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
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. For more information, please contact 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

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 | APOE | Gender | Education | AgeLastExam | rs1801133 | rs3818361 | rs7561528 | rs744373 | rs6943822 | rs4298437 | rs7012010 | rs11136000 | rs10786998 | rs11193130 | rs610932 | rs3851179 | rs3764650 | rs3865444 | Dementia | \\n|----|------|--------|-----------|-------------|----------|----------|----------|----------|----------|----------|----------|-----------|------------|-----------|----------|--------|---------|----------|----------|----------|
| 101| 0    | 0      | 1         | 2           | 1        | 1        | 1        | 2        | 1        | 2        | 1        | 0         | 1          | 1         | 2        | 2      | 1       | 2        | 1        |
| 103| 0    | 0      | 1         | 2           | 1        | 1        | 2        | 2        | 1        | 1        | 2        | 2        | 1          | 0         | 1        | 0      | 1       | 0        | 1        |
| 111| 0    | 1      | 2         | 1           | 2        | 1        | 1        | 2        | 1        | 2        | 1        | 2        | 0          | 1         | 0        | 1      | 1       | 0        | 2        |
| 112| 0    | 0      | 2         | 2           | 2        | 2        | 1        | 1        | 2        | 1        | 1        | 2        | 2          | 0         | 0        | 0      | 2       | 0        | 2        |
| 118| 0    | 1      | 0         | 2           | 2        | 2        | 2        | 0        | 0        | 1        | 1        | 1          | .          | 1         | 1       | 0      | 2        | 0        | 0        |
| 120| 0    | 1      | 2         | 1           | 2        | 1        | 1        | 2        | 1        | 1        | 2        | 1          | 0          | 1         | 1       | 1      | 0        | 2        | 0        |
| 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        | 2        | 0        | 0          | 2          | 2         | 0        | 1      | 1        | 1        | 1        |
| 123| 0    | 0      | 2         | 2           | 2        | 2        | 2        | 0        | 1        | 1        | 0        | 0          | 2          | 0         | 2        | 1      | 0        | 1        | 1        |

**DEMENTIA EXAMPLE**

Z = 0 no disease; Z = 1 disease
**DATA CASES INDEXED BY TIME**

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>t-4</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>t-3</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>t-2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>t-1</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>t</td>
<td>9</td>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
</tbody>
</table>

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

Moore neighborhood

\[ E = DV \]
\[ A, B, C, D, F, G, H, I = IVs \]

IVs & DV have 14 possible states

<table>
<thead>
<tr>
<th>#A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
</tr>
<tr>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
</tr>
<tr>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
</tr>
<tr>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
</tr>
<tr>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
</tr>
<tr>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
</tr>
<tr>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
</tr>
<tr>
<td>71</td>
<td>71</td>
<td>71</td>
<td>95</td>
<td>71</td>
<td>95</td>
<td>71</td>
<td>71</td>
<td>71</td>
</tr>
<tr>
<td>95</td>
<td>71</td>
<td>95</td>
<td>95</td>
<td>71</td>
<td>95</td>
<td>71</td>
<td>71</td>
<td>71</td>
</tr>
<tr>
<td>95</td>
<td>95</td>
<td>95</td>
<td>95</td>
<td>71</td>
<td>95</td>
<td>71</td>
<td>71</td>
<td>71</td>
</tr>
<tr>
<td>71</td>
<td>95</td>
<td>95</td>
<td>90</td>
<td>95</td>
<td>95</td>
<td>95</td>
<td>95</td>
<td>95</td>
</tr>
<tr>
<td>95</td>
<td>95</td>
<td>90</td>
<td>90</td>
<td>71</td>
<td>95</td>
<td>95</td>
<td>95</td>
<td>95</td>
</tr>
<tr>
<td>95</td>
<td>90</td>
<td>90</td>
<td>90</td>
<td>95</td>
<td>90</td>
<td>95</td>
<td>95</td>
<td>90</td>
</tr>
</tbody>
</table>

...
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)

<table>
<thead>
<tr>
<th>Neutral</th>
<th>df #</th>
<th>Directed</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABC*</td>
<td>7</td>
<td>ABZ*</td>
</tr>
<tr>
<td>AB:AC:BC</td>
<td>6</td>
<td>AB:AZ:BZ</td>
</tr>
<tr>
<td>AB:AC</td>
<td>5</td>
<td>AB:BZ</td>
</tr>
<tr>
<td>AB:C</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>AC:B</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>BC:A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A:B:C*</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Reference model is data or independence

