Confidence Regions and Averaging for Trees
Examples

Susan Holmes
Statistics Department, Stanford
and INRA- Biométrie, Montpellier,France
susan@stat.stanford.edu
http://www-stat.stanford.edu/~susan/

Phylogenetic Tree Example

Family trees or phylogenetic trees whose leaves are different evolutionary entities (species, genes, populations).

DNA Data for 12 species of primates

Mitochondria, 898 characters on 12 species.

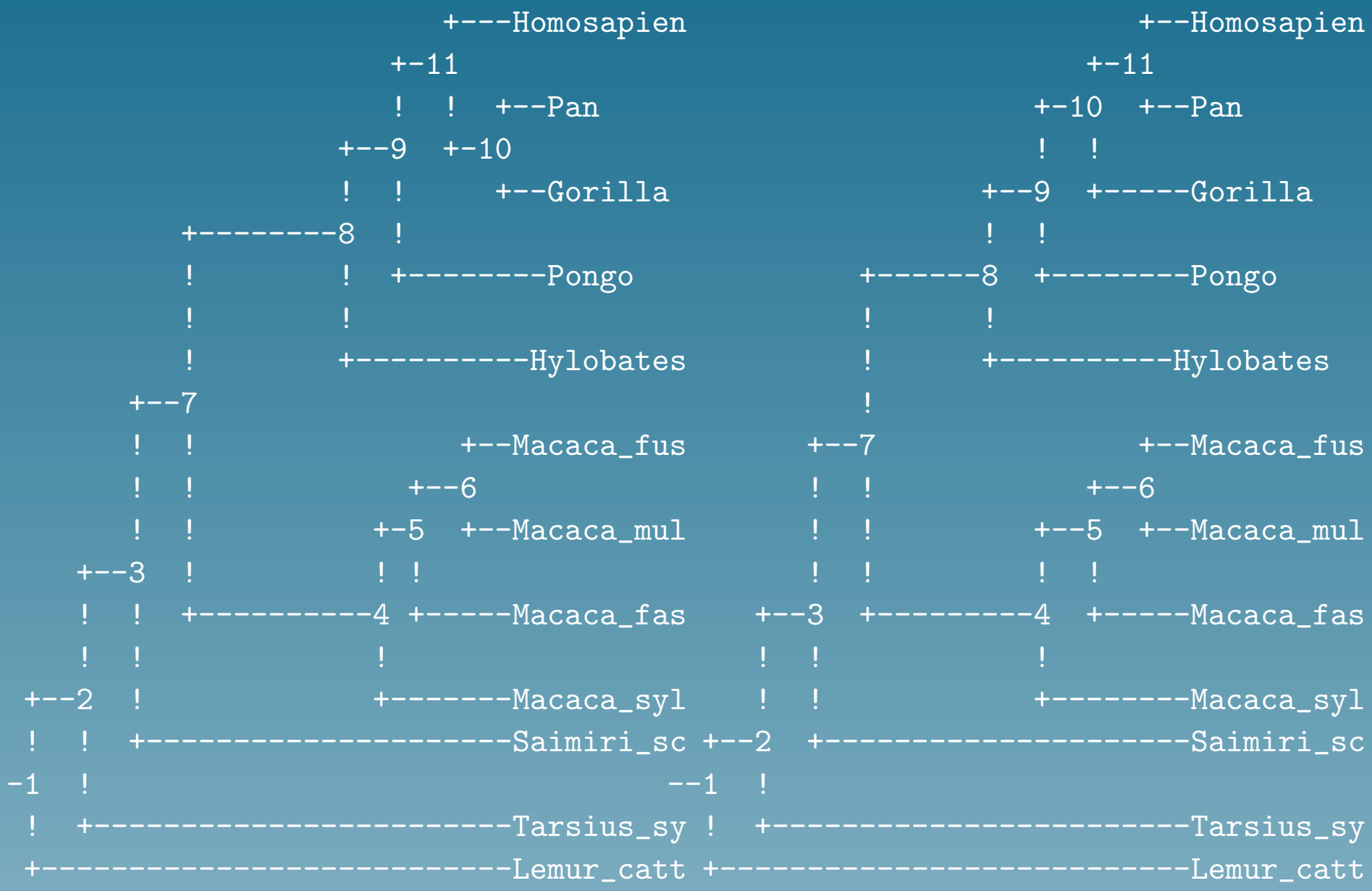
Hayasaka, K., T. Gojobori, and S. Horai. 1988.

Trees are built from DNA data such as the following:

