Probabilistic Foundations for Combining Information

Kenneth Baclawski
Northeastern University
March 4, 2005

Happy Exelauno Day!

Sources of Uncertainty

- Measurement (sensor) error
- Nondeterministic processes
- Unmodeled variables (ontological commitment)
- Subjective probabilities (judgement, belief, trust, etc.)

Theories of Uncertainty

- Techniques for representing uncertainty
  - Classical set theory (e.g., interval math)
  - Probability theory
  - Fuzzy set theory
  - Fuzzy measure theory
  - Rough set theory
- Every theory has a means of combining uncertainty information.
- Only probability theory has an emprical basis.

Stochastic Models and Processes

- A *stochastic model* (or theory) is a set of random variables.
- The model is completely specified by the *joint probability distribution* (JPD) of the RVs.
- A *compound stochastic model* is constructed by using the result of one RV as a parameter for another RV.
- As the number of RVs increases, the complexity of the JPD increases rapidly.
- A *stochastic process* is a sequence of (possibly dependent) stochastic models.

Measurement

- Fundamental problem in science and engineering
- Physical constants
  - Speed of light, Hubble constant,...
- Dynamic systems
  - Position of an asteroid, incoming missle,...
- Probability distributions
  - Probability of benefit due to a drug
- Stochastic dynamic systems
  - CPU load, network congestion,...

Terminology for Combining Information

- Meta-Analysis or Quantitative Research Synthesis
  - Social and Behavioral Sciences
  - Medicine
- Pooling of Results or Creating an Overview
  - Medicine
- Reviewing the results of research
  - Physics
- Critical evaluation
  - Physical Chemistry
  - Thermochemistry
- Data Fusion or Information Fusion
  - Sensors
  - Military

Example of Combining Information

- Consider the the problem of disease diagnosis when the possibilities are known to be one of these following: concussion, meningitis and tumor.
- Two independent assessments are made:

Doctor	Concussion	Meningitis	Tumor
A	0.7	0.2	0.1
B	0.5	0.3	0.2

- How should these be combined?

Information Combination Theorem

If two measurements *A* and *B* of the *same phenomenon* are *independent* and *consistent* then the combination of the two measurements *C* has the distribution

$$Pr(C=x) = k\, Pr(A=x) Pr(B=x)$$

where *k* is chosen so that the probabilities add to 1.

Doctor	Concussion	Meningitis	Tumor
A	0.7	0.2	0.1
B	0.5	0.3	0.2
Combined	0.81	0.14	0.05

Proof

Two random variables that represent the same phenomenon are combined by conditioning on the event $(A=B)$.

The combined random variable has distribution

$$Pr(C=x) = Pr(A=x,B=x|A=B)$$
$$= Pr(A=x,B=x)/Pr(A=B).$$

By the independence of A and B,

$$Pr(A=x,B=x) = Pr(A=x)Pr(B=x).$$

Therefore,

$$Pr(C=x) = Pr(A=x)Pr(B=x)/Pr(A=B).$$ QED

Paradox?

Doctor	Concussion	Meningitis	Tumor
A	0.9	0.0	0.1
B	0.0	0.9	0.1
Combined	0.0	0.0	1.0

- The seemingly unlikely last alternative now becomes certain!
- Zadeh argued that something is wrong with this. Is there a problem?
- Sherlock Holmes: “When you have eliminated all which is impossible, then whatever remains, however improbable, must be the truth.”

Examining the Hypotheses

- It is important to verify the hypotheses before applying the theorem:
  - The observations must be independent.
  - The observations must be of the same phenomenon.
  - The observations must be consistent.
- Independence may fail because the observations are derived from common data.
- The phenomena may be different due to calibration inconsistencies.
- If all possibilities are eliminated, then the observations are inconsistent.

A priori and A posteriori

Bayesian reasoning:

Prior + Observation produces Posterior

Combining information:

Observation1 + Observation2 produces Combined

Mathematically these are the same.

Frequentist reasoning is a special case of Bayesian where the prior distribution is uniform.

Regression (least squares) is yet another example of the same process.

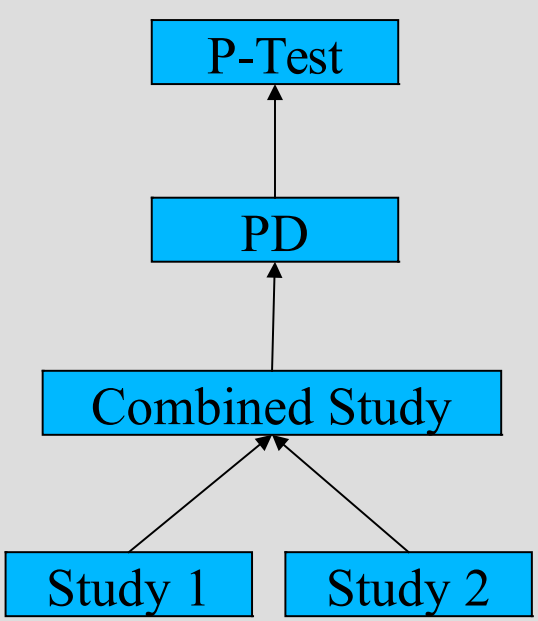
The iterative form of regression is known as the Kalman filter.

