

## Lecture 2 Entity-Relationship Model

Eugene Wu  
Fall 2015

## HW 0

VM vs virtualenv vs folder vs python shell

Maybe talk about GIT workflows

How to ask for help

<https://github.com/w4ll1/syllabus#help>

The TA Situation

## HW0: Why the different results?

### Result: 69

```
import csv
file = open('iowa-liquor-sample.csv')
file_reader = csv.reader(file)
n = 0
for row in file_reader:
    for el in row:
        if "single malt scotch" in el.lower():
            n += 1
print n
```

### Result: 51

```
file = open('iowa-liquor-sample.csv', 'r')
n = 0
for line in file:
    temp = line.lower()
    if 'single malt scotch' in temp:
        n += 1
    print n
```

## HW0 Stats

enrolled 79

on waitlist 84

<http://eugenewu.net/students.html>

## Steps for a New Application

### Requirements

what are you going to build?

### Conceptual Database Design

pen-and-pencil description

### Logical Design

formal database schema

### Schema Refinement:

fix potential problems, normalization

### Physical Database Design

use sample of queries to optimize for speed/storage

### App/Security Design

prevent security problems

## Steps for a New Application

### Requirements

what are you going to build?

### Conceptual Database Design

pen-and-pencil description

ER Modeling

### Logical Design

formal database schema

### Schema Refinement:

fix potential problems, normalization

### Physical Database Design

use sample of queries to optimize for speed/storage

### App/Security Design

prevent security problems

## Database Apps Are Complicated

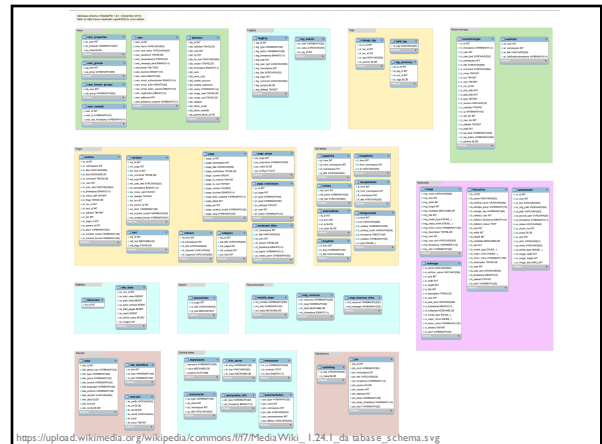
Typical Fortune 100 Company

~10k different information (data) systems

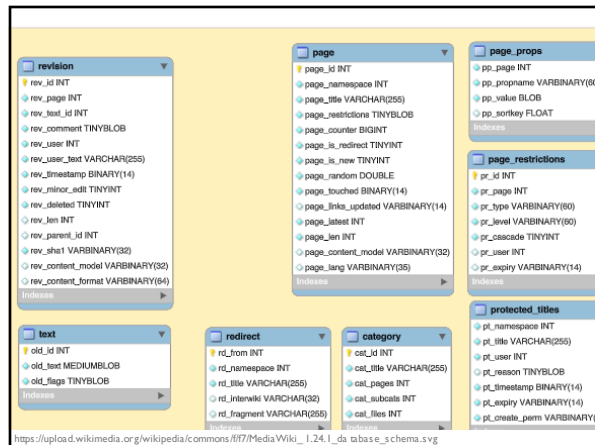
90% relational databases (DBMSes)

Typical database has >100 tables

Typical table has 50 – 200 attributes



[https://upload.wikimedia.org/g/wikipedia/commons/0/07/MediaWiki\\_1.24.1\\_database\\_schema.svg](https://upload.wikimedia.org/g/wikipedia/commons/0/07/MediaWiki_1.24.1_database_schema.svg)



[https://upload.wikimedia.org/g/wikipedia/commons/0/07/MediaWiki\\_1.24.1\\_database\\_schema.svg](https://upload.wikimedia.org/g/wikipedia/commons/0/07/MediaWiki_1.24.1_database_schema.svg)

## Inconsistencies/Constraint Violations

Huge amount of effort to avoid inconsistencies

DBLP is the site for computer science publications

The screenshot shows a search result for 'dblp: eugene wu'. It displays about 116,000 results. The results list multiple entries for Eugene Wu, all with the same affiliation: 'University of Trier'. The entries are:
 

- dblp: Eugene Wu 0002
- dblp: Eugene Wu
- dblp: Eugene Wu 0001

 This illustrates inconsistencies in the database where multiple entries exist for the same person without clear differentiation.

