

Graphical Models for Inference and Decision Making

Instructor: Kathryn Blackmond Laskey
Spring 2019

Unit 1: Introduction and Overview

This course is dedicated to the memory of journalist Danny Pearl, murdered in February 2002, and to the pioneering research of his father Judea Pearl. Judea Pearl's research has the potential to create unprecedented advances in our ability to anticipate and prevent future terrorist incidents. May Judea's research help to bring about Danny's vision of a world where people of all cultures live together in peace, harmony, and mutual respect.

Learning Objectives for Course

- Apply intellectual tools of decision theory to problems of inference and decision-making under uncertainty
- Use graphical probability models to:
 - Develop computationally efficient representations of problems of inference and decision making under uncertainty
 - Use these models to answer questions and/or solve problems requiring reasoning under uncertainty
 - Explain results of models
 - Understand model assumptions, limitations of results, inner workings of models and tools
- Become conversant in state of the art in computationally efficient methods for probabilistic inference and decision making
- Acquire basis for moving state of the art forward

Basic Information

Instructor: Kathryn B. Laskey

- Phone 703-993-1644
- Fax 703-993-1521
- Email klaskey@gmu.edu
- Office 2214 ENGR on GMU Fairfax campus
- Office hours Wednesday 3:30 - 5:30 and by appointment

Place and time: Tuesday 4:30-7:10 PM, ENGR 1110

Recommended text:

- *Probabilistic Graphical Models: Principles and Techniques* (2nd edition) Daphne Koller and Nir Friedman. MIT University Press, 2015

IT Supports:

- Web site <http://mason.gmu.edu/~klaskey/GraphicalModels/>
- Blackboard Upload assignments, solutions, recordings of all lectures

Requirements:

- Assignments 30%
- Take-home midterm 20%
- Take-home final exam 25%
- Project 25%

Prerequisites:

- Strong skills in probability and statistics
- Mathematical maturity
- Computational ability

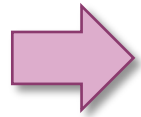
Topics

- Unit 1: Course Overview and Introduction
- Unit 2: Graph Theory, Conditional Independence and Graphical Probability Models
- Unit 3: Representing Knowledge in an Uncertain World
- Unit 4: Inference: Junction Tree Algorithm
- Unit 5: Learning Graphical Models from Data
- Unit 6: Knowledge Engineering and Modeling
- Unit 7: Inference: Other Algorithms
- Unit 8: Planning and Decision Making
- Unit 9: Causality

Learning Objectives - Unit 1

- Describe historical evolution of approaches to modeling uncertainty in intelligent systems
- Learn basic terminology of graphical probability models
 - Conditional independence
 - Graphs to represent conditional dependence structure
 - Using graphical models to represent probability and decision models
- Refresh knowledge of elementary probability theory
- Interpret a graphical probability or decision model

Unit 1 Outline



- Uncertainty and Intelligent Systems
- Decision Theory
- Probability Theory: Review and Fundamental Concepts
- Graphical Probability Models

Information Processing Needs

- Advances in information processing have not kept pace with advances in sensor and computing technology
 - Data deluge
 - Information overload
 - Knowledge “underload”
- Information processing requirements
 - Extract key task-relevant conclusions from huge volumes of data
 - Respond rapidly to previously unknown types of situation
 - Cope with uncertainty and ambiguity
 - Learn from experience
 - Support control of
 - » Allocation of computational resources
 - » Choice of information to display and format of display
 - » Management of collection resources



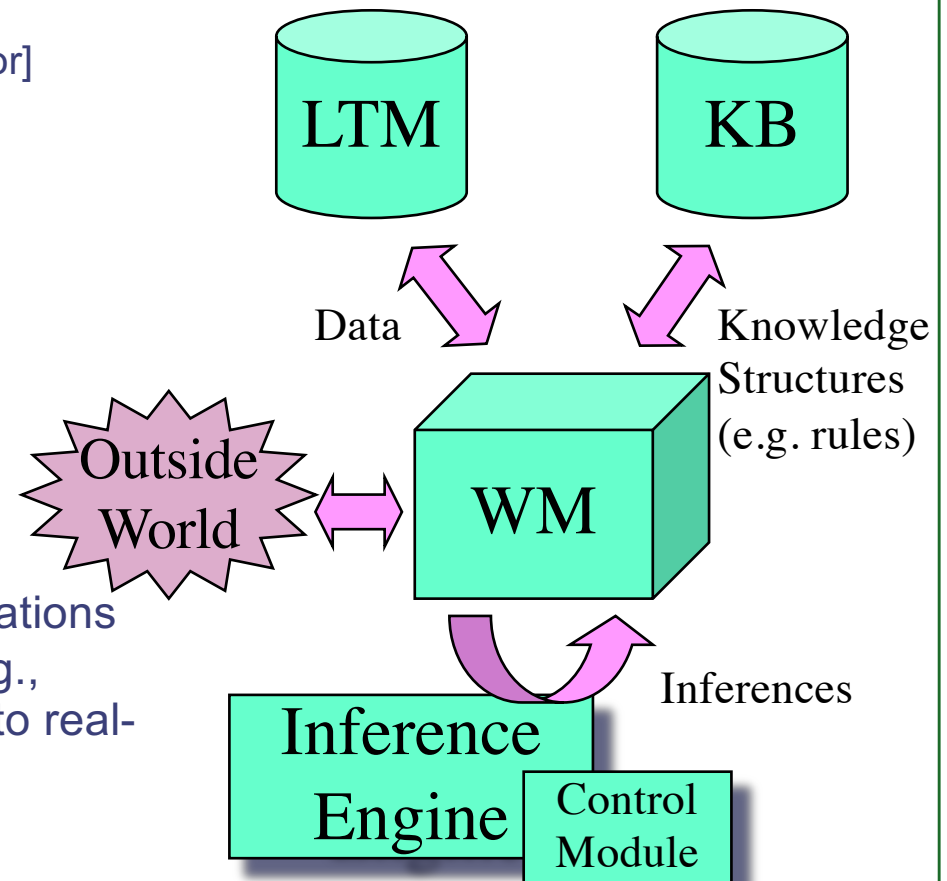
Data, data everywhere, and not the time to think...

