

Data Mining with Information Theory (DMIT)

SySc 531
Jan 4, 2022

Prof. Martin Zwick
zwick@pdx.edu

https://works.bepress.com/martin_zwick/

1

Jan 4, 2022

zwick@pdx.edu

1

Terminology

- DATA MINING = **exploratory** modeling (machine learning): finding relationships between variables in data
- INFORMATION THEORY + graph theory = **Reconstructability Analysis** (RA), a.k.a discrete multivariate modeling (**DMM**)
- **OCCAM** = RA software for course

2

- 2

Projects

- All projects must use OCCAM
- Projects can either (a) use OCCAM alone, or (b) use OCCAM + another data mining method & compare results. (a) is standard for this course. If you do (b), state-based (SB) OCCAM analysis is not required
- Support of other methods is not part of this course

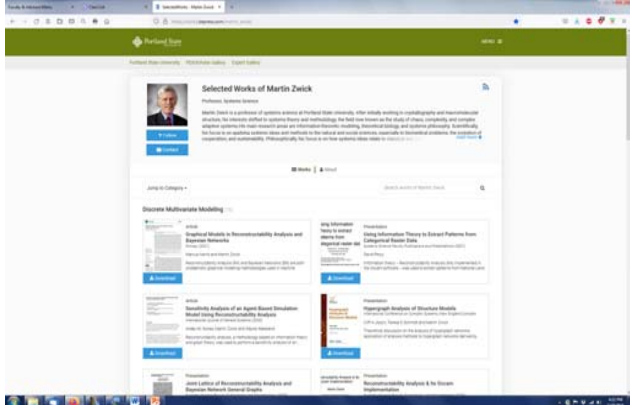
3

- 3

DMM section of MZ Selected Works:

resource for this course

https://works.bepress.com/martin_zwick/



4

[illegible]

More information on RA

- Review articles on SW page & Canvas
 - “Wholes & Parts in General Systems Methodology”
https://works.bepress.com/martin_zwick/52/
 - “An Overview of Reconstructability Analysis”
https://works.bepress.com/martin_zwick/57/
- Krippendorff, Klaus (1986). *Information Theory. Structural Models for Qualitative Data* (Quantitative Applications in the Social Sciences Monograph #62). New York: Sage Publications (at PSU Bookstore).
- *International Journal of General Systems*
- *Kybernetes*, Vol. 33, No. 5/6 2004: special RA issue

5

Past/present applications (some on DMM page)

- **BIOMEDICAL**
Gene-disease association, disease risk factors, gene expression, health care use & outcomes, medical records (dementia, diabetes, heart disease, prostate cancer, brain injury, primate health, surgery)
- **FINANCE-ECONOMICS-BUSINESS**
Stock market, bank loans, credit decisions, apparel analyses, market segmentation
- **SOCIAL-POLITICAL-ENVIRONMENTAL**
Socio-ecological interactions, wars, urban water use, rainfall, forest attributes
- **MATH-ENGINEERING**
Logic circuits, automata dynamics, optimization, neural networks, chip manufacturing, pattern recognition, decision analysis
- **OTHER**
Textual analysis, language analysis

6

Relation to other methods

- RA *related to* **log-linear** methods, **graphical modeling**, e.g., **Bayesian networks**, **logistic regression**, etc., competitive with NN, other machine learning methods
- RA is *explicitly designed for* **exploratory** search & has **unique features** not available in those methods
- RA is transparent: readily **interpretable**

7

Objectives & deliverables

ESSENTIAL GOALS

- To give you **concrete experience** of data mining
- To teach you **to use OCCAM** to analyze data
- To **analyze your data**

DELIVERABLE

- **Report & presentation** of your analysis

If you're ambitious, consider (this is totally optional!)

