Bayesian Networks Workshop

Brescia, 18 5 2015

On Bayesian networks and information quality (InfoQ)

Ron S. Kenett

KPA Ltd., Raanana, Israel

Univ. of Torino, Torino, Italy

Center for Risk Engineering, NYU Poly, New York, USA

ron@kpa-group.com

Joint work with Silvia Salini, Federica Cugnata and Galit Shmueli

KPA Insights through analytics

Applied statistics is about meeting the challenge of **solving real world problems** with **mathematical tools** and **statistical thinking**

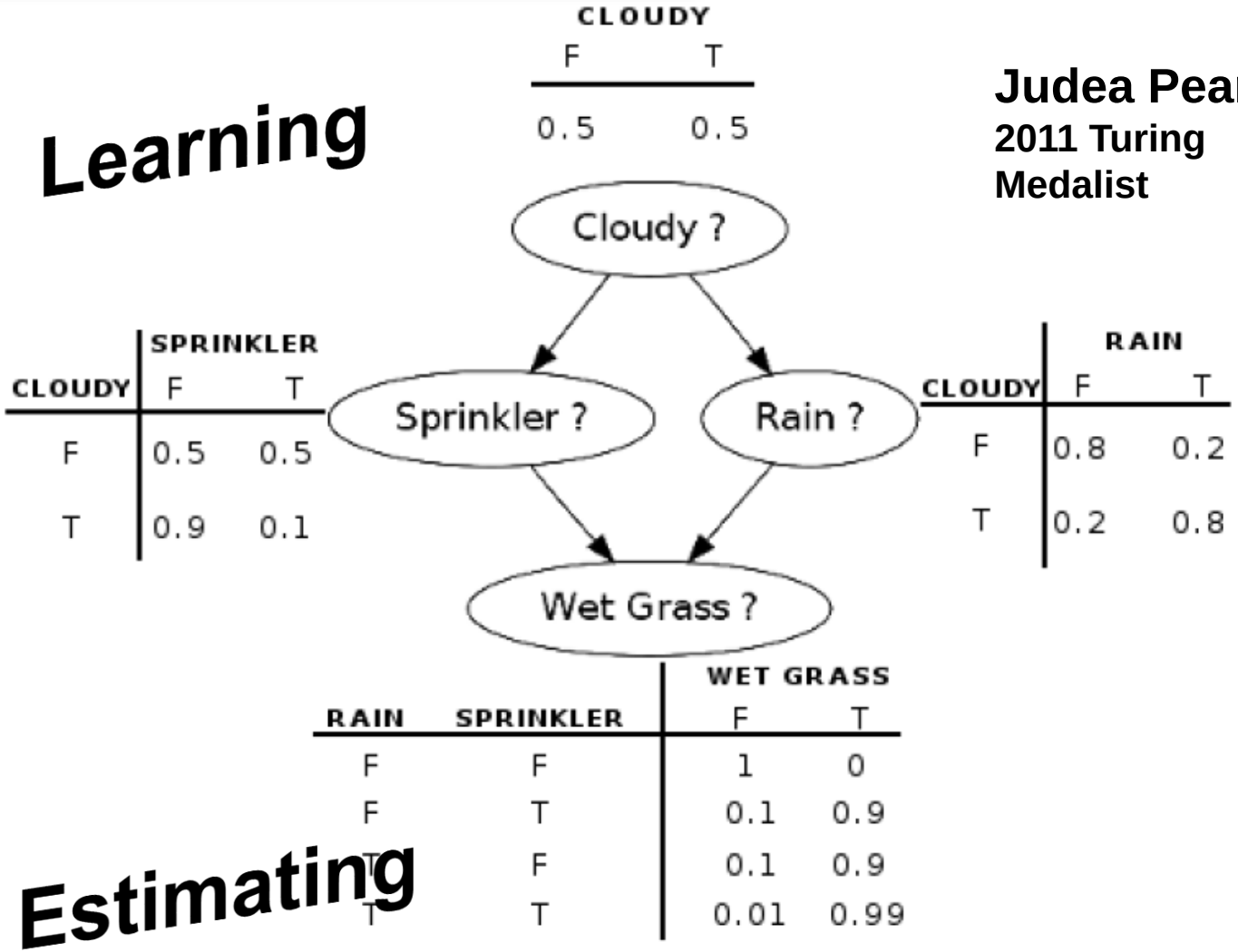
"Much fine work in statistics involves minimal mathematics; some bad work in statistics gets by because of its apparent mathematical content."

David Cox (1981),
Theory and general principle in statistics, *JRSS(A)*, 144, pp. 289-297.

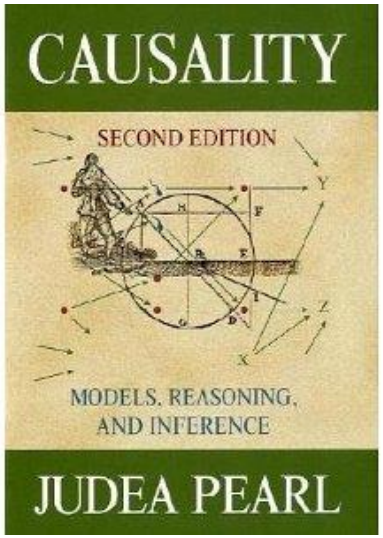
Why Bayesian Networks?

Learning

Estimating



Judea Pearl
2011 Turing Medalist

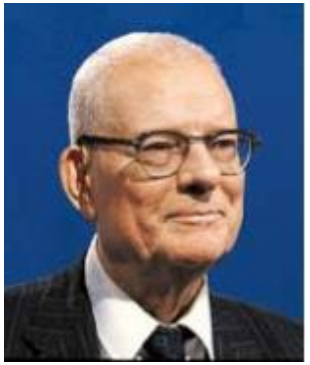


Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*, 2nd ed., Cambridge University Press, UK.

W. Edwards Deming (1900-1993)

"Tests of variables that affect a process are useful only if they **predict what will happen if this or that variable is increased or decreased**.

Statistical theory, as taught in the books, is valid and leads to operationally verifiable tests and criteria for an **enumerative study**. Not so with an **analytic problem**, as the conditions of the experiment will not be duplicated in the next trial.



Unfortunately, most problems in industry are analytic."*

*From preface to *The Economic Control of Quality of Manufactured product* by W. Shewhart, 1931

"Statistics is important because it is conceived as contributing to a **causal understanding** ...
Statistics can indicate causality *even in the absence of a mechanistic understanding*.
But the traditional self-conception of statistics is that it can rarely say anything about causality.
This is a ***paradox***."

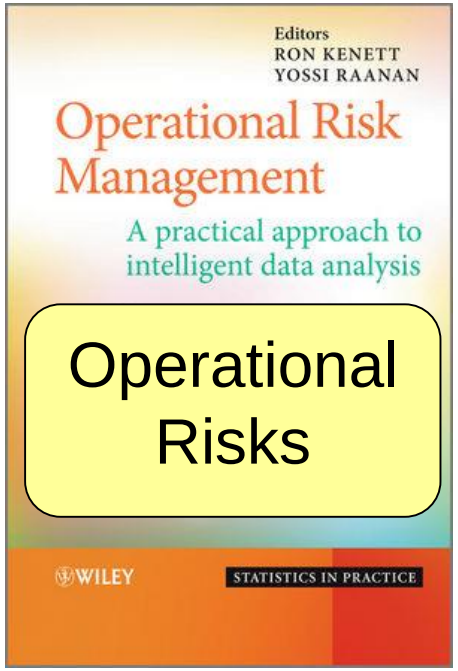
*Statistikk 50 År! Some remarks on causality**

Odd O. Aalen

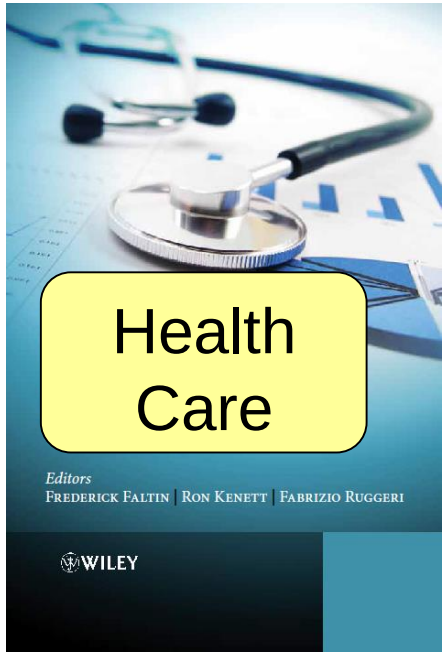
*From a presentation celebrating 50 years to the establishment of a Masters Degree in Statistics in Norway, May 22, 2006

Customer Surveys

Quality Report

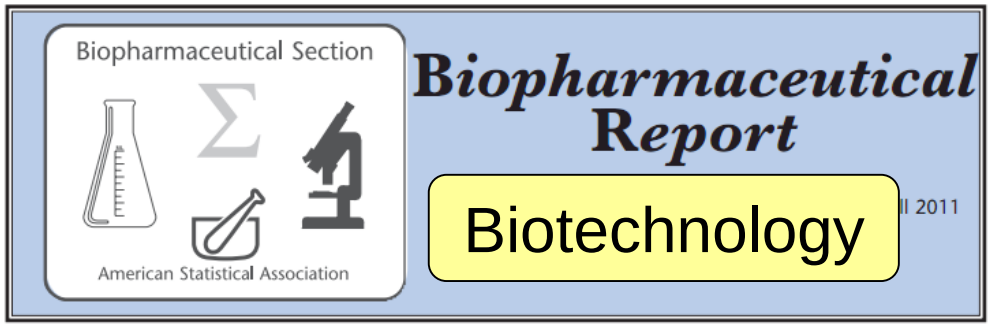


Operational Risks

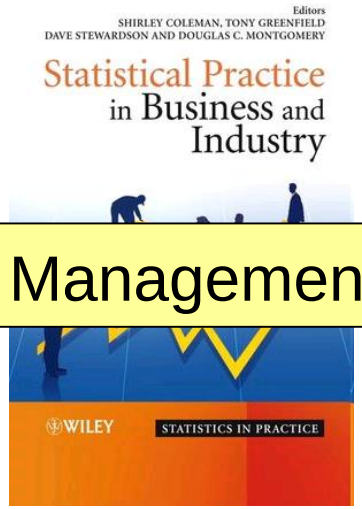


Health Care

Web Usability

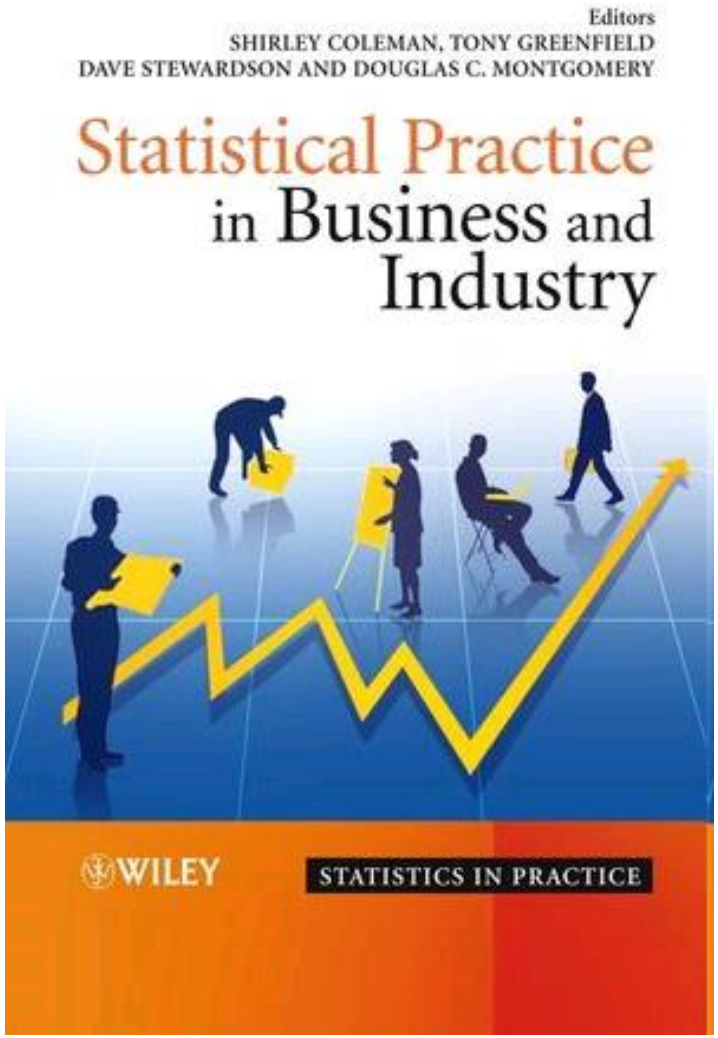


Biotechnology



Management

The 7 case studies

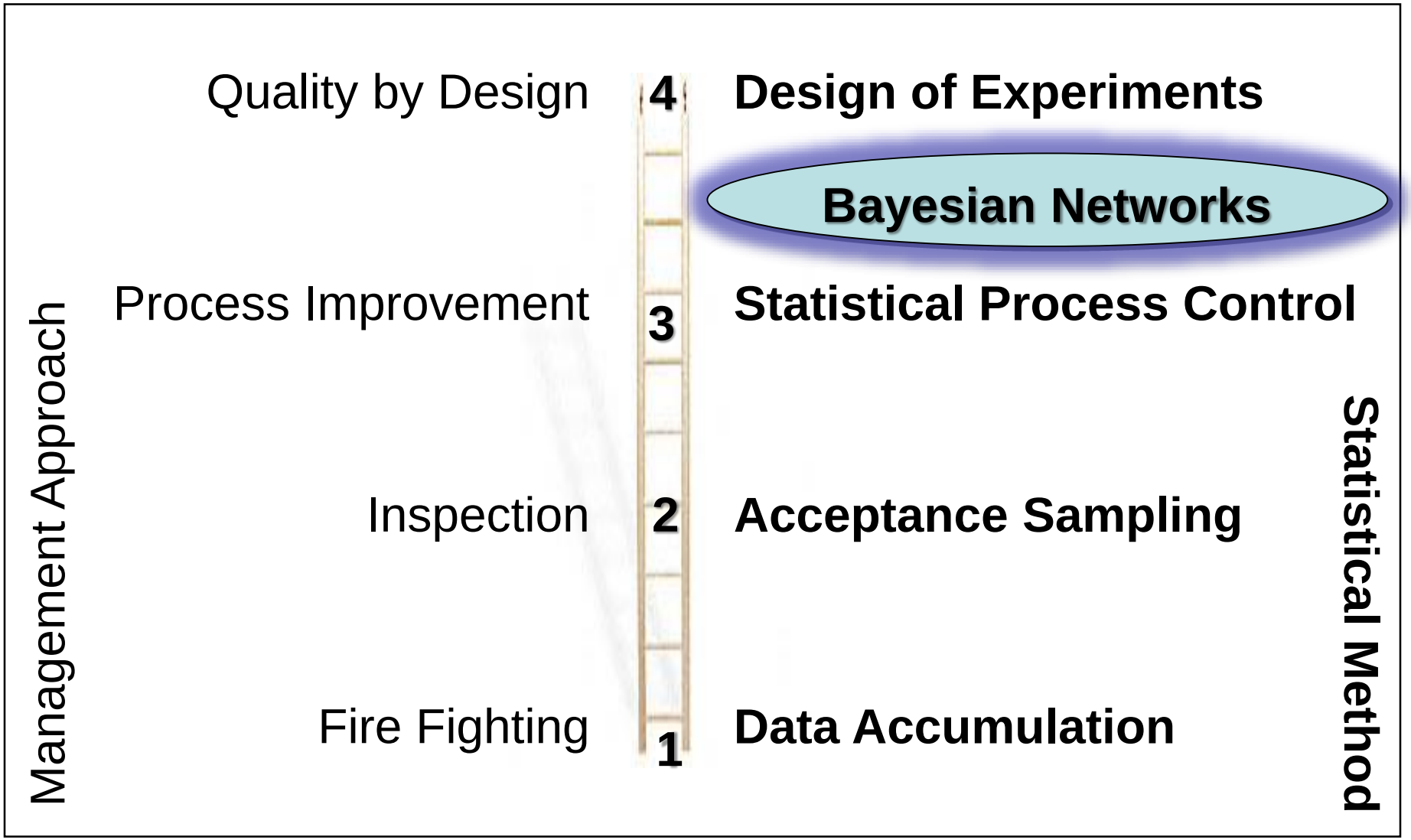


Management

Conjecture Validation

Case study 1

The Quality Ladder



Kenett, R., De Frenne, A., Tort-Martorell, X and McCollin, C. (2008). The Statistical Efficiency Conjecture, in *Applying Statistical Methods in Business and Industry – the state of the art*, Greenfield, T., Coleman, S. and Montgomery, R. (editors), Wiley.

PSE=Practical Statistical Efficiency

PSE = E{R} x T {I} x P {I} x V {PS} x P {S} x V {P} x V {M} x V {D}

PSE component: 1-5

"1"= not very good, "5" = excellent.

Impact

V{D} = **V**alue of the **D**ata actually collected
V{M} = **V**alue of the statistical **M**ethod employed
V{P} = **V**alue of the **P**roblem to be solved
P{S} = **P**robability that the problem actually gets **S**olved
V{PS} = **V**alue of the **P**roblem being **S**olved
P{I} = **P**robability the solution is actually **I**mplemented
T{I} = **T**ime the solution stays **I**mplemented
E{R} = **E**xpected number of **R**eplications

L=Maturity Level on Quality Ladder: 1- 4

1 = Fire fighting (FF), 2 = Inspection (I),

3 = Process Improvement (PI), 4 = Quality by Design (QbD)

Maturity

The Statistical Efficiency Conjecture

Let PSE = PSE of a specific project.

PSE is a random variable with specific realisations for individual projects, 1<= PSE <= 5^8.

E{ PSE } = The expected value of PSE in a given organisation over all projects.

