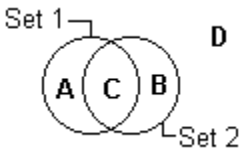
Venn Diagrams and Boolean Operations

Historical Notes:
George Boole and John Venn were 19th century mathematicians. George Boole developed what became known as Boolean algebra or Boolean logic. Boole's work became important when applied to electronic logic circuits in the late 1930s. John Venn is best known for his circle diagrams representing the relationships between sets.

Key Points:

- There are 3 Boolean operators: AND, OR, and NOT.
  - Some systems use AND NOT in place of NOT.
- The Boolean operators are used to combine search terms.
- In search logic, Boolean operators act on sets.
  - In bibliographic database searching the sets are groups of records containing a particular word.
- Complex search statements may include a mix of Boolean operators.

This Venn diagram shows all the possible relationships for records in a database with respect to two sets:



A - the records found exclusively in set 1.
B - the records found exclusively in set 2.
C - the records found in both sets 1 and 2.
D - the records not found in either set.

The Boolean operators allow the searcher to specify the desired combination or combinations of the sets: the intersection of sets (C), the union of sets (A,B,C), etc.

Does computer searching really involve set manipulation?
Yes! You might want to review the section on search mechanics before you proceed with the following example.

Imagine two sets of numbers

- Set A: 1,4,7,10,13,16,19,22,25,28,31,33,36,39,42,45,48,51
- Set B: 2,3,5,7,11,13,17,19,23,29,31,37,41,43,47,53,59,61

How does the computer process the lists given the following Boolean operation:

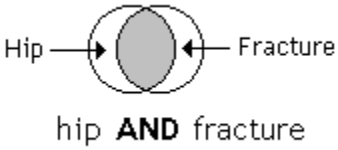
Set A ***AND*** Set B

to create the resulting set, Set C?

Examine *Set A* and *Set B* to find the common numbers, ***then view the illustration***.
At this point you should appreciate that the computer is just crunching numbers when processing search statements!

The Boolean AND operator

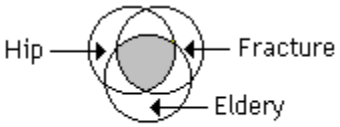
Venn Diagram
(The gray shading represents the outcome of the Boolean operation)



hip **AND** fracture

Description

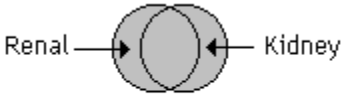
- When terms are combined with the **AND** operator, retrieved records must contain all terms.
- In most search systems, the terms may occur anywhere in the record unless explicitly restricted to one or more fields.
- In most search systems sets of terms may be combined in addition to single terms.
- Multiple AND operators may be used to combine more than two terms. Only records containing all the terms will be retrieved.



hip **AND** fracture **AND** elderly

The Boolean OR operator

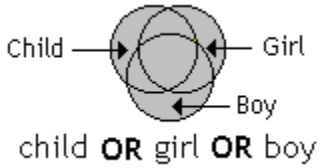
Venn Diagram
(The gray shading represents the outcome of the Boolean operation)



renal **OR** kidney

Description

- When terms are combined with the **OR** operator, retrieved records may contain one or more of the search terms.
- The terms may occur anywhere in the record unless restricted by the IN operator. Sets of terms may be combined in addition to single terms.
- Multiple OR operators may be used to combine more than two terms. Records must contain at least one of the terms to be retrieved.



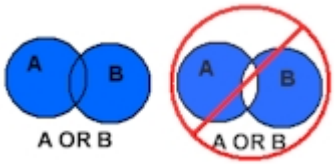
Notes:

1. Rather than combining variations of the same term with the OR operator, consider using truncation or wildcard characters instead:

mouthwash* *in place of*
mouthwash OR mouthwashes OR mouthwashing

2. OR is all inclusive, not exclusive.

- Don't think of it as retrieving what is in one set or the other but not both.
- For example, a search for the terms "Boys OR Girls" will retrieve all of these titles, including 1, 3, and 6 which contain both terms.

