Session 2:
Bayesian Networks

Marek J. Drużdżel

Wydział Informatyki

Politechnika Białostocka

m.druzdzel@pb.edu.pl

http://aragorn.wi.pb.bialystok.pl/~druzdzel/

Session overview

- Bayesian graphical models
- Inference in Bayesian networks
- Extended family of graphical models

What I want you to know after this session?

- Know how to express probabilistic knowledge by means of Bayesian networks.
- Understand reasoning with Bayesian networks.
- Understand the relationship between probability and causality.

Bayesian Graphical Models

Systems and models

Systems - pieces of the real world that can reasonably be studied in separation from the rest of the world

Models - (subjective) abstractions of systems, used in science or everyday thinking

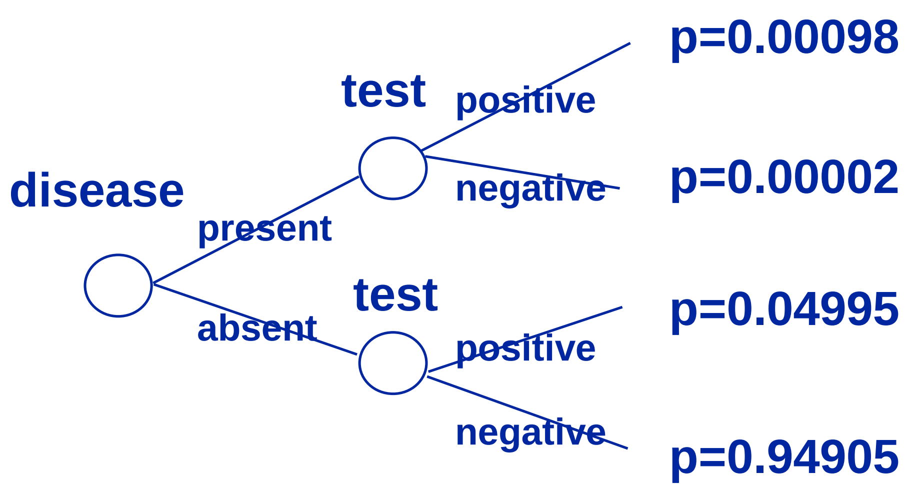
Bayesian networks are models ☺.

Probabilistic knowledge representations

- A probabilistic (Bayesian) model encodes the *joint probability distribution* over its variables.
- Knowledge of the joint probability distribution is sufficient to derive any marginal and conditional probability over the model's variables (and anything else we could possibly be interested in!).

Probability trees

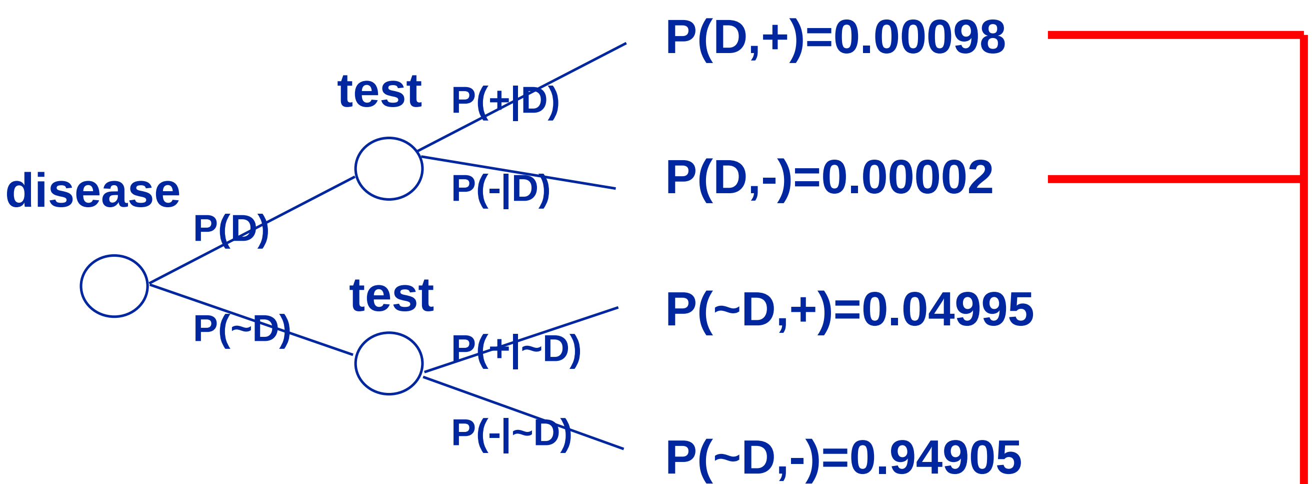
The simplest and quite natural graphical representation of a joint probability distribution over discrete variables



P(disease present $\wedge$ test positive) = P(D $\cap$ +) = 0.00098
P(disease present $\wedge$ test negative) = P(D $\cap$ -) = 0.00002
P(disease absent $\wedge$ test positive) = P(~D $\cap$ +) = 0.04995
P(disease absent $\wedge$ test negative) = P(~D $\cap$ -) = 0.94905

Computation in probability trees

We can calculate any marginal or conditional probability distribution from the joint probability distribution encoded in the tree.

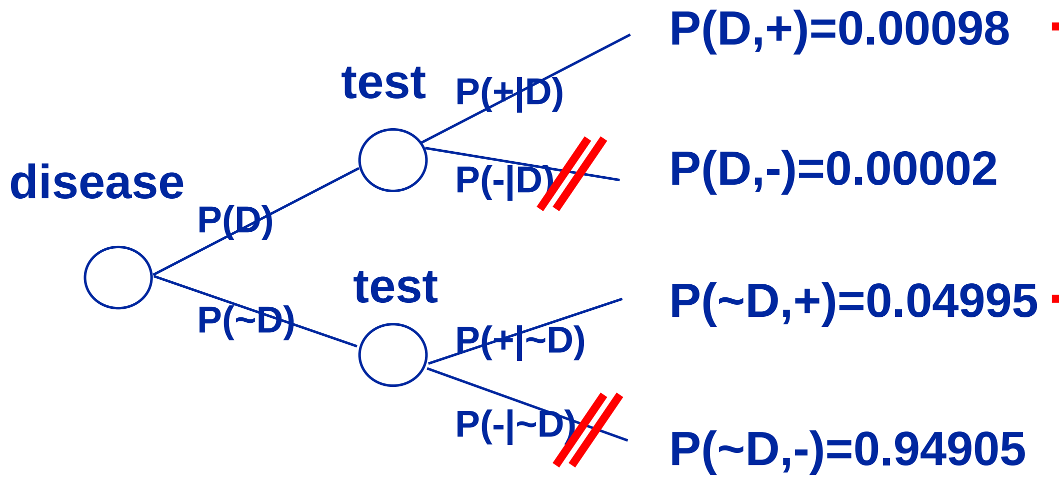


What is the probability of the disease present?

P(D) = 0.00098+0.00002 = 0.001

Computation in probability trees

The simplest and quite natural graphical representation of a joint probability distribution over discrete variables

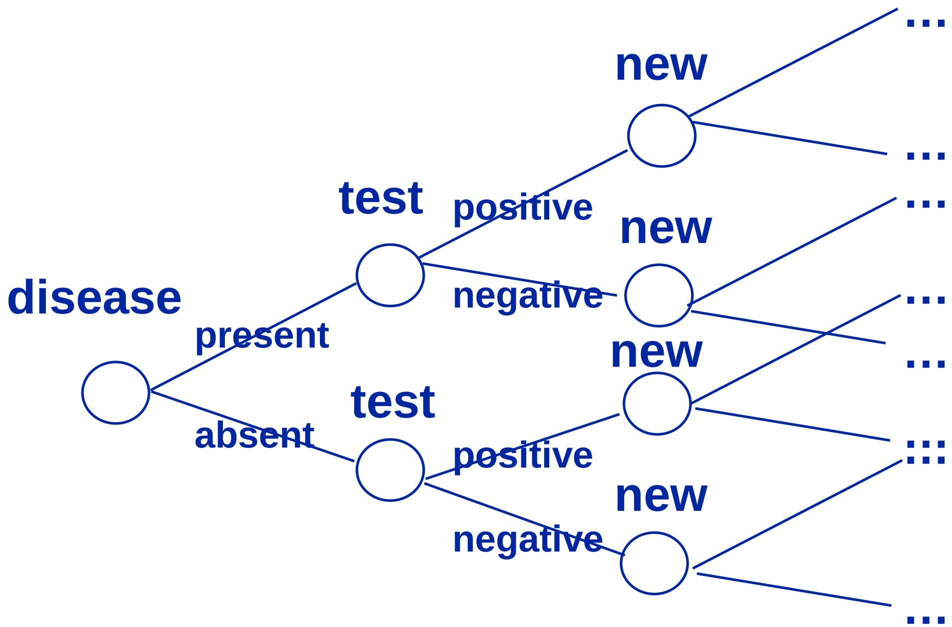


What is the probability of the disease present given a positive test result? Observation of a positive test result makes some of the branches of the tree impossible. What we need to do is just renormalize the remaining, possible (i.e., those that are compatible with the evidence) branches!

$$P(D|+) = 0.00098/(0.00098+0.04995) \approx 0.01924$$

What is wrong with probability trees?

