DATA INTEGRATION

THEORY REVIEW

THE AIM OF DATA INTEGRATION SYSTEMS (DIS) IS TO SET UP A SYSTEM WHERE IT'S POSSIBLE TO QUERY DIFFERENT DATA SOURCES AS THEY WERE A UNIQUE ONE THROUGH A GLOBAL SYSTEM

1) HOMOGENEOUS DATA SOURCES: ALL SOURCES USE THE SAME DATA MODEL
2) HETEROGENEOUS DATA SOURCES: DIFFERENT DATA MODELS

MAIN STEPS

1) SOURCE SCHEMA REVERSE ENGINEERING (LOGIC → CONCEPTUAL SCHEMA)
2) SCHEMA INTEGRATION (RELATED CONCEPT IDENTIFICATION + CONFLICT ANALYSIS AND RESOLUTION)
3) GLOBAL CONCEPTUAL SCHEMA
4) CONCEPTUAL → LOGICAL OF THE GLOBAL SCHEMA
5) (G-AV) MAPPING-DEFINITION AND QUERY ANSWERING-
   a) GAV MAPPING DEFINITION
   b) QUERY FORMULATION (ON GLOBAL SCHEMA)
   c) QUERY REWRITING-
6) CONFLICTS (NAME, TYPE, DATA SEMANTIC, STRUCTURE, CARDINALITY, KEY)

Technologies for Information Systems
Part II

prof. L. Tanca – February 28, 2012

Available Time 2h

Last Name _____

First Name _____

Student ID _____ **Signature**_____

LALuxuryHouses is a real estate agency located in Los Angeles and its business is **exclusively** focused on luxury villas located in the Los Angeles area (State of California). Differently, *USAHouses* is an important real estate agency that rents and sells houses in all the main states of the USA. *USAHouses* wants to increase its business in Los Angeles. Since the Los Angeles area is currently only partially covered by the agencies of *USAHouses*, its management decided to buy *LALuxuryHouses* and founded a new company called *USARealEstateCompany*. The management of *USARealEstateCompany* (the new company) wants to integrate the information available in the two sources (*LALuxuryHouses* and *USAHouses*) in order to be able to query all the available data.

In the following we report the original relational schemas of the two sources.

***LALuxuryHouses*:**

CLIENTS (SSN, Lastname, Firstname, Address, City, State, Age, PhoneNumber)
EMPLOYEE (IDEmployee, Lastname, Firstname, PhoneNumber)
HOUSES (HouseAddress, HouseCity, SizeSquareMeters, Rooms) *// The size of each home is measured in square meters.*
HOUSE-OWNEDBY (HouseAddress, HouseCity, ClientSSN) *// Table House-OwnedBy is used to store the information about the owners of each house.*
RENTAL-CONTRACT (IDRentContract, HouseAddress, HouseCity, StartDate, EndDate, AnnualCost, IDEmployee) *// Each tuple in Table Rental-Contract represents the rental of a house (identified by the pair HouseAddress, HouseCity) for the period from StartDate to EndDate*
RENTEDBY (IDRentContract, ClientSSN) *// Table RentedBy is used to store who are the clients associated to each rental contract (i.e., who rented the house associated to the contract).*
SALE (IDSaleContract, HouseAddress, HouseCity, Date, Cost, IDEmployee) *// Each tuple in Sale corresponds to one sale.*
SOLDTO (IDSaleContract, ClientSSN) *// Table SoldTo is used to store who are the buyers associated to each sale.*

