

Portland State University

PDXScholar

Systems Science Faculty Publications and
Presentations

Systems Science

7-2018

Introduction to Reconstructability Analysis

Martin Zwick

Portland State University, zwick@pdx.edu

Follow this and additional works at: https://pdxscholar.library.pdx.edu/sysc_fac



Part of the [Logic and Foundations Commons](#), and the [Systems Architecture Commons](#)

Let us know how access to this document benefits you.

Citation Details

Zwick, Martin, "Introduction to Reconstructability Analysis" (2018). *Systems Science Faculty Publications and Presentations*. 125.

https://pdxscholar.library.pdx.edu/sysc_fac/125

This Presentation is brought to you for free and open access. It has been accepted for inclusion in Systems Science Faculty Publications and Presentations by an authorized administrator of PDXScholar. Please contact us if we can make this document more accessible: pdxscholar@pdx.edu.

Introduction to Reconstructability Analysis

Martin Zwick

Professor of Systems Science

zwick@pdx.edu

http://www.pdx.edu/sysc/research_dmm.html

ISSS 2018, Corvallis, July 22-27

WHAT IS RA?

- **Reconstructability Analysis** (RA) = a probabilistic graphical modeling methodology
- RA = Info theory + Graph theory
- Graphs, applied to data, are **models**:
- node = variable; link = relationship
- RA uses not only graphs (a link joins 2 nodes), but **hypergraphs** (a link can join **>2** nodes)

WHY RA MIGHT BE OF INTEREST TO YOU 1/2

- Can detect **many-variable** or **non-linear** interactions not hypothesized in advance, i.e., it is explicitly designed for **exploratory** search
- **Transparent** (not black box), easily interpretable
- Designed for **nominal** variables
- Can also analyze **continuous** variables via **binning**
- **Prediction**/classification, **clustering**/network models
- **Time series**, **spatial** analyses
- Overlaps common **statistical** & **machine-learning** methods (but has unique features)

WHY RA MIGHT BE OF INTEREST TO YOU 2/2

- **Web-accessible user-friendly** software (OCCAM)
- Analyses at **3 levels of refinement**:
 - coarse (very fast, *many* variables)
 - fine (slower, 100s of variables)
 - ultra-fine (slow, < 10 variables)
- **Standard application**: frequency data $f(A_i, B_j, C_k, Z_l)$
- Variety of **non-standard capabilities**
 - Data: set-theoretic relations & mappings
 - Predict continuous variables
 - Integrate multiple inconsistent data sets
 - Regression-like Fourier version

PAST/PRESENT RA APPLICATIONS

- ***BIOMEDICAL***

Gene-disease association, disease risk factors, gene expression, health care use & outcomes, **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, genetic algorithm & neural network preprocessing, chip manufacturing, pattern recognition, decision analysis

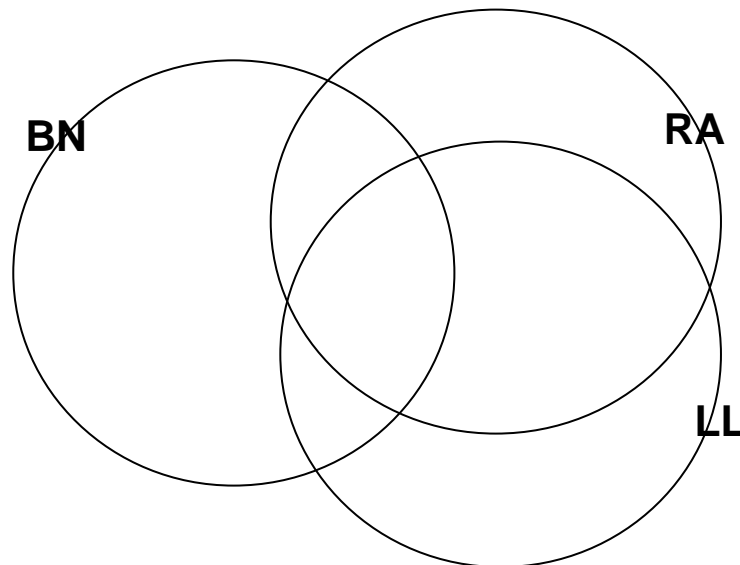
- ***OTHER***

Textual analysis, language analysis

OVERLAP with STATISTICAL, MACHINE LEARNING METHODS

Relation to **log linear** (LL) (& logistic regression)
models & to **Bayesian networks** (BN)

Where methods overlap, they are **equivalent**



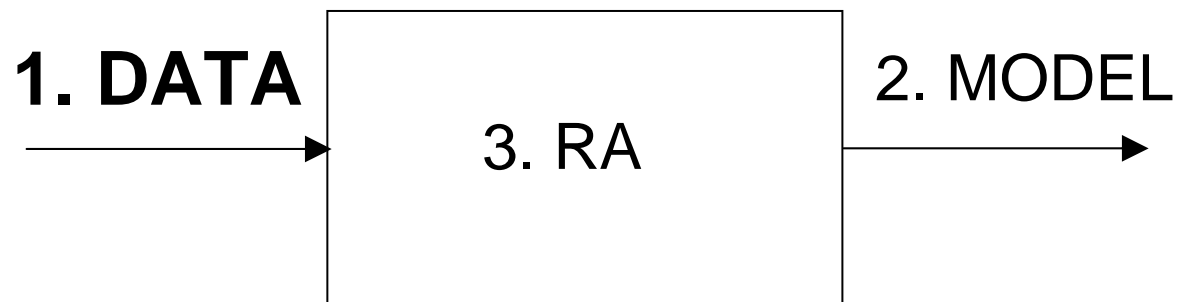
1. input **data** to RA

- **form** of data (cases X variables)
- data cases indexed by **individual, time, space**

2. **model** output from RA

3. **basics** of RA

4. for **more information**



FORM OF DATA

Variables

- Type: **nominal**; **bin** if continuous (continuous DV needn't be binned)
- Number: few variables to 100s (in principle, to 1000s or more)
- Distinctions:

directed system

- *Predict/classify* a DV (output) from IVs (inputs)

neutral system

- No IV-DV distinction: association, **clustering** / network

FORM OF DATA

- frequency(A_i, B_j, C_k, Z_l) or individual cases

				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
...	—
				N

N = sample size

	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 are indexed by
 individual (in a population),
 time, or
 space

$$\text{frequency}(ABCZ) / N = p_{\text{data}}(ABCZ)$$

DATA CASES INDEXED BY *INDIVIDUAL* (#ID)

ID ,0,0,ID
 APOE ,2,1,Ap
 Gender ,2,1,Sx
 Education ,3,1,Ed
 AgeLastExam ,3,1,Ag
 rs1801133 ,3,1,A
 rs3818361 ,4,1,B
 rs7561528 ,3,1,C
 rs744373 ,3,1,D
 rs6943822 ,3,1,E
 rs4298437 ,3,1,F
 rs7012010 ,3,1,G
 rs11136000 ,3,1,H
 rs10786998 ,4,1,J
 rs11193130 ,4,1,K
 rs610932 ,3,1,L
 rs3851179 ,3,1,M
 rs3764650 ,4,1,N
 rs3865444 ,4,1,P
 Dementia ,2,2,Z

DEMENTIA EXAMPLE
 Z = 0 no disease; Z = 1 disease

#ID	Ap	Sx	Ed	Ag	A	B	C	D	E	F	G	H	J	K	L	M	N	P	Z
101	0	0	2	2	1	1	0	1	2	2	1	1	2	0	1	1	2	2	1
103	0	0	2	1	0	2	2	0	1	1	1	2	2	0	1	1	0	1	0
111	0	1	2	1	2	2	1	1	0	1	1	2	1	1	2	2	0	1	0
112	0	0	2	2	2	2	1	1	1	2	1	1	0	2	2	0	0	2	0
118	0	1	0	2	2	2	2	0	0	1	1	1	.	.	1	1	0	2	0
120	0	1	2	2	1	2	1	1	0	1	1	2	1	1	1	2	0	.	1
121	0	0	2	2	2	2	1	1	2	0	0	0	2	0	1	1	1	.	1
122	0	0	1	2	1	2	1	1	2	0	0	2	2	0	1	1	1	1	0
123	0	0	2	2	2	2	2	0	1	1	0	0	2	0	2	1	0	1	1

...