The Statistical Efficiency Conjecture is linking Expected Practical Statistical Efficiency with the maturity of an organisation on the Quality Ladder.

In more formal terms it is stated as:

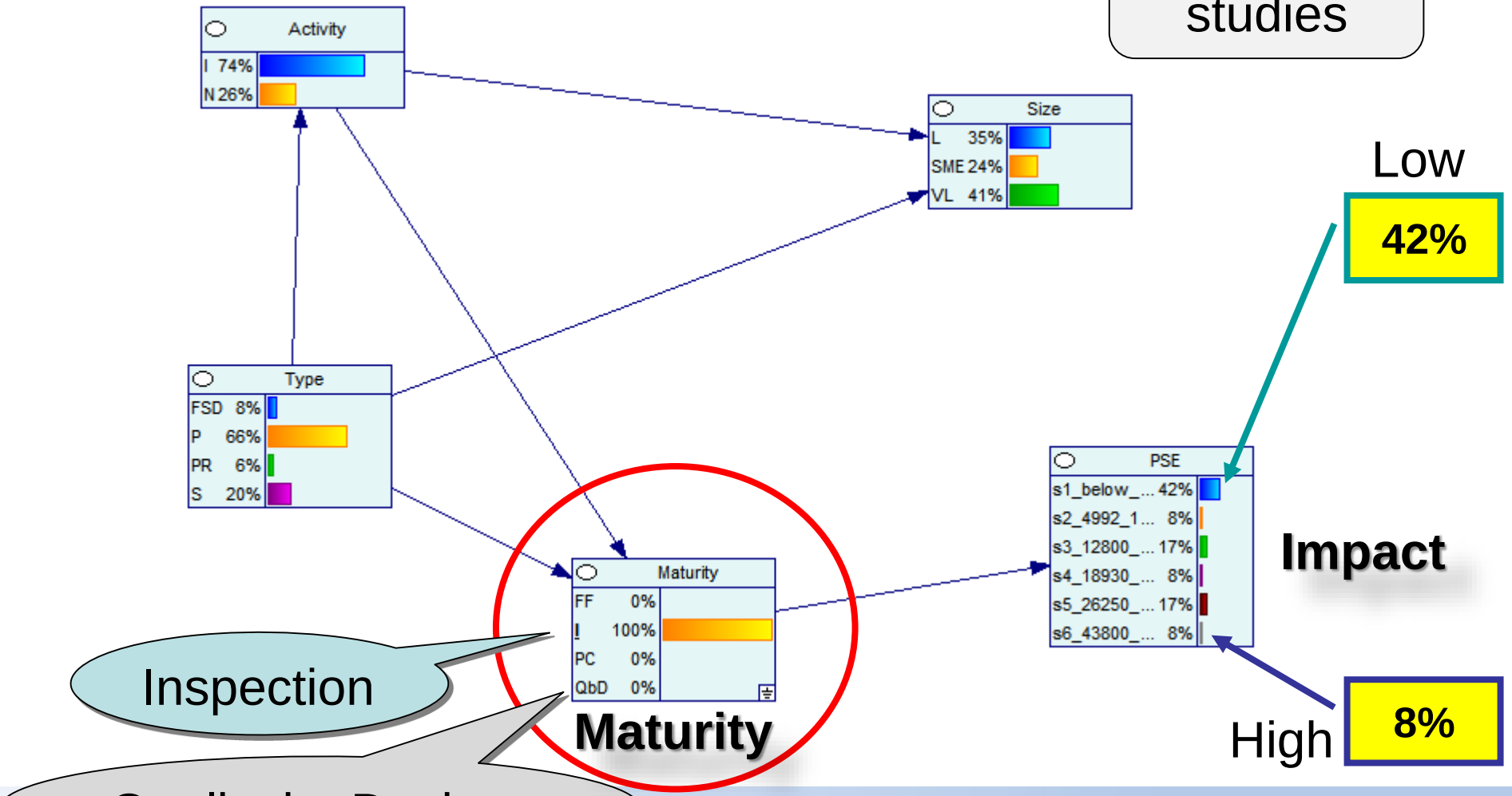
Conjecture:

Conditioned on the right variable,

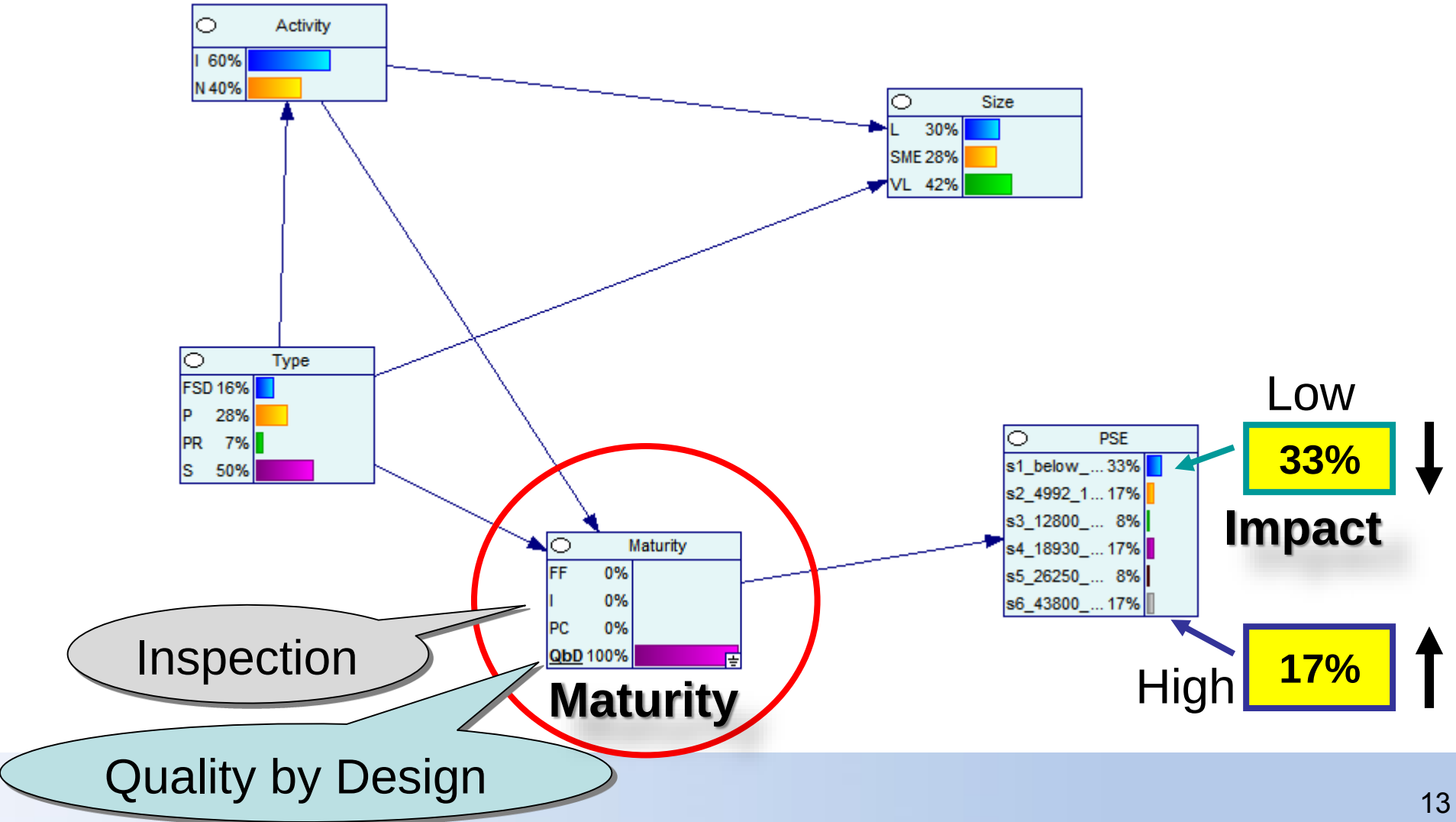
E{ PSE } is an increasing function of L

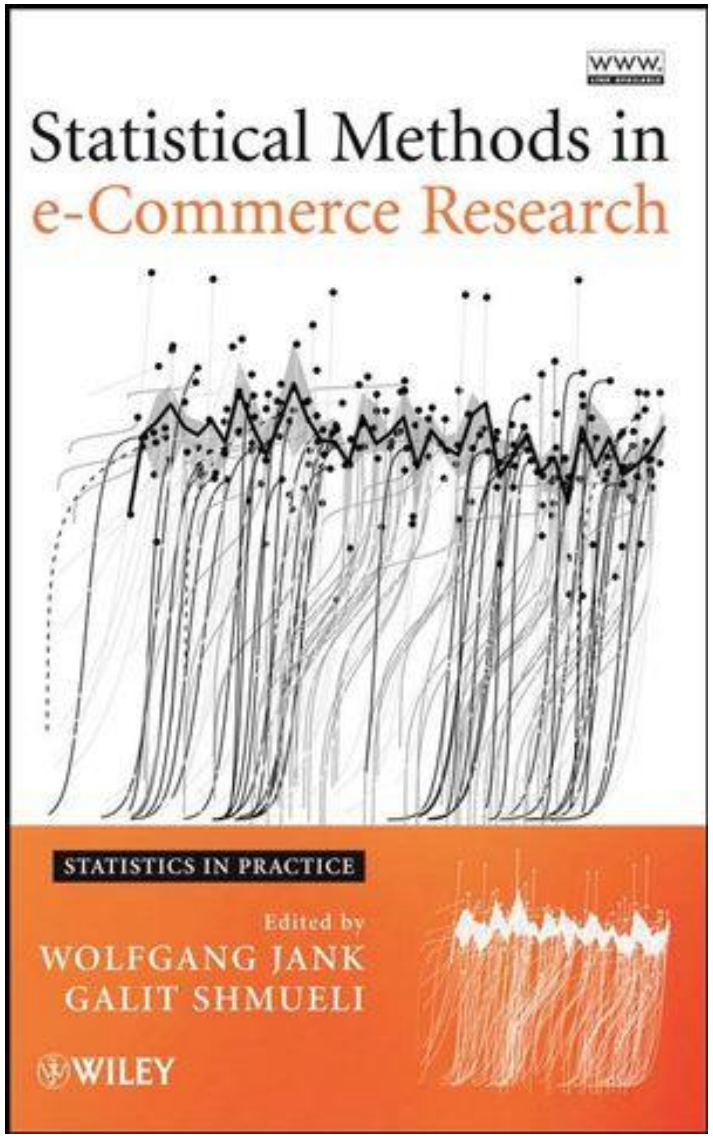
The Statistical Efficiency Conjecture

21 case studies



The Statistical Efficiency Conjecture





Web Usability

Usability Monitoring

Case study 2

The Seven Layers of a *Decision Support for User Interface Design* (DSUID)

The lowest layer – user activity

The second layer – page hit attributes (data)

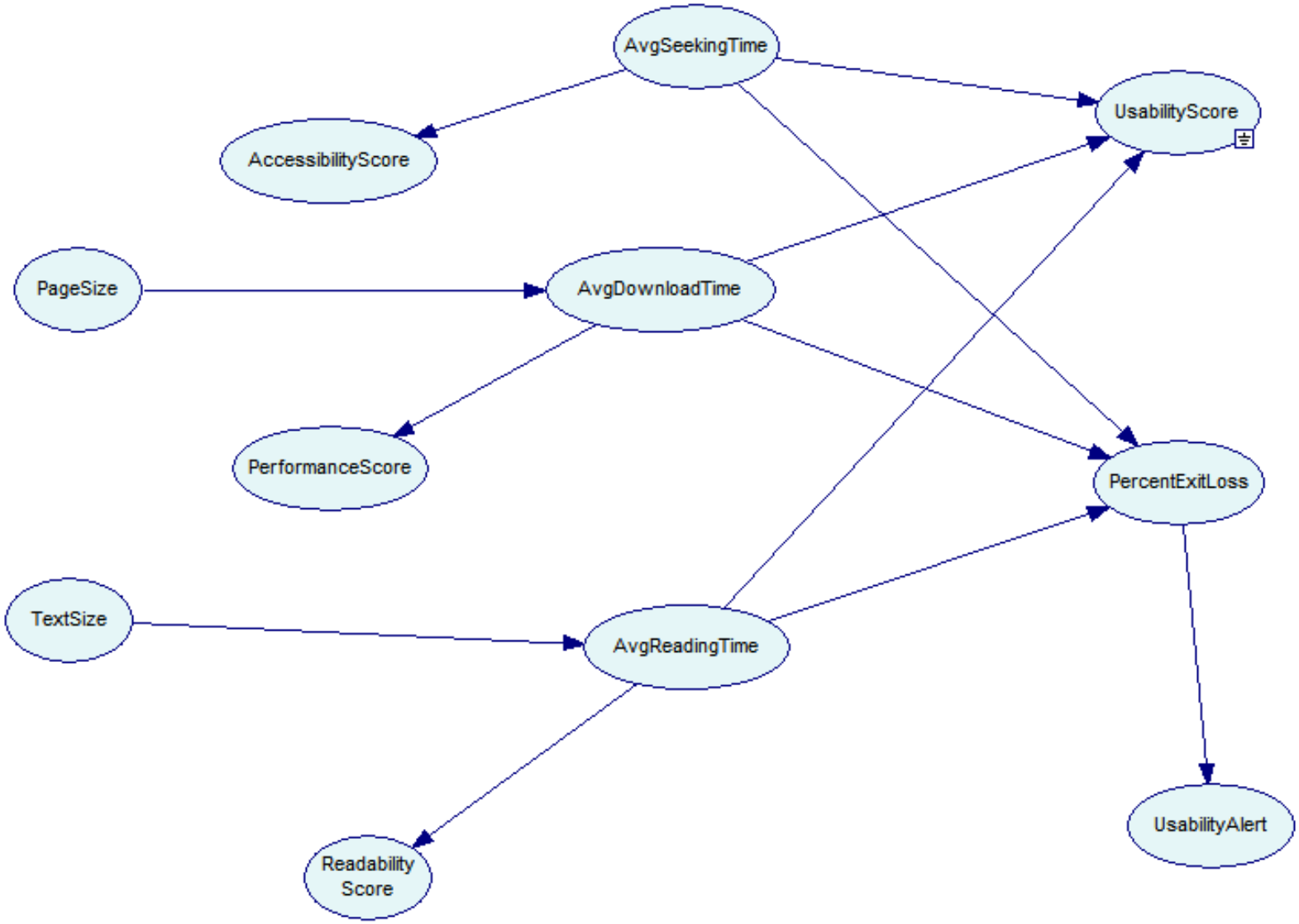
The third layer – transition analysis (dynamic)

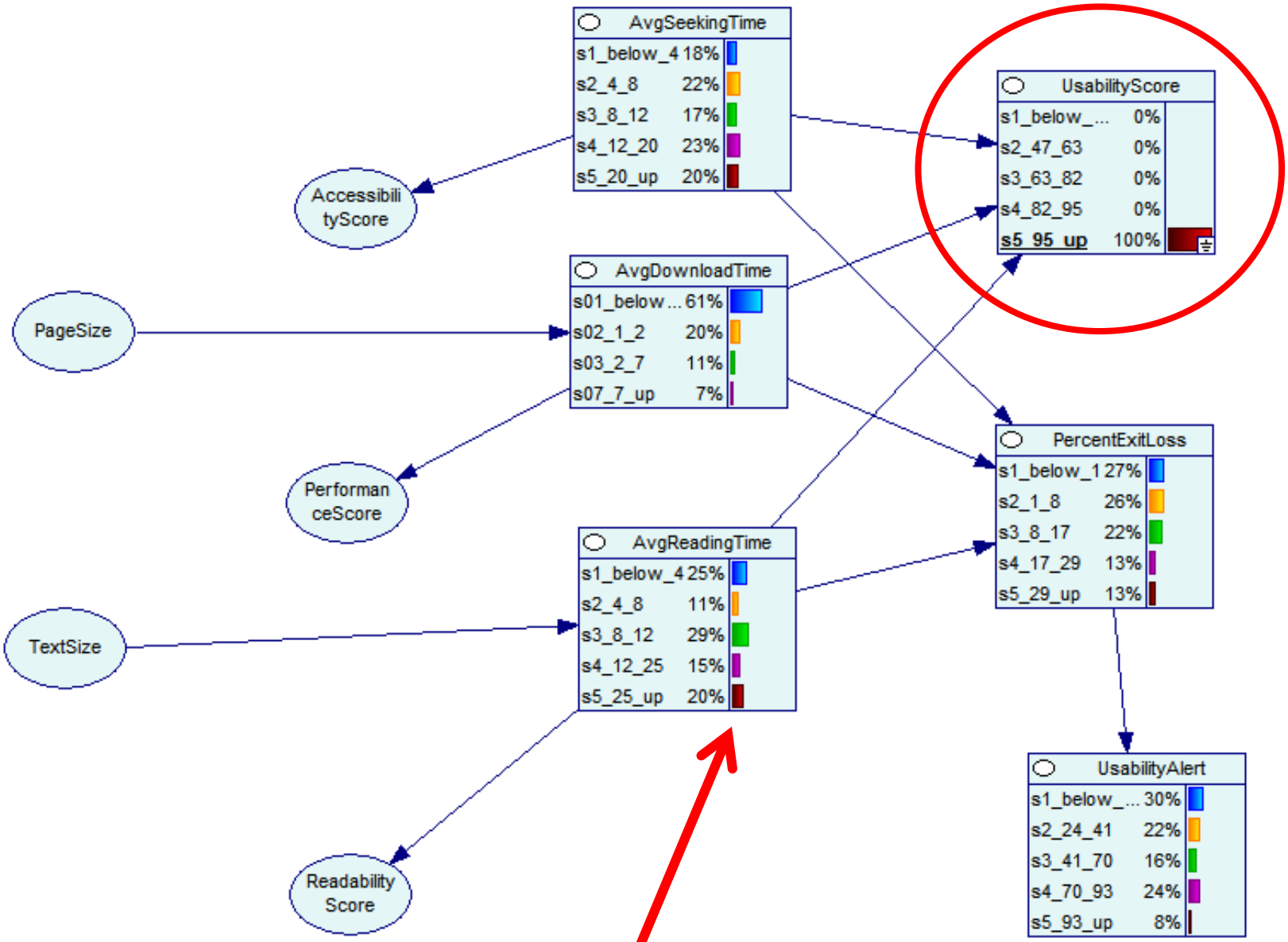
The fourth layer – UPI identification (quant. ➡ subjective)

