Cool Things That One Can Do With Graphical Probabilistic Models

Marek J. Druzdzel

BayesFusion, LLC

marek@bayesfusion.com
http://www.bayesfusion.com/

Overview

- Motivation
- Bayesian probability theory
- Bayesian networks
- Extended family of graphical models
    - Hybrid Bayesian networks
    - Dynamic Bayesian Networks
    - Qualitative Bayesian networks
    - Influence diagrams
- Some applications
- BayesFusion, LLC
- Software demo

Motivation

Why statistics?

"... in this world nothing can be said to be certain, except death and taxes" --- Benjamin Franklin in a letter to his friend M. Le Roy

(*) *The Complete Works of Benjamin Franklin*, John Bigelow (ed.), New York and London: G.P. Putnam's Sons, 1887, Vol. 10, page 170

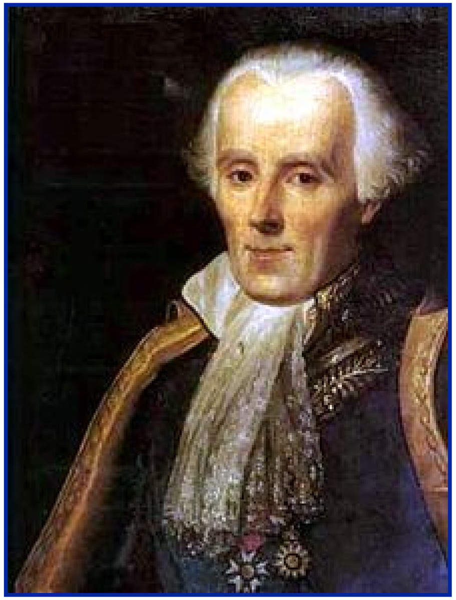
- In other words, "Uncertainty is prominent around us."
- It is an inherent part of all information and all knowledge.
- We need to deal with uncertainty in decision making.

Why probability theory and statistics?

"The theory of probabilities is basically only common sense reduced to a calculus."

("… la théorie des probabilités n'est, au fond, que le bon sens réduit au calcul.")

— Pierre-Simon Laplace, "Philosophical Essay on Probabilities" (1814)



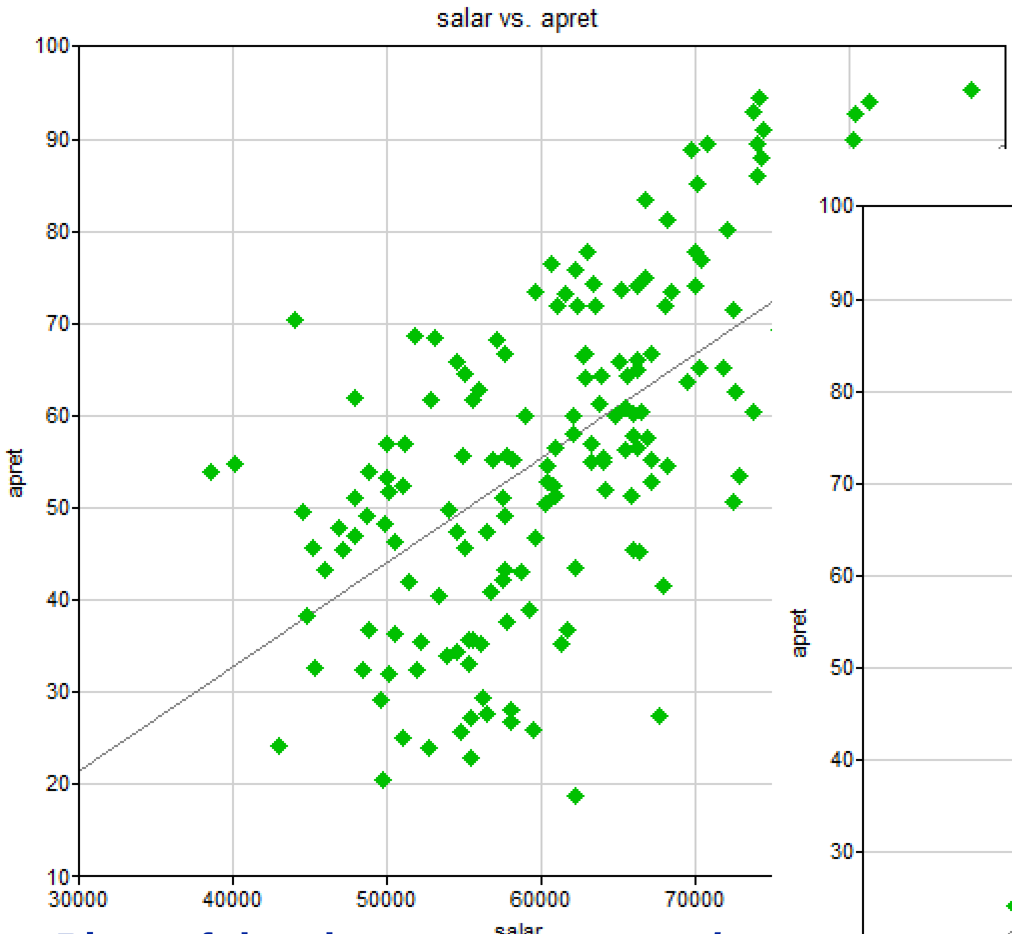
Bayesian Probability Theory

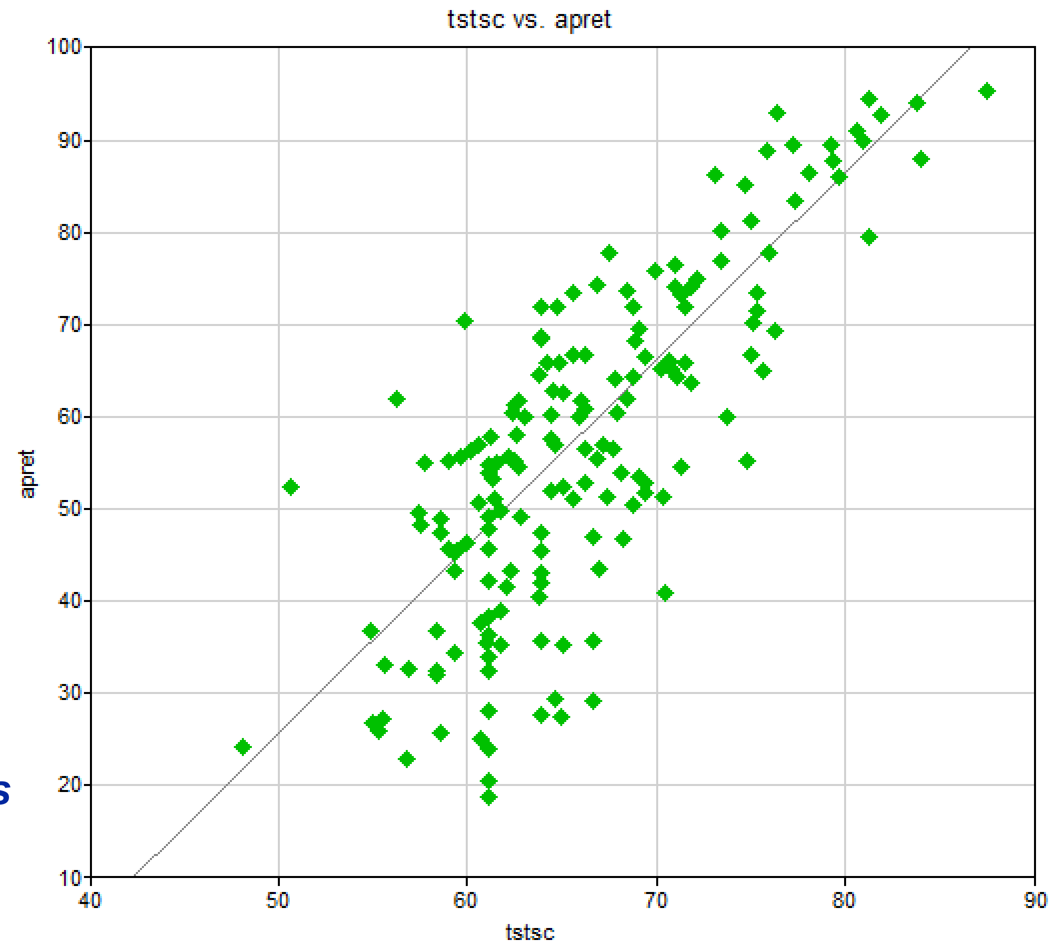
Joint probability distribution

Expresses the probability of events defined over several random variables

Source: http://postrecession.wordpress.com/tag/risk-aversion/

Joint probability distributions





Plots of data known as *scatter plots* give an idea of the joint probability distribution between two variables.

Joint probability distribution

Joint probability distributions are much more interesting than probability distributions over single variables

Why?

Given the value of some of the variables in the join probability distribution, we can estimate the probability distributions over the remaining variables.

e.g., we can predict the grade distribution in a university course given the amount of work that we put into the course

Bayes theorem

An easy to prove theorem, obtained from the definition of conditional probability:

From

$$P(A|B) = P(A,B) / P(B)$$

and

$$P(B|A) = P(A,B) / P(A)$$

we have

$$P(A|B) = P(B|A) / P(B)\ P(A)$$

Posterior (a.k.a. a-posteriori) probability

Prior (a.k.a. a-priori) probability

Bayes theorem gives us a mechanism for changing our opinion in light of new evidence!

Bayes theorem example

Let the prevalence of syphilis in the population of young people planning to get married in Pennsylvania be 0.001.

Let a (mandatory) test, required for obtaining the marriage license have sensitivity of 0.98 and specificity of 0.95.

What is the probability that your fiancée, who tested positive for syphilis, has syphilis?

$P(S|+) = P(+|S)/P(+)\ P(S)$ (Bayes theorem)

$P(+) = P(+|S)\ P(S) + P(+|\sim S)\ P(\sim S)$ (theorem of total probability)

$P(+) = 0.98\ 0.001 + 0.05\ 0.999 = 0.05093$

$$P(S|+) = 0.98\ 0.001\ /\ 0.05093\ 0.001$$

Posterior (a.k.a. a-posteriori) probability

Prior (a.k.a. a-priori) probability

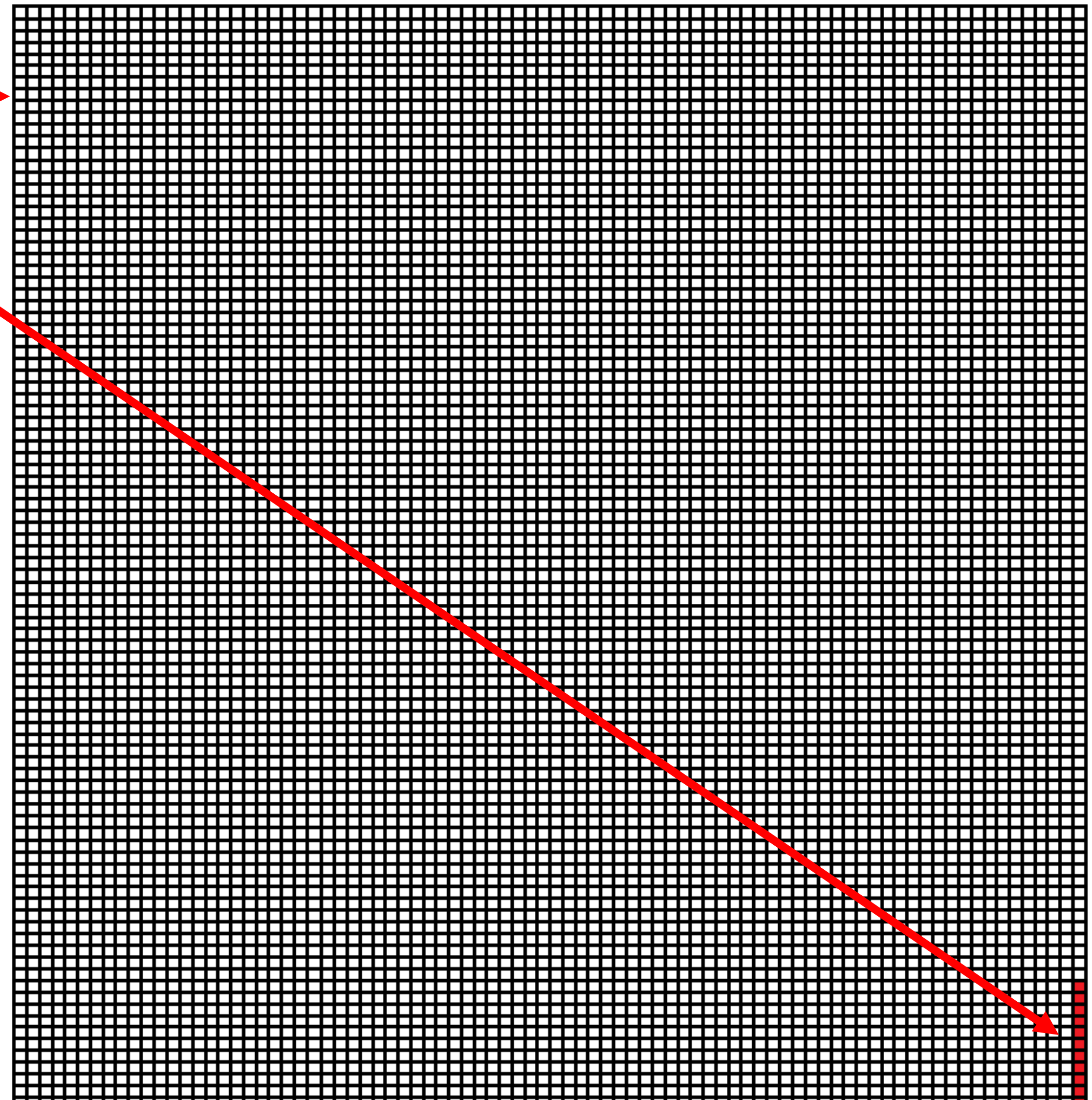
0.01924

A better human interface to the same problem

Imagine a population of 10,000 individuals.

Prevalence of 0.001 means that 10 out of the 10,000 will have the disease.

Let us screen them all.