Knowledge-Based Systems

- A knowledge based system (also called an expert system or intelligent system) is a computer system designed to behave intelligently in some domain
- Basic premise: Intelligent problem solving requires flexible application of knowledge
- Key feature: Separate representation of knowledge from application to solve problems
- Advantages:
 - opportunistic application of knowledge
 - flexibility to try different solution strategies
 - facilitates modularity, maintainability, ability to explain reasoning to users

Elements of a Generic Knowledge Based System

- Knowledge base
 - Stores generalizations about the domain
R1: “IF [Income < 30,000] & [CreditHistory = Poor]
THEN [deny application]”
- Long term memory
 - Stores facts about the world
F1: “Income_Apl#7 = 21,000”
F2: “CreditHistory_Apl#7 = Poor”
- Inference engine
 - Derives conclusions
R1 + F1 + F2 → “deny application of Apl#7”
- Working memory
 - Holds on to intermediate results of computations
 - Receives inputs from the outside world (e.g., request to process query; streaming input to real-time system)
 - Reads from and writes to LTM and KB
- Control module
 - Sets priorities
 - Decides what order to do which tasks



History: Uncertainty in Knowledge-Based Systems

- Early AI stressed the importance of symbolic reasoning as distinct from “number crunching”
- Application engineers recognized the need to represent degrees of plausibility
- Perhaps the most famous expert system that explicitly allowed for degrees of belief was MYCIN, a system for diagnosing bacterial infections and prescribing treatment (mid-1980's)
 - Certainty factors represented partial degrees of belief
 - Isomorphic to probabilities with untenable independence assumptions (Heckerman)
- Initially the AI community resisted the use of probabilities
 - Artificial intelligence is about symbols not numbers
 - Probabilities are too computationally complex
 - Experts can't supply the required inputs
- Probability is now regarded as essential
 - Practical successes
 - Arguments and existence proofs counter perceived problems
 - Researchers in alternative theories encountering the same tough issues

Extensional versus Intensional Uncertainty Processing

EXTENSIONAL SYSTEM

- Certainty combines truth-functionally (certainty of a formula is a function of the certainties of its subformulas)
 - Most early expert systems (e.g., MYCIN) used extensional uncertainty processing
- $A \Rightarrow_p B$ means "whenever you see A, you are licensed to update the certainty in B by an amount which is a function of the rule strength p."

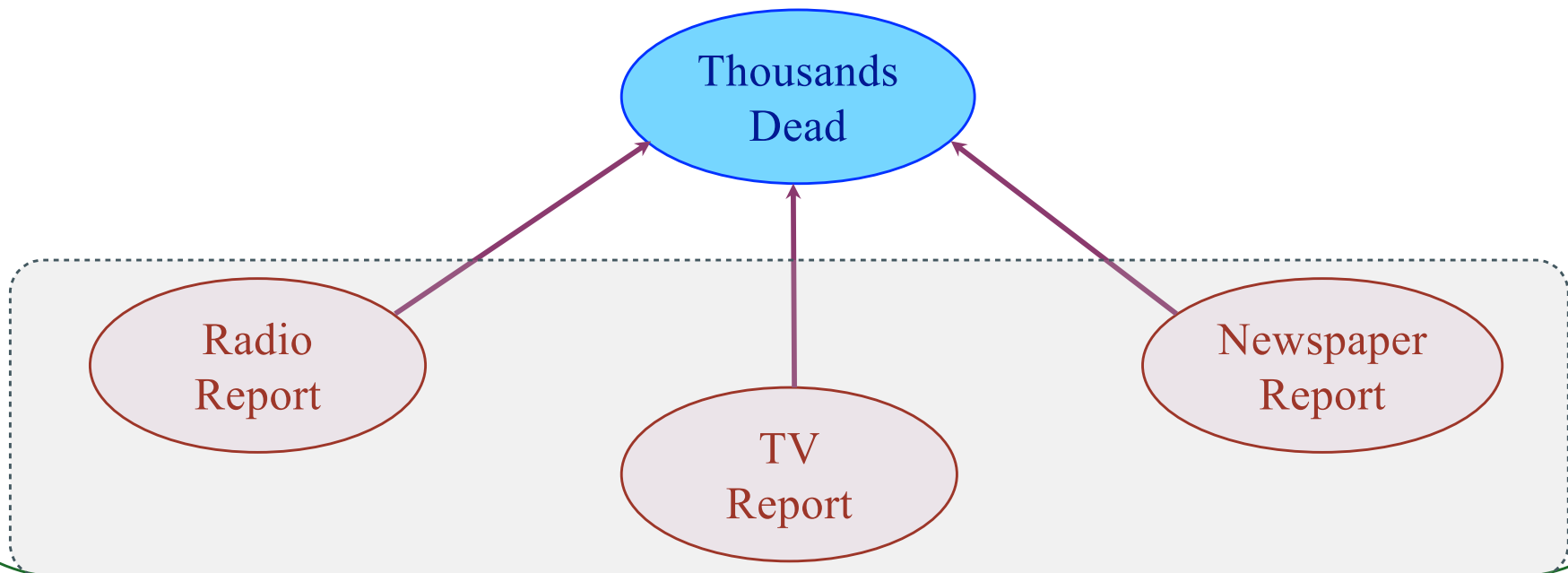
INTENSIONAL SYSTEM

- Assigns probabilities to sets of possible worlds, and the certainties on formulas combine by applying set theoretic operations to the sets of worlds represented by the formulas.
- $A \Rightarrow_p B$ is interpreted as a conditional probability statement: the probability of B conditional on the truth of A is p, or $P_{KB}(B|A) = p$.

A probabilistic system can be either intensional or extensional. Extensional approaches to probabilistic reasoning are justified only if certain independence assumptions are met.