## Inconsistencies/Constraint Violations

This screenshot shows a search result for Eugene Wu, filtered by the year 2014. It displays two entries:
 

- Entry [8]: Eugene Wu, Lellani Battle, Samuel R. Madden: The Case for Data Visualization Management Systems. PVLDB 7(10): 903-906 (2014)
- Entry [7]: Alekh Jindal, Praynaa Rawlani, Eugene Wu, Samuel Madden, Amol Deshpande, Mike Stonebraker: VERTEXICA: Your Relational Friend for Graph Analytics! PVLDB 7(13): 1669-1672 (2014)

 A large '≠' symbol is placed between the two entries, indicating an inconsistency or constraint violation. Below this, a search result for the year 1994 is shown, listing:
 

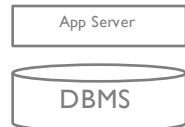
- Entry [2]: James Hwang, Eugene Wu, Alan Bell, Andy Cordell, LeBarian Stokes, Scott Hankins: Design of a SPDM-Like Robotic Manipulator System for Space Station On-Orbit Replaceable Unit Ground Testing - An Overview of the System Architecture. ICRA 1994: 1286-1291
- Entry [1]: Eugene Wu, James Hwang, Scott Hankins: Design of the Control System for a Robotic Manipulator for Space Station On-Orbit Replaceable Unit Ground Testing. ICRA 1994: 1415-1420

## Inconsistencies/Constraint Violations

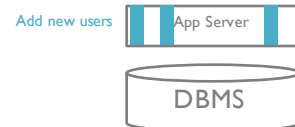
Giving me eugenewu@gmail would violate constraints

The screenshot shows a Google search result for 'eugenewu@gmail'. It displays a message from Google: 'Someone already has that username. Try another?'. Below the message, it says 'Available: eugenewu861'. This illustrates a constraint violation where a username is already taken, even though the user is trying to create a new account.

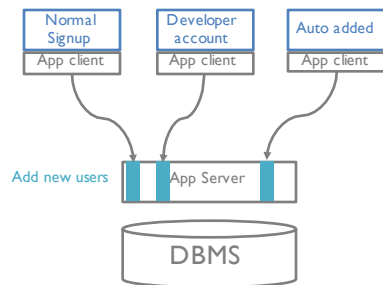
## It is Hard to Design Applications



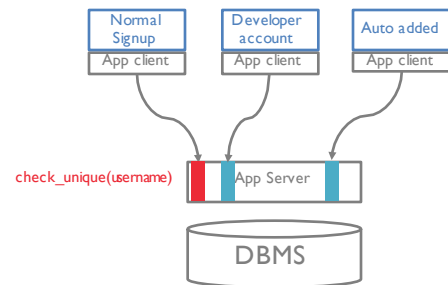
## It is Hard to Design Applications



## It is Hard to Design Applications

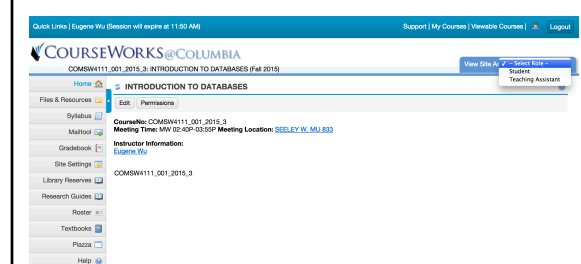


## It is Hard to Design Applications



Let's make a ~~webapp~~ \$\$\$

live exercise time



## Entity-Relationship Modeling

Entities (objects) to store and their attributes  
 Relationships between entities and their attr.  
 Integrity constraints & business rules  
 Visually modeled, easy to turn into DB schema

