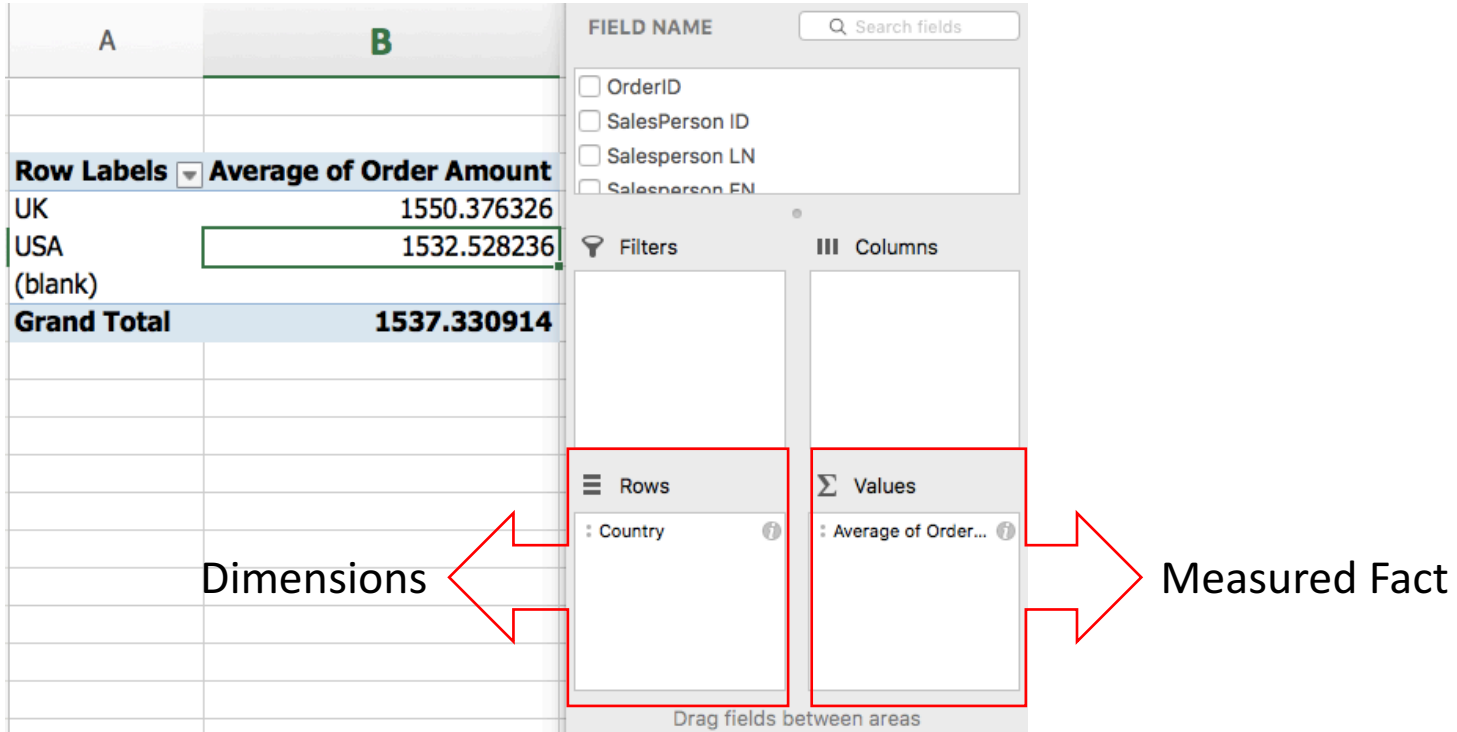
- Choose the business process
  - ✓ what you are interested in
- Identify the fact
  - ✓Quantifiable information for each business event
  - ✓Numeric numbers such as quantity sold, total price
- Decide on the level of granularity
  - ✓Level of detail for each event
    - Choices for time: yearly, quarterly, monthly, daily
    - Choices for store: store, city, state
- Identify the dimensions
  - ✓Time, store and product

Advantages of Data Cube

Speed	• Fast response to give you the information you have previously designed in the cube
Analysis	• The data multi-dimensional data structure allows the data to be analyzed in the most logical way.