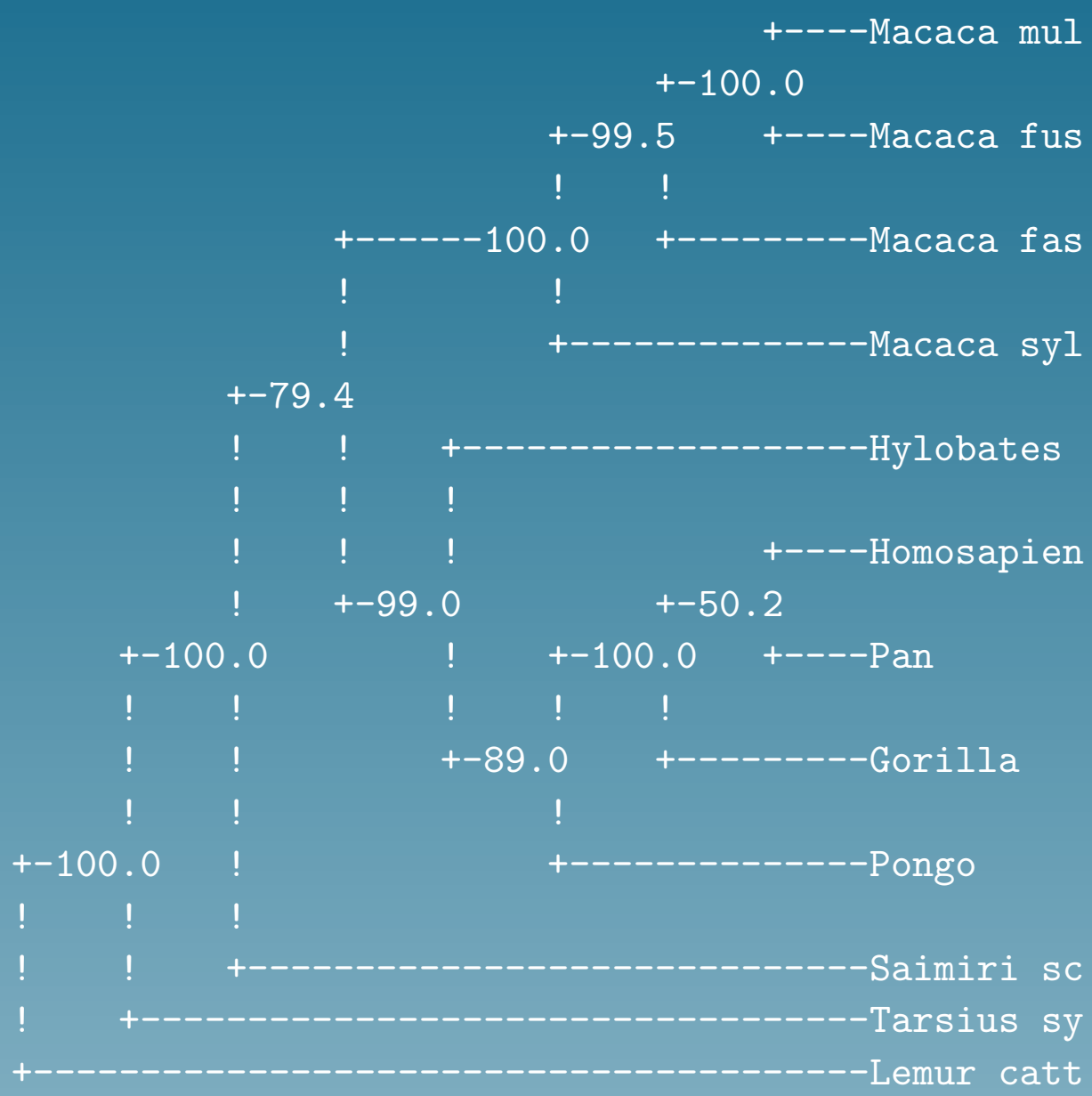
```
  12  60
Lemur_cat AAGCTTCATA GGAGCAACCA TTCTAATAAT CGCACATGGC CTTACATCAT CCATATTATT
Tarsius_s AAGTTTCATT GGAGCCACCA CTCTTATAAT TGCCCATGGC CTCACCTCCT CCCTATTATT
Saimiri_s AAGCTTCACC GGCGCAATGA TCCTAATAAT CGCTCACGGG TTTACTTCGT CTATGCTATT
Macaca_sy AAGCTTCTCC GGTGCAACTA TCCTTATAGT TGCCCATGGA CTCACCTCTT CCATATACTT
Macaca_fa AAGCTTCTCC GGCGCAACCA CCCTTATAAT CGCCCACGGG CTCACCTCTT CCATGTATTT
Macaca_mu AAGCTTTTCT GGCGCAACCA TCCTCATGAT TGCTCACGGA CTCACCTCTT CCATATATTT
Macaca_fu AAGCTTTTCC GGCGCAACCA TCCTTATGAT CGCTCACGGA CTCACCTCTT CCATATATTT
 Hylobate AAGCTTTACA GGTGCAACCG TCCTCATAAT CGCCCACGGA CTAACCTCTT CCCTGCTATT
Pongo     AAGCTTCACC GGCGCAACCA CCCTCATGAT TGCCCATGGA CTCACATCCT CCCTACTGTT
Gorilla   AAGCTTCACC GGCGCAGTTG TTCTTATAAT TGCCCACGGA CTTACATCAT CATTATTATT
Pan       AAGCTTCACC GGCGCAATTA TCCTCATAAT CGCCCACGGA CTTACATCCT CATTATTATT
Homosapie AAGCTTCACC GGCGCAGTCA TTCTCATAAT CGCCCACGGG CTTACATCCT CATTACTATT
```

Color Coded version of the data, after alignment

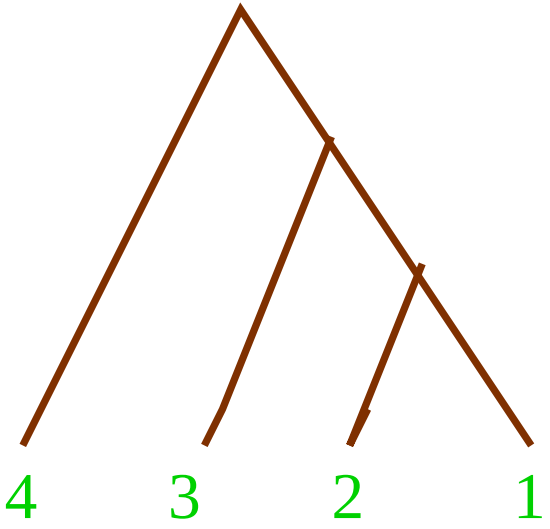
TwoTrees Built with this data by parsimony