- **Compare** RA to another method
- Make your report into a **conference or journal paper**

8

Course prerequisites

- YOU HAVE DATA TO ANALYZE of a type suitable for OCCAM (rectangular, adequate sample size)
- Some knowledge of probability/statistics
- Discrete Multivariate Modeling (DMM, SySc 551) NOT prerequisite.
- DMM will probably be taught in academic year 2022-2023.
- DMIT presents only theory needed to understand OCCAM output

9

Software tool: OCCAM: dmit.sysc.pdx.edu

10

OCCAM server

- [dmit](http://dmit.sysc.pdx.edu) is the OCCAM server for this course; openly accessible over the web
- If class overloads dmit, another server will be made available
- OCCAM is now open source; at <https://github.com/occam-ra/occam>

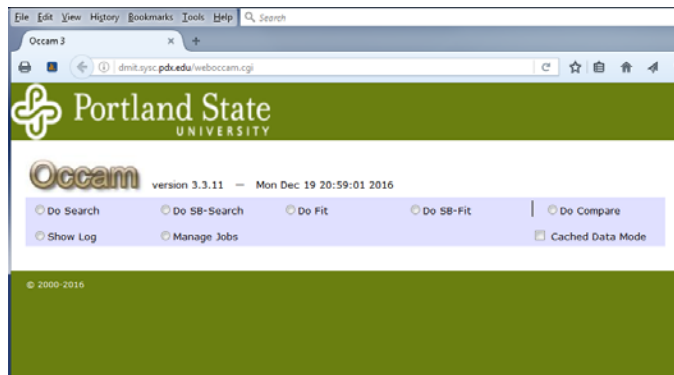
11

OCCAM manual

OCCAM: A Reconstructability Analysis Program (Organizational Complexity Computation and Modeling)	Occam User's Manual	1/19/2021	2
Project Director: Martin Zwick zwick@pdx.edu	Table of Contents		
Past Programmers: Heather Alexander, Joe Fuson, Kenneth Willett ¹	I. FOR INFORMATION ON RECONSTRUCTABILITY ANALYSIS		3
Systems Science Program Portland State University Portland OR 97207	II. ACCESSING OCCAM		3
This manual was last revised on 19 January 2021. Occam version 3.4.1, copyright 2006-2020.	III. SEARCH INPUT		4
	IV. SEARCH OUTPUT		15
	V. STATE-BASED SEARCH		19
	VI. FIT INPUT		20
	VII. FIT OUTPUT		24
	VIII. STATE-BASED FIT		31
	IX. SHOW LOG		31
	X. MANAGE JOBS		31
	XI. FREQUENTLY ASKED QUESTIONS		31
	XII. ERROR AND WARNING MESSAGES		36
	XIII. KNOWN BUGS & INFELICITIES: LIMITATIONS		37
	XIV. PLANNED BUT NOT-YET-IMPLEMENTED FEATURES		38
	APPENDIX 1. REBINNING (RECORDING)		40
	APPENDIX 2. MISSING VALUES IN THE DATA		42
	APPENDIX 3. ADDITIONAL PARAMETERS IN THE INPUT FILE		42
	APPENDIX 4. ZIPPING THE INPUT FILE		43
	APPENDIX 5. COMPARE MODE		43
	APPENDIX 6. CACHED DATA MODE		45
			12

¹ Ken Willett totally rewrote earlier versions of Occam. His version was originally called "Occam2" to distinguish it from three earlier Occam incarnations. The "2" has sadly been dropped in this manual.

OCCAM initial screen



13

OCCAM actions

- **Search** = **exploratory** modeling, examine many models, find best or good ones (OCCAM actions Search, SB-Search)
- **Fit** = look at one model in detail & use for prediction; **confirmatory** modeling (OCCAM actions Fit, SB-Fit)

Compare: specific to one PhD project

Show log, Manage jobs: managerial functions

14

RA model types (1/3)