Pivot Table and Data Cube

- The fields in the ROWS box correspond to dimensions in a data cube
- The fields in the VALUES box correspond to measured facts in a data cube



SQL In/Out

Putting Data into a Database: Tables

- Create, Update and Drop Tables
  - ✓CREATE TABLE schema_name.table_name (
    columnName1 datatype [NULL][NOT NULL],
    columnName2 datatype [NULL][NOT NULL],
    PRIMARY KEY (KeyName),
    FOREIGN KEY (KeyName) REFERENCES
    schema_name.table_name (KeyName1));
  - ✓DROP TABLE schema_name.table_name;

Putting Data into a Database: Columns

- Changing a table's metadata

```
ALTER TABLE schema_name.table_name
ADD COLUMN column_name datatype
[NULL][NOT NULL];
```

or

```
ALTER TABLE schema_name.table_name
DROP COLUMN column_name;
```

or

```
ALTER TABLE schema_name.table_name
CHANGE COLUMN old_column_name
new_column_name datatype
[NULL][NOT NULL];
```

Adds a column to the table

Removes a column from the table

Changes a column in the table

Data Types

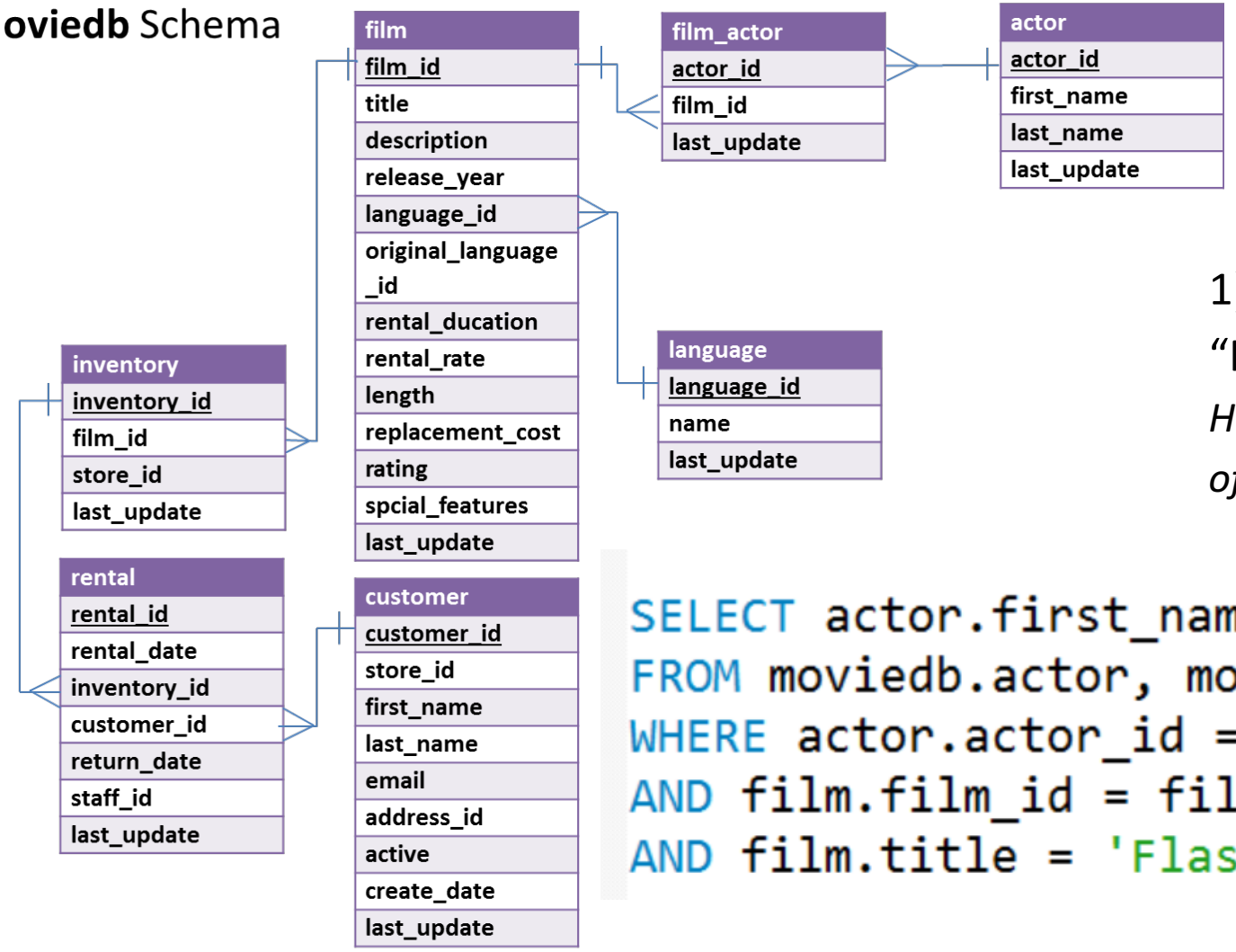
Data type	Description	Examples
INT	Integer	3, -10
DECIMAL(p,s)	Decimal. Example: **decimal(5,2)** is a number that has 3 digits before decimal and 2 digits after decimal (like 123.45)	3.23, 3.14159
VARCHAR(n)	String (numbers and letters) with maximum length n	'Hello', 'I like pizza', 'MySQL!'
DATETIME, DATE	Date/Time, or just Date	'2011-09-01 17:35:00', '2011-04-12'
BOOLEAN	Boolean value	0 or 1

