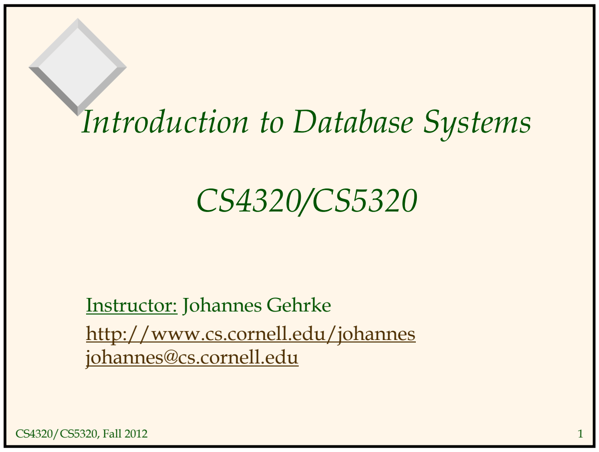
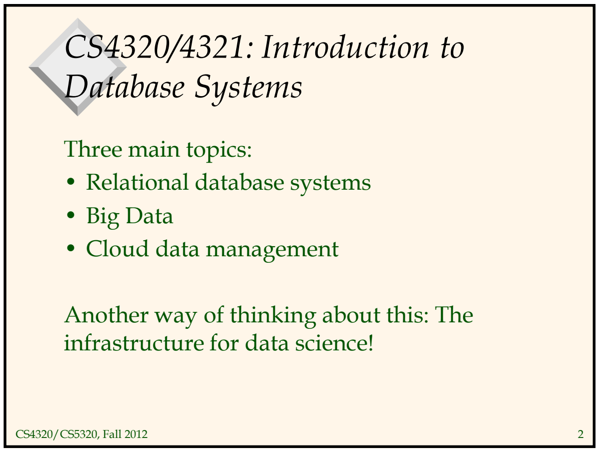


Introduction to Database Systems

CS4320/CS5320

Instructor: Johannes Gehrke
http://www.cs.cornell.edu/johannes
johannes@cs.cornell.edu

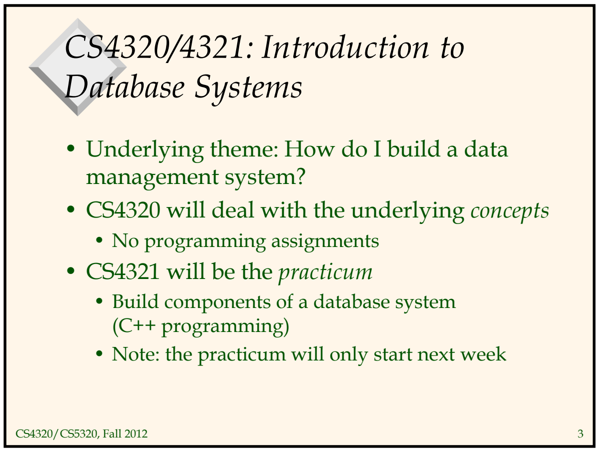


CS4320/4321: Introduction to Database Systems

Three main topics:

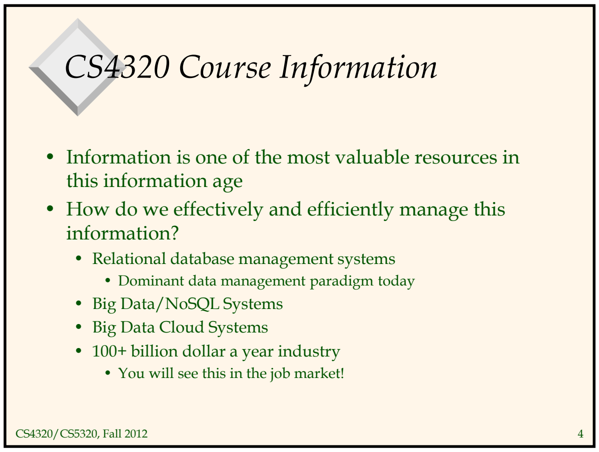
- Relational database systems
- Big Data
- Cloud data management

Another way of thinking about this: The infrastructure for data science!