***USAHouses*:**

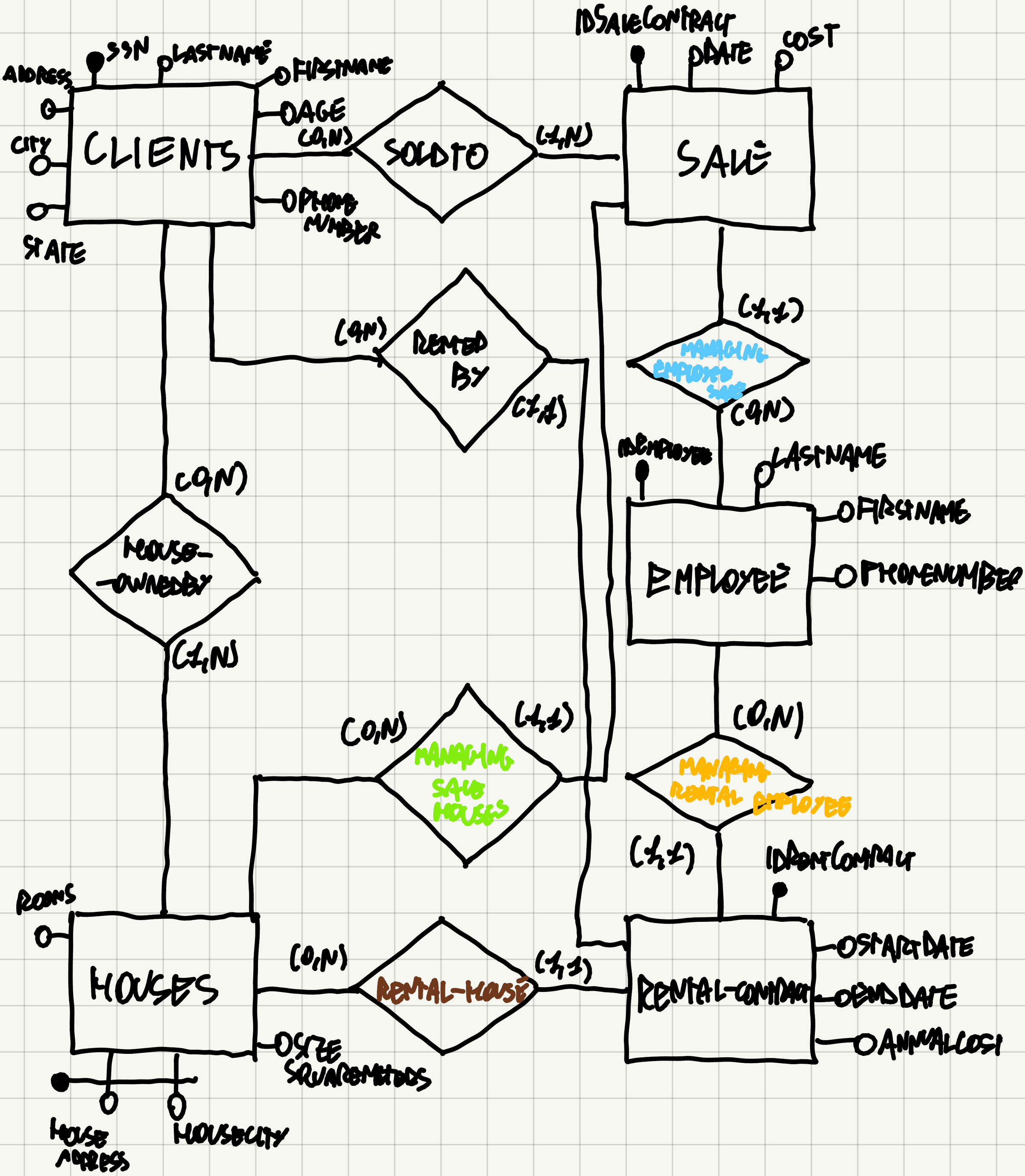
BUYERS (BuyerID, Name, Surname, Address, City, State, YearOfBirth, SSN, PhoneNumber)
// Each tuple in Table Buyers represents someone who bought or rented a real estate
OWNERS (OwnerID, Name, Surname, Address, City, State, YearOfBirth, SSN, PhoneNumber)
// Each tuple in Table Owners represents someone who owns a real estate
AGENTS (AgentID, Name, Surname, MobilePhoneNumber, OfficePhoneNumber)
REALESTATES (IDRE, Address, City, State, NumOfRooms, Size_SquareFeet, NumberOfFloors, OwnerID) *// The size of each real estate is measured in square feet.*
REALESTATE-RENTAL (IDRE, StartDate, EndDate, BuyerID, AgentID, MonthlyCost)
REALESTATE-SALE (IDRE, Date, BuyerID, AgentID, Price)