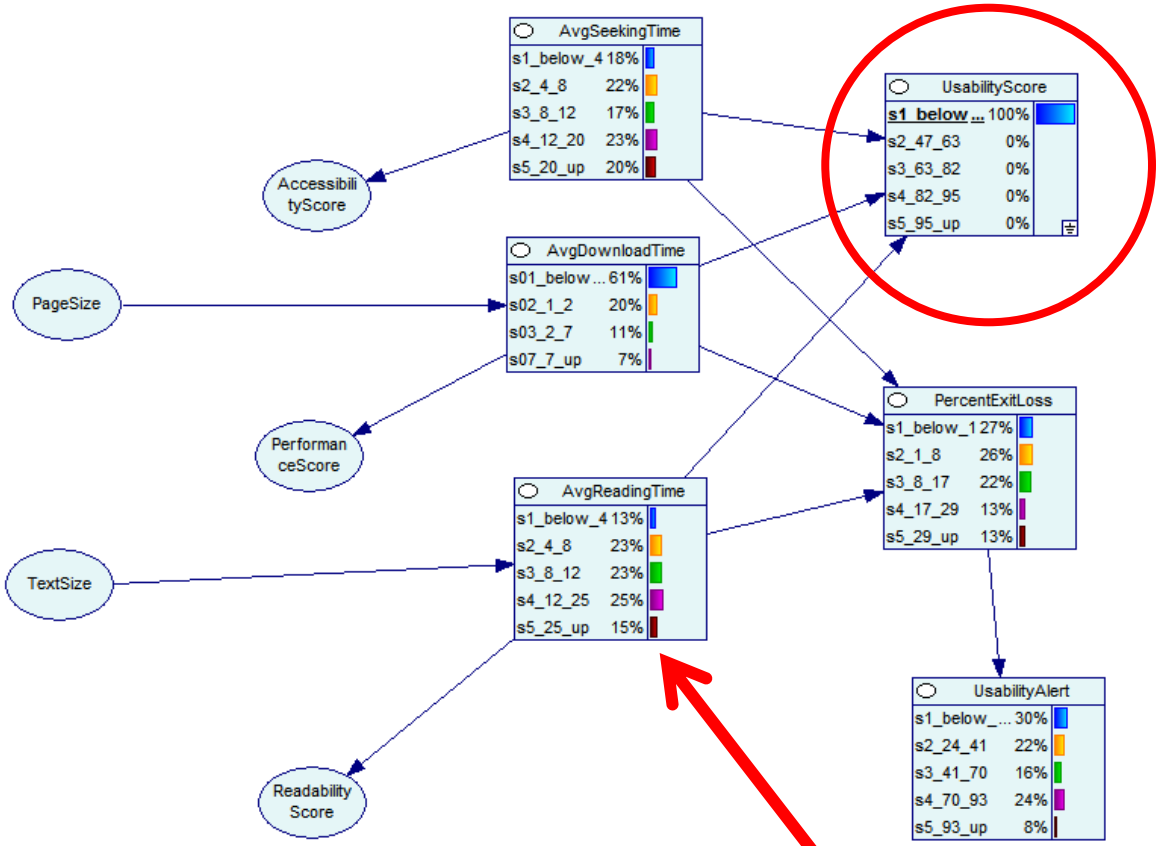
The fifth layer – usage statistics (descriptive)

The sixth layer – statistical decision (t-test, BN,…)

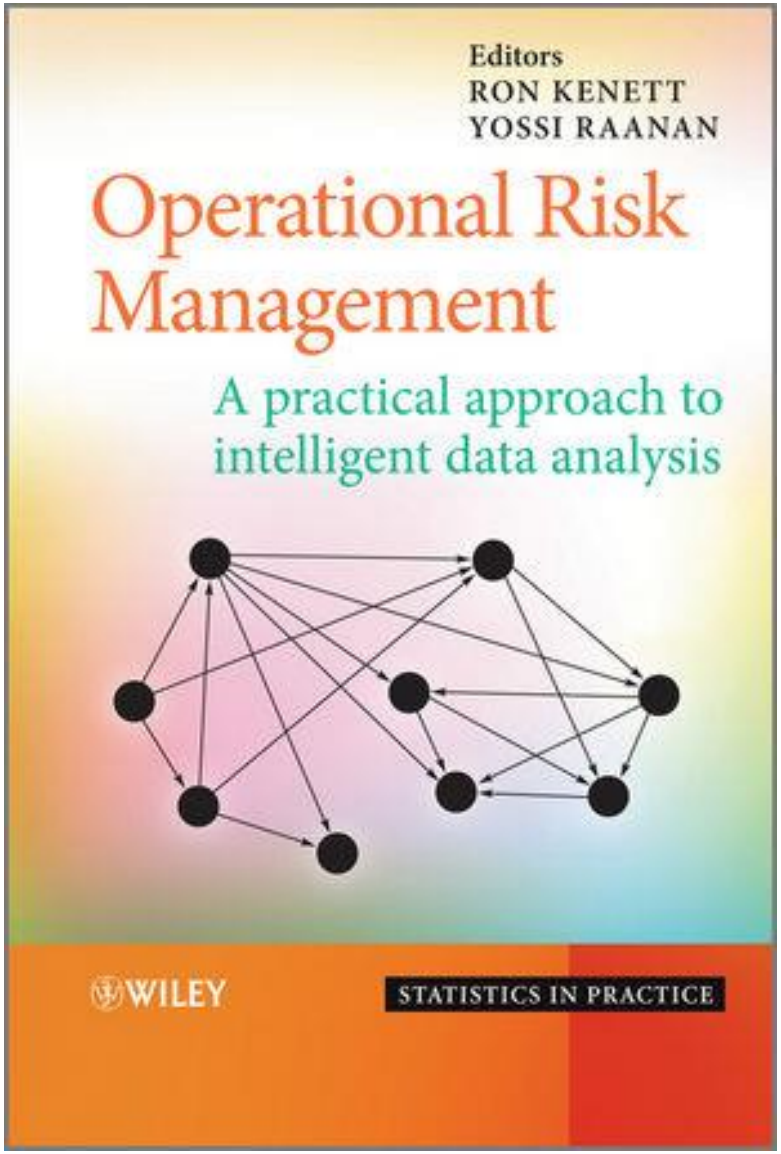
The top layer – interpretation


AvgSeekingTime
AccessibilityScore
UsabilityScore
PageSize
AvgDownloadTime
PerformanceScore
PercentExitLoss
TextSize
AvgReadingTime
UsabilityAlert
Readability Score


AvgSeekingTime
s1_below_4 18%
s2_4_8 22%
s3_8_12 17%
s4_12_20 23%
s5_20_up 20%
UsabilityScore
s1_below_... 0%
s2_47_63 0%
s3_63_82 0%
s4_82_95 0%
s5_95_up 100%
Accessibili tyScore
AvgDownloadTime
s01_below... 61%
s02_1_2 20%
s03_2_7 11%
s07_7_up 7%
PageSize
Performan ceScore
PercentExitLoss
s1_below_1 27%
s2_1_8 26%
s3_8_17 22%
s4_17_29 13%
s5_29_up 13%
AvgReadingTime
s1_below_4 25%
s2_4_8 11%
s3_8_12 29%
s4_12_25 15%
s5_25_up 20%
TextSize
UsabilityAlert
s1_below_... 30%
s2_24_41 22%
s3_41_70 16%
s4_70_93 24%
s5_93_up 8%
Readability Score
20% high reading time
high usability



Lower high reading time (drop to 15% from 20%) ⇒ low usability



Operational Risks

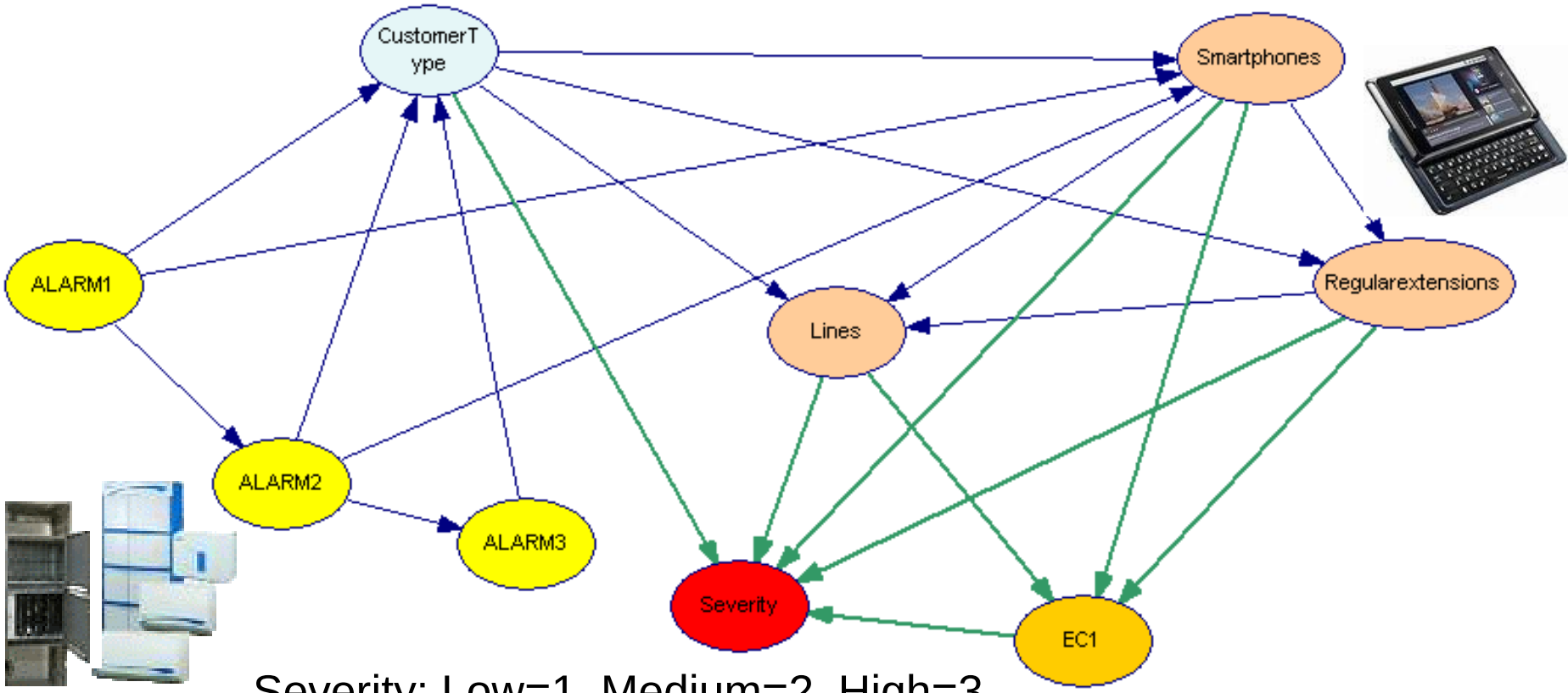
Predictive Application

Case study 3

A: Lines - 4 levels for lines

B: Extensions - 5 levels for regular extension

C: Smart phones - 6 levels for smart phones



Severity: Low=1, Medium=2, High=3

Statistics from the empirical GoF distribution

Run number	Lines	Extensions	Smart phones	GoF mean	GoF Std	5th%	95th%
1	1	1	1	56,60941	0,059105706	56,62155	57
2	1	1	2	59,63072	0,02490878	59,62382	59,88322
3	1	1	3	61,82233	0	61,82233	61,82233
4	1	1	4	61,58676	0,003056467	61,58734	61,66847
5	1	1	5	61,58476	0,002951751	61,58734	61,66847
6	1	1	6	61,82233	0	61,82233	61,82233
7	1	2	1	57,39454	0,048655559	57,39338	57,75812
8	1	2	2	58,16235	0,043181378	58,16356	58,48932
9	1	2	3	61,82233	0	61,82233	61,82233
10	1	2	4	60,37945	0,018637201	60,38212	60,58919
11	1				0	61,82233	61,82233
12	1				0	61,82233	61,82233
13	1				8454	55,29014	55,76737
14	1				0002	58,88973	59,20304
15	1				0	61,82233	61,82233
16	1				0	61,82233	61,82233
17	1				0	61,82233	61,82233
18	1				0	61,82233	61,82233
19	1				5877	54,76769	55,22726
20	1				1381	57,38467	57,76721
21	1				0	61,82233	61,82233
22	1				0	61,82233	61,82233
23	1				6506	61,58734	61,66847
24	1				3469	61,58734	61,67658
25	1	5	1	61,58135	0,002950407	61,57922	61,66847
26	1	5	2	61,58483	0,003111372	61,57922	61,67658
27	1	5	3	61,5814	0,002882963	61,57922	61,66847
28	1	5	4	61,58221	0,003007665	61,58734	61,66847
29	1	5	5	61,58598	0,002768952	61,58734	61,66847
30	1	5	6	61,82233	0	61,82233	61,82233
31	2	1	1	58,63324	0,037792725	58,63446	58,95804
32	2	1	2	58,39079	0,042223195	58,39521	58,7367
33	2	1	3	59,87708	0,025846298	59,87487	60,11655
34	2	1	4	55,81579	0,065592685	55,80323	56,25833
35	2	1	5	61,82233	0	61,82233	61,82233

Robustness of Prediction

From a given BN one can generate simulated outcomes. Goodness of fit (GoF) of a BN model is computed using a distance measure between simulated data and the real data

Classification error

$$\sum_{i=1}^{runs} I_{(s_i \neq \ddot{s}_i)}(x)$$

Indicator function of the subset of severities, s, of the set *X*

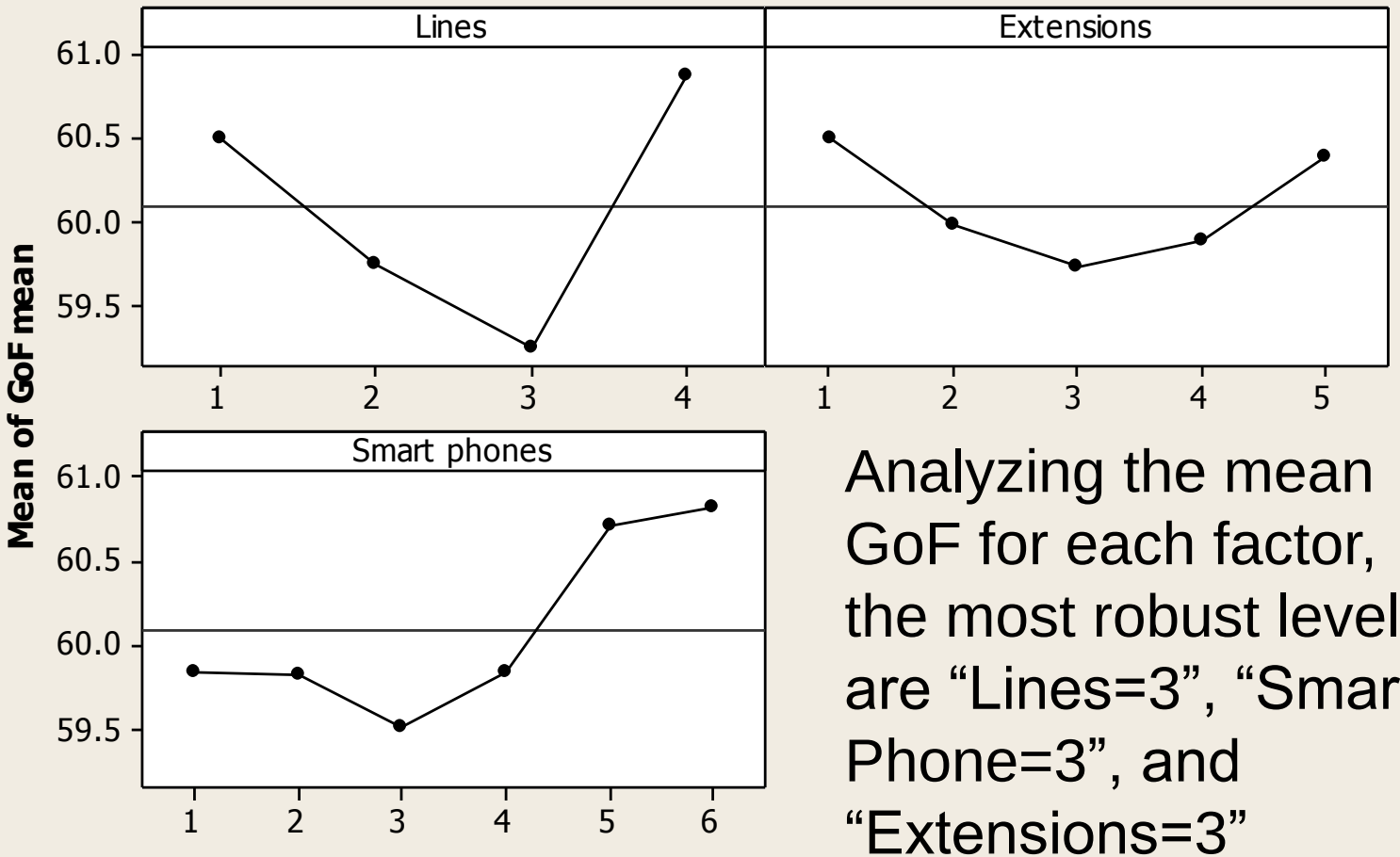
$$I = \begin{cases} 1 \text{ if } x \in \{s : s_i \neq \ddot{s}_i\} \\ 0 \text{ if } x \notin \{s : s_i \neq \ddot{s}_i\} \end{cases}$$

