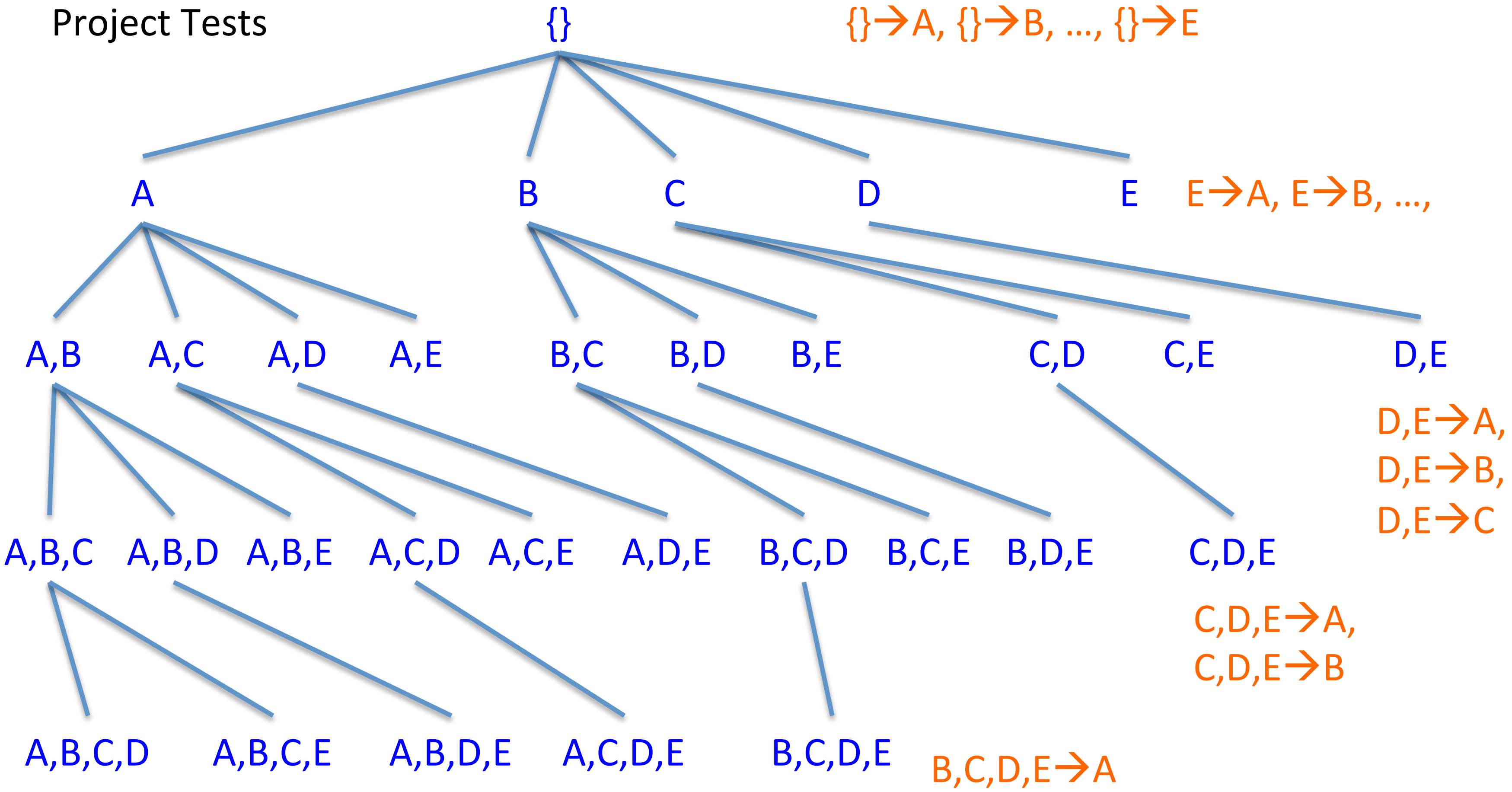
Data Mining
Constructing test sets for individual project

