Modeling Interaction in a Two-Way Layout, with Application to Medicinal Chemistry

R. Daniel Meyer, Bruce Lefker

Pfizer Inc.

Historical Note

Seminal paper written ~50 years ago

Renowned statistician collaborates with a chemist named Wilson

Methodology forms basis for optimization of chemical matter

Can you name that paper?

Outline

- Background
- Roots of the problem – medicinal chemistry
- Statistical problem
- Prototype algorithm
- Example
- Summary / further work

Statistics in Pfizer R&D

- Clinical Statistics
  - Clinical trials of investigational drugs
  - New drug application (NDA)
- Nonclinical Statistics
  - Drug discovery
  - Product development / manufacture
  - Preclinical toxicology/safety
  - Some human studies (genetic association, methodology studies)

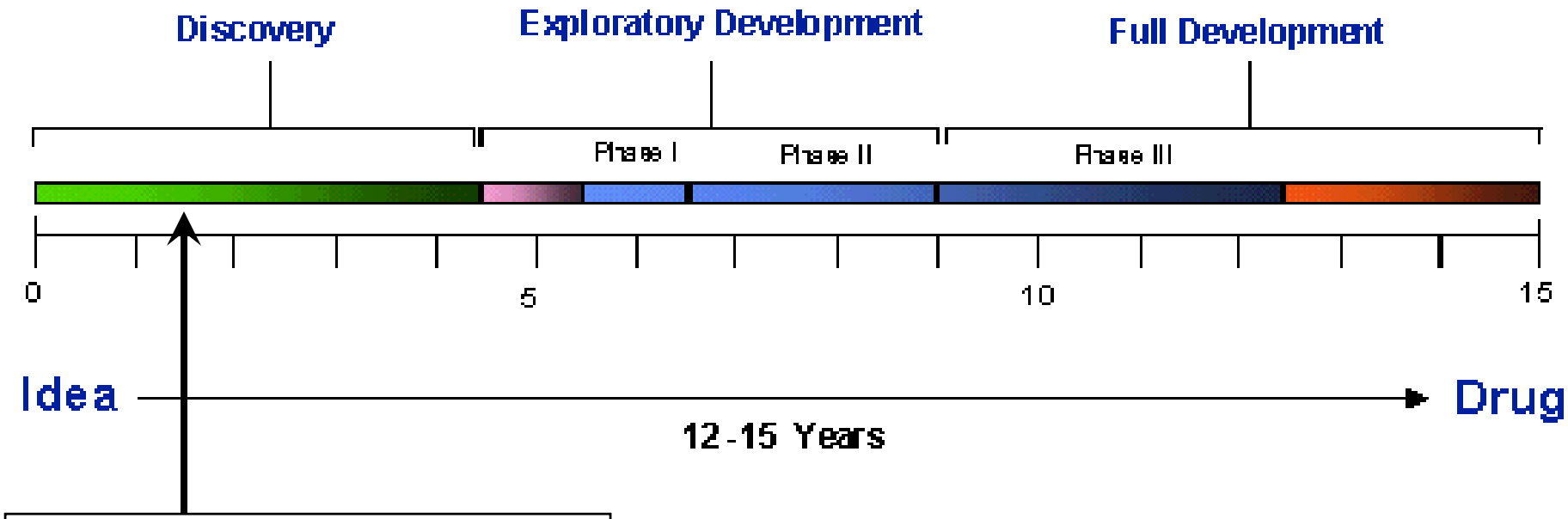
Drug Discovery

- Biology:
  - select disease-relevant targets
  - assays to evaluate new compounds
- **Medicinal Chemistry**:
  - create compounds to be evaluated for biological activity
- Chemistry starting point:
  - Approved drug, natural ligand, HTS, target crystal structure

What is a Drug?

- *Pharmacologically active ingredient* in a...
- *Dosage form* designed to deliver it to the appropriate physiological tissue
- *Drug discovery* is the process of identifying new pharmacologically active chemicals

Drug Development Sequence



Today's talk focuses here: discovering new chemical entities (NCEs)

Required Properties of Drugs

- Potent (binds to desired target)
- Selective (doesn’t bind to non-targets)
- Readily absorbed by the body
- Soluble in body fluids
- Nontoxic
- Metabolizes at right rate for convenient dosing
- Metabolism/excretion pathways benign

How Do Drugs Work?

