These notes come from

```
@article{Demsar06,
        author    = {Janez Dem\v{s}ar},
        title     = {Statistical Comparisons of Classifiers
                     over Multiple Data Sets},
        journal   = {J. Mach. Learn. Res.},
        volume    = {7},
        year      = {2006},
        issn      = {1533-7928},
        pages     = {1--30},
        publisher = {MIT Press},
        address   = {Cambridge, MA, USA},
        url     =
"http://iccle.googlecode.com/svn/trunk/share/pdf/demsar06.pdf"
 }
```

R*N >= 20
Often R=N=10 (for large data sets)
Can be R=10 N=3 for slow processes or small data sets

```
Criteria= {pd, pf, precision}
For d in  data sets
      Out = “d.dat”
      Remove Out (if it exists)
      For  r in R repeats
            Reandomized order of data
            Divide data N ways
            For n in N
                  Test = N[n]
                  Train = Data – Test
                  For l in Learners
                        B4= time()
                        Model=Learn(Train)
                        After =time()
                        For C in Criteria
                              Print d,r, n,l,c,Criteria(Model(Test)) >> Out
                              Print d,r, n,l,”runtime”, After – B4  >> Out
```

System issue: if this dies half way through, need to pick where left off.

Quartile charts for performance

```
For each criteria do
        For each learner do,
                Write down the min, 25%, 50%, 75%, max value (percentiles)
```

Chart it as follows: “|” is median, “+” is upper quartile “-“ is lower quartile

```
==========================================
Criteria X

----| Overall , |----------------------------------------

  bc,   0.3,   2.7,   9.4,  13.4,  88.8, [-  | ++++++++++++++++++++++++++++++++++++++       ]
 bfc,   0.6,   8.4,  11.8,  17.0, 103.6, [--- |  +++++++++++++++++++++++++++++++++++++++++++]
  bf,   0.9,  10.3,  13.0,  21.4,  98.8, [---- |   +++++++++++++++++++++++++++++++++++++++  ]
  fc,   1.4,  12.2,  25.5,  79.3, 117.2, [-----      |                          ++++++++++++]
```

This is the overall pattern. Repeat the above for each data set

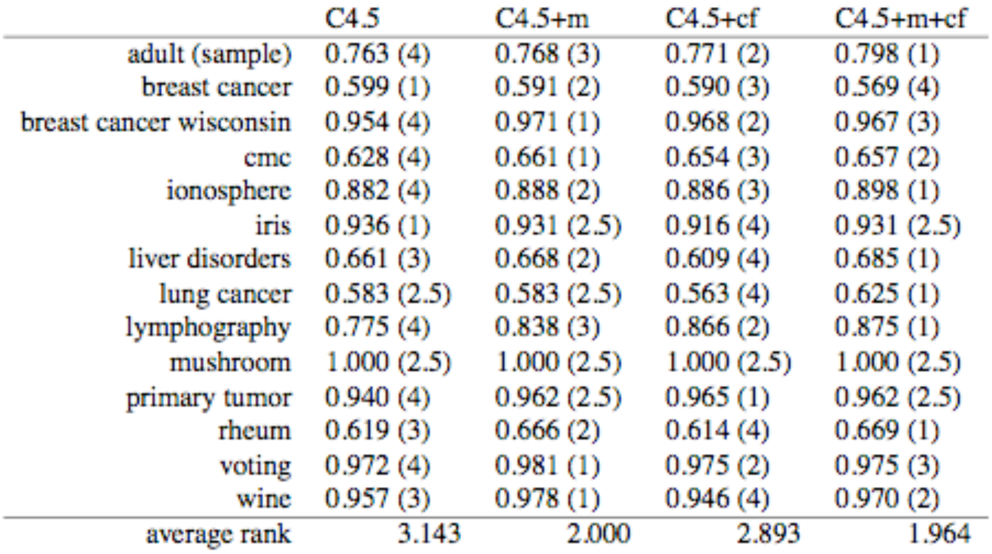
```
----| data set 1 |----------------------------------------

  bc,   0.3,   0.3,   0.5,   0.6,  16.2, [+++++++                                           ]
 bfc,   0.6,   0.6,   0.9,   1.3,   3.9, [+                                                 ]
  fc,   1.4,   1.5,   2.6,   5.3,   7.1, [|+                                                ]
  bf,   0.9,   1.1,   2.7,   4.7,   5.7, [|+                                                ]


----| data set 2 |----------------------------------------

  bc,   1.8,   1.8,   2.0,   2.2,  11.0, [|++++                                             ]
  bf,   2.2,   6.1,  10.6,  11.6,  12.3, [--  |                                             ]
 bfc,   6.3,   6.9,  10.9,  12.2,  12.9, [  - |+                                            ]
  fc,   8.1,   8.6,  12.2,  12.7,  14.3, [   - |                                            ]
```

Only show overall if the individuals are not strikingly different to the overall.