EXAMPLE OF WEAK ENTITY, DRAWN AS 

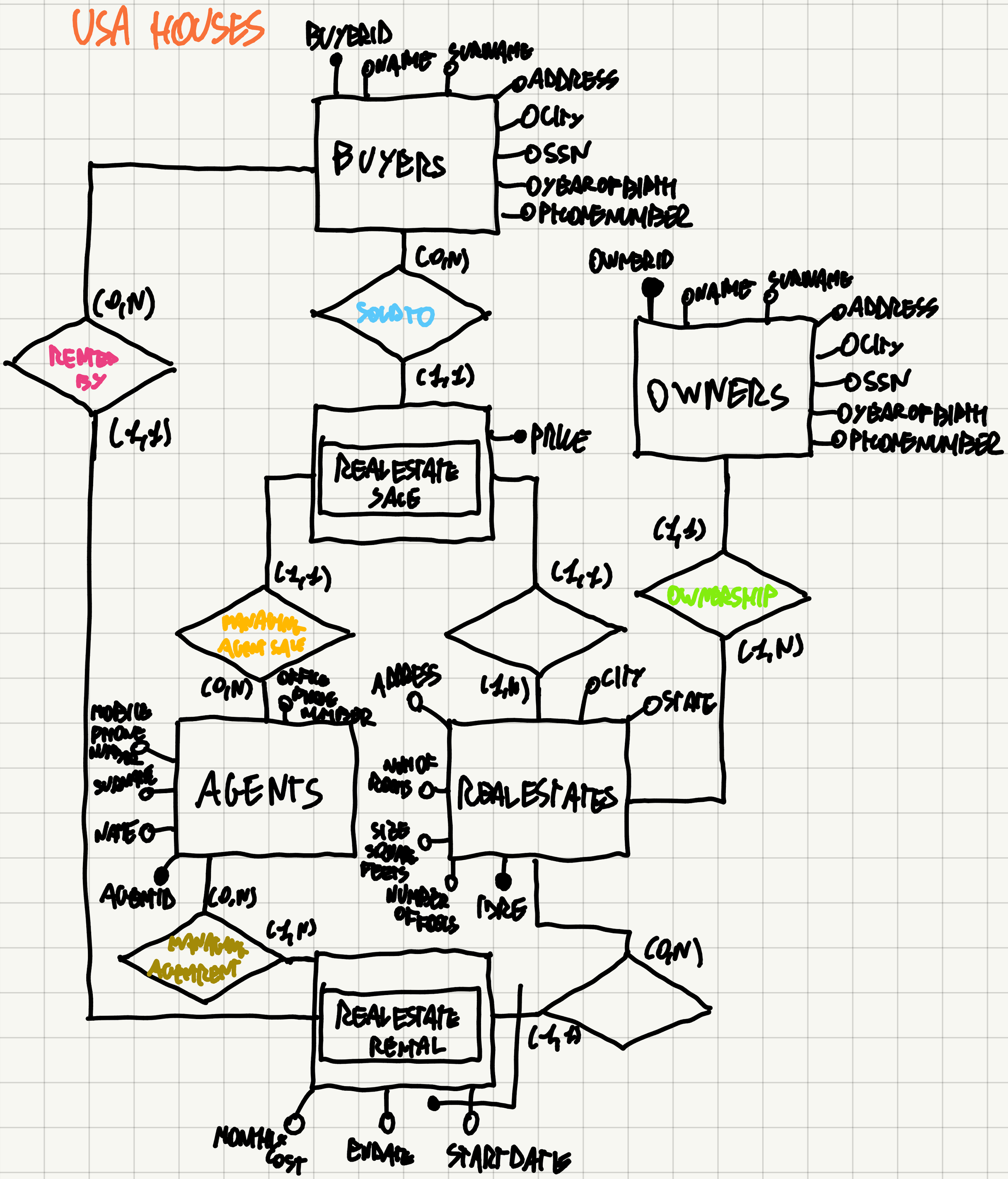
1. Provide, **for each** input data source, the reverse engineering from the logical to the conceptual schema (ER graph). (5 points)
2. Design an integrated global conceptual schema (ER graph) for *USARealEstateCompany* capturing **all** the data coming from both *LALuxuryHouses* and *USAHouses*, and provide the corresponding logical schema. (8 points)
3. Consider the query Q "Find the name and surname of the buyers who live in the city of Los Angeles and have bought at least one house larger than 100 square meters located in the city of Beverly Hills".
    a. Write GAV mappings between the schema of *USARealEstateCompany* and the two sources either in Datalog or SQL. **Write the mappings for the tables used to answer to query Q.** (3 points)
    b. Consider query Q posed on *USARealEstateCompany*'s schema and write it either in Datalog or SQL. (3 points)
    c. Show the rewriting of Q on the data sources either in Datalog or SQL. (3 points)

Important: *1)* Spell out all your assumptions.
2) Avoid information loss as much as possible when defining the new schema.
3) List clearly all conflicts you detect during schema integration, if any.