s_i the real data

$\ddot{s}_i$ the i-th simulated values of severity

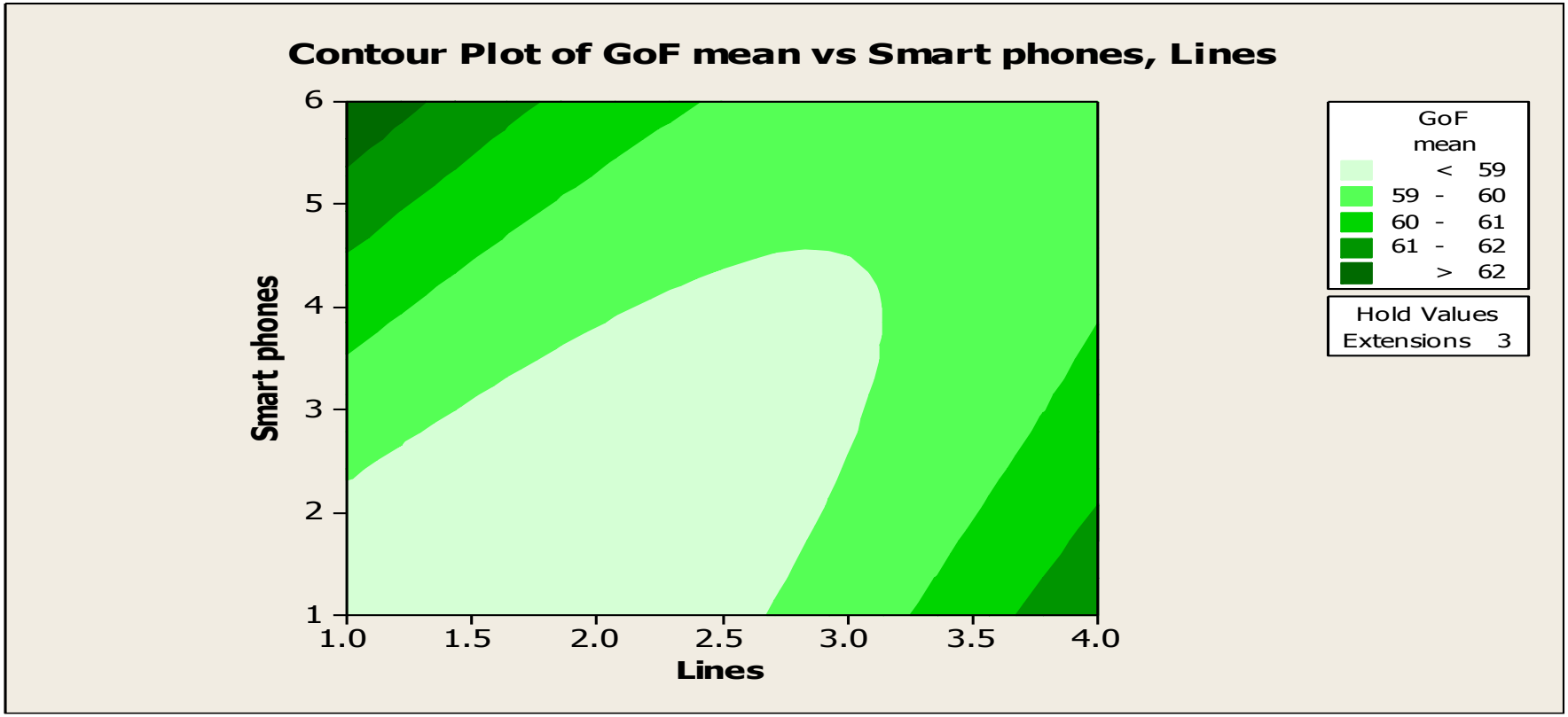
Cornalba, C., Kenett, R.S. and Giudici, P. (2007) Sensitivity Analysis of Bayesian Networks with Stochastic Emulators, ENBIS-DEINDE proceedings, University of Torino, Turin, Italy,

Analysis: Main effect plots of mean GoF

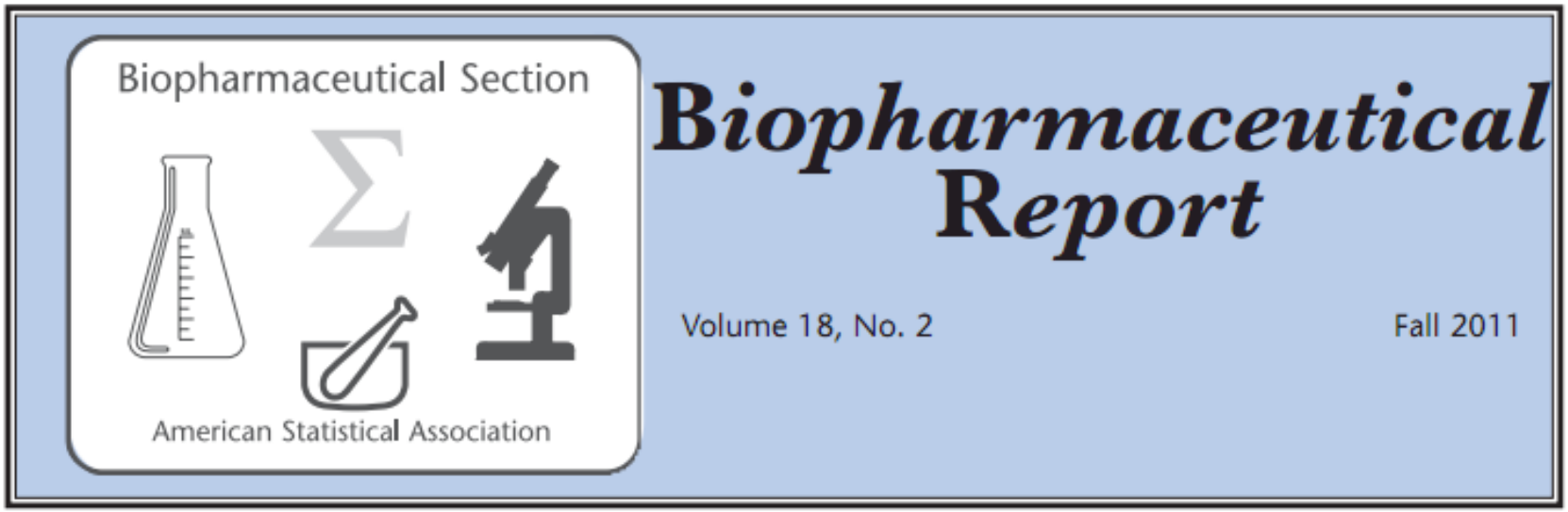


Analyzing the mean GoF for each factor, the most robust levels are “Lines=3”, “Smart Phone=3”, and “Extensions=3”

Analysis: Interaction plots of mean GoF



A limited number of lines and a limited level for smart phone is associated with a low mean GoF