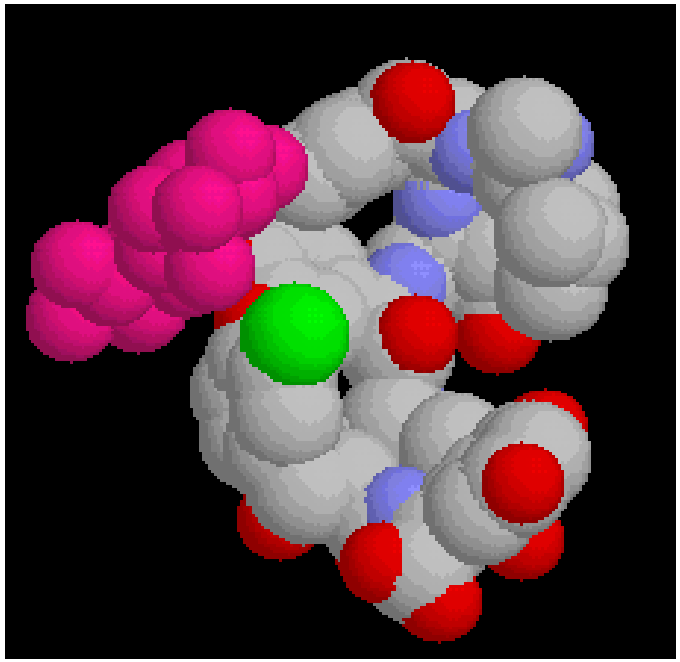
Paul Ehrlich

Corpora non agunt nisi fixata

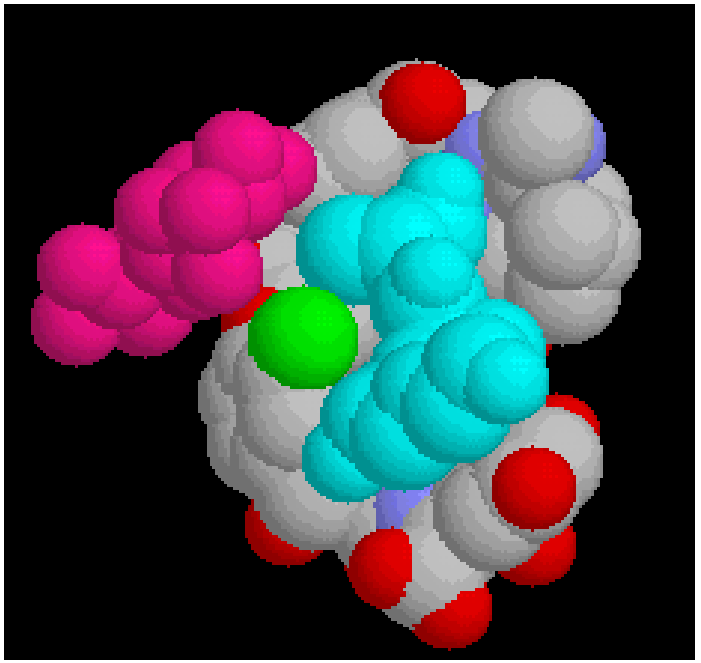
(substances do not act unless bound)

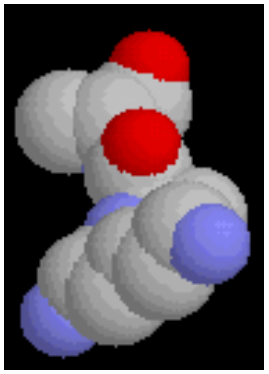
Physical Binding to Target

Vancomycin



Vancomycin-L-LYS-D-ALA-D-ALA





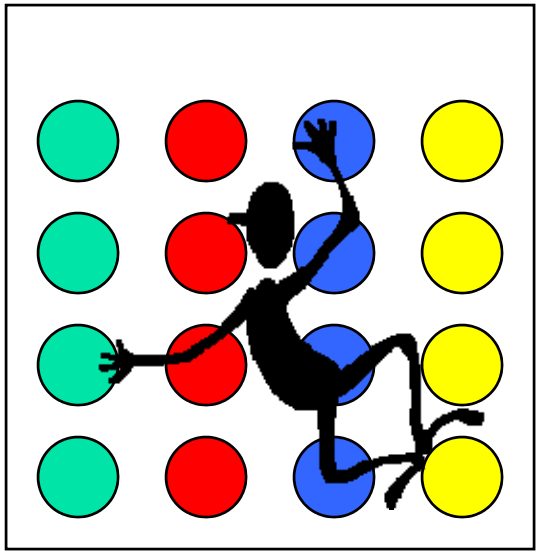
L-LYS-D-ALA-D-ALA

Physical Binding to Target

- 3-dimensional shape of the drug molecule must conform to 3D shape of binding site
- Charge (+/-) on the molecule surface is important to achieve binding strength
- Hydrogen-bonding also contributes to interaction
- Lipophilicity important too

“Twister” Analogy

- Compound must contort to protein pattern, just like I must contort to Twister pattern

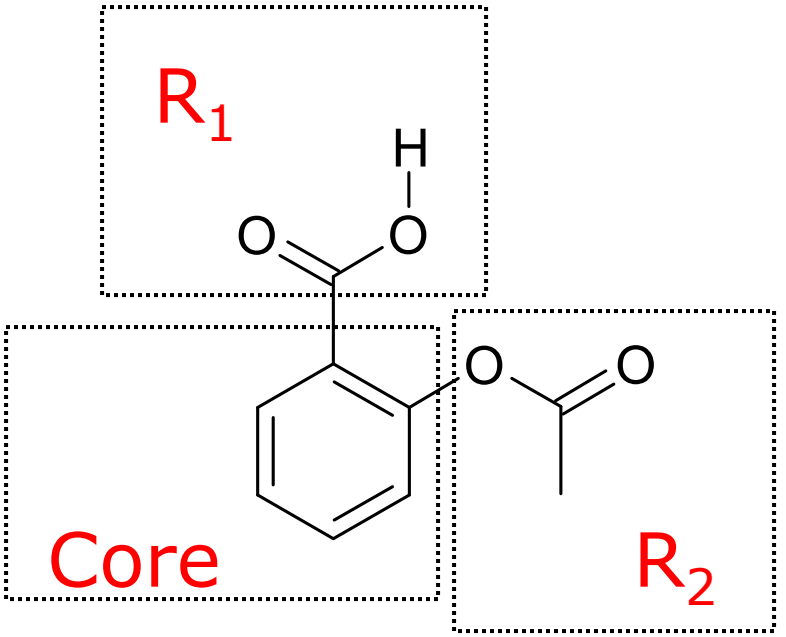


- Compound can bind if contortion not too extreme

Med Chem: Lead Optimization

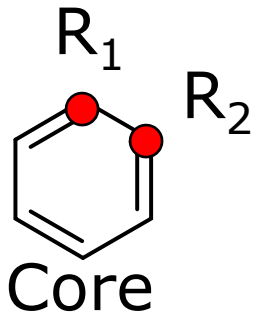
Initial exploration eventually produces a lead compound (looks like a drug)