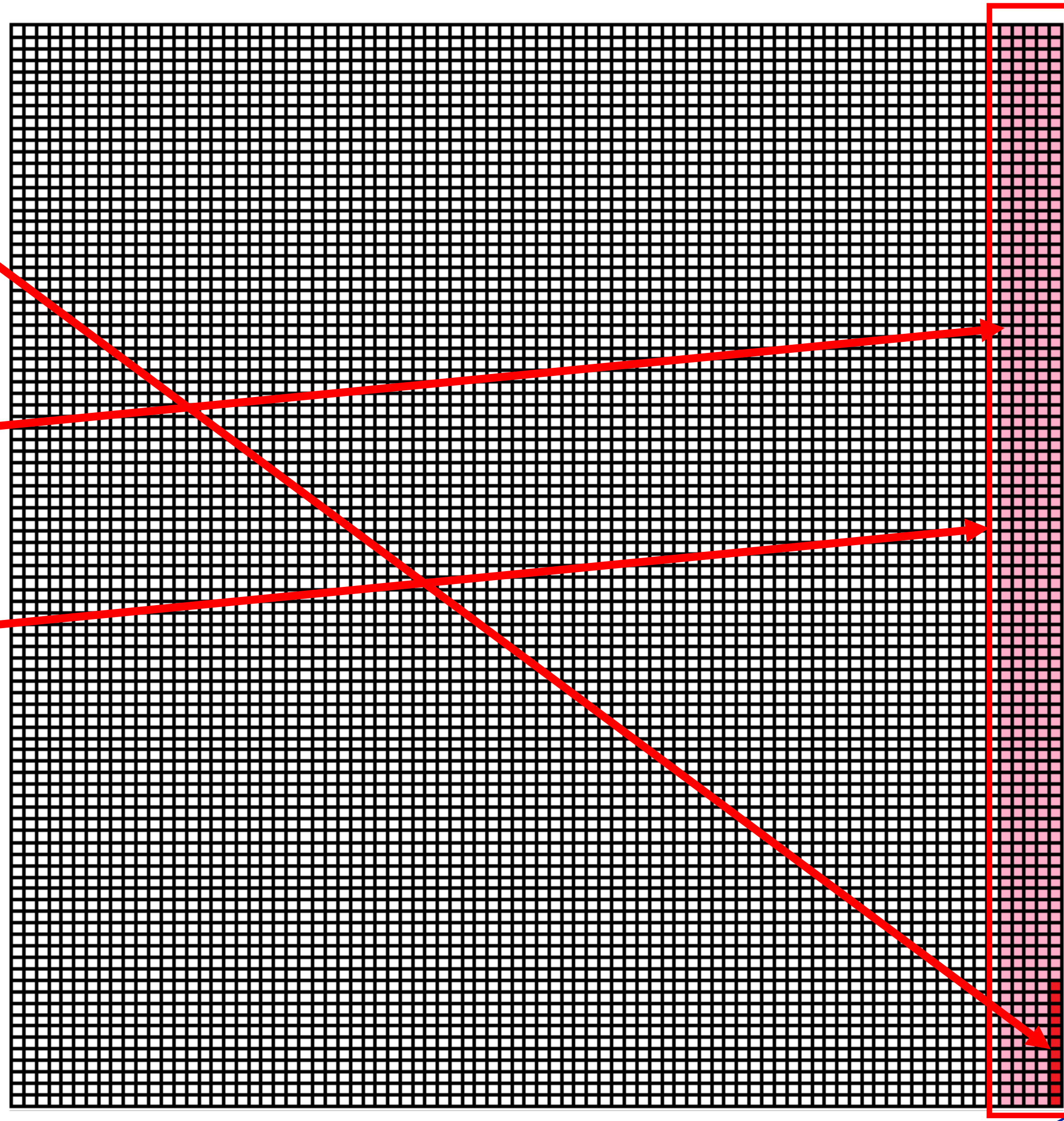
A better human interface to the same problem

With sensitivity of 98%, 9.8 of the 10 diseased will be correctly detected.

With specificity of 95%, we will have 5% (of 9,990), which is 499.5 false positives.

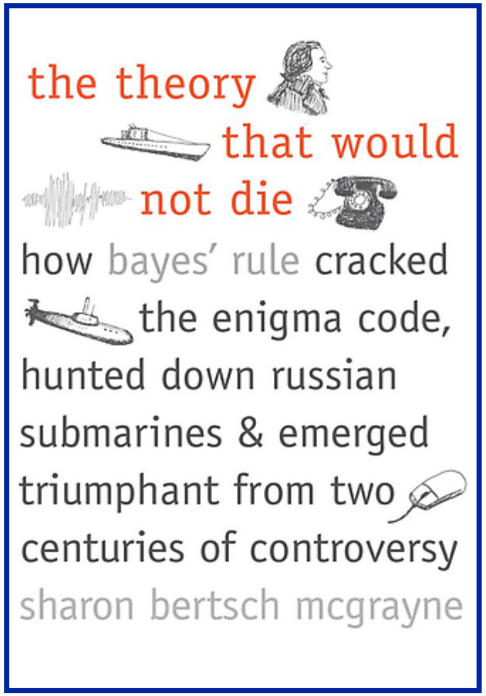
Now, among all those who tested positive, roughly 9.8/(9.8+499.5)≈2% will be diseased.

Is it easier to understand ☺?



Bayes theorem and Bayesian statistics

A versatile and powerful theory that seems to solve a variety of problems, originating from an 18th century English mathematician, Rev. Thomas Bayes (http://en.wikipedia.org/wiki/Thomas_Bayes)



Bayes Theory is so "hot" that a lightly written book "The Theory That Would Not Die," published in 2011, has become a bestseller

Recommended video:
http://www.youtube.com/watch?v=8oD6eBkjF9o

Bayesian modeling is reliable and it solves hard problems.

It can use both, data and expert knowledge.

What is the relation of Bayesian statistics to classical statistics?

What is p()?

Classical statisticians: "We have no clue ☹. Probability is a limiting frequency. A nuclear war is not a repetitive process."

Bayesians: "0.24 ☺. Probability is a measure of belief"

What is the relation of Bayesian statistics to classical statistics?

- Bayesians: "Probability is a measure of belief" (as opposed to "limiting frequency"), so it is subjective!
- Classical statisticians accuse Bayesians of "hocus pocus" with the prior distributions ("How do you know them?").
- Bayesian statistics comes with so called "limit theorems," which say that no matter what distribution you choose for your prior, you will eventually converge to the true distribution if you observe enough evidence.
- Of course, there is nothing wrong with starting with "the right distribution" in the beginning (In other words, it would be unwise to ignore available statistics).
- But even if you don't have them, you can still do useful work, even if you have to just guess the priors.

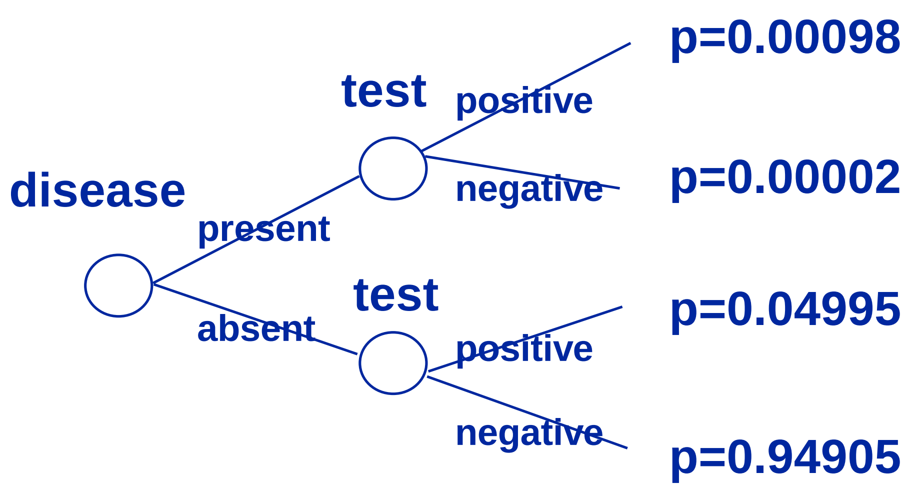
Bayesian Networks

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

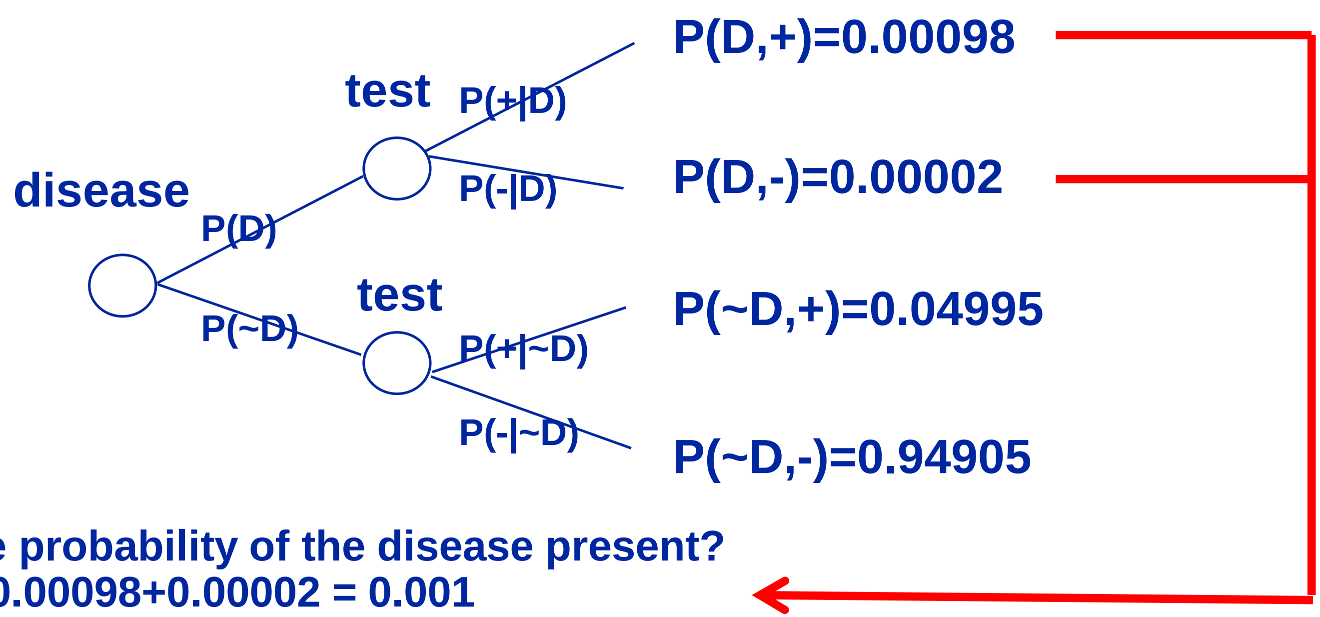
The simplest and quite natural graphical representation of a joint probability distribution over discrete variables



$P(\text{disease present} \wedge \text{test positive}) = P(D \cap +) = 0.00098$
$P(\text{disease present} \wedge \text{test negative}) = P(D \cap +) = 0.00002$
$P(\text{disease absent} \wedge \text{test positive}) = P(\sim D \cap +) = 0.04995$
$P(\text{disease absent} \wedge \text{test negative}) = P(\sim D \cap +) = 0.94905$

Computation in probability trees

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

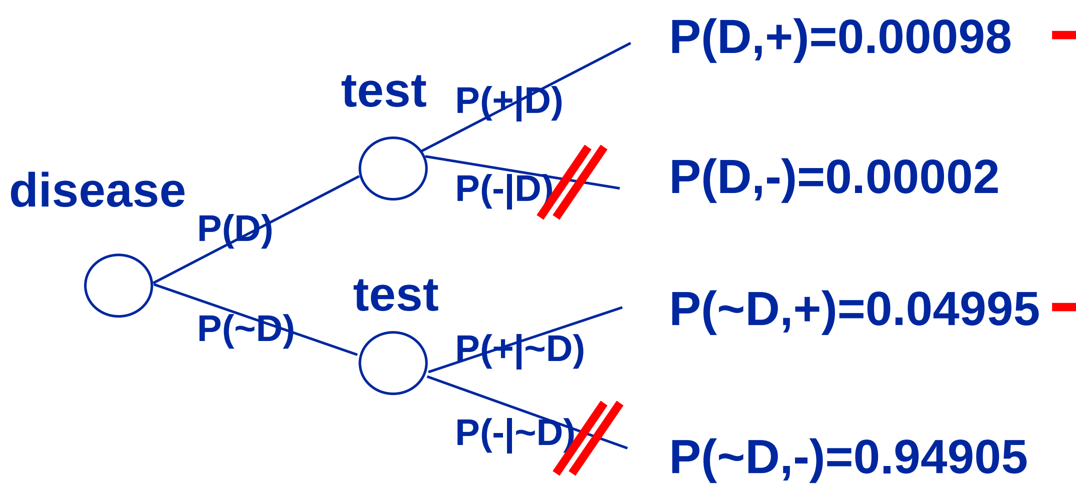


What is the probability of the disease present?

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

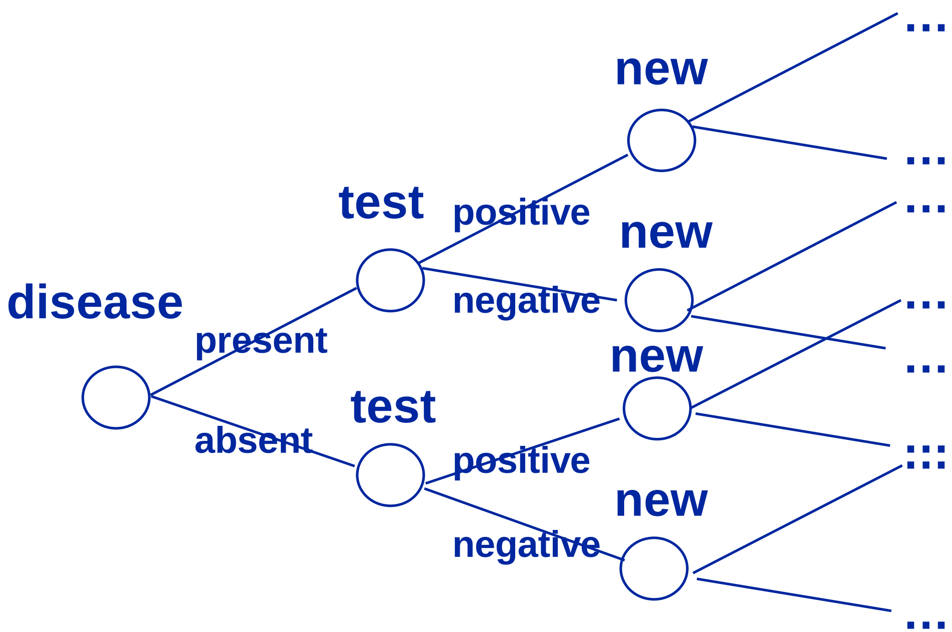


What is the probability of the disease present given a positive test result? Observation of a positive test result makes some of the branches of the tree impossible. What we need to do is just renormalize the remaining, possible (i.e., those that are compatible with the evidence) branches!