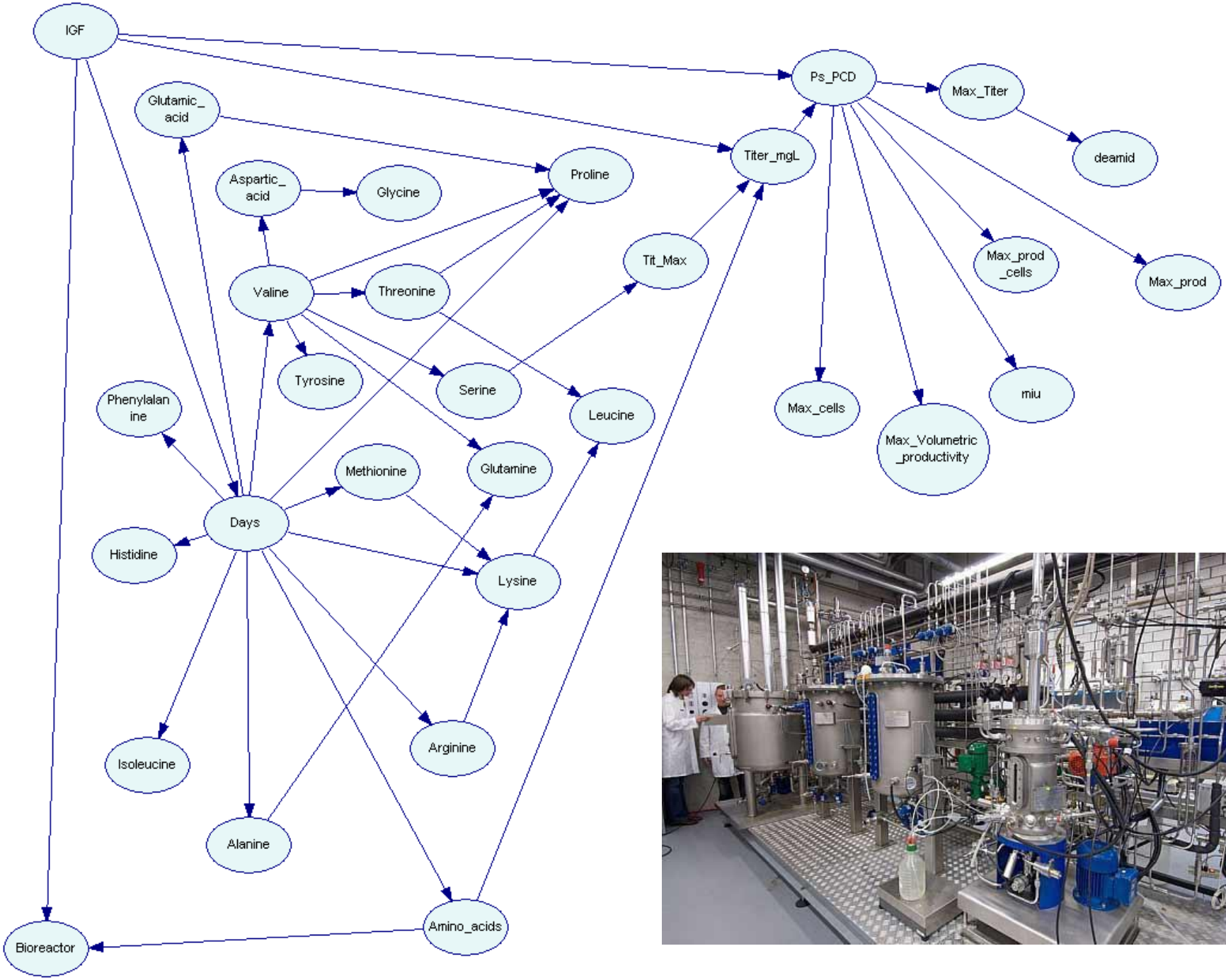


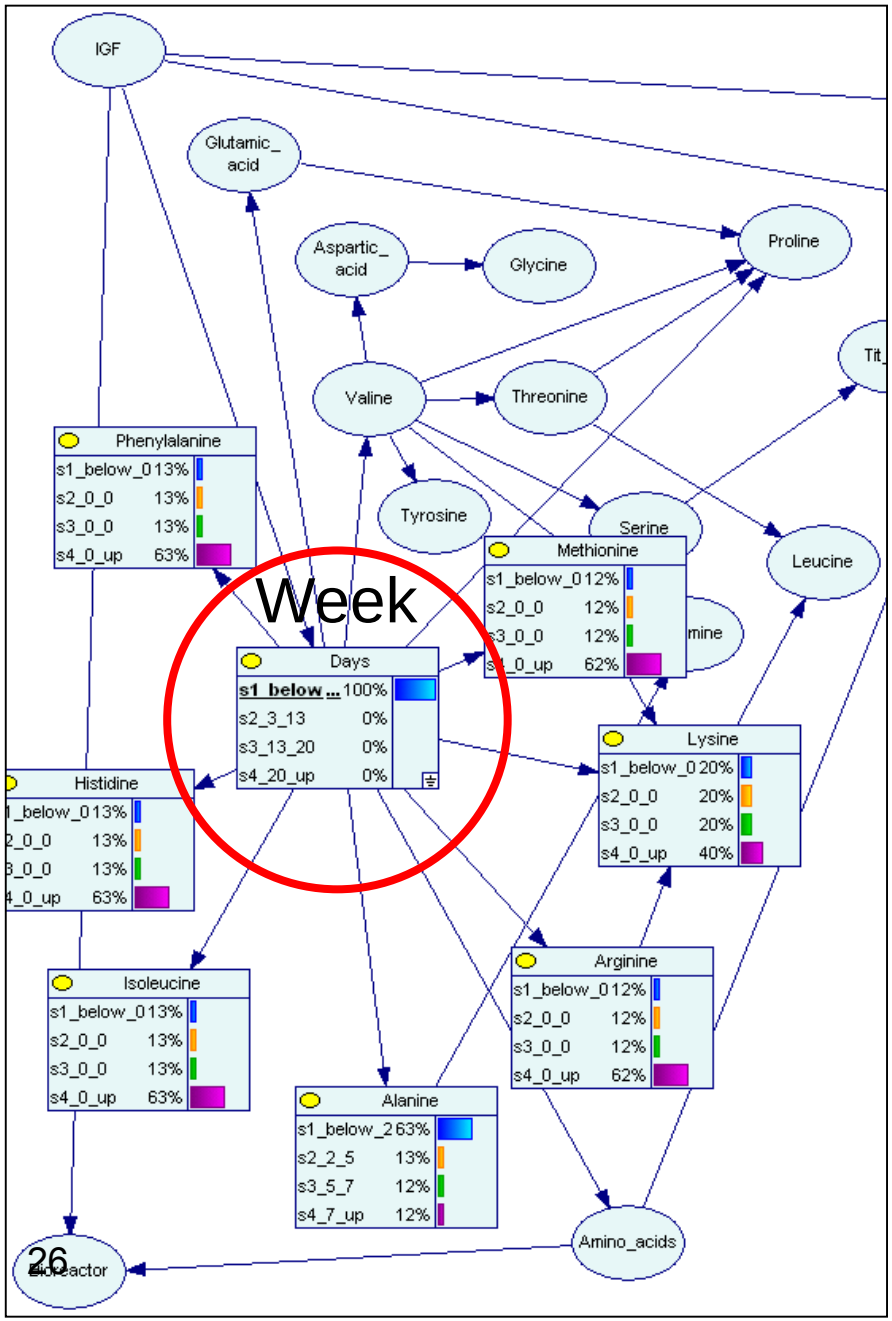
Biotechnology

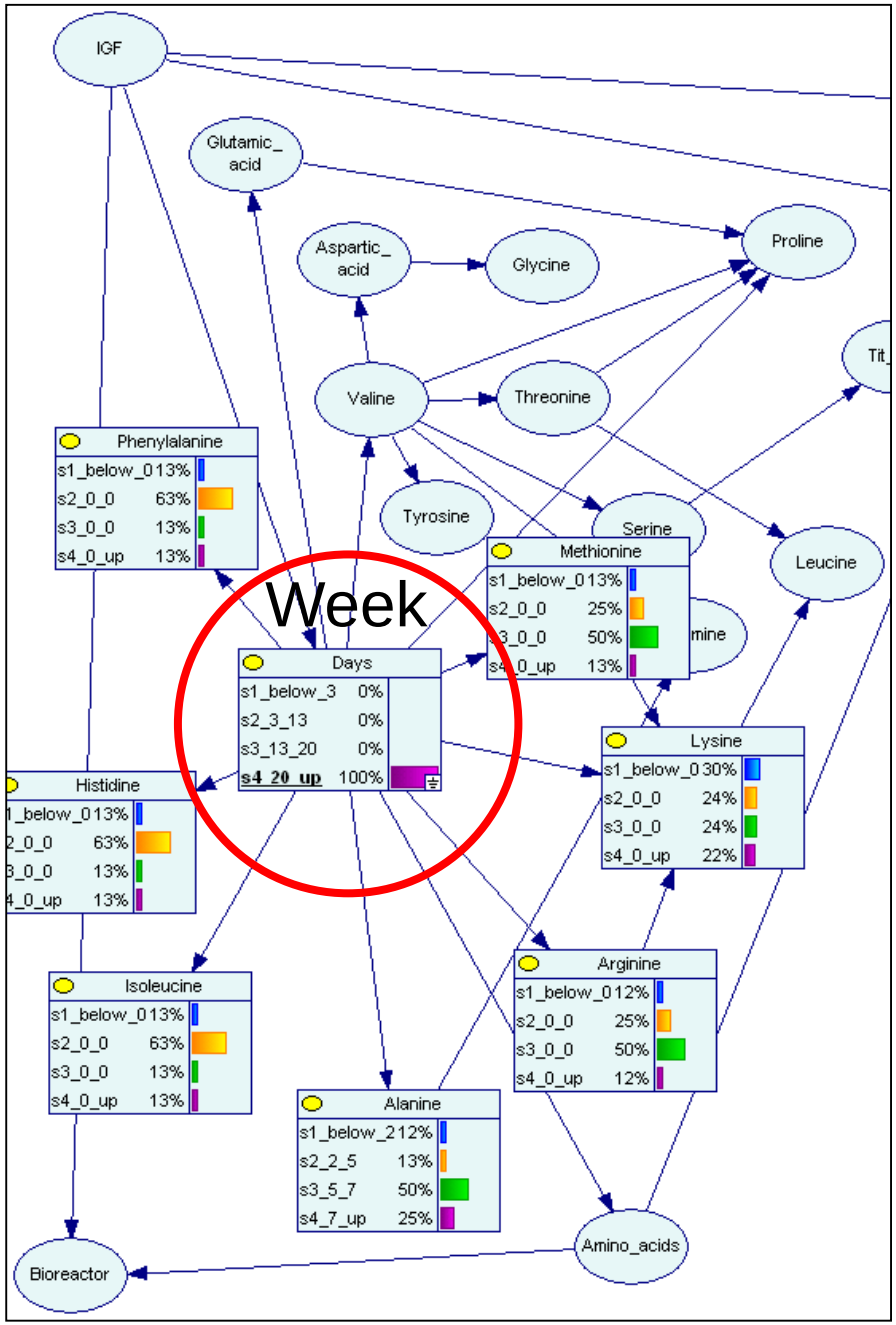
Modeling Effect of Time

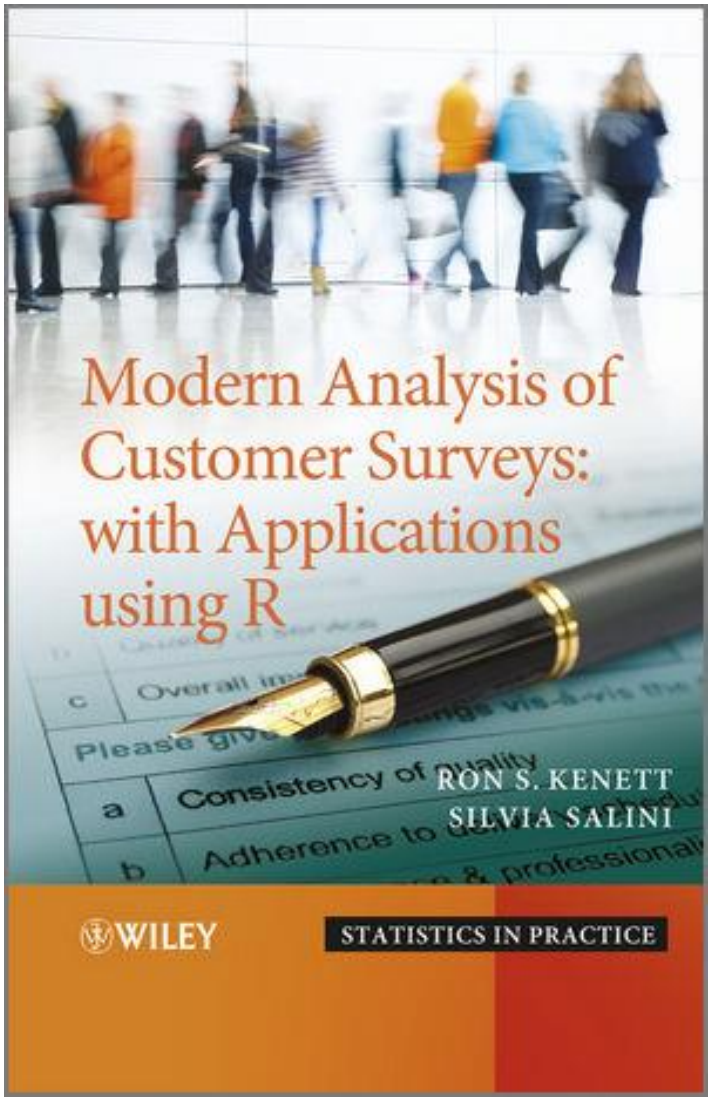
Case study 4

Peterson, J. and Kenett, R.S. (2011), Modelling Opportunities for Statisticians Supporting Quality by Design Efforts for Pharmaceutical Development and Manufacturing, *Biopharmaceutical Report*, ASA Publications, Vol. 18, No. 2, pp. 6-16.


IGF
Ps_PCD
Max_Titer
deamid
Glutamic_
acid
Titer_mgL
Proline
Aspartic_
acid
Glycine
Tit_Max
Max_prod
_cells
Max_prod
Valine
Threonine
Tyrosine
Serine
Leucine
miu
Max_cells
Max_Volumetric
_productivity
Phenylalan
ine
Methionine
Glutamine
Days
Histidine
Lysine
Arginine
Isoleucine
Alanine
Amino_acids
Bioreactor


IGF
Glutamic_ acid
Aspartic_ acid
Glycine
Proline
Valine
Threonine
Tyrosine
Serine
Leucine
Phenylalanine
s1_below_0 13%
s2_0_0 13%
s3_0_0 13%
s4_0_up 63%
Week
Days
s1_below_... 100%
s2_3_13 0%
s3_13_20 0%
s4_20_up 0%
Methionine
s1_below_0 12%
s2_0_0 12%
s3_0_0 12%
s4_0_up 62%
Lysine
s1_below_0 20%
s2_0_0 20%
s3_0_0 20%
s4_0_up 40%
Histidine
Isoleucine
s1_below_0 13%
s2_0_0 13%
s3_0_0 13%
s4_0_up 63%
Arginine
s1_below_0 12%
s2_0_0 12%
s3_0_0 12%
s4_0_up 62%
Alanine
s1_below_2 63%
s2_2_5 13%
s3_5_7 12%
s4_7_up 12%
Amino_acids
Bioreactor


IGF
Glutamic_ acid
Aspartic_ acid
Glycine
Proline
Valine
Threonine
Tyrosine
Serine
Leucine
Phenylalanine
s1_below_0 13%
s2_0_0 63%
s3_0_0 13%
s4_0_up 13%
Week
Days
s1_below_3 0%
s2_3_13 0%
s3_13_20 0%
s4_20_up 100%
Methionine
s1_below_0 13%
s2_0_0 25%
s3_0_0 50%
s4_0_up 13%
Lysine
s1_below_0 30%
s2_0_0 24%
s3_0_0 24%
s4_0_up 22%
Histidine
Isoleucine
s1_below_0 13%
s2_0_0 63%
s3_0_0 13%
s4_0_up 13%
Arginine
s1_below_0 12%
s2_0_0 25%
s3_0_0 50%
s4_0_up 12%
Alanine
s1_below_2 12%
s2_2_5 13%
s3_5_7 50%
s4_7_up 25%
Amino_acids
Bioreactor



Customer Surveys

Diagnostic Application

Case study 5

The ABC 2010 Annual Customer Satisfaction Survey

Company: _____

Completed by: _____

Title/Position: 1. Owner 2. Management 3. Technical Management 4. Technical Sta
5. Operator 6. Administrator 7. Other, please specify: _____

Dear Customer,

For each of the following statements, please select a number indicating the extent of your agreement with the statement concerning your experience with ABC during 2010. Then, under "Importance Level", select another number indicating the importance of the statement to you. If a certain statement is not relevant or not applicable, please select N/A.

Overall Satisfaction from ABC

		Evaluation					
		Very low				Very high	
1.	Overall satisfaction level from ABC:	1	2	3	4	5	
2.	Overall satisfaction level from ABC's improvements during 2010:	1	2	3	4	5	
3.	Is ABC your best supplier?	a. Yes b. No					
		Very unlikely				Very likely	
4.	Would you recommend ABC to other companies?	1	2	3	4	5	
5.	If you were in the market to buy a PRODUCT, how likely would it be for you to purchase an ABC product again?	1	2	3	4	5	

Equipment and System

		Evaluation					Importance Level			
		Strongly disagree				Strongly agree	Low		High	
6.	The equipment's features and capabilities meet your needs.	1	2	3	4	5	1	2	3	N/A
7.	Improvements and upgrades provide value.	1	2	3	4	5	1	2	3	N/A
8.	Output quality meets or exceeds expectations.	1	2	3	4	5	1	2	3	N/A
9.	Uptime is acceptable.	1	2	3	4	5	1	2	3	N/A
10.	For customers who purchased a system during 2010: ABC's equipment proposal met your requirements.	1	2	3	4	5	1	2	3	N/A
		Very low				Very high				
11.	Overall satisfaction level from the **equipment**:	1	2	3	4	5				

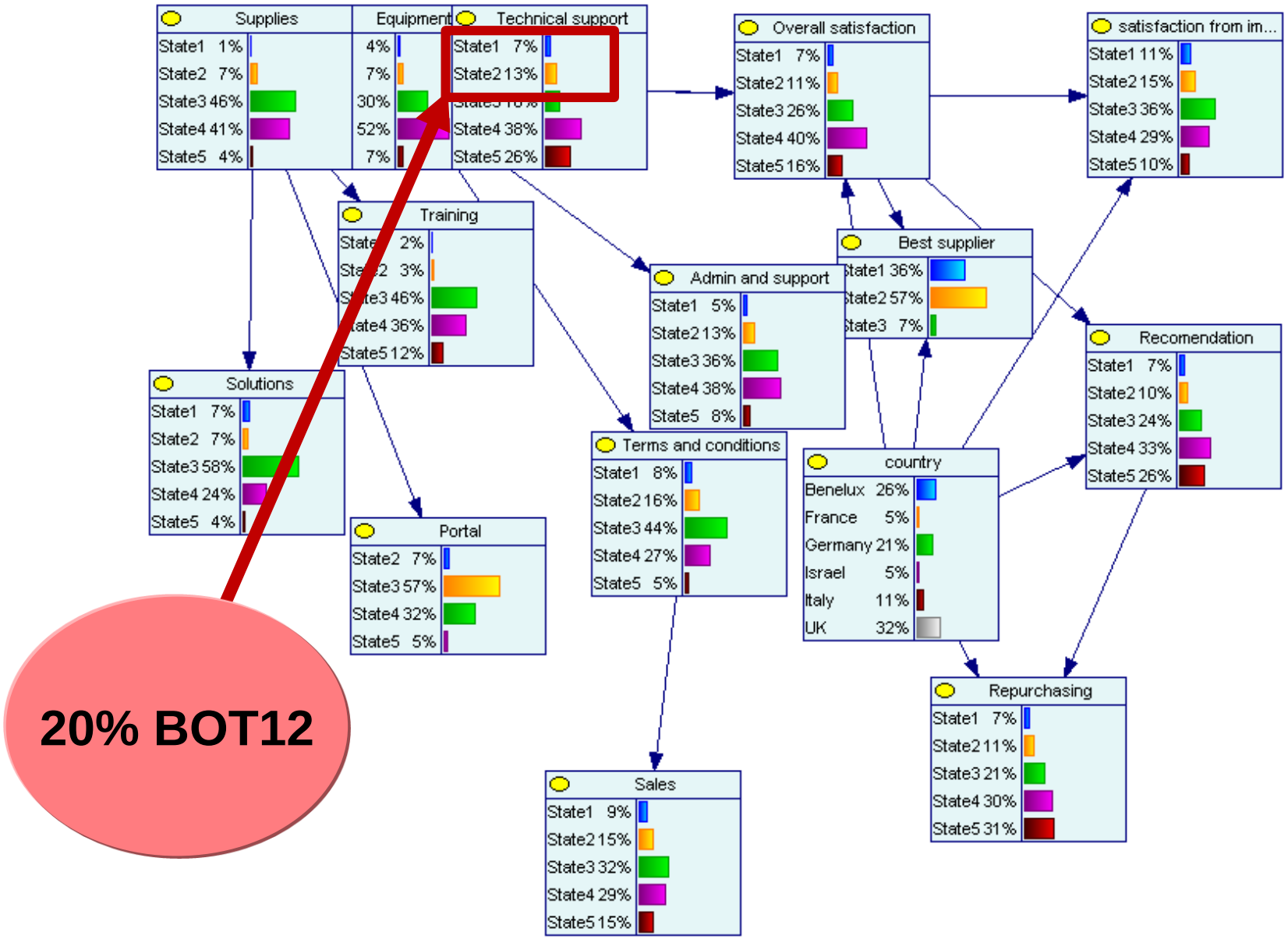
Sales Support

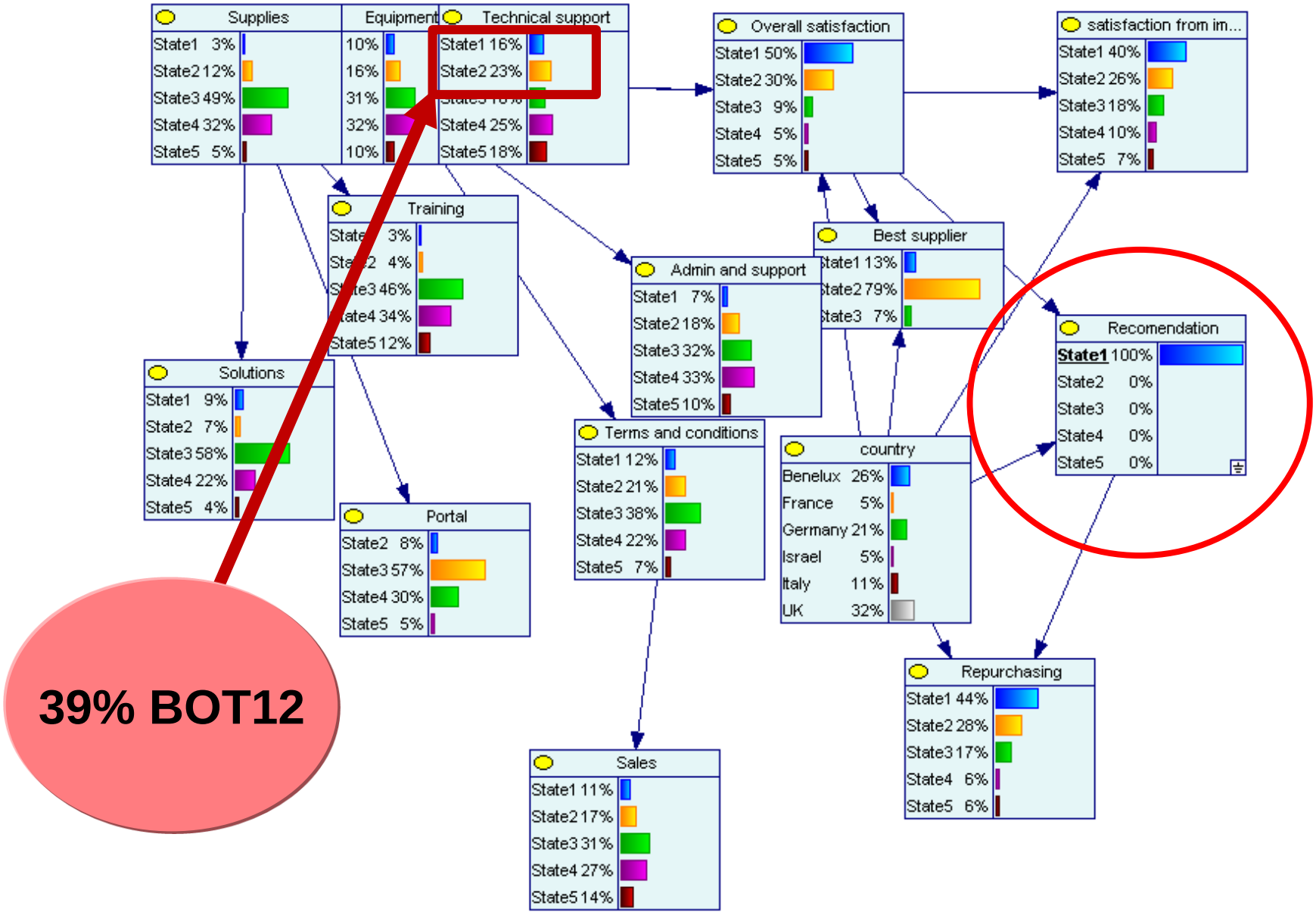
		Evaluation					Importance Level			
		Strongly disagree				Strongly agree	Low		High	
12.	Verbal promises have been honored.	1	2	3	4	5	1	2	3	N/A
13.	Sales personnel communicate frequently enough with you.	1	2	3	4	5	1	2	3	N/A
14.	Sales personnel respond promptly to requests.	1	2	3	4	5	1	2	3	N/A
15.	Sales personnel are knowledgeable about equipment.	1	2	3	4	5	1	2	3	N/A
16.	Sales personnel are knowledgeable about market opportunities.	1	2	3	4	5	1	2	3	N/A
		Very low				Very high				
17.	Overall satisfaction level from **sales support**:	1	2	3	4	5				

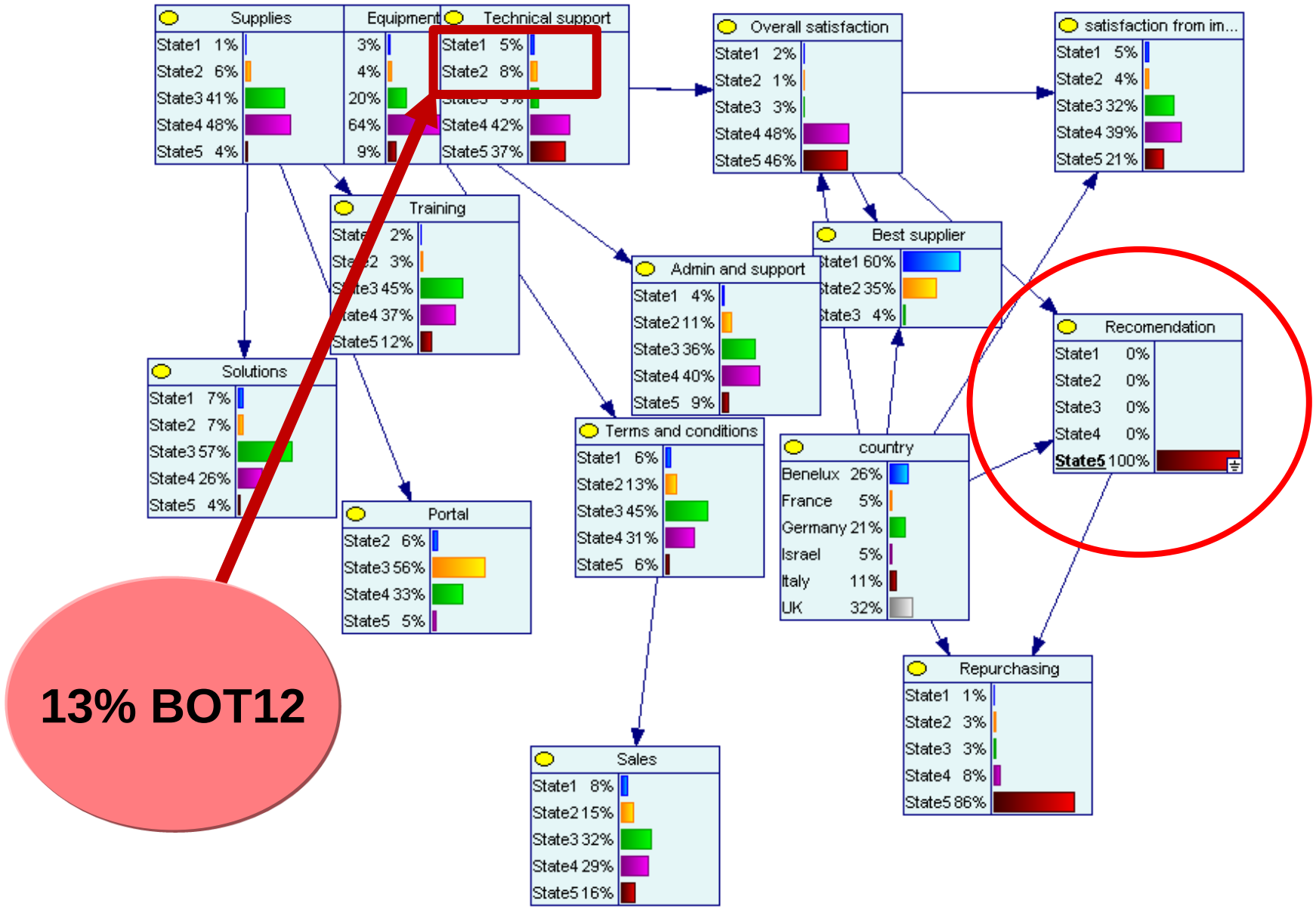
Technical Support

		Evaluation					Importance Level			
		Strongly disagree				Strongly agree	Low		High	
18.	Technical support is available when needed.	1	2	3	4	5	1	2	3	N/A
19.	The technical staff is knowledgeable.	1	2	3	4	5	1	2	3	N/A
20.	The technical staff is well informed about the latest equipment updates/enhancements.	1	2	3	4	5	1	2	3	N/A
21.	Parts are available when needed.	1	2	3	4	5	1	2	3	N/A
22.	The remote support care center is valuable and meets your expectations.	1	2	3	4	5	1	2	3	N/A
23.	Problems are resolved within the required time frame.	1	2	3	4	5	1	2	3	N/A
24.	The technical staff is courteous and helpful.	1	2	3	4	5	1	2	3	N/A
		Very low				Very high				
25.	Overall satisfaction level from **technical support**:	1	2	3	4	5				

Kenett R.S. and Salini S. (2009). New Frontiers: Bayesian networks give insight into survey-data analysis, *Quality Progress*, pp. 31-36, August.