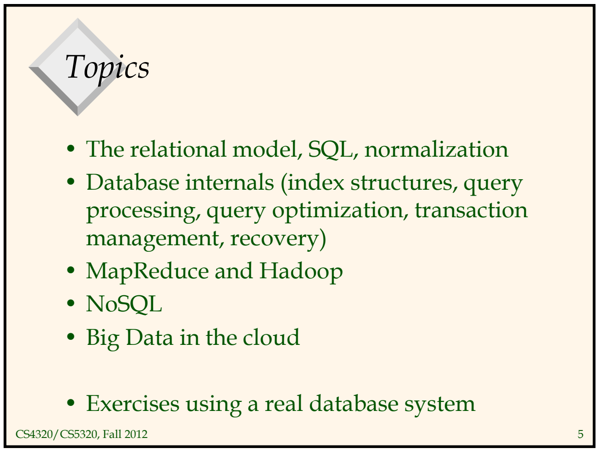
CS4320/4321: Introduction to Database Systems

- Underlying theme: How do I build a data management system?
- CS4320 will deal with the underlying *concepts*
  - No programming assignments
- CS4321 will be the *practicum*
  - Build components of a database system (C++ programming)
  - Note: the practicum will only start next week



CS4320 Course Information

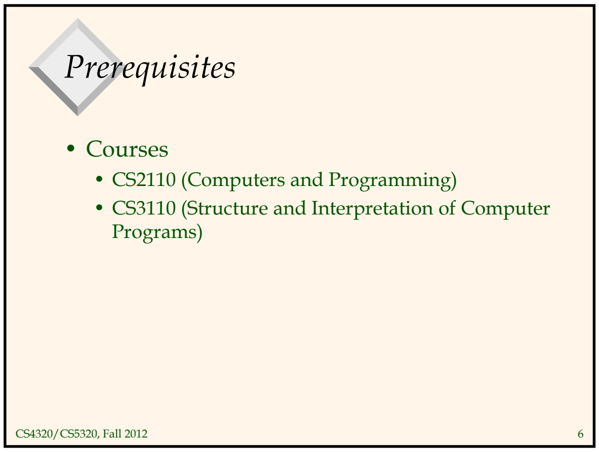
- Information is one of the most valuable resources in this information age
- How do we effectively and efficiently manage this information?
  - Relational database management systems
    - Dominant data management paradigm today
  - Big Data/NoSQL Systems
  - Big Data Cloud Systems
  - 100+ billion dollar a year industry
    - You will see this in the job market!



Topics

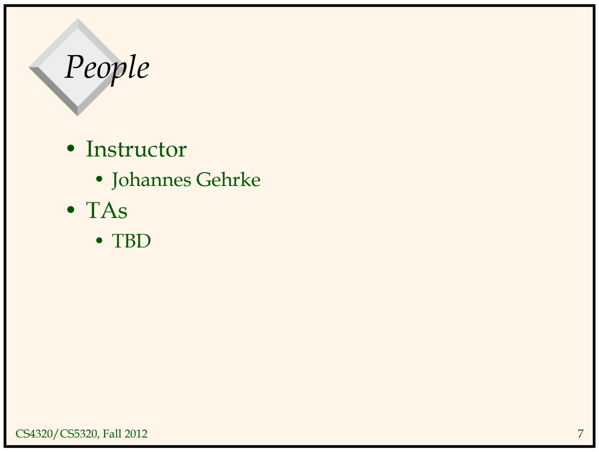
- The relational model, SQL, normalization
- Database internals (index structures, query processing, query optimization, transaction management, recovery)
- MapReduce and Hadoop
- NoSQL
- Big Data in the cloud

- Exercises using a real database system



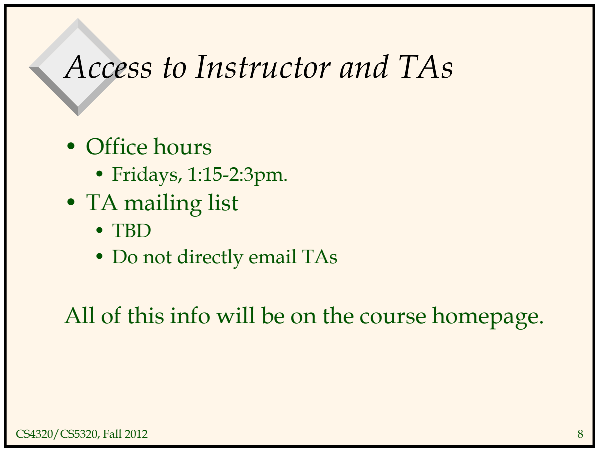
Prerequisites

- Courses
  - CS2110 (Computers and Programming)
  - CS3110 (Structure and Interpretation of Computer Programs)