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

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 ☺)

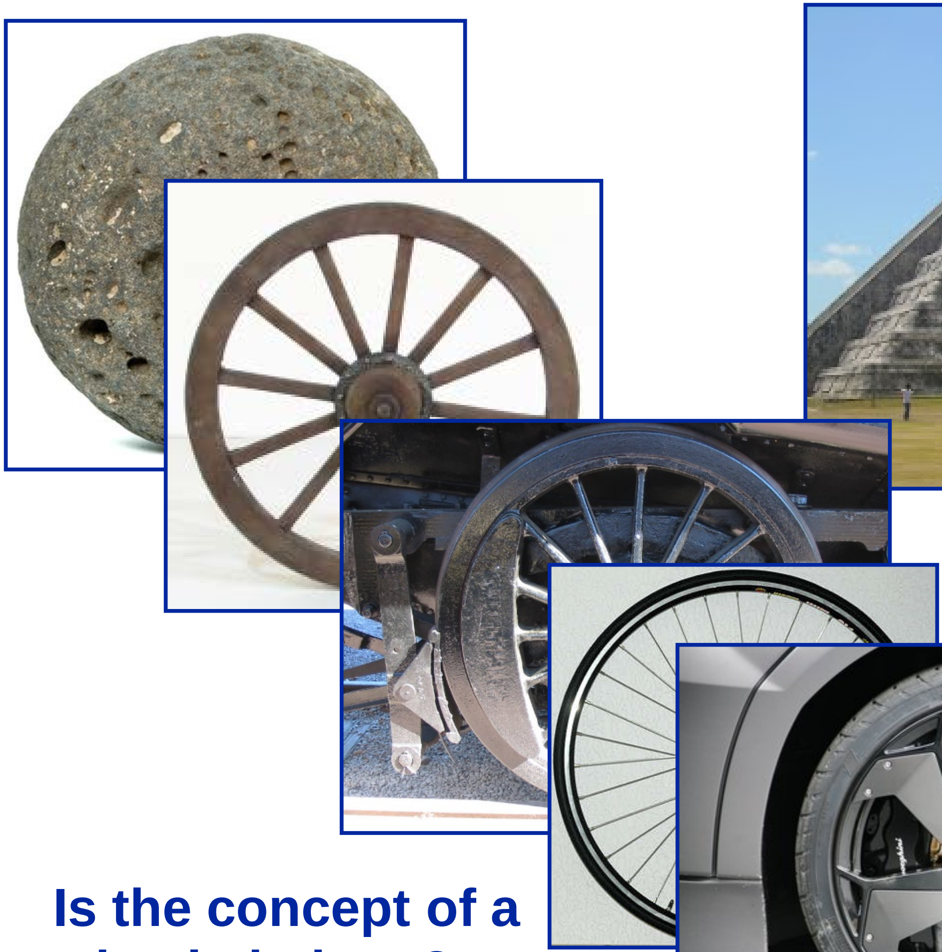
Great idea (only 30-40 years old)

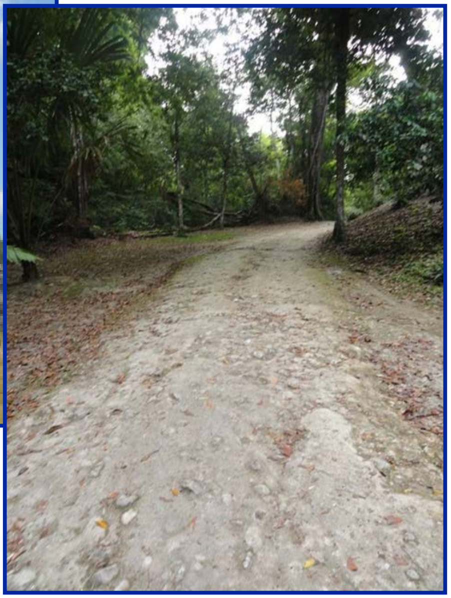
Use independences among variables in the joint probability distribution to reduce the number of parameters in its representation!

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

BAYESFUSION

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





Is the concept of a wheel obvious?

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

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

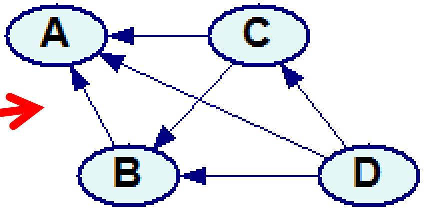
$$P(A,B,C,D)=P(B|A,C)\ P(D|C)\ P(A)\ P(C)$$

BAYESFUSION

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)

...

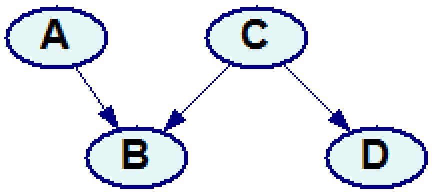
P(A,B,C,D)=P(B|A,C,D) P(D|A,C) P(A|C) P

...

B⊥D|C, D⊥A|C, A⊥C

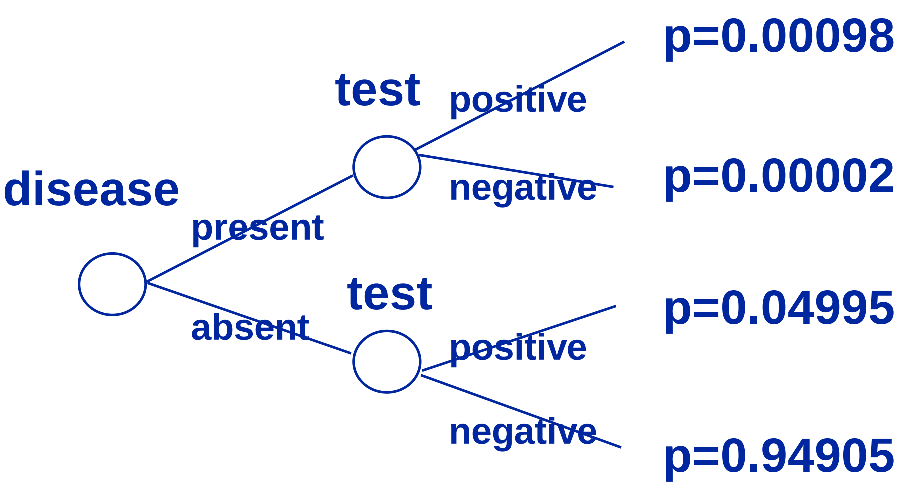
P(A,B,C,D)=P(B|A,C) P(D|C) P(A) P(C)

Absence of an arc is a graphical representation of independence!

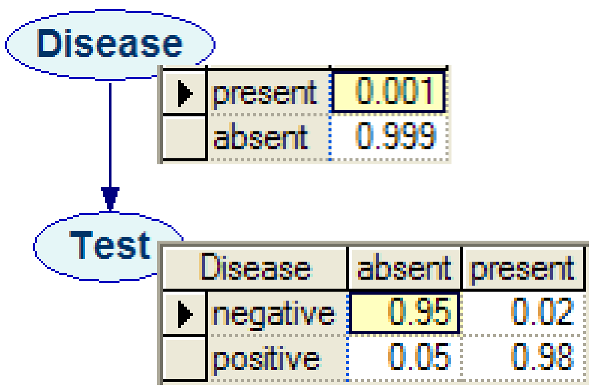


Probability trees and Bayesian networks

probability tree



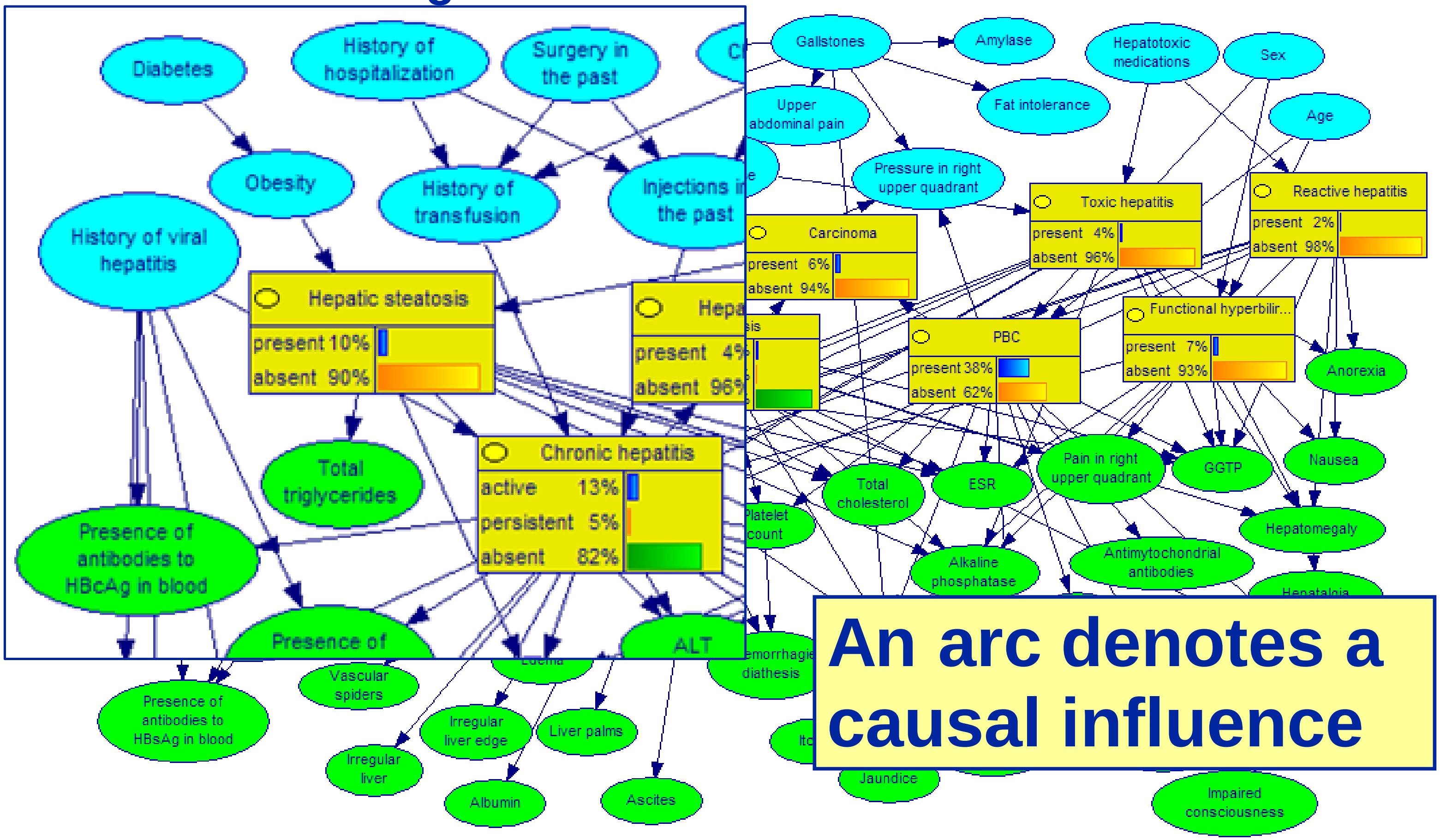
Bayesian network



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!

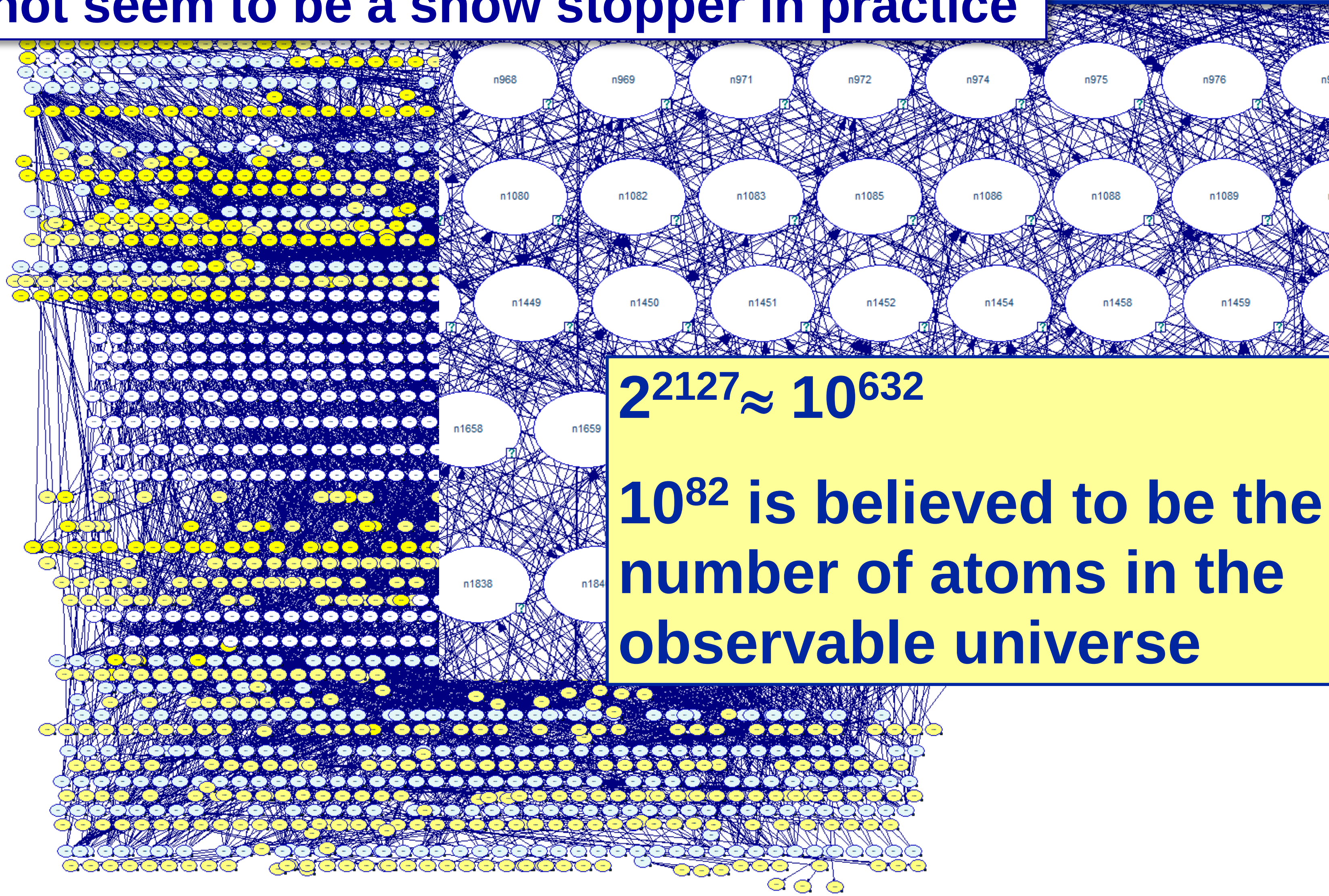
The graphical part of a Bayesian network is a picture of causal relations among the model variables.



An arc denotes a causal influence

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

Complexity of Bayesian networks does not seem to be a show stopper in practice



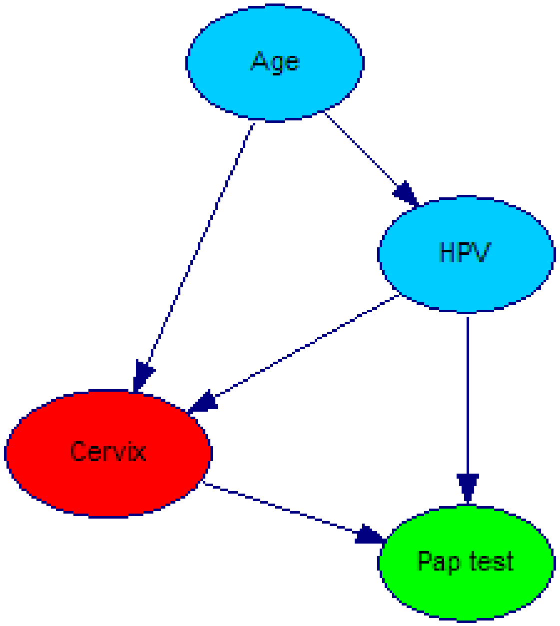
$2^{2127} \approx 10^{632}$

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 $2^{2,127} \approx 10^{632}$!)

Bayesian networks

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

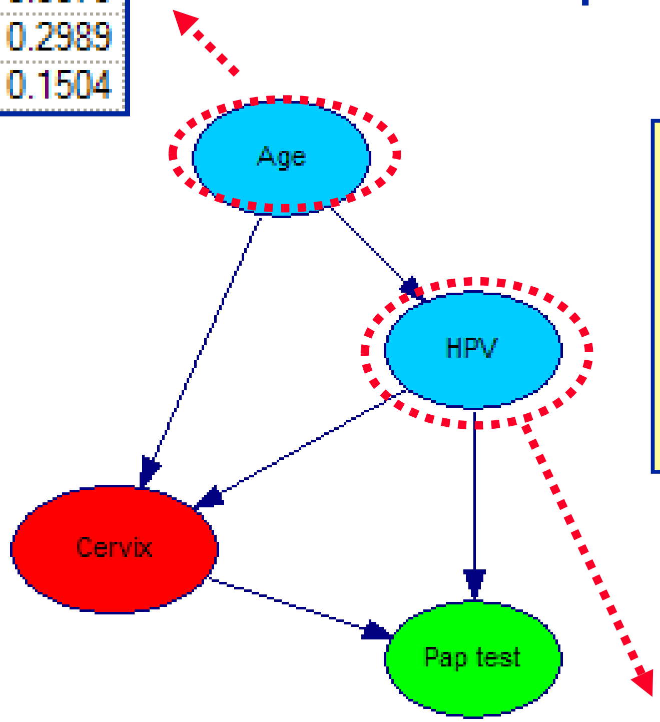


- The **qualitative part**, encoding a domain's variables (nodes) and the probabilistic (usually causal) influences among them (arcs).
- The **quantitative part**, encoding the joint probability distribution over these variables.

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)



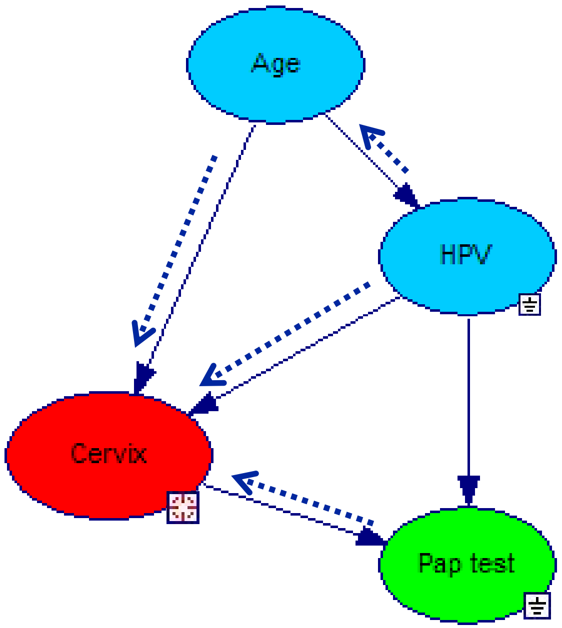
Please note that each absence of an arc (i.e., each independence modeled) is means one less dimension in the corresponding conditional probability table!

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: 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)

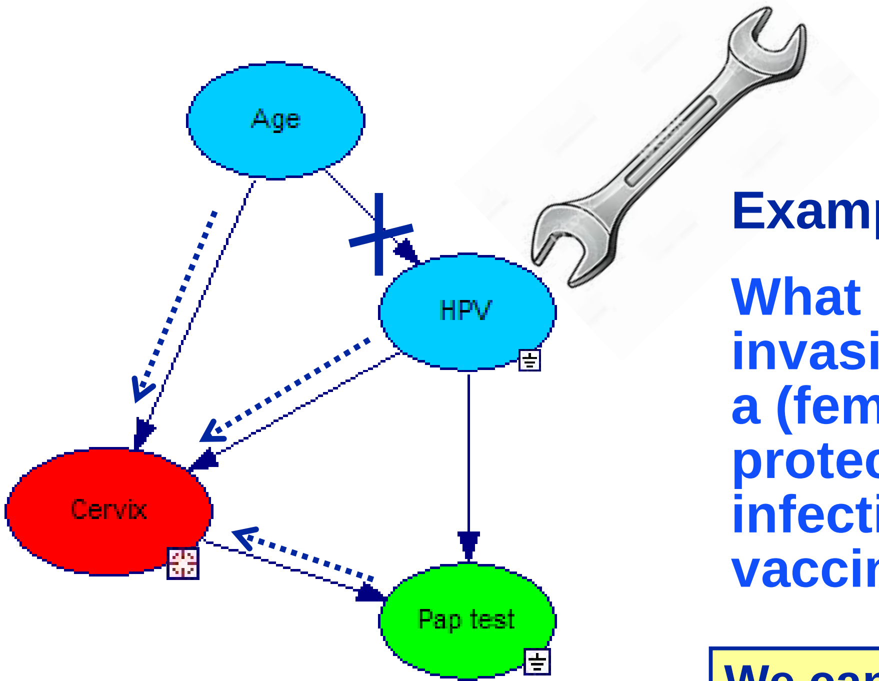
Example:

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.

Inference 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

Equation-based systems and graphical models

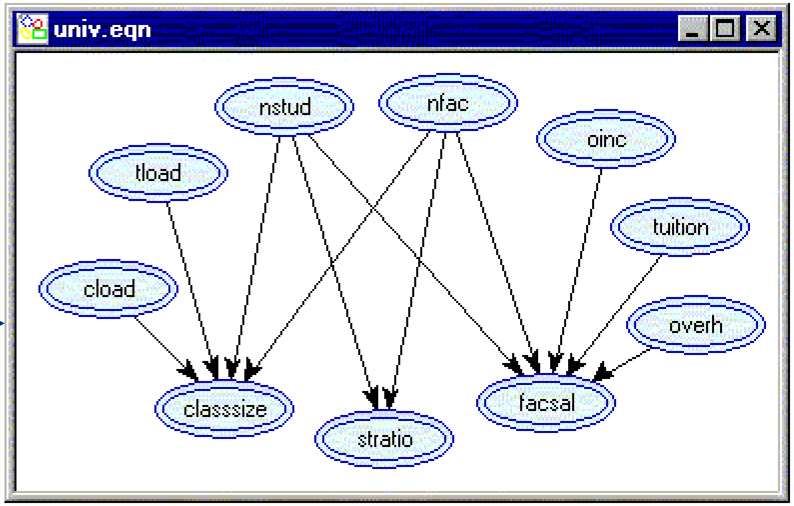
$$classsize = (nstud * cload) / (nfac * tload)$$
$$facsal = (oinc + tuition * nstud) / (nfac * (1 + overh))$$
$$stratio = nstud / nfac$$

← **Core equations**

$$cload = 15$$
$$tload = 6$$
$$nstud = 22102$$
$$nfac = 3006$$
$$oinc = 30000000$$
$$tuition = 12000$$
$$overh = 0.48$$

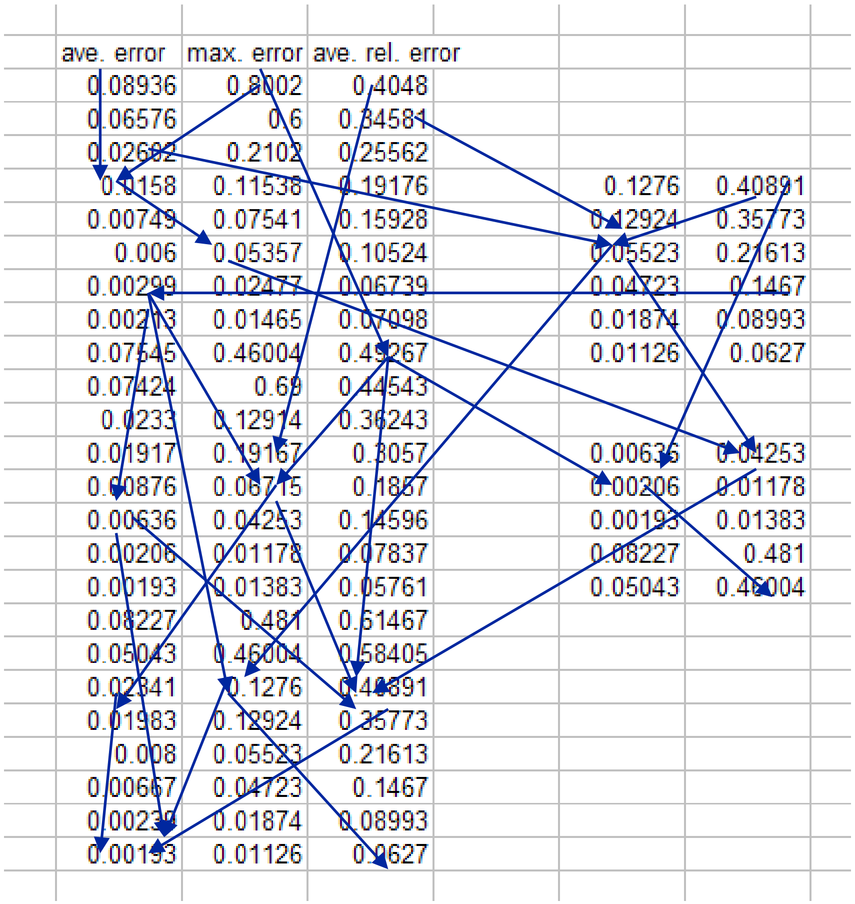
← **Equations for exogenous variables**

Together they determine the structure of the model →



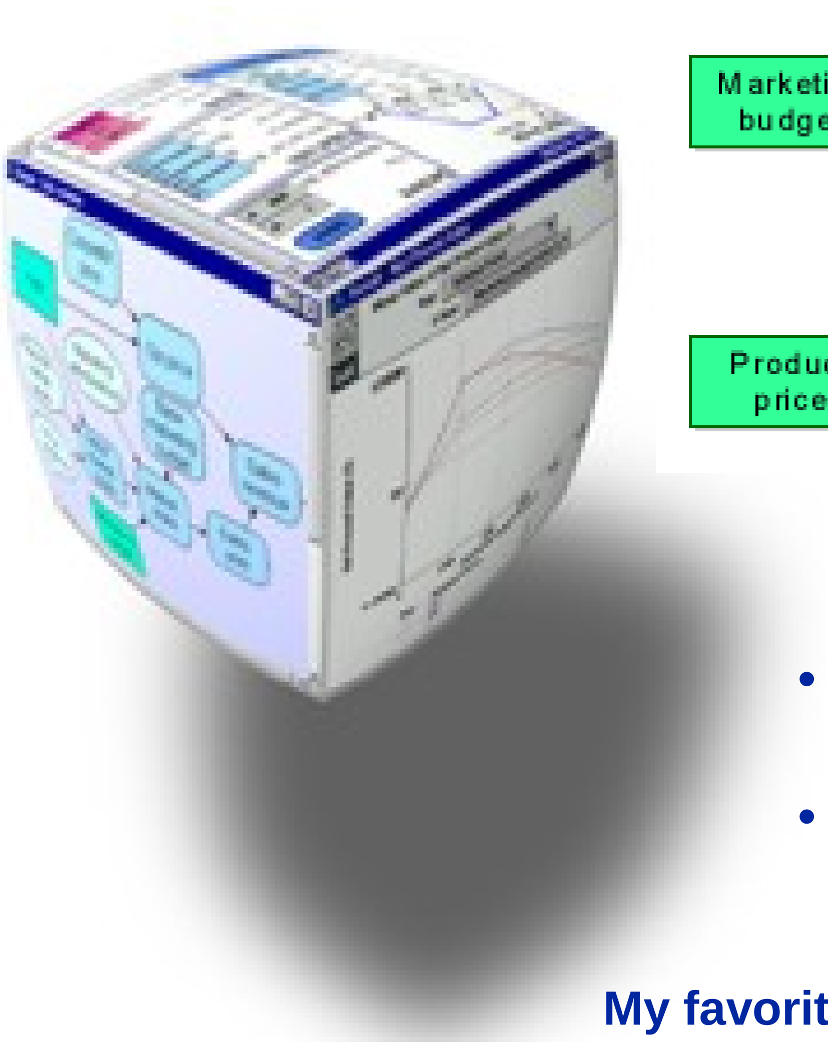
Explication of the asymmetries due to Herb Simon (early 1950s)

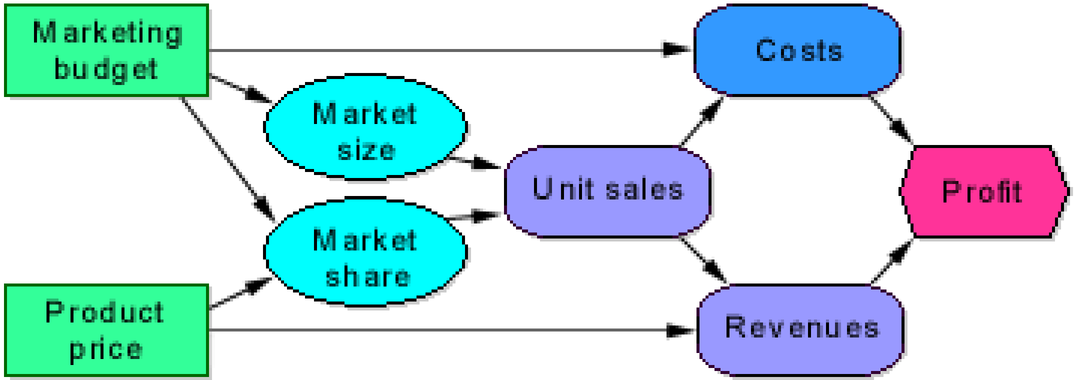
Spreadsheet models



- **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





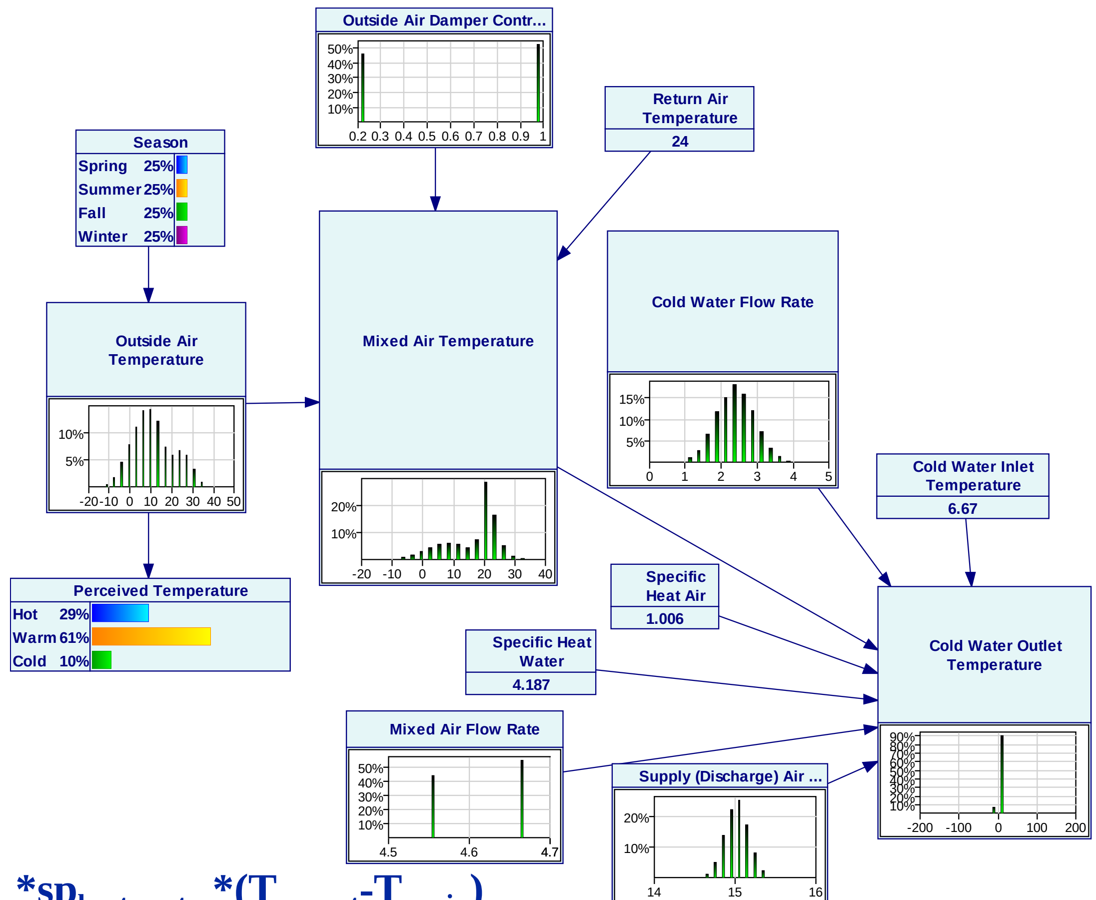
- **Fix almost everything that has been wrong with spreadsheets**
- **Great, but I believe that they could still be improved upon!**

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)

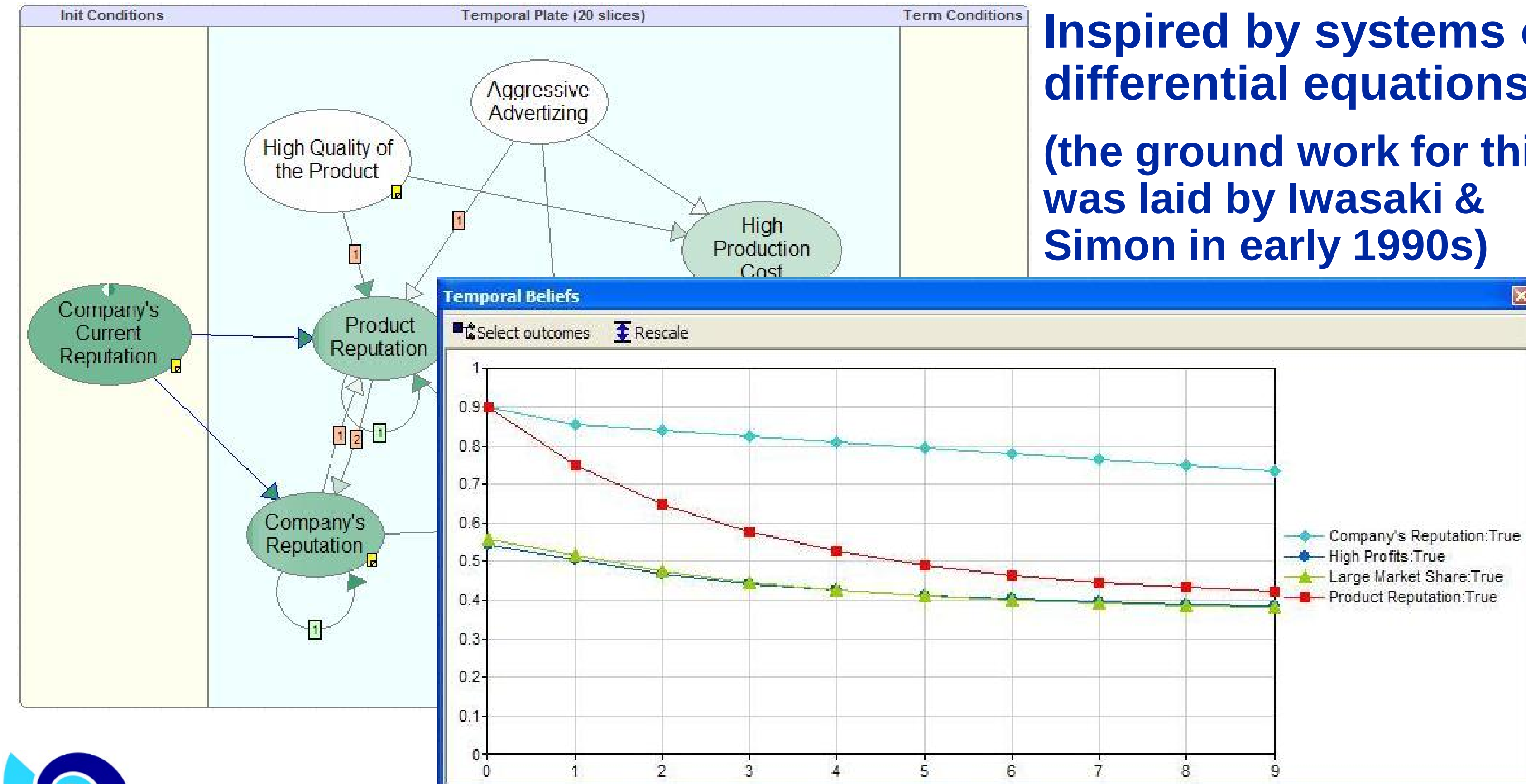
$$m_{flow\_ma}*sp_{heat\_air}*(T_{sa}-T_{ma}) = mdot_{cw}*sp_{heat\_water}*(T_{cw\_out}-T_{cw\_in})$$

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

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

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.



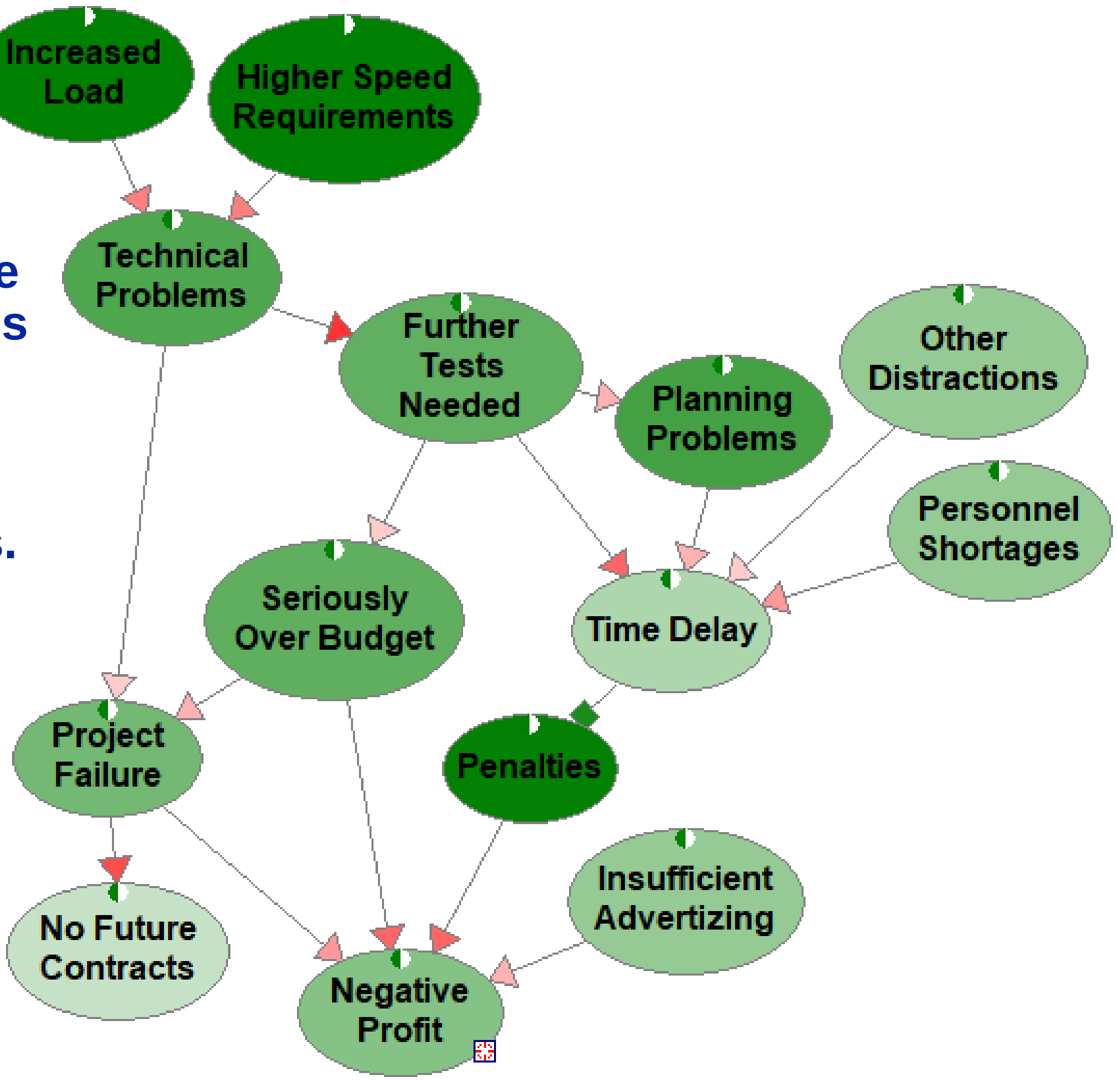
Inspired by systems of differential equations

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

Qualitative Bayesian networks

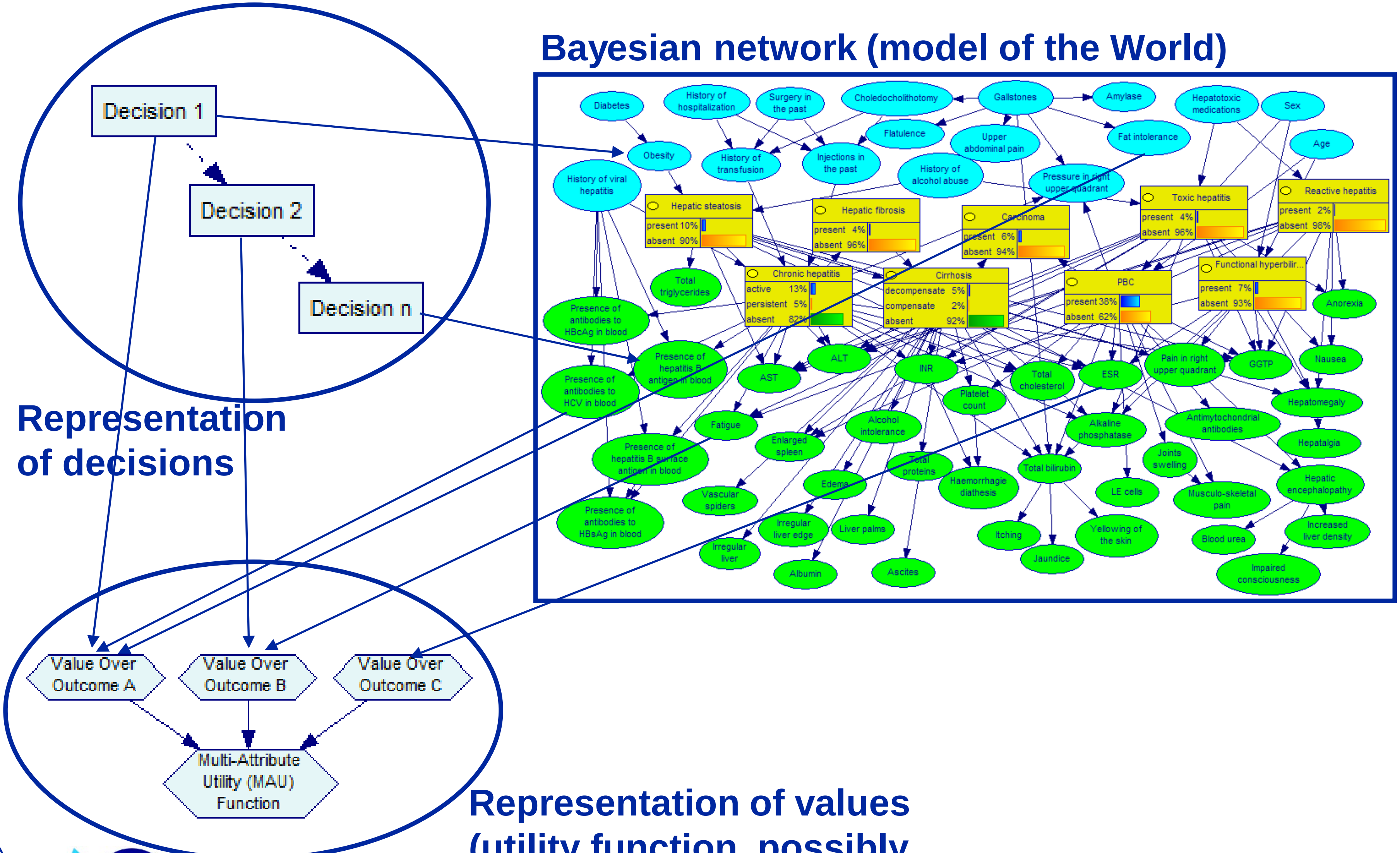
Qualitative interface to Bayesian networks require few numerical probabilities and allow for rapid model building and analysis.

They are great for group decision making sessions.



Decision Making: Influence Diagrams

Bayesian network (model of the World)



Representation of decisions

Representation of values (utility function, possibly multi-attribute)

BAYESFUSION

Learning/Data Mining

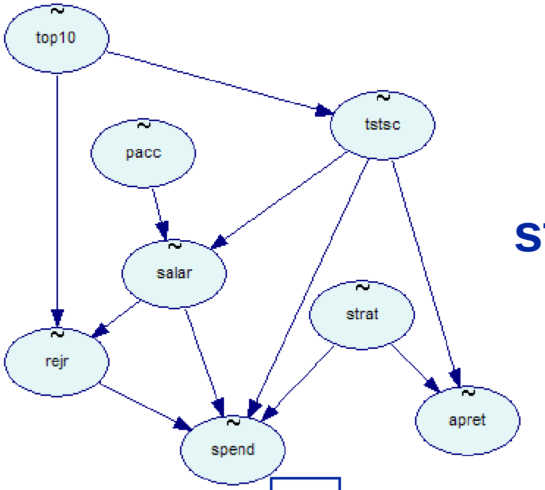
There exist algorithms with a capability to analyze data, discover causal patterns in them, and build models based on these data.

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

data



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

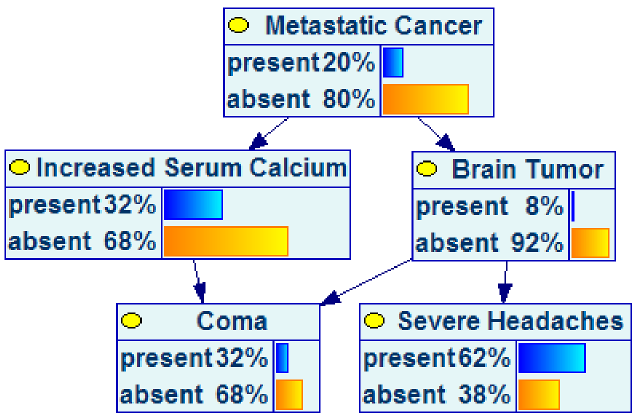
Some Applications

Diagnosis, prediction, prognosis

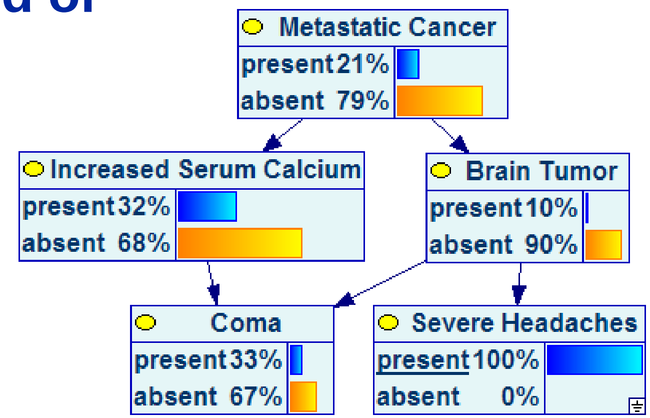
- This is a distinction based on the causal understanding of the World and relates to the direction of reasoning
- Finding out what happened is typically diagnosis
- Calculating what will or may happen is prediction/prognosis

Probability theory does not make this distinction: Bayesian updating

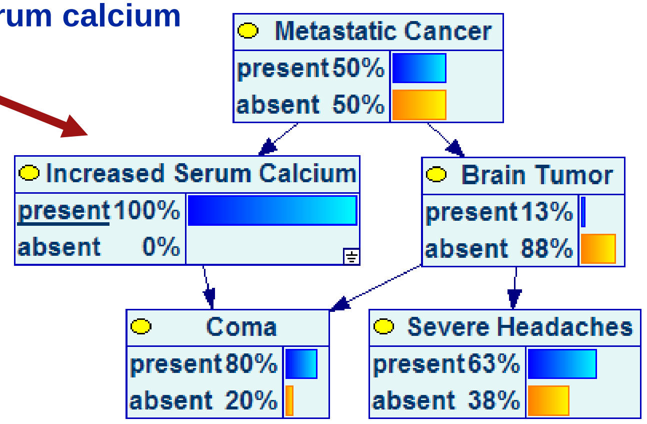
We can compute the impact of observations within a model, whether it goes forward or backward in time



Observation of severe headaches

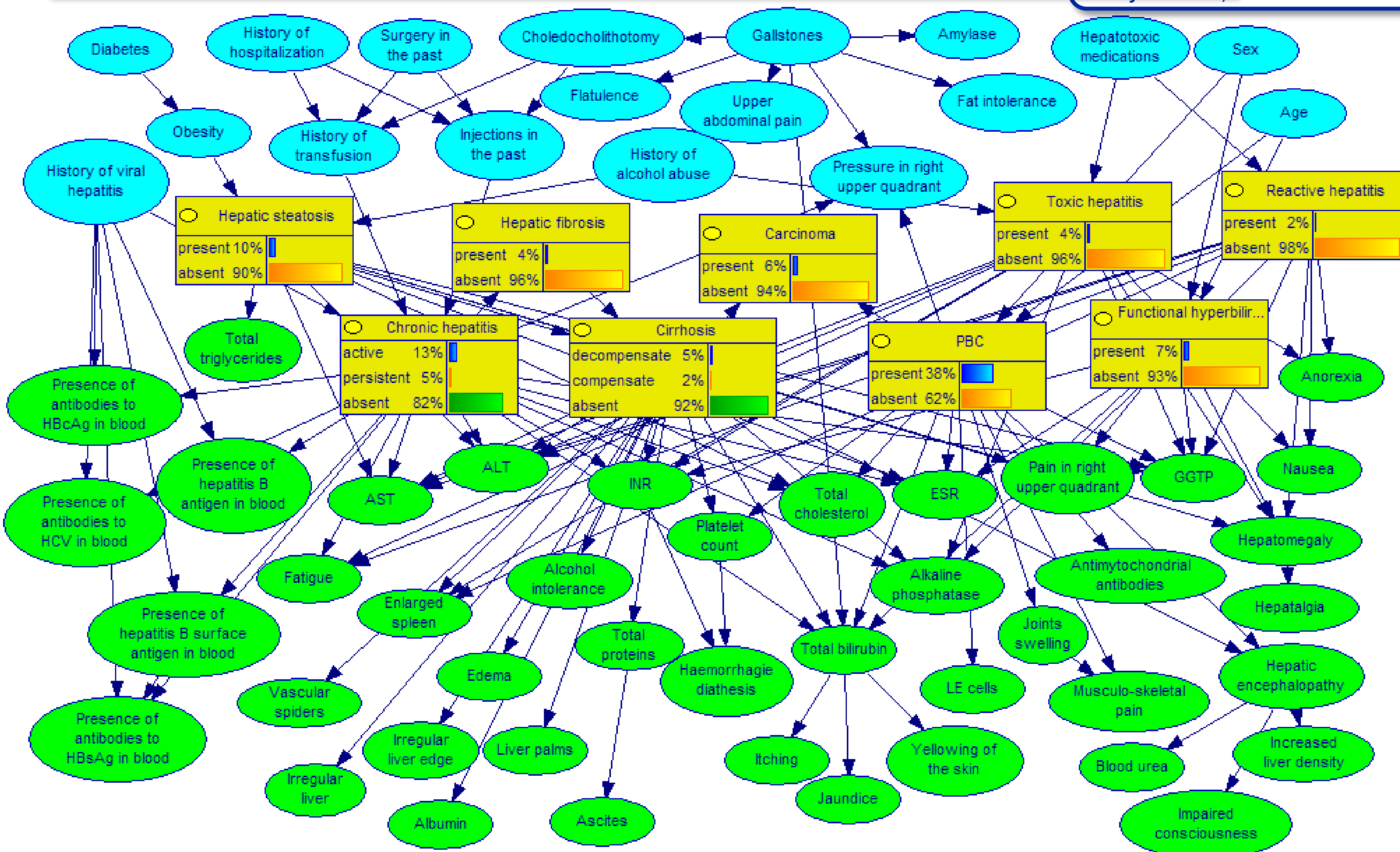


Observation of increased serum calcium



BAYESFUSION

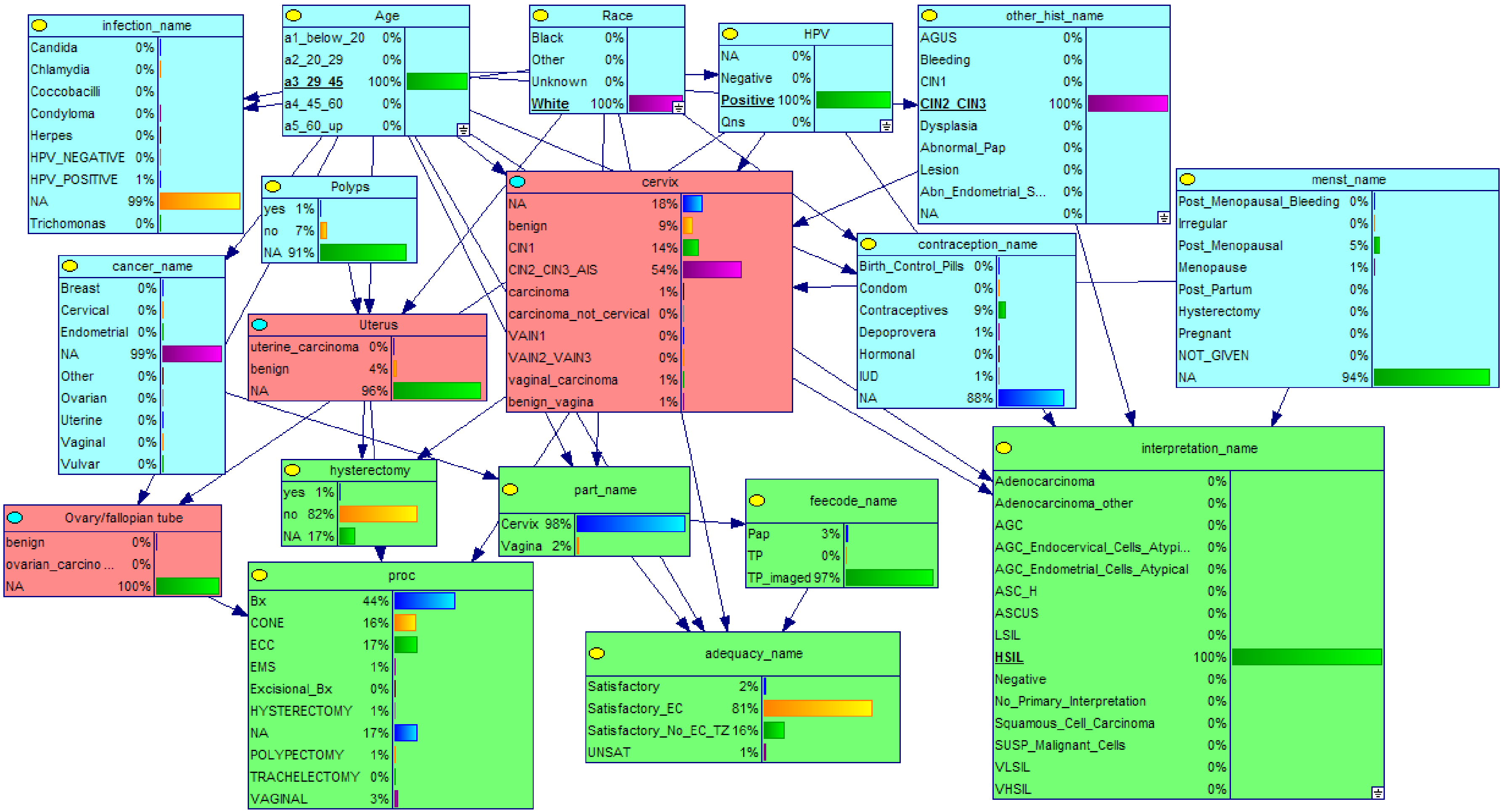
Diagnosis of liver disorders (the HEPAR II model)



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

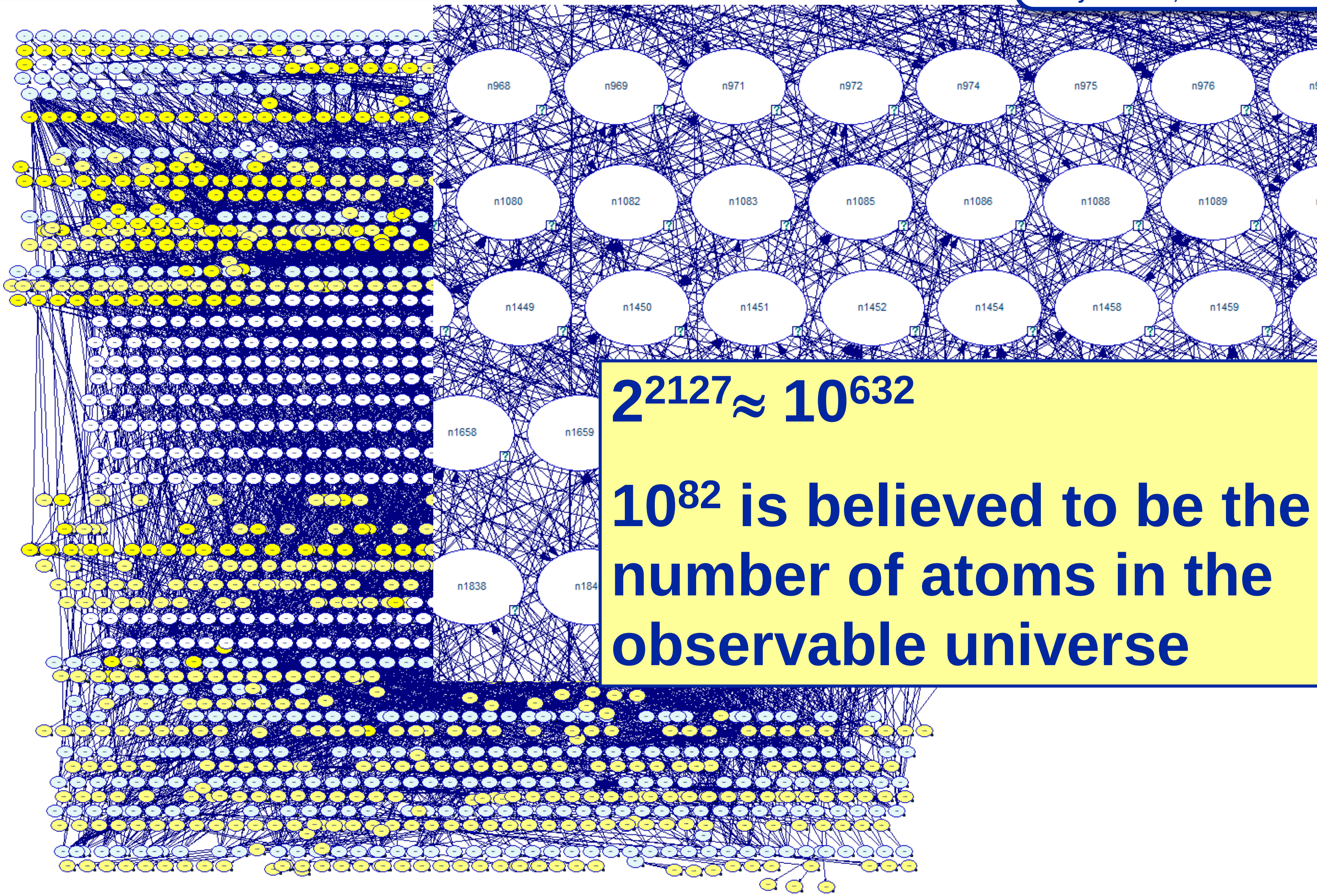
BAYESFUSION

Diagnosis and prognosis of cervical cancer (Pittsburgh Cervical Cancer Screening Model)



[Oniśko et al.] 18 variables; 295,163 numerical parameters (instead of over 10^{13}!)

Diagnosis of Diesel locomotives

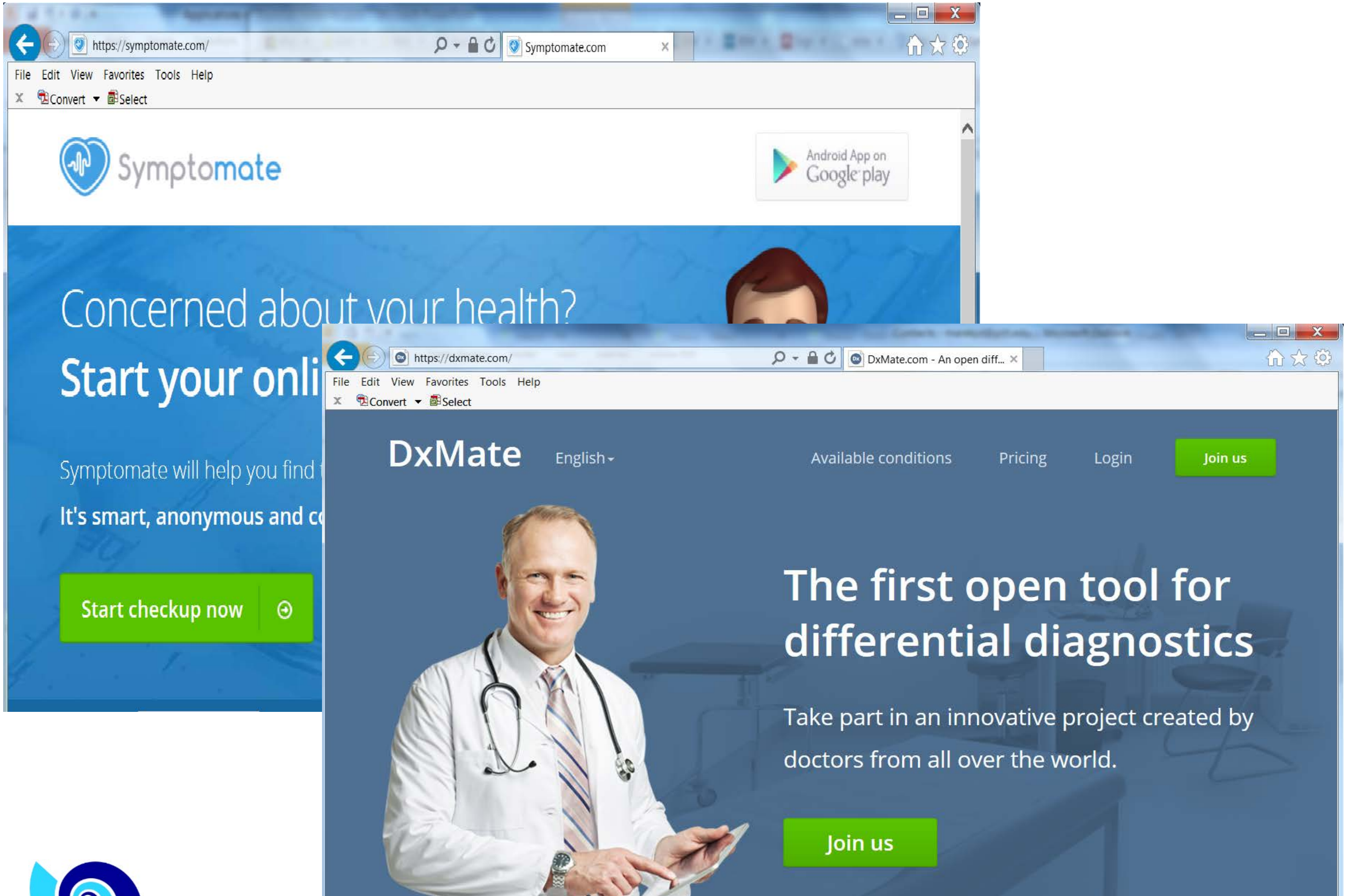


$2^{2127} \approx 10^{632}$

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

[Przytula et al.] 2,127 variables; 12,351 numerical parameters (instead of 2^{2127}!)

Symptomate: An intelligent medical consultant



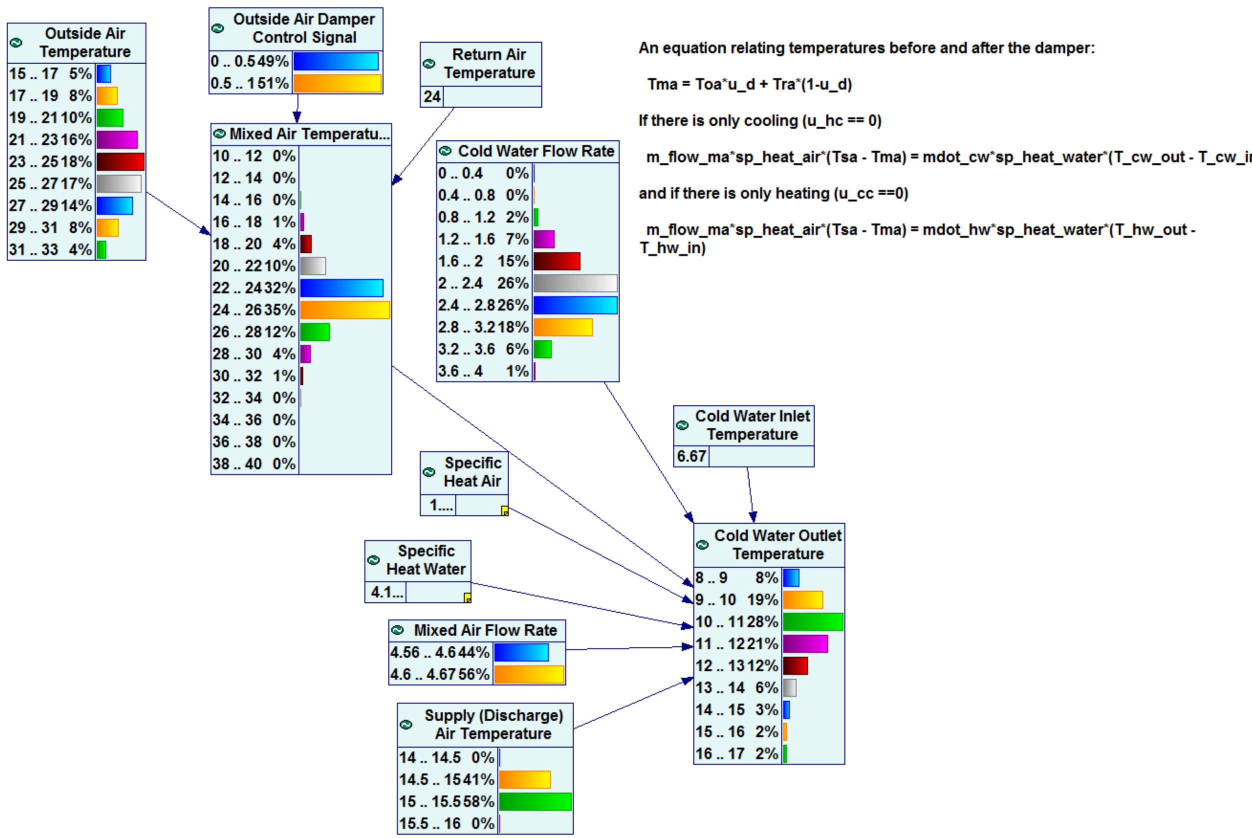
Other diagnostic applications

- Machine prognosis (in the context of machine maintenance)
- Diagnosis of database servers (Oracle)
- Diagnosis of airplanes (Boeing)
- Diagnosis of IC “baking” devices (Intel)

Modeling engineering and financial processes

Continuous Bayesian networks allow for modeling equation-based systems.

Diagnostic inference in such models is hard but we can discretize the variables for the purpose of inference.



Classification

http://guides.wikinut.com/img/23pyri.8.frh-6iy/Classification-of-Organisms

The problem of identifying to which set of categories (sub-populations) a new observation belongs on the basis of a training set of data containing observations (or instances) whose category membership is known.

Detection

- Spam detection
- Fraud detection
- Detection of conflicting medicine

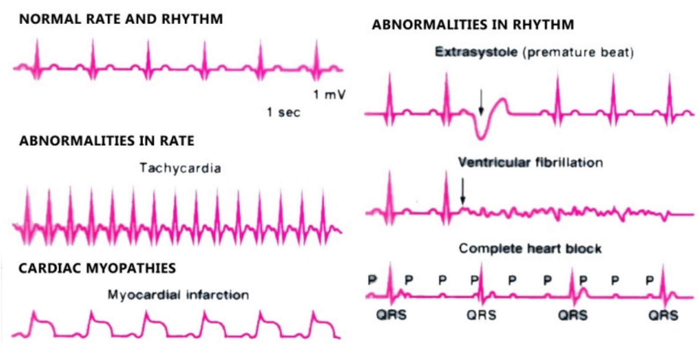
http://sciencebasedpharmacy.wordpress.com/tag/drug-regulation/

http://californialoanfind.com/what-and-who-is-teletrack/

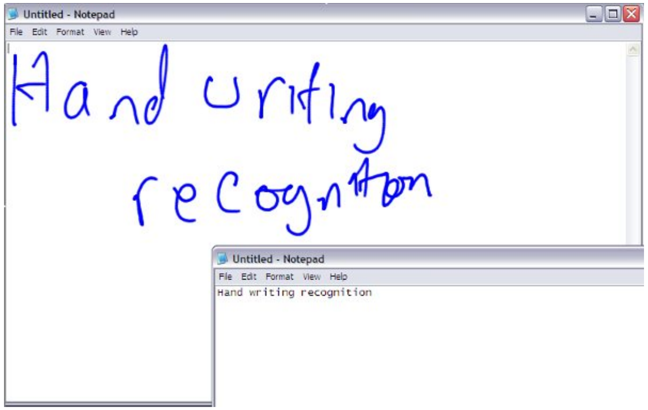
Recognition

- Handwriting recognition
- Face recognition
- Optical character recognition
- Pattern recognition
- Speech recognition

http://www.stanford.edu/class/cs224s/



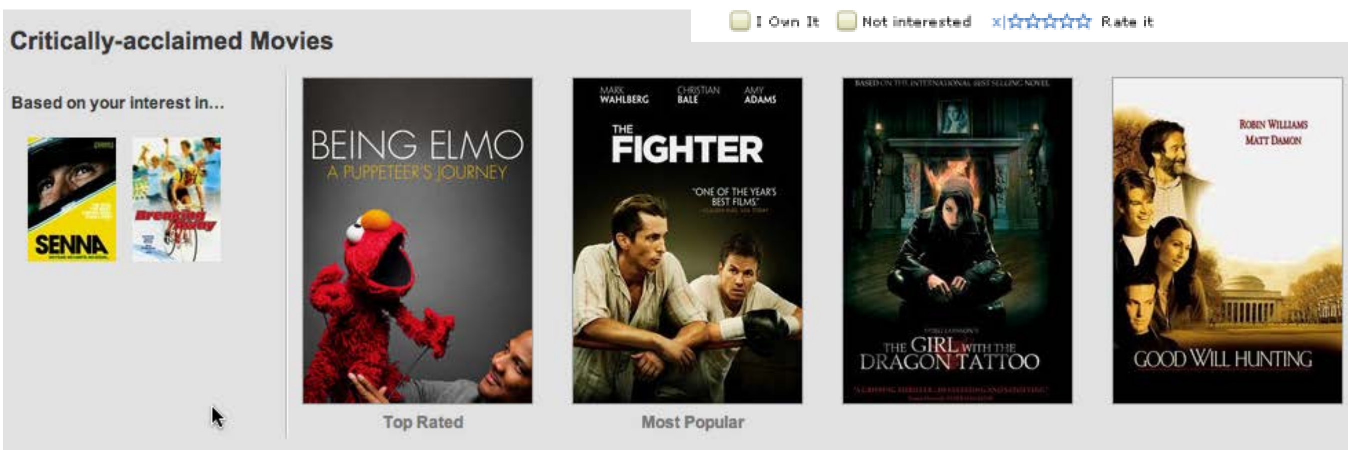
http://www.ivline.info/2010/05/quick-guide-to-ecg.html



http://www.l1id.com/pages/116-face http://networkprogramming.wordpress.com/2009/09/

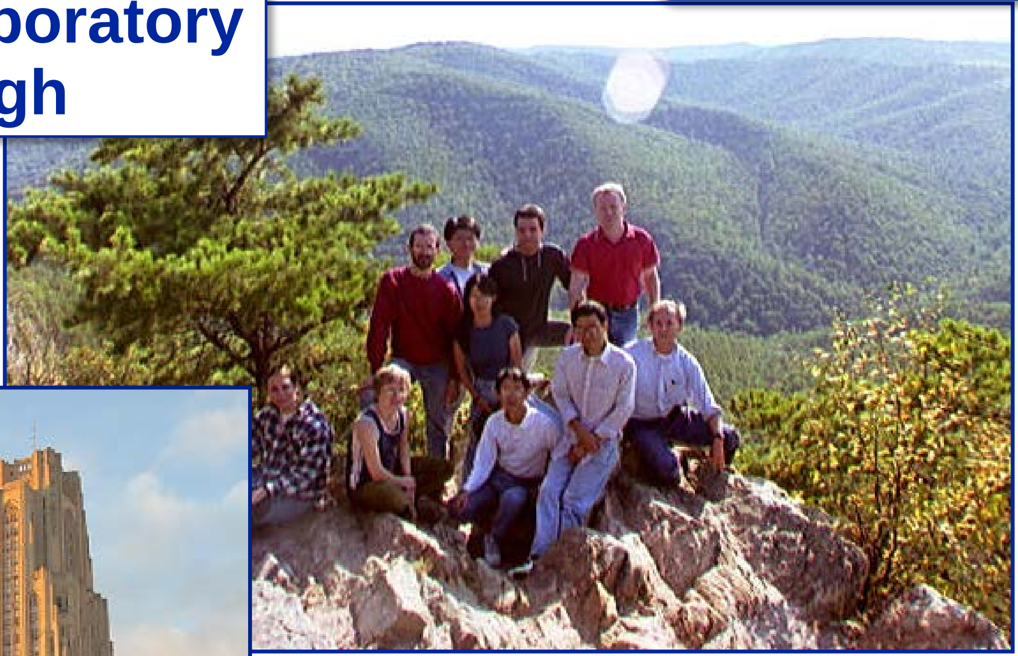
Recommender systems

An effective way to enhance customer shopping experience and increase sales



BayesFusion, LLC

The Origins: Decision Systems Laboratory University of Pittsburgh



BayesFusion, LLC

A Pennsylvania Limited Liability Corporation (LLC), formed 5 May 2015

http://www.bayesfusion.com/

Acquired license for GeNIe, QGeNIe and SMILE, from the University of Pittsburgh 22 June 2015

Partners: Tomek Sowinski & Marek Druzdzel

Research:
Pittsburgh, PA, USA & Bialystok, Poland

Development:
Bialystok, Poland

- **Self-funded, free and flexible in decision making.**
- **The primary source of our income is software sales …**
- **… however, we also offer custom software development, training, scientific consulting, and problem solving.**

The Architecture of GeNIe and SMILE☺

Developed between 1995 and 2015
Made available to the community in 1997
Reliable, fast, thousands of users

Qualitative interface: QGeNIe

Learning and discovery module

Diagnosis: Diagnosis

Model developer module: GeNIe.

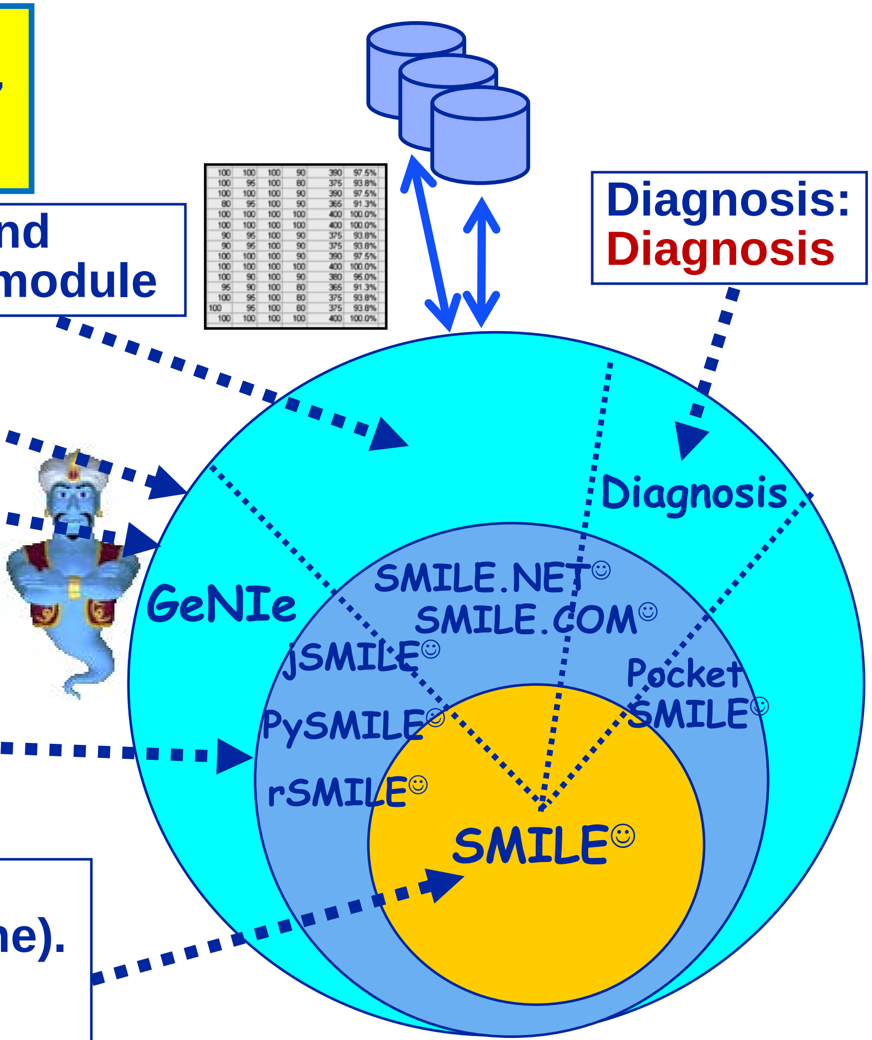
Implemented in Visual C++ in Windows environment.

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

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

Reasoning engine: SMILE☺ (Structural Modeling, Inference, and Learning Engine).

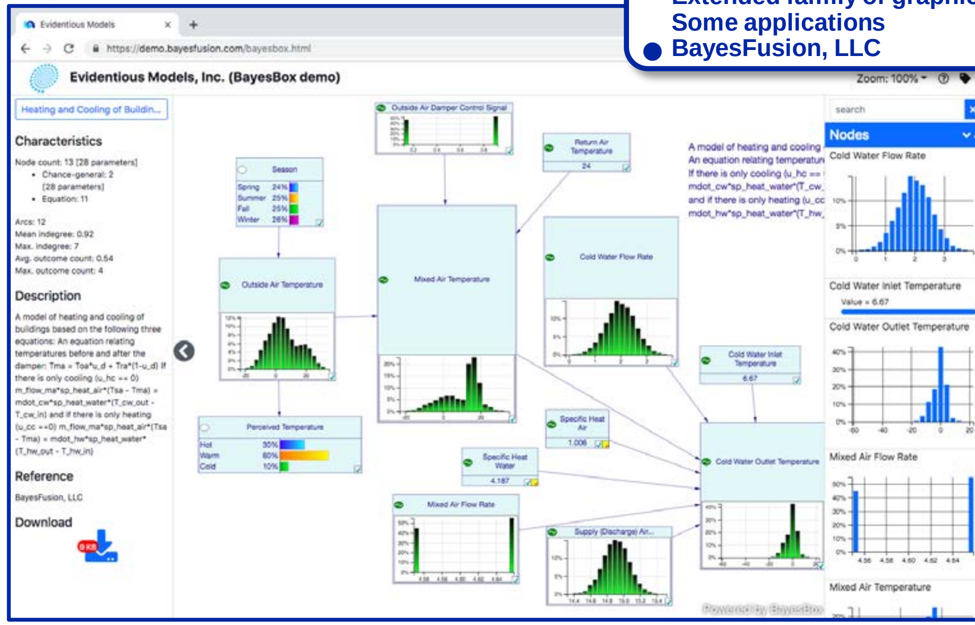
A platform independent library of C++ classes for graphical models.



New Products

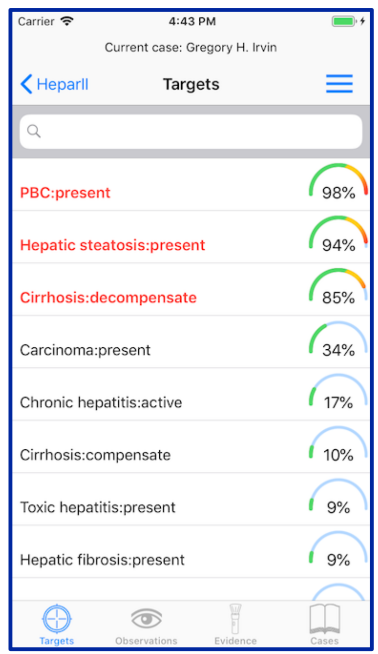
BayesBox:

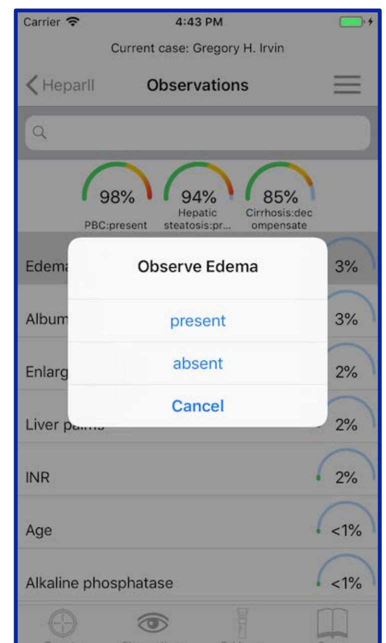
Fully customizable interactive cloud model repository



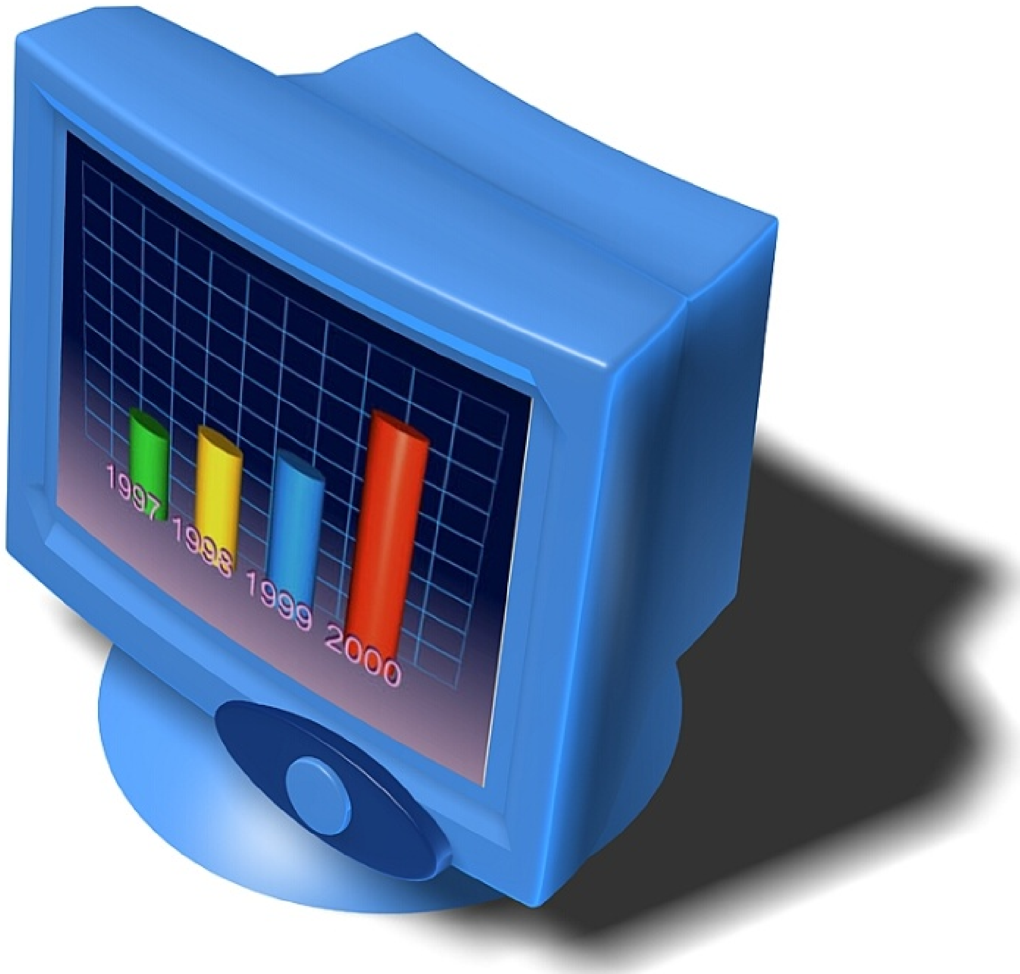
BayesMobile:

Diagnostic interface on portable devices





The rest



Thank you for your attention!

http://www.bayesfusion.com/

bayesfusion

bayesfusion

bayesfusion

BayesFusion

bayesfusion-llc

bayesfusion

For technical inquiries please visit http://support.bayesfusion.com/forum/