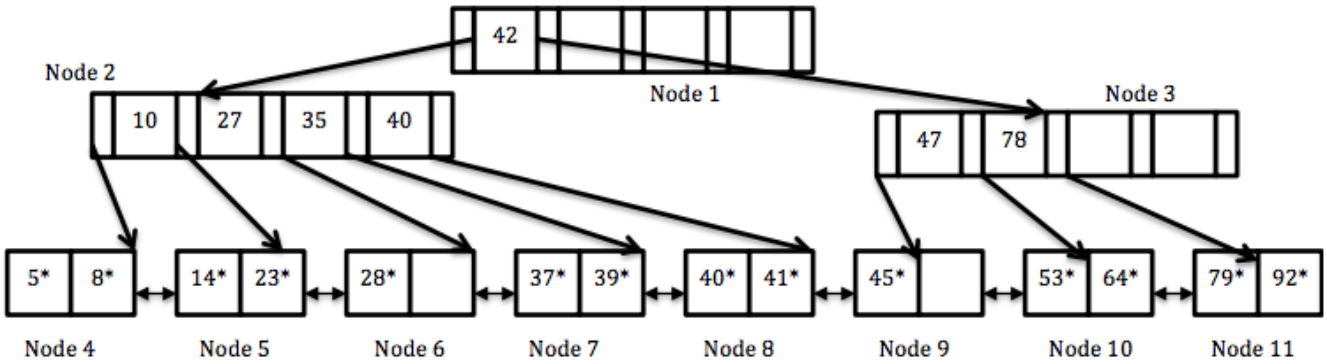
Midterm 2, CS186

Prof. Hellerstein, Spring 2012

Please put all answers on the answer sheets.
Make sure your name is on each answer sheet!

B+-Trees

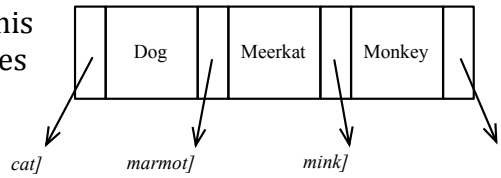
Assume we have the following B+ Tree of order 2 in which each *index* node must have between 2 and 4 entries. The *leaf* nodes can only hold only 2 entries, however. The asterisk on data entries represents the fact that the leaf pages will store additional information.



1. [2 points] On the answer sheet, list all nodes that are accessed when you perform a lookup on the range "37-65".

2. [4 points] A row is inserted into your table with the indexed value "19". We want you to **draw the B+ tree after this insert** occurs. We have started drawing it for you **on the answer sheet**. For any node X that does not change, simply write down "Node X" (as we did for Node 4). For any node that you do change, please fill it in completely. You do not need to assign node numbers to any nodes you generate.

3. [1 point] Consider the index page on the right. This is taken from a different B+-tree, which can fit 32 bytes per index page. In this index, each character takes one byte and each pointer takes 4 bytes.



In the diagram **on the answer sheet, draw an alternative index page with prefix-compressed index entries.** *Where relevant, we have shown the top of the value-range stored in the left-hand subtree.*

4. [2 points] Given your answer to part 3, **write on the answer sheet**:
a) the number of bytes occupied on your compressed page, and
b) the maximum number of additional index entries we could fit in that page.

Information Retrieval

5. **[2 points]** Consider the keyword query "Berkeley Database Research". We said in class that one could treat that query as a document, and compute a TF×IDF score for it. Now it is time to do so! Suppose we know that:
   - **a.** our corpus has 10^9 documents.
   - **b.** "berkeley" appears in 10^5 documents
   - **c.** "database" appears in 10^4 documents
   - **d.** "research" appears in 10^3 documents

   To help create the vector, **on the answer sheet please write down the (unnormalized!) TF×IDF score** for each of the 3 terms above.