People

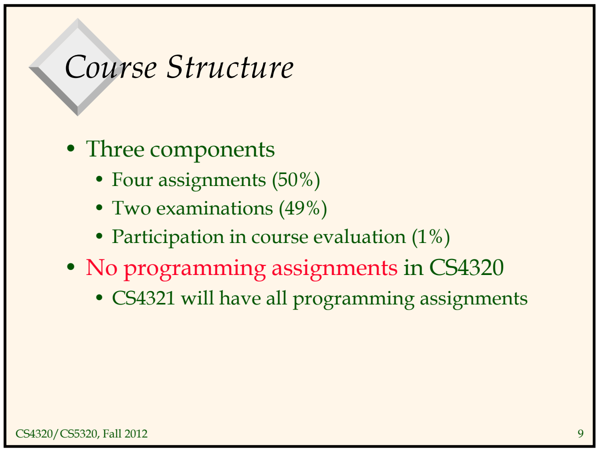
- Instructor
  - Johannes Gehrke
- TAs
  - TBD



Access to Instructor and TAs

- Office hours
  - Fridays, 1:15-2:3pm.
- TA mailing list
  - TBD
  - Do not directly email TAs

All of this info will be on the course homepage.



Course Structure

- Three components
  - Four assignments (50%)
  - Two examinations (49%)
  - Participation in course evaluation (1%)
- No programming assignments in CS4320
  - CS4321 will have all programming assignments

Class Lectures

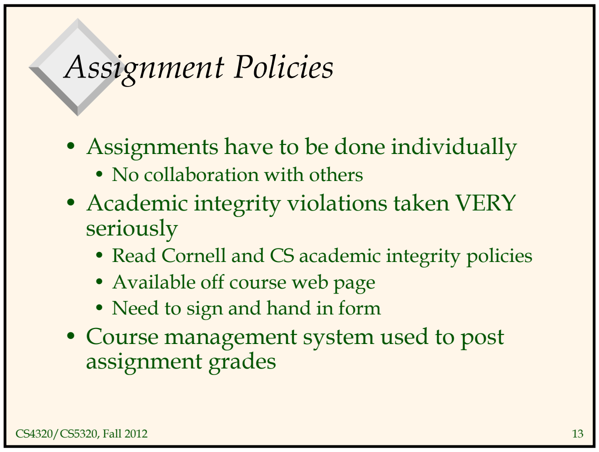
- Textbook: “Database Management Systems” (3rd Edition)
  - By Raghu Ramakrishnan and Johannes Gehrke
  - Required textbook
- Syllabus
  - Defined by class lectures, will be online in CMS
  - Not defined by textbook

Grading

- Three components
  - Assignments (50%)
  - Exams (49%)
  - Course evaluation (1%)