Trees grow exponentially with the number of variables



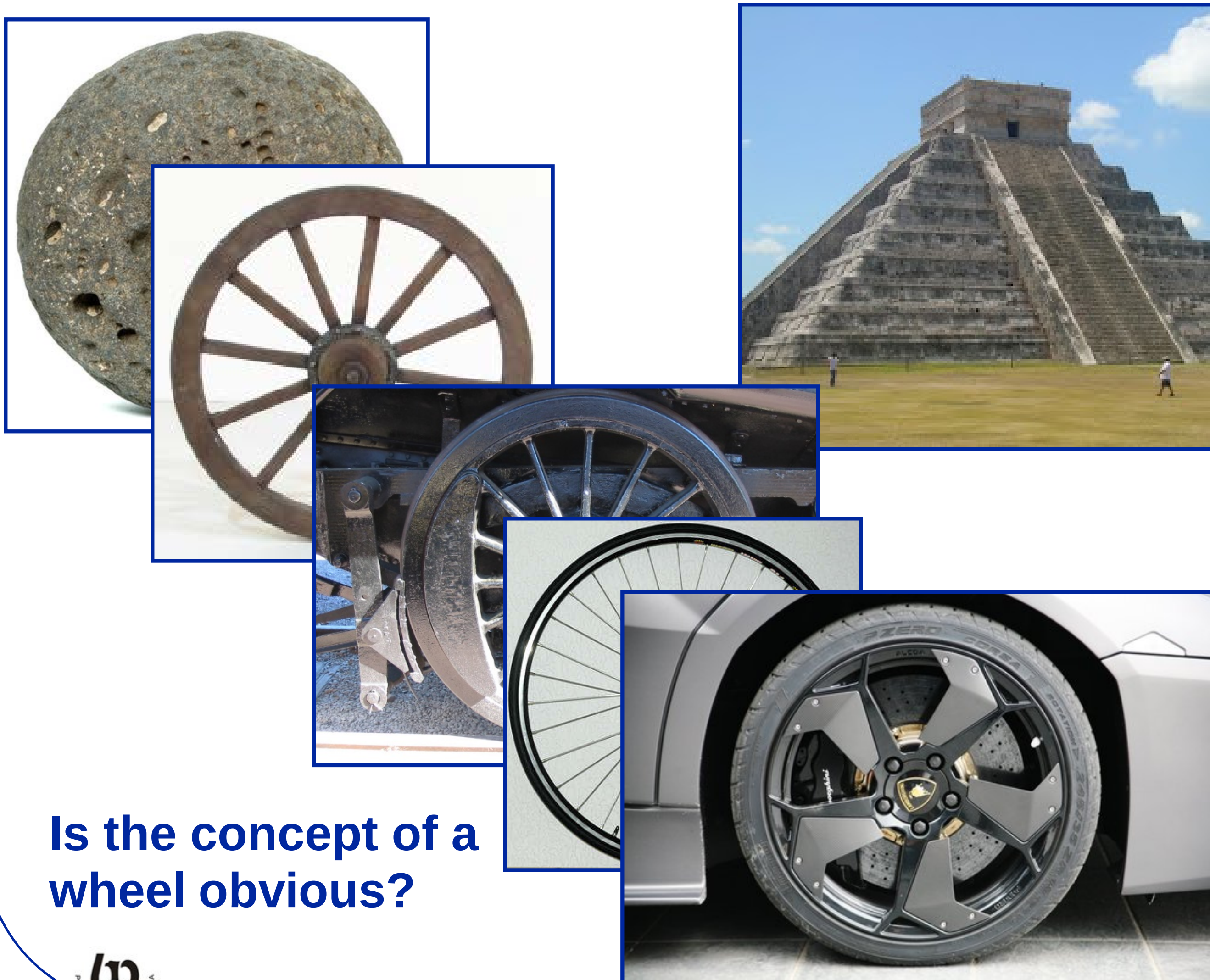
For n binary variables, we will have 2^n branches.
When n=10, the total number of branches is 1,024
When n=11, it is 2,048
...
When n=20, it is 1,048,576 (which is a lot ☺)

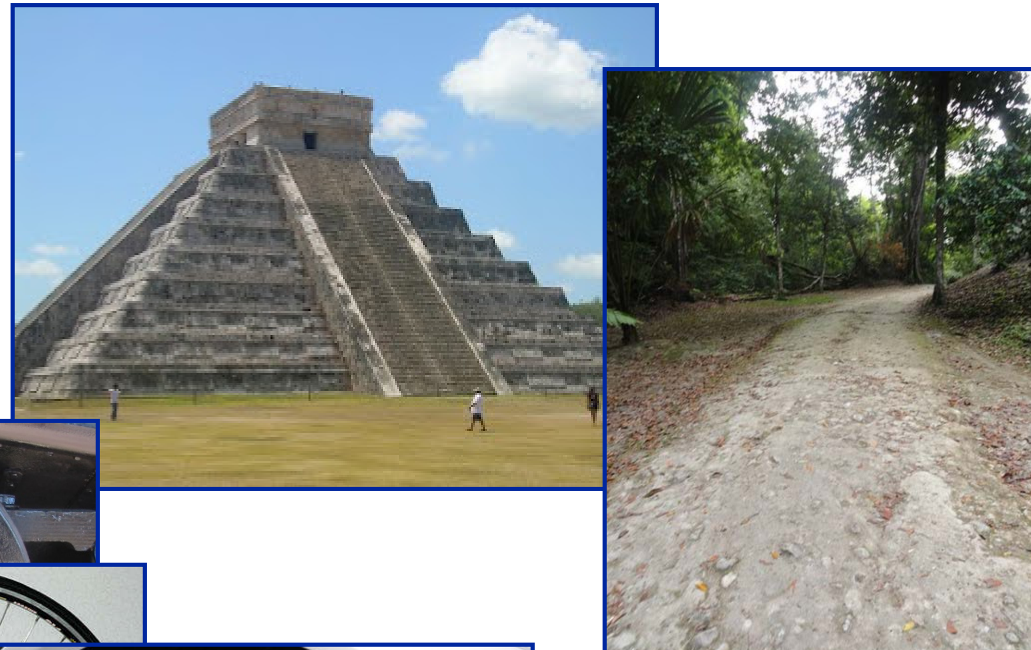
Great idea (only 30-40 years old)

Use independences among variables in the joint probability distribution to reduce the number of parameters in its representation!

Due to seminal work on probabilistic independence by A. Philip Dawid and Judea Pearl

All brilliant ideas are obvious (once we have them ☺)





Then why none of the civilizations in the Americas had it?

Is the concept of a wheel obvious?

Factorability of the joint probability distribution

Every joint probability distribution can be factorized, i.e., rewritten as a product of prior and conditional probability distributions of each of the model's variables

$$f(X_1, X_2, \ldots, X_n) = f(X_1 \mid X_2, X_3, \ldots, X_n)\, f(X_2 \mid X_3, \ldots, X_n) \ldots f(X_{n-2} \mid X_{n-1}, X_n)\; f(X_{n-1} \mid X_n)\, f(X_n)$$

e.g., four variables (a, b, c, d), we have:

$$P(A,B,C,D)=P(A|B,C,D)\ P(B|C,D)\ P(C|D)\ P(D)$$
$$P(A,B,C,D)=P(A|B,C,D)\ P(B|C,D)\ P(D|C)\ P(C)$$
...
$$P(A,B,C,D)=P(B|A,C,D)\ P(D|A,C)\ P(A|C)\ P(C)$$
...

There are n! different directed graphs corresponding to various ways of factorizing a joint probability distribution over n variables.

For n=4, we have 4!=24 different factorizations.

Factorability of the joint probability distribution

- **Any factorization can be simplified if we consider independencies among variables.**
- **Those factorizations that become the simplest are better than others in terms of efficiency of representation.**

e.g., suppose we know that $B \perp D | C$, $D \perp A | C$, and $A \perp C$

We can simplify

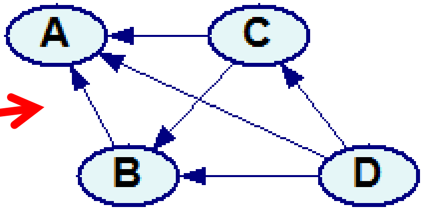
$$P(A,B,C,D)=P(B|A,C,D)\, P(D|A,C)\, P(A|C)\, P(C)$$

into

$$P(A,B,C,D)=P(B|A,C)\, P(D|C)\, P(A)\, P(C)$$

Bayesian networks

- This underlies the very idea of Bayesian networks.
- We draw a directed graph with arc from the conditioning variables to the variables in the factorization.

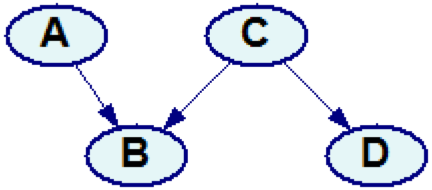


P(A,B,C,D)=P(A|B,C,D) P(B|C,D) P(C|D) P(D)
P(A,B,C,D)=P(A|B,C,D) P(B|C,D) P(D|C) P(C)
...
P(A,B,C,D)=P(B|A,C,D) P(D|A,C) P(A|C) P
...