and

bettering your grade

Project Tests



{}

{}→A, {}→B, …, {}→E

A B C D E

E→A, E→B, …,

A,B A,C A,D A,E B,C B,D B,E C,D C,E D,E

D,E→A,
D,E→B,
D,E→C

A,B,C A,B,D A,B,E A,C,D A,C,E A,D,E B,C,D B,C,E B,D,E C,D,E

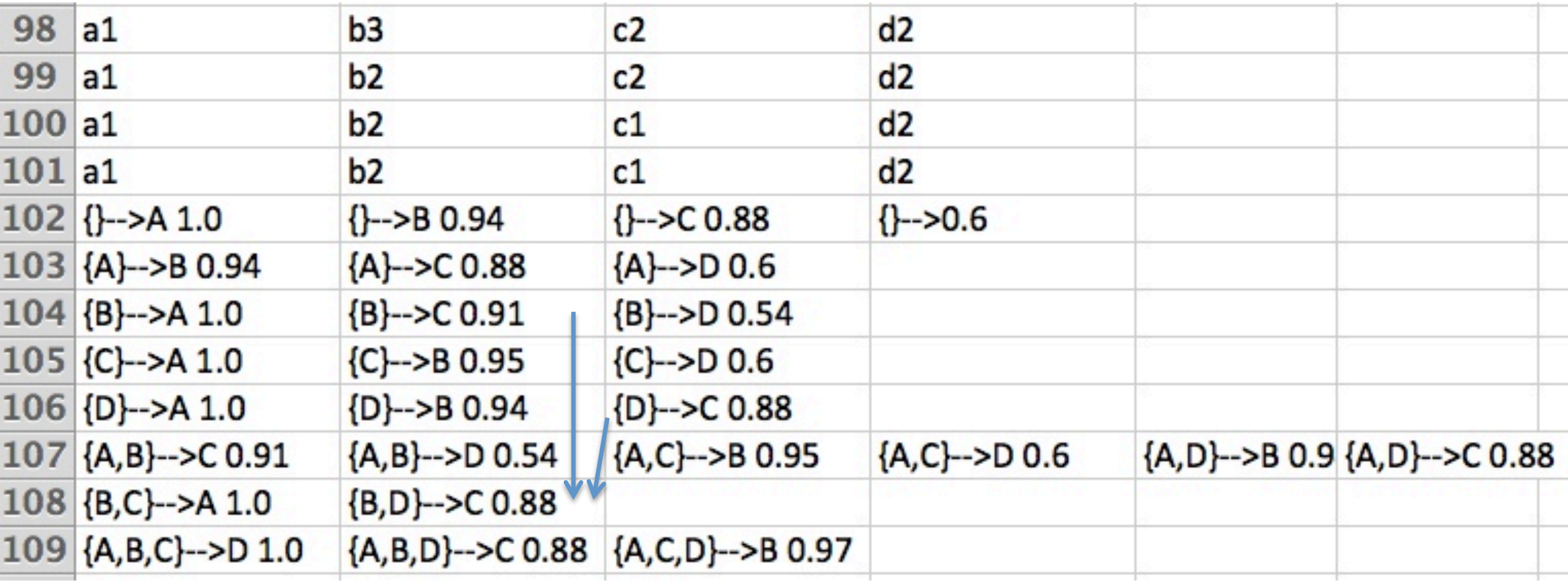
C,D,E→A,
C,D,E→B

A,B,C,D A,B,C,E A,B,D,E A,C,D,E B,C,D,E

B,C,D,E→A

Select approximate FDs that vary in domain cardinality and in degree of support X→Y, where

- $|X| = 0$, $|X| = 1$, $|X| = 2$, $|X| = 3$, … (Depth = 0, Depth = 1, Depth = 2, Depth = 3, …)
- Min-sup = 1.0, Min-sup >= 0.95, Min-sup >= 0.90, Min-sup = 0.85, …