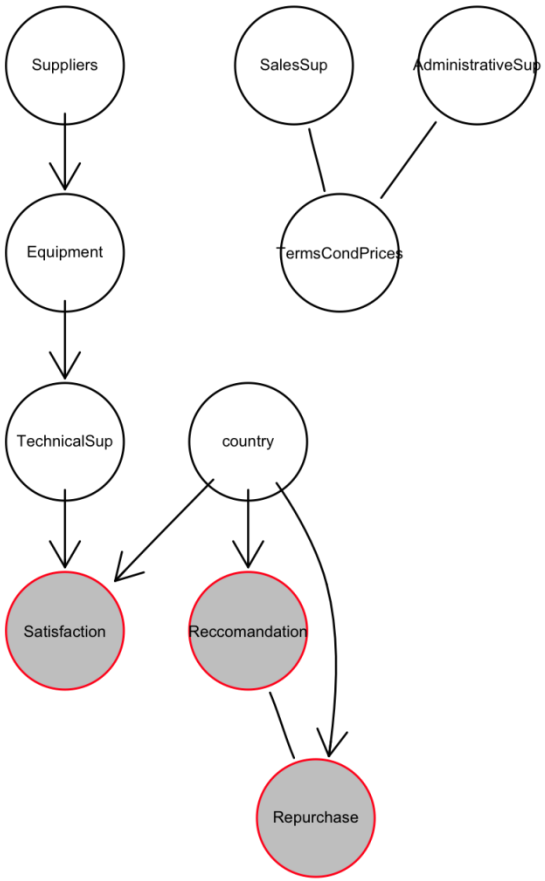

Supplies
State1 1%
State2 7%
State3 46%
State4 41%
State5 4%
Equipment
4%
7%
30%
52%
7%
Technical support
State1 7%
State2 13%
State4 38%
State5 26%
Overall satisfaction
State1 7%
State2 11%
State3 26%
State4 40%
State5 16%
satisfaction from im...
State1 11%
State2 15%
State3 36%
State4 29%
State5 10%
Training
2%
3%
46%
36%
State5 12%
Best supplier
State1 36%
State2 57%
State3 7%
Admin and support
State1 5%
State2 13%
State3 36%
State4 38%
State5 8%
Recomendation
State1 7%
State2 10%
State3 24%
State4 33%
State5 26%
Solutions
State1 7%
State2 7%
State3 58%
State4 24%
State5 4%
Terms and conditions
State1 8%
State2 16%
State3 44%
State4 27%
State5 5%
country
Benelux 26%
France 5%
Germany 21%
Israel 5%
Italy 11%
UK 32%
Portal
State2 7%
State3 57%
State4 32%
State5 5%
Repurchasing
State1 7%
State2 11%
State3 21%
State4 30%
State5 31%
Sales
State1 9%
State2 15%
State3 32%
State4 29%
State5 15%
20% BOT12


Supplies
State1 3%
State2 12%
State3 49%
State4 32%
State5 5%
Equipment
10%
16%
31%
32%
10%
Technical support
State1 16%
State2 23%
State4 25%
State5 18%
Overall satisfaction
State1 50%
State2 30%
State3 9%
State4 5%
State5 5%
satisfaction from im...
State1 40%
State2 26%
State3 18%
State4 10%
State5 7%
Training
3%
4%
State3 46%
State4 34%
State5 12%
Best supplier
State1 13%
State2 79%
State3 7%
Admin and support
State1 7%
State2 18%
State3 32%
State4 33%
State5 10%
Recomendation
State1 100%
State2 0%
State3 0%
State4 0%
State5 0%
Solutions
State1 9%
State2 7%
State3 58%
State4 22%
State5 4%
Terms and conditions
State1 12%
State2 21%
State3 38%
State4 22%
State5 7%
country
Benelux 26%
France 5%
Germany 21%
Israel 5%
Italy 11%
UK 32%
Portal
State2 8%
State3 57%
State4 30%
State5 5%
Repurchasing
State1 44%
State2 28%
State3 17%
State4 6%
State5 6%
Sales
State1 11%
State2 17%
State3 31%
State4 27%
State5 14%
39% BOT12

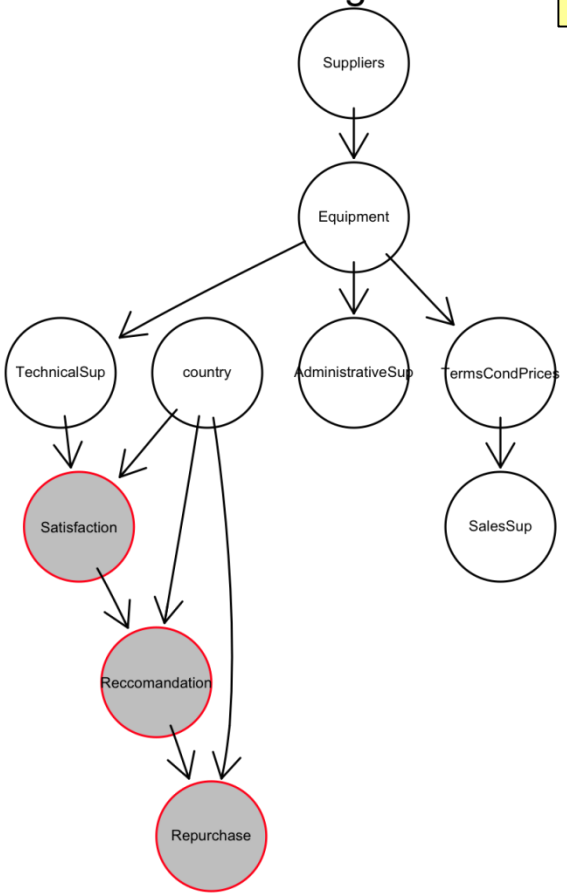

Supplies
State1 1%
State2 6%
State3 41%
State4 48%
State5 4%
Equipment
3%
4%
20%
64%
9%
Technical support
State1 5%
State2 8%
State4 42%
State5 37%
Overall satisfaction
State1 2%
State2 1%
State3 3%
State4 48%
State5 46%
satisfaction from im...
State1 5%
State2 4%
State3 32%
State4 39%
State5 21%
Training
2%
3%
45%
37%
State5 12%
Best supplier
60%
35%
4%
Admin and support
State1 4%
State2 11%
State3 36%
State4 40%
State5 9%
Recomendation
State1 0%
State2 0%
State3 0%
State4 0%
State5 100%
Solutions
State1 7%
State2 7%
State3 57%
State4 26%
State5 4%
Terms and conditions
State1 6%
State2 13%
State3 45%
State4 31%
State5 6%
country
Benelux 26%
France 5%
Germany 21%
Israel 5%
Italy 11%
UK 32%
Portal
State2 6%
State3 56%
State4 33%
State5 5%
Repurchasing
State1 1%
State2 3%
State3 3%
State4 8%
State5 86%
Sales
State1 8%
State2 15%
State3 32%
State4 29%
State5 16%
13% BOT12

Network learning

Constraint-based algorithms

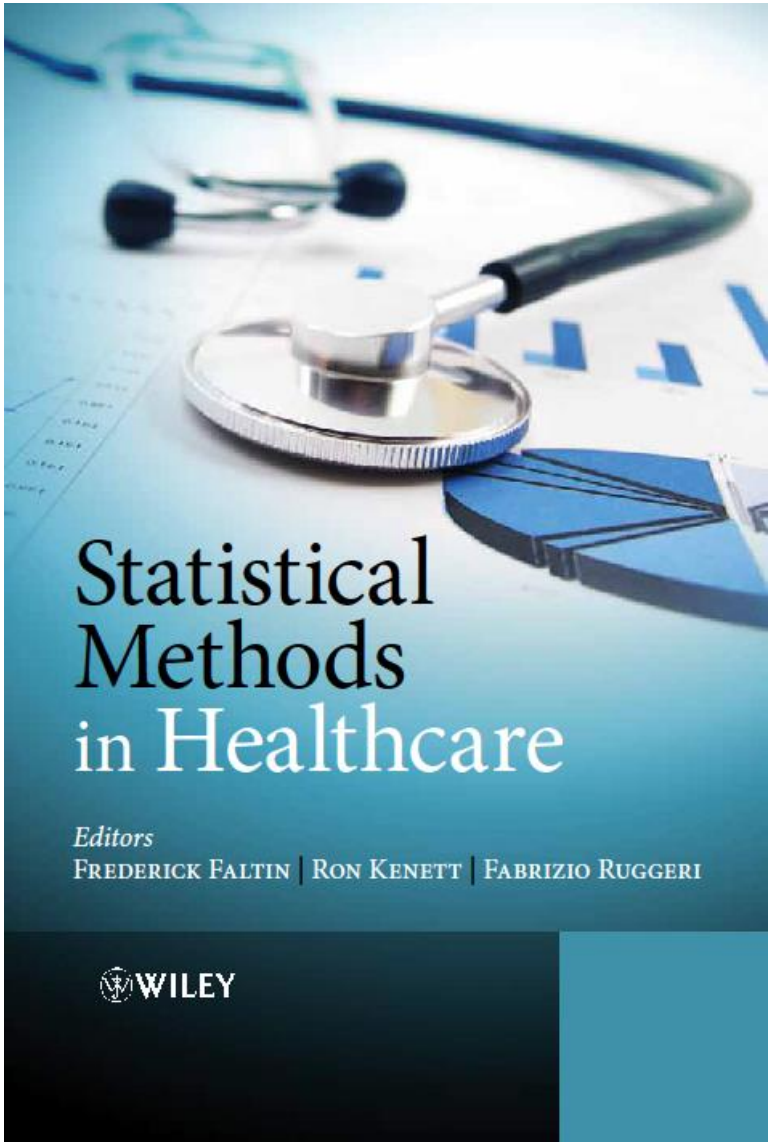


Scored-based algorithms



Robustness of Network

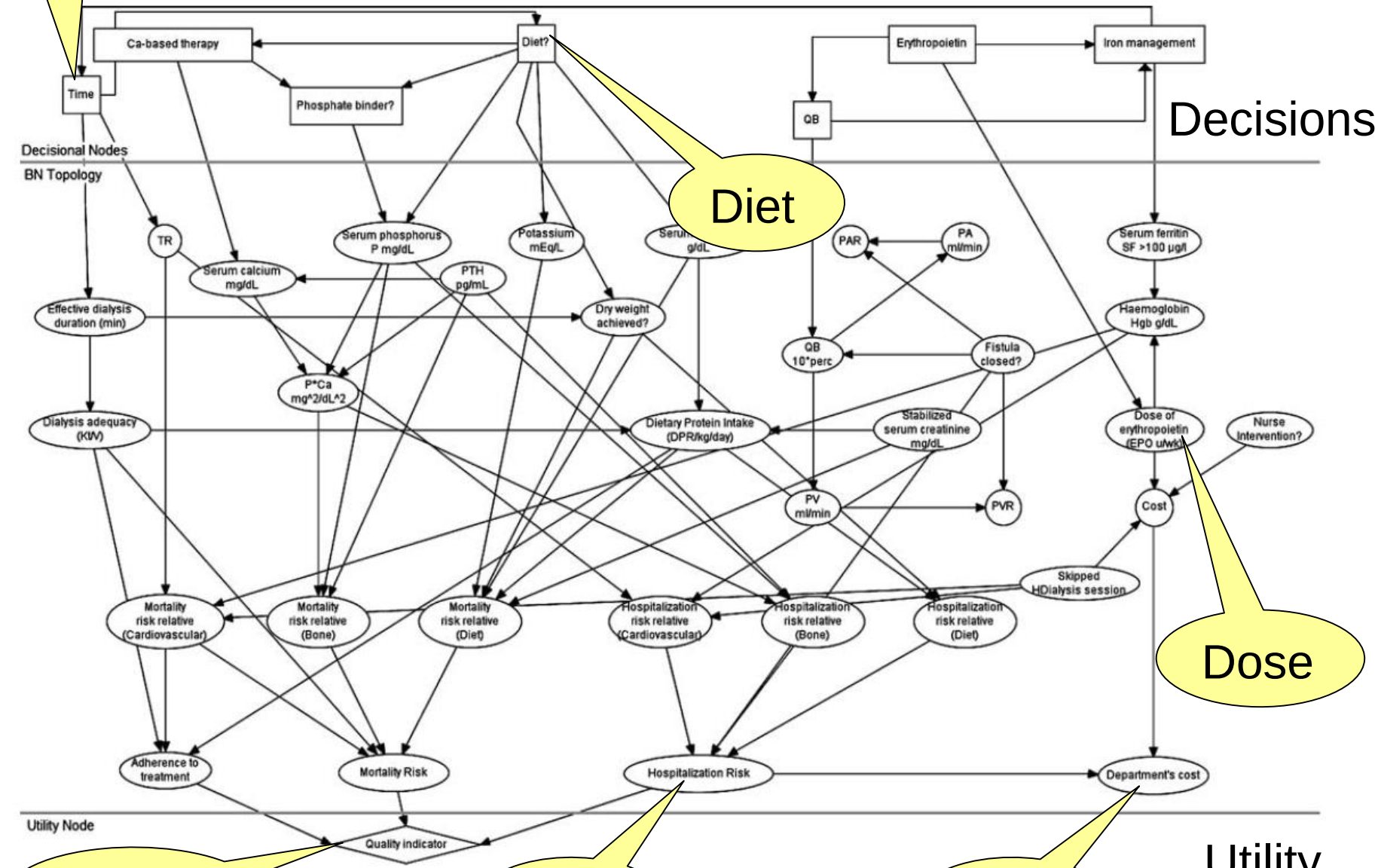
bnlearn: 1) Grow-Shrink (GS) algorithm, 2) Incremental Association (IAMB) algorithm, 3) Interleaved-IAMB (Inter-IAMB) algorithm, 4) Fast-IAMB (Fast-IAMB) algorithm, 5) Max-Min Parents and Children (MMPC) algorithm, 6) ARACNE and Chow-Liu algorithms, 7) Hill- Climbing (HC) greedy search algorithm, 8)Tabu Search (TABU) algorithm, 9) Max-Min Hill-Climbing (MMHC) algorithm and 10) two-stage Restricted Maximization (RSMAX2) algorithm for both discrete and Gaussian networks