1) LA LUXURY HOUSES



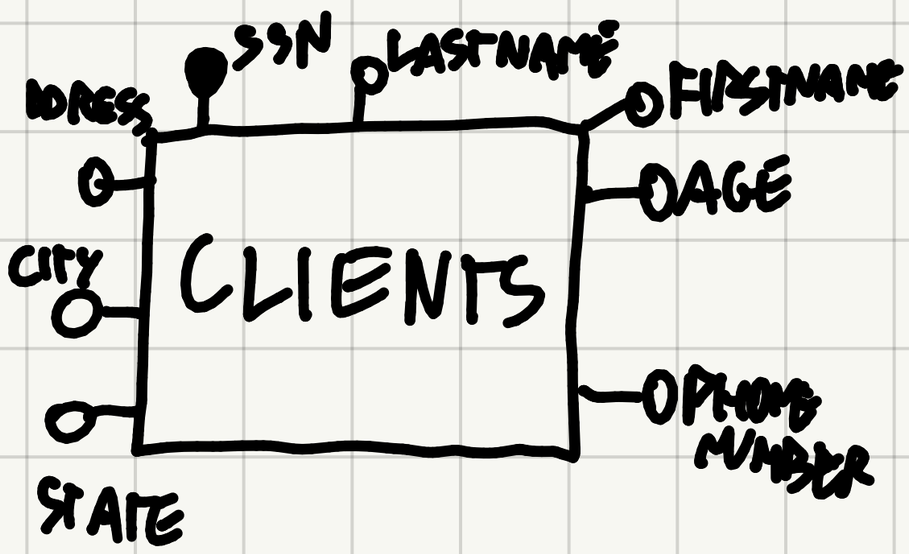
USA HOUSES


BUYERID
ONAME
SURNAME
OADDRESS
OCity
OSSN
OYEAROFBIRTH
OPHONENUMBER
BUYERS
(0,N)
SOLDTO
(1,1)
(0,N)
RENTED BY
(1,1)
REALESTATE SALE
PRICE
(1,1)
(1,1)
PAYMENT AGENT SALE
(0,N)
OFFICE PHONE NUMBER
(1,N)
OWNERID
ONAME
SURNAME
OADDRESS
OCity
OSSN
OYEAROFBIRTH
OPHONENUMBER
OWNERS
(1,1)
OWNERSHIP
(1,N)
ADDRESS
(1,N)
OCity
OSTATE
MOBILE PHONE NUMBER
SURNAME
NAME
AGENTID
AGENTS
NUM OF ROOMS
REALESTATES
SIZE SQUARE FEETS
NUMBER OF POOLS
IDRE
(0,N)
PAYMENT AGREEMENT
(1,N)
REALESTATE RENTAL
(1,1)
(0,N)
MONTHLY COST
ENDATE
STARTDATE

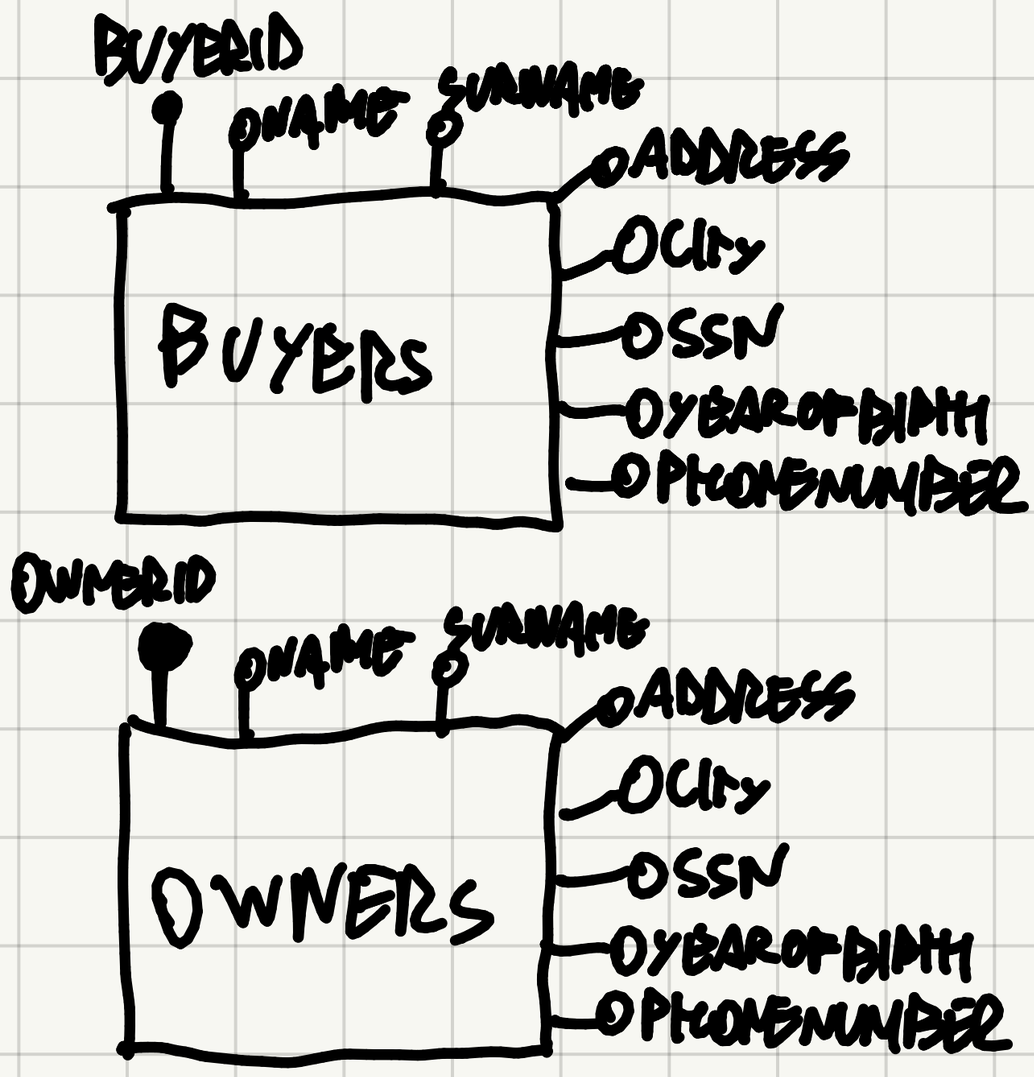
2)

TABLES

LA LUXURY HOUSES



USA HOUSES



BEST PRACTICE: CHOOSE THE MOST GENERAL ENTITY. IN THIS CASE, WE UNITE BUYERS AND OWNERS IN A GENARAL ENTITY CLIENTS

CONFLICTS

SOLUTION

NAME

ENTITIES (CLIENTS; BUYERS, OWNERS) ⟶ CLIENTS

ATTRIBUTES (LAST NAME; SURNAME) ⟶ LAST NAME } *

(FIRST NAME; NAME) ⟶ FIRST NAME

* THE CHOICE IS NOT IMPORTANT, EVEN "SURNAME - NAME" IS AN ACCEPTABLE SOLUTION. THE IMPORTANT IS TO REMAIN COHERENT.