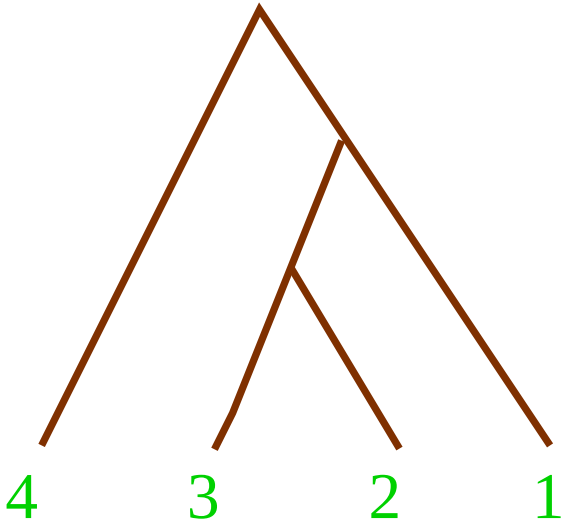


With the Bootstrap "Confidence" Values

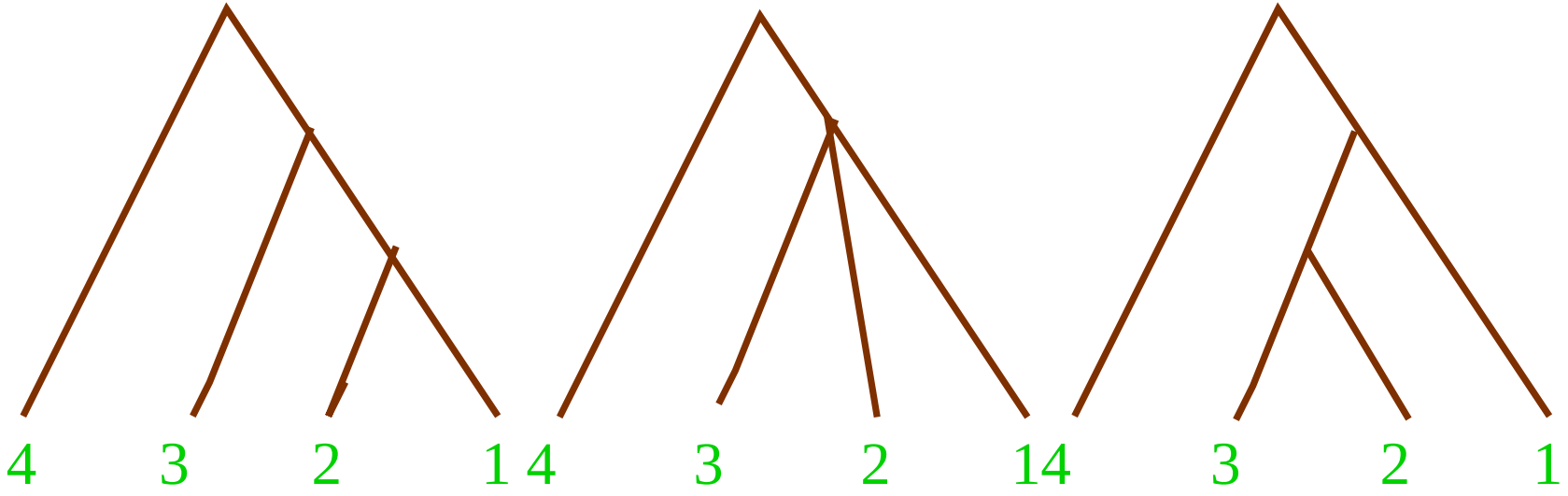


Using the geometry

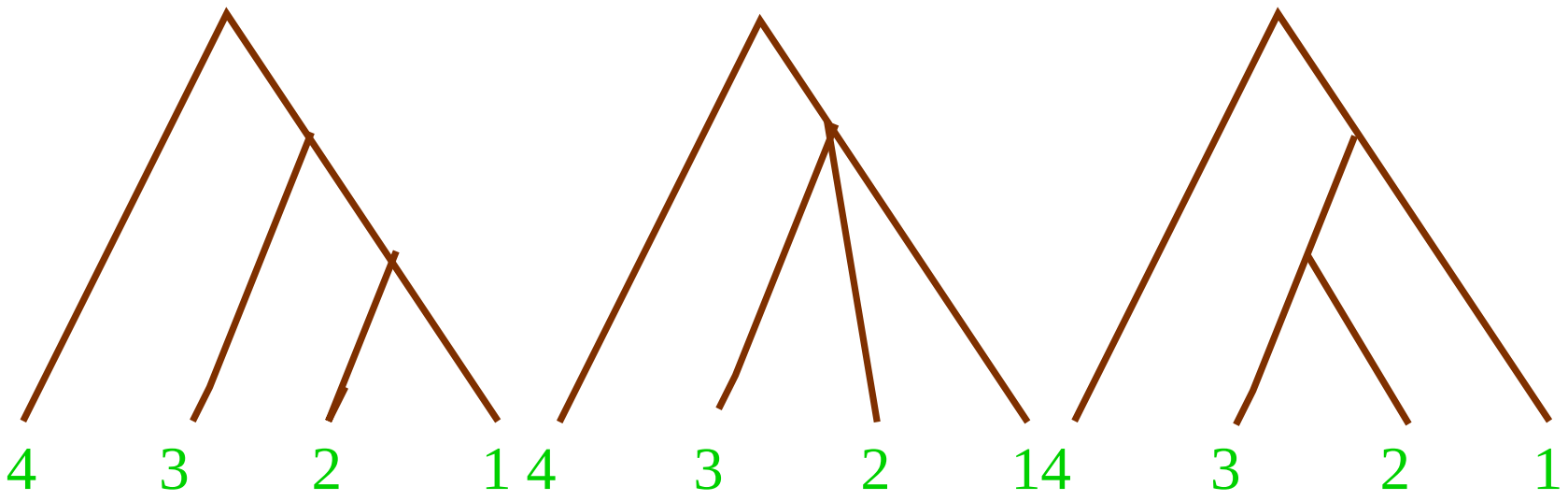