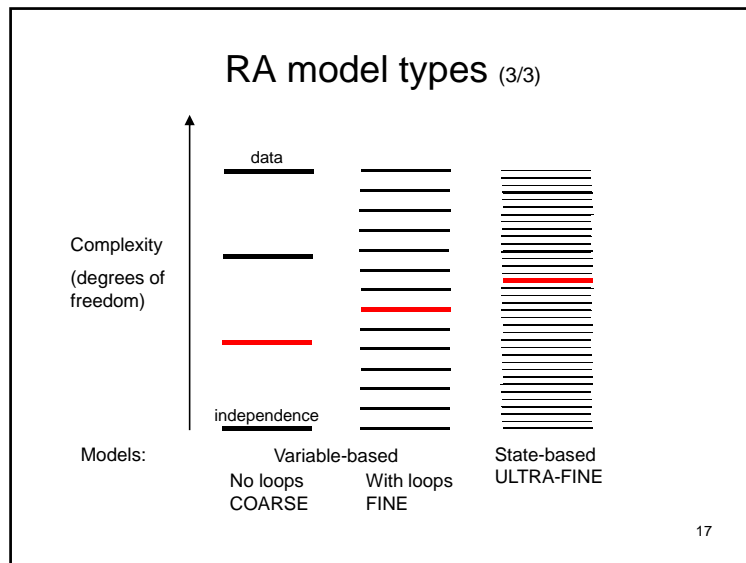
- Models are
 - **Variable-based** (VB), or
 - **State-based** (SB)
- A VB or SB model
 - Is **loopless**, or
 - Has **loops**

15

RA model types (2/3)

- **Variable-based**
 - loopless **many variables COARSE**
ABC:ABZ simple prediction, *feature selection*
FAST
 - with loops **up to 100s variables FINE**
ABC:AZ:BZ better prediction, *detailed analysis*
SLOW (exercise caution in runs)
- **State-based** **< 10 variables ULTRA-FINE**
 - usually w' loops best (most precise) prediction
ABC:Z: A₁Z:B₂Z₁ **V. SLOW** (exercise super caution)

16



- ### What a model is
- Some models:
 - ABC:ABZ VB loopless
 - ABC:AZ:BZ VB with loops
 - ABC:Z: A₁Z:B₂Z₁ SB (usually with loops)
 - A model = **set of relations**, separated by colon “:”
 - Relations are **predictive** (ABZ) or **not** (ABC = “IV”)
 - Multiple** predictive relations must be **integrated**
 - A relation is a **joint probability** distribution
- 18

- ### Lattice of models
- Top** model = data = “saturated model” = most complex, e.g., ABCZ
 - Bottom** model = “independence model” = least complex
 - With IV | DV distinction (directed): ABC:Z
 - Without IV | DV distinction (neutral) : A:B:C: ...
 - Or, bottom could be uniform distribution
 - Search up** from bottom, **or down** from top,
 - Or up/down from some other model
 - Starting** model (& search direction) vs. **reference** model
- 19

OCCAM search (VB, SB) input page

Portland State

Occam

version 3.3.11 — Fri Jan 1 15:50:55 2016

Do Fit Do Search Do SB-Fit Do SB-Search Do Compare Show Log Manage Jobs

Data File: No file selected.

Starting Model:

Composition Method: ☒ Standard ☐ Back Projection (Fourier)

Reference Model: ☒ Default ☐ Top ☐ Bottom ☐ Starting Model

Models to Consider: ☒ All ☐ Loopless ☐ Dependent ☐ Chain

Search Direction: ☐ Up ☐ Down ☐ Chain

Search Settings: ☒ BIC ☐ AIC ☐ Information ☐ Alpha ☐ % Correct

When Searching, Prefer: ☒ Larger Values ☐ Smaller Values

Search Width: 3

Search Levels: 7 (leave blank to use settings from data file)

Report Settings: ☒ Information ☐ Alpha ☐ dDF ☐ Level ☐ % Correct ☐ BIC ☐ AIC

In Report, Sort: ☒ Descending ☐ Ascending

Include in Report: ☒ In ☒ dA ☒ Alpha ☒ % dDF ☒ dAIC ☒ dBIC

☐ SP-based Transmission ☒ Incremental Alpha

☒ % Correct ☒ % Coverage of Data ☒ % Missing in Test

☐ Return data in spreadsheet format

☒ Print option settings (but don't print variable definitions)

☐ Use inverse notation for models

Run in Background, Email Results To:

Subject line for email (optional):