6. **[2 points]** Suppose you have a relation stored as a heap file F representing an inverted file with the following schema:

```
(docId,  // an integer  ID representing this document
 termID,  // an integer ID representing this term
 position,  // the word offset of this term occurrence in the doc
 surrounding_text,  // a snippet of text containing this term occurrence
 DocTermRank  // the TFxIDF score).
```

   You will build a standard B+-tree index with "posting list" entries in the leaves as we discussed in class. The search key of the index will contain all attributes *except* for `surrounding_text`. In order for the index to be clustered, what should be the (lexicographic) sort order of the file F? **Write the answer on the answer sheet.**

7. **[2 points]** Your mom tells you that the schema in part (2) is not in BCNF. What implicit functional dependency makes her say that? **Write your answer on the answer sheet.**

8. **[2 points]** In class we discussed the idea of a "spider trap" web service that could foil a web crawler by generating interlinked pages dynamically. For each of the following arguments, **mark T (true) or F (false) on the answer sheet**:

   a. The spider trap can change the IDF scores of other pages.
   b. Spider traps can increase the search frontier of the crawler.
   c. The spider trap can attract lots of search hits by generating pages with arbitrarily high IDF.

Buffer Management

It's time to play buffer manager! For each of the following replacement policies, record the eviction for each access number (if any). Assume a **three page buffer pool** and remember to only record **evictions** in the frame number they occur.

For example, if record Q is evicted during access number 5 and record R is evicted in access 10, put Q in the box marked 5 and R in the box marked 10. **Do not** place Q in box 1 and R in box 2; you need to record the access number of the eviction. Some boxes may be left blank.

Access #	1	2	3	4	5	6	7	8	9	10
Record Accessed	D	B	S	A	B	C	D	S	B	S

9. [2.5 points] Most Recently Used. (**Fill in the eviction history on answer sheet**).

10. [2.5 points] Clock. (**Fill in the eviction history on answer sheet**).

Notes on 10: As in Bunny #5…assume a maximum reference count of one. Do not advance the clock hand when you access a page—only advance it when choosing a page to evict. Do **not** advance the clock hand after you replace a page, and assume each page is unpinned before the next access #.

For the next 2 questions, **mark T (true) or F (false) on the answer sheet**:

11. [1 point] Pages evicted from the buffer pool must be written back to disk.

12. [1 point] Sequential flooding occurs when **both** of the following hold:

a. repeated sequential scans of a file and

b. the number of unpinned buffer frames exceeds the number of file pages

Functional Dependencies.

Consider the relation: A B C D E
with the functional dependencies F= {A → BC, B → A, DE → C, C → B, E → D }

13. [2 points] There is a unique minimal superkey. What is it?
Write your answer on the answer sheet.

For the next 2 questions, **mark T (true) or F (false) on the answer sheet**:

14. [2 points] Consider the decomposition of the relation into two tables: ABED, EC

a) **True/False:** This decomposition is in BCNF.

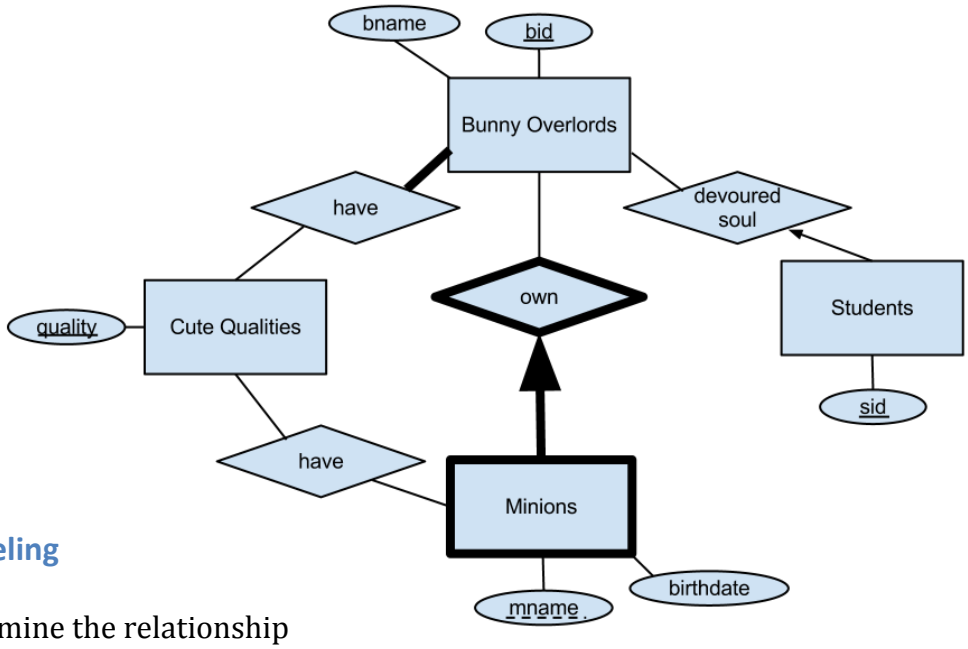
b) **True/False:** The decomposition is Lossless Join.