Putting Data into a Database:Rows

- Adding a row
  - ✓**INSERT INTO** schema_name.table_name (columnName1, columnName2, columnName3) **VALUES** (value1, value2, value3);
- Changing a row
  - ✓**UPDATE** schema_name.table_name **SET** columnName1=value1, columnName2=value2 **WHERE** condition;
    - We can change multiple rows at a time
- Deleting a row
  - ✓**DELETE FROM** schema_name.table_name **WHERE** condition;
    - We can delete multiple rows at a time

Join Tables and Sub-select

moviedb Schema

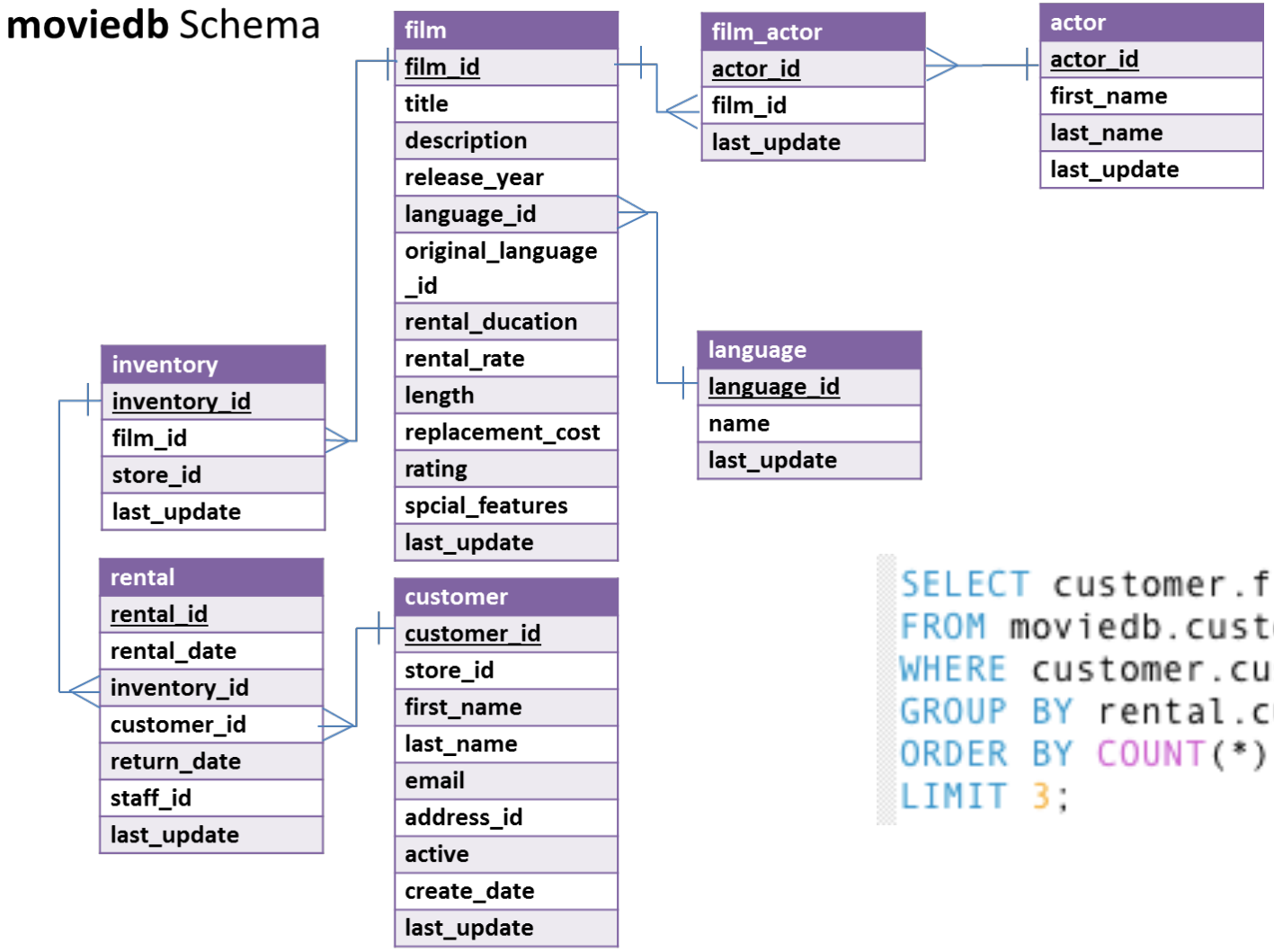


1). Who are the actors of the movie "Flash Wars"?

Hint: display the first name and last name of the actors.

```
SELECT actor.first_name, actor.last_name
FROM moviedb.actor, moviedb.film, moviedb.film_actor
WHERE actor.actor_id = film_actor.actor_id
AND film.film_id = film_actor.film_id
AND film.title = 'Flash Wars';
```

Join Tables and Sub-select



2). Please identify the 3 most value customers who rented most often.