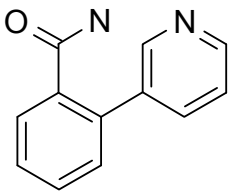
Aspirin



- **Basic idea**: Substitute other chemical fragments (substituents) at the R_1 and R_2 sites

Lead Optimization



R_2 \ R_1	C_3H_7	carbonyl	$CONH_2$	. . .
pyridine				
pyrimidine				
. . .				

Virtual library

Lead Optimization

R_2 \ R_1	a	b	c	d	e	f	g
A							
B			120				
C				200			
D			10	2.2	5		
E							
F			40				

Compound	R_1	R_2	IC_{50}
1	c	B	120
2	d	C	200
3	c	D	10
4	d	D	2.2
5	e	D	5
6	c	F	40

- Large 2-way (k-way) layout; common to have >100 levels
- Expensive to fill in a cell → requires making, testing the compound → many empty cells
- No ordering of the rows and columns

Footnote: Descriptors

Compound	R_1	R_2	IC_{50}	X_1	X_2	...	X_k
1	c	B	120	0	2.345		1
2	d	C	200	1	6.54		3
3	c	D	10	1	7.805		2
4	d	D	2.2	1	5.435		5
5	e	D	5	0	3.905		4
6	c	F	40	0	5.983		7

- Descriptors are computed variables that describe the chemical structure; *k* can be > 1000
- Model $Y = f(X_1, \ldots, X_k)$; numerous approaches to approximating f(•)
- But what can we do without descriptors?

Statistical Models

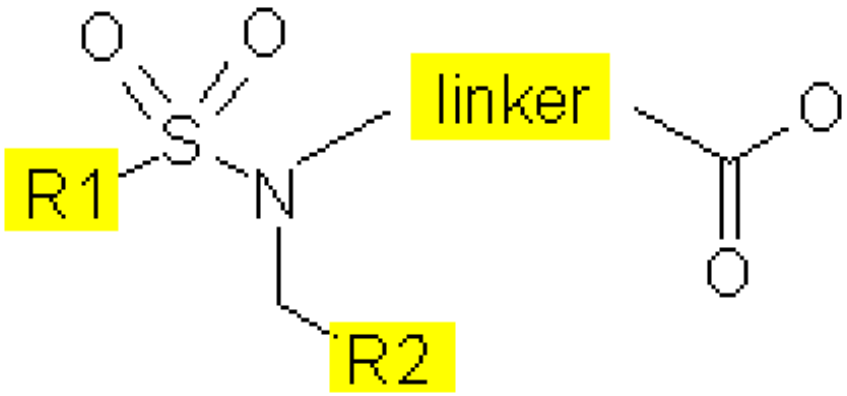
- Free and Wilson (1964) *J. Med. Chem*

Response = average +

effect of R_1 substituent + effect of R_2 substituent

- Main effects model
- R1 and R2 are independent variables
- Their levels are labels of substituents

EP2 Project



- Bone-healing / osteoporosis (died in Phase II)
- Free-Wilson worked well at first
- One compound that didn't fit the model was re-tested . . .