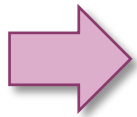
Example: Common Sources Reporting an Event

- You receive 3 reports from usually reliable sources that there has been a terrible nuclear accident near Kiev and thousands are dead
- Then you learn that all 3 reporters talked to the same source
- An intensional system can model the impact of this knowledge on your belief that thousands are dead; an extensional system cannot



Unit 1 Outline

- Uncertainty and Intelligent Systems



- Decision Theory
- Probability Theory: Review and Fundamental Concepts
- Graphical Probability Models

History: Statistical Inference

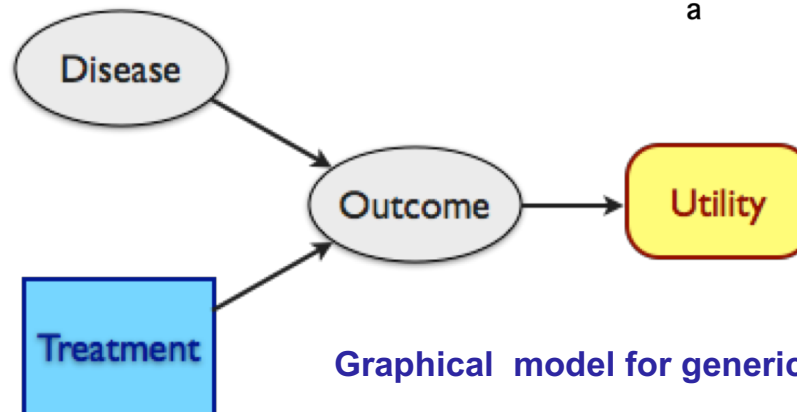
- Traditionally, statistical inference was focused on problems with very few parameters
 - Limit theorems examined behavior of estimators as sample size grew large and number of parameters remained fixed
- Emphasis is growing on modeling very high-dimensional problems, e.g.
 - Marketing tailored to customer
 - Healthcare tailored to patient
 - Time-varying spatial processes
 - Region-specific weather and/or climate models
- Statistical inference methods must be adapted to deal with parameter spaces that grow very large to adapt flexibility to variability in the data
 - Number of parameters grows with size of data set

What is Decision Theory?

- Formal, structured, scientifically sound approach for
 - Representing knowledge about decision making under uncertainty
 - Applying knowledge to solve problems
- "Divide and conquer" principle
 - Human provides judgments on simple sub-problems
 - Computer aggregates responses consistently
- Heuristic approaches can be judged by how well they approximate decision theoretic ideal
- Decision theory provides a foundation for learning systems that:
 - Protect against "overfitting"
 - » Decision theoretic systems have a built in bias against complexity for small data sets ("natural Occam's razor") but add complexity as needed to explain data
 - Provide clear semantics for "bias"
 - Integrate top-down expert knowledge with bottom-up case-based learning

Decision Theory

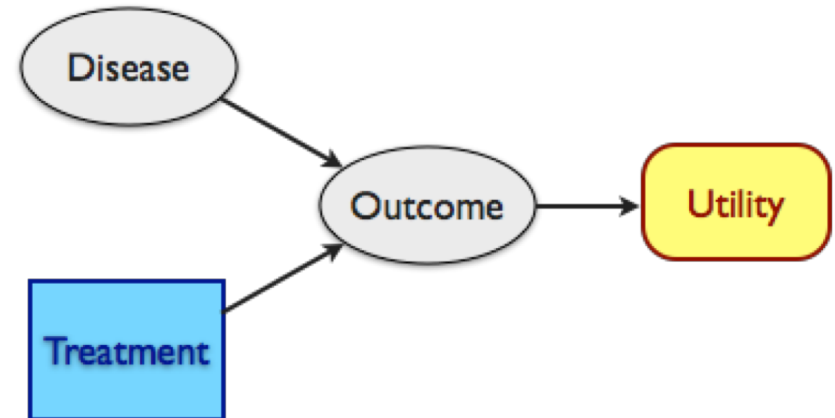
- Elements of a decision problem
 - Possible actions: $\{a\}_{a \in A}$
 - States of the world (usually uncertain): $\{s\}_{s \in S}$
 - Possible consequences: $\{c\}_{c \in C}$
 - » Consequences $c(s,a)$ are functions of states and actions
- Question: What is the best action?
- Ingredients of a decision theoretic model
 - Utility function $u(c)$ expresses preference for consequences
 - Probability $p(s)$ expresses knowledge/uncertainty about state of world
 - Best action maximizes expected utility: $a^* = \underset{a}{\operatorname{argmax}} \{E[u(c) | a]\}$



Graphical model for generic medical decision problem

Simplified Example

- Actions: diagnoses and treatment plans
- States of world: Actual disease patient has
- Consequences: Outcome of treatment plan
- Utilities: Measure "goodness" of outcomes



Action	State	Outcome	Utility
Treat	Has disease	Disease free Side effects	90
Treat	Doesn't have disease	Disease free Side effects	90
Don't treat	Has disease	Long illness No side effects	0
Don't treat	Doesn't have disease	Disease free No side effects	100

Probabilities: $P(\text{Has disease}) = .6$

Expected utilities:

$$\text{Treat: } .6 \times 90 + .4 \times 90 = 90$$

$$\text{Don't treat: } .6 \times 0 + .4 \times 100 = 40$$

Optimal action: Treat patient

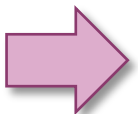
Decision Theory as a Foundation for Intelligent Systems

- Justifications
 - Axiomatics
 - Clear semantics
 - Practical success
 - Built in learning theory
- Some reasons for growing popularity:
 - Computational tools now exist for problems of realistic complexity
 - Systems based on decision theory have achieved success
 - Benefits of decision theoretic thinking and pitfalls of shortcuts are becoming more widely understood
 - Evolutionary arguments: Adaptive agents situated in environments that reward “rational” behavior will evolve toward approximate rationality
- The best working systems may not be decision theoretic
 - Engineering tradeoff between cost/tractability and theoretical purity
 - Beware of any system that is not approximately decision theoretic!