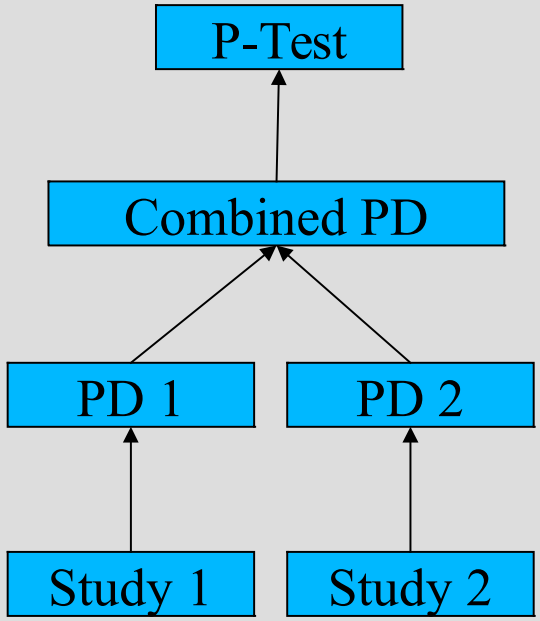
Mathematically these are all the same.

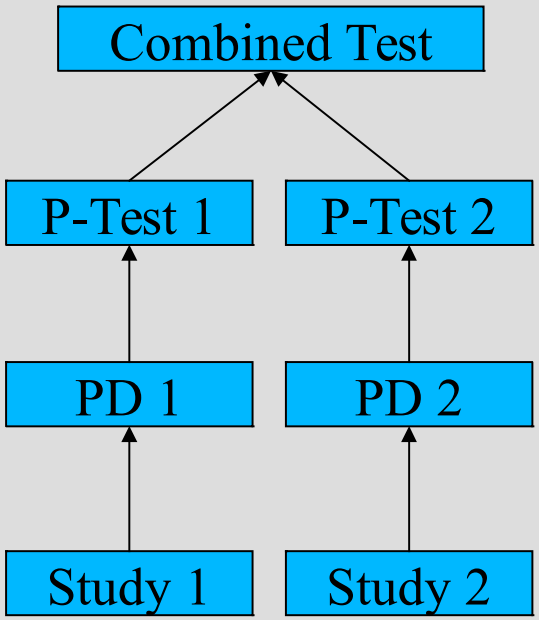
Decision Fusion

Decision fusion is the process of combining decisions rather than probability distributions.

Which one of these is better?







Asprin/MI Study Combination

Studies of the effects of aspirin following MI.

	Aspirin		Placebo		Comparison	
Study	**No. Patients**	**Mortality %**	**No. Patients**	**Mortality %**	**Diff %**	**stderr %**
UK-1	615	7.97	624	10.74	2.77	1.65
CDPA	758	5.80	771	8.30	2.50	1.31
GAMS	317	8.52	309	10.36	1.84	2.34
UK-2	832	12.26	850	14.82	2.56	1.67
PARIS	810	10.49	406	12.81	2.31	1.98
AMIS	2267	10.85	2257	9.70	-1.15	0.90
Total	5599	9.88	5217	10.73	0.86	0.59

One could pool the data at three different levels:

Level 0: At the patient level (as on the **Total** line)

Level 1: At the study level (combining distributions)

Level 2: At the test level (combining P-tests)

Standard Model

Level	Name	Process	Estimation	Product	Medicine
0	Signal Assessment	Identify features	Detection	Signal State	Observation
1	Object Assessment	Identify entities	Attributive State	Entity State	Symptom
2	Situation Assessment	Relationships among entities	Relation	Situation State	Diagnosis
3	Impact Assessment	Evaluation	Game Theory	Situation Utility	Prognosis

There used to be a level 4 dealing with decisions.
It corresponds roughly to “treatment” in medicine.