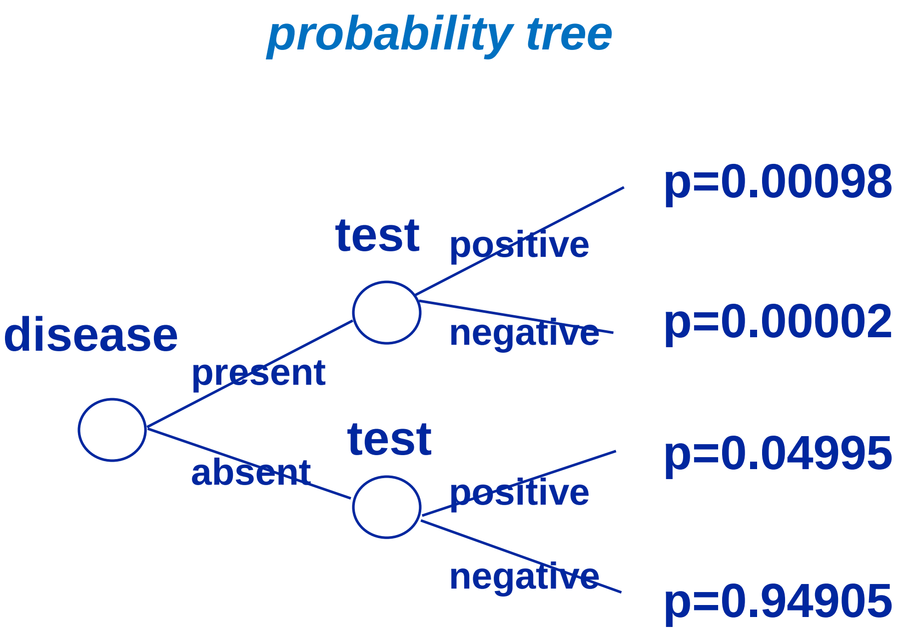
B⊥D|C, D⊥A|C, A⊥C

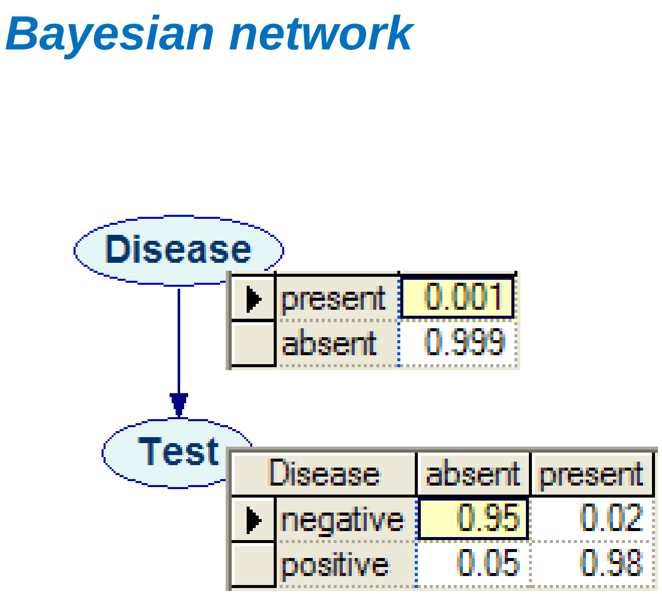
P(A,B,C,D)=P(B|A,C) P(D|C) P(A) P(C)

Absence of an arc is a graphical representation of independence!



Probability trees and Bayesian networks

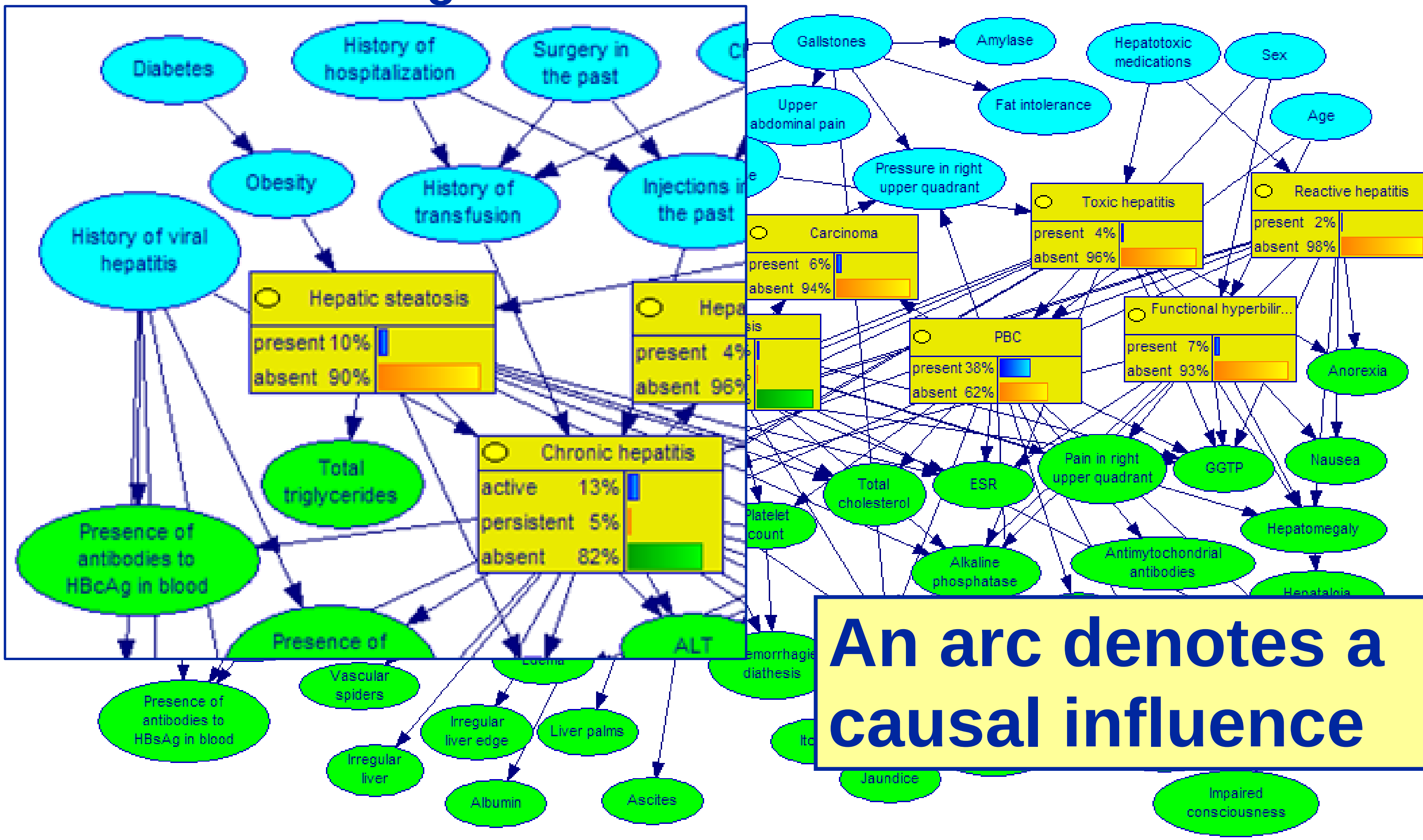




The two representations are equivalent
But, when there are independences in the domain,
Bayesian networks are much, much more efficient!

Bayesian networks: An alternative view (quite consistent with the theoretical view, it turns out!

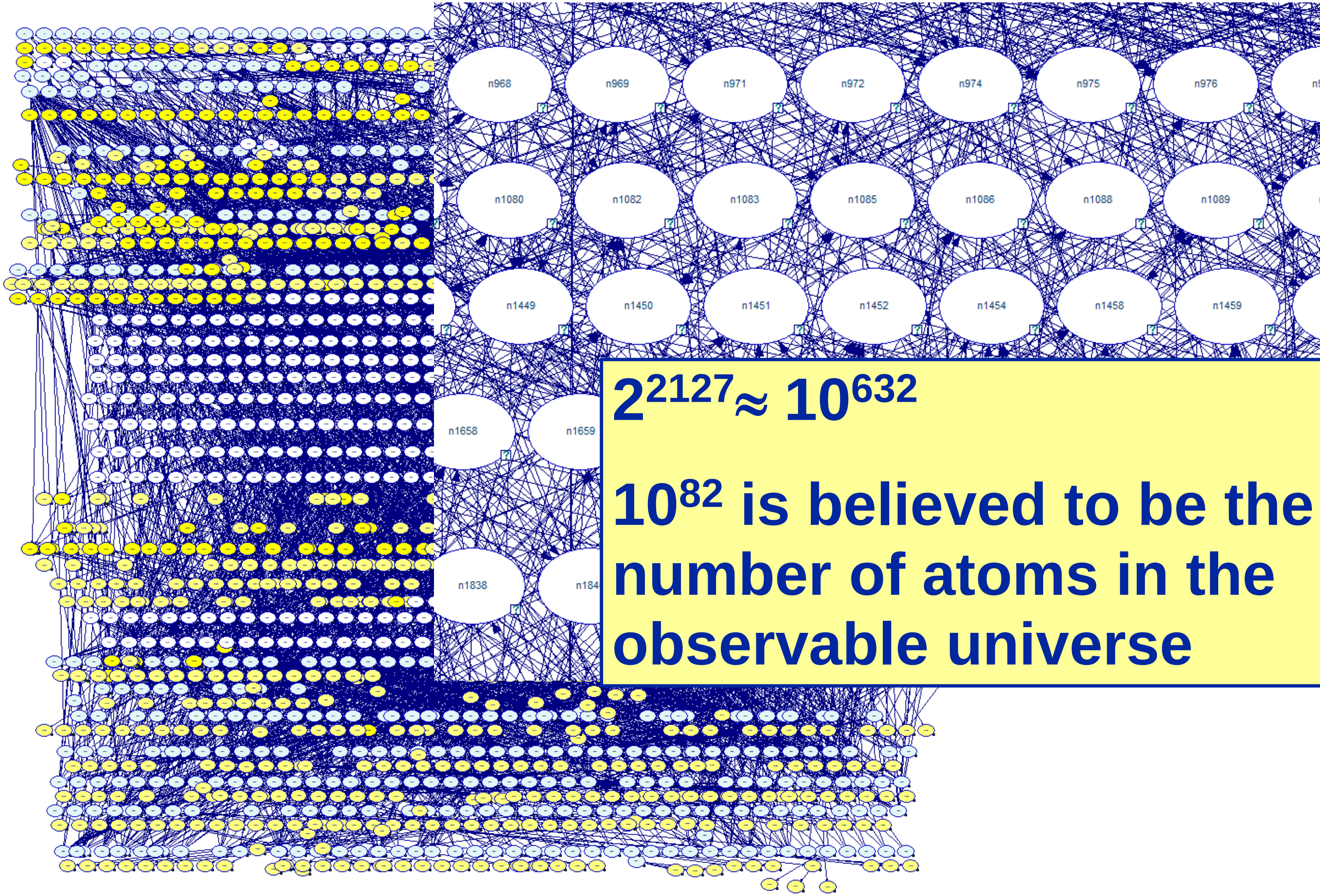
The graphical part of a Bayesian network is a picture of causal relations among the model variables.