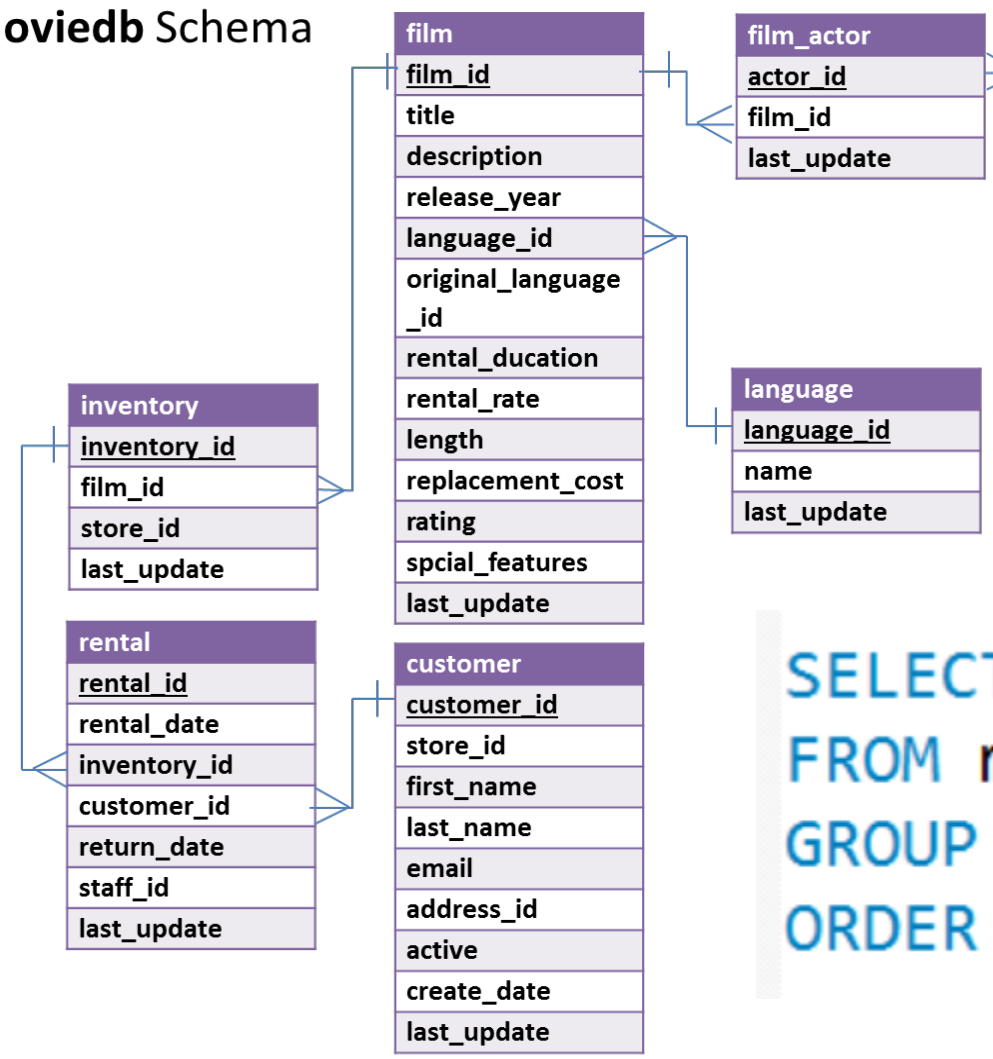
Hint 1: display first name, last name and number of rentals;

Hint 2: there is more than 1 film that has the highest rental rate

```sql
SELECT customer.first_name, customer.last_name, COUNT(*)
FROM moviedb.customer, moviedb.rental
WHERE customer.customer_id=rental.customer_id
GROUP BY rental.customer_id
ORDER BY COUNT(*) DESC
LIMIT 3;
```