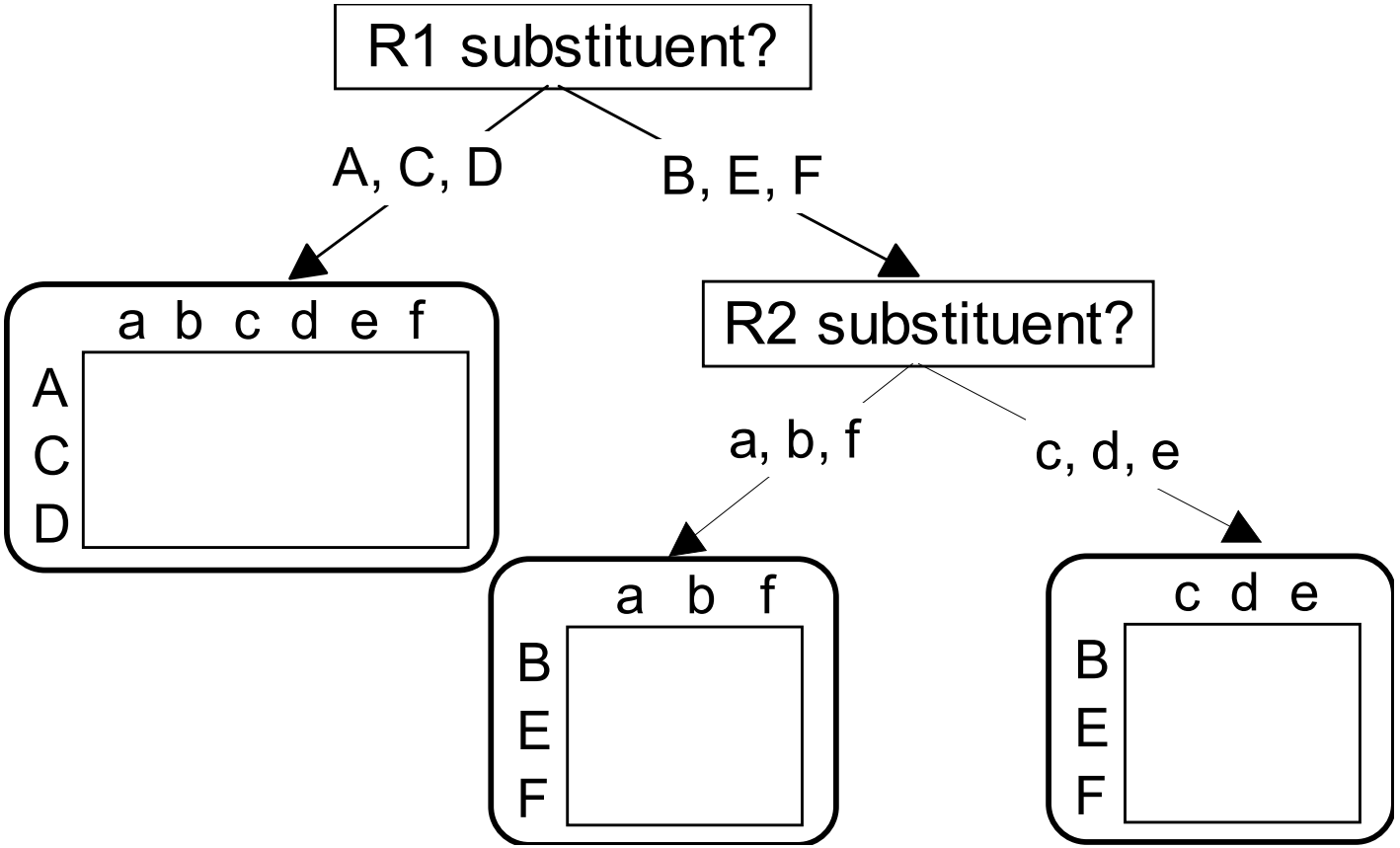
EP2 Project

- Eventually 6 linkers, 67 R1's, 242 R2's
- As series grew, model deteriorated
- Chemist suggested partitioning the table by chemical group → It worked!

Model	s.d.	R-Square	# of Param
. Main effects	1.38	0.70	315
. Lefker partition	0.70	0.94	514

- If statisticians could automatically find groupings . . .

IDEA: ANOVA tree

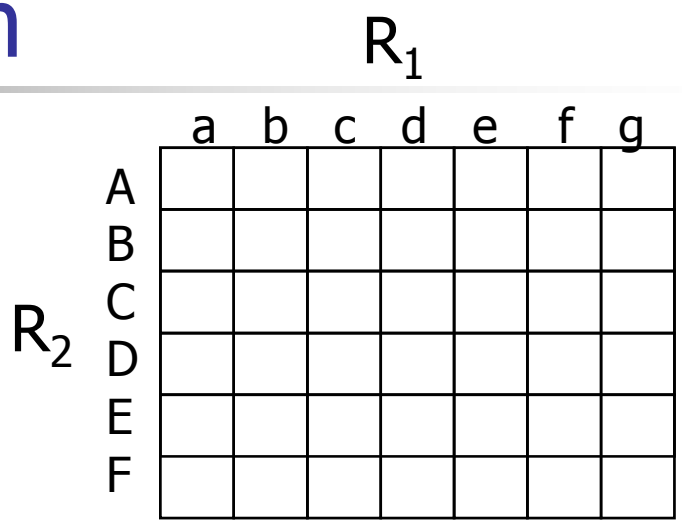


Model the 2-way interaction → within a terminal node, no interaction → able to predict the empty cells

Barriers: Data / Tools

- Chemical structures not stored in R-group format
  - R-group representation is not unique
- Tools to reconstruct data in R-group format did not exist
- Did not pursue further development of the algorithm
- Tools are improving and value of algorithm has increased

Statistical Problem



- No ordering of levels → Large space of models to navigate
- Standard **recursive partitioning** algorithms
  - Sort levels based on mean(Y); best partition must be along that sequence
  - No statistic analogous to the mean to apply to this problem