### ⚙ NEXT SEMESTER COURSES

Fall 2015 – Spring 2016 Courses

| Course Number        | Course Title              |
|----------------------|---------------------------|
| COMSE6910.024.2015.3 | FIELDWORK                 |
| COMSW4111.001.2015.3 | INTRODUCTION TO DATABASES |

Reflects Registrar changes through Mar-06-2015 2:02:13AM

### Courses

Course Number  
 Course Title  
 Year  
 Semester

Eugene Wu test test again just then [Clear](#)

Say something [Say it](#)

[Profile](#) [Wall](#)

#### Basic Information

Nickname

Birthday

Personal summary   
 B / I U ABC | x, x\* | [Link](#) [Image](#) [Video](#) [Audio](#) [Text](#) [HTML](#)

[Save changes](#) [Cancel](#)

#### Contact Information

Email

Home page

Work phone

Home phone

Mobile phone

Facsimile

### Users

Nickname  
 Name  
 Birthday  
 Summary  
 Email  
 ...

## Basics: Entities

**Entity** e.g., intro to databases  
 real-world object distinguishable from other objects  
 described as set of attribute & the values  
 (think one record)

**Entity Set** e.g., all courses  
 collection of similar entities  
 all entities have same attributes (unless Is-A)  
 must have one or more keys  
 attributes have domains  
 ≈ table

## Example: Entity

Keys (cid, uid) are underlined  
 Values must be unique  
 (think: can use as hashtable key to lookup table)



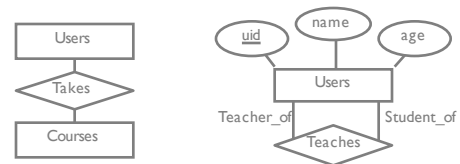
## Basics: Relationships

**Relationship:** association between 2 or more entities  
 e.g., alice **is taking** Introduction to DBs

**Relationship Set:** collection of similar relationships  
 N-ary relationship set R relates N entity sets  $E_1 \dots E_n$   
 Each  $r \in R$  involves entities  $e_1 \dots e_n$   
 An  $E_i$  can be part of diff. relationship sets or diff. roles in same set

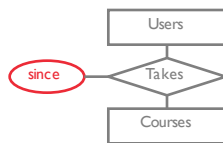
## Basics: Relationships

Users takes diff roles in same relationships set



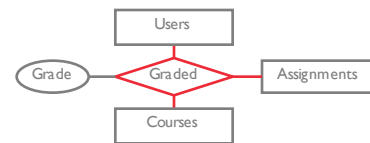
## Basics: Relationships

Relationships sets can have descriptive attributes  
e.g., the *since* attribute of *Instructs*



## Basics: Ternary Relationships

Connects three entities  
N-ary relationships possible too.



## Constraints

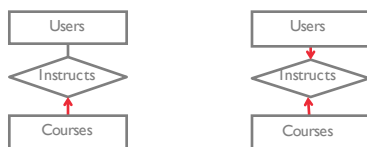
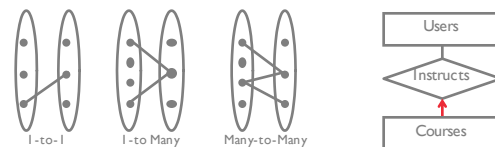
Help avoid corruption, inconsistencies

Key constraints  
Participation constraints  
Weak entities  
Overlap and covering constraints

## Key Constraints

Defines cardinality requirements on relationships

**Many to many** e.g., consider *Takes*  
a user can take many courses  
a course can have many users that take the course  
**One to Many** e.g., consider *Instructs*  
a course has at most one instructor



A course has at most one instructor

???

