# Graphical Models for Inference and Decision Making

## Knowledge Engineering

Instructor: Kathryn Blackmond Laskey

Spring 2019

# Learning Objectives

- Describe knowledge elicitation as a problem in system lifecycle engineering
  - Describe the stages in building a Bayesian network and/or decision graph model
  - Describe the activities that occur at each stage
  - Describe the products produced at each stage
- Describe how the KE process is managed
- Be prepared to carry out the process of developing, implementing and testing a Bayesian network or decision graph model for a problem of interest to you

# Unit 7 Outline

- The Knowledge Acquisition Lifecycle
- Building the Model
- Managing and Evaluating the Model
- Knowledge Engineering for Relational Graphical Models

# Importance of Structured KE Process

- Graphical models have become well established tools for representing and reasoning under uncertainty
- Applications are growing more complex
- A formal, repeatable process for knowledge engineering is becoming more important
  - Early work on elicitation of probability models (1970's) focused on eliciting single probabilities or univariate probability distributions
  - Early work in graphical models tended to assume that structure elicitation was relatively straightforward
  - As models become more complex the KE process must be managed
- Knowledge elicitation for large Bayesian networks is a problem in systems engineering

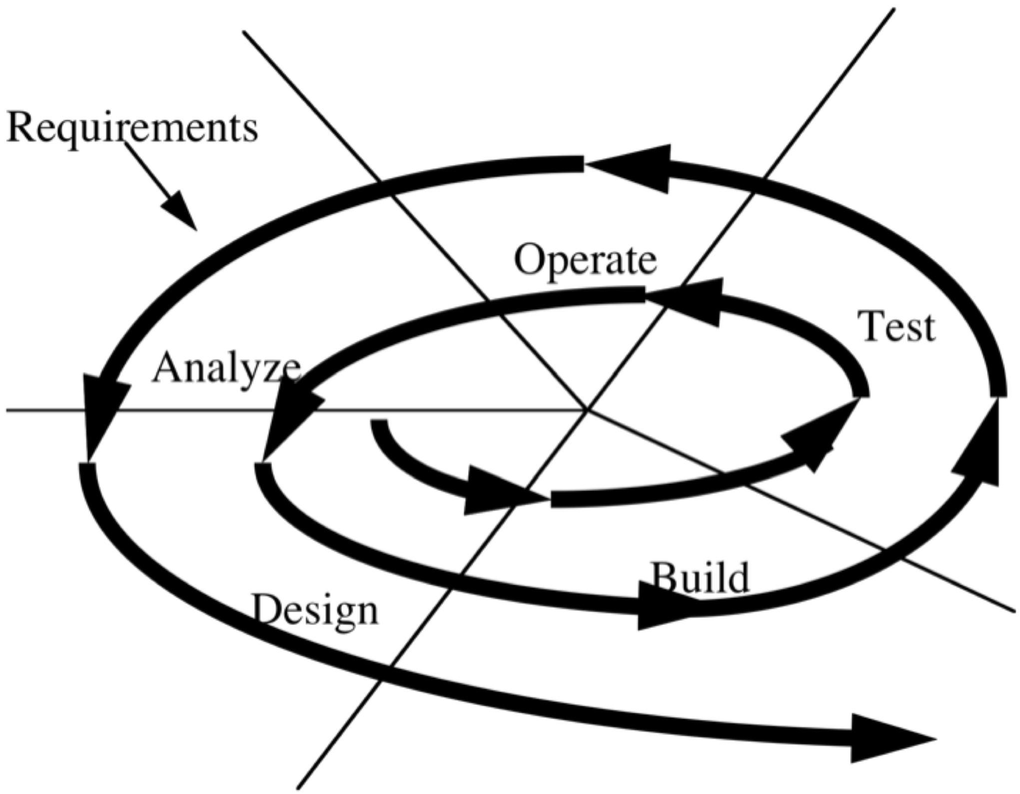
# What is Knowledge Acquisition?