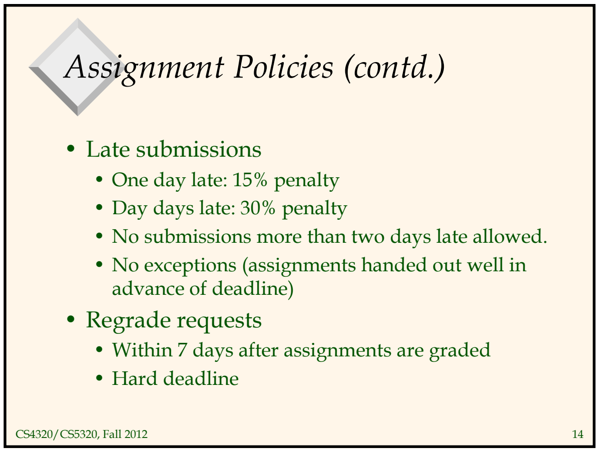
Assignments

- Four assignments
- Each assignment worth 12.5% of total grade



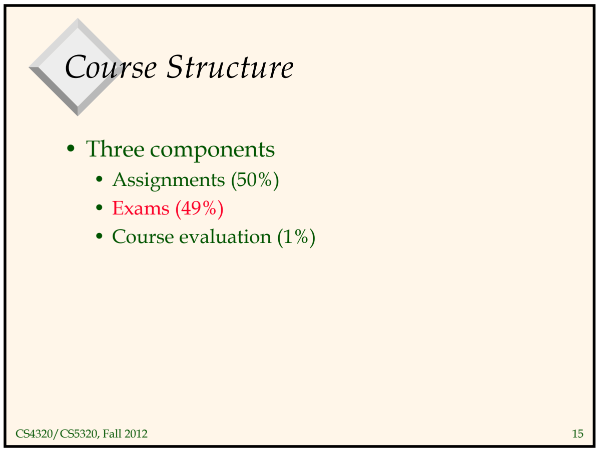
Assignment Policies

- Assignments have to be done individually
  - No collaboration with others
- Academic integrity violations taken VERY seriously
  - Read Cornell and CS academic integrity policies
  - Available off course web page
  - Need to sign and hand in form
- Course management system used to post assignment grades



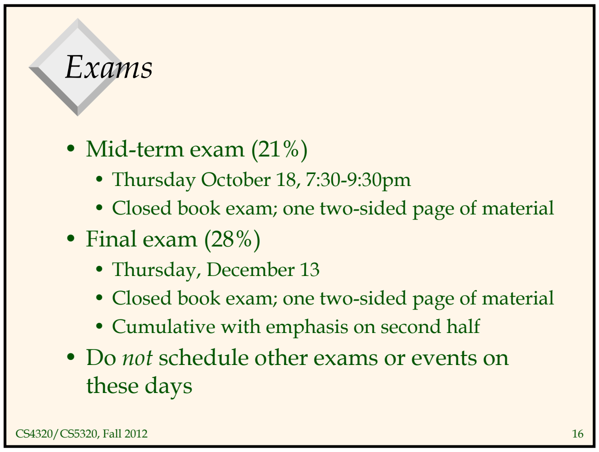
Assignment Policies (contd.)

- Late submissions
  - One day late: 15% penalty
  - Day days late: 30% penalty
  - No submissions more than two days late allowed.
  - No exceptions (assignments handed out well in advance of deadline)
- Regrade requests
  - Within 7 days after assignments are graded
  - Hard deadline



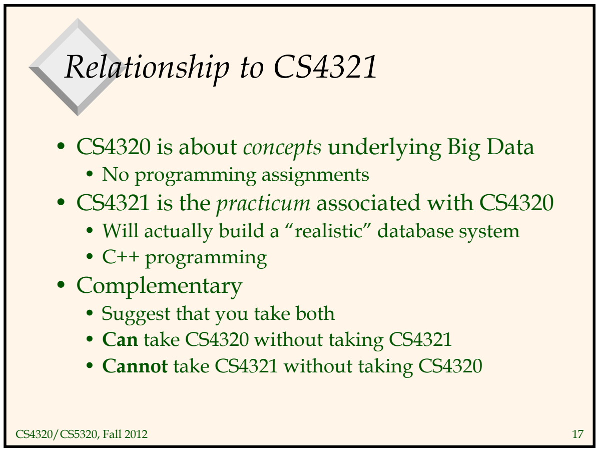
Course Structure

- Three components
  - Assignments (50%)
  - Exams (49%)
  - Course evaluation (1%)



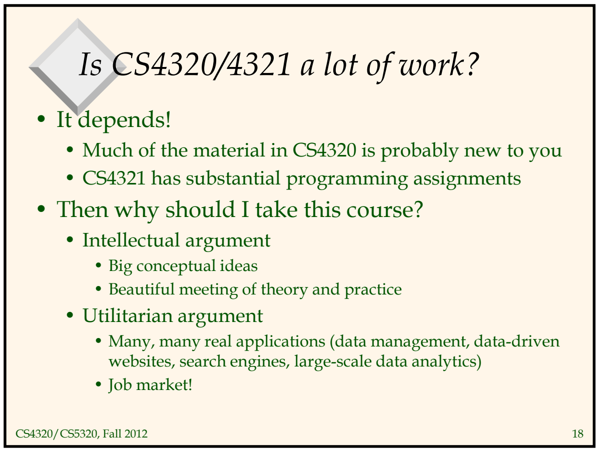
Exams

- Mid-term exam (21%)
  - Thursday October 18, 7:30-9:30pm
  - Closed book exam; one two-sided page of material
- Final exam (28%)
  - Thursday, December 13
  - Closed book exam; one two-sided page of material
  - Cumulative with emphasis on second half
- Do *not* schedule other exams or events on these days