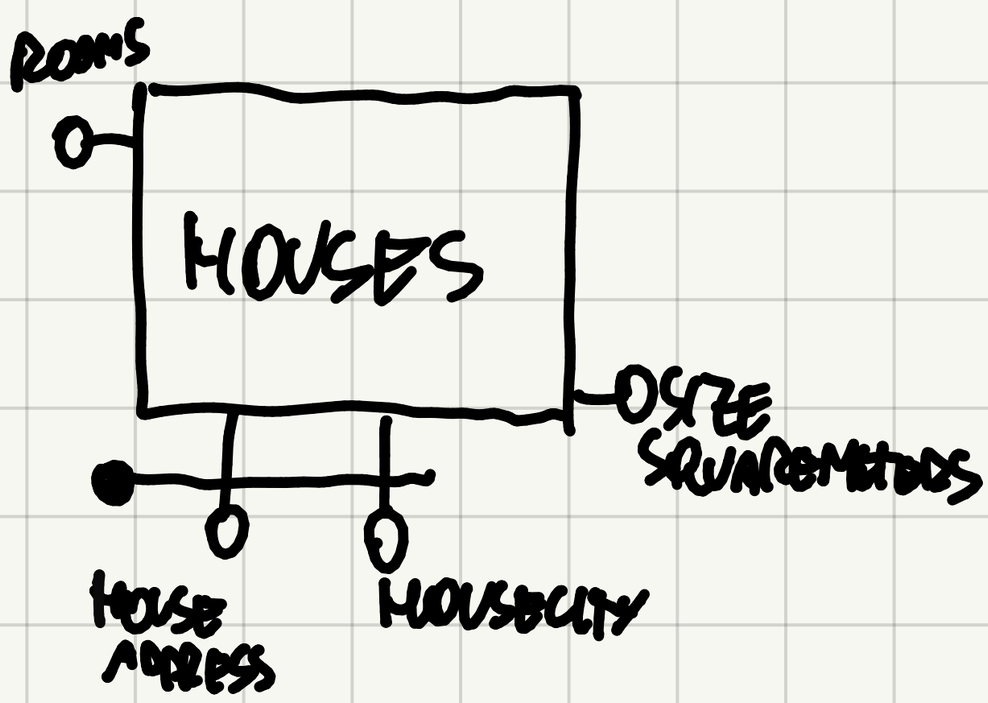
IN GENERAL, IT IS PREFERRED TO CHOSE THE NAME OF THE ATTRIBUTE THAT MATCHES ITS RESPECTIVE ENTITY, AS WE DID NOW

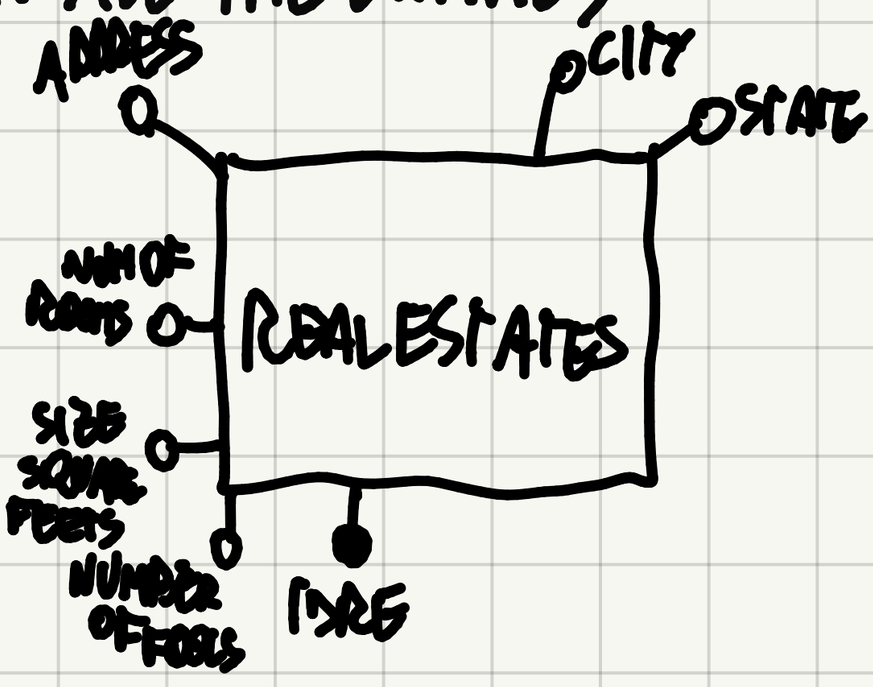
DATA SEMANTIC (AGE; YEAR OF BIRTH) ⟶ YEAR OF BIRTH

