Tree-based methods:

AID Morgan + Sonquist (1963)
Journal of Amer. Stat. Association
58, 415–434

Sonquist + Morgan (1964)
Monograph 35, ISR, U. of Michigan

THAID Messenger and Mandell (1972)
Journal of Amer. Stat. Association
67, 768–772

Morgan and Messenger (1973)
THAID, SRC-ISR, U. of Michigan

CHAID Kass, G. V. (1980),
Applied Statistics, 29, 119–127

CART Brieman, et al. (1984)
Classification and Regression Trees,
Wadsworth

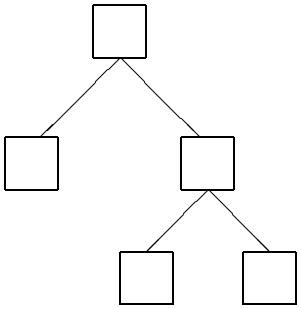
CHAID

- categorical response variable

  Credit Rating:

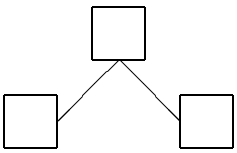
  "bad" "poor" "good" "very good"

- categorical explanatory variabales
- create a decision tree



Algorithm:

Dividing the cases that reach a certain node in the tree.



(Step 1) Cross tabulate the response variable with each of the explanatory variables.

	NT=0	NT≥1
Bad		
Poor		
Good		
V.Good		

	Fico < 700	700-750	Fico > 750
Bad			
Poor			
Good			
V.Good			

When there are more than two columns, find the "best" subtable formed by combining column categories.

(Step 2) This is applied to each table with more than 2 columns.

Compute Pearson X^2 tests for independence for each allowable subtable

Fico

	< 700	700-750	700-750	> 750
bad				
poor				
good				
v.good				

X_1^2 X_2^2

Look for the smallest X^2 value. If it is not significant, combine the column categories.

	< 750	> 750
bad		
poor		
good		
v.good		

Repeat step 2 if the new table has more than two columns

(Step 3) Allows categories combined at step 2 to be broken apart.

For each compound category consisting of at least 3 of the original categories,

- find the "most significant" binary split
- if X^2 is significant, implement the split and return to step 2.
- otherwise retain the compound categories for this variable, and move on to the next variable.

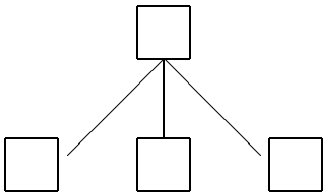
(Step 4) You have now completed the "optimal" combining of categories for each explanatory variable.

Find the "most significant of these "optimally" merged explanatory variables.

	C1+C2	C3	C4+C5+C6
bad			
poor			
good			
v.good			

Compute a "Bonferroni" adjusted chi-squared test of independence for the reduced table for each explanatory variable.

(Step 5) Use the "most significant" variable in step 4 to split the node with respect to the "merged" categories for that variable.



C1+C2 C3 C4+C5+C6

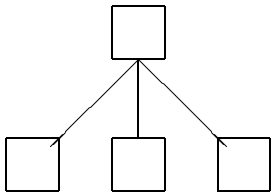
repeat steps 1-5 for each of the offspring nodes.

Stop if

- no variable is significant in step 4.
- the number of cases reaching a node is below a specified limit.

Summary:

- CHAID is an algorithm
- Must categorize every variable
  - ordinal variables
  - nominal variables
- At each node it tries to find
  - best explanatory variable
  - best merger of categories



Try to make the distributions of cases across the response categories as different as possible in the "offspring" nodes.