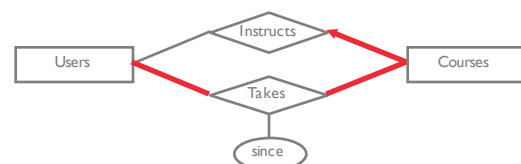
## Participation Constraints

Does every course need an instructor?

If yes, it's a **participation constraint**

e.g., participation of *Courses* in *instructs* is **Total**

Otherwise, **partial** participation constraint

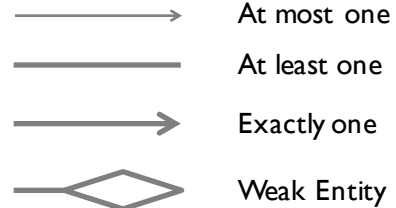
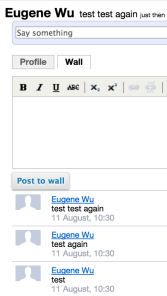
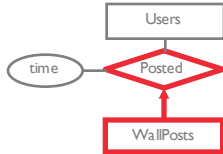


## Weak Entities

A **weak entity** can only be uniquely identified by using the primary key of its owner entity

Owner and weak entity sets must be in one to many relationship set

Weak entity set must have total participation in this **identifying** relationship set



## ISA (is a) Hierarchies

Inheritance rules similar to programming languages

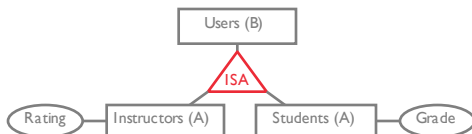
A ISA B  $\rightarrow$  every A also considered a B

When querying for Bs, must consider As (unlike e.g., C++)

Why use ISA?

add descriptive attributes specific to a subclass e.g., grade

identify entities that participate in a relationship



## ISA (is a) Hierarchies

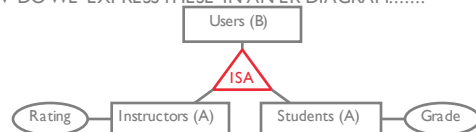
**Overlap Constraint**

can eugene be an instructor and a student? (allow/disallow)

**Covering Constraint**

must every user be an instructor or student? (yes/no)

HOW DO WE EXPRESS THESE IN AN ER DIAGRAM??????



## Update on the Course

Enrollment hijinks are over

5 Tas!  $\rightarrow$  Increased cap to 120.

123 enrolled

Waitlist auto-maintained (you can see status)

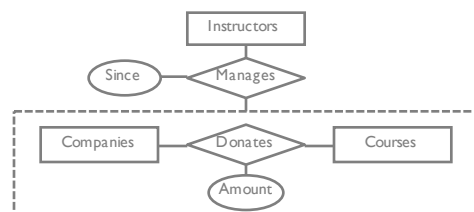
Did you get a github invite?

Project I Part I out

## Aggregation

Relationships between (entities – relationships)

Lets us treat a Relationship Set like an Entity Set so it can participate in other relationships



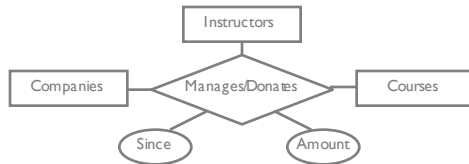
## Aggregation vs Ternary Relationships

Why use aggregation?

Manages and Donates are distinct relationships with own attrs

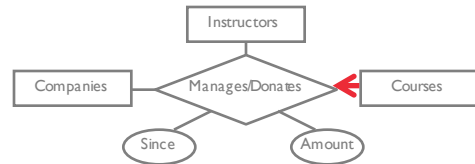
Can define constraints on relationship sets

e.g., a donation can be managed by at most one instructor



## Aggregation vs Ternary Relationships

Constraints apply to all connected entity sets



## Using the ER Model

Design Choices for a concept

Entity or Attribute?

Entity or Relationship?

Binary or Ternary relationship?

Aggregation or Ternary relationship?

## Entity or Attribute?

Is **users.address** an attribute of Users or an entity connected to Users by a relationship?

Depends (and may change over time!)