DATA CASES INDEXED BY TIME

	X	Y	Z
t-4	--	--	--
t-3	0	1	2
t-2	3	4	5
t-1	6	7	8
t	9	10	11

original data

A	B	C	X	Y	Z
--	--	--	--	--	--
--	--	--	--	--	--
0	1	2	3	4	5
3	4	5	6	7	8
6	7	8	9	10	11

transformed data

Values are labels for variable states at particular times

XYZ = **generating variables**

Apply **mask** (here # lags = 2) to data

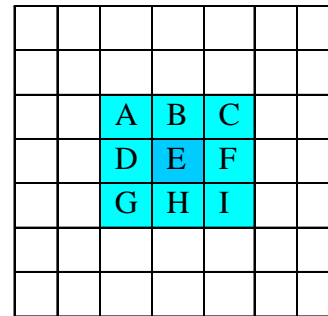
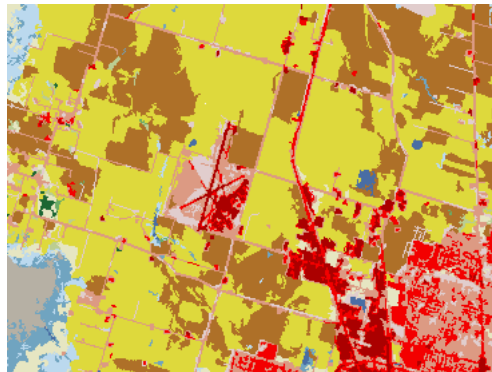
Mask adds lagged variables, $ABC(t) = XYZ(t-1)$

E.g., $A(t-1) = X(t-2)$, labeled 3

Masking: time series → **atemporal** sample

DATA CASES INDEXED BY SPACE : 1 generating variable

A,14,1,A
 B,14,1,B
 C,14,1,C
 D,14,1,D
E,14,2,E
 F,14,1,F
 G,14,1,G
 H,14,1,H
 I,14,1,I



Moore neighborhood

E = DV

A,B,C,D,F,G,H,I = IVs

IVs & DV have 14 possible states

#A	B	C	D	E	F	G	H	I
71	71	71	71	71	71	71	71	71
71	71	71	71	71	71	71	71	71
71	71	71	71	71	71	71	71	71
71	71	71	71	71	71	71	71	71
71	71	71	71	71	71	71	71	71
71	71	71	71	71	71	71	71	71
71	71	71	71	71	71	71	71	71
71	71	71	71	71	71	71	71	71
71	71	71	95	71	95	71	71	71
95	71	95	95	71	95	71	71	71
95	95	95	95	95	71	71	71	95
71	95	95	90	95	95	71	95	95
95	95	90	90	71	95	95	95	95
95	90	90	90	95	90	95	95	90

...

1. input data to RA

2. *model output from RA*

model = structure (hypergraph) applied to data (GT)

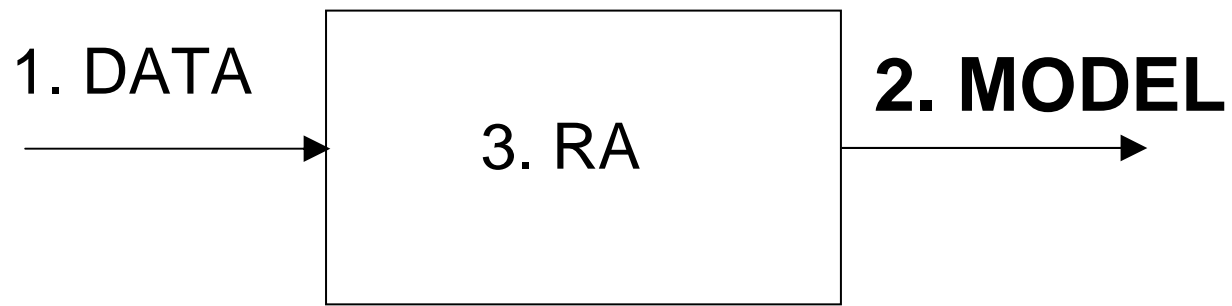
types of structures (GT)

selecting a model (IT)

model = (conditional) probability distribution (IT)

3. basics of RA

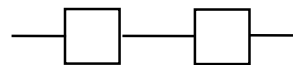
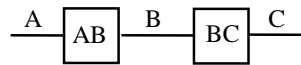
4. for [more information](#)



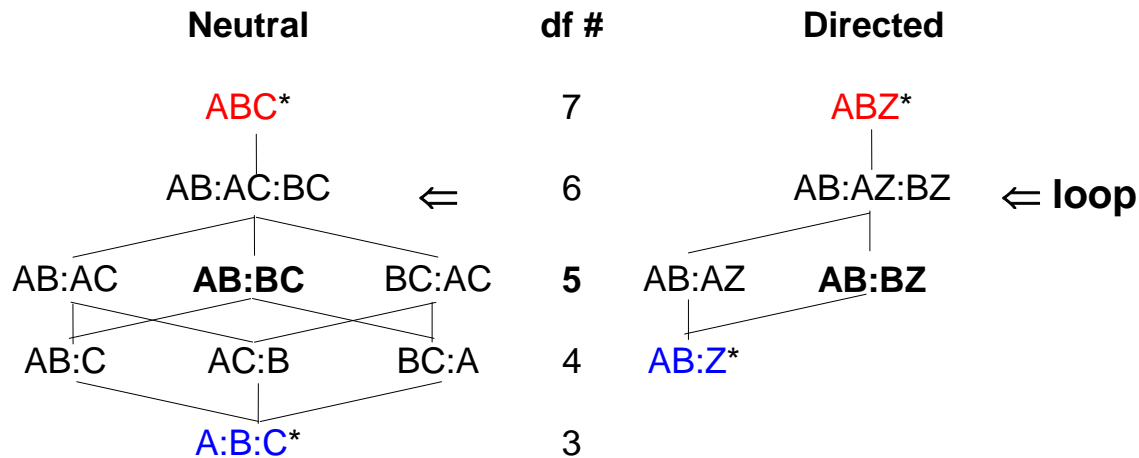
MODEL = STRUCTURE APPLIED TO DATA

A structure (graph or hypergraph) is a set of relationships (GT)