Relationship to CS4321

- CS4320 is about *concepts* underlying Big Data
  - No programming assignments
- CS4321 is the *practicum* associated with CS4320
  - Will actually build a “realistic” database system
  - C++ programming
- Complementary
  - Suggest that you take both
  - **Can** take CS4320 without taking CS4321
  - **Cannot** take CS4321 without taking CS4320



Is CS4320/4321 a lot of work?

- It depends!
  - Much of the material in CS4320 is probably new to you
  - CS4321 has substantial programming assignments
- Then why should I take this course?
  - Intellectual argument
    - Big conceptual ideas
    - Beautiful meeting of theory and practice
  - Utilitarian argument
    - Many, many real applications (data management, data-driven websites, search engines, large-scale data analytics)
    - Job market!

CS5300: Architecture of Large-Scale Information Systems

- How do you build e-commerce websites such as amazon.com?
- How do you build a reliable web service that scales to millions of users?

CS5300: Architecture of Large-Scale Information Systems

- Underlying theme: How do I build *applications* on top of a database system?
- Will combine coverage of fundamental concepts with “hands-on” experience on Amazon EC2
- Prerequisite: CS4320

CS5300: Material Covered

- Three-tier architectures
- Edge caches
- Distributed transaction management
- Web services
- Content management

Instructor

Personal:

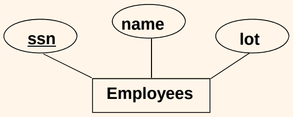
- Ph.D. from U of Wisconsin-Madison (CS, marketing) in 1999; joined Cornell right afterwards
- Chief Scientist at Fast Search and Transfer; acquired by Microsoft in 2008
- Technical advisor to Microsoft and other companies, consulting in Big Data

Research:

- Big Data Infrastructure
- Big Data Analytics

The Entity-Relationship Model

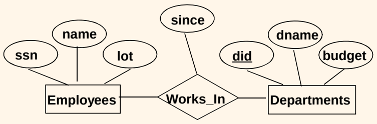
Entities



ER Model Basics

- *Entity*: Real-world object distinguishable from other objects. An entity is described (in DB) using a set of *attributes*
- *Entity Set*: A collection of similar entities. E.g., all employees
  - All entities in an entity set have the same set of attributes
  - Each entity set has a *key*
  - Each attribute has a *domain*

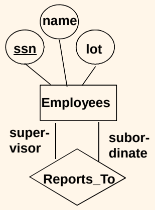
Relationships



ER Model Basics (Contd.)

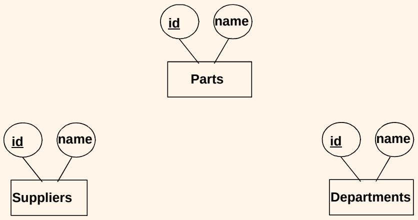
- *Relationship*: Association among two or more entities.
  - E.g., Attishoo works in Pharmacy department.
- *Relationship Set*: Collection of similar relationships.
  - An n-ary relationship set R relates n entity sets E1 ... En
  - Each relationship in R involves entities e1 in E1, ..., en in En

Relationships (Contd.)



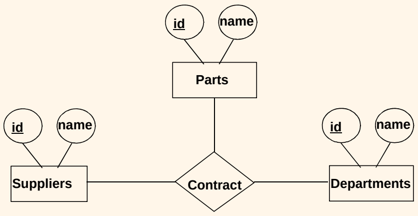
- Want to capture supervisor-subordinate relationship

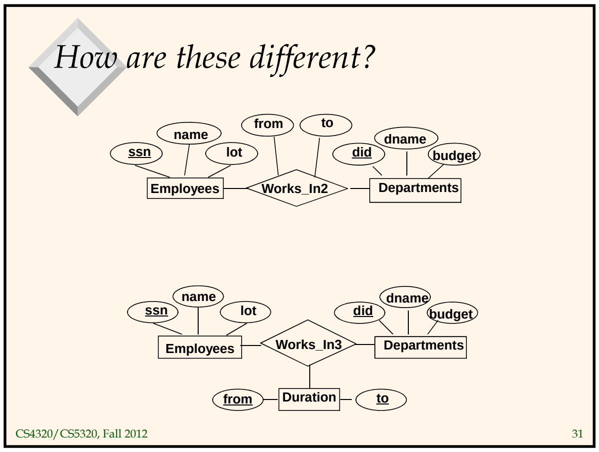
Relationships (Contd.)