Statistical tests for differences

```
for c in Criteria
    write one table
        whose columns are each learner and
        whose rows are each data set and
        whose celles are the median value for Criteria on data with learner
            (across all r repeats and rNways)
```

e,g four variants on C4.5 . criteria-= accuracy.



	C4.5	C4.5+m	C4.5+cf	C4.5+m+cf
adult (sample)	0.763 (4)	0.768 (3)	0.771 (2)	0.798 (1)
breast cancer	0.599 (1)	0.591 (2)	0.590 (3)	0.569 (4)
breast cancer wisconsin	0.954 (4)	0.971 (1)	0.968 (2)	0.967 (3)
cmc	0.628 (4)	0.661 (1)	0.654 (3)	0.657 (2)
ionosphere	0.882 (4)	0.888 (2)	0.886 (3)	0.898 (1)
iris	0.936 (1)	0.931 (2.5)	0.916 (4)	0.931 (2.5)
liver disorders	0.661 (3)	0.668 (2)	0.609 (4)	0.685 (1)
lung cancer	0.583 (2.5)	0.583 (2.5)	0.563 (4)	0.625 (1)
lymphography	0.775 (4)	0.838 (3)	0.866 (2)	0.875 (1)
mushroom	1.000 (2.5)	1.000 (2.5)	1.000 (2.5)	1.000 (2.5)
primary tumor	0.940 (4)	0.962 (2.5)	0.965 (1)	0.962 (2.5)
rheum	0.619 (3)	0.666 (2)	0.614 (4)	0.669 (1)
voting	0.972 (4)	0.981 (1)	0.975 (2)	0.975 (3)
wine	0.957 (3)	0.978 (1)	0.946 (4)	0.970 (2)
average rank	3.143	2.000	2.893	1.964

Convert each row to ranks. Largest numbers get topped ranked (unless you looking something like probability of failure in which case bottom numbers get top ranked).

Same values get same ranks e.g,

1,1,1,1 get ranks 1,2,3,4 which sums to 10 which averages to 2.5

Compute mean rank of each column (see last line)

Apply Friedman test at 95% confidence to see if any of these are different

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right]$$

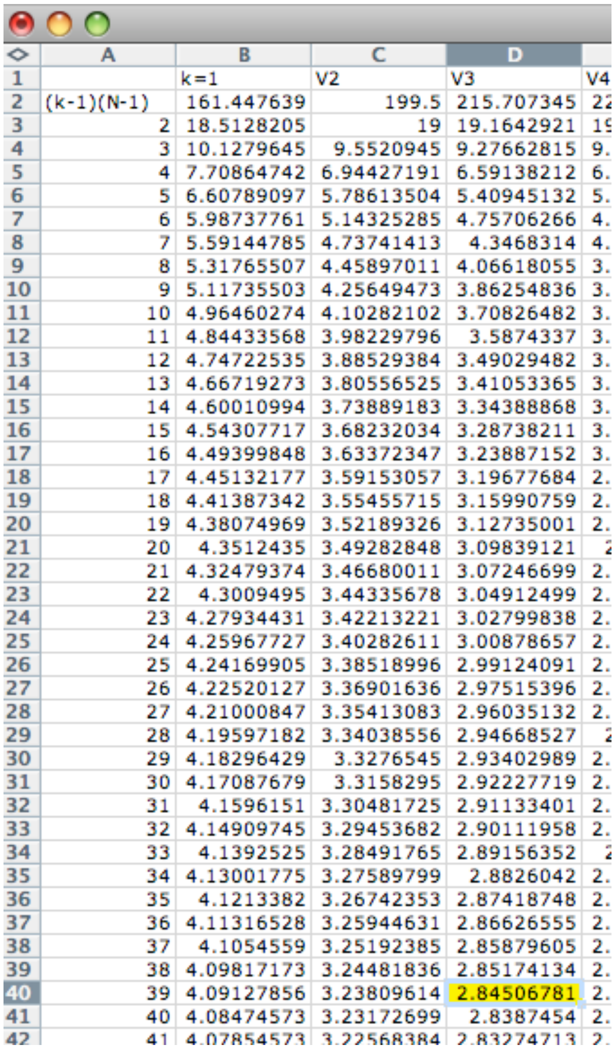
$$F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2}$$

Here, N= # rows (data sets) and k= # columns (learners). E.g.

$$\chi_F^2 = \frac{12\cdot 14}{4\cdot 5}\left[(3.143^2+2.000^2+2.893^2+1.964^2)-\frac{4\cdot 5^2}{4}\right]=9.28$$

$$F_F = \frac{13\cdot 9.28}{14\cdot 3-9.28}=3.69.$$

Look up critical value for F(k-1, (k-1)*(N-1))= F(3,39) in iccle/trunk/share/cvs/Fvalue95percent.csv