IT IS A BETTER CHOICE BECAUSE, WHILE A PERSON'S AGE CHANGES EVERY YEAR, HIS YEAR OF BIRTH DO NOT AND, SO, IT DOESN'T NEED CONTINUOUS UPDATES

KEY CONFLICT (SSN; BUYERID, OWNERID) ⟶ SSN

WE NOTICE THAT SSN COMPARES IN ALL THE ENTITIES





CONFLICTS

SOLUTION

ENTITIES (HOUSES; REALESTATE) NAME ⟶ REALESTATE

ATTRIBUTES (HOUSEADDRESS, HOUSECITY; ADDRESS, CITY) → ADDRESS, CITY

(ROOMS; NUMBEROFROOMS) ⟶ ROOMS

DATA SEMANTIC CONFLICTS (SIZESQUAREMETERS; SIZESQUAREFEETS)

⟶ SIZESQUAREMETERS *

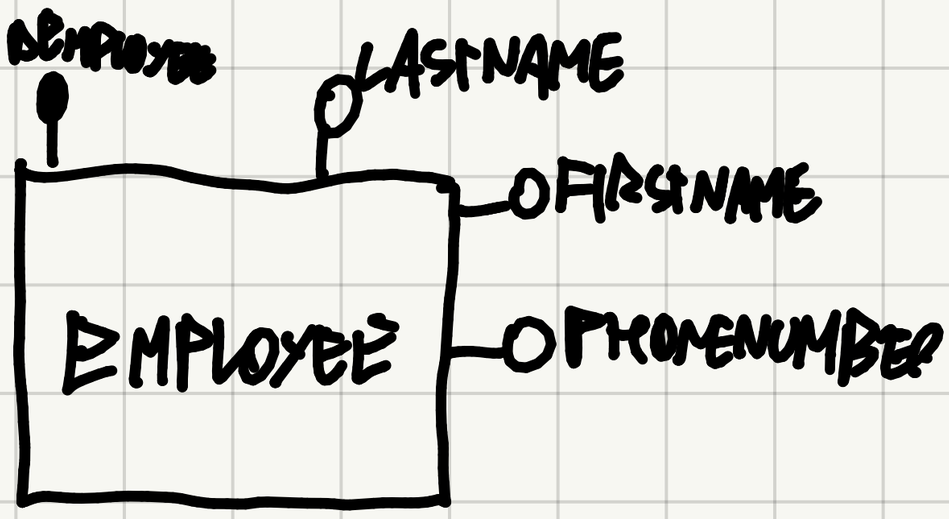
* AN ADVICE IS TO SEE IF THERE ARE SOME SPECIFIC REQUIREMENTS FROM THE EXERCISE'S TRACK. IN THIS CASE, THE REQUESTED QUERY CALLS FOR "... SQUARE METERS..."

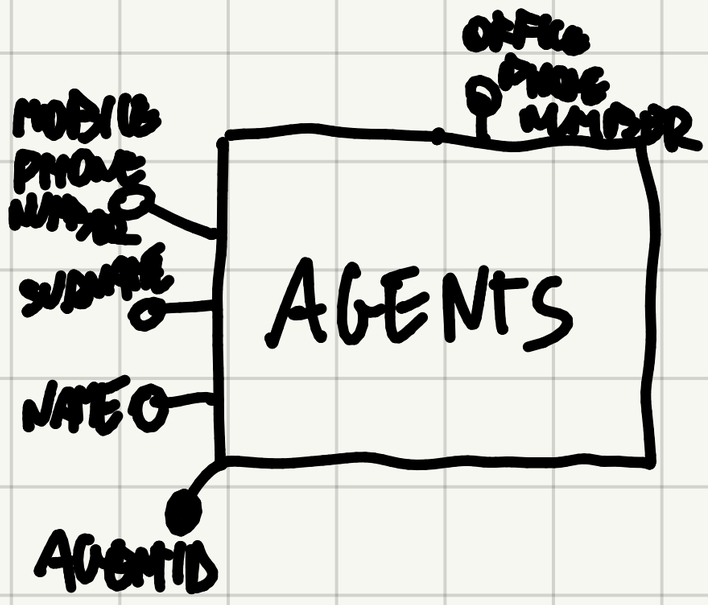
KEY CONFLICT (HOUSEADDRESS, HOUSECITY; IDRE) → ADDRESS, CITY

CARDINALITY CONFLICTS

- NUMBER OF OWNERS ((1,N); (1,1)) ⟶ (1,N)

WE PRETEND TO STAY AS MUCH GENERAL AS POSSIBLE.