Join Tables and Sub-select

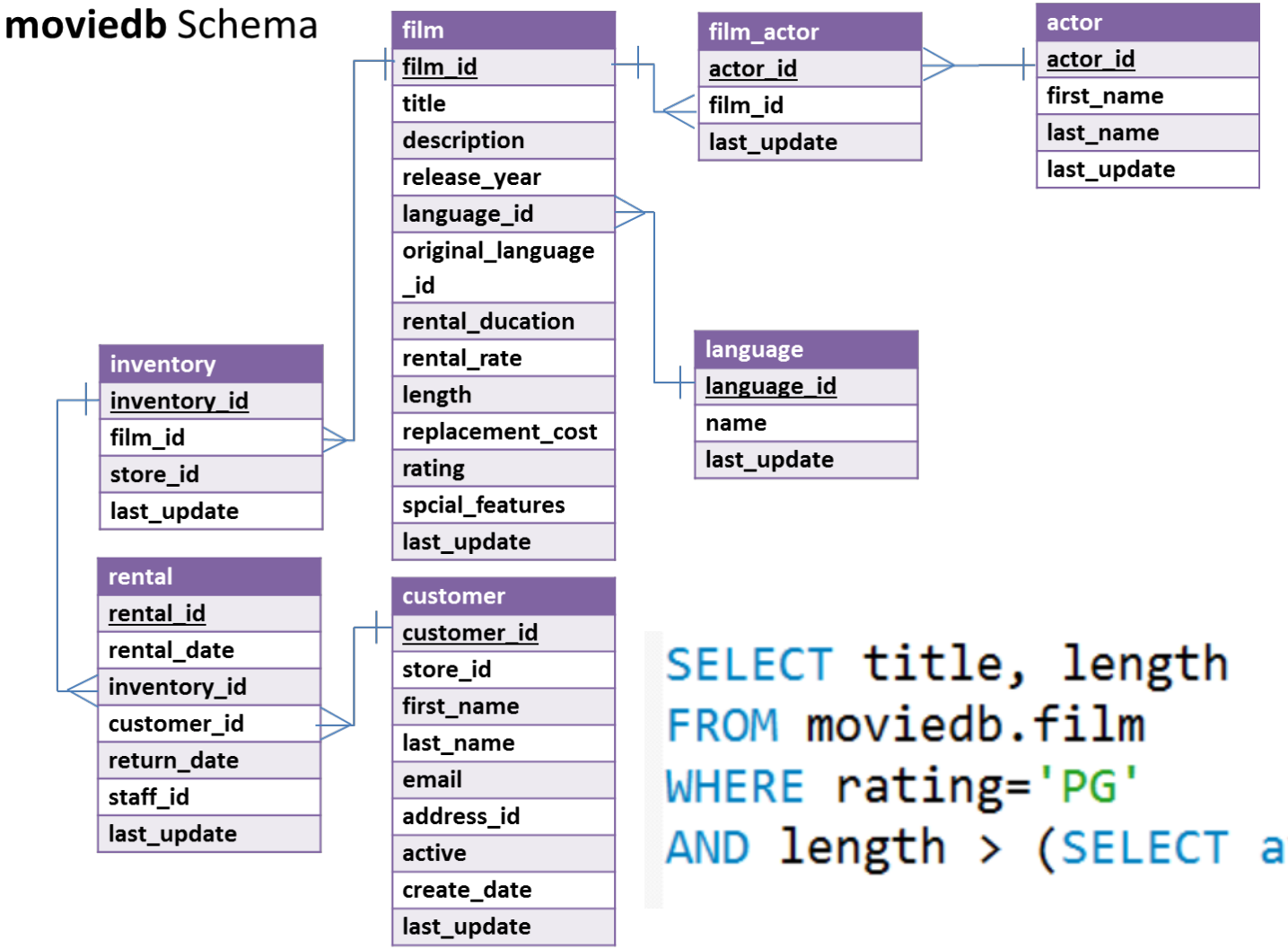
moviedb Schema



3). For different film ratings (i.e., G, PG, R, NC-17), which rating has the highest average rental rate?

```
SELECT rating, AVG(rental_rate)
FROM moviedb.film
GROUP BY rating
ORDER BY AVG(rental_rate) DESC LIMIT 1;
```