Send

20

Variables & cases

- Variables: so far up to about 500 IVs
Need at least 5
Best is bigger, but not too big (>10, <100)
IV/DV (input/output) distinction: directed
No distinction: neutral This course: **directed** only
- Will usually (not always) name the DV “Z” (or “X”)
- Cases (sample size): so far up to 6.5×10^6
Need at least 100
Best is bigger, but not too big (>1K, <<1M)

21

Data (e.g., IVs = A,B,C; DV = Z)

- Contingency table or No frequency, just
frequency (A_i, B_j, C_k, Z_l) **cases X variables**

				frequency
A ₀	B ₀	C ₀	Z ₀	13
A ₀	B ₀	C ₀	Z ₁	2
A ₀	B ₀	C ₁	Z ₀	9
A ₀	B ₀	C ₁	Z ₁	11
...

	A	B	C	Z
case ₁	A ₀	B ₀	C ₀	Z ₀
case ₂	A ₁	B ₂	C ₃	Z ₁
...				
case _N	A ₀	B ₀	C ₀	Z ₀

- Cases = individuals, instances, time or space values
- Variables = IVs & DV are **nominal**
If continuous, **bin** (e.g., with Excel binning program)
(In *variants* of standard RA, DV can be continuous)

22

Sample input file: weisdorf.txt) (for demo#1)

```

:action
:search

:nominal
season, 2,1,s
absentee, 2,0,b
squest, 8,1,c
valued, 12,1,d
yrbuilt, 12,1,e
zip, 40,0,f
ref, 16,0,w
normcof, 16,2,w

:inc-frequency

:data
0 0 1 2 9 39 1 1
0 1 2 3 9 19 0 0
0 1 1 3 5 19 2 2
0 1 1 4 7 19 0 0
0 0 1 2 0 1 3 2
0 1 1 1 0 1 3 2
0 1 1 3 9 39 1 1
0 1 3 5 7 19 1 0
0 1 1 2 1 39 0 0
0 0 0 1 5 39 0 0
0 1 1 1 1 39 1 1
0 1 2 3 6 19 1 1
0 0 2 2 0 1 1 1
0 1 2 4 5 19 2 2
0 1 2 3 1 39 2 2
0 1 2 3 6 19 2 2
0 1 0 1 3 39 1 0
0 1 1 3 5 39 1 1
0 1 0 2 0 39 1 1
0 1 1 2 1 39 0 0
0 1 1 4 3 19 2 2

```

23

Preparing an OCCAM input file

- Include all variables you have (you can always tell OCCAM to ignore some); make DV the *last* variable
- In binning variables (by yourself or with **utility program**), **3 bins** is a reasonable default, but it's better to use **more bins** (e.g., **12**) since you can rebin (aggregate bins) on the fly in the OCCAM input file; see Appendix of OCCAM manual on rebinning
- Code **missing values** as “.”
- Use **comments** “#” in input file

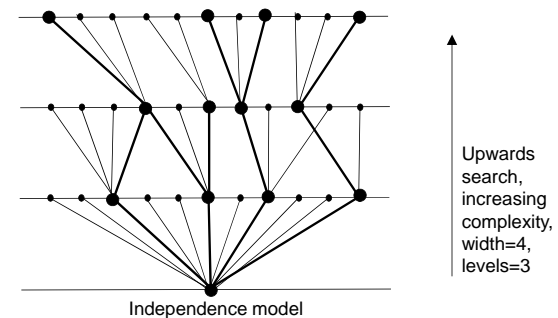
24

OCCAM search output (1/4)

- Search many models: select best **w** ('width') models at each of **l** 'levels'
- Output all selected **w*l** models
- OCCAM **summarizes** the best of these

25

OCCAM search output (2/4)



26

OCCAM search output (3/4)

- OCCAM **summarizes**:
- Of width*levels models found in search, it indicates the **best three** by different criteria:
 - (i) highest ΔBIC
 - (ii) highest ΔAIC
 - (iii) highest information with OK **p-values** (both **cumulative** & **incremental**)