An arc denotes a causal influence

[Oniśko et al.] 70 variables, 123 arcs, 2,415 independences, 2,139 numerical parameters (instead of over $2^{70} \approx 10^{21}$!)

Diagnosis of Diesel locomotives



$2^{2127} \approx 10^{632}$

10^{82} is believed to be the number of atoms in the observable universe

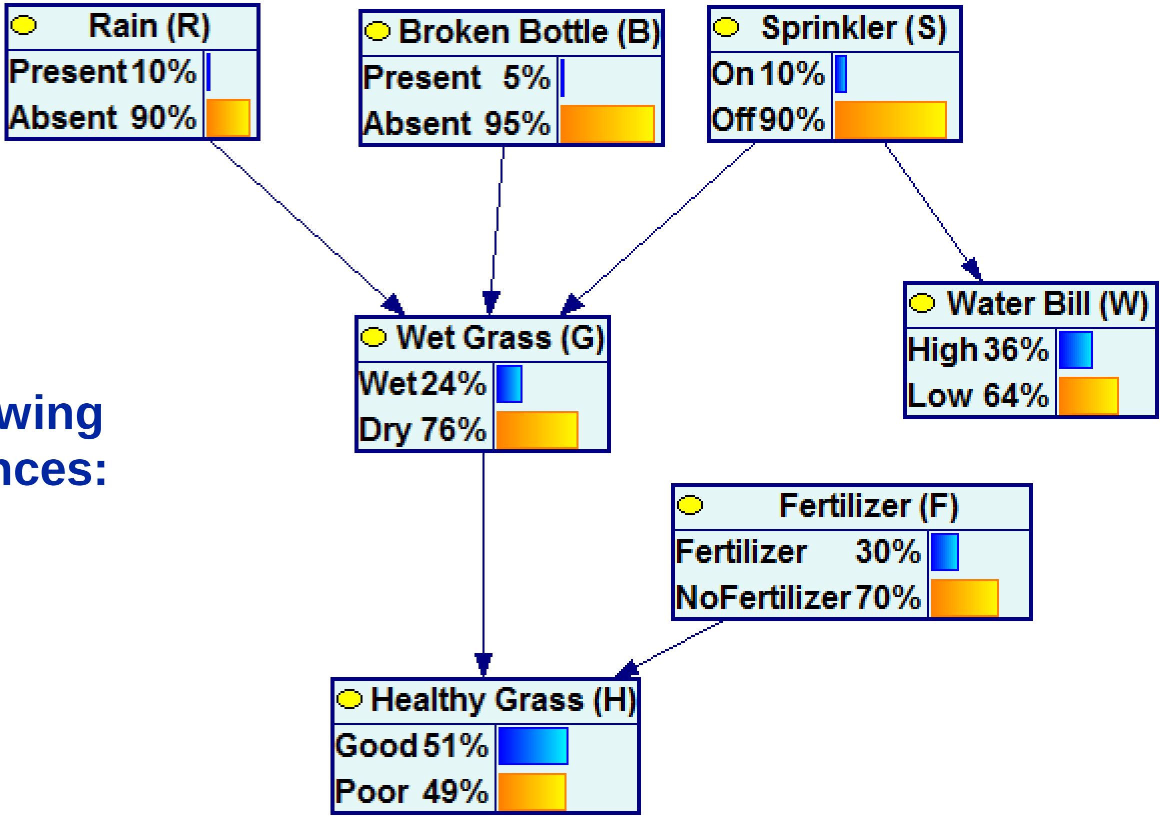
[Przytula et al.] 2,127 variables, 3,595 arcs, 2,261,001 independences, 12,351 numerical parameters (instead of $2^{2,127} \approx 10^{632}$!)

Independences: Markov condition

- Allows to read back dependences and independences from the graph.
- Informally speaking, it is an assumption that ties directed probabilistic graphs with probability, specifying how a directed graphs represents independence.
- A node is independent of its non-descendants given its predecessors.

Markov condition: Example

$$P(H,G,W,R,B,S,F)=P(H|G,F)\,P(G|R,B,S)\,P(W|S)\,P(R)\,P(B)\,P(S)\,P(F)$$



This graph implies the following (conditional) independences:

$R\perp B$, $R\perp S$, $B\perp S$, $R\perp F$, $B\perp F$, $S\perp F$

$R\perp W$, $B\perp W$, $W\perp F$, $G\perp F$

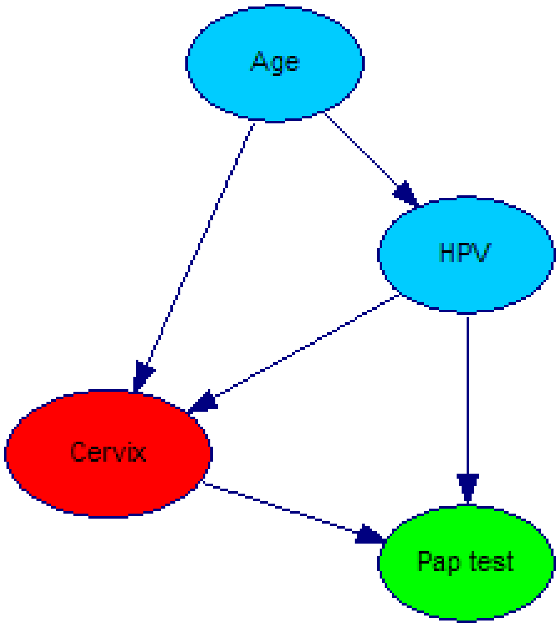
$R\perp H|G$, $B\perp H|G$, $S\perp H|G$, $W\perp H|G$

$W\perp *|S$

$R\perp W|G,S$, $B\perp W|G,S$

Bayesian networks

A **Bayesian network** [Pearl 1988] is an acyclic directed graph consisting of:

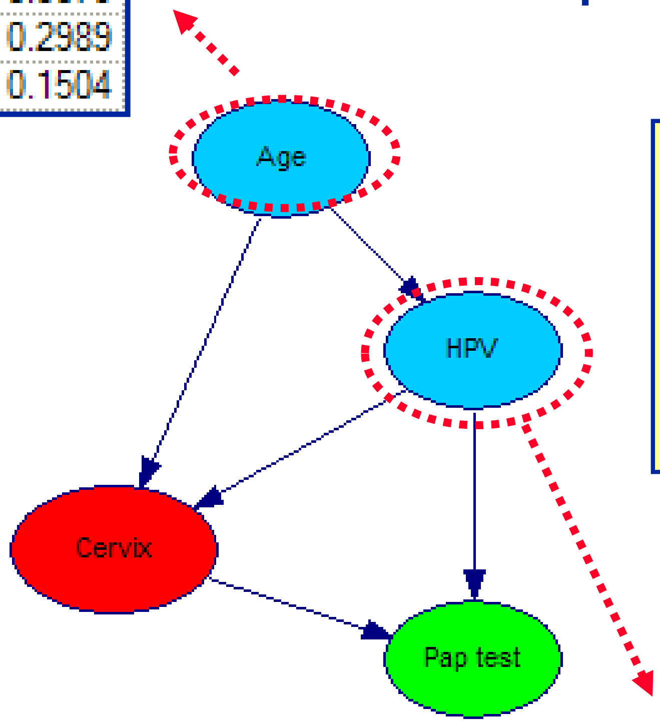


- The **qualitative part**, encoding a domain's variables (nodes) and the probabilistic (usually causal) influences among them (arcs).
- The **quantitative part**, encoding the joint probability distribution over these variables.

Bayesian networks: Numerical parameters

a1_below_20	0.0416
a2_20_29	0.2012
a3_29_45	0.3079
a4_45_60	0.2989
a5_60_up	0.1504

Prior probability distribution tables for nodes without predecessors (Age)



Please note that each absence of an arc (i.e., each independence modeled) is means one less dimension in the corresponding conditional probability table!

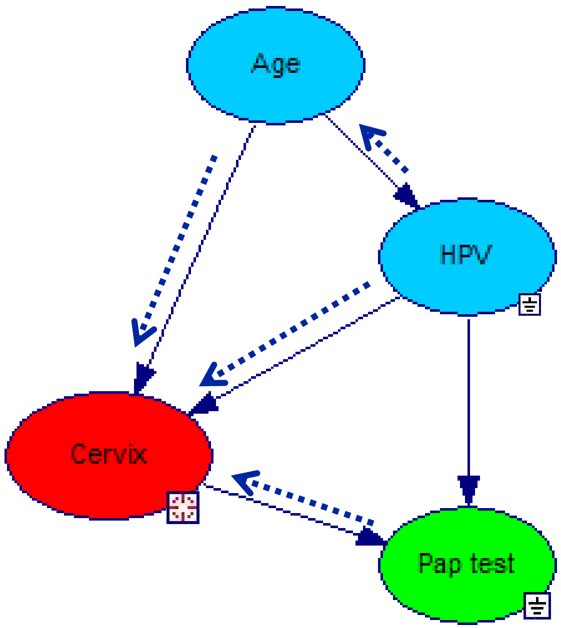
Conditional probability distributions tables for nodes with predecessors (HPV, Pap test, Cervix)

Age	a1_below_20	a2_20_29	a3_29_45	a4_45_60	a5_60_up
NA	0.8652	0.8387	0.7904	0.8055	0.8851
Negative	0.069	0.0901	0.1782	0.1765	0.1012
Positive	0.0613	0.0667	0.0282	0.0142	0.0082
Qns	0.0045	0.0045	0.0032	0.0038	0.0055

Inference in Bayesian Networks

Reasoning in Bayesian networks: Bayesian updating

The most important type of reasoning in Bayesian networks is updating the probability of a hypothesis (e.g., a diagnosis) given new evidence (e.g., medical findings, test results).