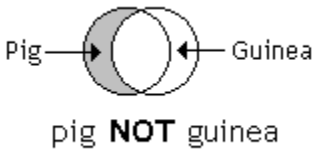


1. Mediated Activity in the Primary Classroom: **Girls**, **Boys** and Computers.
2. Fifth- and Seventh-Grade **Girls**' Decisions about Participation in Physical Activity.
3. Reading Disability in **Boys** and **Girls**: No Evidence for a Differential Genetic Etiology.
4. Learning and Motivational Characteristics of **Boys** with AD/HD and/or Giftedness.
5. Raising **Boys**' Attainment in Reading: Some Principles for Intervention.
6. Schooling of **Girls** and **Boys** in A West African Country: The Effects of Parental Education, Income, and Household Structure.
7. Supporting Communication of **Girls** with Rett Syndrome and Their Mothers in Storybook Reading.

- Furthermore ...
  A search for just the term "***girls***" would retrieve **1, 2, 3, 6, 7**.
  A search for just the term "***boys***" would retrieve **1, 3, 4, 5, 6**.

The Boolean NOT operator

Venn Diagram
(The gray shading represents the outcome of the Boolean operation)



Description

- **NOT** finds records containing one term but not another.
- The terms may occur anywhere in the record unless restricted by the IN operator. Sets of terms may be combined in addition to single terms.

Notes:

1. The NOT operator excludes some records which contain the desired search term.
    - Don't use NOT unless you are confident the excluded term always results in the retrieval of irrelevant records.
    - For example: If the topic being searched may yield articles about infants and children, but the searcher isn't interested in the literature on infants, he should resist the temptation to search *children NOT infants*. Some papers may focus on the subject in children but note that the subject also occurs in infants. Papers of this sort could be eliminated by this use of the NOT operator.
    - The following record from the Medline database illustrates this problem. The paper is about children, but the author uses the word "infants" when mentioning other studies:

TI: Crossing the midline: a study of four-year-old **children**.
AU: Screws-DP; Eason-BL; Surburg-PR
SO: Percept-Mot-Skills. 1998 Feb; 86(1): 201-3
AB: Midline crossing refers to behavior that results in reaching, stepping, or looking, across the body's midline. Several studies have indicated that *infants*, young **children**, and individuals with disability make more errors on midline-crossing tasks than on similar tasks placed at the ipsilateral side. Until recently, assessment of midline crossing has used a spatial protocol and has been criticized for not having a temporal component. The purpose of this study was to assess midline crossing by 9 4-yr.-old **children** within an information processing context. Analysis indicated that contralateral tasks required more processing time than similar tasks placed ipsilaterally.

2. A common use of NOT is to remove duplication between sets. The purpose in this case is to avoid printing or viewing duplicate records:

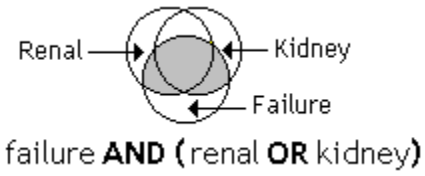
Set	Records	Search Statement
#1	16	ochratoxin **and** coffee
#2	43	ochratoxin **and (**cereal **or** cereals **or** grain **or** grains**)**
#3	32	#2 **NOT** #1

In this example the searcher knew that some of the records he had already seen in set #1 would also occur in set #2. He eliminated those records from set #2 using the **NOT** operator.

Mixed Boolean operations

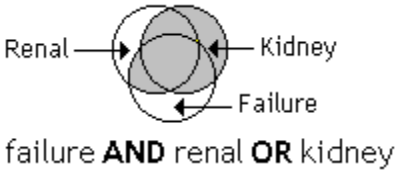
Mixing operators is allowed in most search systems. Since the order of operation may vary from system to system, **u**se **parentheses to specify the proper combination**. The order in which operations are processed varies between systems. For example, one system may process search statements from left to right while another system may process ANDs before ORs.