If a user has >1 addresses, must be an entity

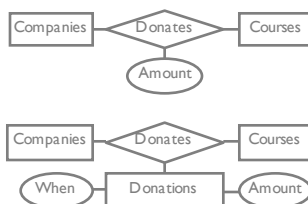
If an address has attrs (structure), must be entity

e.g., want to search for users by city, state, or zip

## Entity or Attribute?

A company can't donate multiple amounts (top fig)

Use ternary relationship (bottom fig)



## Entity or Relationship?

OK if company donates to courses individually

What if company donates to school for all data-related courses?

**Redundancy** of amount, need to remember to update every one

**Misleading** implies amount tied to each donation individually

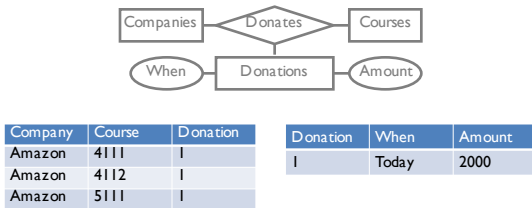


| Company | Course | Amount |
|---------|--------|--------|
| Amazon  | 4111   | 2000   |
| Amazon  | 4112   | 2000   |
| Amazon  | 5111   | 2000   |

} These amounts are logically the same (redundant)!

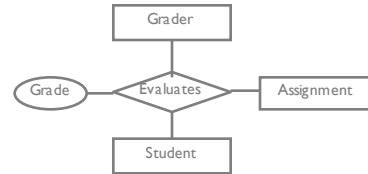
## Entity or Relationship?

If company donates once to school for data related courses.  
Refactor amount into an entity



## Binary or Ternary Relationship?

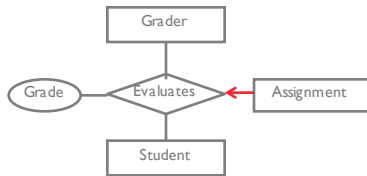
What if assignments have at most one grader?



## Binary or Ternary Relationship?

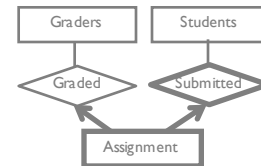
What if assignments have at most one grader?

Only one student can complete HW0!  
Actually two separate relationships



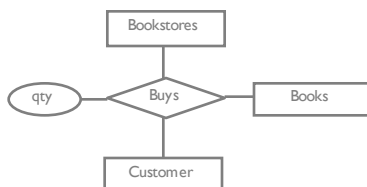
## Binary or Ternary Relationship?

Binary relationships allows additional constraints



## Binary or Ternary Relationship?

Sometimes have true ternary relationship that is defined by all three entities.



## Using ER Modeling

### Constraints in ER Modeling

Many types of data semantics can be captured using ER  
Some constraints not captured (discuss limitations later)

### Need further schema refinement

ER Model is still subjective, need further refinement after translated into relational schema



## Summary

### Requirements

what are you going to build?

### Conceptual Database Design

pen-and-pencil description

(Today) ER Modeling

### Logical Design

formal database schema

### Schema Refinement

fix potential problems, normalization

### Physical Database Design

use sample of queries to optimize for speed/storage

### App/Security Design

prevent security problems

## Summary

Conceptual design follows *requirements analysis*

ER model helpful for conceptual design

constraints are expressive

matches how we often think about applications

Core constructs

entity, relationship, attribute

weak entities, ISA, aggregation

Many variations beyond today's discussion

## Summary

ER design is subjective based on usage+needs

Today we saw multiple ways to model same idea

ER design is not complete/perfect

Developed in an enterprise-oriented world (ER First)

Doesn't capture semantics (what does "instructor" mean?)

Doesn't capture e.g., processes/state machines

How to combine multiple ER models automatically?

Limitation of imagination when designing webapp

Open problems!

ER design is a useful way of thought

## Next Time

Relational Model: de-facto DBMS standard

Set up for ER diagrams → Relational models

- Maybe talk about abstractions