Specific structure **AB:BC** General structure

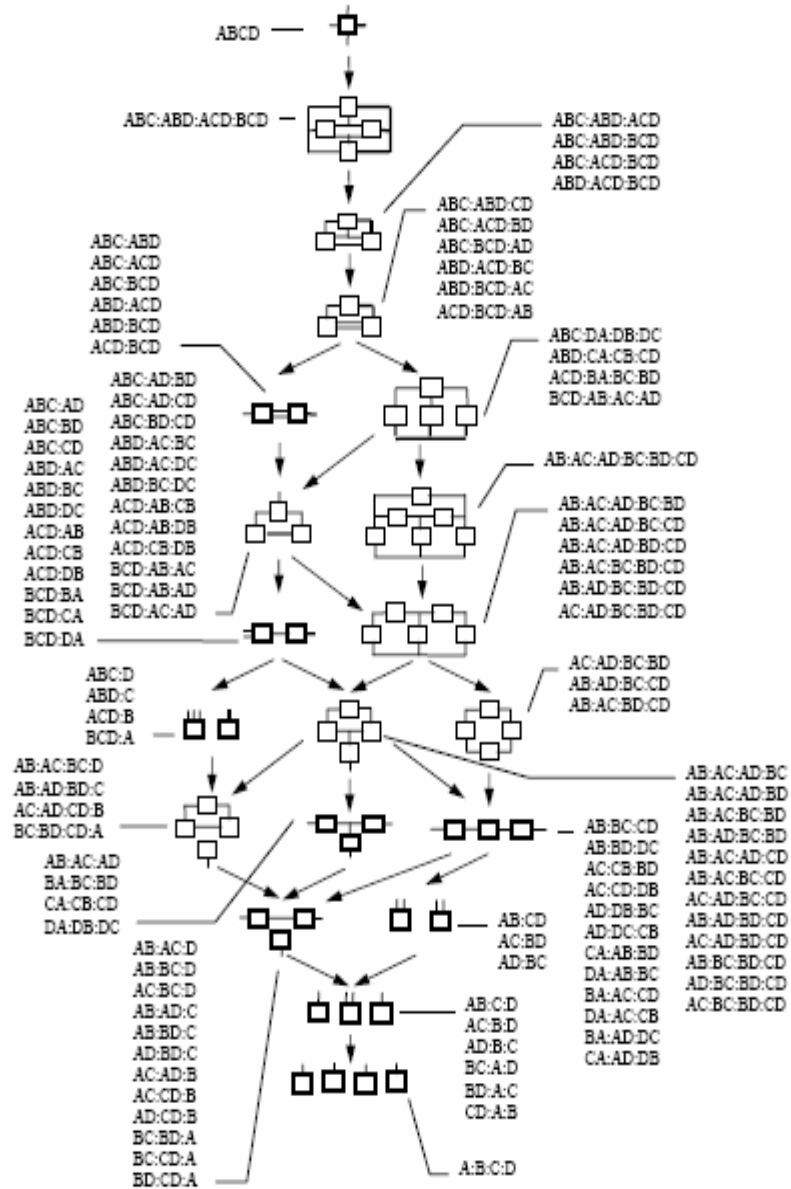
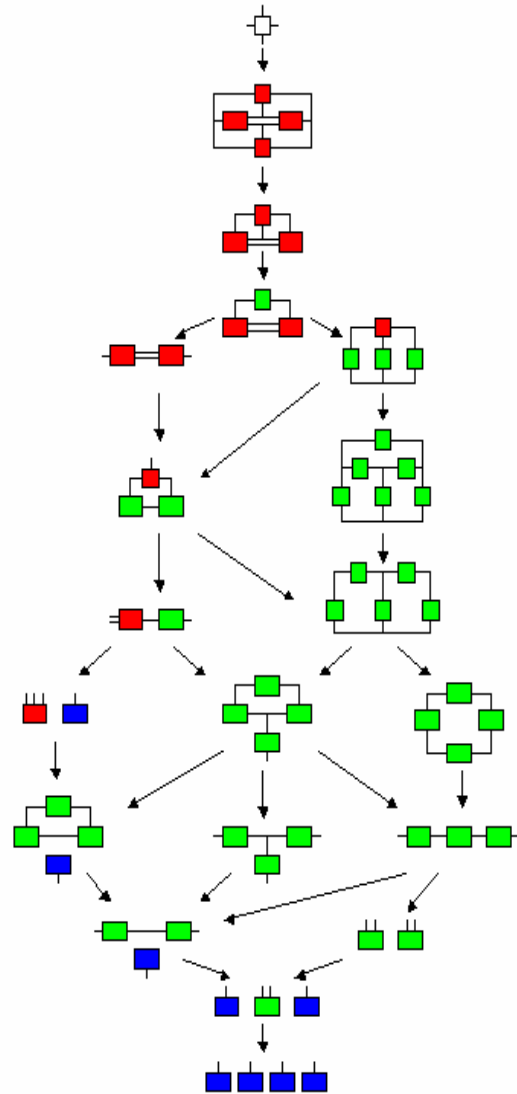


LATTICE OF SPECIFIC STRUCTURES (3 variables)



* Reference model is **data** or **independence**
 # df (degrees of freedom) values are for binary variables

STRUCTURES 4 variables (GT)



STRUCTURES (GT)

Combinatorial explosion

# variables	3	4	5	6
# general structures	5	20	180	16,143
# specific structures	9	114	6,894	7,785,062
(where 1 variable is DV)	5	19	167	7,580
(1 DV, no loops)	4	8	16	32

NEED **INTELLIGENT HEURISTICS** TO SEARCH LATTICE

Can analyze 100s of variables, & for simple models, many more.

TYPES OF STRUCTURES (GT)

FOR **PREDICTION / CLASSIFICATION** (directed system)

- **Variable-based**

- no loops

- IV:ACZ

- many* variables (**fast**) [*coarse*]

- simple prediction, **feature selection**

- with loops

- IV:ABZ:BCZ

- up to 100s of variables (slow) [*fine*]