P(CxCa | HPV=positive, HSIL=yes)

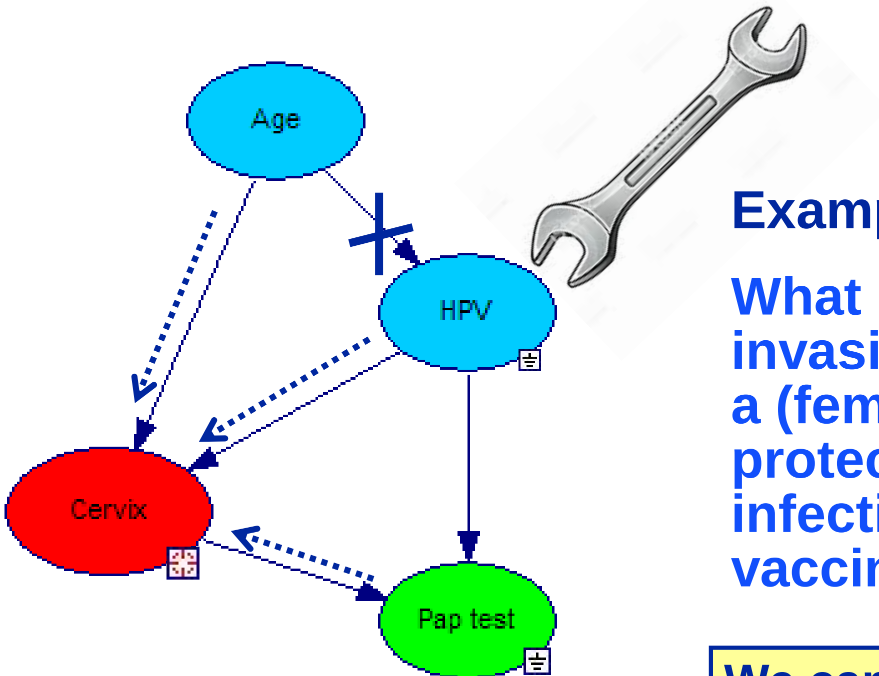
Example:

What is the probability of invasive cervical cancer in a (female) patient with high grade dysplasia with a history of HPV infection?

Generally, the more sparse the structure of your network, the fewer parameters, the faster inference in the Bayesian network.

Reasoning in Bayesian networks: Changes in structure

Changes in structure is an economic/econometric terms used for predicting the effects of manipulation of a modeled system



Example:

What is the probability of invasive cervical cancer in a (female) patient protected from an HPV infection by a (perfect) vaccine?

P(CxCa | HPV=negative, HSIL=yes)

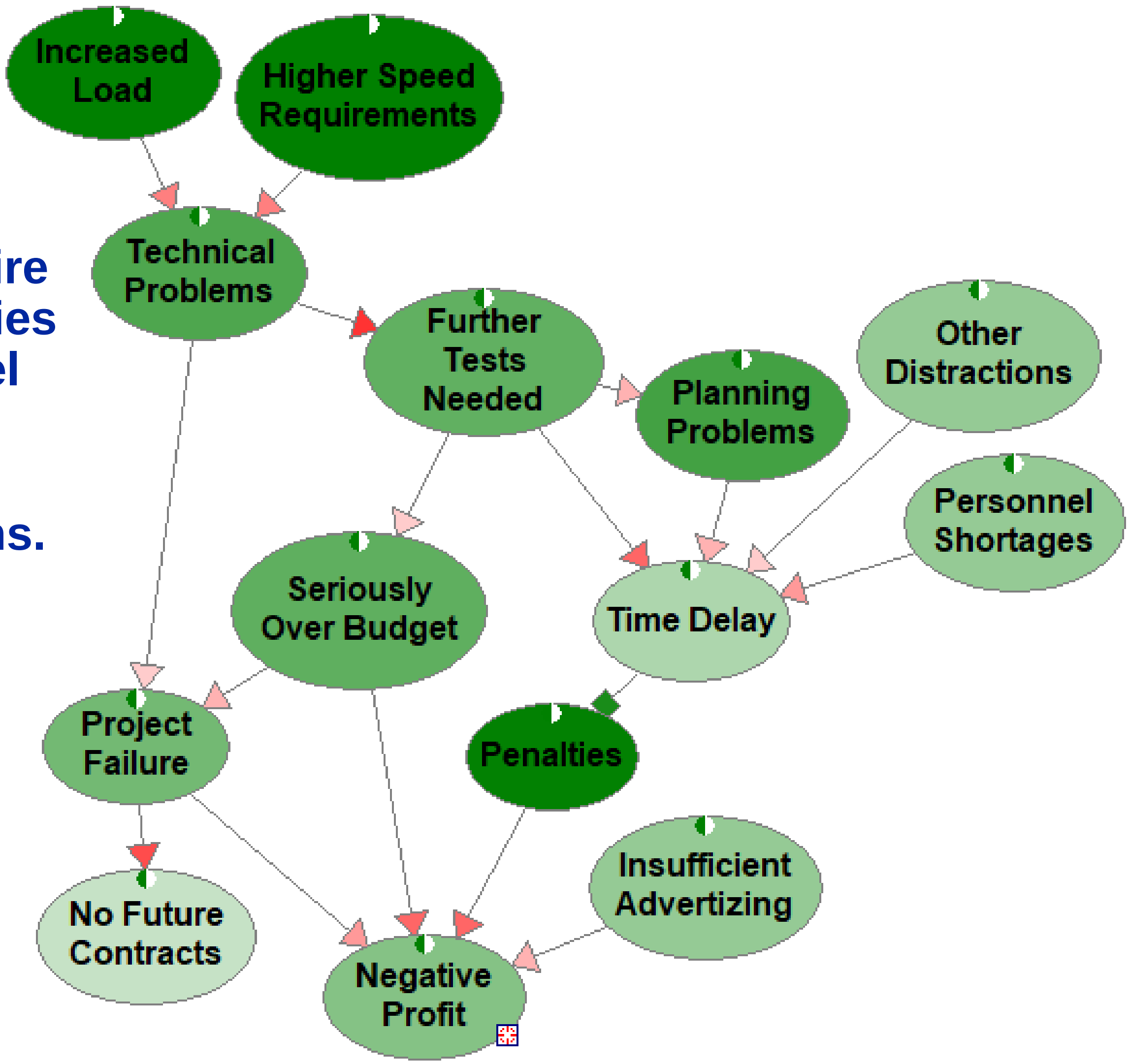
We can calculate the effects of changes in structure only if we have a causal model of the system

Extended Family of Bayesian Graphical Models

Qualitative Bayesian networks

Qualitative interface to Bayesian networks require few numerical probabilities and allow for rapid model building and analysis.

They are great for group decision making sessions.



Equation-based systems and graphical models

classsize = (*nstud* * *cload*) / (*nfac* * *tload*)

facsal = (*oinc* + *tuition* * *nstud*) / (*nfac* * (1 + *overh*))

stratio = *nstud* / *nfac*

Core equations

cload = 15

tload = 6

nstud = 22102

nfac = 3006

oinc = 30000000

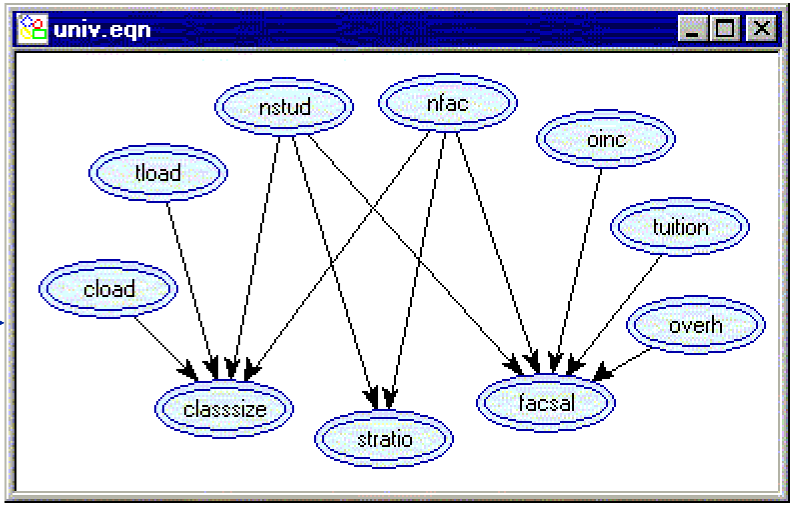
tuition = 12000

overh = 0.48

Equations for exogenous variables

Together they determine the structure of the model

Explication of the asymmetries due to Herb Simon (early 1950s)