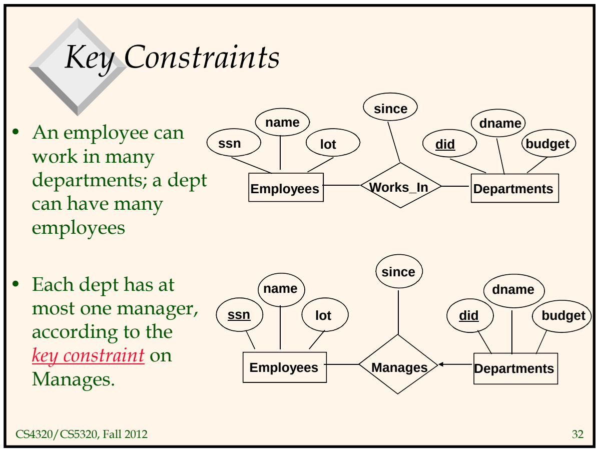
- Want to capture information that a Supplier s supplies Part p to Department d

Ternary Relationship



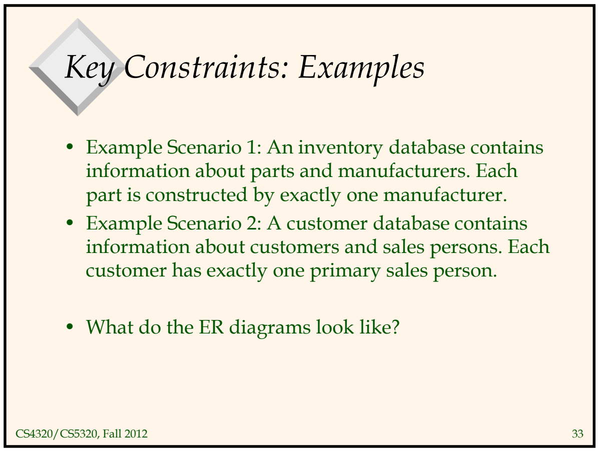


How are these different?



Key Constraints

- An employee can work in many departments; a dept can have many employees
- Each dept has at most one manager, according to the *key constraint* on Manages.

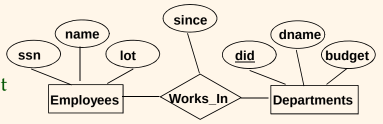


Key Constraints: Examples

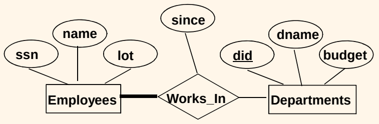
- Example Scenario 1: An inventory database contains information about parts and manufacturers. Each part is constructed by exactly one manufacturer.
- Example Scenario 2: A customer database contains information about customers and sales persons. Each customer has exactly one primary sales person.
- What do the ER diagrams look like?

Participation Constraints

- An employee can work in many departments; a dept can have many employees



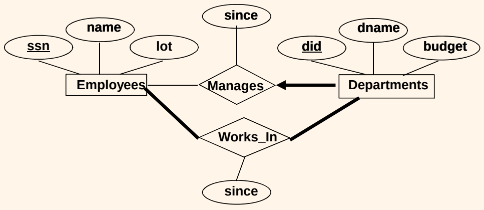
- Each employee works in at least one department according to the *participation constraint* on Works_In



Participation Constraints: Examples

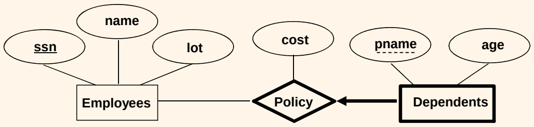
- Example Scenario 1 (Contd.): Each part is constructed by exactly one or more manufacturer.
- Example Scenario 2: Each customer has exactly one primary sales person.

What does this mean?