# df (degrees of freedom) values are for binary variables

Reference model is data or independence

# df (degrees of freedom) values are for binary variables

Reference model is data or independence

# df (degrees of freedom) values are for binary variables
STRUCTURES 4 variables (GT)
### STRUCTURES (GT)

**Combinatorial explosion**

<table>
<thead>
<tr>
<th># variables</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td># general structures</td>
<td>5</td>
<td>20</td>
<td>180</td>
<td>16,143</td>
</tr>
<tr>
<td># specific structures</td>
<td>9</td>
<td>114</td>
<td>6,894</td>
<td>7,785,062</td>
</tr>
<tr>
<td>(where 1 variable is DV)</td>
<td>5</td>
<td>19</td>
<td>167</td>
<td>7,580</td>
</tr>
<tr>
<td>(1 DV, no loops)</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
</tr>
</tbody>
</table>

**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    many variables (fast) \([coarse]\)
    \(IV:ACZ\)    simple prediction, feature selection
  – with loops  up to 100s of variables (slow) \([fine]\)
    \(IV:ABZ:BCZ\)   better prediction

• State-based    < 10 variables \((v.\ slow); \ [ultra-fine]\)
    \(IV:Z: A_1B_1Z : B_2C_3Z_1\)   best prediction; detailed models

“IV” = ABC (all IVs); Z = DV
All directed system models include an IV component
TYPES OF STRUCTURES (GT)

<table>
<thead>
<tr>
<th></th>
<th>COARSE</th>
<th>FINE</th>
<th>ULTRA-FINE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complexity (degrees of freedom)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>No loops</td>
<td>With loops</td>
<td>State-based</td>
</tr>
</tbody>
</table>

- No loops
- With loops
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., \( p_{A_i B_j C_k} (Z | Z_l) \)

- 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 $\Delta H$, reduction of uncertainty of DV
   
   – Generic measure: high %correct, accuracy of prediction

2. Low complexity: df, degrees of freedom

3. Information ↔ 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: $IV=A; DV=Z; T(A:Z)=\text{mutual information (association)}$

- *Uncertainty reduction* is like variance explained
  
  Model $AZ = \text{predict } Z$, i.e., reduce $H(Z)$, by knowing $A$

- Uncertainty *reduced* = $T(A:Z)$; uncertainty *remaining* = $H(Z|A)$
  
  $\Delta H = \frac{T(A:Z)}{H(Z)}$ *fractional uncertainty reduction* (will express in %)

![Venn Diagram](image-url)
UNCERTAINTY REDUCTION: SIMPLE EXAMPLE

- $p(Z_1)/p(Z_0) = 1:1$, 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

Criterion model $\Delta H(\%)$ $\Delta df$ $\% c$ $\Delta BIC$

Variable-based (with loops)

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

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 = $A_p E_d C K L...$ (all the independent variables); $\% c( IV:Z ) = 52$
## PROBABILITY DISTRIBUTION DEMENTIA EXAMPLE

| IV  | Ap | Ed | K | freq | obs p(Z | IV) | calc p(Z | IV) | (p-value) | correct | (p-value) |
|-----|----|----|---|------|---------|------------|------------|----------|---------|----------|
|     |    |    |   |      | Z₀  | Z₁       | rule      | p<sub>rule</sub> | #   |          |
|     |    |    |   |      | Z₀  | Z₁       | rule      | p<sub>rule</sub> | #   |          |
| 0   | 0  | 0  | 0 | 4    | 0.0 | 1.000    | .122      | .878     | 1     | .131     | 4  | 100.0   | 0.028 |
| 0   | 0  | 0  | 1 | 8    | .125| .875     | .124      | .876     | 1     | .033     | 7  | 87.5    | 0.002 |
| 0   | 0  | 2  | 4 | 1    | .250| .750     | .294      | .706     | 1     | .409     | 3  | 75.0    | 0.138 |
| 0   | 1  | 0  | 31| 1    | .645| .355     | .616      | .384     | 0     | .198     | 20 | 64.5    | 0.707 |
| 0   | 1  | 1  | 37| 1    | .622| .378     | .619      | .381     | 0     | .147     | 23 | 62.2    | 0.714 |
| 0   | 1  | 2  | 23| 1    | .783| .217     | .827      | .173     | 0     | .002     | 18 | 78.3    | 0.072 |
| 0   | 0  | 0  | 66| 1    | .636| .364     | .640      | .360     | 0     | .023     | 42 | 63.6    | 0.894 |
| 0   | 2  | 1  | 61| 1    | .656| .344     | .644      | .357     | 0     | .025     | 40 | 65.6    | 0.942 |
| 0   | 2  | 2  | 33| 1    | .848| .152     | .842      | .158     | 0     | .000     | 28 | 84.8    | 0.020 |
| 0   | -- | -- | 267| 1    | .648| .352     | .648      | .352     | 0     |          |    |         |        |
| 1   | 0  | 0  | 1 | 1    | .000| 1.000    | .026      | .974     | 1     | .343     | 1  | 100.0   | 0.571 |
| 1   | 0  | 1  | 7 | 1    | .143| .857     | .026      | .974     | 1     | .012     | 6  | 85.7    | 0.134 |
| 1   | 0  | 2  | 2 | 1    | .000| 1.000    | .074      | .926     | 1     | .228     | 2  | 100.0   | 0.514 |
| 1   | 1  | 0  | 13| 1    | .308| .692     | .234      | .766     | 1     | .055     | 9  | 69.2    | 0.709 |
| 1   | 1  | 1  | 24| 1    | .167| .833     | .237      | .763     | 1     | .010     | 20 | 83.3    | 0.633 |
| 1   | 1  | 2  | 11| 1    | .545| .455     | .478      | .522     | 1     | .884     | 5  | 45.5    | 0.146 |
| 1   | 2  | 0  | 32| 1    | .219| .781     | .254      | .746     | 1     | .005     | 25 | 78.1    | 0.732 |
| 1   | 2  | 1  | 39| 1    | .256| .744     | .256      | .744     | 1     | .002     | 29 | 74.4    | 0.735 |
| 1   | 2  | 2  | 17| 1    | .529| .471     | .504      | .496     | 0     | .973     | 9  | 52.9    | 0.040 |
| 1   | -- | -- | 146| 1    | .281| .719     | .281      | .719     | 1     |          |    |         |        |
|     | 413| 1    |   | 0.518| 0.482 | 0.518 | 0.482 | 0.291 | 0.705 |         |    |         |        |
**PROBABILITY DISTRIBUTION  DEMENTIA EXAMPLE**