15. [2 points] Consider the decomposition of the relation into two tables: ACDE, CB

a) **True/False:** The decomposition is in BCNF.

b) **True/False:** The decomposition is Lossless Join.

16. [2 points] Perform a BCNF decomposition of the relational A B C D E, considering the functional dependencies in F **from left to right**. **Show the final decomposed schema on the answer sheet.**



ER Modeling

Let's examine the relationship between *Bunny Overlords*, *Students*, their *Minions* and their *Cute Qualities*. Here are the basics of the relationship.

a) Bunny Overlords devour the souls of Students.
b) Bunny Overlords own Minions, who do their bidding.
c) Bunny Overlords have cute qualities (such as fluffiness or cuddliness).
d) Minions also have cute qualities.

The ER-diagram above specifies the exact nature of the relationships. Please answer the following questions using the ER diagram and clarifications above.

For the next 2 questions, **mark T (true) or F (false) on the answer sheet**:

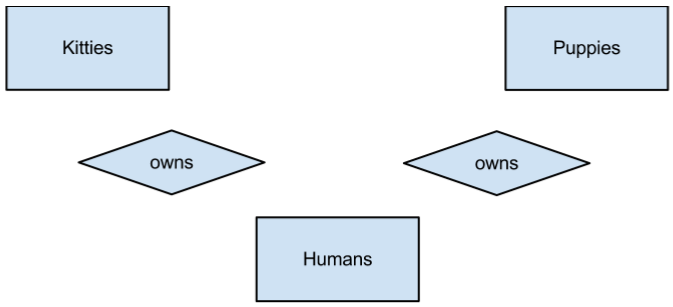
17. **[1 point] True / False**. A student may have their soul devoured by more than one Bunny Overlord.

18. **[1 point] True / False**. A particular Bunny Overlord cannot have two minions with the same name.

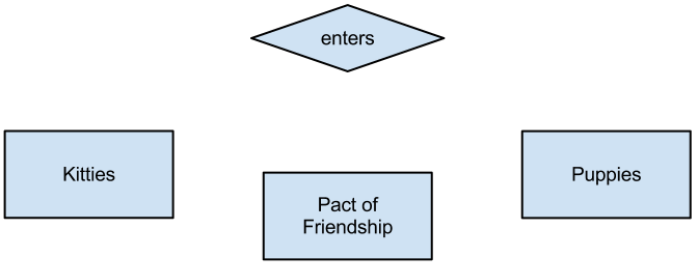
19. **[1 point] Multiple Choice**. Consider a Bunny Overlord, Mipsy. Mipsy has the fewest possible cute qualities that a Bunny Overlord can have according to the ER diagram. Mark the most general correct answer on the answer sheet.
    a. Mipsy must have more cute qualities than some Minion.
    b. Mipsy must have more cute qualities than all Minions.
    c. Mipsy must have fewer cute qualities than some Minion.
    d. Mipsy must have fewer cute qualities than all Minions.
    e. None of the above.

Now it's your turn to design your own ER diagram, modeling the following relationship between *Kitties*, *Puppies*, *Humans*, and *Pacts of Friendship*. For this part, we recommend you draw the solution out in the diagrams provided. **In question 20 you will copy the whole thing over to the answer sheet**. Be sure to clearly distinguish normal and bold lines/arrows!