Health Care

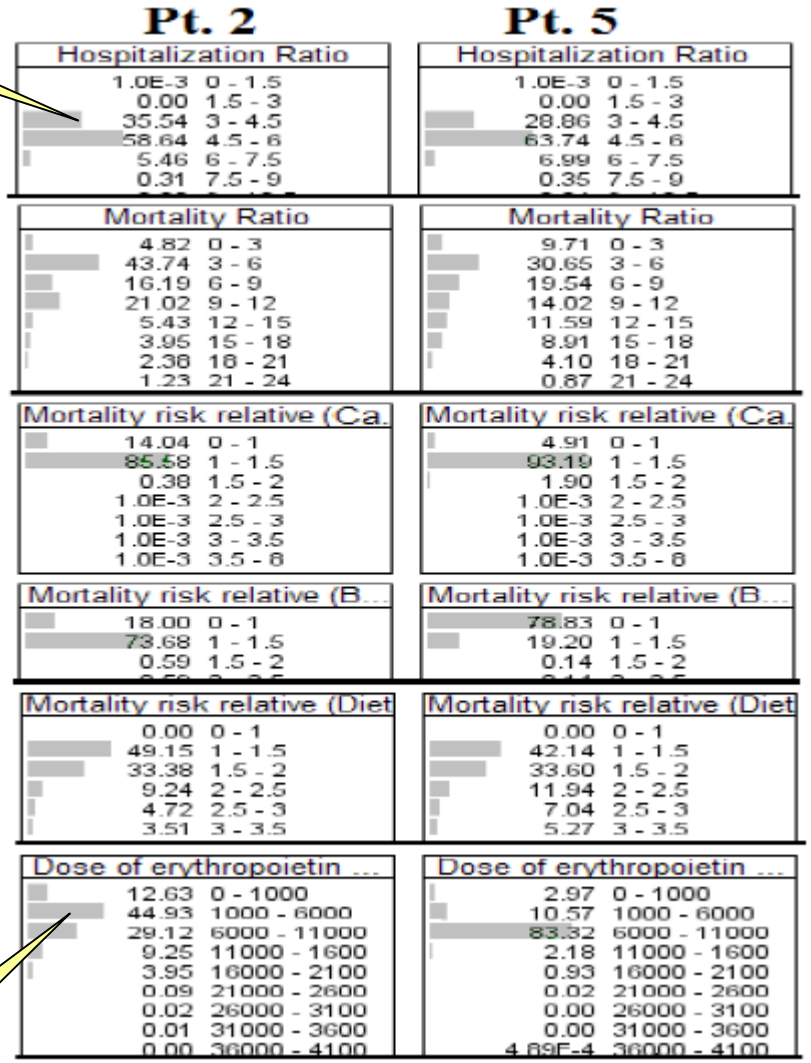
Decision Support Systems

Case study 6

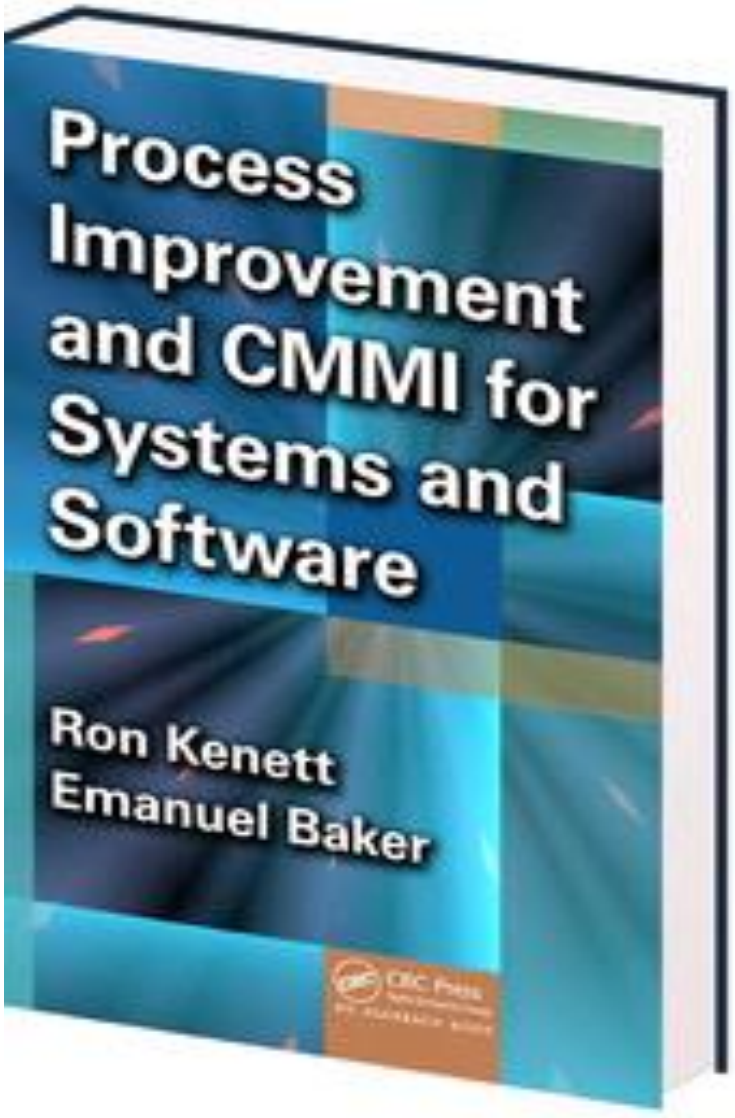

Time
Ca-based therapy
Diet?
Erythropoietin
Iron management
Time
Phosphate binder?
QB
Decisions
Decisional Nodes
BN Topology
Diet
TR
Serum phosphorus P mg/dL
Potassium mEq/L
PAR
PA ml/min
Serum ferritin SF >100 µg/l
Serum calcium mg/dL
PTH pg/mL
Effective dialysis duration (min)
Dry weight achieved?
Haemoglobin Hgb g/dL
QB 10*perc
Fistula closed?
P*Ca mg^2/dL^2
Dialysis adequacy (Kt/V)
Dietary Protein Intake (DPR/kg/day)
Stabilized serum creatinine mg/dL
Dose of erythropoietin (EPO u/wk)
Nurse intervention?
PV ml/min
PVR
Cost
Skipped HDialysis session
Mortality risk relative (Cardiovascular)
Mortality risk relative (Bone)
Mortality risk relative (Diet)
Hospitalization risk relative (Cardiovascular)
Hospitalization risk relative (Bone)
Hospitalization risk relative (Diet)
Dose
Adherence to treatment
Mortality Risk
Hospitalization Risk
Department's cost
Utility Node
Quality indicator
Utility
Quality
Risk
Cost

Risk

Decision	State of Actions
Ca-based Therapy	* Treat hypercalcemia; * Continue current therapy; * Decrease vitamin D dose to achieve ideal Ca; Decrease Ca-based . . . * Decrease or discontinue vitamin D dose or Ca-based phosphate binders; . . .
Phosphate binder?	* Assess nutrition, discontinue phosphate binder if . . . * Being dietary counseling and restrict dietary phosphate; start or increase . . . * Being short-term Al-based phosphate binder use, then increase . . .
QB	* Increase QB; * Keep QB; * Decrease QB.
Erythro-poietin	* Keep the current dose; . . .
Iron Management	* Keep the treatment; . . .
Time	. . .
Diet?	. . .



Dose



Quality Report

Software Testing

Case study 7

Definitions

Unit number: The software development team that developed the software version

PM: The total number of person months used to develop and test the software version

Escaping Defects: The number of Severity 1-3 defects detected during the first three months after release of software version divided by PM -the lower the better.

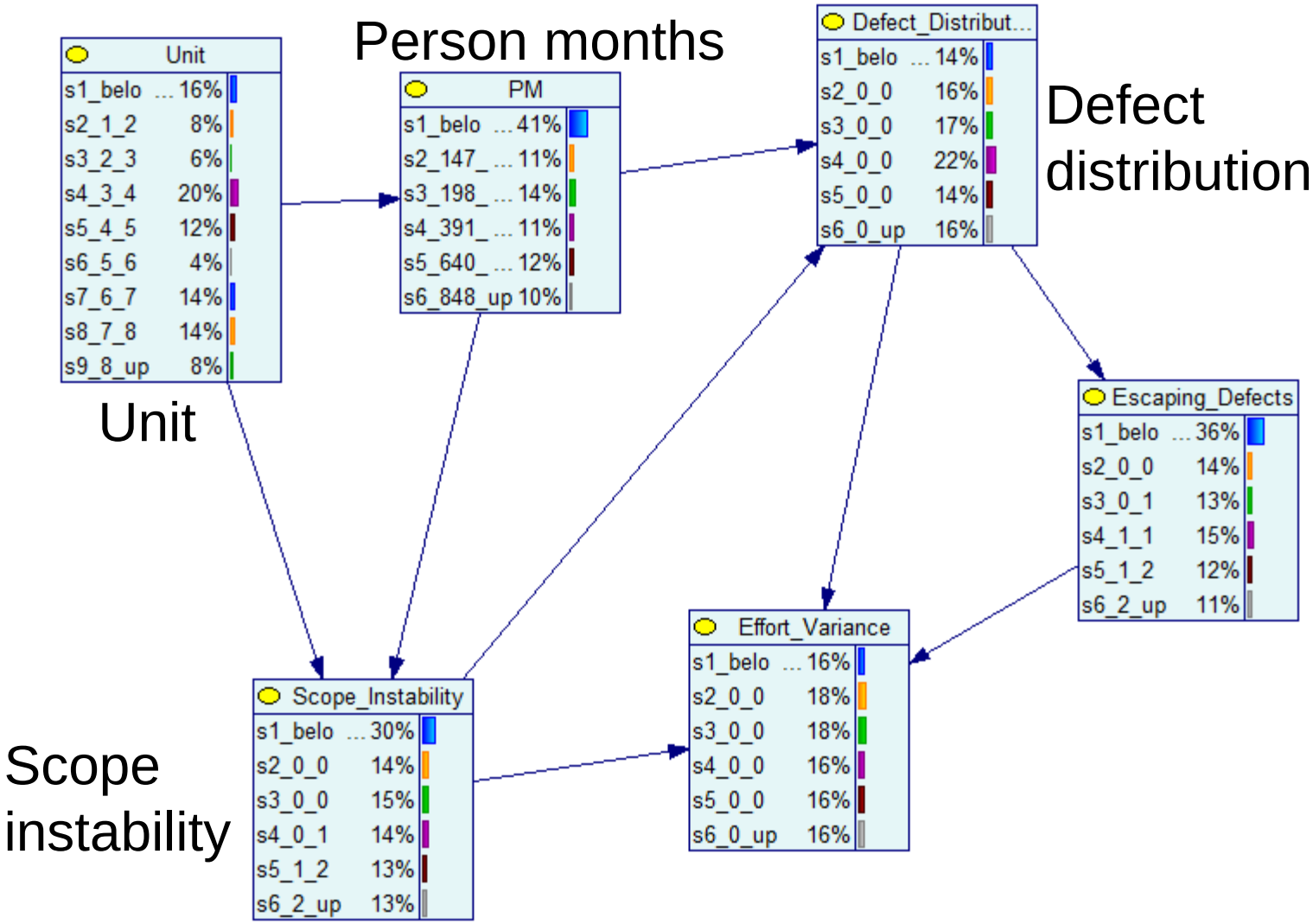
Defect Distribution: The proportion of defects detected during the various internal testing phases versus defects detected by customers during the first three months after release of software version -the higher the better.

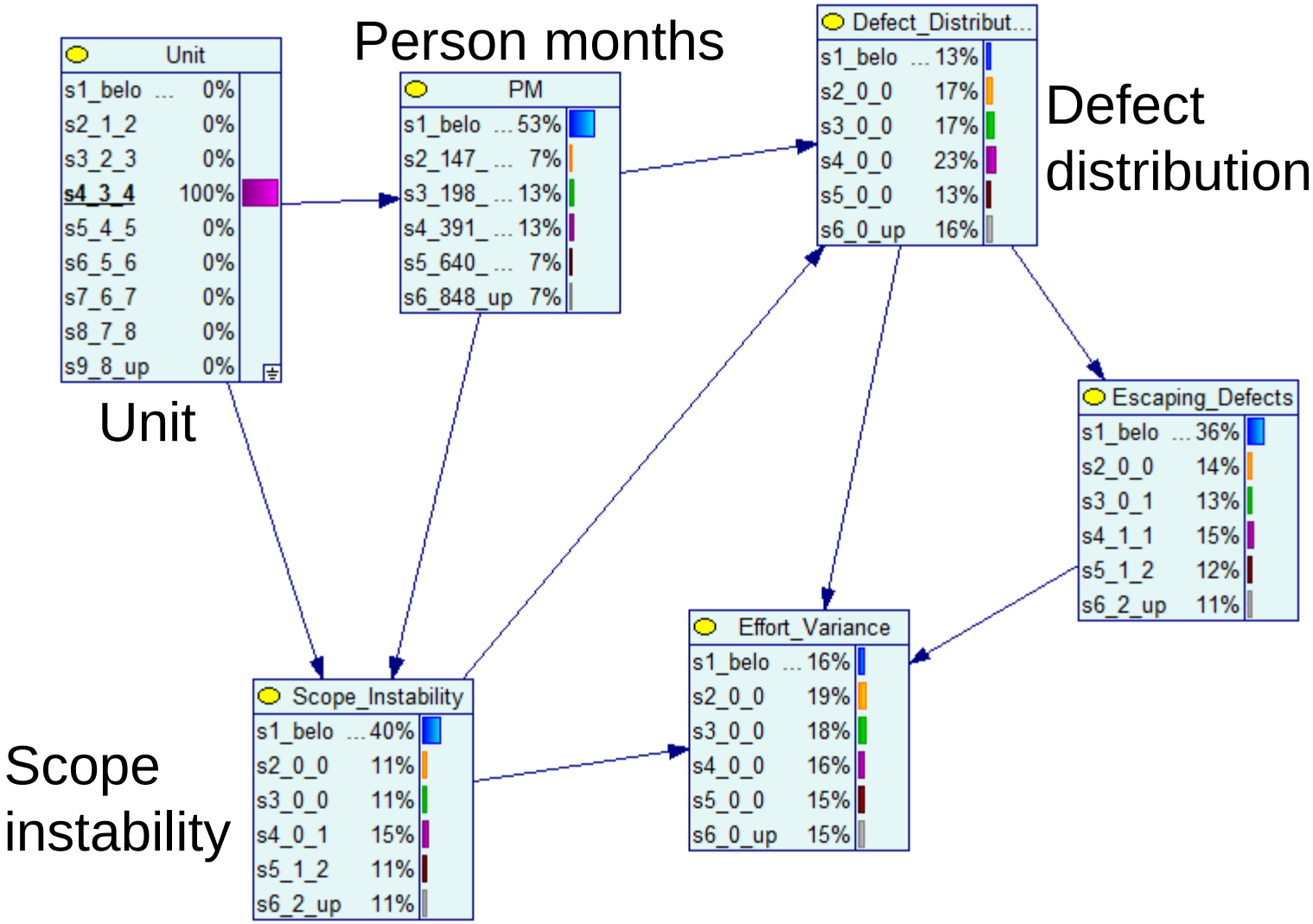
Effort Variance: The effort variance, in PM, from the budget baseline of the software version during the release development period

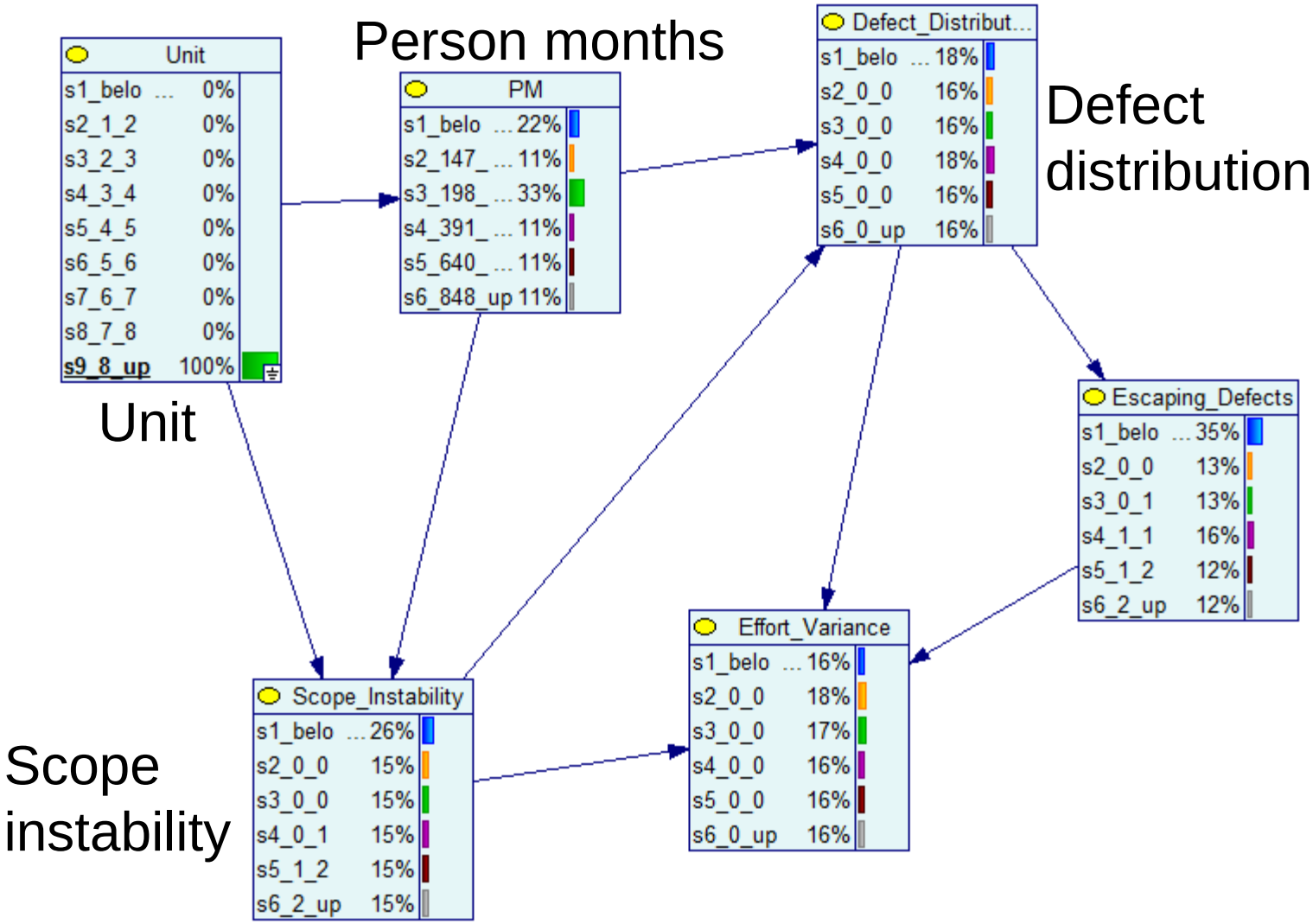
Scope Instability: Represents changes to requirements scoped to the release and stability of the software development baseline. i.e. number of changed, deleted and added requirements after scope sign-off date divided by the number of requirements in the release scope at release sign-off date.