- better prediction

- **State-based**

- IV:Z: A₁B₁Z : B₂C₃Z₁

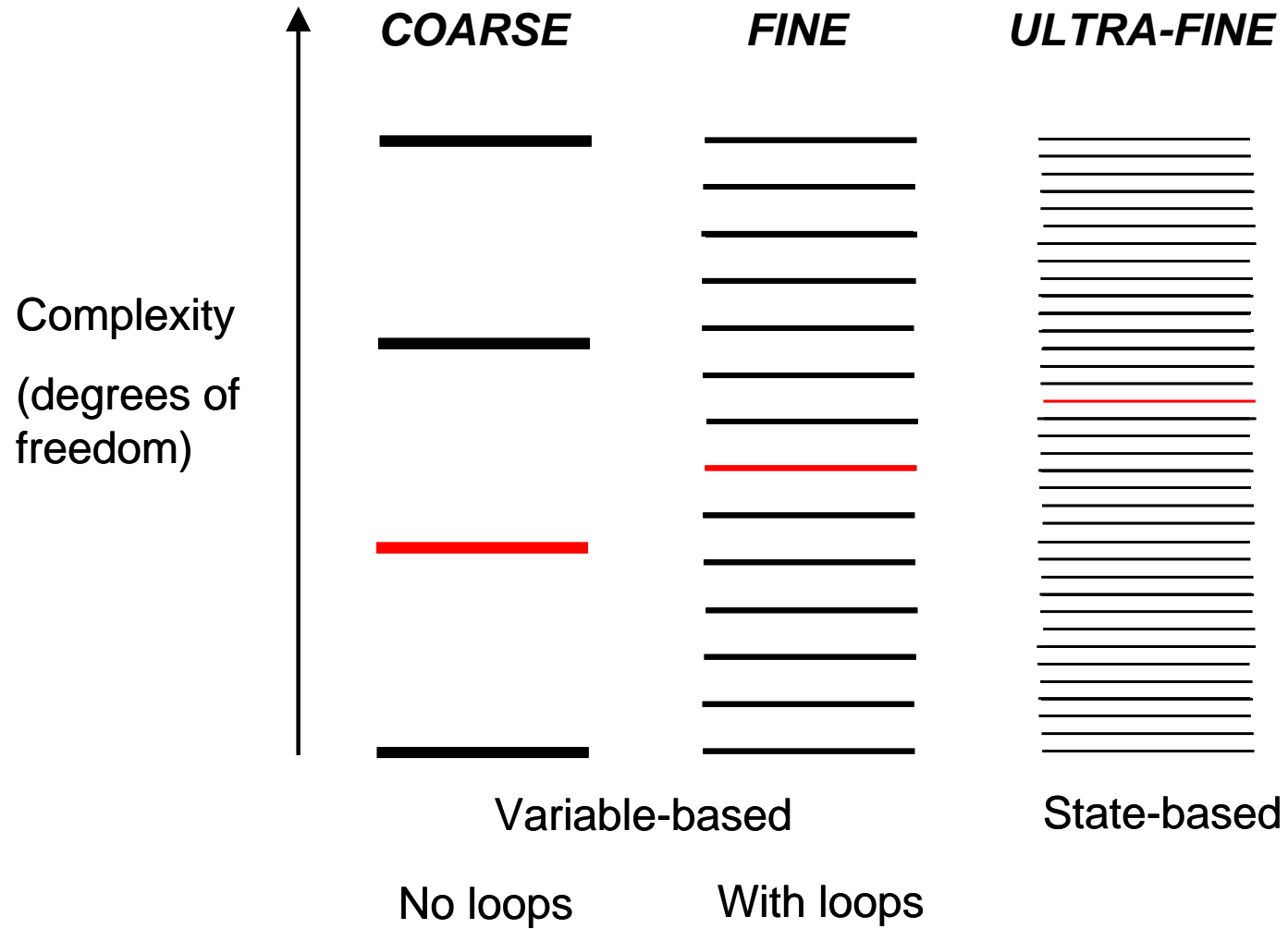
- < 10 variables (**v. slow**); [*ultra-fine*]

- best prediction; detailed models

“IV” = ABC (all IVs); Z = DV

All directed system models include an IV component

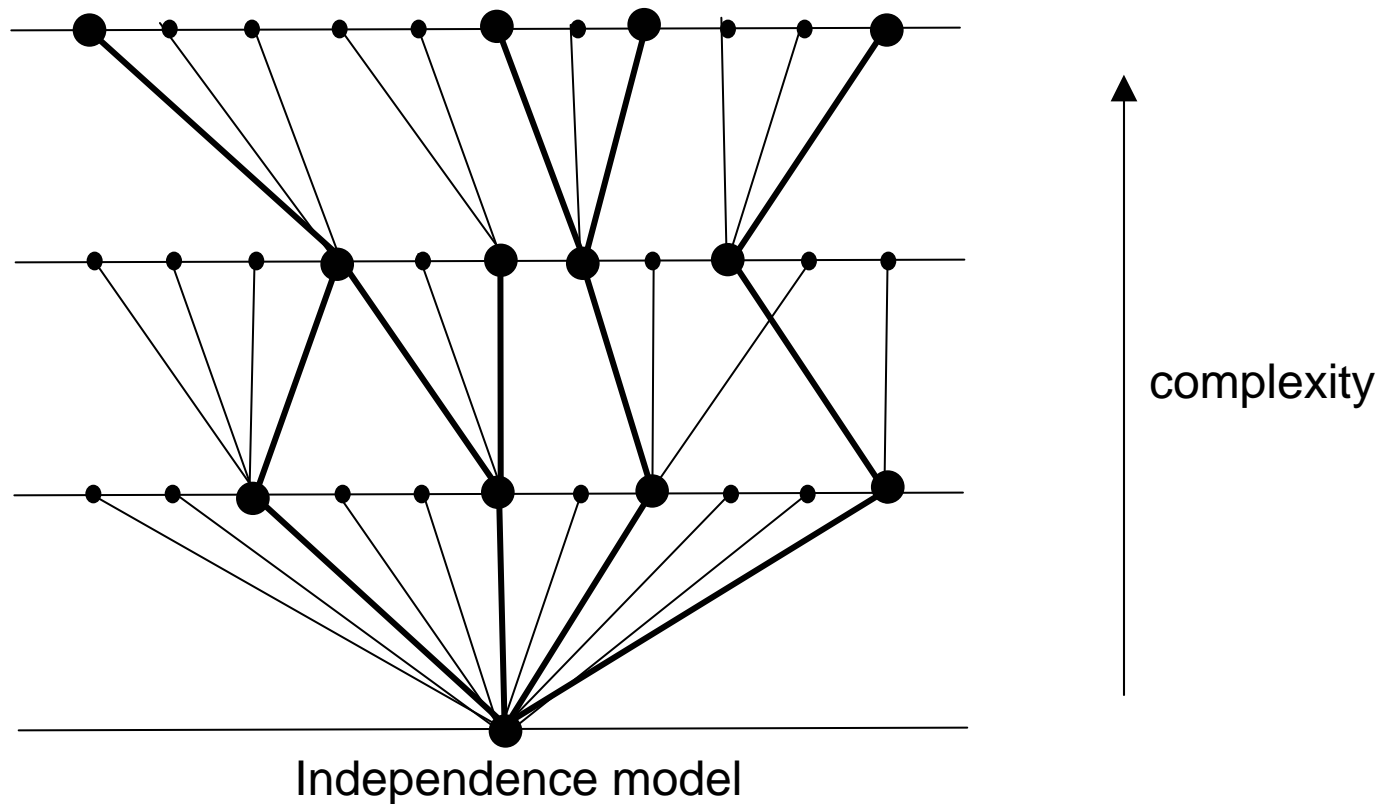
TYPES OF STRUCTURES (GT)



SEARCHING LATTICE OF STRUCTURES

beam search, levels = 3, width = 4 (node = model)

(there are many other search algorithms)



MODEL = PROBABILITY DISTRIBUTION (IT)

for **directed** system: *conditional* distribution

for **neutral** system: *joint* distribution

gotten by **applying data** to a **structure**

Directed system:

- Model = **calculated** *conditional* probability distribution, e.g., $p_{IV:AZ:BZ}(Z_i | A_i B_j C_k)$
- Distribution gives **rule** to **predict** DV (Z) from IVs (A,B,C) (e.g., rule = 0 means predict Z_0)

SELECTING A MODEL (IT)