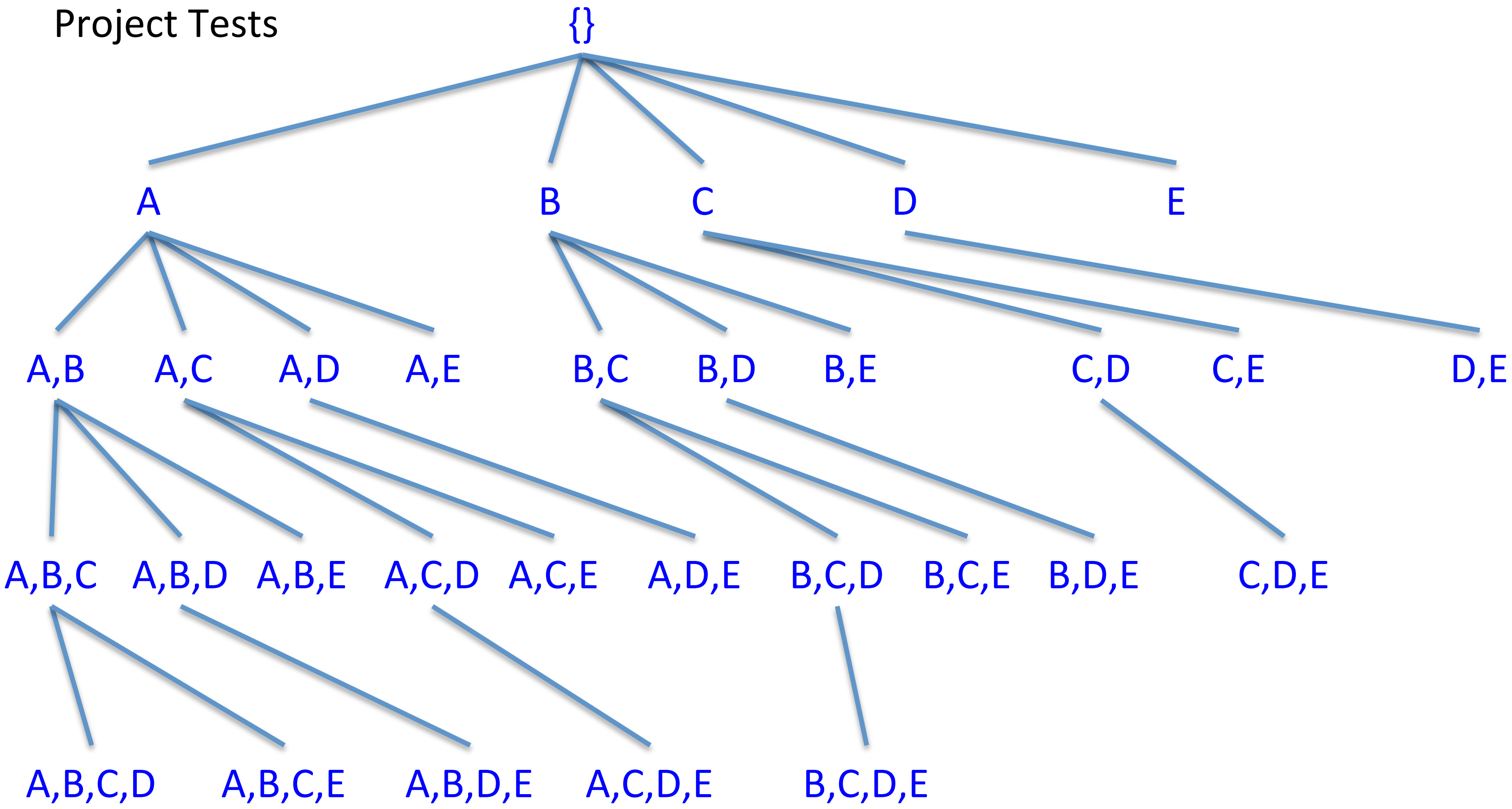
98	a1	b3	c2	d2		
99	a1	b2	c2	d2		
100	a1	b2	c1	d2		
101	a1	b2	c1	d2		
102	{}-->A 1.0	{}-->B 0.94	{}-->C 0.88	{}-->0.6		
103	{A}-->B 0.94	{A}-->C 0.88	{A}-->D 0.6			
104	{B}-->A 1.0	{B}-->C 0.91	{B}-->D 0.54			
105	{C}-->A 1.0	{C}-->B 0.95	{C}-->D 0.6			
106	{D}-->A 1.0	{D}-->B 0.94	{D}-->C 0.88			
107	{A,B}-->C 0.91	{A,B}-->D 0.54	{A,C}-->B 0.95	{A,C}-->D 0.6	{A,D}-->B 0.9	{A,D}-->C 0.88
108	{B,C}-->A 1.0	{B,D}-->C 0.88				
109	{A,B,C}-->D 1.0	{A,B,D}-->C 0.88	{A,C,D}-->B 0.97			

B→C 0.91 There are

- 86 rows with B=b1 and C=c1,
- 6 rows with B=b1 and C=c3,
- 2 rows with B=b1 and C=c2,
- 3 rows with B=b3 and C=c2,
- 2 rows with B=b2 and C=c1,
- 1 row with B=b2 and C=c2

(86+3+2)/100 = 0.91

Project Tests



{}

A B C D E

A,B A,C A,D A,E B,C B,D B,E C,D C,E D,E

A,B,C A,B,D A,B,E A,C,D A,C,E A,D,E B,C,D B,C,E B,D,E C,D,E

A,B,C,D A,B,C,E A,B,D,E A,C,D,E B,C,D,E

Select approximate FDs that vary in domain cardinality and in degree of support X→Y, where

- $|X| = 0$, $|X| = 1$, $|X| = 2$, $|X| = 3$, … (Depth = 0, Depth = 1, Depth = 2, Depth = 3, …)
- Min-sup = 1.0, Min-sup >= 0.95, Min-sup >= 0.90, Min-sup = 0.85, …

If you

- correctly implement find-approximate-functional-dependencies and all your auxiliary functions, for both on data sets you will be given ahead of time and those we use to grade;
- nicely format and comment your code with comprehensible and informative function header comments;

then you will receive an A- score (90%).

If, in addition,

- you implement some efficiency enhancement (such as pruning), and explain it clearly in comments at the top of the submission file (perhaps comparing runtime before and after enhancement); or
- simply instrument the code (top level function find_fds) and report runtime results as depth-limit varies for a fixed data set and minimal-support, and as minimal-support varies for a fixed data set and depth-limit; or
- Give a short write-up (e.g., one page) on results on an additional "real-world" data set, such as the "happiness" data set (already formatted for you +5%) or translate, test, and write up results with another real-world data set, such as Congressional Voting Records https://www.congress.gov/roll-call-votes, which has not been translated for you +10%);

then you can receive up to 100%.