Kenett, R.S. and Baker, E. (2010). *Process Improvement and CMMI for Systems and Software,* Taylor and Francis, Auerbach CRC Publications

Unit	PM	Escaping Defects	Defect Distribution	Effort Variance	Scope Instability
1.0	727	0.31	0.71	-0.03	0.01
1.0	250	1.63	0.37	-0.09	0.66
1.0	773	0.12	0.92	0.04	0.14
1.0	49	0.05	0.89	-0.07	0.14
1.0	52	0.23	0.74	-0.09	0.03
1.0	923	0.08	0.92	-0.04	0.00
1.0	21	1.27	0.52	0.11	0.00
2.0	7	0.43	0.76	0.00	0.00
2.0	60	0.20	0.61	0.05	0.05
2.0	105	0.40	0.74	0.02	0.15
3.0	747	0.60	0.53	0.31	0.80
3.0	25	1.87	0.57	0.09	0.21
4.0	230	1.27	0.16	0.36	0.00
4.0	553	3.31	0.14	0.92	0.00
4.0	57	0.93	0.66	0.06	0.20
4.0	29	0.75	0.58	-0.10	0.03
4.0	60	0.63	0.66	-0.18	0.73
4.0	16	0.06	0.92	-0.03	0.00
4.0	36	0.90	0.41	0.00	0.00
4.0	37	0.38	0.39	0.09	0.01
4.0	86	0.69	0.73	0.08	0.00
5.0	157	0.31	0.68	0.02	0.57
5.0	182	0.99	0.57	0.12	2.25
5.0	35	0.46	0.56	0.09	3.46


Unit
s1_belo ... 16%
s2_1_2 8%
s3_2_3 6%
s4_3_4 20%
s5_4_5 12%
s6_5_6 4%
s7_6_7 14%
s8_7_8 14%
s9_8_up 8%
Person months
PM
s1_belo ... 41%
s2_147_ ... 11%
s3_198_ ... 14%
s4_391_ ... 11%
s5_640_ ... 12%
s6_848_up 10%
Defect_Distribut...
s1_belo ... 14%
s2_0_0 16%
s3_0_0 17%
s4_0_0 22%
s5_0_0 14%
s6_0_up 16%
Defect distribution
Escaping_Defects
s1_belo ... 36%
s2_0_0 14%
s3_0_1 13%
s4_1_1 15%
s5_1_2 12%
s6_2_up 11%
Effort_Variance
s1_belo ... 16%
s2_0_0 18%
s3_0_0 18%
s4_0_0 16%
s5_0_0 16%
s6_0_up 16%
Scope_Instability
s1_belo ... 30%
s2_0_0 14%
s3_0_0 15%
s4_0_1 14%
s5_1_2 13%
s6_2_up 13%
Scope instability

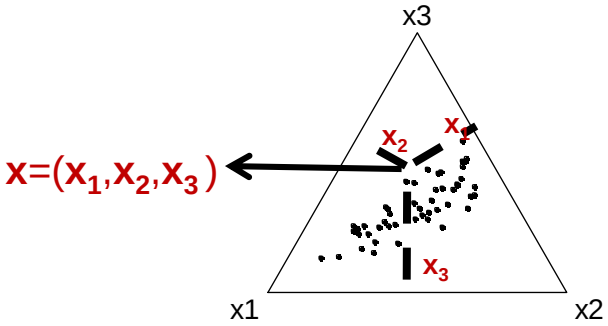

Unit
s1_belo ... 0%
s2_1_2 0%
s3_2_3 0%
s4_3_4 100%
s5_4_5 0%
s6_5_6 0%
s7_6_7 0%
s8_7_8 0%
s9_8_up 0%
Unit
Person months
PM
s1_belo ... 53%
s2_147_... 7%
s3_198_... 13%
s4_391_... 13%
s5_640_... 7%
s6_848_up 7%
Defect_Distribut...
s1_belo ... 13%
s2_0_0 17%
s3_0_0 17%
s4_0_0 23%
s5_0_0 13%
s6_0_up 16%
Defect distribution
Escaping_Defects
s1_belo ... 36%
s2_0_0 14%
s3_0_1 13%
s4_1_1 15%
s5_1_2 12%
s6_2_up 11%
Effort_Variance
s1_belo ... 16%
s2_0_0 19%
s3_0_0 18%
s4_0_0 16%
s5_0_0 15%
s6_0_up 15%
Scope_Instability
s1_belo ... 40%
s2_0_0 11%
s3_0_0 11%
s4_0_1 15%
s5_1_2 11%
s6_2_up 11%
Scope instability


Unit
s1_belo ... 0%
s2_1_2 0%
s3_2_3 0%
s4_3_4 0%
s5_4_5 0%
s6_5_6 0%
s7_6_7 0%
s8_7_8 0%
s9_8_up 100%
Person months
PM
s1_belo ... 22%
s2_147_ ... 11%
s3_198_ ... 33%
s4_391_ ... 11%
s5_640_ ... 11%
s6_848_up 11%
Defect_Distribut...
s1_belo ... 18%
s2_0_0 16%
s3_0_0 16%
s4_0_0 18%
s5_0_0 16%
s6_0_up 16%
Defect distribution
Escaping_Defects
s1_belo ... 35%
s2_0_0 13%
s3_0_1 13%
s4_1_1 16%
s5_1_2 12%
s6_2_up 12%
Effort_Variance
s1_belo ... 16%
s2_0_0 18%
s3_0_0 17%
s4_0_0 16%
s5_0_0 16%
s6_0_up 16%
Scope_Instability
s1_belo ... 26%
s2_0_0 15%
s3_0_0 15%
s4_0_1 15%
s5_1_2 15%
s6_2_up 15%
Scope instability

Compositional data (CoDa)

- CoDa: vectors of non-negative components showing the *relative importance of a set of* ***parts*** *in a total*.
- Sample space: the simplex

  with k = 1, 100, 10^6, 10^9 (proportions, %, ppm, ppb) .
- The total "k": is considered irrelevant, not informative.



S^3 : Ternary diagram

$$S^D=\{\mathbf{x}\in R_+^D : x_1+x_2+\cdots+x_D = k\}$$

Two principles:

- *scale invariance* and
- *subcompositional coherence*

A four-hour workshop

Conference in honor of Corrado Gini
(Padua, 7-8 September, 2015)

Compositional Data (CoDa) methods in the analysis of customer surveys

M. Vives-Mestres[1], J.A. Martín-Fernández[1], and R.S. Kenett[2]

[1]Universitat de Girona, Spain; *marina.vives@udg.edu*
[2] KPA Group, Israel; University of Turin, Italy; NYU Center for Risk Engineering, USA

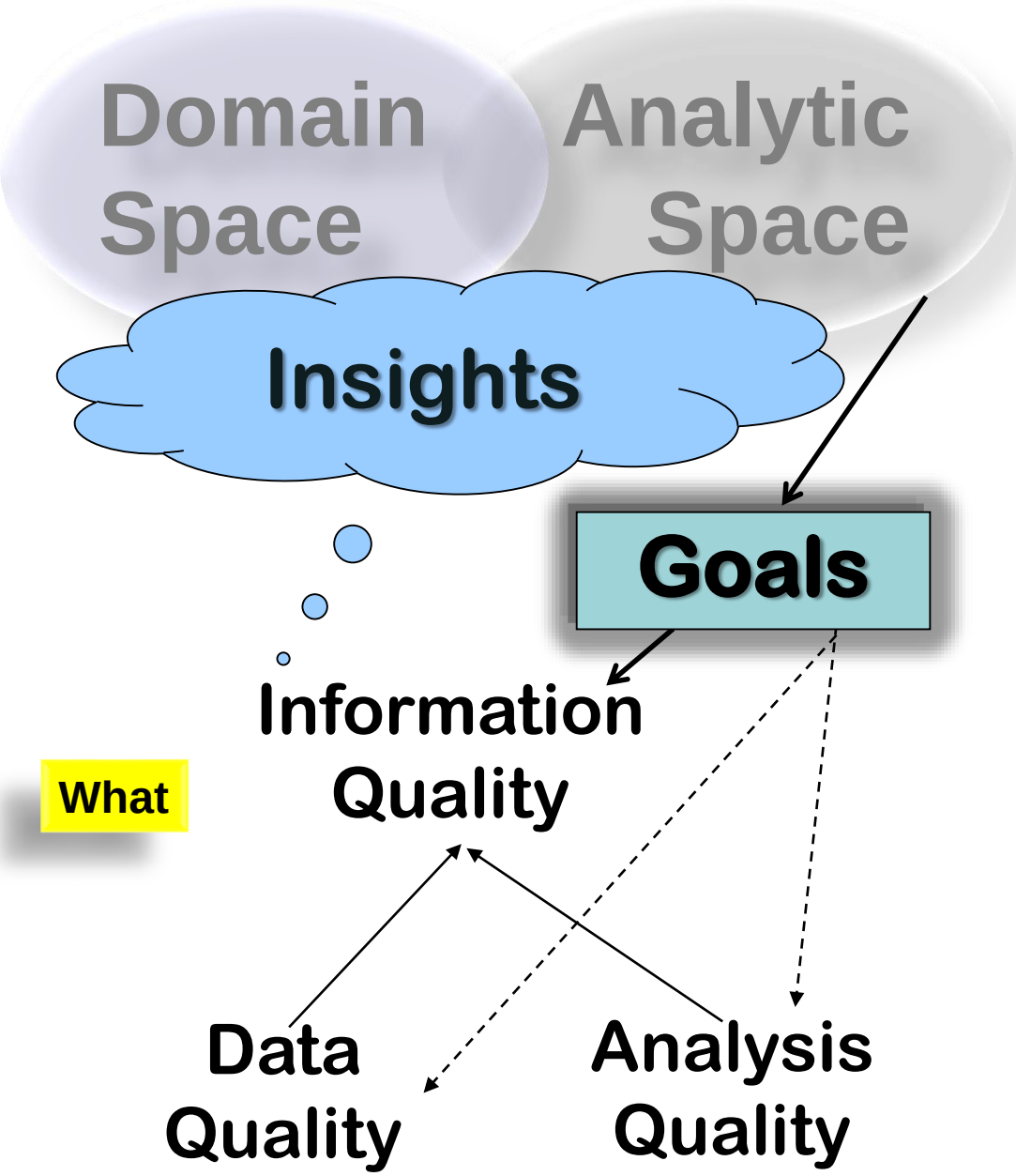
September 8th, 2015
14:30-18:30

http://convegnogini.stat.unipd.it/en/index.php#workshop

$$InfoQ(f, X, g) = U(f(X|g))$$

g	A specific analysis goal
X	The available dataset
f	An empirical method
U	A utility measure

Information Quality (InfoQ)



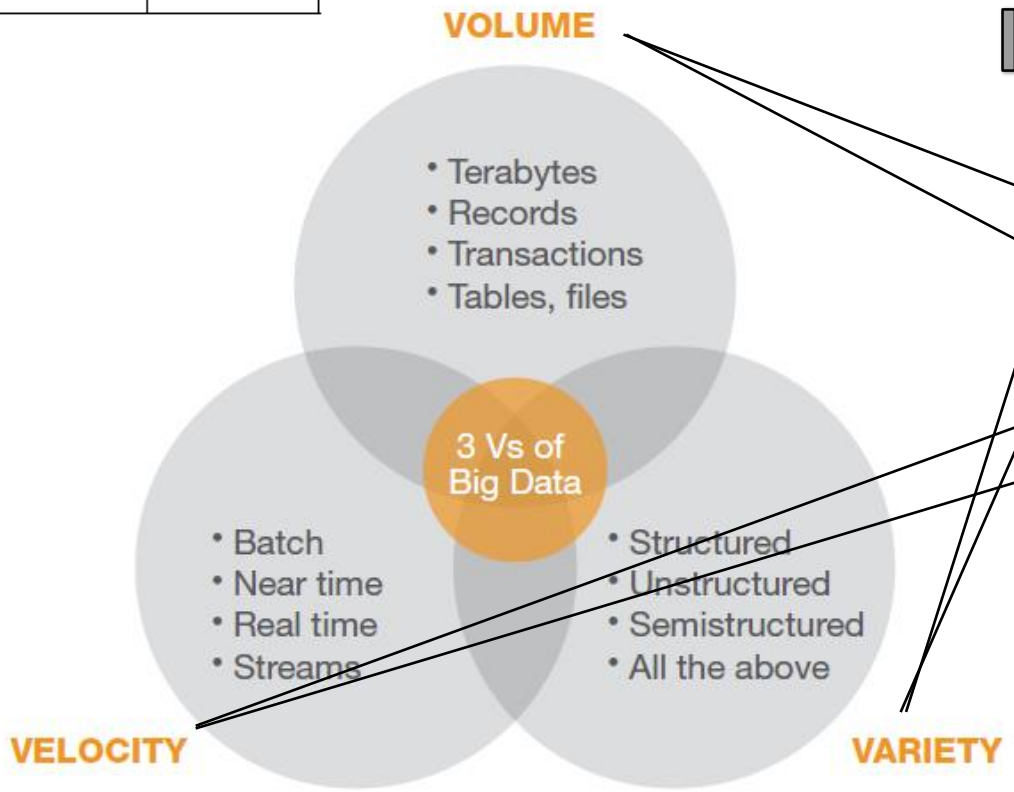
1. Data resolution
2. Data structure
3. Data integration
4. Temporal relevance
5. Chronology of data and goal
6. Generalizability
7. Operationalization
8. Communication

How

#	Dimension	Note	Value	Index
1	Data resolution		5	1.0000
2	Data structure		4	0.7500
3	Data integration		5	1.0000
4	Temporal relevance		5	1.0000
5	Generalizability		3	0.5000
6	Chronology of data and goal		5	1.0000
7	Concept operationalization		2	0.2500
8	Communication		3	0.5000
	InfoQ Score = 0.68			

Power	Prefix
10^{9}	Giga
10^{12}	Tera
10^{15}	Peta
10^{18}	Exa
10^{21}	Zetta
10^{24}	Yotta

Big Data Analytics



InfoQ Dimensions

1. Data resolution
2. Data structure
3. Data integration
4. Temporal relevance
5. Chronology of data and goal
6. Generalizability
7. Operationalization
8. Communication

Russom, P., Big Data Analytics, TDWI Best Practices Report, Q4 2011

Why Bayesian Networks?

InfoQ Dimension	1 Mgmt	2 Web Usability	3 ICT Risks	4 Biotec	5 Surveys	6 Health	7 Testing	Dimension score
Data Resolution	1	4	5	4	5	4	4	**3.50**
Data Structure	2	3	5	4	4	5	4	**3.71**
Data Integration	3	4	4	4	4	5	5	**4.09**
Temporal Relevance	5	4	4	5	4	5	4	**4.40**
Chronology of Data and Goal	4	4	5	5	5	5	4	**4.54**
Generalizability	5	4	4	4	5	5	3	**4.22**
Operationalization	4	3	4	4	4	5	4	**3.96**
Communication	4	4	5	5	5	4	5	**4.54**
Use case InfoQ score	**3.15**	**3.72**	**4.47**	**4.35**	**4.47**	**4.73**	**4.08**	

- Ben Gal, I. (2007). Bayesian Networks, in *Encyclopaedia of Statistics in Quality and Reliability*, Ruggeri, F., Kenett, R. S. and Faltin, F. (editors in chief), Wiley, UK.
- Cornalba, C., Kenett, R.S. and Giudici, P. (2007) Sensitivity Analysis of Bayesian Networks with Stochastic Emulators, ENBIS-DEINDE proceedings, University of Torino, Turin, Italy,
- Cugnata, F., Kenett, R.S., and Salini, S. (2014). Bayesian Network Applications to Customer Surveys and InfoQ, *Procedia Economics and Finance*, 17, 3-9.
- Harel. A., Kenett, R.S. and Ruggeri, F. (2009). Modeling Web Usability Diagnostics on the basis of Usage Statistics, in *Statistical Methods in eCommerce Research*, W. Jank and G. Shmueli (editors), Wiley.
- Kenett, R.S. (2007). Cause and Effect Diagrams, in *Encyclopedia of Statistics in Quality and Reliability*, Ruggeri, F., Kenett, R.S. and Faltin, F. (editors in chief), Wiley, Online version, Wiley InterScience, UK.
- Kenett, R.S., Harel, A. and Ruggeri, F. (2009). Controlling the Usability of Web Services, *International Journal of Software Engineering and Knowledge Engineering*, Vol. 19, No. 5, pp. 627-651.
- Kenett R.S. and Salini S. (2009). New Frontiers: Bayesian networks give insight into survey-data analysis, *Quality Progress*, pp. 31-36, August.
- Kenett, R.S. and Raanan, Y. (2010). *Operational Risk Management: a practical approach to intelligent data analysis*, Wiley and Sons. http://www.wiley.com/WileyCDA/WileyTitle/productCd-047074748X.html
- Kenett, R.S. and Baker, E. (2010). *Process Improvement and CMMI for Systems and Software,* Taylor and Francis, Auerbach CRC Publications
- Kenett R.S. and Salini S. (2012). *Modern Analysis of Customer Satisfaction Surveys: with applications using R*, John Wiley and Sons, Chichester: UK, .http://www.wiley.com/WileyCDA/WileyTitle/productCd-0470971282.html
- Kenett, R.S. and Salini, S. (2012). Modern Analysis of Customer Surveys: comparison of models and integrated analysis", with discussion*, Applied Stochastic Models in Business and Industry*, 27, pp. 465–475.
- Kenett, R.S. (2012). Risk Analysis in Drug Manufacturing and Healthcare, in *Statistical Methods in Healthcare*, Faltin, F., Kenett, R.S. and Ruggeri, F. (editors in chief), John Wiley and Sons.
- Kenett, R. S., Shmueli, G.. On information quality. *Journal of the Royal Statistical Society - Series A: Statistics in Society*, 177(1):3–38, 2014.
- Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*, 2nd ed., Cambridge University Press, UK.
- Peterson, J. and Kenett, R.S. (2011), Modelling Opportunities for Statisticians Supporting Quality by Design Efforts for Pharmaceutical Development and Manufacturing, *Biopharmaceutical Report*, ASA Publications, Vol. 18, No. 2, pp. 6-16.
- Salini, S. and Kenett, R.S. (2009). Bayesian Networks of Customer Satisfaction Survey Data, *Journal of Applied Statistics*. 36(11) pp. 1177-1189.