1. High **information** (or low **error**) in model

For directed system

- *Info-theory measure: high ΔH , reduction of uncertainty of DV*
- *Generic measure: high %correct, accuracy of prediction*

2. Low **complexity**: df, degrees of freedom

3. Information \leftrightarrow complexity **tradeoff**

- Statistical **significance** (Chi-square p-values)
- **Integrated** measures: AIC, BIC
(Akaike & Bayesian Information Criteria)
- BIC a **conservative** selection criterion

UNCERTAINTY REDUCTION: SIMPLE EXAMPLE

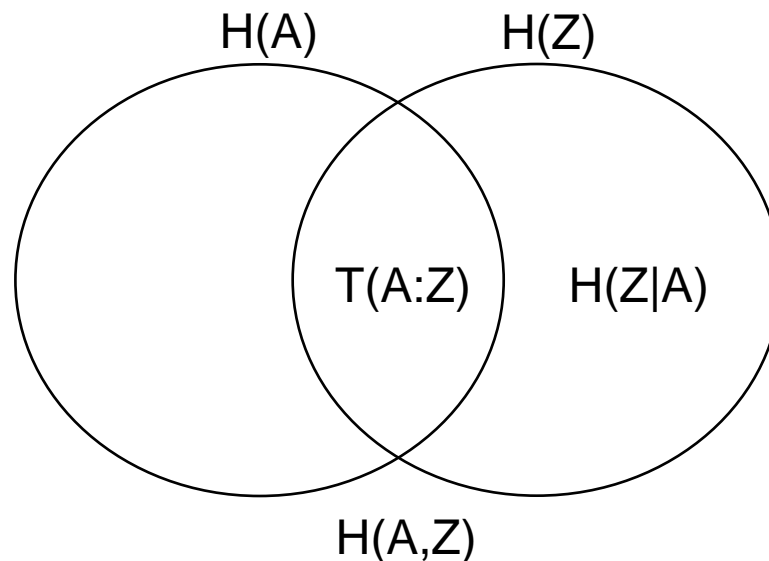
2 variables: $IV=A$; $DV = Z$; $T(A:Z)$ =mutual information (*association*)

- *Uncertainty reduction* is like variance explained

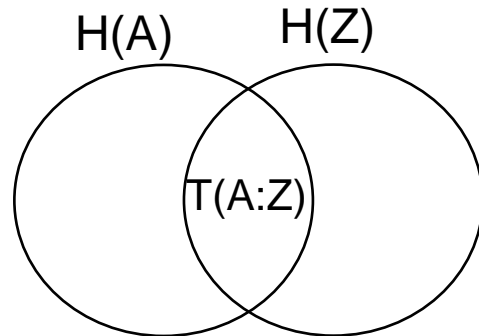
Model AZ = predict Z , i.e., reduce $H(Z)$, by knowing A

- Uncertainty *reduced* = $T(A:Z)$; uncertainty *remaining* = $H(Z|A)$

$\Delta H = T(A:Z) / H(Z)$ *fractional uncertainty reduction* (will express in %)



UNCERTAINTY REDUCTION: SIMPLE EXAMPLE



	Z ₀	Z ₁	
A ₀	.67*.5	.33*.5	.5
A ₁	.33*.5	.67*.5	.5
df=3	.5	.5	

- $p(Z_1)/p(Z_0) = 1:1$, not knowing A \rightarrow 2:1 or 1:2, knowing A
- $\Delta H(Z) = T(A:Z) / H(Z) = 8\%$
- 8% reduction in uncertainty is *large* (unlike variance!)

SELECTING A MODEL DEMENTIA EXAMPLE

<u>Criterion</u>	<u>model</u>	<u>ΔH(%)</u>	<u>Δdf</u>	<u>%c</u>	<u>ΔBIC</u>
------------------	--------------	--------------	------------	-----------	-------------

Variable-based (with loops)

BIC	IV: $A_p Z : E_d Z : K Z$	16	5	70	59
-----	---------------------------	----	---	----	----

p-value	IV: $A_p Z : E_d Z : K Z : C Z : L Z$	18	9	71	
---------	---------------------------------------	----	---	----	--

AIC	IV: $B A_p Z : E_d Z : K Z : C Z$	20	11	72	
-----	-----------------------------------	----	----	----	--

State-based

BIC	(model below; each interaction = 1 df)	20	6	72	81
-----	--	----	---	----	----

IV:Z: $A_{p_1} Z : E_{d_0} Z : K_2 Z : A_{p_0} E_{d_2} C_2 Z : A_{p_0} E_{d_1} C_2 K_1 Z : A_{p_0} E_{d_1} C_0 K_1 Z$

Models integrate multiple predicting interactions

IV = $A_p E_d C K L \dots$ (all the independent variables);

$\%c(IV:Z) = 52$

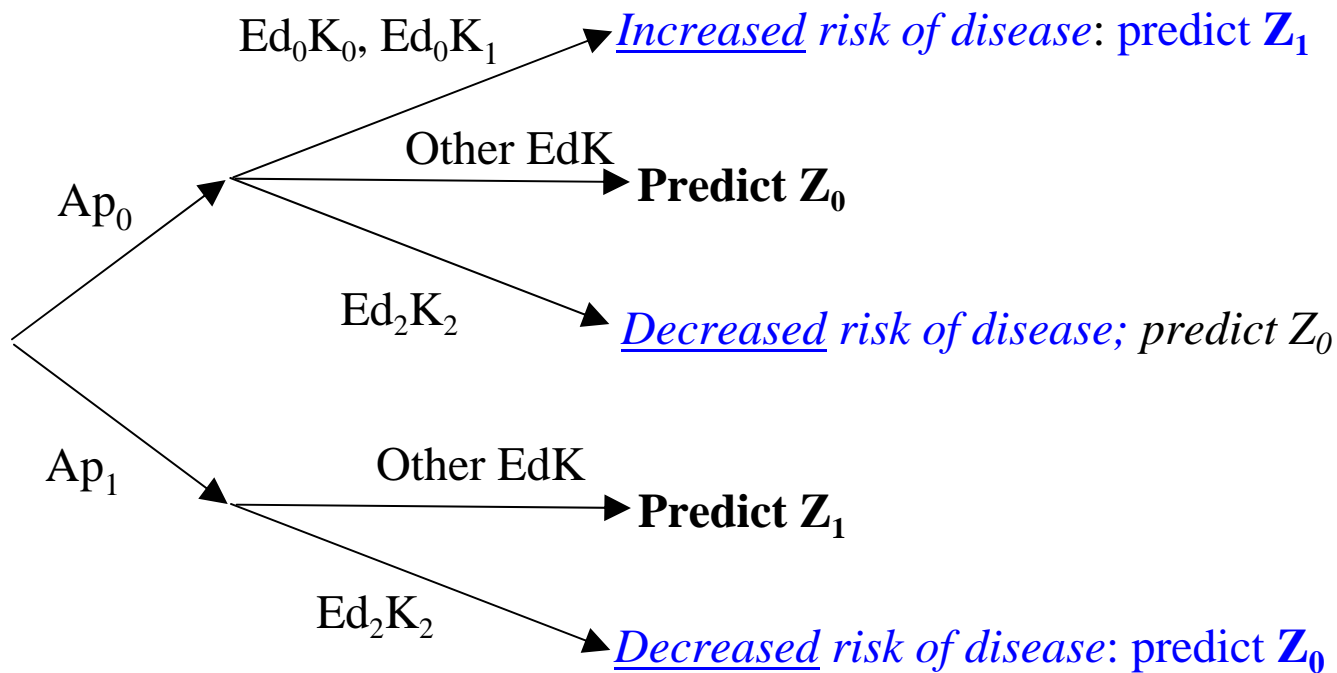
PROBABILITY DISTRIBUTION DEMENTIA EXAMPLE