Weak Entities

- A *weak entity* can be identified uniquely only by considering the primary key of another (*owner*) entity.
  - Owner entity set and weak entity set must participate in a one-to-many relationship set (one owner, many weak entities).
  - Weak entity set must have total participation in this *identifying* relationship set.



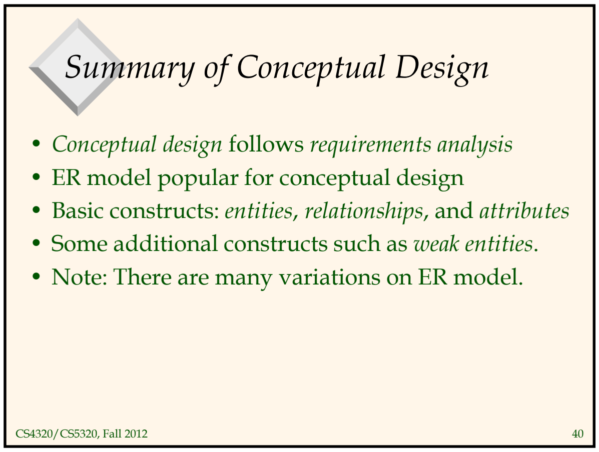
Exercise

- Give two real-life examples where each of the following would occur:
  - A key constraint
  - A participation constraint
  - A weak entity set

ER Modeling: Case Study

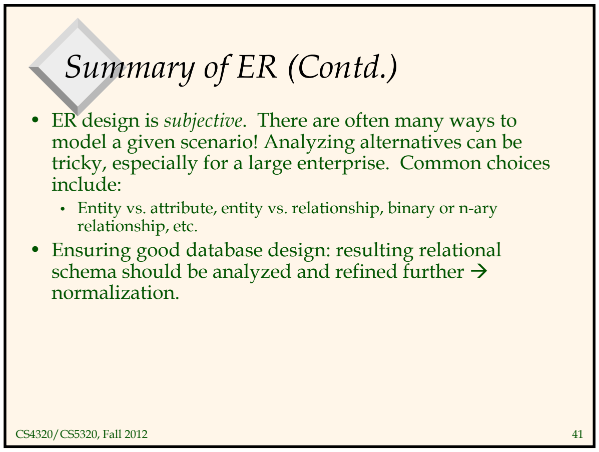
Drugwarehouse.com has offered you a free life-time supply of prescription drugs (no questions asked) if you design its database schema. Given the rising cost of health care, you agree. Here is the information that you gathered:

- Patients are identified by their SSN, and we also store their names and age.
- Doctors are identified by their SSN, and we also store their names and specialty.
- Each patient has one primary care physician, and we want to know since when the patient has been with her primary care physician.
- Each doctor has at least one patient.



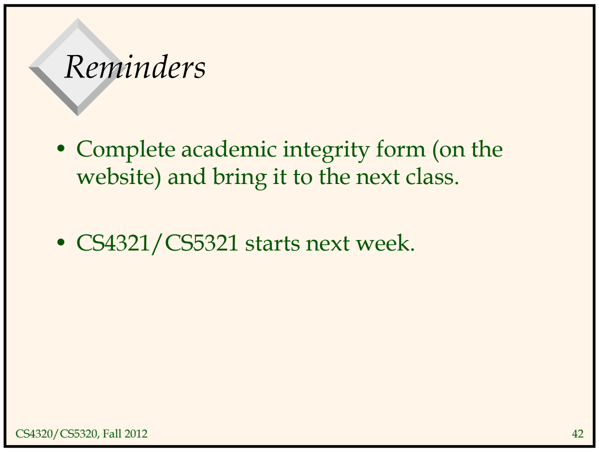
Summary of Conceptual Design

- *Conceptual design* follows *requirements analysis*
- ER model popular for conceptual design
- Basic constructs: *entities*, *relationships*, and *attributes*
- Some additional constructs such as *weak entities*.
- Note: There are many variations on ER model.



Summary of ER (Contd.)

- ER design is *subjective*. There are often many ways to model a given scenario! Analyzing alternatives can be tricky, especially for a large enterprise. Common choices include:
  - Entity vs. attribute, entity vs. relationship, binary or n-ary relationship, etc.
- Ensuring good database design: resulting relational schema should be analyzed and refined further → normalization.



Reminders

- Complete academic integrity form (on the website) and bring it to the next class.
- CS4321/CS5321 starts next week.