Using the geometry



Using the geometry

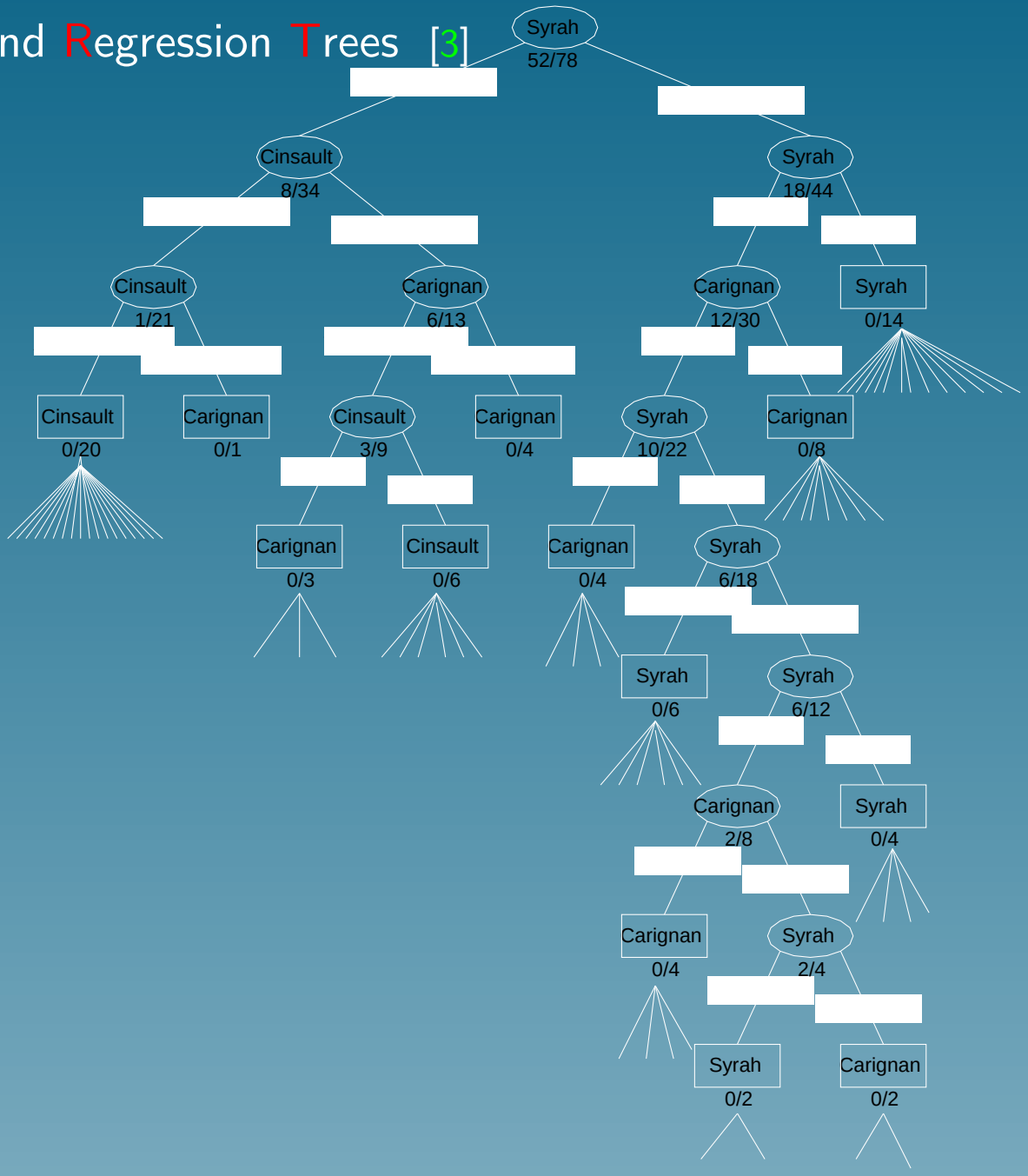


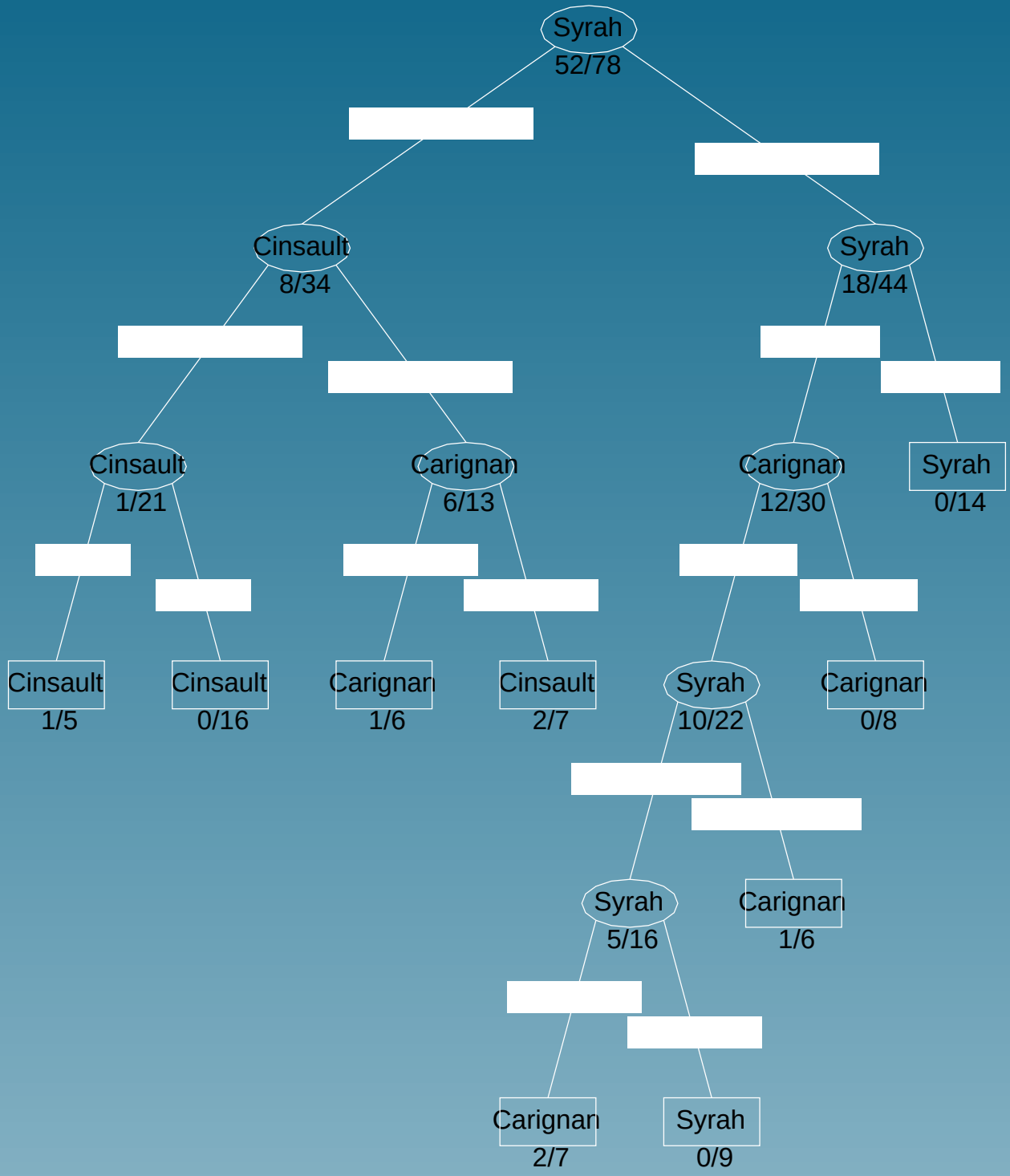
Already used by biolgists to indicate 'unresolved' branchings.