Decision Theory for Intelligent Systems

- Traditional decision theory:
 - Problem context is assumed known
 - Fixed, known set of possible actions, possible outcomes
 - Known (or pre-existing to be elicited) probabilities and utilities
 - Representation is simple and unstructured
 - Emphasis is on solving for posterior distribution or optimal decision
- Traditional decision analysis:
 - Human-intensive, largely manual with some computer assistance
 - Emphasis is on one-time solution
- For knowledge-based systems
 - Constructing a *knowledge representation* is as important as problem solution
 - Possible actions, possible outcomes are constructed at solution time from implicit internal representation
 - Emphasis is on repeatable performance on a class of problems

Unit 1 Outline

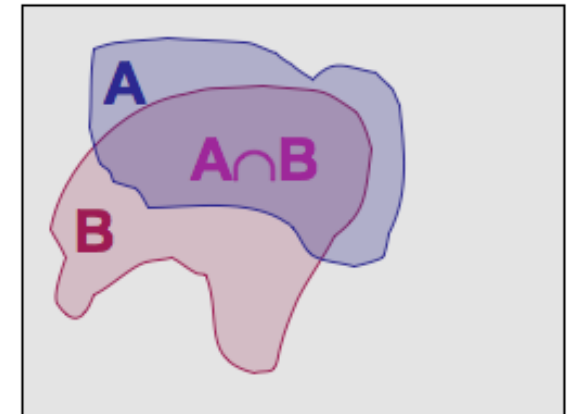
- Uncertainty and Intelligent Systems
- Decision Theory
-  • Probability Theory: Review and Fundamental Concepts
- Graphical Probability Models

Probability Theory

- Probability theory is a body of mathematical theory which has been applied to problems of reasoning with uncertainty
- Probability can be used to model:
 - Problems with natural symmetries (physics; games of chance; explicit randomization in statistical experiments)
 - Problems characterized by stable long-run frequencies
 - Degrees of belief of rational agents for propositions about which they are uncertain
- There has been active debate over “interpretations of probability” and the types of problem to which it can meaningfully be applied

Probability: Basic Definition

- A finitely additive probability distribution is a function applied to events, or subsets of a universal set Ω , such that:
 1. $P(A) \geq 0$ for all sets A
 2. $P(\Omega) = 1$
 3. If $A \cap B = \emptyset$ then $P(A \cup B) = P(A) + P(B)$
- A probability distribution is countably additive if it satisfies:
 3. If $A_i \cap A_j = \emptyset$ for all i, j , then $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$
- From this basic definition one can deduce:
 - $P(\emptyset) = 0$
 - $P(A)$ lies between 0 and 1 for all subsets A of Ω
 - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ for any A and B
 - and other familiar identities of probability theory



Fundamental Concepts of Probability

- Conditional probability
- Independence and dependence
- Conditional independence
- Law of total probability
- Bayes rule
- Random variable
- Joint probability distribution

Conditional Probability

- The conditional probability of an event given another event is the probability that the first occurs if the second has occurred
 - $P(A|B, C)$ means probability of event A given that events B and C have occurred.
- Example:
 - In Figure 1, the probability space consists of the whole square. Each event is represented by a shape inside the square. The probability of an event is represented by the area of its shape.
 - In Figure 2, we know that event A is true, so everything outside A is false. This is shown by “expanding” A to have area 1 and setting $P(B | A)$ to the area of the part of B that is inside of A

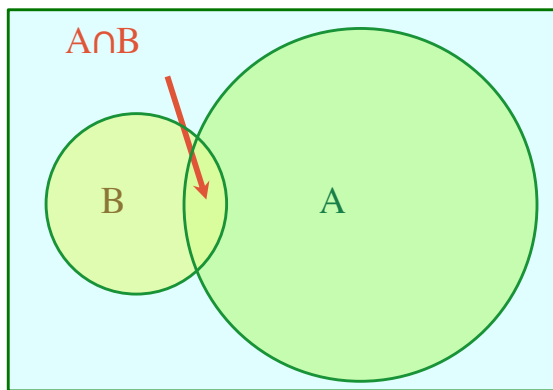


Figure 1

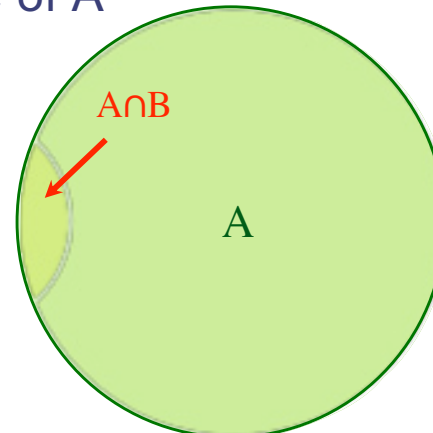


Figure 2

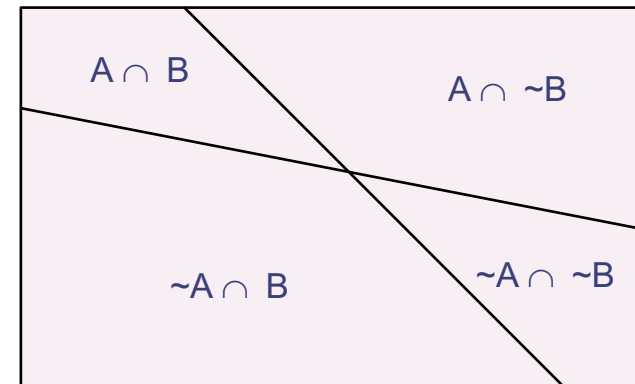
$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Independence

- Events A and B are independent if the conditional probability of B given A is equal to the unconditional probability of B
- That is:
 - $P(A|B) = P(A)$, or $P(B|A) = P(B)$.
 - This means that knowing B doesn't change our beliefs about A

$A \cap B$	$A \cap \sim B$
$\sim A \cap B$	$\sim A \cap \sim B$

A and B are independent



A and B are not independent

Conditional Independence

- Events A and B are conditionally independent given event C if learning the value of C makes A and B independent
 - $P(A|C) = P(A|B, C)$
 - This means that if C is known, then learning about B doesn't change our beliefs about A

$A \cap B \cap C$	$A \cap \sim B \cap C$	$A \cap B \cap \sim C$	$A \cap \sim B \cap \sim C$
$\sim A \cap B \cap C$	$\sim A \cap \sim B \cap C$	$\sim A \cap B \cap \sim C$	$\sim A \cap \sim B \cap \sim C$