In any system regardless of the order of operation, this strategy will retrieve records containing the pair of words *renal & failure*, or the pair of words *kidney & failure*, or all three terms.



In a system that processes ANDs before ORs, or simply processes statements from left to right the same search but without the parentheses will yield very different results because the AND operation is done first.

This strategy will retrieve records containing either the pair of words *renal & failure* or any record with the word *kidney*. If you exam the titles below, which meet the criteria for retrieval, you will see that those that have the word kidney are not necessarily about kidney failure:



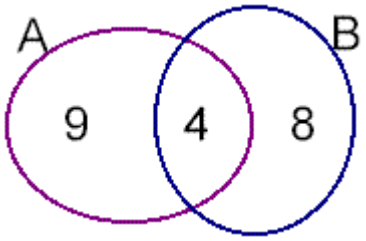
TI: Differential gene expression in nickel(II)-treated normal rat **kidney** cells.
TI: Interleukin-2 regulatory effect on P-selectin and interleukin-8 production in patients with chronic **renal failure**.
TI: Chronic **kidney** disease prevalence and rate of diagnosis.
TI: Early diagnosis of acute **kidney** injury.
TI: Endovascular aneurysm repair in a patient with a horseshoe **kidney** and impaired **renal** function.
TI: Hereditary hypophosphatemias: new genes in the bone-**kidney** axis.
TI: Fatal right-sided endocarditis due to Aspergillus in a **kidney** transplant recipient.
TI: **Kidney** cancer: energy ablation.
TI: Comprehensive molecular diagnostics in autosomal dominant polycystic **kidney** disease.
TI: Effect of age, gender, and diabetes on excess death in end-stage **renal failure**.
TI: Oxalosis presenting as early **renal** allograft **failure**.
TI: Carbamoylation of glomerular and tubular proteins in patients with **kidney failure**: a potential mechanism of ongoing **renal** damage.

Boolean "Arithmetic"

Look at the following Venn diagrams. The number of items in each discrete **area** *(not in each set)* of the diagram is given.

In each case, do you understand why the number of items resulting from the OR combination is what it is?

Diagram 1

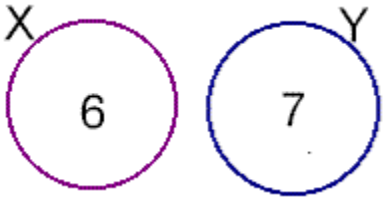


Set **A** contains **13** items.
Set **B** contains **12** items

The total number of items in set A OR set B is 21.

Because the result of the OR operation is not 25 items (13+12), but 21, there must be 4 items (25-21) which occur in both sets.

Diagram 2

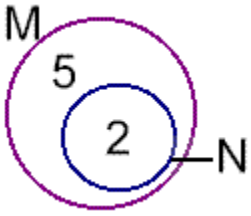


Set **X** contains **6** items.
Set **Y** contains **7** items

The total number of items in set X OR set Y is 13.

Because the result of the OR operation is the same as the sum of the number of items in the two sets, the sets have no items in common.

Diagram 3



Set **M** contains **7** items.
Set **N** contains **2** items

The total number of items in set M OR set N is 7.

Because the result of the OR operation is the same as the number of items in the larger set, the smaller set must be a subset of the larger.

Practice:

Using the three diagrams above as models, draw a Venn diagram to represent the relationship between sets #1 and #2 in each of the following four searches. (*Calculate the sum of sets #1 and #2 in each case to help you determine the nature of the relationship.)*

Set #	Search Statement	Items
#1	salbutamol	411
#2	albuterol	234
#3	#1 OR #2	620

Set #	Search Statement	Items
#1	pennsylvania	351
#2	nebraska	113
#3	#1 OR #2	464

Set #	Search Statement	Items
#1	tylenol and headache	12
#2	acetaminophen and headache	109
#3	#1 OR #2	109

Set #	Search Statement	Items
#1	men	20,856
#2	women	31,369
#3	#1 OR #2	40,526