TREEDISC macro is SAS

- modified version of CHAID
- now part of the data mining package
- application to the Wisconsin Driver data
  - response: traffic violations in 1974
    (1) at least one
    (0) none
  - explanatory variables:
    sex
    age
    history of cardiovascular disease
    place of residence
- missing values are treated as another category


```
/*  This program uses the TREEDISC
    macro in SAS to apply a modified
    CHAID algorithm to the Wisconsin
    driver data. This code is stored
    in the file    chaidwis.sas   */

/*  Fisrt set some graphics options */
/*  To print postscipt files in UNIX */
/*
goptions cback=white ctext=black
     targetdevice=ps300 rotate=landscape;
*/
/* To print postscript files from Windows */
goptions cback=white ctext=black
     device=WIN target=ps
     rotate=landscape ;

DATA SET1;
  INFILE 'c:\courses\st557\sas\drivall.dat';
  INPUT AGE  SEX  D  V  R  X;
  LABEL    AGE = AGE GROUP
             D = DRIVER GROUP
             V = VIOLATION STATUS
```



```
             R = RESIDENTIAL AREA
             X = COUNT;
  run;

proc format; value sex 1 = 'Male'
                       2 = 'Female';
             value age 1 = '16-36'
                       2 = '36-55'
                       3 = 'over 55';
             value d   1 = 'Disease'
                       2 = 'Control';
             value v   1 = 'Some'
                       2 = 'None';
             value r   1 = '> 150000'
                       2 = '39-150000'
                       3 = '10-39000'
                       4 = '< 10000'
                       5 = 'rural';
run;

proc print data=set1;
  run;
```

```
/*  Load in the xmacros file */

%inc 'c:\courses\st557\sas\xmacro.sas';


/*  Load in the TREEDISC macro  */

%inc 'c:\courses\st557\sas\treedisc.sas';

/* Compute a tree for predicting
   violation status (V) from age, sex,
   disease stauts(D) and residence(R) */

%treedisc(data=set1, depvar=v, freq=x,
       ordinal=age: r:,nominal=d: sex:,
       outtree=trd, options=noformat,
       trace=long);

/*  Draw the tree on one page */

%treedisc(intree=trd, draw=graphics);
```



```
/*  Draw a larger tree on several
    pages */

goptions cback=white ctext=black
   device=WIN target=ps rotate=portrait;

%treedisc(intree=trd,
            draw=graphics, pos=90 120);
```

```
TREEDISC Analysis

Values of  AGE :  1  2  3

Values of  R :  1  2  3  4 5

Values of  D :  1  2

Values of  SEX :  1  2


Dependent variable (DV):  V

DV values:  1  2
```



```
Splits Considered for Node      1

Predictor       Type     Chi-Square Adjusted p

AGE             Ordinal       57.39     0.0001
SEX             Nominal       36.80     0.0001
D               Nominal        4.40     0.0359
R               Ordinal        2.53     0.4458


Best split:  AGE Ordinal with p = 0.0000

   New node: 3    AGE = 2  3
                  DV count:   147   1864

   New node: 2    AGE = 1
                  DV count:   133    656
```

Splits Considered for Node 2

Predictor	Type	Chi-Square	Adjusted p
SEX	Nominal	41.59	0.0001
D	Nominal	0.01	0.9193
R	Ordinal	0.15	0.9975

Best split: SEX Nominal with p = 0.0000

New node: 5 SEX = 1
DV count: 102 302

New node: 4 SEX = 2
DV count: 31 354

Splits Considered for Node 20

Predictor	Type	Chi-Square	Adjusted p
R	Ordinal	1.41	0.7031
D	Nominal	0.06	0.8101

Best split: R Ordinal with p = 0.7031
*** Reject split

TREEDISC Analysis of Dependent
Variable (DV) V

V value(s): 1 2
DV counts: 280 2520
Best p-value(s): 0.0001 0.0001

AGE value(s): 1
DV counts: 133 656
Best p-value(s): 0.0001 0.9193

SEX value(s): 2
DV counts: 31 354
Best p-value(s): 0.6064 0.8571

SEX value(s): 1
DV counts: 102 302
Best p-value(s): 0.7334 0.9703

AGE value(s): 2 3
DV counts: 147 1864
Best p-value(s): 0.0001 0.0221

SEX value(s): 2
DV counts: 20 563
Best p-value(s): 0.0856 0.5368

AGE value(s): 2
DV counts: 14 284
Best p-value(s): 0.8083 0.8990

AGE value(s): 3
DV counts: 6 279
Best p-value(s): 0.0264 0.1102

D value(s): 2
DV counts: 0 127

```
        D value(s): 1
        DV counts: 6  152
        Best p-value(s): 0.0592

            R value(s): 1
            DV counts: 3  22

            R value(s): 2  3  4
            DV counts: 1  111
            Best p-value(s): 0.5928

            R value(s): 5
            DV counts: 2  19

SEX value(s): 1
DV counts: 127  1301
Best p-value(s): 0.0232 0.1940

    AGE value(s): 2
    DV counts: 58  462
    Best p-value(s): 0.0215 0.7310
```



```
        D value(s): 2
        DV counts: 18  217
        Best p-value(s): 0.1317

        D value(s): 1
        DV counts: 40  245
        Best p-value(s): 0.3814

AGE value(s): 3
DV counts: 69  839
Best p-value(s): 0.0363 0.8254

    R value(s): 1
    DV counts: 20  139
    Best p-value(s): 0.8899

    R value(s): 2  3  4  5
    DV counts: 49  700
    Best p-value(s): 0.7031 0.8101
```