DATA				MODEL IV:ApZ:EdZ:KZ									
IV				obs p(Z IV)		calc p(Z IV)		(p-value)		correct		(p-value)	
Ap	Ed	K	freq	Z ₀	Z ₁	Z ₀	Z ₁	rule	p _{rule}	#	%	P _{Ap}	
0	0	0	4	0.0	1.000	.122	.878	1	0.131	4	100.0	0.028	
0	0	1	8	.125	.875	.124	.876	1	0.033	7	87.5	0.002	
0	0	2	4	.250	.750	.294	.706	1	0.409	3	75.0	0.138	
0	1	0	31	.645	.355	.616	.384	0	0.198	20	64.5	0.707	
0	1	1	37	.622	.378	.619	.381	0	0.147	23	62.2	0.714	
0	1	2	23	.783	.217	.827	.173	0	0.002	18	78.3	0.072	
0	2	0	66	.636	.364	.640	.360	0	0.023	42	63.6	0.894	
0	2	1	61	.656	.344	.644	.357	0	0.025	40	65.6	0.942	
0	2	2	33	.848	.152	.842	.158	0	0.000	28	84.8	0.020	
0	--	--	267	.648	.352	.648	.352	0					
1	0	0	1	.000	1.000	.026	.974	1	0.343	1	100.0	0.571	
1	0	1	7	.143	.857	.026	.974	1	0.012	6	85.7	0.134	
1	0	2	2	.000	1.000	.074	.926	1	0.228	2	100.0	0.514	
1	1	0	13	.308	.692	.234	.766	1	0.055	9	69.2	0.709	
1	1	1	24	.167	.833	.237	.763	1	0.010	20	83.3	0.633	
1	1	2	11	.545	.455	.478	.522	1	0.884	5	45.5	0.146	
1	2	0	32	.219	.781	.254	.746	1	0.005	25	78.1	0.732	
1	2	1	39	.256	.744	.256	.744	1	0.002	29	74.4	0.735	
1	2	2	17	.529	.471	.504	.496	0	0.973	9	52.9	0.040	
1	--	--	146	.281	.719	.281	.719	1					
			413	.518	.482	.518	.482	0		291	70.5		

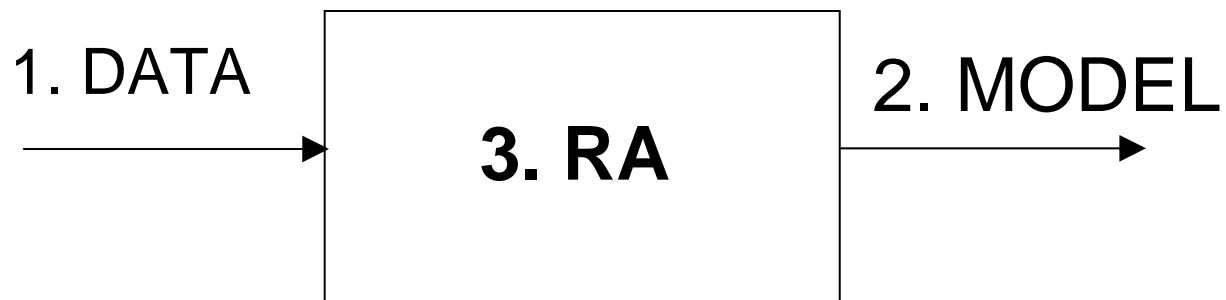
PROBABILITY DISTRIBUTION *DEMENTIA EXAMPLE*

Decision tree from conditional probability distribution

(Increase or decrease of risk given by **odds ratios**.)



1. input data to RA
2. model output from RA
3. basic RA algorithms (*IT, inside the black box*)
 - generate model
 - evaluate model
4. for more information

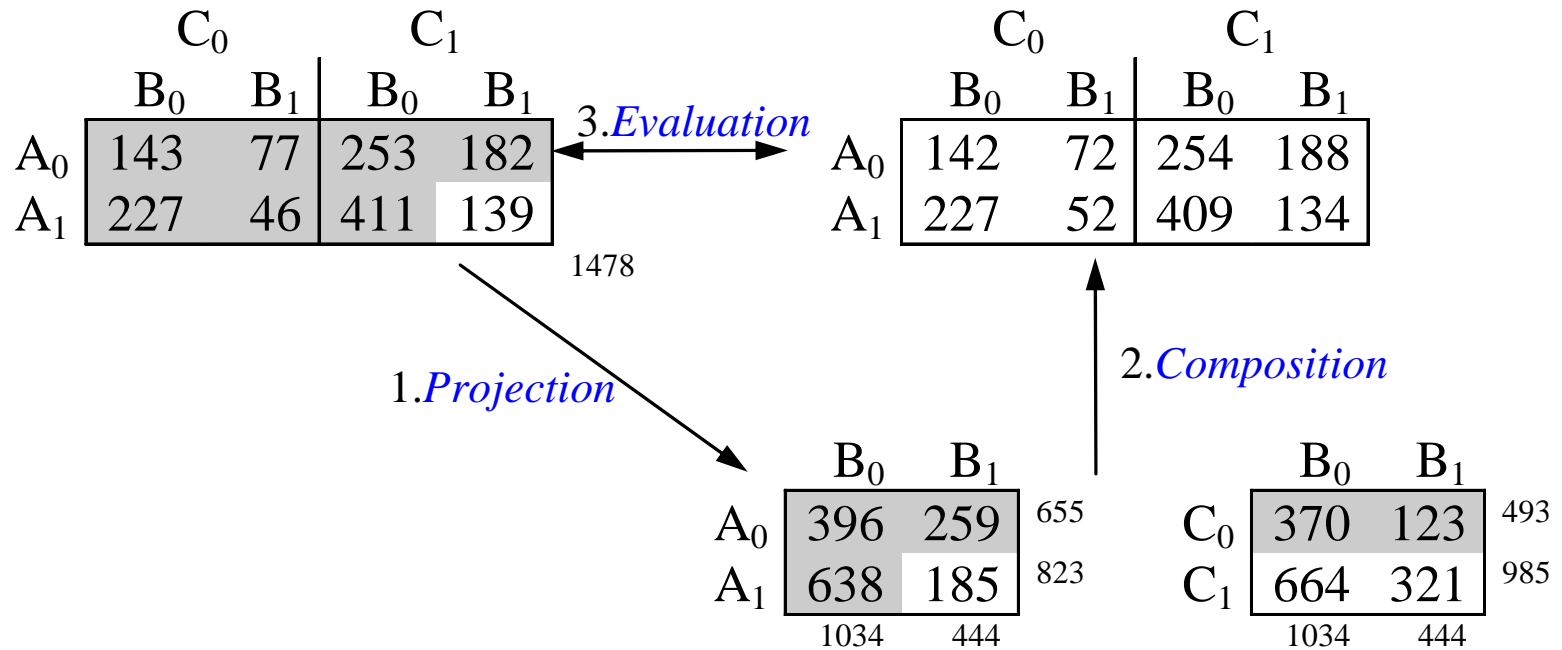


GENERATE MODEL

frequencies shown, not probabilities

data: observed ABC (df=7)

model: calculated ABC_{AB:BC}



model: AB:BC (df=5)