- Objective:
  - Construct a model to perform defined task
  - Develop knowledge base for use in solving problems in defined class
    - » Modularity
    - » Modifiability and reusability
- Participants: Collaboration between problem expert(s) and modeling expert(s)
- Process: Iterate until done
  - Define task objective
  - Construct model
  - Evaluate model

# Systems Engineering

- System
  - A set of interacting components organized to serve a specified objective
- Systems engineering
  - The technical and managerial process by which a user need is translated into an operational system
- System life cycle
  - Systems evolve through predictable phases
    - Design
    - Development
    - Operation
    - Retirement
  - Systems engineering is organized around life cycle
    - Support current phase
    - Anticipate and plan for next phase

# Spiral Model of Lifecycle Engineering



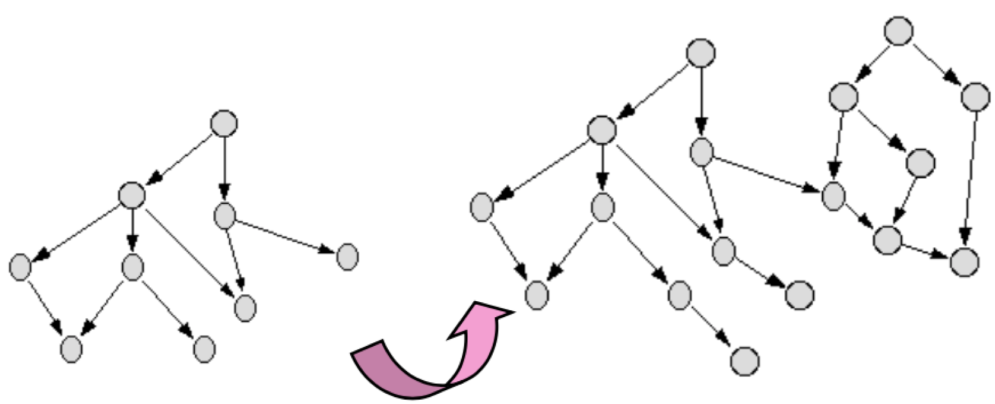

- System development viewed as repeating cycles of design, implementation, operation, evaluation
- Evaluation used to plan next cycle
- Early phases develop prototype for planning and risk mitigation
- Later phases develop operational versions

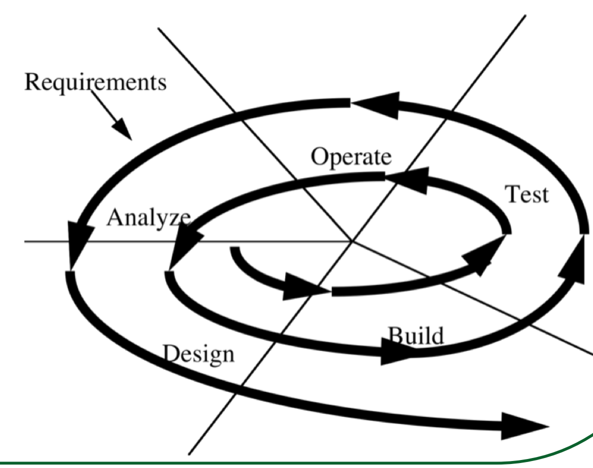
# Agile Development

- Developed by software engineers; being applied by systems engineers
  - http://www.incose.org/chesapek/Docs/2011/Presentations/2011_09_21_Johnson_AgileEngineering.pdf
- Principles:
  - Continuous collaboration with customer
  - Continuous updates
  - New delivery on very short cycle (often weekly)
  - Value participants and interactions
  - Emphasize simplicity
- Difficult to do well but good implementation can provide major benefits on projects emphasizing interaction between developers and customers

# Agile Process for Graphical Model Engineering

- Goal of knowledge engineering
  - Discovery and construction of appropriate model
  - Not extraction of pre-existing model
- Agile approach is necessary for systems in which requirements are discovered as development progresses
- KE spirals
  - Construct series of prototype models
  - Explore behavior of prototype model on sample problems
  - Evaluate prototypes and restructures as necessary
- KE changes both expert and elicitor
  - Understanding of expert and elicitor deepen as KE proceeds
  - Improves communication between elicitor and expert