A and B are independent
with high probability given C

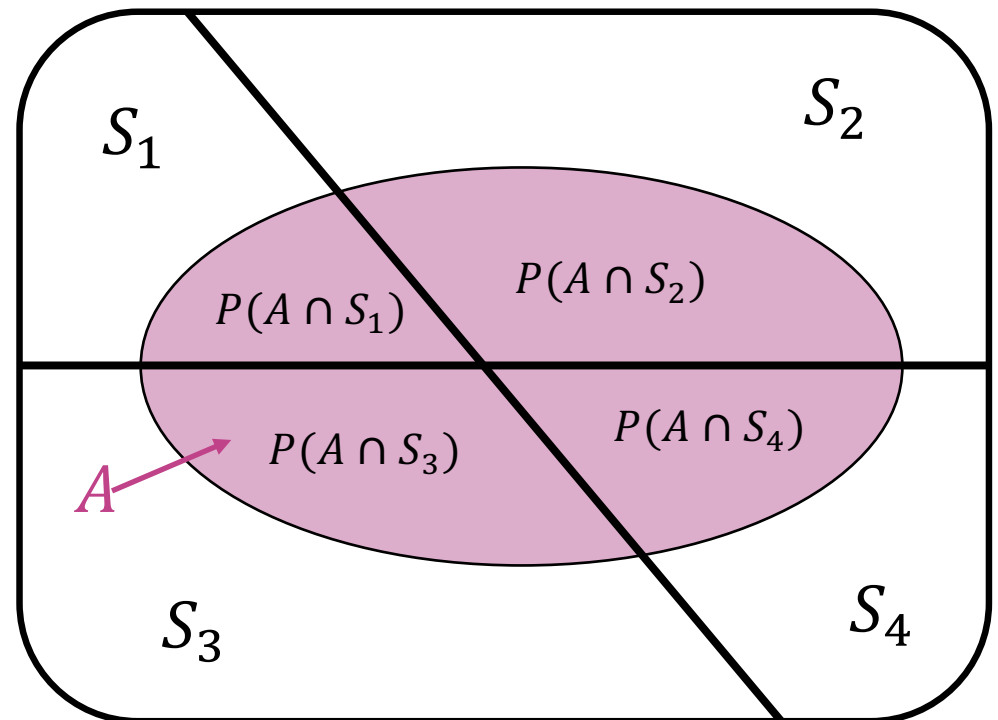
A and B are independent
with low probability given $\sim C$

Law of Total Probability

$$P(A) = \sum_{i=1}^n P(A \cap S_i) = \sum_{i=1}^n P(A|S_i)P(S_i)$$

if $S_i \cap S_j = \emptyset$ for all i, j

and $\bigcup_{i=1}^n S_i = \Omega$



$$P(A) = P(A \cap S_1) + P(A \cap S_2) + P(A \cap S_3) + P(A \cap S_4)$$

Bayes Rule: The Law of Belief Dynamics

- Objective: use new evidence E to update beliefs probability of hypothesis H
 - H: patient has (does not have) disease
 - E: evidence from test
- From the definition of conditional probability it is easy to derive Bayes rule:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} \quad [P(E)>0] \quad (\text{standard form})$$

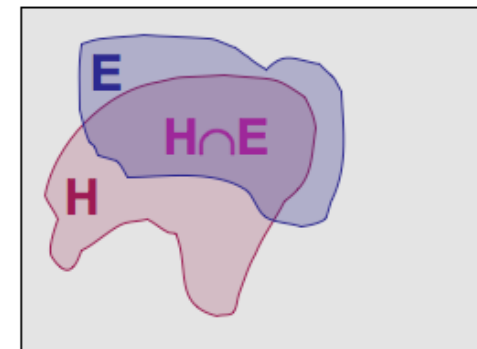
$$\frac{P(H_1 | E)}{P(H_2 | E)} = \frac{P(E|H_1)P(H_1)}{P(E|H_2)P(H_2)} \quad [P(H_2)>0] \quad (\text{odds-likelihood form})$$

– Terminology:

- » $P(H)$ - The prior probability of H
- » $P(E)$ - The predictive probability of E
- » $\frac{P(E|H_1)}{P(E|H_2)}$ - The likelihood ratio for H_1 versus H_2

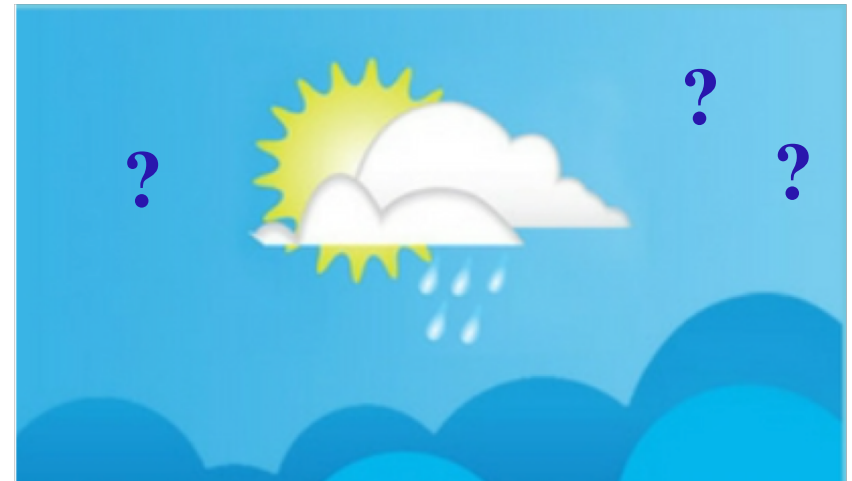
- $P(E|H)$ - The likelihood for E given H
- $P(H|E)$ - The posterior probability of H given E
- $\frac{P(H_1)}{P(H_2)}$ - The prior odds ratio for H_1 versus H_2

The posterior probability of H_1 increases relative to H_2 if the evidence is more likely given H_1 than given H_2



Bayes Rule Example

- At this time of the year in this location, 80% of the days have rain
- However, the weather forecaster says tomorrow will be sunny.
- We know:
 - She is right 90% of the time when she predicts a sunny day, and
 - She is right 70% of the time when she predicts a rainy day.
- What is the probability of rain for tomorrow?