27

Search output for weisdorf.txt VB models without loops (search sort on bic)(demo#1)

```

Search output for weisdorf.txt
VB models without loops (search sort on bic)(demo#1)

Search output summary
Sample Size      150018
B(Data)          9.61708
P(27)            7.48711
S(DV)            2.35084
T(EV(DV))        0.200864
DV in use (k)    A C D E
DV               A

Resolving levels:
L = 4 new models, 0 kept; 0 total kept; 0 kb memory used; 0.0 seconds, 0.0 total
L = 4 new models, 0 kept; 11 total models; 7 total kept; 820 kb memory used; 0.1 seconds, 0.1 total
L = 3 new models, 0 kept; 10 total models; 10 total kept; 838 kb memory used; 0.1 seconds, 0.1 total
L = 1 new models, 1 kept; 10 total models; 11 total kept; 850 kb memory used; 0.1 seconds, 0.1 total
L = 0 new models, 7 kept; 10 total models; 11 total kept; 850 kb memory used; 0.1 seconds, 0.1 total

ID  MODEL  Level  B      dfP  dfR      Alpha  Inf      HSB(DV)  HSBIC  dBIC      Inc.Alpha  Prop.  MC(Data)  Resever
13  ACCEX  4      9.6170  12440  44901.1157  0.0000  1.00000000  9.8444  -19329.8947  -335043.4345  1.0000  0  39.7733  12.9326
10* IVIACEX 3      9.6613  2865  34314.0074  0.0000  0.77984404  6.6633  28584.0074  18.4274  0.0000  5  39.1474  98.9853
8* IVIACEX 3      9.6654  2426  33417.8688  0.0000  0.75940049  6.4393  28147.8688  1995.5202  0.0000  6  39.1448  77.2727
9* IVIACEX 3      9.6676  2943  32943.0094  0.0000  0.74860550  6.3970  28053.0094  -14205.4855  0.0000  3  39.2012  100.0000
7* IVIACEX 2      9.4785  225  30539.7243  0.0000  0.69404710  5.9304  30089.7243  27644.3718  0.0000  3  39.8595  100.0000
6* IVIACEX 2      9.4921  315  27574.9048  0.0000  0.62466834  5.3544  26944.9048  23804.2504  0.0000  2  39.9013  100.0000
5* IVIACEX 2      9.7456  245  15641.4391  0.0000  0.34002256  3.0762  15151.4391  11711.6279  0.0000  3  39.7437  100.0000
4* IVIACEX 1      9.7470  100  13554.9322  0.0000  0.39240016  2.5126  14504.9322  14595.0398  0.0000  1  39.7752  100.0000
3* IVIACEX 1      9.7604  15  12561.7145  0.0000  0.25546228  2.4393  12531.7145  12382.1576  0.0000  1  39.7437  100.0000
2* IVIACEX 1      9.7617  150  12309.9942  0.0000  0.27976330  2.3904  12009.9942  10514.4246  0.0000  1  39.8070  100.0000
1* IVIACEX 0      9.6179  0  0.0000  1.0000  0.60000000  0.0000  0.0000  0.0000  0.0000  0  39.7437  100.0000
ID  MODEL  Level  B      dfP  dfR      Alpha  Inf      HSB(DV)  HSBIC  dBIC      Inc.Alpha  Prop.  MC(Data)  Resever

Best Model(s) by dBIC:
7* IVIACEX 2      9.4785  225  30539.7243  0.0000  0.69404710  5.9304  30089.7243  27644.3718  0.0000  3  39.8595  100.0000
Best Model(s) by HSBIC:
7* IVIACEX 2      9.4785  225  30539.7243  0.0000  0.69404710  5.9304  30089.7243  27644.3718  0.0000  3  39.8595  100.0000
Best Model(s) by Information, with all Inc. Alpha < 0.05:
10* IVIACEX 3      9.6613  2865  34314.0074  0.0000  0.77984404  6.6633  28584.0074  18.4274  0.0000  5  39.1474  98.9853

Run time: 31.267388 seconds
    
```

28

[illegible]