# Applying Agile Knowledge Engineering

- Begin with a small sub-problem
  - Self-contained
  - Can be completed in short time
  - Interesting in its own right
  - Reasonably representative of global problem
- Build and test model for sub-problem
  - Look for common structures and processes that will recur
  - Think about more efficient ways to structure KE
  - Develop and document conventions (“style guide”) to be followed as models are expanded
- Expand to more complex problems

# Selecting a Subproblem

- Initial model or expansion of existing model
- Characteristics
  - Manageable size
  - Interesting in its own right
  - Path to expansion
  - Risk mitigation
- How to restrict
  - Focus or target variables - variables of direct interest to client
    - Restrict to subset of variables of interest
    - Restrict to subset of values
  - Evidence variables - variables for which information will be available; used to draw inferences about the focus variables
    - Restrict to subset of evidence sources
  - Context variables - variables that will be assumed known and will be set to definite values
    - Restrict to subset of contextual conditions (sensing conditions, background casual conditions; assignment of objects to sensors; number of objects)

# The Participants

- Naive view
  - Put problem experts and modeling experts in a room together and magic will happen”
- Realistic view
  - Pure “problem experts” and pure “modeling experts” will often talk past each other
  - Modeling experts must learn about the problem and problem experts must learn what models can do
  - This process can be time consuming and frustrating
  - Team will be more productive if both sides expect and tolerate this process
- Training
  - The most productive way of training modelers and problem experts is to construct very simple models of stylized domain problems
  - Goal is understanding and NOT realism or accuracy!
  - Beware: the training phase can seem pointless and frustrating
  - It is important to get expert buy-in

# The Domain and the Expert

- Domains well suited to reliably and measurably good performance
  - Tasks are repeatable
  - Outcome feedback is available
  - Problems are decomposable
  - Phenomena are inherently predictable
  - Human behavior/"gaming" not involved
- Characteristics to look for in an expert
  - Expertise acknowledged by peers
  - Articulate
  - Interest and ability to reason about reasoning process
  - Tolerant of messy model-building process
- Note: some of the best experts as measured by performance on the problem are not very good experts for knowledge elicitation
  - Can do it but cannot articulate how they do it
  - Become frustrated with elicitation process

# Unit 7 Outline

- The Knowledge Acquisition Lifecycle
- Building the Model
- Managing and Evaluating the Model
- Knowledge Engineering for Relational Graphical Models

# Model Components

- What are the variables?
  - Random variables
  - Action and utility nodes
- What are their states?
- What is the graph structure?
  - Is there repeated structure?
- What is the structure of the local distributions?
- What are the parameters?
  - Probability distributions
  - Utility functions

# The Clarity Test

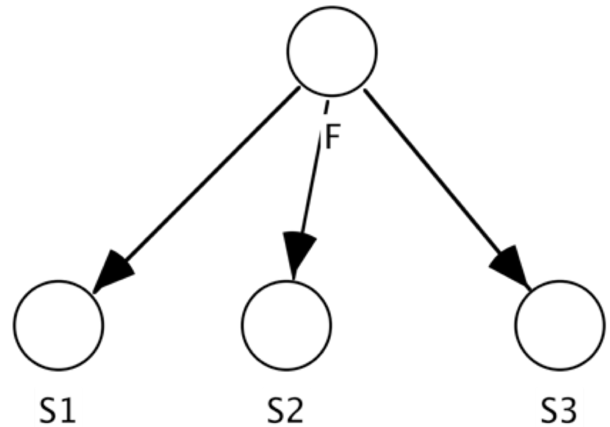
- Usually begin with vague structure to develop understanding of problem
- Final model should have clear operational meaning for all components
- Clarity test:
  - Could a clairvoyant unambiguously specify value of all nodes and states?
  - "Fever is high" does not pass clarity test
  - "Fever ≥ 103° F" passes clarity test