CONFLICTS

SOLUTION

NAME

ENTITIES (EMPLOYEE; AGENTS) ⟶ EMPLOYEE *

* IN THIS CASE, BOTH CHOICES ARE ACCEPTABLE

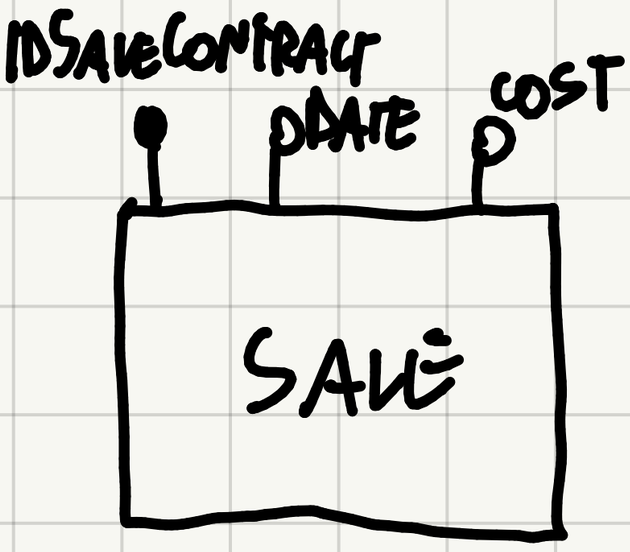
ATTRIBUTES (IDEMPLOYEE; AGENTID) ⟶ EMPLOYEE ID

(LASTNAME; SURNAME) ⟶ LASTNAME

(FIRSTNAME; NAME) ⟶ FIRSTNAME

(PHONENUMBER; MOBILEPHONENUMBER, OFFICEPHONENUMBER) → OFFICEPHONENUMBER

(WE ASSUME THAT PHONENUMBER=OFFICEPHONENUMBER)



CONFLICTS

SOLUTION

NAME

ENTITIES (SALE; REALESTATENAME) ⟶ SALE

ATTRIBUTES (COST; PRICE) ⟶ COST

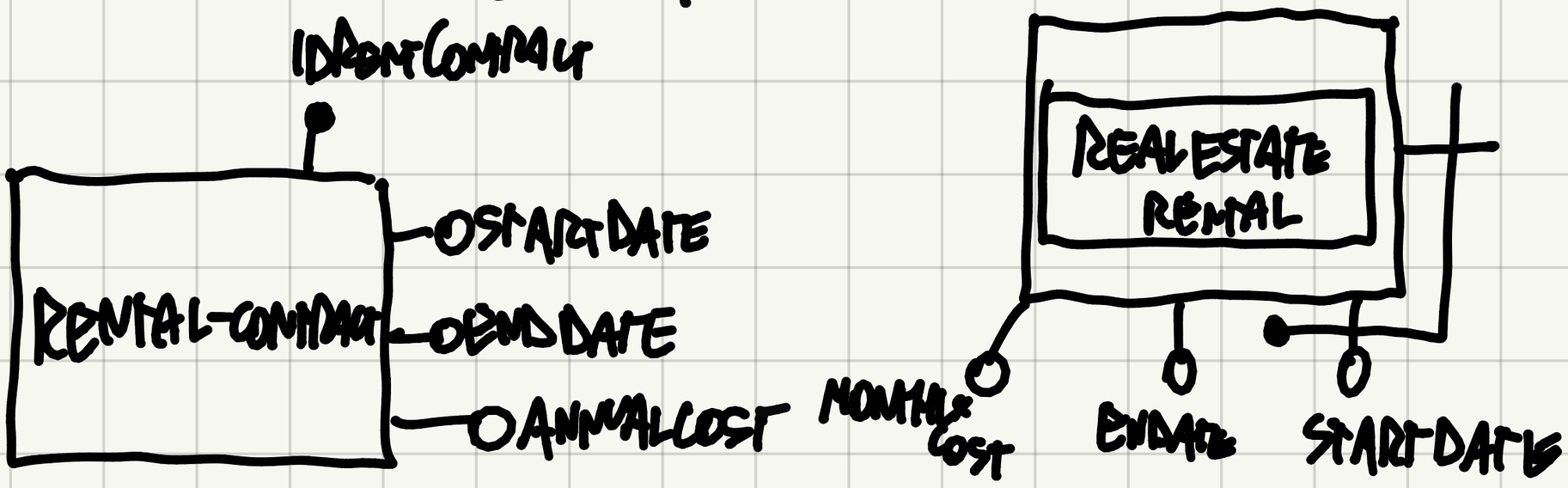
KEY CONFLICT (IDSALECONTRACT; DATE+IDRE) ⟶ IDSALECONTRACT

[KEYGEN(ID, LALUXURYHOUSES); KEYGEN(DATE+IDRE, USAHOUSES)] **

* WE CAN GENERATE A PRIMARY KEY FROM THE UNION OF THOSE

CARDINALITY CONFLICTS

- NUMBER OF BUYERS (CO,N); (1,1) ⟶ (CO,N)