Continuous Combinations

If two *independent* continuous random variables *A* and *B* measure the *same phenomenon* and have probability densities *f(x)* and *g(x)*, then the density of the combination is

$$\frac{f(x)g(x)}{\int f(y)g(y)dy}$$

If *A* and *B* are independent normally distributed RVs with means *m* and *n* and variances *v* and *w*, then the combined random variable is normal with mean and variance:

$$\frac{wm+vn}{v+w} \quad \text{and} \quad \frac{vw}{v+w}$$

Example of Measurements

For example, suppose that two independent measurements of a temperature are:

$30.5° \pm 0.4°C$, and $30.2° \pm 0.3°C$.

Their combination is: $30.3° \pm 0.24°C$.

The combination is closer to the second measurement because that one is more accurate.

The combination for the aspirin studies is: 1.44 ± 0.50

The combination theorem for normal distributions is the basis for the Kalman filter and modern multi-sensor tracking algorithms.

Examining the Hypotheses

- It is important to verify the hypotheses before applying the theorem:
  - The observations must be independent.
  - The observations must be of the same phenomenon.
  - The observations must be consistent.
- Independence may fail because the observations are derived from common data.
- The phenomena may be different due to calibration inconsistencies.
- If all possibilities are eliminated, then the observations are inconsistent.

Compound Models

A *fixed effects model* assumes that all of the studies are dealing with the same phenomenon. In this case, a normal distribution *Normal(diff, var)*.

To model the differences between the studies, one can use a compound stochastic model:

- First choose a number *b* from *Normal(0,w)*.
- Then perform a study yielding *Normal(diff+b,var)*.

A compound stochastic model is also known as a *random effects model*. Such a model accounts for calibration inconsistencies at the cost of estimating one additional parameter.

For the aspirin studies, a RE model yields this probability distribution:

and b 1.51 ± 0.87

Measuring Distributions

Measuring a Bernoulli (dichotomous) distribution produces in a Binomial distribution

For large samples, the Binomial is close to being a normal distribution.

A normal distribution has two parameters: the mean and the variance.

Measuring a normal distribution produces a bivariate distribution consisting of a t distribution and a χ^2 distribution. The two distributions are independent, so they can be measured separately.

Measuring a Normal Distribution

For large samples, both of these distributions are close to being normal distributions. So the measurement of a normal distribution produces a bivariate normal distribution.

Let X be a random variable from a normally distributed population with mean μ and variance σ^2, For a random sample of N independent measurements of X, the sample mean and variance are a bivariate normal distribution with mean $\begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}$ and variance $\begin{pmatrix} \frac{\sigma^2}{N} & 0 \\ 0 & \frac{\sigma^4}{N-1} \end{pmatrix}$.

Simulation

- Easily programmed
- Improves understanding
- Helps determine sensitivity and robustness
- Simulating the Aspirin/MI example showed the following:
  - Even with no difference between the two groups, the estimated difference for the simulation will sometimes be as high as the estimated difference for the actual studies.
  - The P-value is about 0.04, so one would expect this to happen once in 25 simulations.