# Defining the Variables

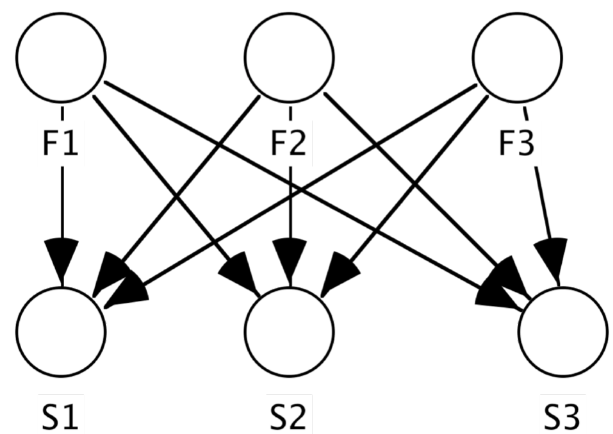
- Begin with “focus variable” and spread out to related variables
- Ask about causally related variables
  - Variables that could cause a state to be true
  - Variables that could prevent a state from being true
- Ask about enabling variables
  - Conditions that permit, enhance or inhibit operation of a cause
- Ask about effects of a variable
- Ask about associated variables
  - Knowing value provides information about another variable
- Ask about observables
  - What evidence could be observed that would enable you to infer state of a variable

# Target or Focus Variable in Diagnosis

- Diagnosis problem: goal is to infer "fault," "disease," "problem" from a set of "findings," "symptoms" or "indicators"
  - Direction of inference is usually from effect to cause
- Modeling issue: single or multiple fault?
- Single fault
  - Collect all faults as states of a single node
  - Modeling simplicity and inference tractability
- Applicable domains:
  - Pathology- one disease/slide
  - Pediatrics- acute diseases
  - Highly maintained mechanical systems
- Modified single-disease hypothesis:
  - Include common combinations as explicit hypotheses




**Single Fault Model**




**Multiple Fault Model**

# Target or Focus Variable in Prediction

- Prediction:
  - Objective is to predict a variable that has not yet occurred or is not known
  - Direction of inference is usually from cause to effect
- Applications:
  - Planning
  - Intelligence analysis
  - Policy modeling
  - Strategic decision making

# States of Variables

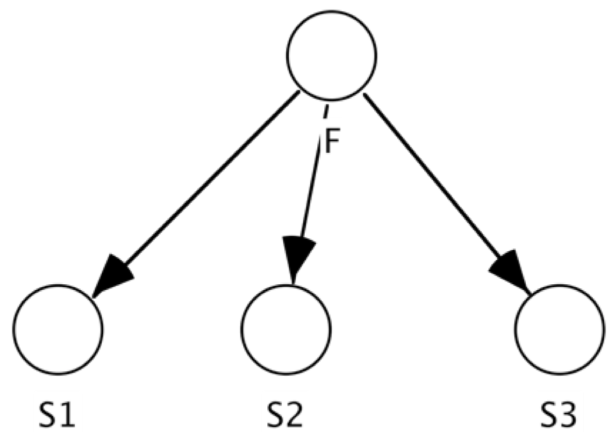
- States must be exclusive and exhaustive
  - Naive modelers sometimes create separate variables for different states of the same variable
- Types of variable
  - Binary (2-valued)
  - Qualitative
  - Numeric discrete
  - Numeric continuous
- Dealing with infinite and continuous state sets
  - Standard Bayesian network software requires finitely many states per random variable
    - » Continuous random variables must be grouped into bins
    - » Bin boundaries should represent meaningful differences in effect on related variables
    - » Different resolutions may be appropriate for different purposes
  - Exact inference algorithms exist for linear Gaussian and conditional linear Gaussian BNs
    - » Software support is limited
  - Monte Carlo inference can be used for BNs with continuous variables
    - » Software support is limited

# Graph Structure

- Goal: develop model that is good enough for task
- Criteria to consider
  - Parameter parsimony
    - » Fewer nodes, fewer arcs, smaller state spaces, coarser partitions simplifies elicitation makes learning more efficient (fewer observations required)
  - Fidelity of model to problem
    - » Greater fidelity often requires more nodes, arcs, states, finer partitions
    - » Balance benefit against cost of additional modeling
    - » Too much detail can decrease accuracy
  - Expert comfort with probability assessments
- Direction of arcs
  - Causal direction can increase:
    - » Conditional independence
    - » Ease of probability elicitation
    - » Efficiency of learning
  - Causal direction is required if modeling effects of interventions (planning)
  - It may be helpful to show user a graph with arcs in inferential direction even if BN has causal arcs