GENERATE MODEL

- *Projection* = sum frequencies or probabilities
- *Composition*

Maximize model entropy *subject to* model constraints

Model entropy: $H(p_{\text{model}}) = - \sum p_{\text{model}} \log_2 p_{\text{model}}$

E.g., for model AB:BC, *maximize* $H(p_{\text{AB:BC}})$ *subject to*

$$p_{\text{AB:BC}}(\text{AB}) = p_{\text{data}}(\text{AB})$$

$$p_{\text{AB:BC}}(\text{BC}) = p_{\text{data}}(\text{BC})$$

Composition is **critical computational step**; done

(a) Algebraically (very fast)

loopless models

(b) **Iteratively** (Iterative Proportional Fitting)

models with loops

EVALUATE MODEL (1/2)

- *Evaluation* (1 = data dependent; 2 = data independent)

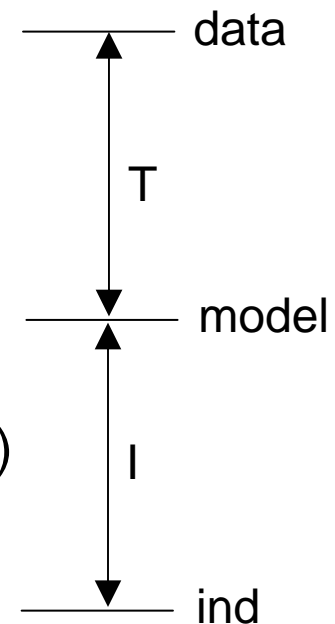
1. [ref=data]

error, T_{model} $= H_{\text{model}} - H_{\text{data}}$
 $= \sum p_{\text{data}} \log_2(p_{\text{data}}/p_{\text{model}})$

[ref=independence]

information, I_{model} $= H_{\text{ind}} - H_{\text{model}}$
 $= \sum p_{\text{data}} \log_2(p_{\text{model}}/p_{\text{ind}})$

uncertainty reduction $= H(\text{DV}) - H_{\text{model}}(\text{DV} | \text{IV})$



2. [ref=independence]

complexity $= \Delta df = df_{\text{model}} - df_{\text{ind}}$

EVALUATE MODEL (2/2)

Trade off information (or error) & complexity, define **best model** criterion, via:

Use likelihood ratio Chi-square, $LR = k N T$

- **p-values** from ΔLR , Δdf , Chi-square table

Or linear combinations of information & complexity

- **$\Delta AIC = \Delta LR + 2 \Delta df$**
- **$\Delta BIC = \Delta LR + \ln(N) \Delta df$**

1. input data to RA
2. model output from RA
3. basic RA algorithms
4. for **more information**
 - DMM (RA) web page
 - Software: OCCAM
 - MORE INFORMATION ON RA

DMM (RA) WEB PAGE

<http://pdx.edu/sysc/research-discrete-multivariate-modeling>

The screenshot shows a web browser window with the URL www.pdx.edu/sysc/research-discrete-multivariate-modeling. The page header includes the Portland State University logo and the text "Systems Science Graduate Program". A navigation menu contains "Courses", "Program", "Faculty", "Students", "Research", and "Resources". The "Research" menu is active, and the page title is "Research: Discrete Multivariate Modeling".

Artificial Life
Computational Intelligence
Discrete Multivariate Modeling
System Dynamics and Simulation
Neural Nets and Fuzzy Systems
Systems Theory and Philosophy

PSU » System Science » Research » Research: Discrete Multivariate Modeling

Research: Discrete Multivariate Modeling

The methods used are also known in the systems literature as "reconstructability analysis" (RA). RA overlaps significantly with the fields of logic design and machine learning and with log-linear statistical modeling. The papers "Wholes and Parts in General Systems Methodology" and "An Overview of Reconstructability Analysis" listed below offer a concise review of RA methodology.

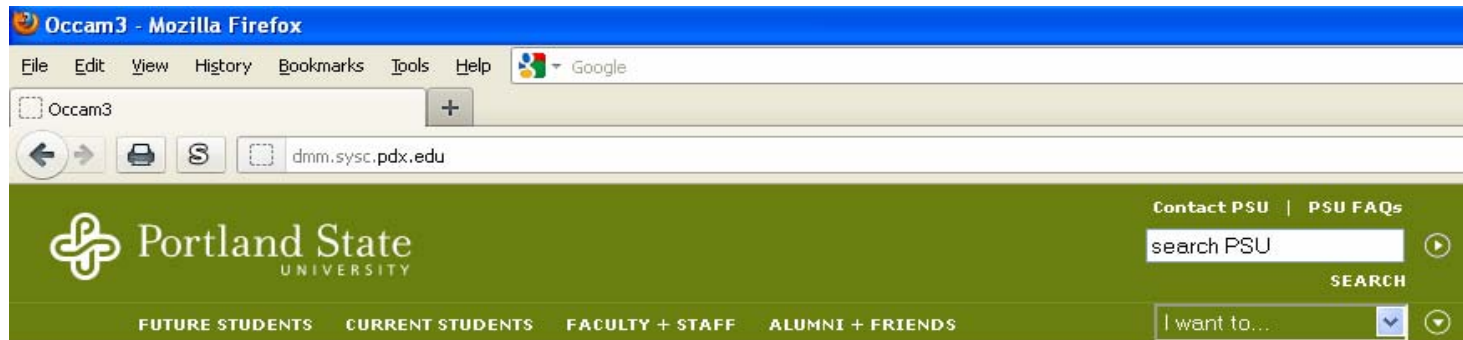
Projects

Theory/Methodology

- OCCAM: RA software for data analysis & data mining**
 - [Occam3 \(web accessible; try it out\)](#)
 - [User manual \(PDF\)](#)
- EDA: Extended Dependency Analysis**
 - Heuristic RA search for loopless models.
 - [Download executable, sample files, and documentation \(for Windows\)](#)
- RA utility programs**

Below is the lattice of structures for a 4-variable *directed* system with 1 dependent variable (output).
Boxes = relations; lines = variables;
bold lines = the dependent variable.

SOFTWARE: OCCAM (access on DMM page)



Occam

Occam is a Discrete Multivariate Modeling (DMM) tool based on the methodology of Reconstructability Analysis (RA). Its typical usage is for analysis of problems involving large numbers of discrete variables. *Models* are developed which consist of one or more *components*, which are then evaluated for their fit and statistical significance. Occam can search the lattice of all possible models, or can do detailed analysis on a specific model.

In *Variable-Based Modeling (VBM)*, model components are collections of variables. In *State-Based Modeling (SBM)*, components identify one or more specific states or substates.

Occam provides a web-based interface, which allows uploading a data file, performing analysis, and viewing or downloading results.