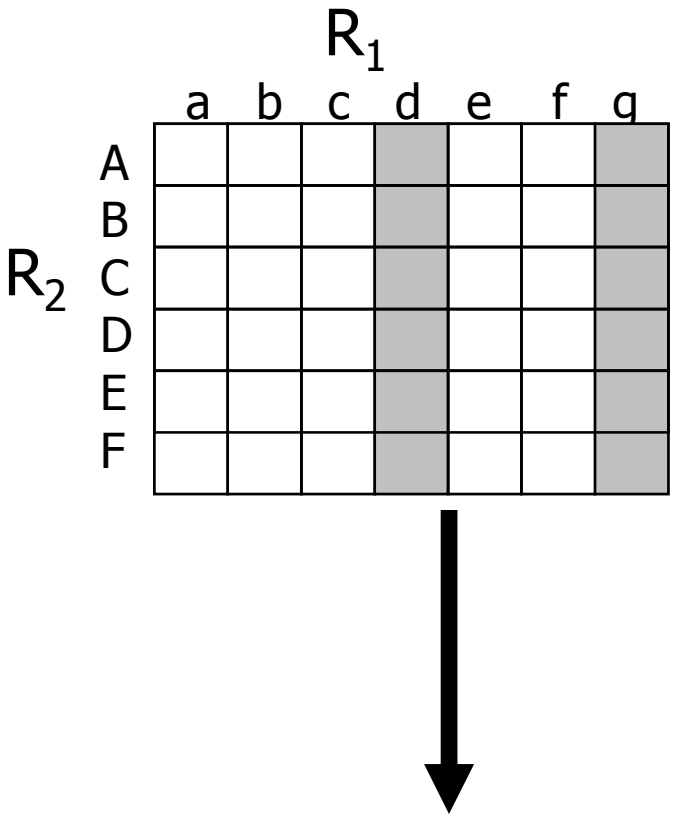
Relevant Literature

- Loh W-Y (2002) *Statistica Sinica* "Regression Trees With Unbiased Variable Selection and Interaction Detection."
  - Algorithm based on residuals
- Alexander WP, Grimshaw SD (1996) *JCGS* "Treed Regression."
  - Simple linear regression at each terminal node
- Friedman (1991) *Annals of Statistics* "Multivariate Adaptive Regression Splines."
- Chipman (2001) "Bayesian Treed Models."
  - MCMC probabilistic model selection

Possible algorithms

- Heuristic – simulated annealing, genetic algorithms
- Stochastic – Bayesian model selection
- Greedy - stepwise

Algorithm



- Build tree from the **bottom up** (as in agglomerative clustering)
- At each step, merge the two nodes that are "closest"
- Distance measure similar to Ward (1963) clustering algorithm

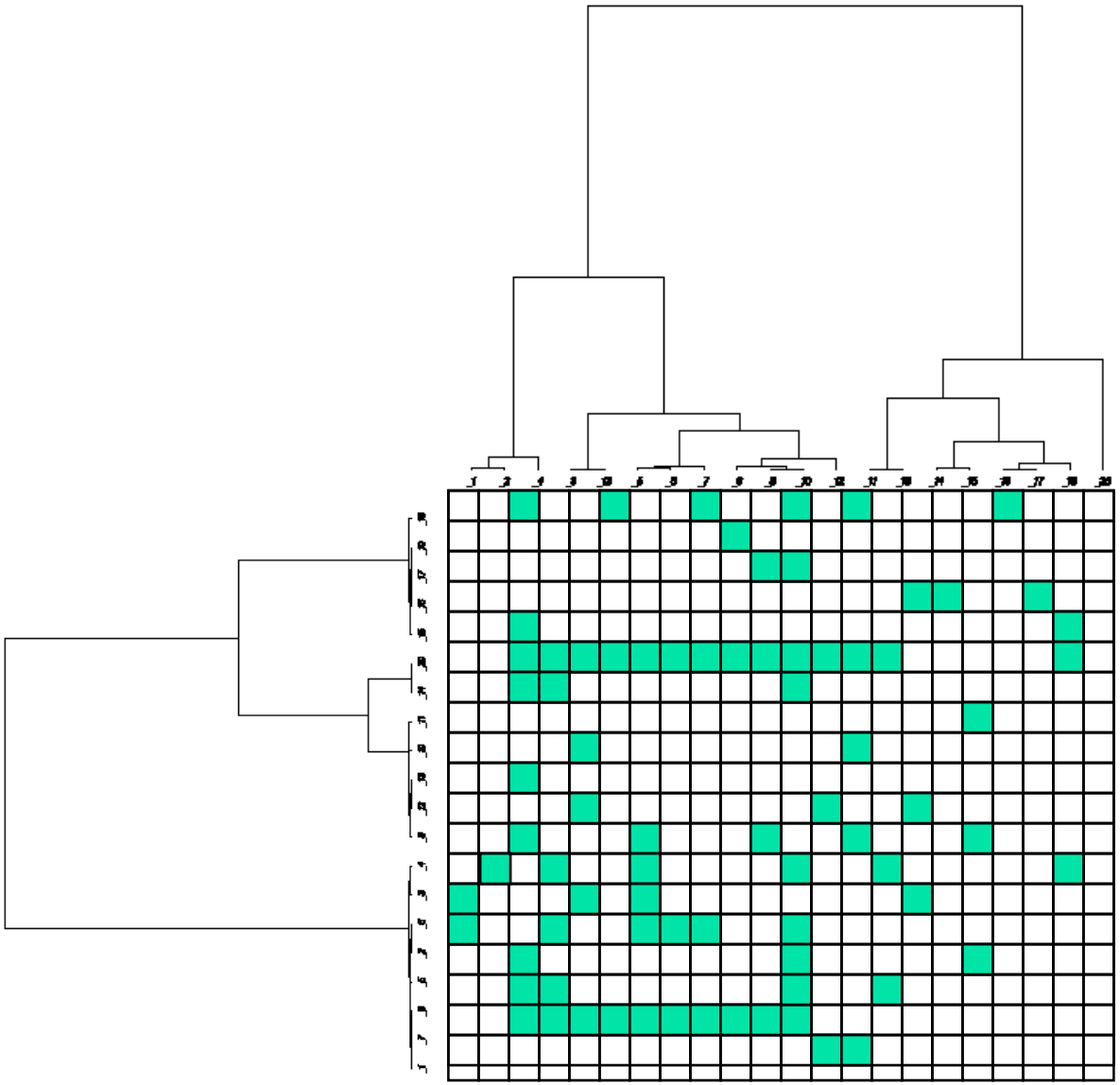
Distance(*d*,*g*) = (measure of fit from main effects ANOVA model on columns *d* and *g* only)

Algorithm details

$$D(C_i, C_j) = \frac{RSS(C_i + C_j) - RSS(C_i) - RSS(C_j)}{p_i + p_j - p_{ij}}$$