The Wine Data

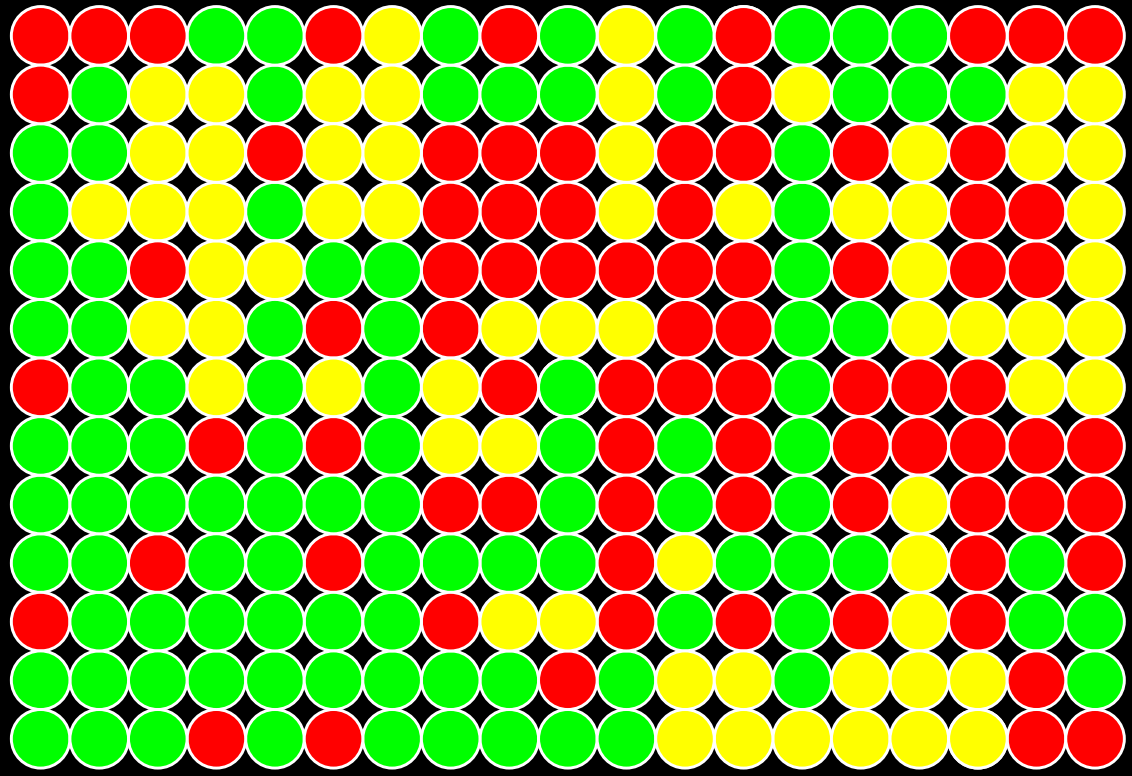
wine	D28	TAN	IGEL	IHCL	COL	JAU	RGE	BLE	DFLV	TNT	ANT	PVP	ION	PH	cepag
1	1.09	2.09	45	13	1.06	64	52	14	54	0.64	413	40	26	3.80	Syral
2	0.61	1.27	35	20	0.65	40	45	16	38	0.89	220	48	22	3.80	Syral
3	0.90	1.68	38	29	0.84	36	49	15	48	0.74	422	46	22	3.84	Syral
4	1.11	2.00	28	26	0.74	37	51	12	51	0.74	527	27	11	3.83	Syral
5	0.85	1.50	37	40	0.66	39	49	12	48	0.80	369	38	17	3.78	Syral
6	0.98	2.00	41	26	0.99	36	51	13	51	0.72	316	67	16	3.75	Syral
7	1.24	2.41	58	27	0.83	38	50	12	49	0.77	316	44	16	3.69	Syral
8	0.92	1.63	50	18	0.58	40	47	12	45	0.85	229	15	15	3.70	Syral
9	1.18	2.00	59	30	0.84	39	48	13	46	0.80	316	39	15	3.80	Syral
10	1.13	2.32	62	25	0.84	40	48	12	46	0.82	281	41	9	3.82	Syral
11	1.04	2.36	61	30	0.69	40	47	13	44	0.85	299	53	7	3.84	Syral
12	1.27	2.69	69	38	0.91	40	47	13	43	0.85	352	68	11	3.85	Syral
13	1.23	3.09	62	22	0.88	42	46	12	41	0.92	255	62	10	3.82	Syral
14	1.34	3.00	64	28	0.89	42	46	12	42	0.91	141	44	11	3.76	Syral
15	1.28	3.00	75	33	1.00	42	46	12	41	0.91	202	83	18	3.82	Syral
16	1.17	2.23	68	20	1.06	42	45	13	39	0.92	105	83	57	3.76	Syral
17	1.31	2.81	65	58	1.00	42	45	13	39	0.94	123	71	43	3.71	Syral
18	0.91	1.63	39	21	0.59	47	42	11	32	1.10	132	73	25	3.76	Syral
19	0.93	1.63	64	12	0.62	46	42	12	32	1.08	149	76	19	3.88	Syral
20	1.08	2.01	44	30	0.79	44	43	13	35	1.02	141	38	21	3.90	Syral
21	1.11	2.27	30	9	0.69	46	42	12	30	1.09	229	46	13	4.08	Syral
22	0.90	1.76	54	23	0.80	46	41	13	29	1.11	88	70	37	4.08	Syral
23	1.09	2.06	55	26	1.21	43	42	15	31	1.03	158	94	27	4.28	Syral
24	0.80	1.76	35	19	0.62	47	41	12	28	1.15	62	86	36	3.94	Syral
25	0.76	1.62	31	24	0.55	47	42	11	31	1.13	202	78	12	4.04	Syral
26	0.86	1.71	39	20	0.67	47	41	12	28	1.14	44	40	16	4.13	Syral
27	0.63	1.41	33	26	0.59	39	49	12	47	0.80	290	45	6	3.66	Carign
28	0.65	2.11	48	26	0.66	37	50	13	49	0.75	387	55	3	3.63	Carign
29	0.69	1.34	22	27	0.60	38	51	11	51	0.74	308	40	7	3.65	Carign
30	0.70	1.25	33	16	0.59	43	45	13	38	0.96	185	19	14	3.85	Carign
31	0.69	1.27	33	19	0.44	43	44	13	37	0.97	246	39	9	3.78	Carign
32	0.78	1.68	39	17	0.65	41	46	13	40	0.91	229	46	15	3.87	Carign
33	0.59	1.47	29	20	0.45	42	44	14	38	0.94	123	71	16	3.72	Carign
34	0.58	1.24	33	34	0.41	41	47	12	43	0.89	141	63	13	3.69	Carign
35	0.66	1.45	19	10	0.31	44	46	10	41	0.96	202	70	7	3.66	Carign
36	0.79	1.41	72	28	0.54	47	42	10	32	1.11	44	80	17	3.91	Carign

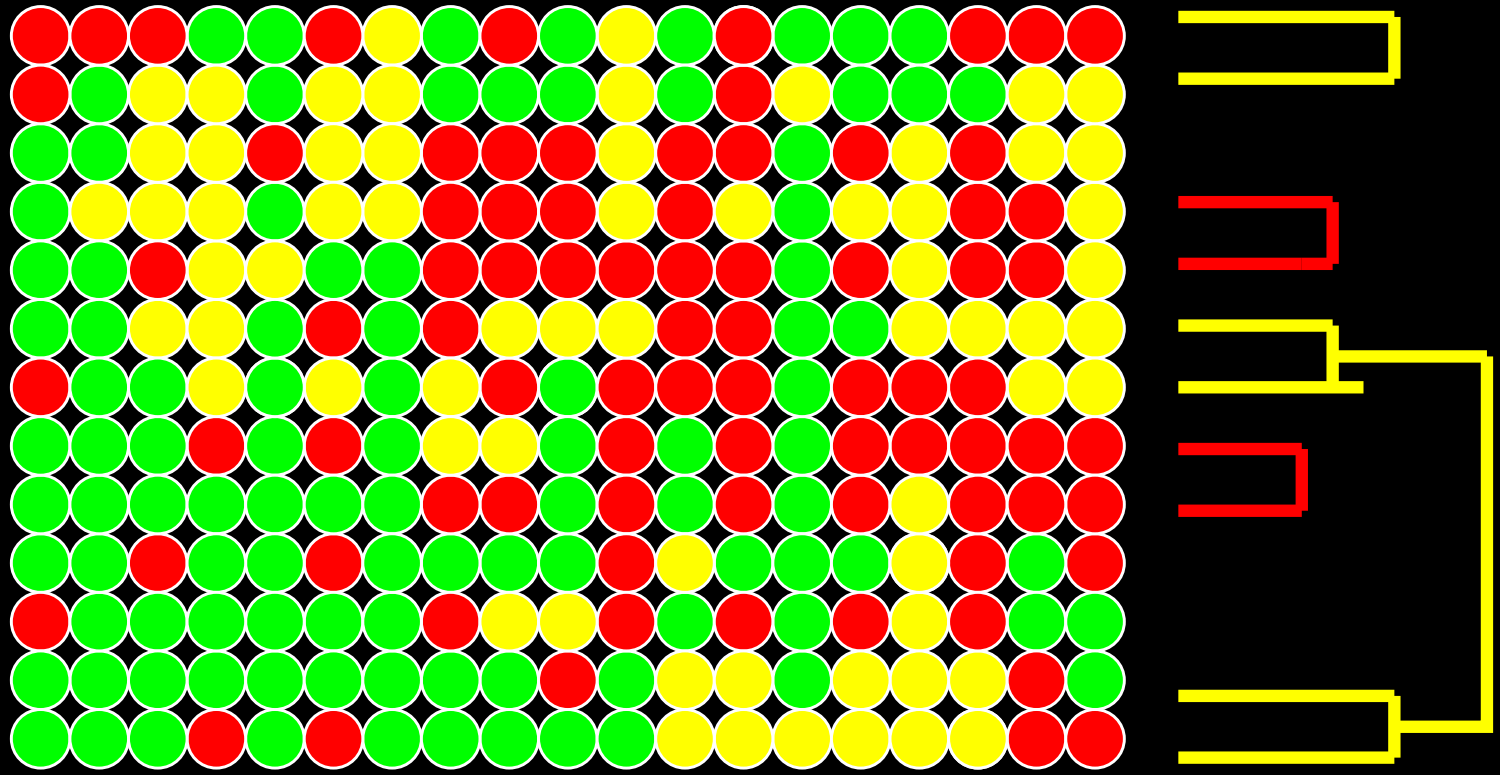
Classification And Regression Trees [3]