- [Run Occam](#)
- For basic operation instructions, please see the manual: [PDF](#)
- Sample data files. You can download these to local files on your computer, then upload them via the Occam Web interface.
[A Neutral System](#)
[A Directed System](#)
- Links:
[Dr. Zwick's DMM Research Page](#)
[Systems Science Graduate Program](#)
[Occam-users mailing list \(discussion\)](#)
[Occam-news mailing list \(announcements\)](#)
- Contacts:
[Occam feedback email address](#)
[Dr. Martin Zwick, Systems Science](#)
[Joe Fusion, Graduate Assistant, Systems Science](#)

OCCAM Initial Screen

The screenshot shows a web browser window with the URL `dmit.sysc.pdx.edu/weboccam.cgi`. The header features the Portland State University logo and name. Below this, the word "Occam" is displayed in a stylized font, followed by "version 3.4.0" and the timestamp "Tue Jun 19 14:41:08 2018". A light blue control bar contains several buttons: "Do Search", "Do SB-Search", "Do Fit", "Do SB-Fit", "Do Compare", "Show Log", "Manage Jobs", and "Cached Data Mode". The footer of the page shows the copyright notice "© 2000-2017".

BASIC OCCAM ACTIONS

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

(OCCAM actions: Show Log, Manage Jobs = managerial functions)

INFORMATION ON RA

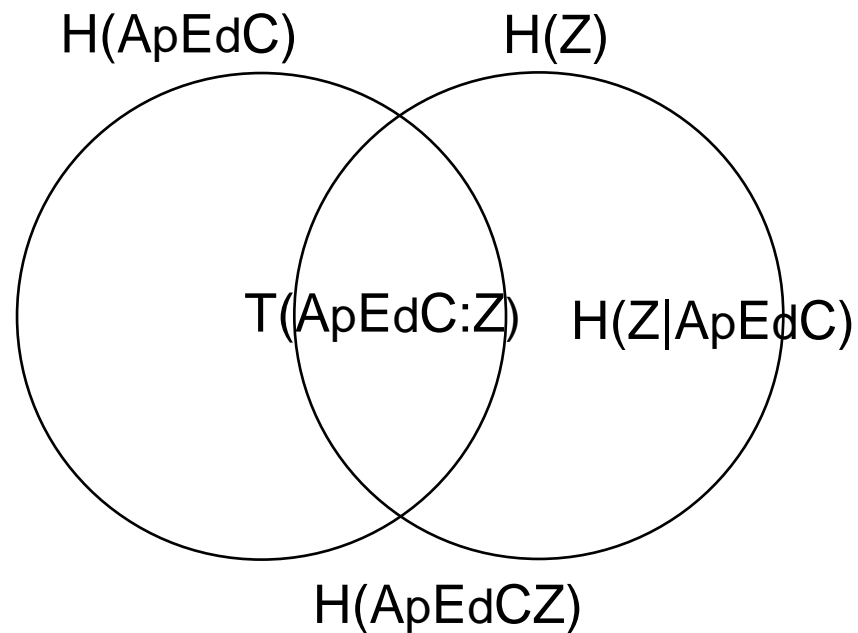
- Review articles on DMM page
 - “Wholes & Parts in General Systems Methodology” (accessible)
 - “An Overview of Reconstructability Analysis” (encompassing)
- Krippendorff, Klaus (1986). *Information Theory. Structural Models for Qualitative Data* (Quantitative Applications in the Social Sciences Monograph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

- OCCAM is available for use
(but consult with me before doing anything other than variable-based models without loops)
- Plan to make OCCAM [open-source](#); contact me if you would like to be involved
- zwick@pdx.edu
- *Thank you.*

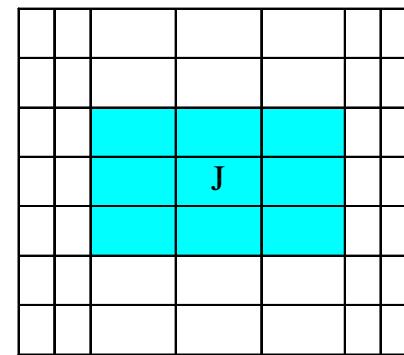
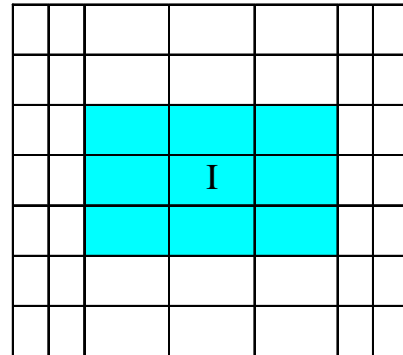
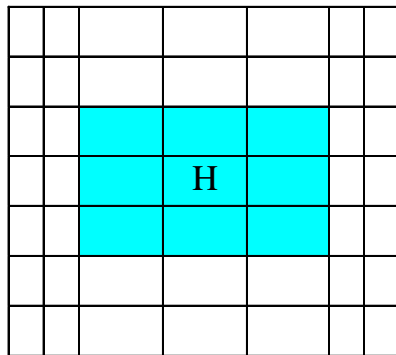
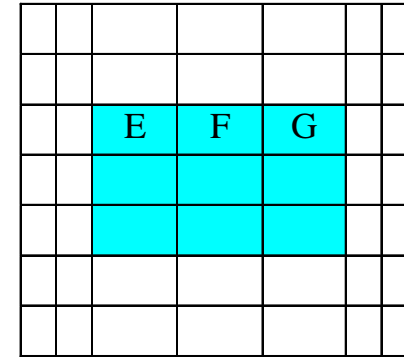
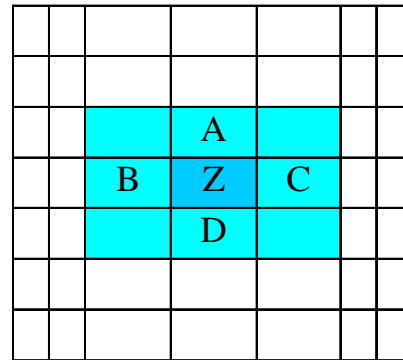
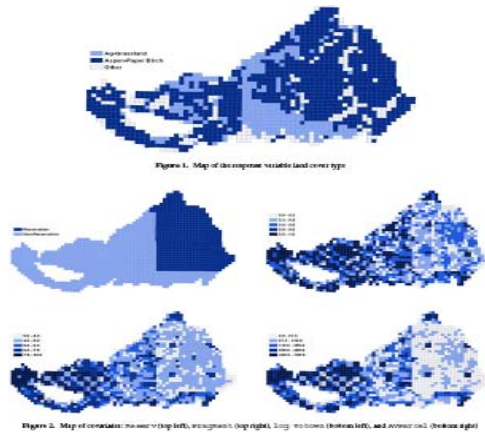
UNCERTAINTY REDUCTION: DEMENTIA EXAMPLE

<u>Criterion</u>	<u>model</u>	<u>$\Delta H(\%)$</u>	<u>Δdf</u>	<u>$\%c$</u>
BIC	IV:ApZ:EdZ:CZ	16	5	70

$$\Delta H = T_{IV:ApZ:EdZ:CZ}(ApEdC:Z) / H(Z) = 14\%$$



HYPOTHETICAL MODEL SPATIAL EXAMPLE



- In 5-generating-variables spatial example, model *could* be:
- IV: ABCD Z:EFG Z:HIJ Z