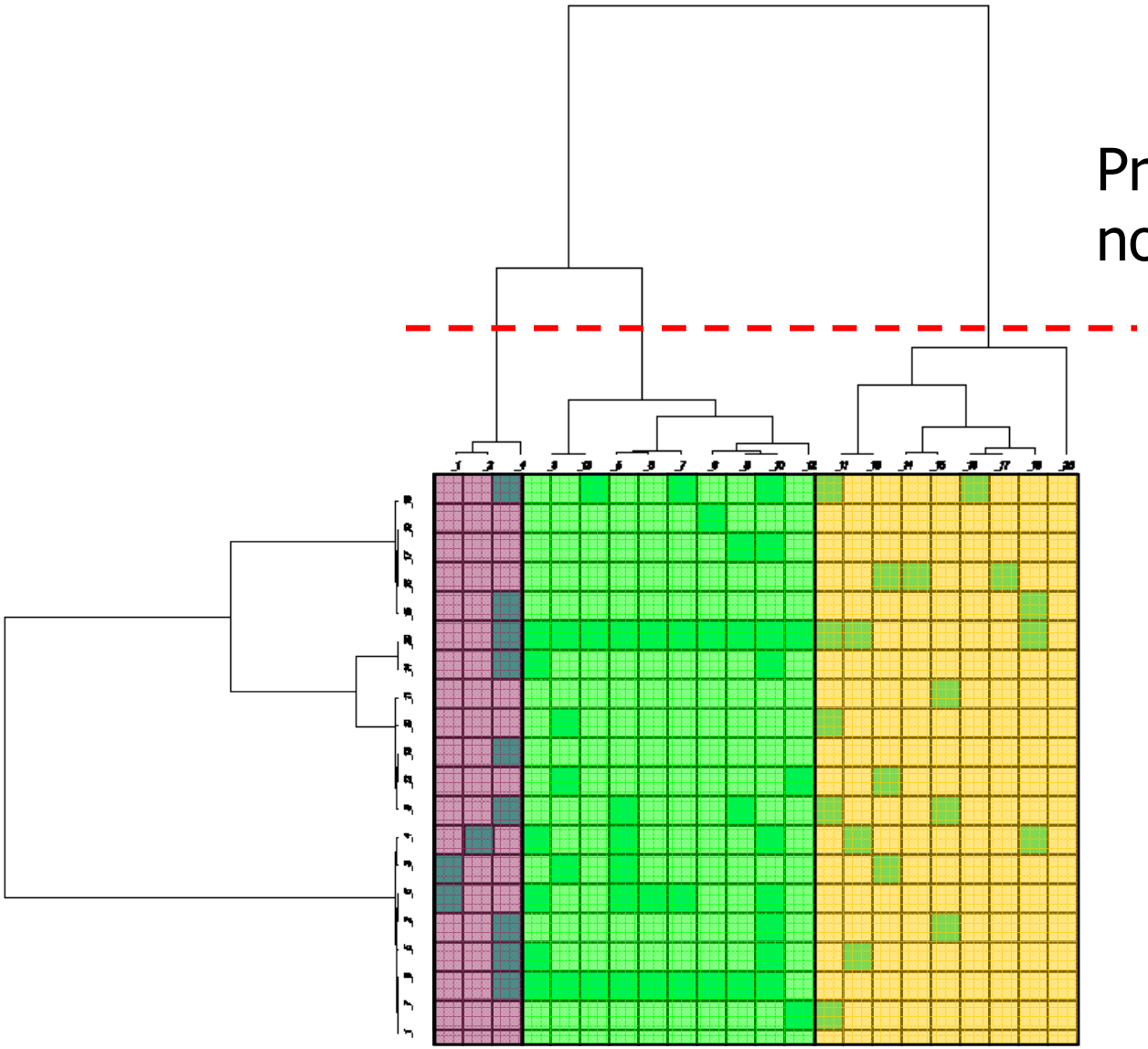
- C_i = Current cluster of one or more columns
- p_i = no. of parameters in main effects model on C_i
- $C_i + C_j$ ➔ New merged cluster from C_i and C_j
- D(C_i, C_j) = Numerator of F-test comparing simpler model $C_i + C_j$ with more complex model with C_i and C_j separate

ANOVA tree structure



- Current algorithm builds tree separately for rows and columns
- Prune the tree by cross-validation (leave out data and predict)