Equation-based systems: Reversibility of causal ordering

classsize = (*nstud* * *cload*) / (*nfac* * *tload*)

facsal = (*oinc* + *tuition* * *nstud*) / (*nfac* * (1 + *overh*))

stratio = *nstud* / *nfac*

cload = 15

tload = 6

nstud = 22102

~~*nfac* = 3006~~ stratio = 10

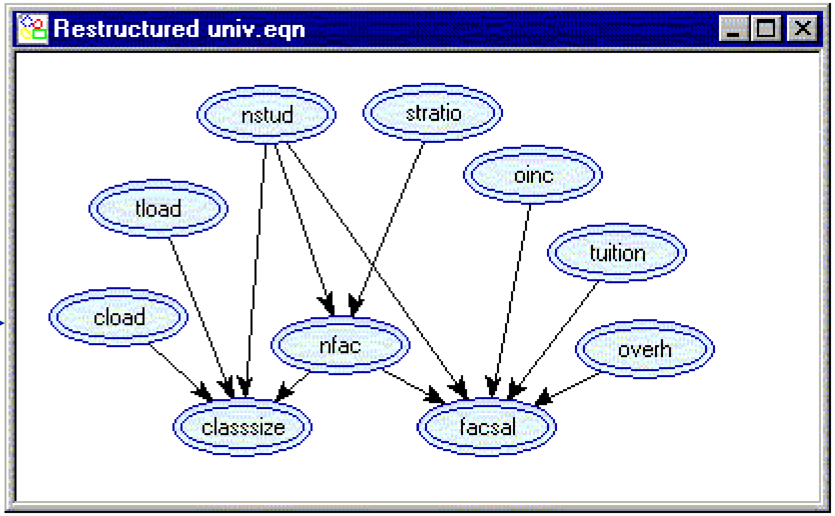
oinc = 30000000

tuition = 12000

overh = 0.48

Setting *stratio* to be exogenous at the expense of *nfac*

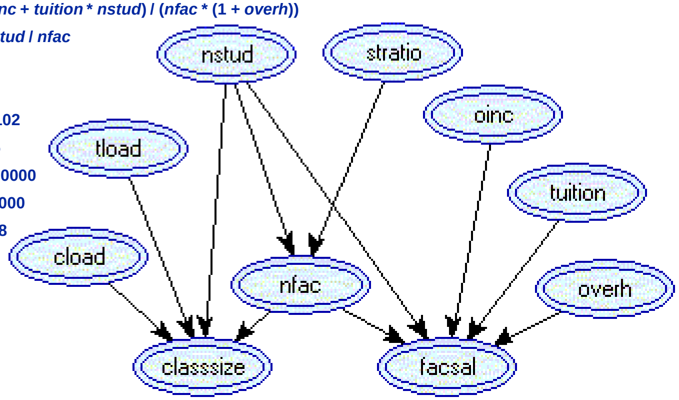
The new model structure

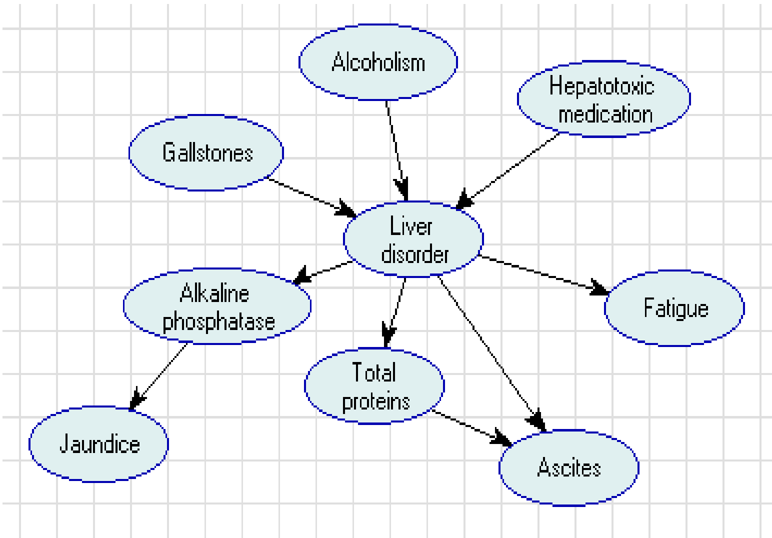


Family of directed graphs (a bigger picture)

(a.k.a. “influence nets,” “causal diagrams,” etc.)

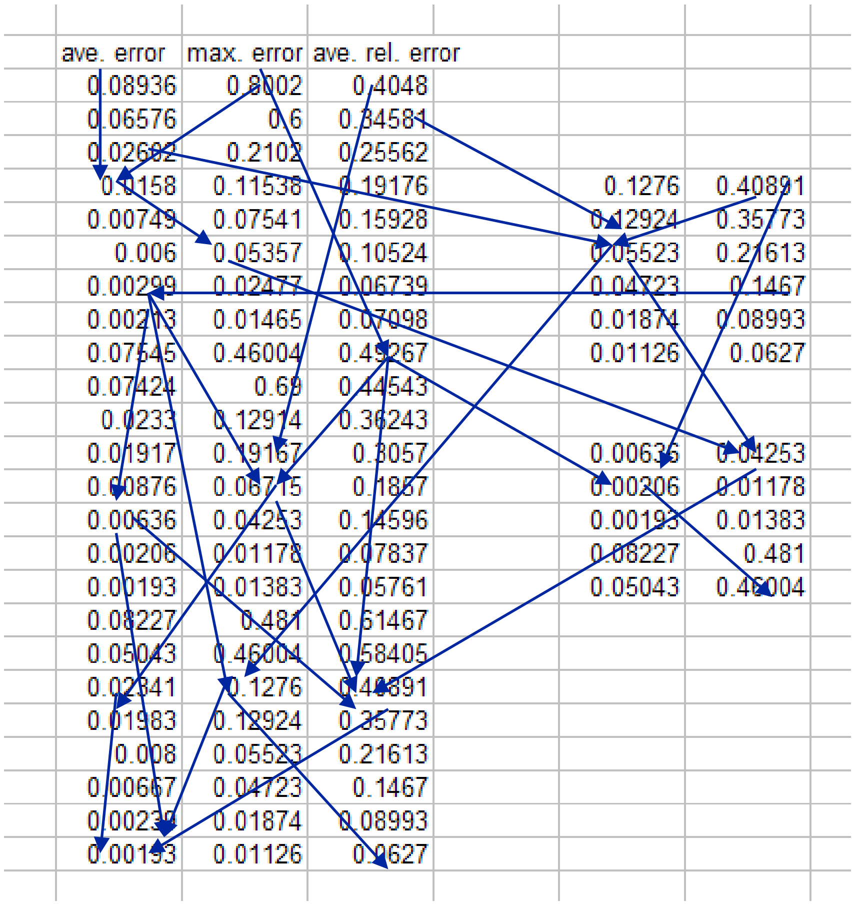
*classsize = (nstud * cload) / (nfac * tload)*
*facsal = (oinc + tuition * nstud) / (nfac * (1 + overh))*
stratio = nstud / nfac
cload = 15
tload = 6
nstud = 22102
nfac = 3006
oinc = 30000000
tuition = 12000
overh = 0.48





Both, systems of equations and joint probability distributions can be pictured by acyclic directed graphs.

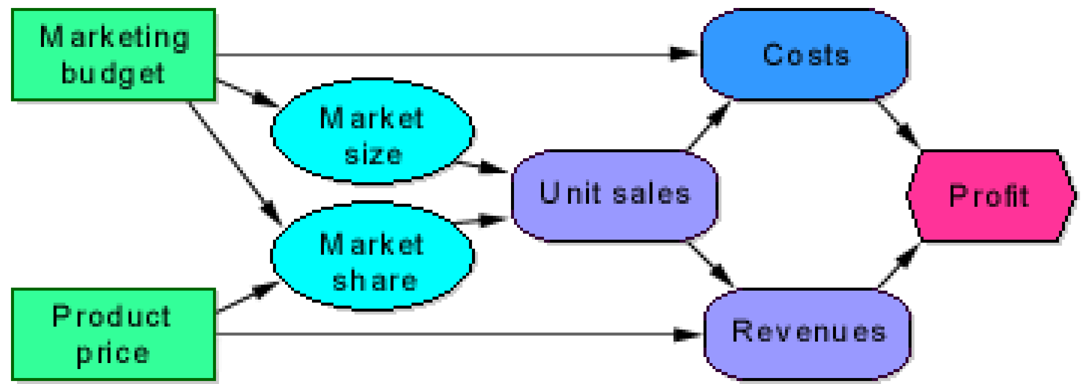
Spreadsheet models



ave. error	max. error	ave. rel. error		
0.08936	0.8002	0.4048		
0.06576	0.6	0.34581		
0.02682	0.2102	0.25562		
0.0158	0.11538	0.19176	0.1276	0.40891
0.00749	0.07541	0.15928	0.12924	0.35773
0.006	0.05357	0.10524	0.05523	0.21613
0.00299	0.02477	0.06739	0.04723	0.1467
0.00213	0.01465	0.07098	0.01874	0.08993
0.07545	0.46004	0.49267	0.01126	0.0627
0.07424	0.69	0.44543		
0.0233	0.12914	0.36243		
0.01917	0.19167	0.3057	0.00636	0.04253
0.00876	0.06715	0.1857	0.00206	0.01178
0.00636	0.04253	0.14596	0.00193	0.01383
0.00206	0.01178	0.07837	0.08227	0.481
0.00193	0.01383	0.05761	0.05043	0.46004
0.08227	0.481	0.51467		
0.05043	0.46004	0.58405		
0.02341	0.1276	0.40891		
0.01983	0.12924	0.35773		
0.008	0.05523	0.21613		
0.00667	0.04723	0.1467		
0.00239	0.01874	0.08993		
0.00193	0.01126	0.0627		