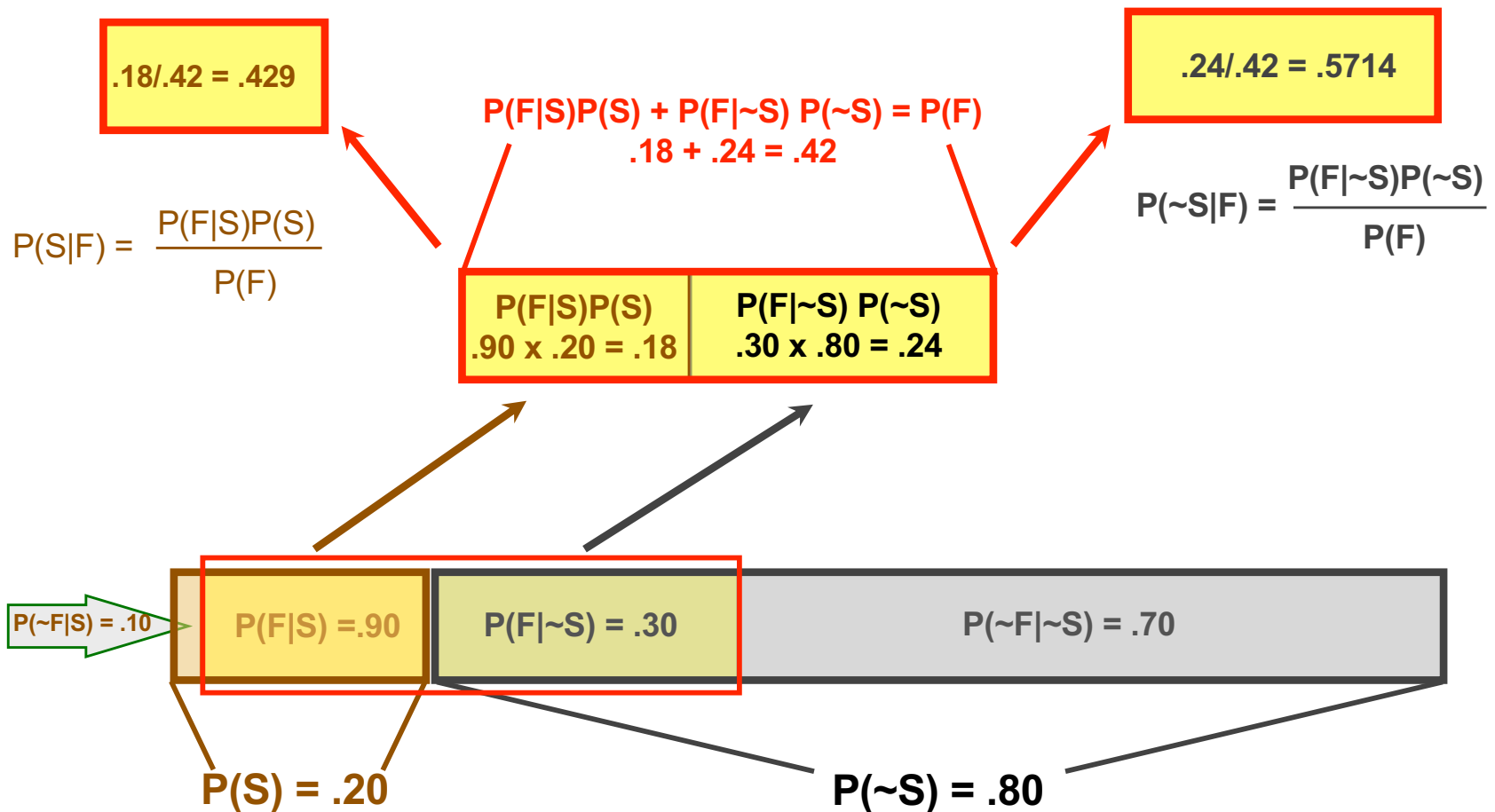
Bayes Rule Example – With Numbers

- Notation:

- S: sunny day.
- $\sim S$: (not S) rainy day
- F: a sunny weather forecast.
- $\sim F$: a rainy weather forecast.

- Data:

- $P(\sim S) = 0.80 \longrightarrow P(S) = 0.20$
- $P(F|S) = 0.90 \longrightarrow P(\sim F|S) = 0.10$
- $P(\sim F|\sim S) = 0.70 \longrightarrow P(F|\sim S) = 0.30$

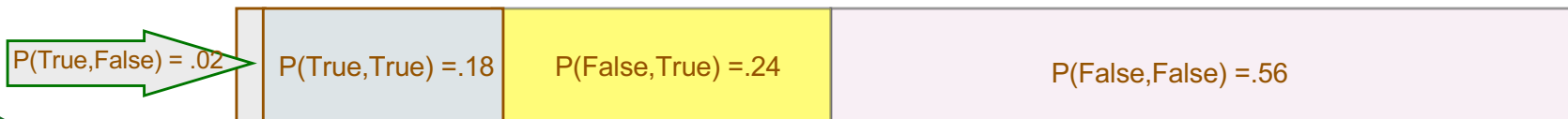


Random Variable

- A *random variable* represents a property or feature of the world
 - Property can take on one of a set of allowable values
 - Probability quantifies likelihoods of different values
 - Likelihood of different values may depend on values of other properties
 - Statistics texts often assume values are numbers; we assume values can be taken from any well-defined set
- Examples:
 - Whether a patient is a smoker (Possible values: True/False)
 - Speed of a car in km/hr (Possible values: Non-negative numbers)
 - Marital status of a loan applicant (Possible values: Single / Married / Divorced / Separated / Widowed)
- Mathematically, a random variable is defined as a function that maps elements of a *sample space* to outcomes of the random variable
 - Smoker : Patients \rightarrow {True, False}
 - Speed : Cars @ times \rightarrow Positive Real Numbers
 - MaritalStatus : Applicants \rightarrow {Single, Married, Divorced,...}
- Probabilities on sample space give rise to probabilities on values