ANOVA tree structure

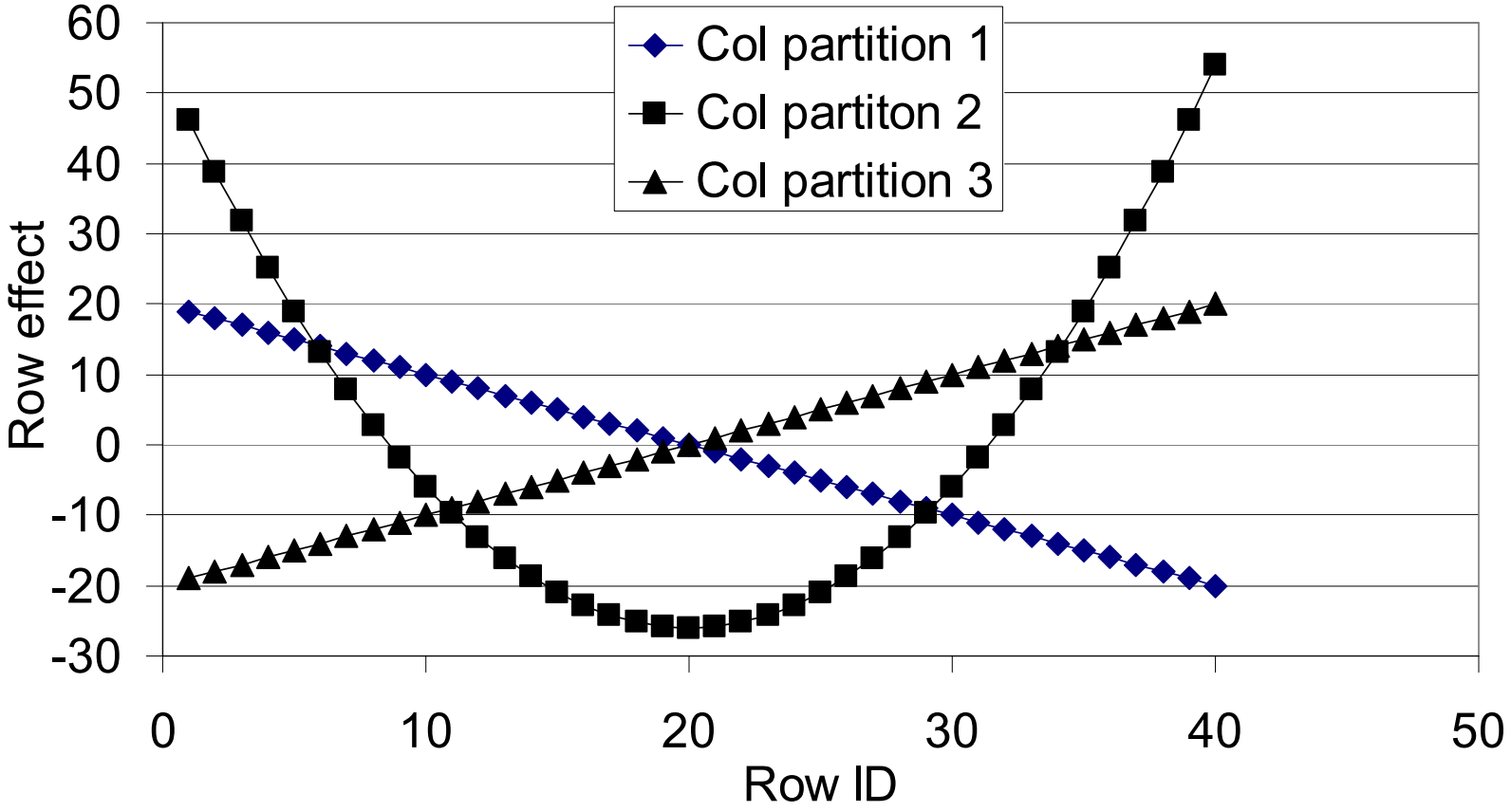


Pruned tree w/ 3 nodes

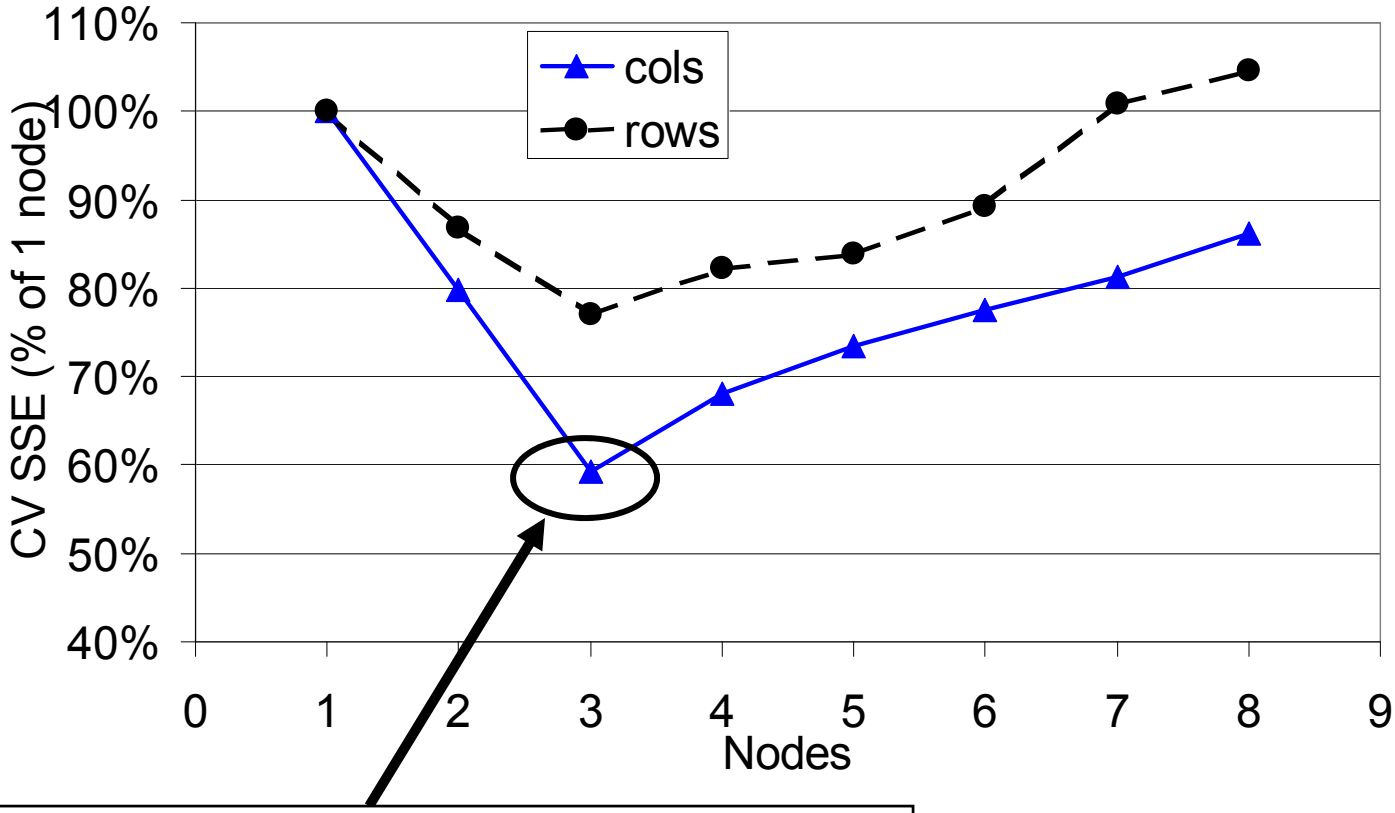
Artificial example

- 40 x 40: Row effects depend on three distinct column partitions
- 50% of cells empty (randomly)
- Will algorithm find the three partitions?

Artificial example

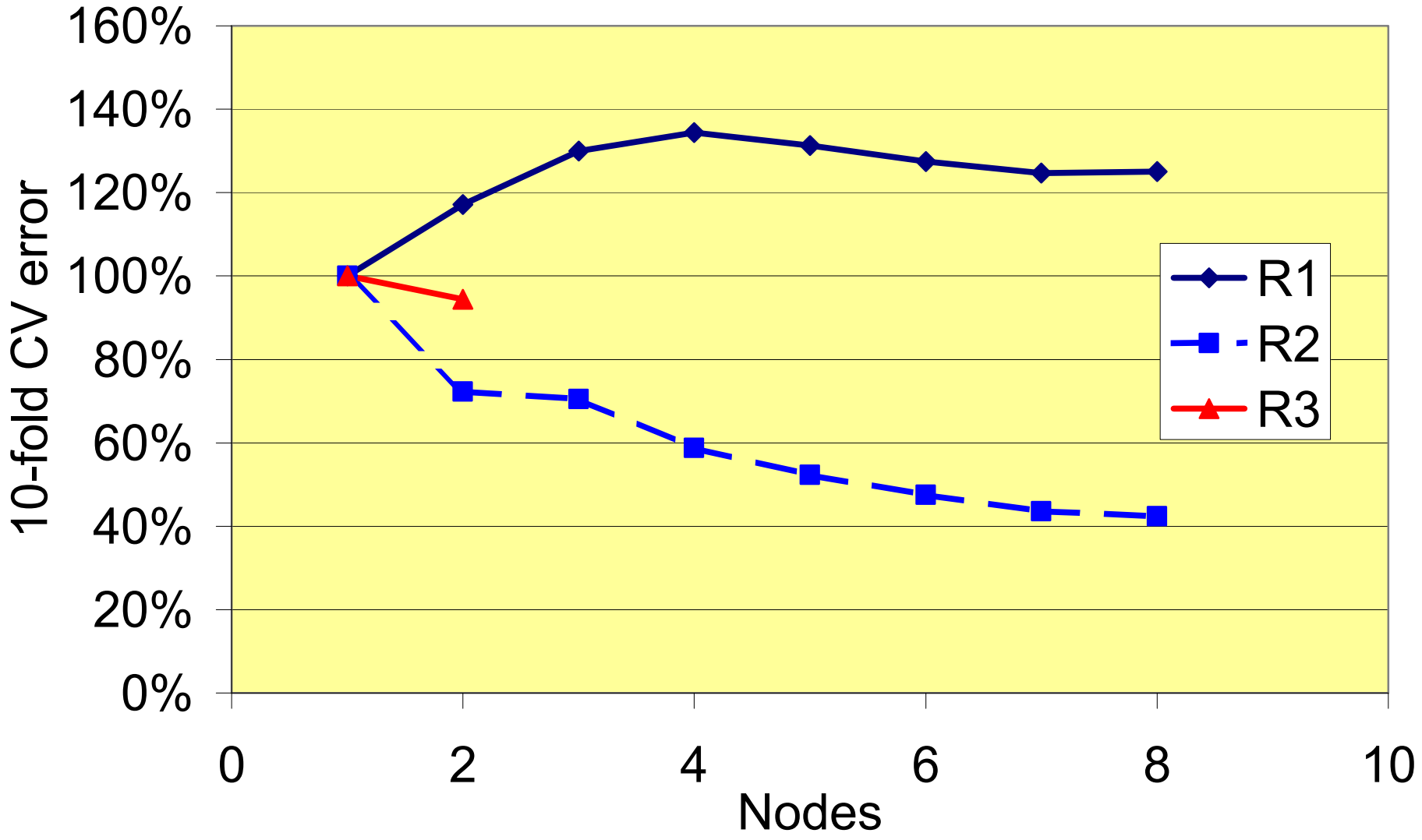


Results – Artificial example



- Prune column tree at 3 nodes
- Resulting partition matches simulation model exactly

Results – EP2 Data



Experimental design implications

- Typically, use model to predict empty cells; make compounds predicted to be good
- Additional compounds to inform the model; How?
  - Minimize entropy – multiple models?

		R_1							
		a	b	c	d	e	f	g	h
R_2	u								
	v								
	w								
	x								
	y								
	z								

Summary

- ANOVAtree an intuitively appealing model for interaction in large 2-way (or k-way) layout
- Need nonstandard fitting algorithm
- Basis for sequential experimental design