Join Tables and Sub-select

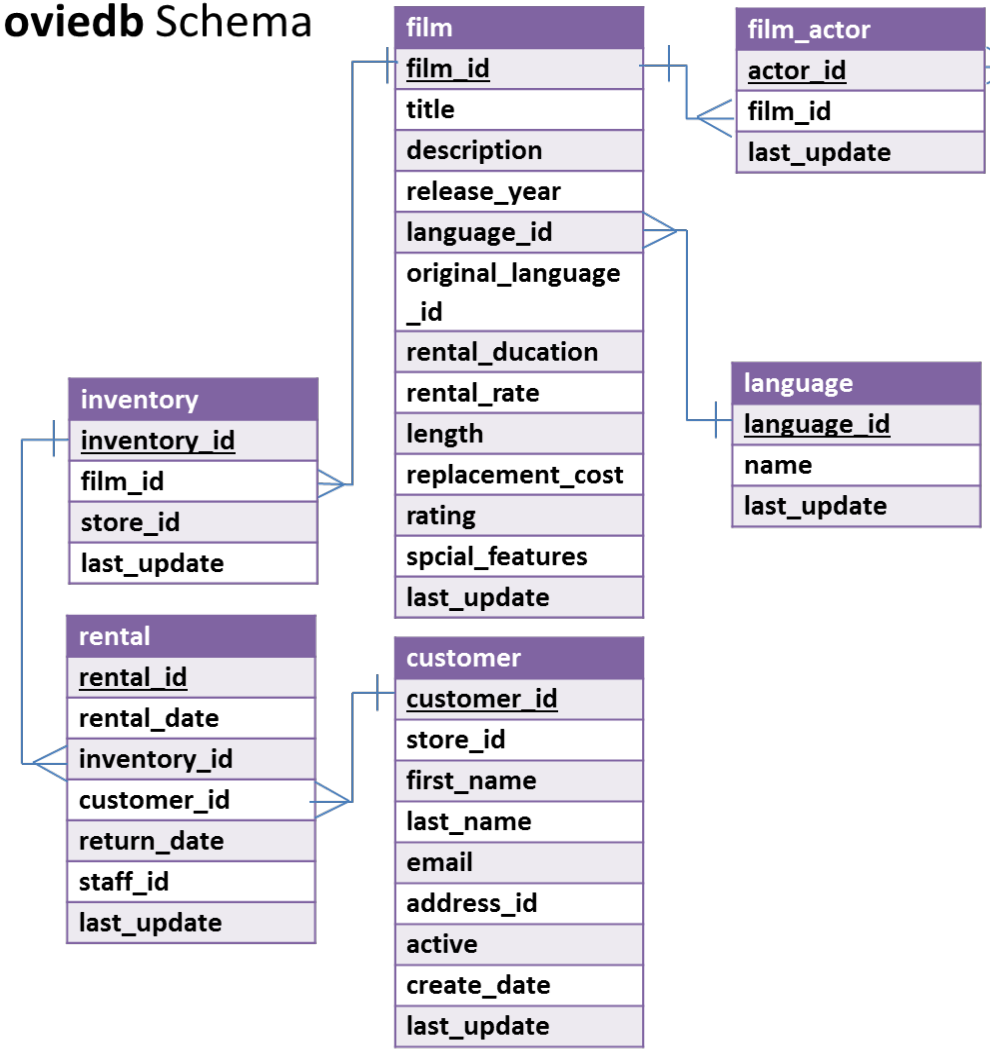


4). What are the title and length for films rated PG and longer than average length?

```
SELECT title, length
FROM moviedb.film
WHERE rating='PG'
AND length > (SELECT avg(film.length) from moviedb.film);
```

Join Tables and Sub-select

moviedb Schema



5). What's the most expensive (in terms of rental rate) film by Salma Nolte? And what is the rental rate?

```
SELECT film.title, film.rental_rate
FROM moviedb.actor, moviedb.film, moviedb.film_actor
WHERE actor.actor_id = film_actor.actor_id
AND film.film_id = film_actor.film_id
AND actor.first_name='Salma'
AND actor.last_name='Nolte'
AND film.rental_rate=(SELECT max(film.rental_rate)
FROM moviedb.actor, moviedb.film, moviedb.film_actor
WHERE actor.actor_id = film_actor.actor_id
AND film.film_id = film_actor.film_id
AND actor.first_name='Salma'
AND actor.last_name='Nolte');
```

Good Luck!