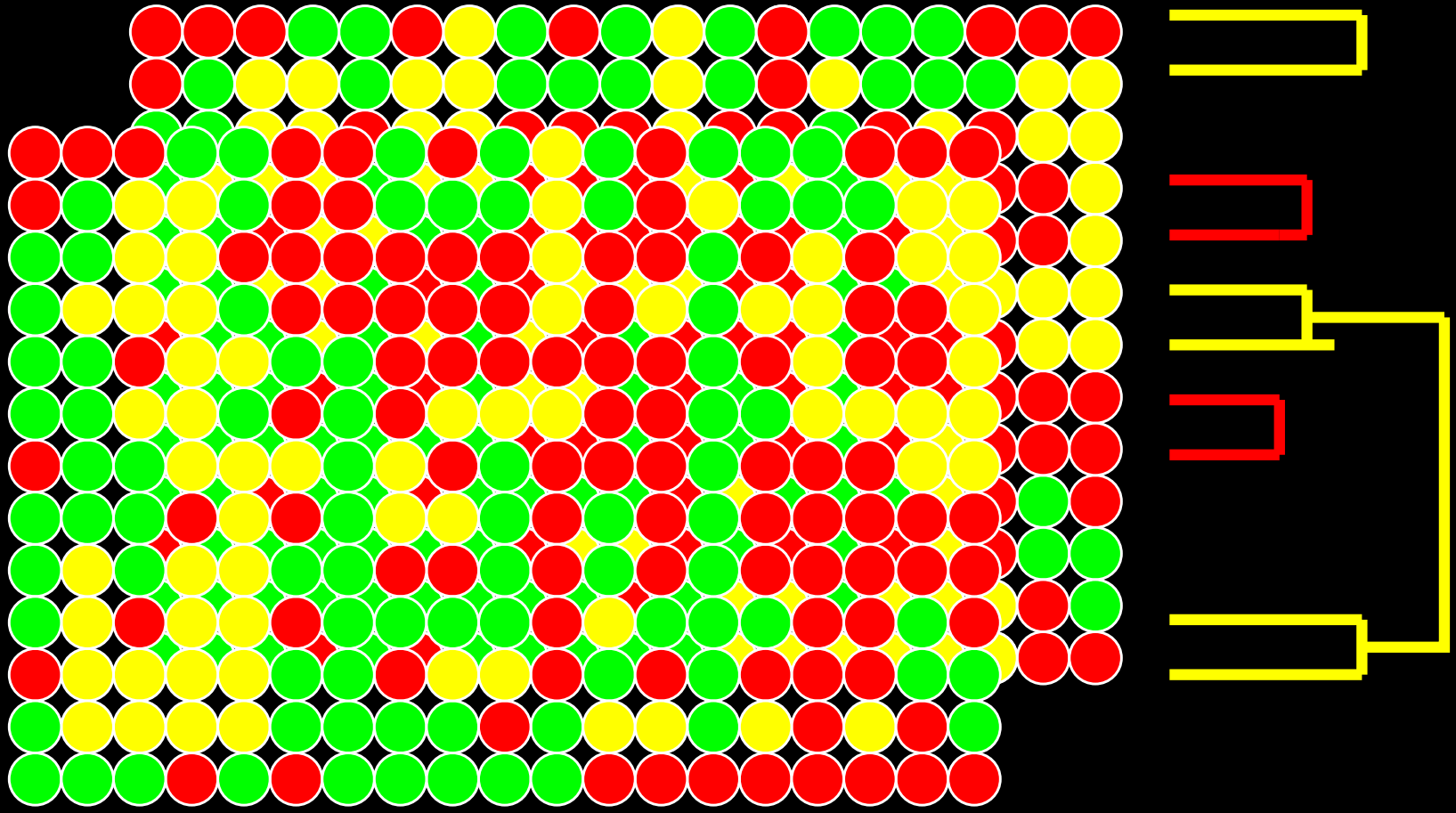


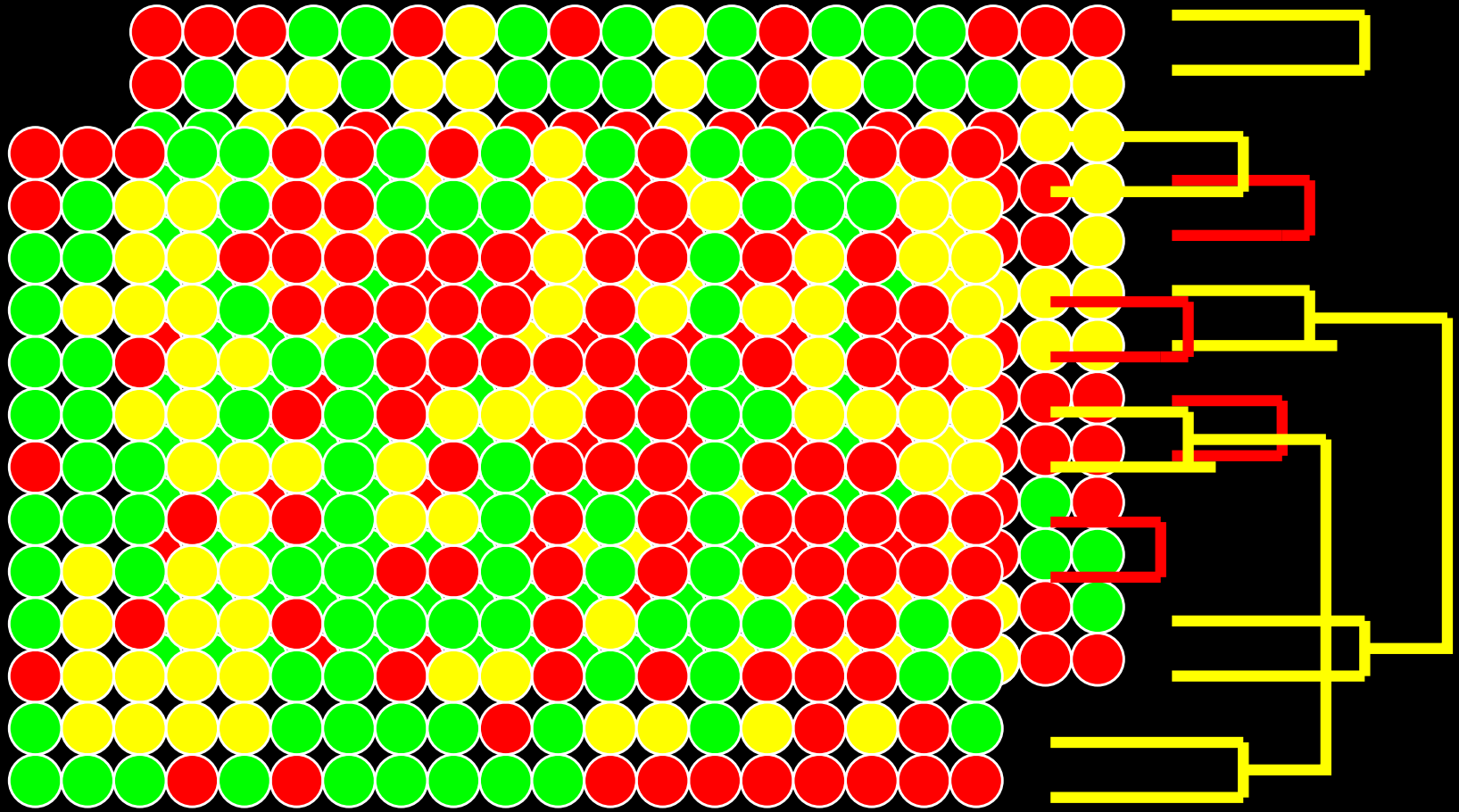

Syrah
52/78
Cinsault
8/34
Syrah
18/44
Cinsault
1/21
Carignan
6/13
Carignan
12/30
Syrah
0/14
Cinsault
1/5