Decision tree from conditional probability distribution
(Increase or decrease of risk given by odds ratios.)

![Decision Tree Diagram]

- **Ap₀**: Predict \( Z₀ \)
- **Ed₀K₀, Ed₀K₁**: **Increased risk of disease:** predict \( Z₁ \)
- **Other EdK**: Predict \( Z₀ \)
- **Ed₂K₂**: **Decreased risk of disease:** predict \( Z₀ \)
- **Ap₁**: Predict \( Z₁ \)
- **Other EdK**: Predict \( Z₁ \)
- **Ed₂K₂**: **Decreased risk of disease:** predict \( Z₀ \)
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
**data:** observed ABC (df=7)  

<table>
<thead>
<tr>
<th>A_0</th>
<th>C_0</th>
<th>B_0</th>
<th>B_1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>143</td>
<td>77</td>
<td>253</td>
</tr>
<tr>
<td>227</td>
<td>46</td>
<td>411</td>
<td>139</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A_1</th>
<th>C_0</th>
<th>B_0</th>
<th>B_1</th>
</tr>
</thead>
<tbody>
<tr>
<td>396</td>
<td>259</td>
<td>655</td>
<td></td>
</tr>
<tr>
<td>638</td>
<td>185</td>
<td>823</td>
<td></td>
</tr>
</tbody>
</table>

**model:** calculated ABC_{AB:BC}

<table>
<thead>
<tr>
<th>A_0</th>
<th>C_0</th>
<th>B_0</th>
<th>B_1</th>
</tr>
</thead>
<tbody>
<tr>
<td>142</td>
<td>72</td>
<td>254</td>
<td>188</td>
</tr>
<tr>
<td>227</td>
<td>52</td>
<td>409</td>
<td>134</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A_1</th>
<th>C_0</th>
<th>B_0</th>
<th>B_1</th>
</tr>
</thead>
<tbody>
<tr>
<td>370</td>
<td>123</td>
<td>493</td>
<td></td>
</tr>
<tr>
<td>664</td>
<td>321</td>
<td>985</td>
<td></td>
</tr>
</tbody>
</table>

**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** \( H(p_{AB:BC}) \) **subject to**

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

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

2. [ref=ind]
   
   information, $I_{model} = H_{ind} - H_{model}$
   
   $= \sum p_{data} \log_2 (p_{model} / p_{ind})$

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

2. [ref=ind]
   
   complexity $= \Delta df = df_{model} - df_{ind}$
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 \( \Delta LR, \Delta df, \) 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
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 "Whales 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
  - Occam8 (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 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 substrates.

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
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)


• *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**

<table>
<thead>
<tr>
<th>Criterion</th>
<th>model</th>
<th>$\Delta H(%)$</th>
<th>$\Delta df$</th>
<th>%c</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIC</td>
<td>IV:ApZ:EdZ:CZ</td>
<td>16</td>
<td>5</td>
<td>70</td>
</tr>
</tbody>
</table>

• In 5-generating-variables spatial example, model *could* be:

• **IV: ABCD Z:EFG Z:HIJ Z**