ID	Model	Level	1	dOF	Df1	Alpha	Inf	wdt(DV)	dRMC	Inc	Alpha	Prog.	IC(Data)	ICover
20	TVAp1-IC1-IC1-IC1-IC1	7	9.3221	20	120.2026	0.00000000	0.20478939	20.4766	0.0206	-0.7921	0.9050	19	72.4057	32.7778
21	TVAp1-IC1-IC1-IC1-IC1	7	9.3224	21	120.0422	0.00000000	0.20448930	20.4489	0.7894	0.9053	0.1393	18	72.6435	38.8437
22	TVAp1-IC1-IC1-IC1-IC1	7	9.3224	20	120.2026	0.00000000	0.20448930	20.4441	0.0236	-0.9713	0.9059	17	71.9340	32.7778
23	TVAp1-IC1-IC1-IC1-IC1	7	9.3224	20	120.2026	0.00000000	0.20448930	20.4441	0.0236	-0.9713	0.9059	17	71.9340	32.7778
10	TVAp1-IC1-IC1-IC1-IC1	6	9.3332	17	112.5347	0.00000000	0.19146949	19.1469	0.7894	0.9052	0.1438	16	72.1699	34.7778
11	TVAp1-IC1-IC1-IC1-IC1	6	9.3332	17	112.5347	0.00000000	0.19146949	19.1469	0.7894	0.9052	0.1438	16	72.1699	34.7778
12	TVAp1-IC1-IC1-IC1-IC1	6	9.3361	16	111.9804	0.00000000	0.19079474	19.0740	0.9056	0.1389	0.0053	14	72.4057	32.7778
13	TVAp1-IC1-IC1-IC1-IC1	5	9.4338	14	107.4649	0.00000000	0.18300000	18.3050	0.7894	0.9052	0.0037	13	71.2244	36.3839
14	TVAp1-IC1-IC1-IC1-IC1	5	9.4338	14	107.4649	0.00000000	0.18300000	18.3050	0.7894	0.9052	0.0037	13	71.2244	36.3839
15	TVAp1-IC1-IC1-IC1-IC1	5	9.4338	14	107.4649	0.00000000	0.18300000	18.3050	0.7894	0.9052	0.0037	13	71.2244	36.3839
16	TVAp1-IC1-IC1-IC1-IC1	5	9.4338	14	107.4649	0.00000000	0.18300000	18.3050	0.7894	0.9052	0.0037	13	71.2244	36.3839
17	TVAp1-IC1-IC1-IC1-IC1	5	9.4338	14	107.4649	0.00000000	0.18300000	18.3050	0.7894	0.9052	0.0037	13	71.2244	36.3839
18	TVAp1-IC1-IC1-IC1-IC1	5	9.4338	14	107.4649	0.00000000	0.18300000	18.3050	0.7894	0.9052	0.0037	13	71.2244	36.3839
19	TVAp1-IC1-IC1-IC1-IC1	5	9.4338	14	107.4649	0.00000000	0.18300000	18.3050	0.7894	0.9052	0.0037	13	71.2244	36.3839
20	TVAp1-IC1-IC1-IC1-IC1	5	9.4338	14	107.4649	0.00000000	0.18300000	18.3050	0.7894	0.9052	0.0037	13	71.2244	36.3839
21	TVAp1-IC1-IC1-IC1-IC1	5	9.4338	14	107.4649	0.00000000	0.18300000	18.3050	0.7894	0.9052	0.0037	13	71.2244	36.3839
22	TVAp1-IC1-IC1-IC1-IC1	5	9.4338	14	107.4649	0.00000000	0.18300000	18.3050	0.7894	0.9052	0.0037	13	71.2244	36.3839
23	TVAp1-IC1-IC1-IC1-IC1	5	9.4338	14	107.4649	0.00000000	0.18300000	18.3050	0.7894	0.9052	0.0037	13	71.2244	36.3839
24	TVAp1-IC1-IC1-IC1-IC1	5	9.4338	14	107.4649	0.00000000	0.18300000	18.3050	0.7894	0.9052	0.0037	13	71.2244	36.3839
25	TVAp1-IC1-IC1-IC1-IC1	5	9.4338	14	107.4649	0.00000000	0.18300000	18.3050	0.7894	0.9052	0.0037	13	71.2244	36.3839
26	TVAp1-IC1-IC1-IC1-IC1	5	9.4338	14	107.4649	0.00000000	0.18300000	18.3050	0.7894	0.9052	0.0037	13	71.2244	36.3839
27	TVAp1-IC1-IC1-IC1-IC1	5	9.4338	14	107.4649	0.00000000	0.18300000	18.3050	0.7894	0.9052	0.0037	13	71.2244	36.3839
28	TVAp1-IC1-IC1-IC1-IC1	5	9.4338	14	107.4649	0.00000000	0.18300000	18.305						