Cinsault
0/16
Carignan
1/6
Cinsault
2/7
Syrah
10/22
Carignan
0/8
Syrah
5/16
Carignan
1/6
Carignan
2/7
Syrah
0/9

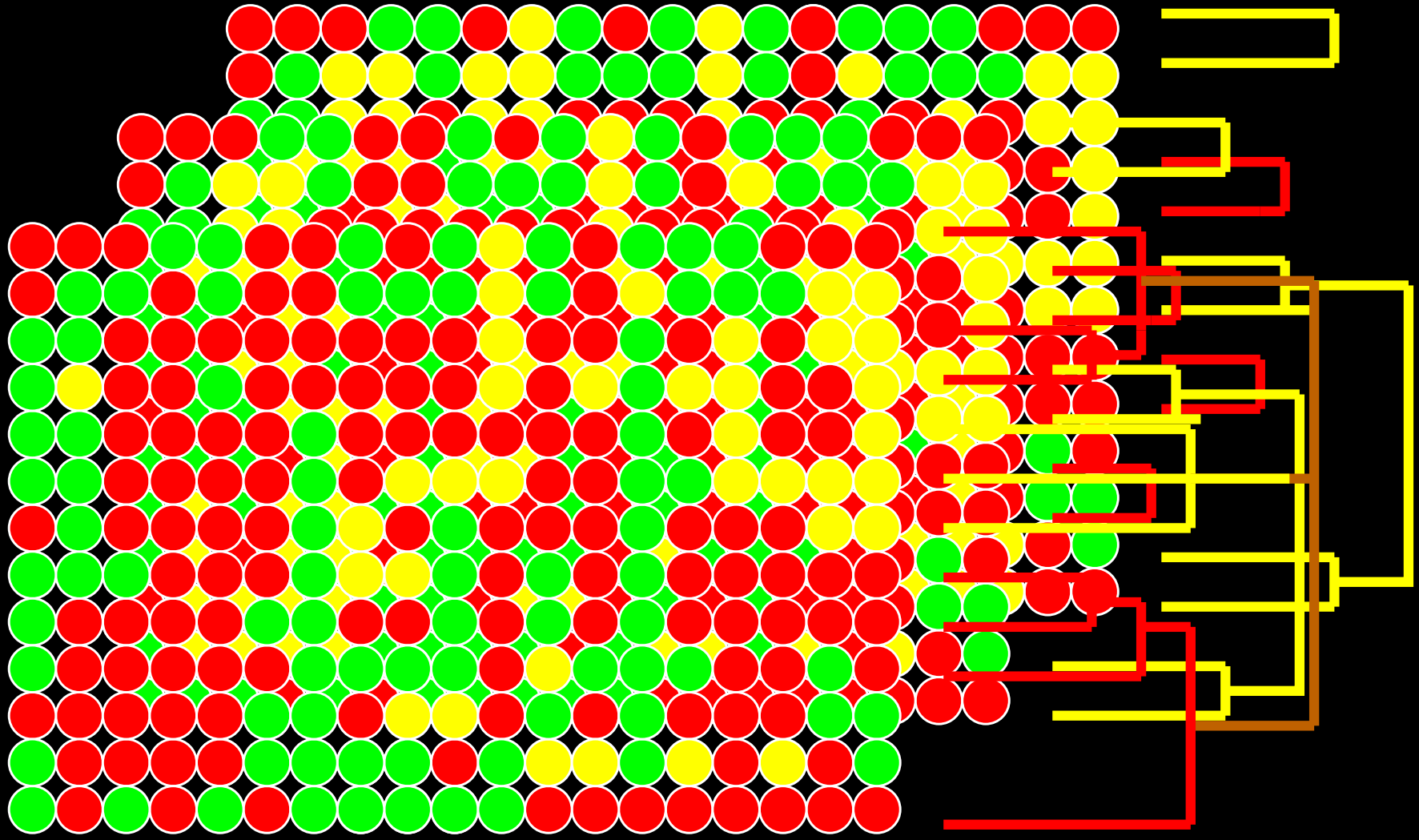
Clustering Trees - For micro-arrays [11]