◇	A	B	C	D	
1		k=1	V2	V3	V4
2	(k-1)(N-1)	161.447639	199.5	215.707345	22
3	2	18.5128205	19	19.1642921	19
4	3	10.1279645	9.5520945	9.27662815	9.
5	4	7.70864742	6.94427191	6.59138212	6.
6	5	6.60789097	5.78613504	5.40945132	5.
7	6	5.98737761	5.14325285	4.75706266	4.
8	7	5.59144785	4.73741413	4.3468314	4.
9	8	5.31765507	4.45897011	4.06618055	3.
10	9	5.11735503	4.25649473	3.86254836	3.
11	10	4.96460274	4.10282102	3.70826482	3.
12	11	4.84433568	3.98229796	3.5874337	3.
13	12	4.74722535	3.88529384	3.49029482	3.
14	13	4.66719273	3.80556525	3.41053365	3.
15	14	4.60010994	3.73889183	3.34388868	3.
16	15	4.54307717	3.68232034	3.28738211	3.
17	16	4.49399848	3.63372347	3.23887152	3.
18	17	4.45132177	3.59153057	3.19677684	2.
19	18	4.41387342	3.55455715	3.15990759	2.
20	19	4.38074969	3.52189326	3.12735001	2.
21	20	4.3512435	3.49282848	3.09839121	2
22	21	4.32479374	3.46680011	3.07246699	2.
23	22	4.3009495	3.44335678	3.04912499	2.
24	23	4.27934431	3.42213221	3.02799838	2.
25	24	4.25967727	3.40282611	3.00878657	2.
26	25	4.24169905	3.38518996	2.99124091	2.
27	26	4.22520127	3.36901636	2.97515396	2.
28	27	4.21000847	3.35413083	2.96035132	2.
29	28	4.19597182	3.34038556	2.94668527	2
30	29	4.18296429	3.3276545	2.93402989	2.
31	30	4.17087679	3.3158295	2.92227719	2.
32	31	4.1596151	3.30481725	2.91133401	2.
33	32	4.14909745	3.29453682	2.90111958	2.
34	33	4.1392525	3.28491765	2.89156352	2
35	34	4.13001775	3.27589799	2.8826042	2.
36	35	4.1213382	3.26742353	2.87418748	2.
37	36	4.11316528	3.25944631	2.86626555	2.
38	37	4.1054559	3.25192385	2.85879605	2.
39	38	4.09817173	3.24481836	2.85174134	2.
40	39	4.09127856	3.23809614	2.84506781	2.
41	40	4.08474573	3.23172699	2.8387454	2.
42	41	4.07854573	3.22568384	2.83274713	2.

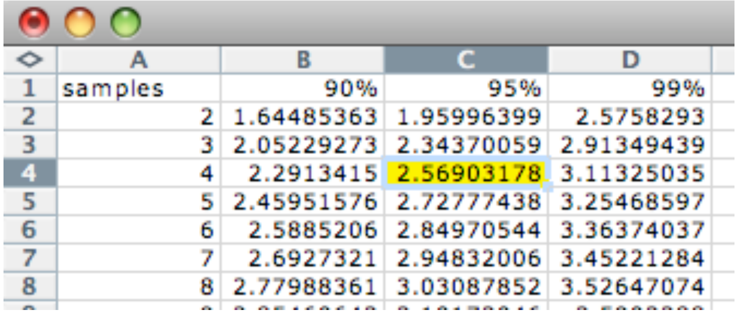
These ranks are different since F (3,39) < F; i..e. 2.845 < 3.69. If otherwise, your conclusion would be that all these perform the same.

Now that we kow that they are different, we seek the “losers”; i.e. the learners that score much less than the top ranked. To define “much less’ we use the Nemenyi critical distance (CD). You are a loser if your average rank is less than topRank - CD

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}}$$

where q comes from iccle/trunk/share/cvs/nemanyiCriticalValue.csv

E.g. Top ranked (above) was 1.964. The q value for 4 learners at 95% confidence is



◇	A	B	C	D
1	samples	90%	95%	99%
2	2	1.64485363	1.95996399	2.5758293
3	3	2.05229273	2.34370059	2.91349439
4	4	2.2913415	2.56903178	3.11325035
5	5	2.45951576	2.72777438	3.25468597
6	6	2.5885206	2.84970544	3.36374037
7	7	2.6927321	2.94832006	3.45221284
8	8	2.77988361	3.03087852	3.52647074

so the critical distance is $2.569\sqrt{\frac{4 \cdot 5}{6 \cdot 14}} = 1.25$

Now, note that ALL our ranks in the above are not losers. I.e. there is no evidence that any of the above methods does better than anything else.

In summary,

Tie if Friedman says there is no different
Else some are losers and the remainder are winners

The final report is the non-losing learners. E.g. at the 955 confidence level a Friendman-Nemenyi test reports no statistical difference between these learners.

Prediction: most of your variants will be statistically indistinguishable.