CONFLICTS

SOLUTION

NAME

ENTITIES(RENTAL-CONTRACT; REALESTATERENTAL) ⟶ RENTAL-CONTRACT

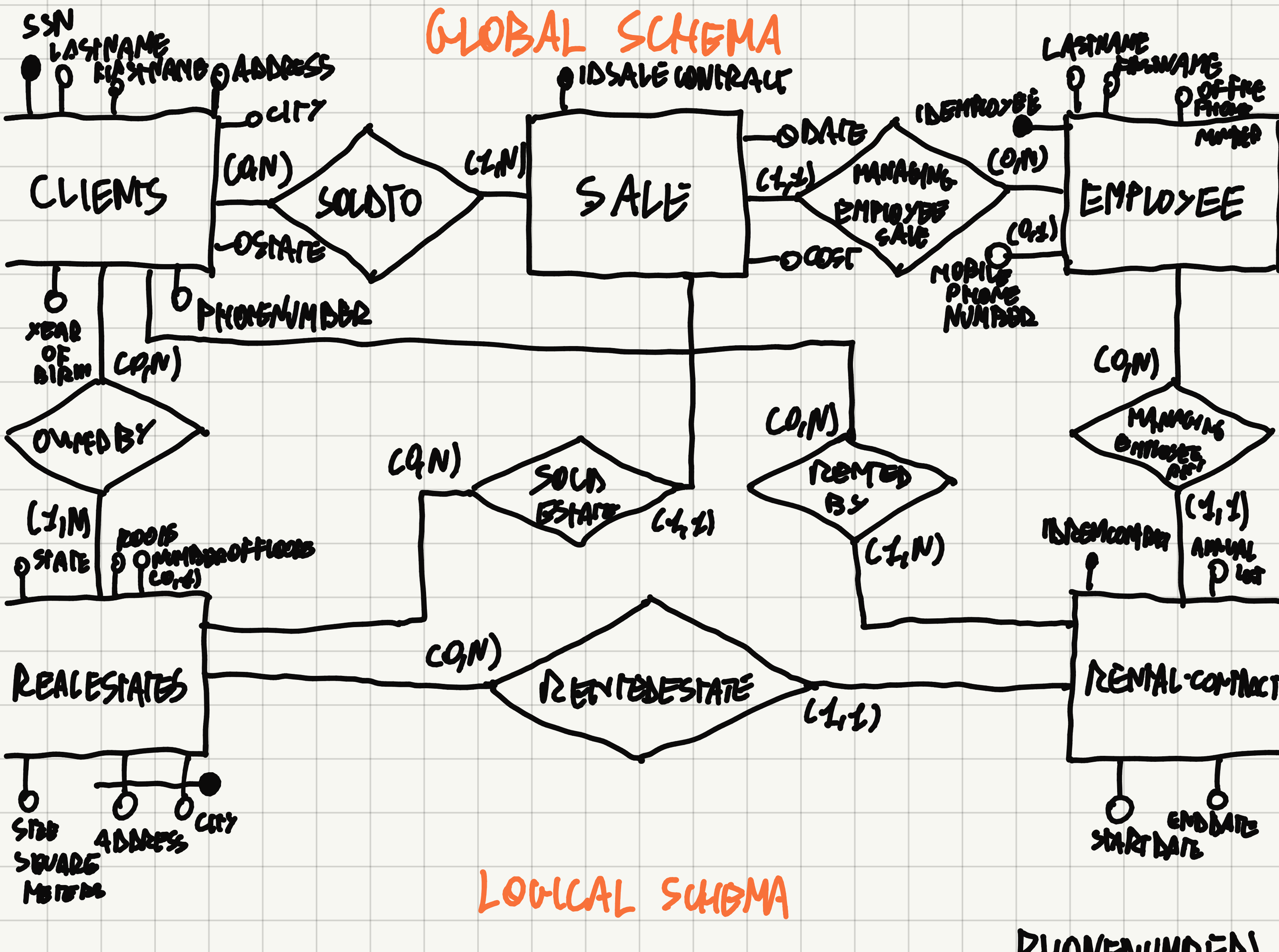
DATASEMANTICCONFLICTS(ANNUALCOST; MONTHLYCOST) ⟶ ANNUALCOST

KEYCONFLICT(IDRENTCONTRACT; IDRE+STARTDATE) ⟶ IDRENTALCONTRACT

CARDINALITY CONFLICT

- NUMBER OF RENTERS (CO,N); (1,1) ⟶ (CO,N)

GLOBAL SCHEMA



LOGICAL SCHEMA

CLIENTS(SSN, LASTNAME, FIRSTNAME, ADDRESS, CITY, STATE, YEAROFBIRTH, PHONENUMBER)

SALE (IDSALECONTRACT, DATE, COST, ESTATEADDRESS, ESTATECITY, IDEMPLOYEE)

SOLDTO (IDSALECONTRACT, CLIENT)

REALESTATES(ADDRESS, CITY, STATE, SIZESQUAREMETERS, ROOMS, NUMBEROFFLOORS*)

RENTAL-CONTRACT (IDRENTCONTRACT, STARTDATE, ENDDATE, ANNUALCOST, IDEMPLOYEE, ESTATEADDRESS, ESTATECITY)

RENTEDBY (IDRENTCONTRACT, CLIENT)

OWNEDBY (ESTATEADDRESS, ESTATECITY, CLIENT)

EMPLOYEE(IDEMPLOYEE, LASTNAME, FIRSTNAME, OFFICEPHONENUMBER, MOBILEPHONENUMBER*)

* OPTIONAL ATTRIBUTE