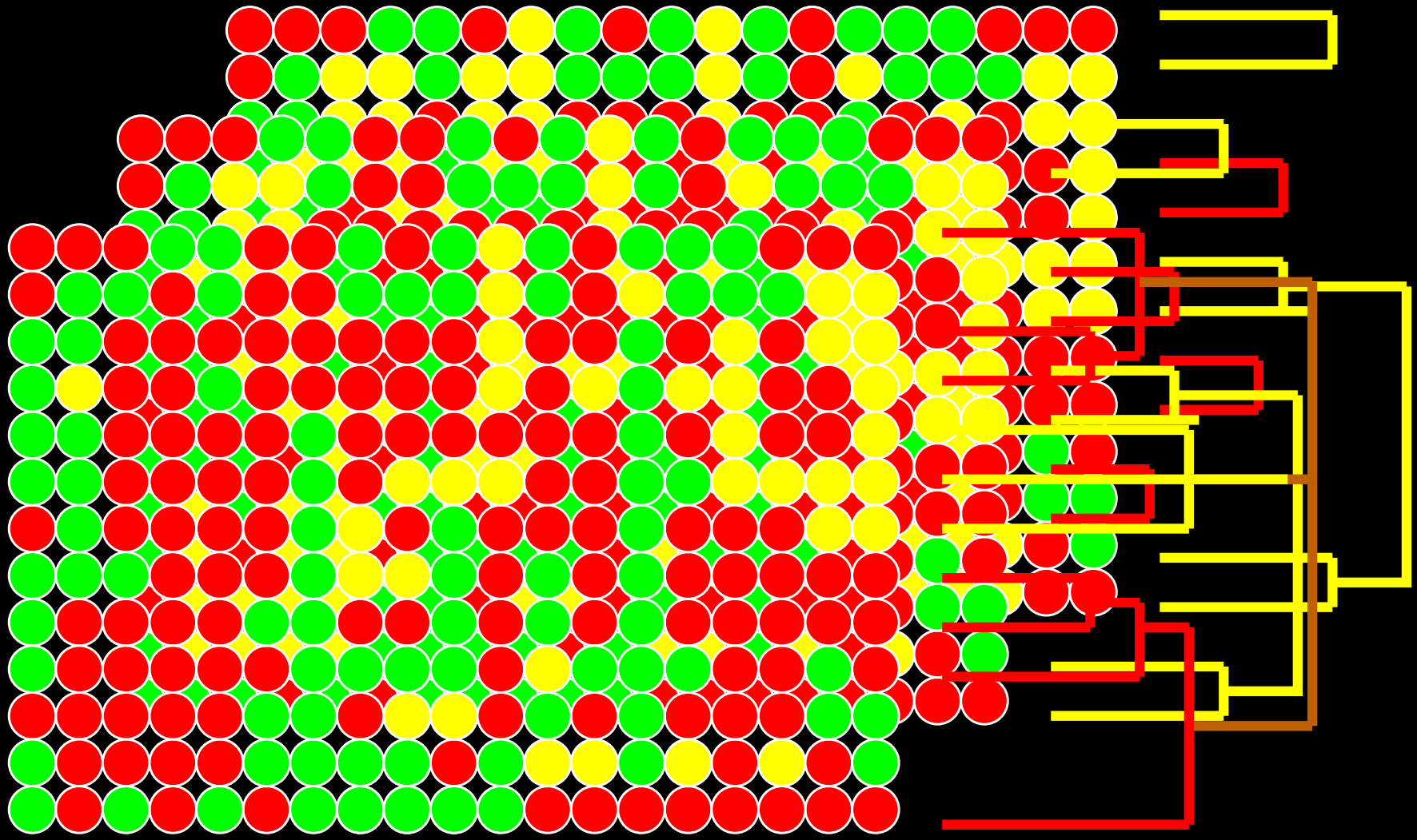












→

How can abstract mathematics help?

- Decompositions that can be generalisable.
- Geometric Picture of Tree Space
  - ⋆ A space for comparisons.
  - ⋆ Ways of *projecting*.
  - ⋆ *Follow* trees as they change, (paths of trees)
  - ⋆ Centroids of trees
  - ⋆ Neighborhoods (convex hulls of trees)....
  - ⋆ Averages of trees

How can abstract mathematics help?

- Decompositions that can be generalisable.
- Geometric Picture of Tree Space
  - ⋆ A space for comparisons.
  - ⋆ Ways of *projecting*.
  - ⋆ *Follow* trees as they change, (paths of trees)
  - ⋆ Centroids of trees
  - ⋆ Neighborhoods (convex hulls of trees)....
  - ⋆ Averages of trees
- Justification of commonsense, ground for generalizations.

References

[1] Y. Amit and D. Geman, *Quantization and recognition with randomized trees*, Neural Computation, 9 (1997), pp. 1545–1588.

[2] L. Billera, S. Holmes, and K. Vogtmann, *Geometry of tree space*, to appear, Statistics, Stanford, 1999.

[3] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Wadsworth, 1984.

[4] M. Charleston, *??*, `http://taxonomy.zoology.gla.ac.uk/~mac/landscape /trees.html.`, ?? (1996), pp. –.

[5] B. Charnomordic and S. Holmes, `Dnaview`*, an interactive viewer for alignment and tree building*, unpublished software, (1997).

[6] P. Diaconis, *Group Representations in Probability and Statistics*, Institute of Mathematical Statistics, 1988.

[7] ——, *A generalization of spectral analysis with application to ranked data*, The Annals of Statistics, 17 (1989), pp. 949–979.

[8] P. W. Diaconis and S. P. Holmes, *Matchings and phylogenetic trees*, Proc. Natl. Acad. Sci. USA, 95 (1998), pp. 14600–14602 (electronic).

[9] B. Efron, E. Halloran, and S. P. Holmes, *Bootstrap confidence levels for phylogenetic trees*, Proc. Natl. Acad. Sci. USA, 93 (1996), pp. 13429–34.

[10] B. Efron and R. Tibshirani, *The problem of regions*, Ann. Statist., 26 (1998), pp. 1687–1718.

[11] M. Eisen, P. Spellman, P. Brown, and D. Botstein, *Cluster analysis and display of genome-wide expression patterns.*, Proc Natl Acad Sci USA, (1998).

[12] J. Friedman, *Greedy function approximation: A gradient boosting machine*, tech. rep., Stanford Statistics Dept., 1999. http://www-stat.stanford.edu/~jhf/trebst.ps.

[13] M. Gromov, *Hyperbolic groups*, in Essays in group theory, Springer, New York, 1987, pp. 75–263.

[14] K. Hayasaka, T. Gojobori, and S. Horai, *Molecular phylogeny and evolution of primate mitochondrial dna.*, Mol. Biol. Evol., 5/6 (1988), pp. 626–644.

[15] S. Holmes, *Phylogenies: An overview*, in Statistics and Genetics, E. Halloran and S. Geisser, eds., no. 81 in IMA, Springer Verlag, NY, 1999.

[16] S. Holmes and P. Diaconis, *Computing with Trees*, Interface Foundation of North America, 1999.

[17] S. Li, D. K. Pearl, and H. Doss, *Phylogenetic tree construction using mcmc*, Journ. American Statistical Association (to appear)., (2000). Ohio Statistics Dept., ().

[18] B. MacFadden and R. Hulbert Jr, *Explosive speciation at the base of the adaptive radiation of miocene grazing horses*, Nature, 336 (1988), pp. 466–468.

[19] B. Mau, M. A. Newton, and L. B., *Bayesian phylogenetic inference via markov chain monte carlo methods*, Biometrics, (1999).

[20] E. Schröder, Zeit. für. Math. Phys., 15 (1870), pp. 361–376.

[21] G. L. Thompson, *Generalized permutation polytopes and exploratory graphical methods for ranked data*, The Annals of Statistics, 21 (1993), pp. 1401–1430.

[22] G. M. Ziegler, *Lectures on polytopes*, Springer-Verlag, New York, 1995.