- **They could also be viewed as graphs**
- **Graphs would show causal dependences among cells (variables)**
- **Of course, for any practical spreadsheet, we would essentially get a spaghetti of connections ☺**
- **Systems of simultaneous equations and spreadsheet models are not the best we can do**
- **Directed graphs seem to be better as a user interface!**

Visual spreadsheets



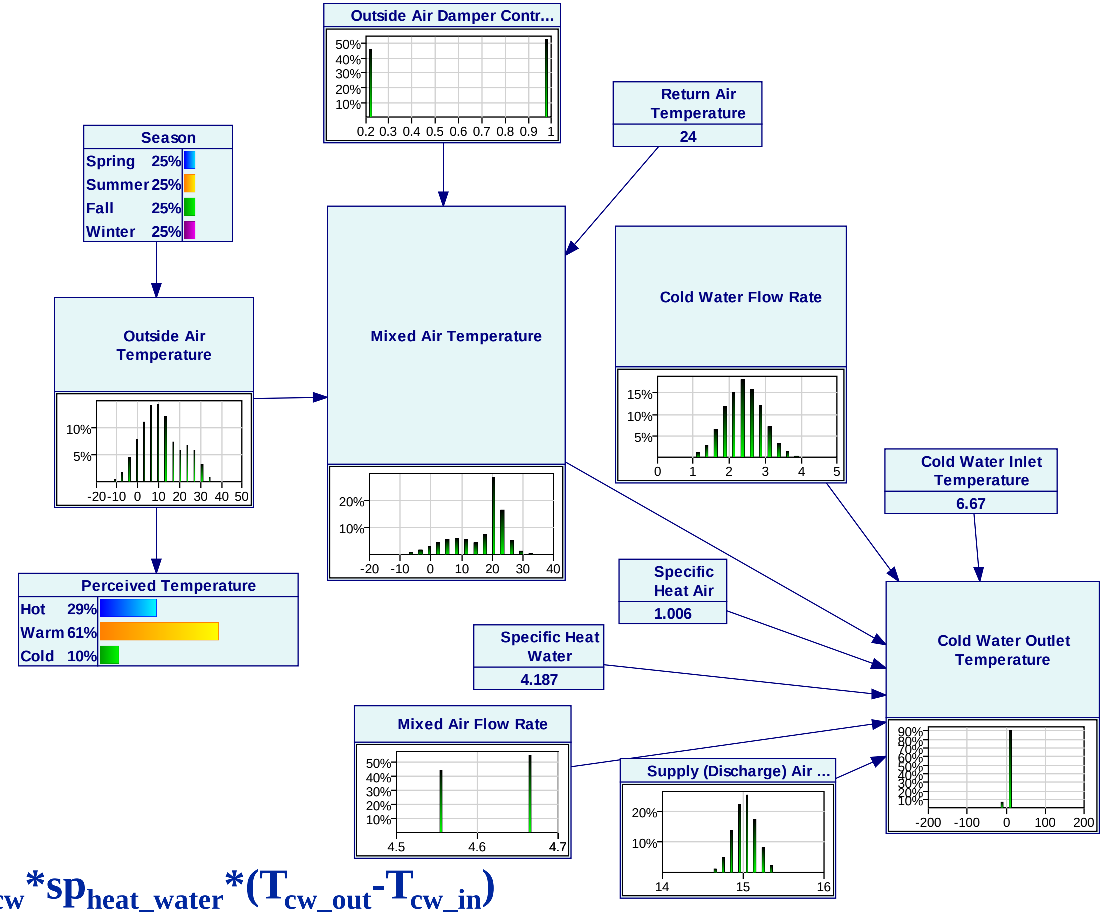
- **Fix almost everything that has been wrong with spreadsheets**
- **Great, but I believe that they could still be improved upon!**

My favorite is Analytica (http://www.lumina.com/)

Example of a simultaneous structural equation-based model turned into a Bayesian network

A model of heating and cooling of buildings.

Two core equations, continuous variables/distributions.



Equations relating temperatures before and after the damper:

$$T_{ma} = T_{oa} * u_d + T_{ra} * (1-u_d)$$

If there is only cooling (u_{hc}=0)

$$m_{flow\_ma} * sp_{heat\_air} * (T_{sa} - T_{ma}) = mdot_{cw} * sp_{heat\_water} * (T_{cw\_out} - T_{cw\_in})$$

and if there is only heating (u_{cc}=0)

$$m_{flow\_ma} * sp_{heat\_air} * (T_{sa} - T_{ma}) = mdot_{hw} * sp_{heat\_water} * (T_{hw\_out} - T_{hw\_in})$$

Advantages of directed graphs

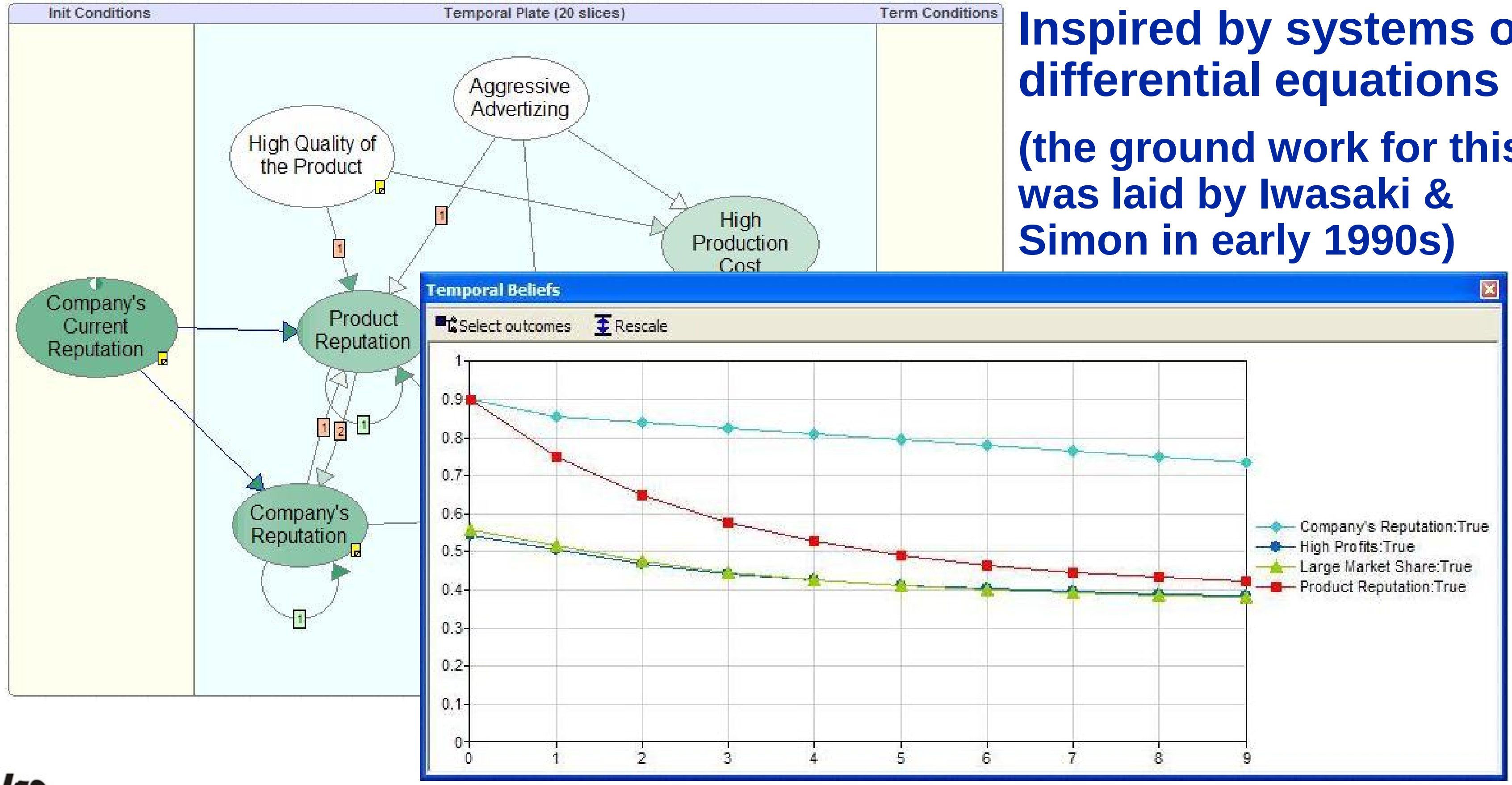
- May be built to reflect the causal structure of a model (helps with obtaining insight into the problem)
- Can accommodate representation of uncertainty
- Can be reconfigured as needed
- Have sound theoretical foundations: We are dealing here with probability theory and decision theory
- We can talk (almost) the same language with statisticians, philosophers, and scientists