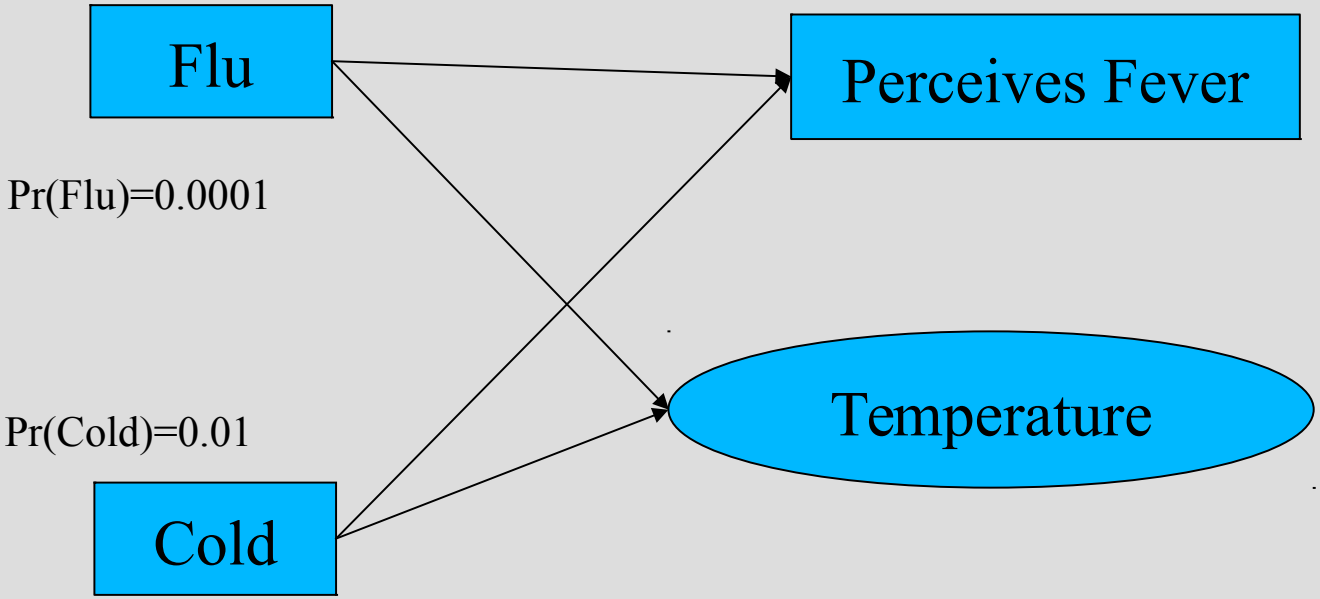
Examining the Hypotheses

- It is important to verify the hypotheses before applying the theorem:
  - The observations must be independent.
  - The observations must be of the same phenomenon.
  - The observations must be consistent.
- Independence may fail because the observations are derived from common data.
- The phenomena may be different due to calibration inconsistencies.
- If all possibilities are eliminated, then the observations are inconsistent.

Bayesian Networks

- Efficient graphical mechanism for representing stochastic models with dependencies among the RVs.
- A Bayesian Network (BN) is a directed graph in which:
  - A node corresponds to a RV.
  - An edge represents a stochastic dependency.
  - The conditional probability distribution (CPD) at each RV conditioned on all incoming RVs.
- It is commonly assumed that the RVs are discrete.

Bayesian Network Specification

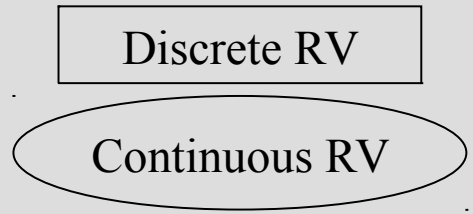


	Perceives Fever TRUE	FALSE
Flu = F & Cold = F	0.01	0.99
Flu = F & Cold = T	0.1	0.9
Flu = T & Cold = F	0.9	0.1
Flu = T & Cold = T	0.95	0.05

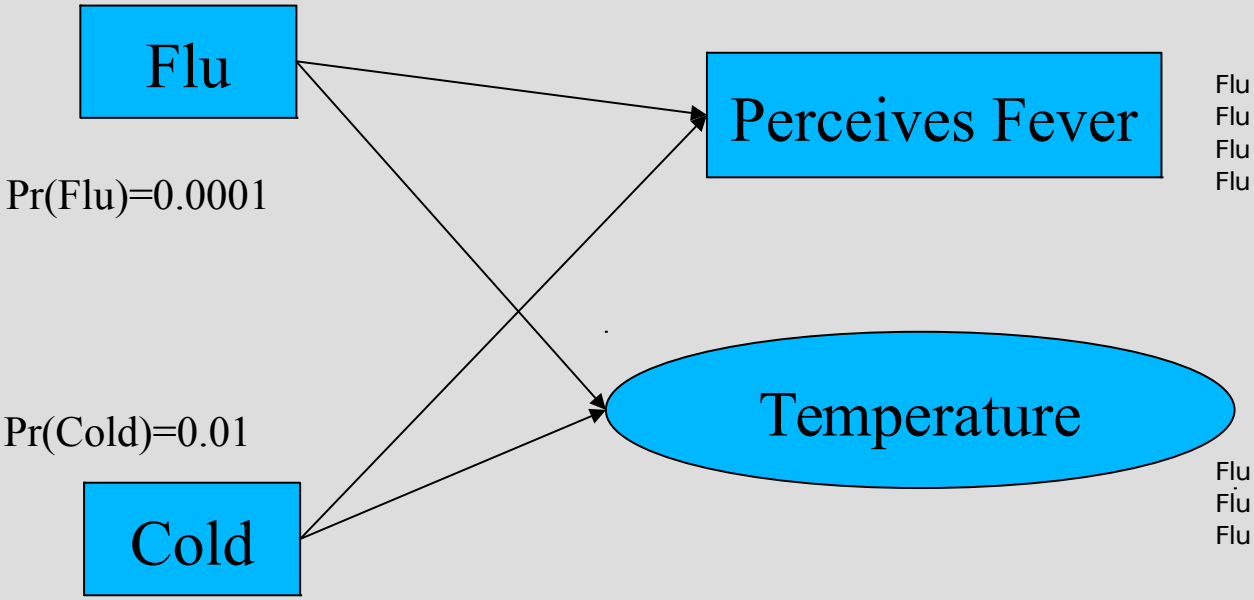
	Temperature Mean	Std Dev
Flu = F & Cold = F	37	0.5
Flu = F & Cold = T	37.5	1
Flu = T & Cold = F	39	2

Required CPDs:

1. Perceives Fever given Flu and/or Cold.
2. Temperature given Flu and/or Cold.
3. Probability of Flu (unconditional).
4. Probability of Cold (unconditional).



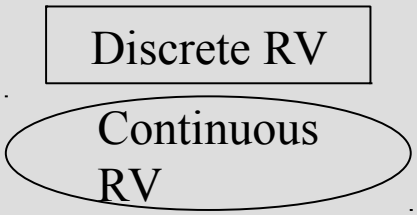
Bayesian Network Specification



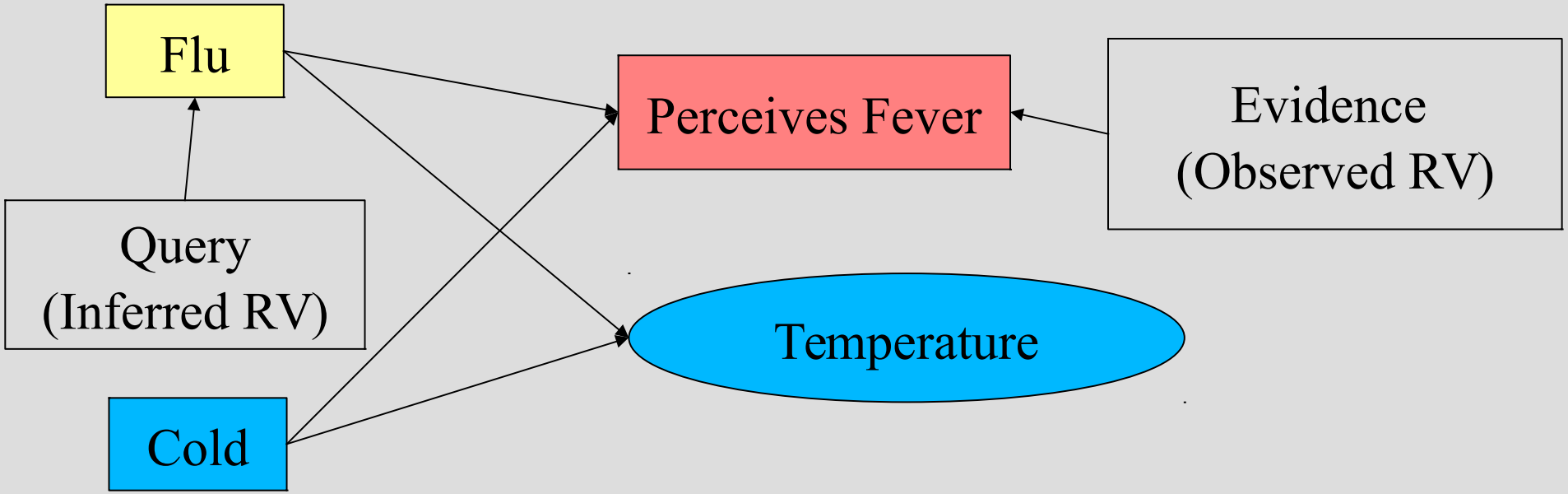
	Perceives Fever TRUE	FALSE
Flu = F & Cold = F	0.01	0.99
Flu = F & Cold = T	0.1	0.9
Flu = T & Cold = F	0.9	0.1
Flu = T & Cold = T	0.95	0.05

	Temperature Mean	Std Dev
Flu = F & Cold = F	37	0.5
Flu = F & Cold = T	37.5	1
Flu = T & Cold = F	39	2

The joint probability distribution is the product of all the CPDs. The probability distribution of any RV (or set of RVs) is obtained by computing the marginal distribution.

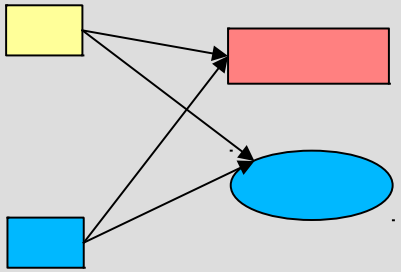


Bayesian Network Inference

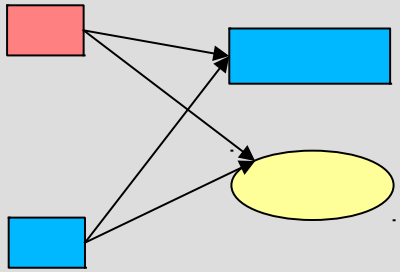


Inference is performed by observing some RVs (evidence) and computing the distribution of the RVs of interest (query).
The evidence can be a value or a probability distribution.
The BN combines the evidence probability distributions even when there are probabilistic dependencies.

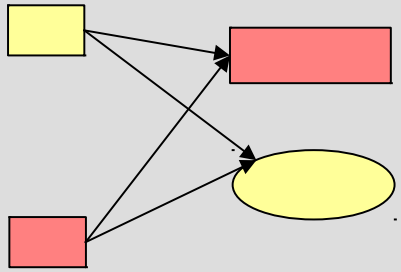
Bayesian Network Inference



Diagnostic Inference



Causal Inference



Mixed Inference

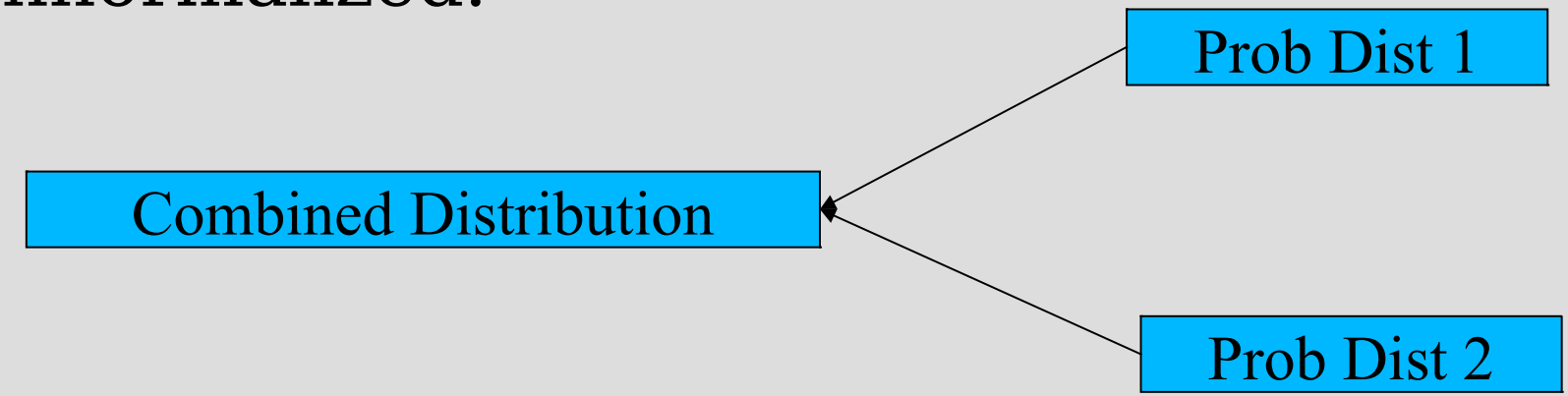
Evidence

Query

- Inference in the same direction as the edges is called ***causal.*** Also called ***deductive.***
- Inference against the direction of the edges is called ***diagnostic.*** Also called ***abductive.***
- Inference in both directions is called ***mixed inference.***
- The evidence is combined with the *prior* distribution. The answer is the marginal distribution of the query RVs.

BNs and Meta-Analysis

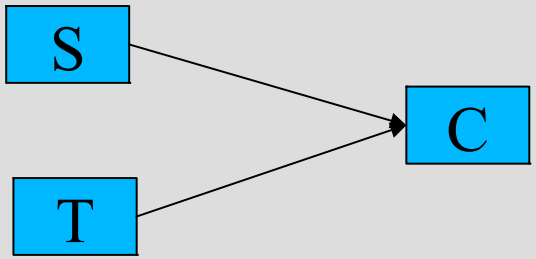
- BN inference combines evidence with the prior distribution.
- Information combination can be represented as a BN. However, the CPDs must be unnormalized.



If this BN is normalized, then the two independent probability distributions become dependent!

Berkson's Paradox

- This paradox is also known as *selection bias*, or "explaining away" in Artificial Intelligence.
- Suppose that a genetic condition C is correlated with two SNPs S and T. The probabilities of S and T are 0.1 and 0.3, and they are independent.



S	T	Pr(C)
Y	Y	1.0
Y	N	0.9
N	Y	0.8
N	N	0.1

- If one knows that a patient has C, then S and T are no longer independent. For example:
  $Pr(S \text{ and } T|C) = 0.081$, but $Pr(S|C)Pr(T|C) = 0.25$
  $Pr(S|T,C) = 0.122$, but $P(S|C) = 0.25$
- Within the population of persons with C, if one has T, then it is *less* likely that one also has S.

Selection Bias

- Meta-analysis is generally based on published research. This can result in implicit selection bias:
  - Publication bias: Only significant results are published
  - Reporting bias: Within published work, only significant results are reported.
  - Retrieval bias: Difficulty in finding relevant research.
  - Stopping bias: Ending study when significance is achieved.
- Approaches for dealing with selection bias:
  - Draw the funnel display
  - Compute the minimum number of unpublished studies necessary to overturn the conclusion.

Types of Bayesian Network

- BNs can be discrete, continuous or hybrid.
  - Discrete is the most commonly supported.
  - Connectionist (neural) networks are examples of continuous BNs.
  - Hybrid BNs:
    - From discrete to continous: mixed Gaussian
    - From continous to discrete: connectionist classifiers

BN Inference Techniques

- Inference is computationally expensive as the size of the BN increases.
- Exact inference
  - Clique
  - OOBN
- Approximate
  - Propagation
  - Monte Carlo (e.g., Gibbs sampling)

BN Software Tools

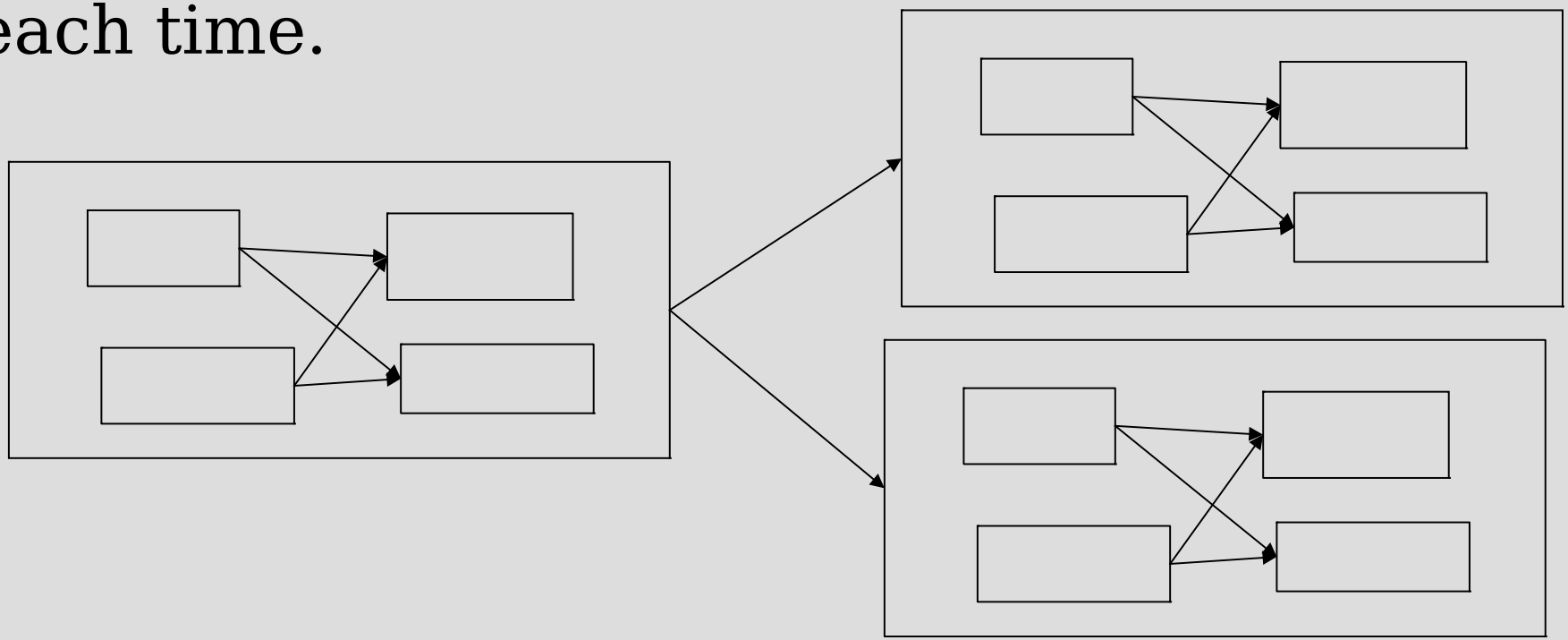
- Many software tools are available, both commercial and free.
  - Commercial: Netica, Hugin, Analytic
  - Free: Smile, Genie, Java Bayes, MSBN
- See www.ai.mit.edu/~murphyk/Bayes/bnsoft.html
- These tools often assume that the RVs are discrete.

BN Development

- Select the important variables.
- Specify the dependencies.
- Specify the CPDs.
- Evaluate.
- Iterate over the steps above.

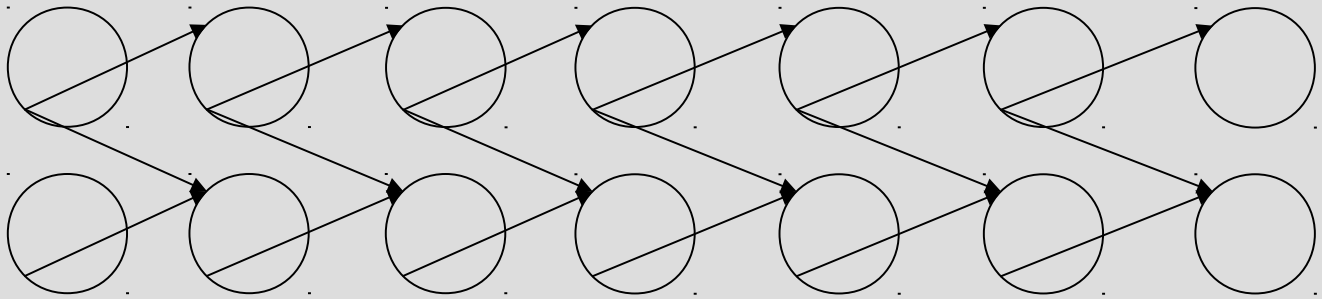
Large Scale BN Development

- The BN is built from smaller parametrized units that have been carefully checked.
- A unit can be used (instantiated) many times within a larger BN, with different parameters each time.



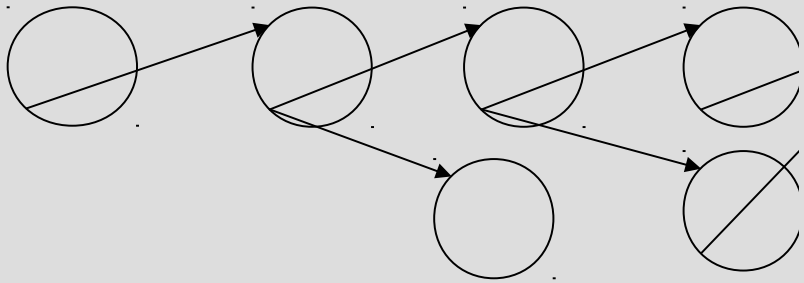
Stochastic Dynamic Systems

- A BN can model a stochastic dynamic system. The structure of the BN does not vary, but nodes represent states at different times. For example, Markov chains.



Structure-Dynamic BN

- A structure-dynamic BN varies its structure in time.
- Much less is known about the behavior of such BNs.



Semantic Web

- Life science professionals use the Web, but the Web is designed for human interaction, not automated processing.
- One can easily access information, but one cannot easily integrate different sources or add analysis tools.
- The Semantic Web addresses these issues by representing the meaning (semantics) of data on the Web.
- Information is annotated using meta-data expressed in the Web Ontology Language (OWL).
- The Semantic Web will also include logical reasoning and retrieval facilities.

Semantic Web Scenario

To illustrate a possible use of the Semantic Web, consider the following hypothetical scenario. A scientist would like to determine whether a novel protein (protein Y) interacts with p21-activated kinase 1 (PAK1). To answer this question, the scientist first goes to the kinase pathway database `kinasedb.ontology.ims.u-tokyo.ac.jp` to obtain a list of all known proteins that interact with PAK1 (e.g., MYLK and BMX). The scientist then writes a set of rules to determine whether the protein Y is structurally similar to any PAK1-interacting proteins. After applying the rules using a Semantic Web enabled protein interaction server, one hit, protein X, is found. This leads to the prediction that protein Y will interact with PAK1. Next, the scientist wishes to relate this interacting pair to a particular signaling pathway. As all the tools used refer to the same ontologies and terminology defined through GO, the researcher can easily map this interacting pair to a relevant signaling pathway obtained from a Semantic Web enabled pathway server. During the information foraging described above, the scientist constantly used literature databases to read relevant articles. Despite the tremendous growth of more than 5000 articles each week, the biologist still managed to quickly find the relevant articles by using an ontology-based search facility.

Bayesian Web

- The Semantic Web is good for logical reasoning.
- It does not support empirical or stochastic reasoning.
- The Bayesian Web (BW) is a proposal to deal with this issue.
  - The BW is built on the SW.
  - Both logical and probabilistic reasoning are supported.
  - Stochastic reasoning is based on BNs.

Bayesian Web facilities

- Common interchange format
- Ability to refer to common variables (diseases, drugs, ...)
- Context specification
- Authentication and trust
- Open hierarchy of probability distribution types
- Component based construction of BNs
- BN inference engines
- Meta-analysis services

Bayesian Web capabilities

- Use a BN developed by another group as easily as navigating from one Web page to another.
- Perform stochastic inference using information from one source and a BN from another.
- Combine BNs from the same or different sources.
- Reconcile and validate BNs.