31

32

OCCAM **Fit** (VB, SB) input page

33

Fit output

- 34

Fit output for dementia05.txt
model IV:ApZ:EdZ:CZ (has loop) (demo#3)

35

Input file for SB-search (for demo#4)
dementia05ApEdC.txt: reduce #IVs to 3

36

SB search output (demo#4)

ID	MODEL	Level	B	dDP	dLR	Alpha	Inf	4dH(DV)	dAIC	dBIC	Inc.Alpha	Prog.	NC(Data)	%cover
16	IV:AgEd0C02:ApEd0C22:ApEd0I:Ap11:C22:2	5	4.3897	5	86.7978	0.0000	0.95071864	14.7861	76.7978	56.5491	0.0949	12	69.5755	100.0000
15	IV:AgEd0C02:ApEd0I:Ap0C02:Ap11:C22:2	5	4.3906	5	86.2578	0.0000	0.94480434	14.6941	76.2878	56.0091	0.1337	12	70.2472	100.0000
14	IV:AgEd0C22:ApEd0C02:ApEd0I:Ap11:C22:2	5	4.3906	5	86.2410	0.0000	0.94480205	14.6912	76.2410	55.9923	0.1362	13	69.5755	100.0000
13*	IV:AgEd0C22:ApEd0I:Ap11:C22:2	4	4.3935	4	84.5686	0.0000	0.92630190	14.4063	76.5686	60.3697	0.0337	9	69.5755	100.0000
12	IV:AgEd0C02:ApEd0I:Ap11:C22:2	4	4.3944	4	84.0051	0.0000	0.92012899	14.3103	76.0051	59.8062	0.1044	10	69.5755	100.0000
11	IV:AgEd0I:Ap0C22:Ap11:Ed0C22:2	4	4.3981	4	83.6256	0.0000	0.91597252	14.2457	75.6256	59.4266	0.0567	8	69.5755	100.0000
10*	IV:AgEd0I:Ap11:C22:2	3	4.3989	3	81.3650	0.0000	0.89121192	13.8606	75.3650	63.2158	0.0002	7	69.5755	100.0000
9*	IV:AgEd0C22:ApEd0I:Ap11:2	3	4.4011	3	80.0648	0.0000	0.87698704	13.6390	74.0648	61.9153	0.0002	6	69.5755	100.0000
8*	IV:AgEd0I:Ap0C22:Ap11:2	3	4.4013	3	79.9909	0.0000	0.87616134	13.6265	73.9909	61.8417	0.0004	7	69.5755	100.0000
7*	IV:AgEd0I:Ap11:2	2	4.4216	2	69.0475	0.0000	0.74534205	11.5919	64.0475	55.9480	0.0000	4	69.5755	100.0000
6*	IV:AgEd0C22:Ap11:2	2	4.4232	2	67.1062	0.0000	0.73503174	11.4316	63.1062	55.0067	0.0000	2	66.9811	100.0000
5*	IV:Ap11:Ed02:2	2	4.4243	2	66.4783	0.0000	0.72815386	11.3246	62.4783	54.3788	0.0001	4	69.5755	100.0000
4*	IV:Ap11:2	1	4.4504	1	51.0910	0.0000	0.55961325	8.7034	49.0910	45.0413	0.0000	1	66.9811	100.0000
3*	IV:Ag0C22:2	1	4.4727	1	37.9865	0.0000	0.41607561	6.4710	35.9865	31.9367	0.0000	1	62.2642	100.0000
2*	IV:Ag0C22:2	1	4.4770	1	35.4733	0.0000	0.38554880	6.0429	33.4733	29.4236	0.0000	1	58.7264	100.0000
1*	IV:2	0	4.5374	0	0.0000	1.0000	0.00000000	0.0000	0.0000	0.0000	0.0000	0	52.1226	100.0000
ID	MODEL	Level	B	dDP	dLR	Alpha	Inf	4dH(DV)	dAIC	dBIC	Inc.Alpha	Prog.	NC(Data)	%cover
Best Model(s) by dBIC:														
10*	IV:AgEd0I:Ap11:C22:2	3	4.3989	3	81.3650	0.0000	0.89121192	13.8606	75.3650	63.2158	0.0002	7	69.5755	100.0000
Best Model(s) by dAIC:														
16	IV:AgEd0C02:ApEd0C22:ApEd0I:Ap11:C22:2	5	4.3897	5	86.7978	0.0000	0.95071864	14.7861	76.7978	56.5491	0.0949	12	69.5755	100.0000
Best Model(s) by Information, with all Inc. Alpha < 0.05:														
13*	IV:AgEd0C22:ApEd0I:Ap11:C22:2	4	4.3935	4	84.5686	0.0000	0.92630190	14.4063	76.5686	60.3697	0.0337	9	69.5755	100.0000