Joint Probability Distribution

- A *joint probability distribution* is a probability distribution defined on a k -dimensional sample space
- Weather example – sample space
 - $S: \text{day} \rightarrow \{\text{True}, \text{False}\}$ True means sunny day
 - $F: \text{day} \rightarrow \{\text{True}, \text{False}\}$ True means forecast sunny
 - $(S,F): \text{day} \rightarrow \{(\text{True},\text{True}), (\text{True},\text{False}), (\text{False},\text{True}), (\text{False},\text{False})\}$
- Weather example – joint probability distribution
 - $P(\text{True}, \text{True}) = 0.2 \times 0.9 = 0.18$
 - $P(\text{True}, \text{False}) = 0.2 \times 0.1 = 0.02$
 - $P(\text{False}, \text{True}) = 0.8 \times 0.3 = 0.24$
 - $P(\text{False}, \text{False}) = 0.8 \times 0.7 = 0.56$
- Weather example – *marginal* distribution for S:
 - $P(S=\text{True}) = 0.2, P(S=\text{False}) = 0.8$
- Weather example – *conditional* distribution for S given F
 - $P(S=\text{True} \mid F=\text{True}) = 0.429, P(S=\text{False} \mid F=\text{True}) = 0.571$
 - $P(S=\text{True} \mid F=\text{False}) = 0.034, P(S=\text{False} \mid F=\text{False}) = 0.966$



Unit 1 Outline

- Uncertainty and Intelligent Systems
- Statistical Decision Theory and Graphical Models
- Probability Theory: Review and Fundamental Concepts
- ➔ • Graphical Probability and Decision Models

Graphical Models for Representing Inference and Decision Problems

- Graphs are a natural representation for links between propositions
- Graphical representations of probabilistic dependencies have become very popular, e.g.
 - Directed graphical models (Bayesian networks)
 - Markov networks (Markov random fields)
 - Graphical models on chain graphs
 - Factor graphs
 - Probability trees
 - Hidden Markov models - special case of Bayesian networks
 - Influence diagrams (decision graphs)
 - Markov decision processes (MDPs) and partially observable Markov decision processes (POMDPs) - special case of influence diagrams
- This course examines graphical models for representing
 - Joint probability distributions on many random variables
 - Decision problems with uncertain outcomes

Uncertain Inference Example

Vehicles travel faster on roads than on smooth terrain and faster on smooth terrain than on rough terrain. Tracked vehicles travel off-road more often than wheeled vehicles. Tracked vehicles can travel on very rough terrain where wheeled vehicles cannot travel. Tracked vehicles travel faster than wheeled vehicles on rough terrain, but wheeled vehicles travel faster than tracked vehicles on smooth terrain and roads. A moving target indicator (MTI) sensor provides approximate position and velocity for vehicles that are moving, but cannot see stationary objects. An imaging sensor usually distinguishes vehicles from other objects, and usually reports correctly whether a vehicle is tracked or wheeled. Cloud cover can interfere with the ability of the imaging sensor to distinguish vehicles from other objects.

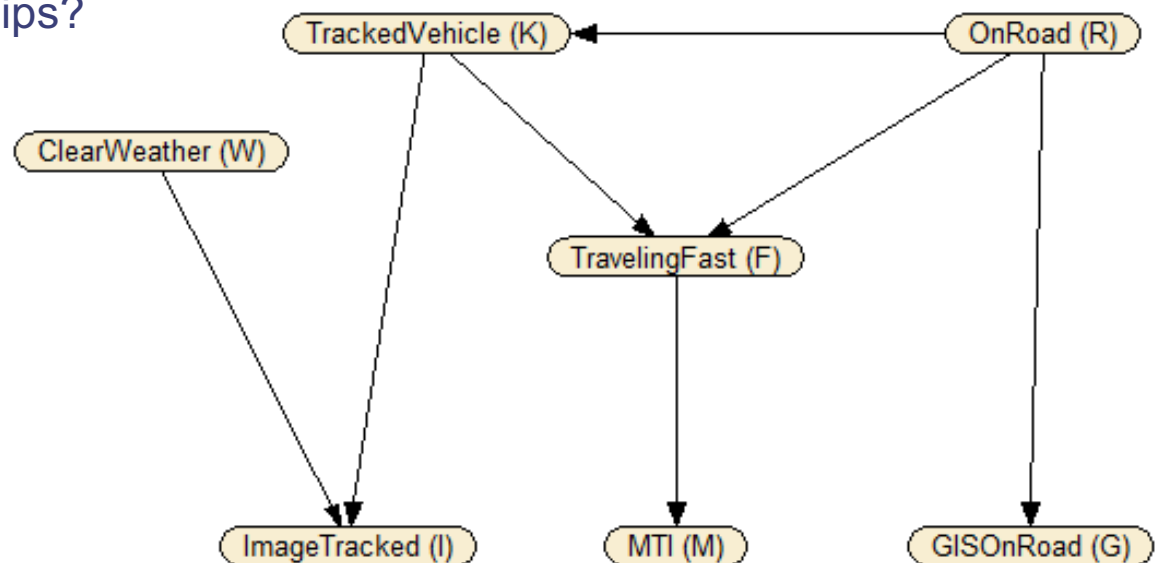
Objective: infer vehicle type from MTI and imaging sensor reports (given information about the road network, terrain, and weather)

Vehicle Identification Random Variables

- **Highly simplified sample space with seven binary RVs**
 - K = Vehicle is tracked (true/false)
 - R = Vehicle is on road (true/false)
 - W = Weather is clear (true/false)
 - F = Vehicle is traveling fast (true/false)
 - I = Image sensor reports tracked vehicle (true/false)
 - M = MTI sensor reports vehicle is moving fast (true/false)
 - G = GIS sensor reports road (true/false)
- **Binary RVs map sample space of $2^7 = 128$ combinations of truth-values to true/false values**
 - e.g. R maps (K, R, W, F, I, M, G) to value of R
 - Joint distribution specifies a probability for each of these 128 elements
- **For a problem with 100 true/false random variables, the sample space would have $2^{100} = 1.3 \times 10^{30}$ elements (more than Avogadro's number)**
 - Reasoning directly with the full joint distribution is not scalable
 - We need a more efficient divide-and-conquer approach

Vehicle ID: Graphical Probability Model

- There are 7 random variables: (R, W, K, F, I, G, M)
 - Each random variable maps an element of the sample space to a value indicating whether the random variable is true or false in that world
- We can use a graphical probability model to represent direct dependencies
- In a Bayesian network (aka directed graphical model) each random variable is conditionally independent of its non-descendants given its parents
 - Example: I is conditionally independent of F given its parents K and W
 - Can you identify some other conditional independence relationships?