Temporal reasoning: Dynamic Bayesian networks

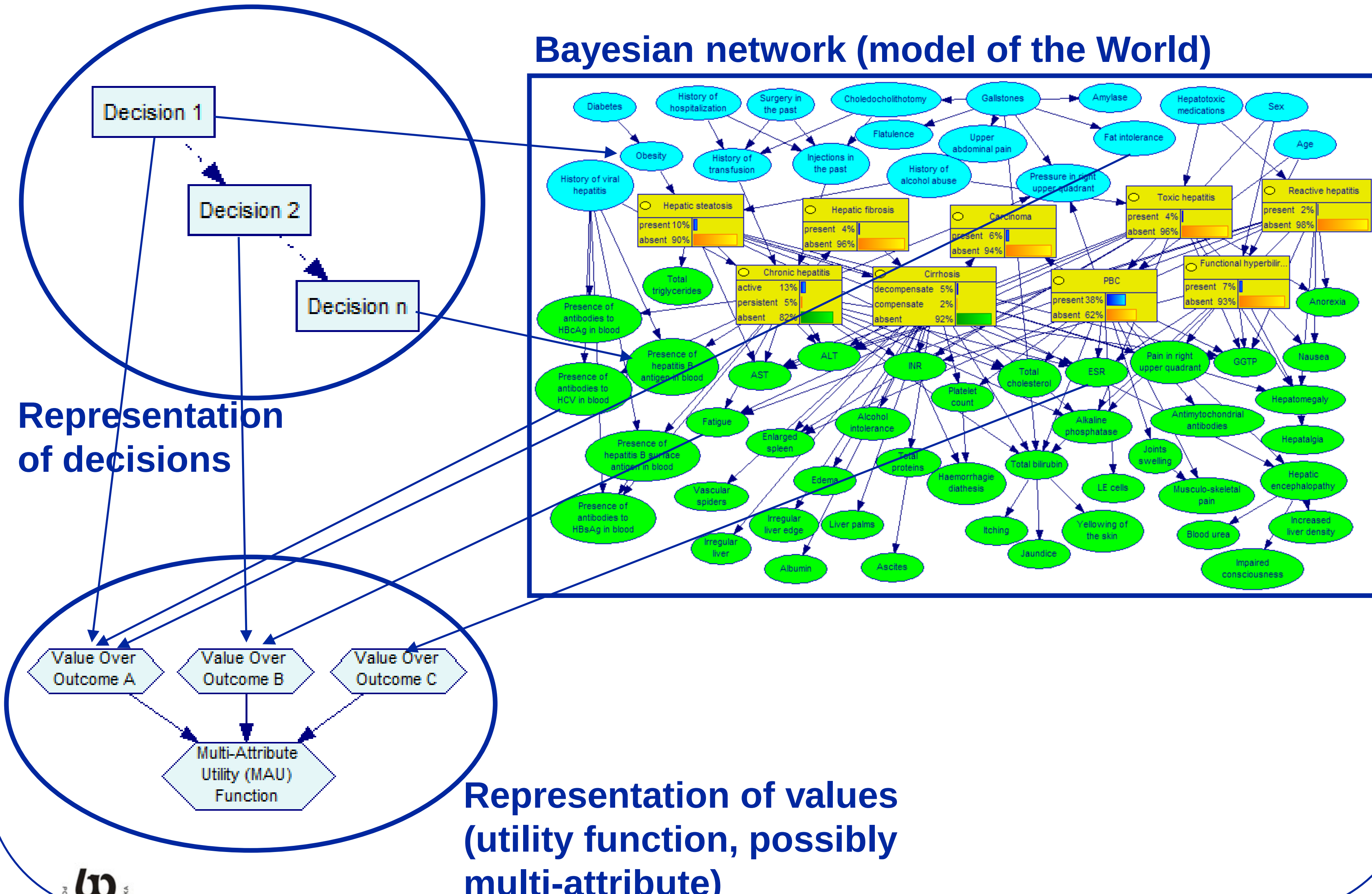
Dynamic Bayesian networks allow for tracking development of a system over time and support decision making in complex environments, where not only the final effect counts but also the system's trajectory.

Inspired by systems of differential equations

(the ground work for this was laid by Iwasaki & Simon in early 1990s)



Decision Making: Influence Diagrams



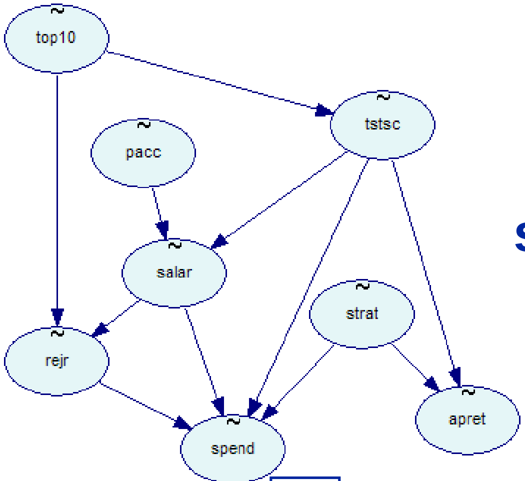
Learning/Data Mining

There exist algorithms with a capability to analyze data, discover causal patterns in them, and build models based on these data.

Retention.txt

spend	apret	top10	rejr	tstsc	pacc	strat	salar
9855	52.5	15	29.474	65.063	36.887	12	60800
10527	64.25	36	22.309	71.063	30.97	12.8	63900
7904	37.75	26	25.853	60.75	41.985	20.3	57800
6601	57	23	11.296	67.188	40.289	17	51200
7251	62	17	22.635	56.25	46.78	18.1	48000
6967	66.75	40	9.718	65.625	53.103	18	57700
8489	70.333	20	15.444	59.875	50.46	13.5	44000
9554	85.25	79	44.225	74.688	40.137	17.1	70100
15287	65.25	42	26.913	70.75	28.276	14.4	71738
7057	55.25	17	24.379	59.063	44.251	21.2	58200
16848	77.75	48	26.69	75.938	27.187	9.2	63000
18211	91	87	76.681	80.625	51.164	12.8	74400
21561	69.25	58	44.702	76.25	26.689	9.2	75400
20667	65	68	22.995	75.625	28.038	11	66200
10684	61.75	26	8.774	66	33.99	9.5	52900
11738	74.25	32	25.449	66.875	27.701	12	63400
10107	74	43	11.315	71	29.096	16.2	66200
7817	65.75	36	33.709	64.25	52.548	17.7	54600
7050	26	11	0	55.313	55.651	18.8	59500
9082	83.5	73	64.668	77.375	43.185	13.6	66700
11706	60	56	16.937	73.75	39.479	12.7	62100
7643	49.25	23	36.635	62.813	39.302	18.7	57700
25734	90	77	67.758	80.938	44.133	10	80200
20155	86	84	69.31	79.688	48.766	17.6	74000
29852	94.5	84	75.009	81.313	51.363	10.6	74100
7980	68.5	34	9.122	63.875	35.294	16.3	53100

Row 1 of 170



structure

Success	0.2
Failure	0.8

Success	Success	Failure
Good	0.4	0.1
Moderate	0.4	0.3
Poor	0.2	0.6

numerical parameters

data

GeNIe

A developer's environment for graphical decision models (https://www.bayesfusion.com/).

Qualitative interface: QGeNIe

Learning and discovery module: SMiner

Diagnosis: Diagnosis

Model developer module: GeNIe.

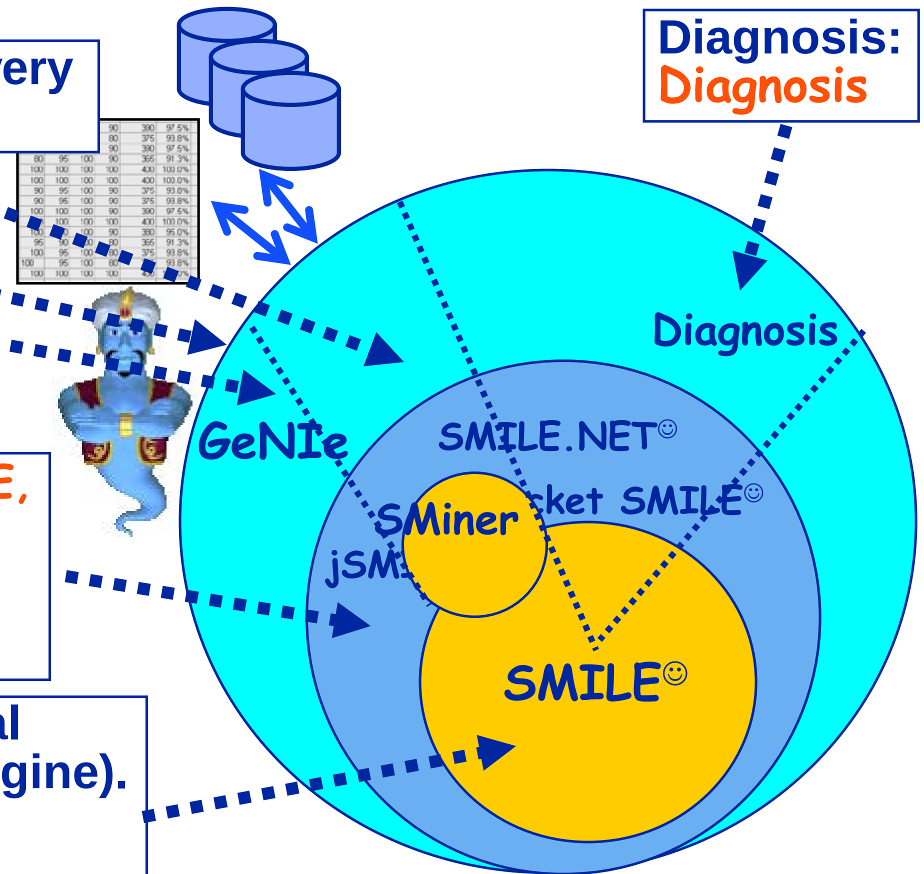
Implemented in Visual C++ in Windows environment.

Wrappers: SMILE.NET jSMILE, rSMILE, PySMILE, SMILE.COM, Pocket SMILE

Allow SMILE☺ to be accessed from applications other than C++compiler

Reasoning engine: SMILE☺ (**S**tructural **M**odeling, **I**nference, and **L**earning **E**ngine).

A platform independent library of C++ classes for graphical models.



The rest