37

SB-search output

- Example of a best model from SB-Search:

Best model by BIC criterion

IV:Z: Ap_0Ed_0Z : Ap_1Z : C_2Z (note interaction effect)

For this data, get about same uncertainty reduction & %correct, but **2 df simpler**

In other data often get **higher uncertainty reduction & %correct** for about same df

38

Non-standard or enhanced RA

Non-standard

- Time series analysis
- Continuous DVs (not for this course)
- Set-theoretic data (not for this course)

Enhanced

- Validation (training-test data splits)
- Inter-method comparison

39

Time series analysis

Can similarly do spatial (e.g., GIS) analysis

	A	B	C		A	B	C		U	V	W	X	Y	Z
t-4	--	--	--		--	--	--		--	--	--	--	--	--
t-3	--	--	--		0	1	2		--	--	--	--	--	--
t-2	--	--	--		3	4	5		0	1	2	3	4	5
t-1	U	V	W		6	7	8		3	4	5	6	7	8
t	X	Y	Z		9	10	11		6	7	8	9	10	11
mask					original data (numbers label variables)				transformed data $XYZ(t) = ABC(t)$ $UVW(t) = ABC(t-1)$					

Enhanced RA (for directed systems)

(Totally optional for this course)

Validation: **training/test** data **splits**
or 3-way splits; 5- or 10-fold validation

Comparison with other data mining **methods**
e.g., Logistic Regression, Support Vector Machines,
Bayesian Networks, ...

41

RA framework (1/2)

bold = typical RA use; **blue** = in OCCAM; **red** = other pgms

1. VARIABLE	nominal (discrete: binary or multi-valued)
	ordinal (discrete)
	quantitative (bin continuous values for IVs)
2. SYSTEM	directed (deterministic/stochastic) (supervised learning)
	neutral (unsupervised learning)
3. DATA	Information-theoretic (IRA)
	frequency/probability distribution
	function ('k-systems' & 'u-systems' RA)
	set-theoretic (SRA) mapping, relation

42

RA framework (2/2)

bold = typical RA use; **blue** = in OCCAM; **red** = other pgms

4. PROBLEM	reconstruction (decomposition)
	confirmatory
	exploratory
	exhaustive (look at all models)
	heuristic (search lattice of models)
	identification (composition)
5. METHOD	variable-based (VB)
	state-based (SB)
	latent variable-based (LVB)

43

Questions?

LOOKING FORWARD

TO AN EXCITING, CHALLENGING, &

PRODUCTIVE COURSE !!

44