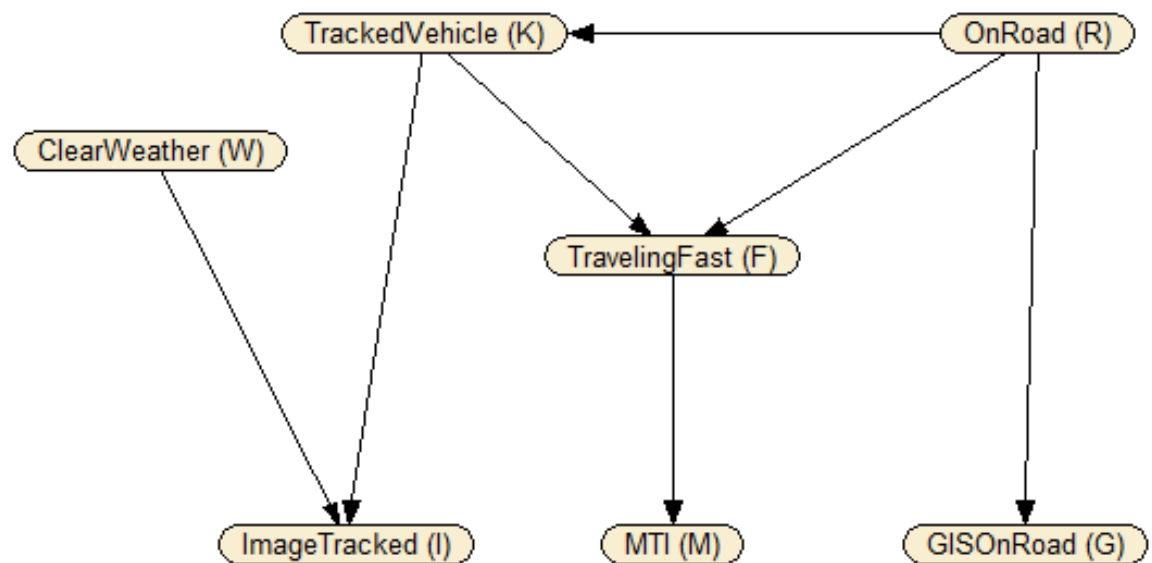


Graphical Probability Model and Parsimony

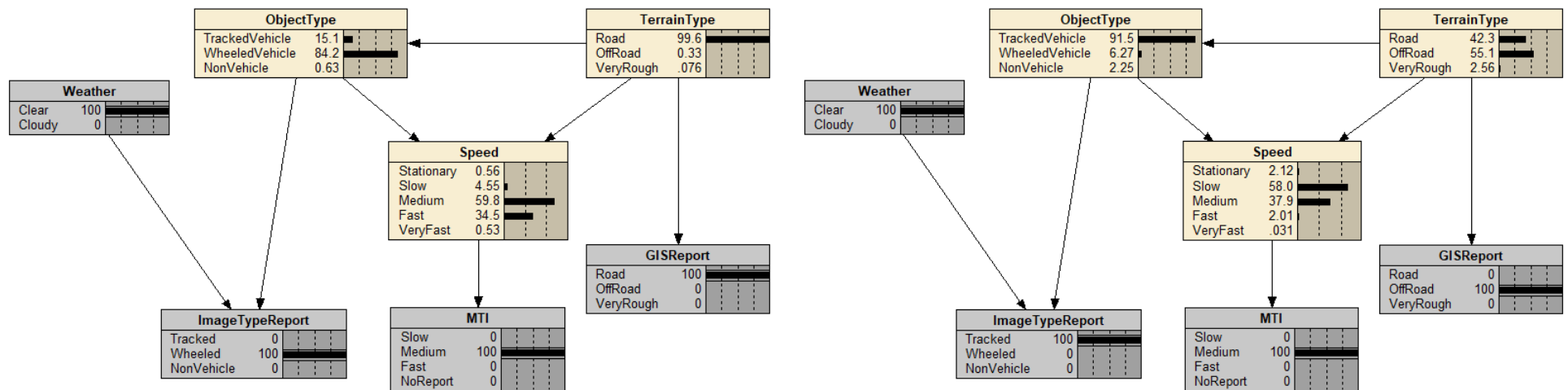
- There are $2^7 = 128$ elements of the sample space (why?)
- 127 numbers are needed to specify the joint distribution (why?)
- The joint distribution specified by this Bayesian network can be written in a factored form:

$$P(R, W, K, F, I, G, M) = P(R)P(W)P(K|R)P(F|K, R)P(I|W, K)P(G|R)P(M|F)$$

- This Bayesian network can be specified with: $1 + 1 + 2 + 4 + 4 + 2 + 2 = 14$ probabilities



Vehicle Example Extended: Many-Valued Random Variables



Netica software: <http://norsys.com>

- Purpose of model: identify object type from multi-sensor reports
- BN simplifies specification:
 - Fully general joint distribution: 3240 probabilities (why?)
 - This Bayesian network (general): 78 probabilities (why?)
- This model can be specified by combination of expert judgment and estimation from data
- A more realistic model would use continuous distribution for speed

Summary: Graphical Probability Model

- Formal language for representing knowledge about uncertain quantities
 - nodes represent random variables
 - arcs represent direct dependence relationships among random variables
 - local numerical functions encode strengths of dependencies
- Computational architecture for computing impact of evidence on beliefs
 - updates beliefs when new evidence is observed
 - exploits independence assumptions to make computation more efficient
- Factored representation of a joint probability distribution on many random variables
 - scalable representation of high-dimensional distribution

Summary and Synthesis

- Probability was viewed with disfavor in early days of artificial intelligence
 - What changed things? Why?
- Very high-dimensional distributions were viewed with disfavor in early days of mathematical statistics
 - What changed things? Why?
- What is a graphical probability model?
- Graphical probability models have:
 - Helped to legitimize probability in artificial intelligence
 - Helped statisticians to build increasingly sophisticated and complex models
 - Helped to tie together symbolic and numeric computing
 - Helped to bridge the sub-symbolic / symbolic gap
- Graphical models support
 - Representation of knowledge as modular components that express both what is known and what is uncertain about a small part of the world
 - Integration of expert knowledge and data
 - Decision making in presence of uncertainty, including decisions about whether to collect information
- The inventors of probability theory thought of it as a logic of enlightened rational reasoning. The information technology revolution is providing “cognitive tools” to support enlightened rational reasoning