- A *human* may **own** at most one *kitty* and one *puppy* (he or she may have one of each, though). All *Kitties* and *Puppies* are **owned** by exactly one *human.*



- *Kitties* and *Puppies* may **enter** into *Pacts of Friendship*. A *kitty* or *puppy* may participate in multiple *Pacts of Friendship*, but each *Pact of Friendship* involves exactly one kitty and one puppy.



- A *Pact of Friendship* may **eternally hate** any number of humans (or none). A single *human*, however, can only be hated by up to one *Pact of Friendship*. We recommend that you draw this relationship here.

20. **[3 points] Copy the answers** from the above diagrams into the complete ER diagram **on the answer sheet.**

21. **[1 point]** Now **add the following attributes** to the answer sheet's ER diagram.
    - *Kitties*, *Puppies*, and *Humans* each have a "name".
    - *Pacts of Friendship* have a "date" for when they were formed.
    - *Humans* have "nicknames" that are assigned by *Pacts of Friendship* when they **eternally hate** a *Human,* so they can talk about the hated human without him/her knowing.

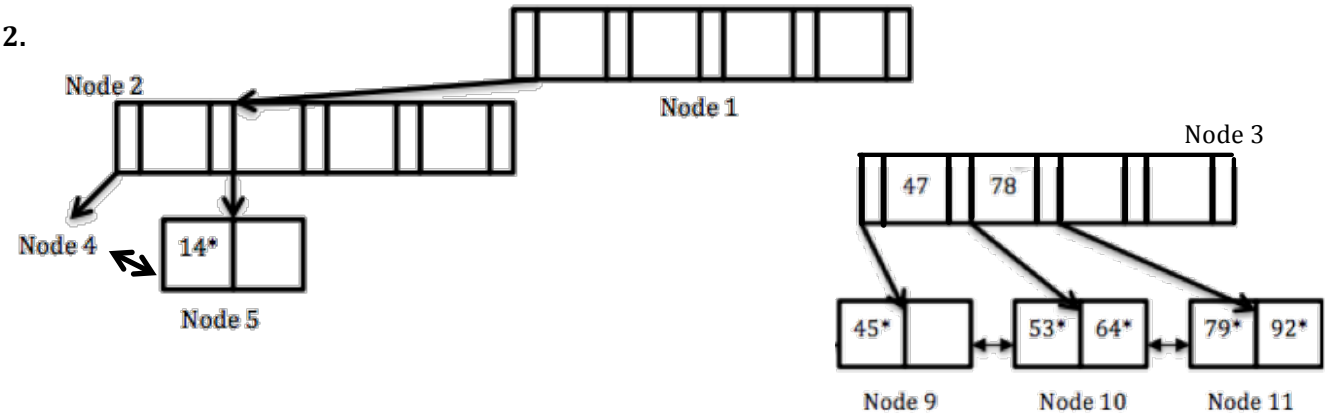
Make sure you've copied all answers over to the answer sheet!

Name:	Class Account: cs186-_____

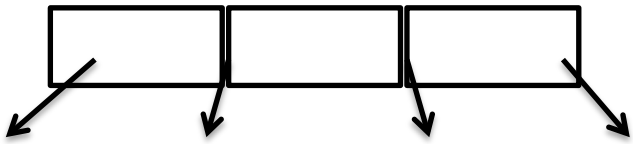
B+-Trees

1. List of nodes accessed:

2.



3.



4a . # bytes occupied:

4b. # additional entries:

Information Retrieval

5. berkeley: **database:** **research:**

6. Sort order is:

7. The FD is:

8. a ○ T ○ F **b** ○ T ○ F **c** ○ T ○ F

Name:	Class Account: cs186-_____

Buffer Management

9. MRU eviction history

Access #	1	2	3	4	5	6	7	8	9	10
Record evicted										

10. Clock eviction history

Access #	1	2	3	4	5	6	7	8	9	10
Record evicted										

11. ○ T ○ F **12.** ○ T ○ F

Functional Dependencies and Normal Forms

13 Superkey:

14a. ○ T ○ F **14b.** ○ T ○ F **15a.** ○ T ○ F **15b** ○ T ○ F

16. Decomposed schema:

ER Modeling

17. ○ T ○ F

18. ○ T ○ F

19. Pick one (a-e): _____

20, 21: