Adapted from AIMA slides

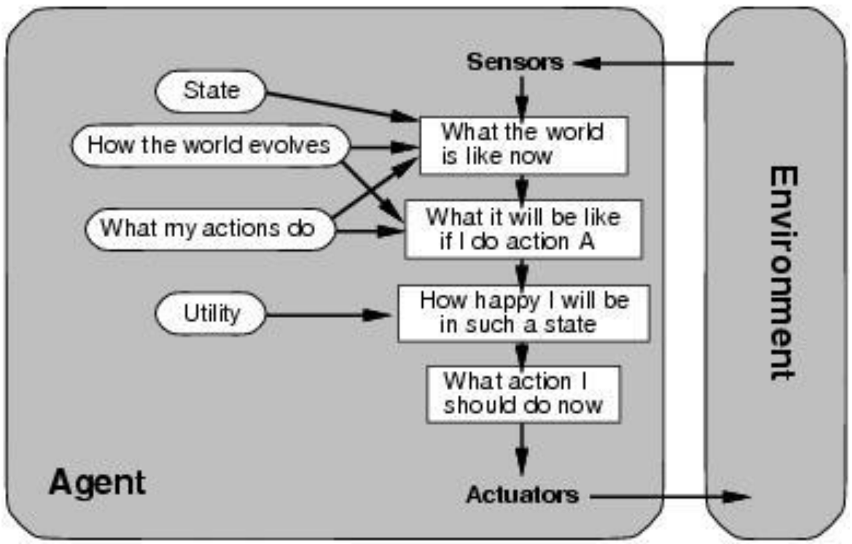
Uncertainty

Peter Antal
antal@mit.bme.hu

Outline

- Reminder
- A real-life example & demo for the homework
- Uncertainty
- Probability
- Syntax and Semantics
- Inference
- Independence and Bayes' Rule

Agent types; utility-based



- Certain goals can be reached in different ways.
  - Some are better, have a higher utility.
- Utility function maps a (sequence of) state(s) onto a real number.
- Improves on goals:
  - Selecting between conflicting goals
  - Select appropriately between several goals based on likelihood of success.

Rationality

- What is rational at a given time depends on four things:
    - Performance measure,
    - Prior environment knowledge,
    - Actions,
    - Percept sequence to date (sensors).
- DEF: *A rational agent chooses whichever action maximizes the expected value of the performance measure given the percept sequence to date and prior environment knowledge.*

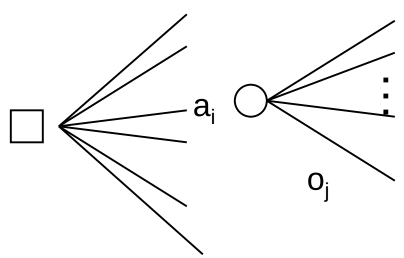
Decision theory
probability theory+utility theory

- Decision situation:
  - Actions a_i
  - Outcomes o_j
  - Probabilities of outcomes $p(o_j \mid a_i)$
  - Utilities/losses of outcomes $U(o_j \mid a_i)$
  - Maximum Expected Utility Principle (MEU) $EU(a_i) = \sum_j U(o_j \mid a_i) p(o_j \mid a_i)$
  - Best action is the one with maximum expected utility $a^* = \arg\max_i EU(a_i)$

Decision theory
probability theory+utility theory

- Decision situation:
  - Actions a_i
  - Outcomes o_j
  - Probabilities of outcomes $p(o_j \mid a_i)$
  - Utilities/losses of outcomes $U(o_j \mid a_i)$

Actions a_i | Outcomes | Probabilities | Utilities, costs | Expected utilities

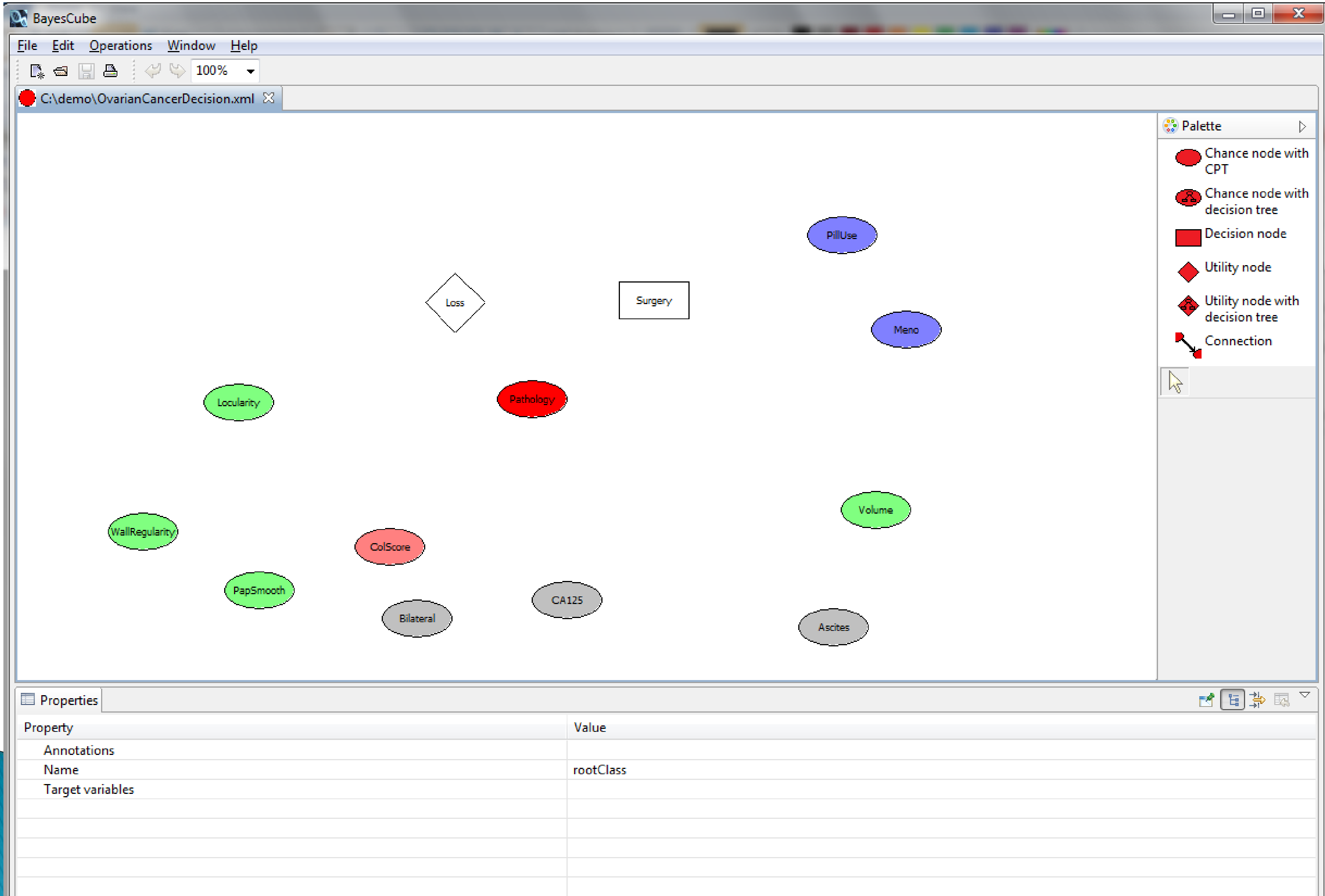


$P(o_j|a_i)$ ⋮ | $U(o_j)$, $C(a_i)$ ⋮ | $EU(a_i) = \sum P(o_j|a_i)U(o_j)$

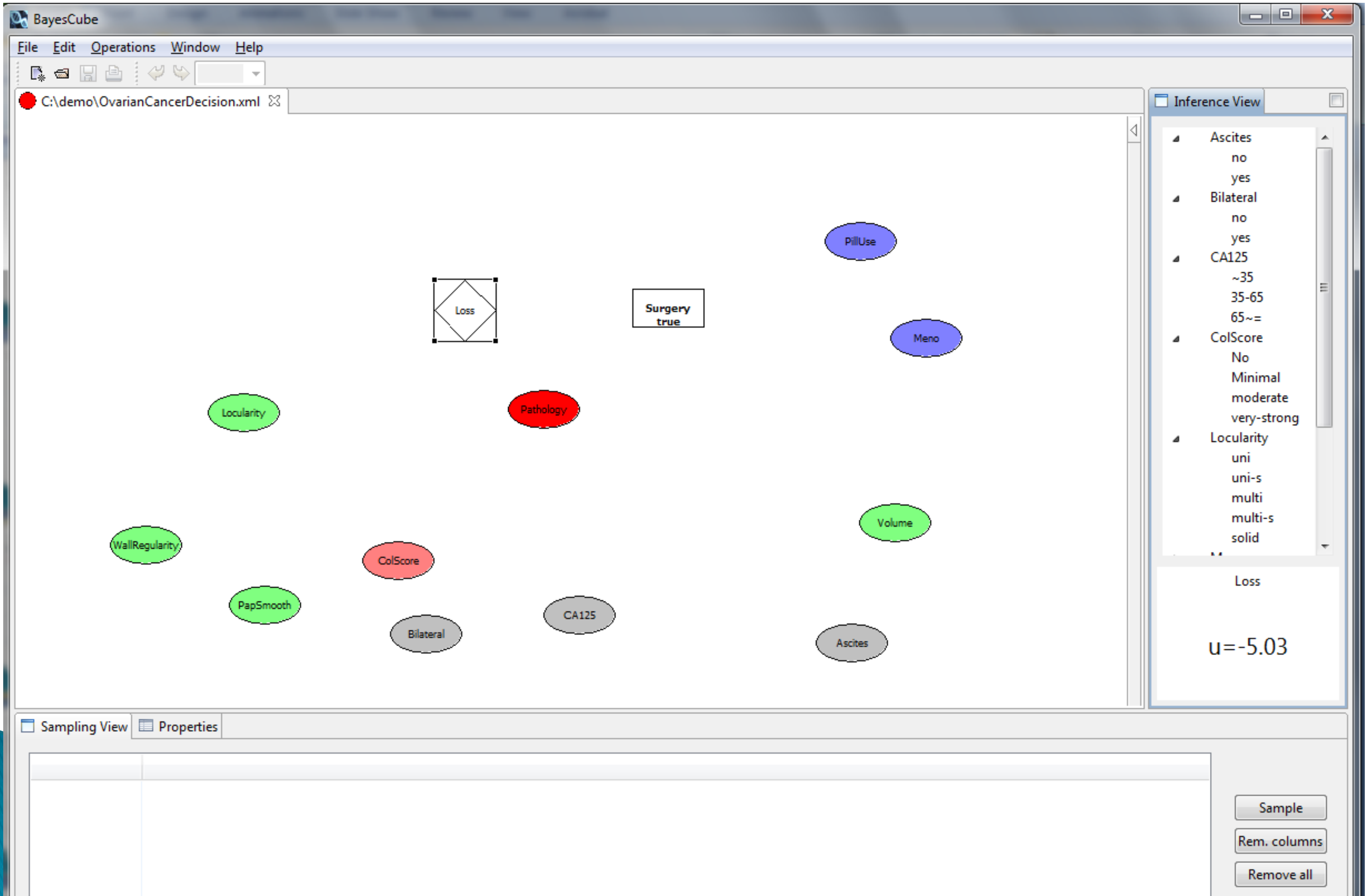
Bayesian network based decision support systems (DSS): Phases of construction

I. Variables/Nodes (concepts)
II. Values (descriptions)
III. Dependencies/Edges
IV. Parameters/Conditional probabilities
V. Utilities/losses
VI. Probabilistic inference
VII. Sensitivity of inference

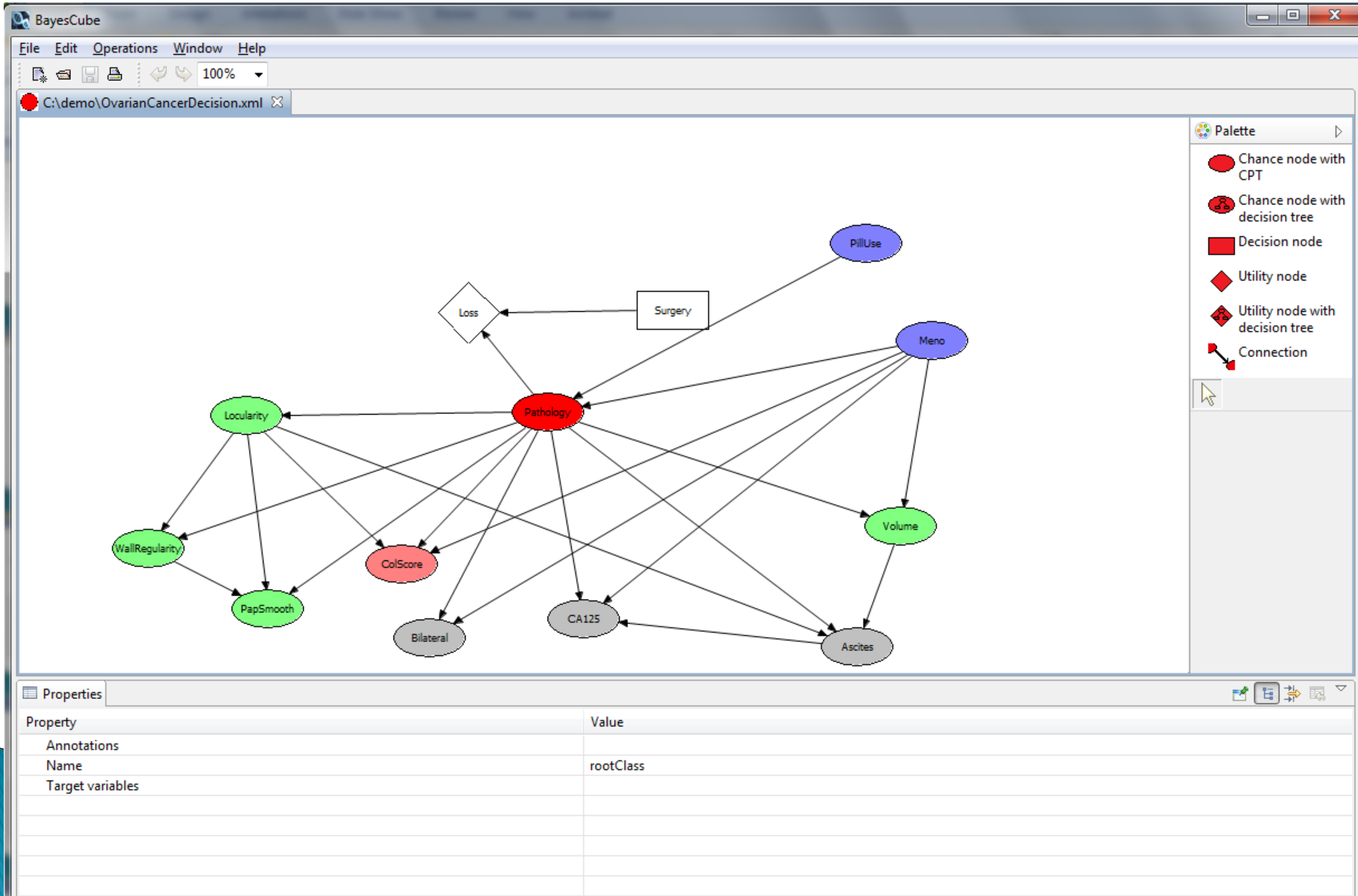
Decision support systems I. variables



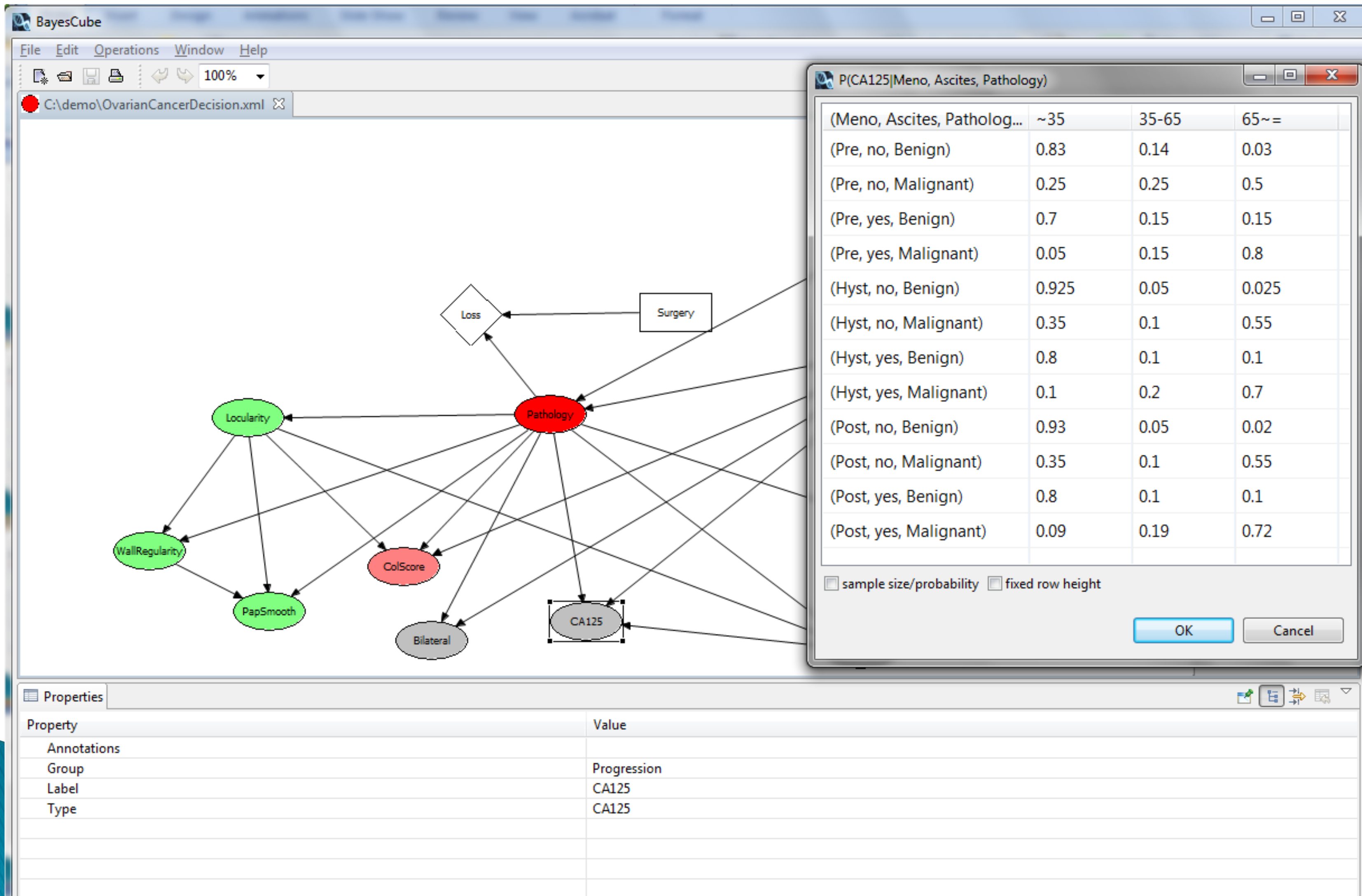
Decision support systems II. values



Decision support systems III. dependencies

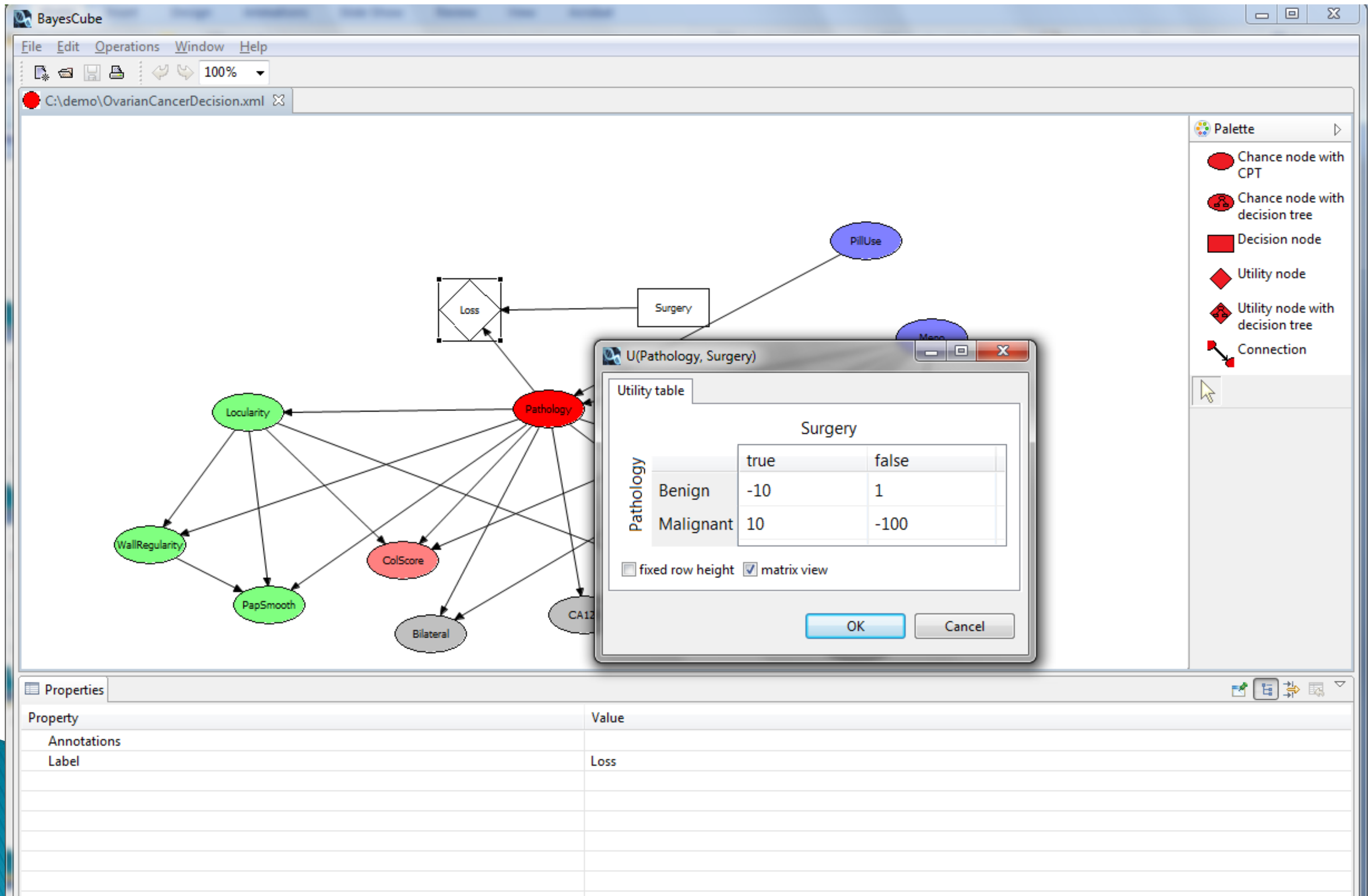


Decision support systems IV. Conditional probabilites

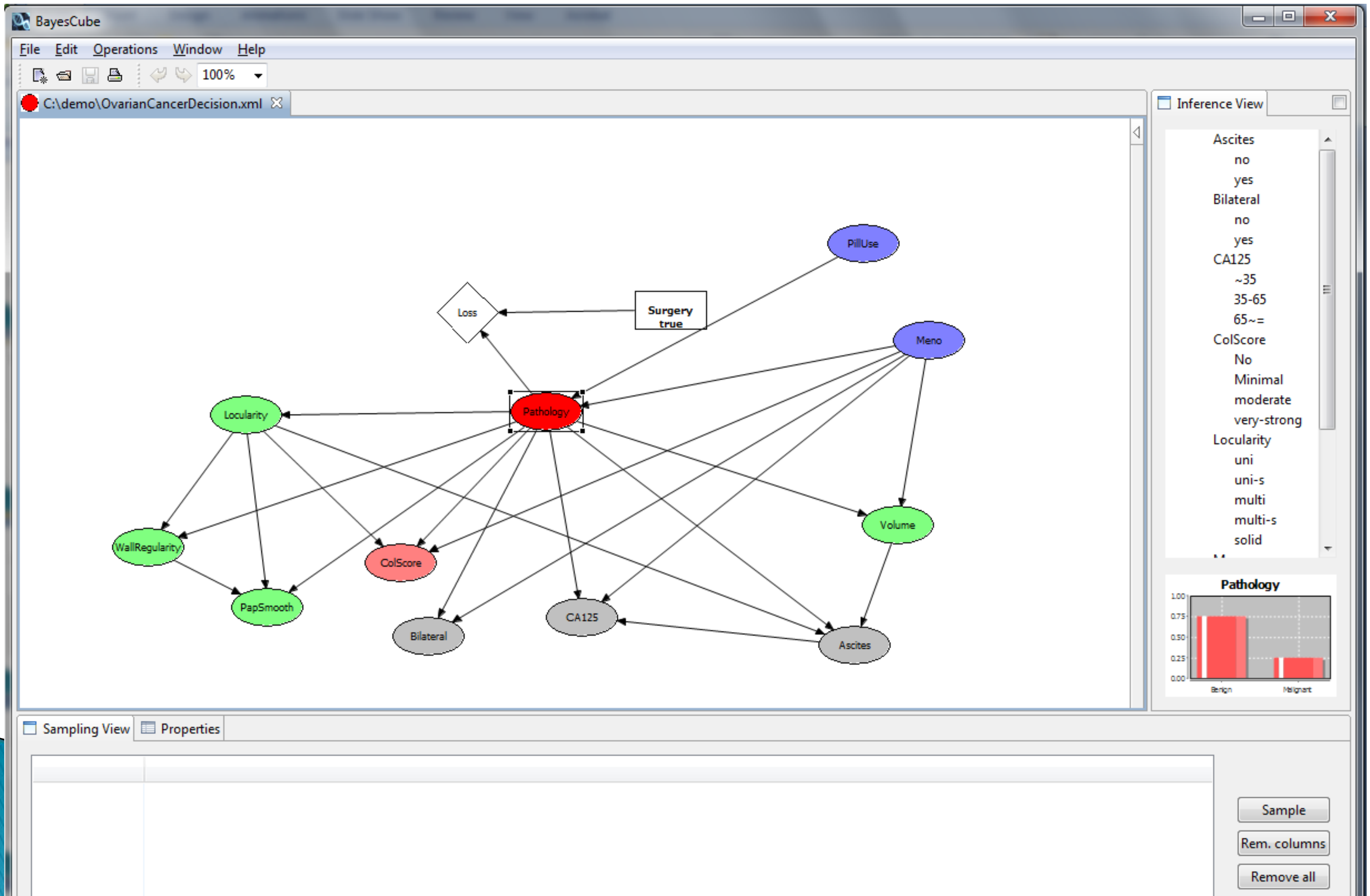


(Meno, Ascites, Patholog...	~35	35-65	65~=
(Pre, no, Benign)	0.83	0.14	0.03
(Pre, no, Malignant)	0.25	0.25	0.5
(Pre, yes, Benign)	0.7	0.15	0.15
(Pre, yes, Malignant)	0.05	0.15	0.8
(Hyst, no, Benign)	0.925	0.05	0.025
(Hyst, no, Malignant)	0.35	0.1	0.55
(Hyst, yes, Benign)	0.8	0.1	0.1
(Hyst, yes, Malignant)	0.1	0.2	0.7
(Post, no, Benign)	0.93	0.05	0.02
(Post, no, Malignant)	0.35	0.1	0.55
(Post, yes, Benign)	0.8	0.1	0.1
(Post, yes, Malignant)	0.09	0.19	0.72

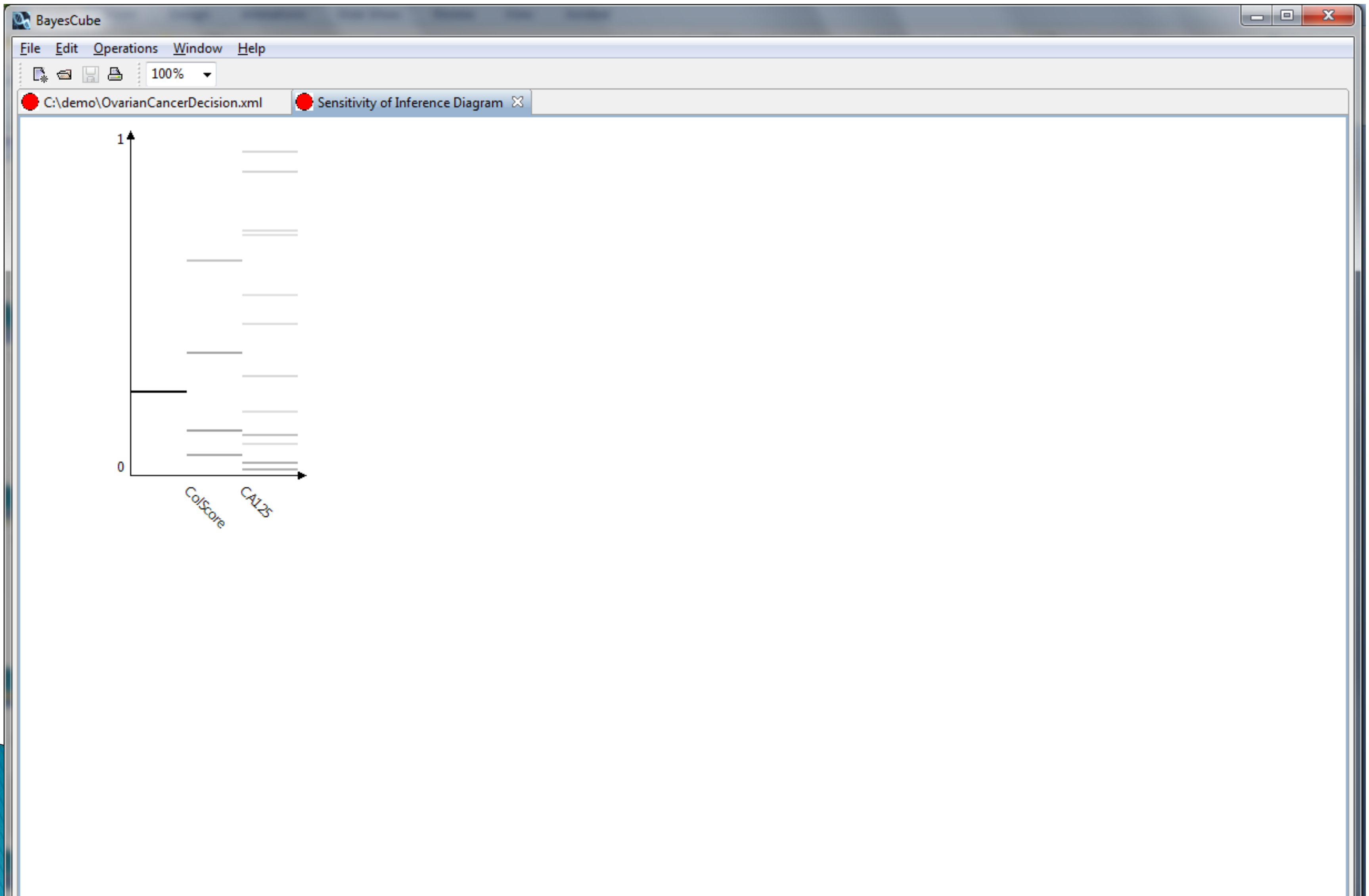
Decision support systems V. utilities/losses



Decision support systems VI. inference



Decision support systems VII. Sensitivity of inference



Uncertainty

Let action A_t = leave for airport $_t$ minutes before flight
Will A_t get me there on time?

Problems:

1. partial observability (road state, other drivers' plans, etc.)
2. noisy sensors (traffic reports)
3. uncertainty in action outcomes (flat tire, etc.)
4. immense complexity of modeling and predicting traffic

Hence a purely logical approach either

1. risks falsehood: "A_{25} will get me there on time", or
2. leads to conclusions that are too weak for decision making:

"A_{25} will get me there on time if there's no accident on the bridge and it doesn't rain and my tires remain intact etc etc."

(A_{1440} might reasonably be said to get me there on time but I'd have to stay overnight in the airport …)

Methods for handling uncertainty

- Default or nonmonotonic logic:
  - Assume my car does not have a flat tire
  - Assume A_{25} works unless contradicted by evidence
- Issues: What assumptions are reasonable? How to handle contradiction?
-
- Rules with fudge factors:
  - A_{25} $\mapsto_{0.3}$ get there on time
  - *Sprinkler* $\mapsto_{0.99}$ *WetGrass*
  - *WetGrass* $\mapsto_{0.7}$ *Rain*
- Issues: Problems with combination, e.g., *Sprinkler* causes *Rain*??
-
- Probability
  - Model agent's degree of belief
  - Given the available evidence,
  - A_{25} will get me there on time with probability 0.04

Probability

Probabilistic assertions summarize effects of

laziness: failure to enumerate exceptions, qualifications, etc.

ignorance: lack of relevant facts, initial conditions, etc.

Subjective (personal, Bayesian) probability (belief):

- Probabilities relate propositions to agent's own state of knowledge

  e.g., $P(A_{25} \mid \text{no reported accidents}) = 0.06$

These are not assertions about the world

Probabilities of propositions change with new evidence:

e.g., $P(A_{25} \mid \text{no reported accidents, 5 a.m.}) = 0.15$

Making decisions under uncertainty

Suppose I believe the following:

$P(A_{25}$ gets me there on time $| \ldots) = 0.04$
$P(A_{90}$ gets me there on time $| \ldots) = 0.70$
$P(A_{120}$ gets me there on time $| \ldots) = 0.95$
$P(A_{1440}$ gets me there on time $| \ldots) = 0.9999$

- Which action to choose?
- Depends on my preferences for missing flight vs. time spent waiting, etc.
  - Utility theory is used to represent and infer preferences
  - Decision theory = probability theory + utility theory

Interpretations of probability

- Sources of uncertainty
  - inherent uncertainty in the physical process;
  - inherent uncertainty at macroscopic level;
  - ignorance;
  - practical omissions;
- Interpretations of probabilities:
  - combinatoric;
  - physical propensities;
  - frequentist;
  - personal/subjectivist;
  - instrumentalist;
- The three „as if” theorems:
  - Uncertainty by probabilities
  - Preferences by utility function
  - Optimal action by maximum expected utility principle

$$\lim_{N\to\infty}\frac{N_A}{N}=\lim_{N\to\infty}\hat{p}_N(A)=p(A)\,?\,p(A|\xi)$$

Note:

• Axioms in probability theory are the same (Kolmogorov)

• “Independence” and convergence of frequencies are empirical observations (e.g., „laws of large numbers” are consequences of some assumptions about independencies).

A chronology

- [1713] Ars Conjectandi (The Art of Conjecture), Jacob Bernoulli
  - **Subjectivist interpretation** of probabilities
- [1718] The Doctrine of Chances, Abraham de Moivre
  - the first textbook on probability theory
  - **Forward predictions**
    - „given a specified number of white and black balls in an urn, what is the probability of drawing a black ball?"
    - his own death
- [1764, posthumous] Essay Towards Solving a Problem in the Doctrine of Chances, Thomas Bayes
  - **Backward questions**: „given that one or more balls has been drawn, what can be said about the number of white and black balls in the urn"
- [1812], Théorie analytique des probabilités, Pierre-Simon Laplace
  - General Bayes rule
- [1921]: **Correlation and causation**, S. Wright's diagrams
- -1950 **Frequentist statistics**
  - Ronald A. Fisher (J. Neyman and E. Pearson)
    - [Bayesianism is a] „fallacious rubbish"
    - His own approach was „Fiducial inference" ~ Bayesian statistics
    - He used informed priors in genetics

A chronology (cont'd)

- [1937], "La prévision: ses lois logiques, ses sources subjectives", B. de Finetti
  - Exchangeability (instead of independency)
- [1939] "Theory of probability„, Harold Jeffreys
- 1950-: „**Bayesian" statistics** (as opposed to the „frequentist" school
  - I.J. Good, B.O. Koopman, Howard Raiffa, Robert Schlaifer and Alan Turing
- [1979] Conditional Independence in Statistical Theory, A.P. Dawid
  - Axiomatization of indepencies in **multivariate** distributions
- [1982] The decomposition of a multivariate distribution, S.Lauritzen
- [1988] Bayesian networks, J.Pearl
  - Representation of independencies
- [1989] Exact general inference methods, S. Lauritzen

- … Markov Chain Monte Carlo methods – GPGPUs…

Bayes–omics

- **Thomas Bayes (c. 1702 – 1761)**
- Bayesian probability
- Bayes’ rule
- Bayesian statistics
- Bayesian decision
- Bayesian model averaging

- Bayesian networks
- Bayes factor
- Bayes error
- Bayesian „communication”
- ...

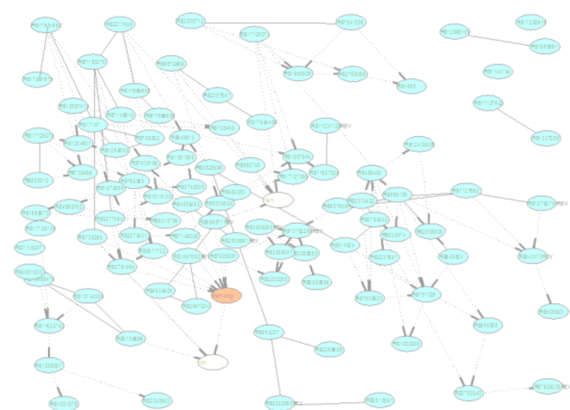
$$p(Cause \mid Effect) \propto p(Effect \mid Cause)\, p(Cause)$$

$$p(\text{Model} \mid \text{Data}) \propto p(\text{Data} \mid \text{Model})\, p(\text{Model})$$

$$a^* = \arg\max_i \sum_j U(o_j)\, p(o_j \mid a_i)$$

$$p(prediction \mid data) =$$

$$= \sum_i p(pred. \mid Model_i)\, p(Model_i \mid data)$$



Probability theory: concepts for the course

- Joint distribution
- Conditional probability
- Independence, conditional independence
- Bayes rule
- Marginalization/Expansion
- Chain rule
- Expectation, variance

Syntax

- Basic element: random variable
- Similar to propositional logic: possible worlds defined by assignment of values to random variables.
- Boolean random variables
- e.g., *Cavity* (do I have a cavity?)
-
- Discrete random variables
- e.g., *Weather* is one of $<$*sunny,rainy,cloudy,snow*$>$
- Domain values must be exhaustive and mutually exclusive
- Elementary proposition constructed by assignment of a value to a
- random variable: e.g., $Weather = sunny$, $Cavity = false$
- (abbreviated as $\neg cavity$)
- Complex propositions formed from elementary propositions and standard logical connectives e.g., $Weather = sunny \vee Cavity = false$

Syntax

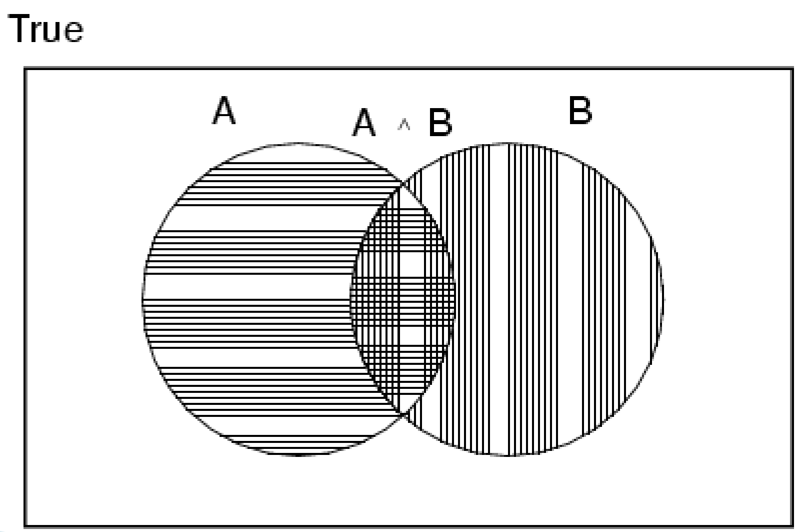
- Atomic event: A complete specification of the state of the world about which the agent is uncertain
- E.g., if the world consists of only two Boolean variables *Cavity* and *Toothache*, then there are 4 distinct atomic events:

$$Cavity = false \wedge Toothache = false$$
$$Cavity = false \wedge Toothache = true$$
$$Cavity = true \wedge Toothache = false$$
$$Cavity = true \wedge Toothache = true$$

- Atomic events are mutually exclusive and exhaustive

Axioms of probability

- For any propositions *A, B*
-
  - $0 \leq P(A) \leq 1$
  - $P(true) = 1$ and $P(false) = 0$
  - $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$
  -



Prior probability

- Prior or unconditional probabilities of propositions
- e.g., $P(Cavity = \text{true}) = 0.1$ and $P(Weather = \text{sunny}) = 0.72$ correspond to belief prior to arrival of any (new) evidence
- Probability distribution gives values for all possible assignments:
- $\mathbf{P}(Weather) = \langle 0.72, 0.1, 0.08, 0.1 \rangle$ (normalized, i.e., sums to 1)
- Joint probability distribution for a set of random variables gives the probability of every atomic event on those random variables
- $\mathbf{P}(Weather, Cavity) =$ a 4×2 matrix of values:

Weather =	sunny	rainy	cloudy	snow
Cavity = true	0.144	0.02	0.016	0.02
Cavity = false	0.576	0.08	0.064	0.08

Conditional probability

- Conditional or posterior probabilities
- e.g., $P(cavity \mid toothache) = 0.8$

  i.e., given that *toothache* is all I know

- (Notation for conditional distributions:
- $\mathbf{P}(Cavity \mid Toothache)$ = 2-element vector of 2-element vectors)

- If we know more, e.g., *cavity* is also given, then we have
- $P(cavity / toothache, cavity) = 1$

- New evidence may be irrelevant, allowing simplification, e.g.,
- $P(cavity / toothache, sunny) = P(cavity \mid toothache) = 0.8$
- This kind of inference, sanctioned by domain knowledge, is crucial
-

Conditional probability

- Definition of conditional probability:
- $P(a \mid b) = P(a \wedge b) / P(b)$ if $P(b) > 0$
-
- Product rule gives an alternative formulation:
- $P(a \wedge b) = P(a \mid b) P(b) = P(b \mid a) P(a)$
-
- A general version holds for whole distributions, e.g.,
- $\mathbf{P}(Weather,Cavity) = \mathbf{P}(Weather \mid Cavity)\, \mathbf{P}(Cavity)$
- (View as a set of 4 × 2 equations, not matrix mult.)
-
- Chain rule is derived by successive application of product rule:
- $\mathbf{P}(X_1, \ldots, X_n) = \mathbf{P}(X_1,\ldots,X_{n-1})\, \mathbf{P}(X_n \mid X_1,\ldots,X_{n-1})$

$$= \mathbf{P}(X_1,\ldots,X_{n-2})\, \mathbf{P}(X_{n-1} \mid X_1,\ldots,X_{n-2})\, \mathbf{P}(X_n \mid X_1,\ldots,X_{n-1})$$

$$= \ldots$$

$$= \pi_{i=1}\hat{}n\; \mathbf{P}(X_i \mid X_1, \ldots ,X_{i-1})$$

Inference by enumeration

- Every question about a domain can be answered by the joint distribution.

- Start with the joint probability distribution:
-

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

- For any proposition ϕ, sum the atomic events where it is true: $P(\phi) = \Sigma_{\omega:\omega \models \phi} P(\omega)$
-

Inference by enumeration

- Start with the joint probability distribution:
-

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

- For any proposition φ, sum the atomic events where it is true: $P(\phi) = \Sigma_{\omega:\omega \models \phi} P(\omega)$
-
- P(*toothache*) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2
-

Inference by enumeration

- Start with the joint probability distribution:
-

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

- For any proposition φ, sum the atomic events where it is true: $P(\phi) = \Sigma_{\omega:\omega \models \phi} P(\omega)$
-
- P(*toothache*) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2
-

Inference by enumeration

- Start with the joint probability distribution:
-

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

- Can also compute conditional probabilities:
-

$$P(\neg cavity \mid toothache) = \frac{P(\neg cavity \wedge toothache)}{P(toothache)}$$

$$= \frac{0.016+0.064}{0.108 + 0.012 + 0.016 + 0.064}$$

$$= 0.4$$

Normalization

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

- Denominator can be viewed as a normalization constant α
-

P(*Cavity* | *toothache*) = α, **P**(*Cavity,toothache*)
= α, [**P**(*Cavity,toothache,catch*) + **P**(*Cavity,toothache*,¬ *catch*)]
= α, [<0.108,0.016> + <0.012,0.064>]
= α, <0.12,0.08> = <0.6,0.4>

General idea: compute distribution on query variable by fixing evidence variables and summing over hidden variables

Inference by enumeration, contd.

Typically, we are interested in the posterior joint distribution of the query variables **Y** given specific values **e** for the evidence variables **E**

Let the hidden variables be **H** = **X** - **Y** - **E**

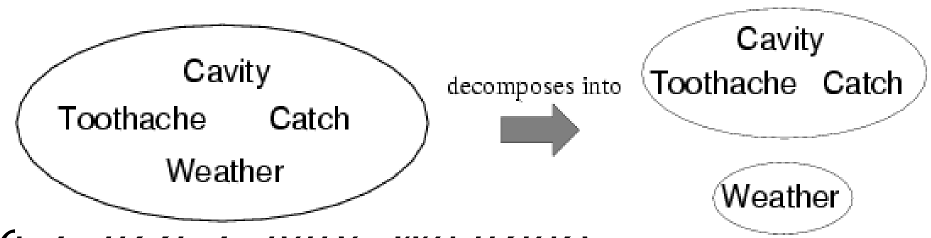
Then the required summation of joint entries is done by summing out the hidden variables:

$\mathbf{P}(\mathbf{Y} \mid \mathbf{E} = \mathbf{e}) = \alpha\mathbf{P}(\mathbf{Y},\mathbf{E} = \mathbf{e}) = \alpha\Sigma_h\mathbf{P}(\mathbf{Y},\mathbf{E}= \mathbf{e}, \mathbf{H} = \mathbf{h})$

- The terms in the summation are joint entries because **Y**, **E** and **H** together exhaust the set of random variables
- Obvious problems:
  1. Worst-case time complexity $O(d^n)$ where d is the largest arity
  2. Space complexity $O(d^n)$ to store the joint distribution
  3. How to find the numbers for $O(d^n)$ entries?

Independence, conditional independence

- *A* and *B* are independent iff
  $\mathbf{P}(A/B) = \mathbf{P}(A)$ or $\mathbf{P}(B/A) = \mathbf{P}(B)$ or $\mathbf{P}(\mathrm{A, B}) = \mathbf{P}(A)\,\mathbf{P}(B)$



$\mathbf{P}(Toothache, Catch, Cavity, Weather)$
$= \mathbf{P}(Toothache, Catch, Cavity)\,\mathbf{P}(Weather)$

- 32 entries reduced to 12; for *n* independent biased coins, $O(2^n) \rightarrow O(n)$
-

- Absolute independence powerful but rare
- *A* and *B* are conditionally independent iff
  $\mathbf{P}(A/B) = \mathbf{P}(A)$ or $\mathbf{P}(B/A) = \mathbf{P}(B)$ or $\mathbf{P}(\mathrm{A, B|C}) = \mathbf{P}(\mathrm{A|C})\,\mathbf{P}(\mathrm{B|C})$

Bayes rule

An algebraic triviality

$$p(X \mid Y) = \frac{p(Y \mid X)\, p(X)}{p(Y)} = \frac{p(Y \mid X)\, p(X)}{\sum_X p(Y \mid X)\, p(X)}$$

A scientific research paradigm

$$p(Model \mid Data) \propto p(Data \mid Model)\, p(Model)$$

A practical method for inverting causal knowledge to diagnostic tool.

$$p(Cause \mid Effect) \propto p(Effect \mid Cause) \times p(Cause)$$

Decision theory= probability theory+utility theory

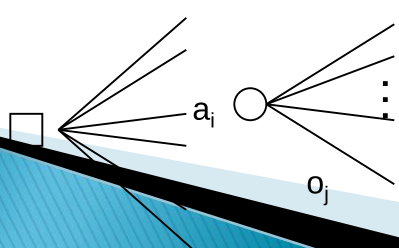
- Decision situation:
  - Actions a_i
  - Outcomes o_j
  - Probabilities of outcomes $p(o_j \mid a_i)$
  - Utilities/losses of outcomes $U(o_j \mid a_i)$
    - QALY, micromort

$$EU(a_i) = \sum_j U(o_j \mid a_i) p(o_j \mid a_i)$$

  - Maximum Expected Utility Principle (MEU)
    - Best action is the one with maximum expected utility

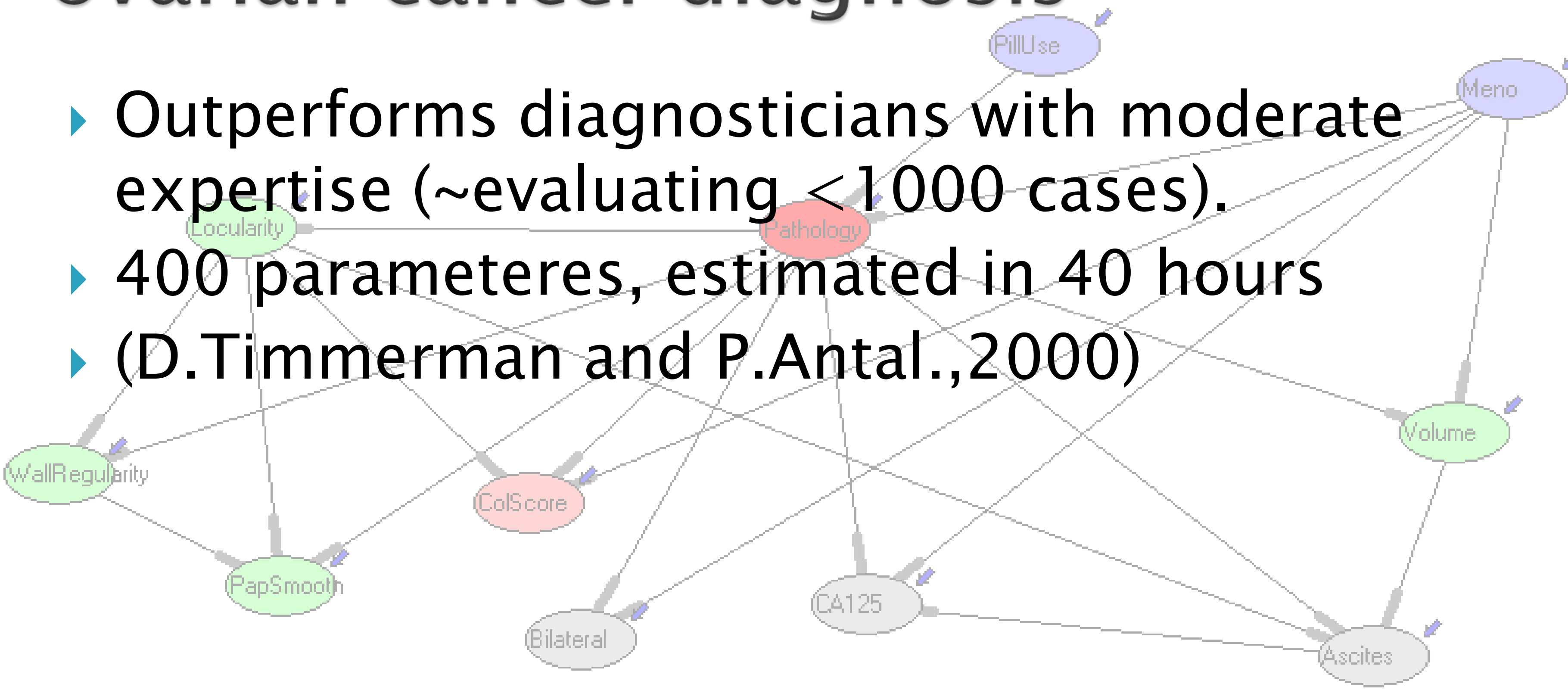
$$a^* = \arg\max_i EU(a_i)$$

Actions a_i (which experiment)	Outcomes (e.g. dataset)	Probabilities	Utilities, costs	Expected utilities
		$P(o_j \mid a_i)$	$U(o_j)$, $C(a_i)$	$EU(a_i) = \sum P(o_j \mid a_i) U(o_j \mid a_i)$
		⋮	⋮	



Medical decision support ovarian cancer diagnosis

- Outperforms diagnosticians with moderate expertise (~evaluating <1000 cases).
- 400 parameteres, estimated in 40 hours
- (D.Timmerman and P.Antal.,2000)



BAYES CUBE (~BAYES EYE)

http://mitpc40.mit.bme.hu/~balazs/BayesCube_131014_win64_alpha.zip

Summary

- Probability is a rigorous formalism for uncertain knowledge.
- The subjective/Bayesian interpretation of probabilities avoids the necessity of repeatability.
- Joint probability distribution specifies probability of every atomic event.
- Queries can be answered by summing over atomic events.

- **Suggested reading:**
  - Malakoff: Bayes Offers a `New' Way to Make Sense of Numbers, Science, 1999
  - Efron: Bayes' Theorem in the 21st Century, Science, 2013
  - Charniak: Bayesian networks without tears, 1991