# Naïve Bayes

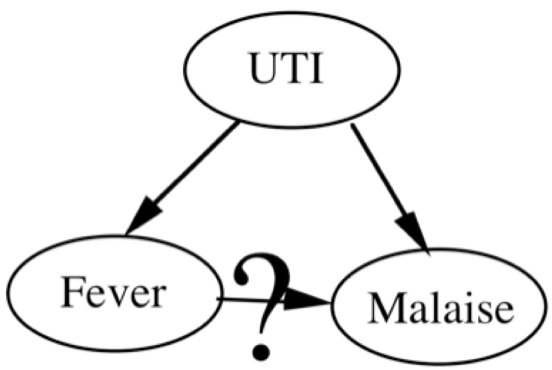
- Commonly applied in diagnosis problems
  - Simplifies elicitation
  - Simplifies inference
  - Simplifies learning
- Single parent node and multiple leaf nodes that are conditionally independent given parent
  - Also known as "idiot Bayes"
  - Simplifies knowledge engineering and speeds up computation
  - Often OK at least approximately




Naïve Bayes Model

# Handling Dependency: Adding States to Parent Variable

- Problem: Symptoms not independent given fault
- Solution: Redefine parent variable to create model with independent symptoms
  - Incorporate into states of parent variable conditions that modify relationship between symptoms
- Example
  - P(Malaise|UTI, fever) > P(Malaise | UTI)
  - Redefine UTI states {absent, mild, moderate, severe}
  - P(Malaise | severe UTI, fever) ≈ P(Malaise | Severe UTI)

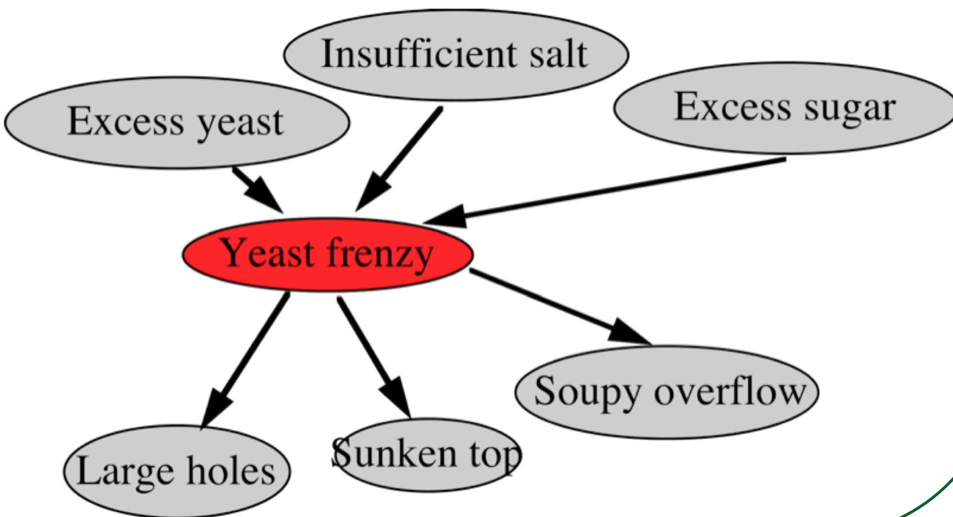



Example courtesy of Mike Shwe

# Handling Dependency: Adding Intermediate Variables

- Intermediate variable is used to model dependency of children given parent
  - Symptoms are independent of fault given intermediate variable
  - Children are dependent given original parents
  - More parsimonious than drawing arcs between symptoms
- Examples:
  - "True state" variable creates conditional independence of sensor reports
  - Intermediate mechanism creates independence among a set of related findings

**Using partitions or ICI can simplify specification of distribution of intermediate random variable**




Example courtesy of Mike Shwe

# Local Distribution Structure

- Local distributions:
  - One distribution for each combination of values of parent variables
  - Assessment is exponential in number of parent variables
  - Assessment can be reduced by exploiting structure
- Examples of local distribution structure
  - Context specific independence: elicitation by partition
  - Independence of causal influence
  - Divorcing

# Context-Specific Independence

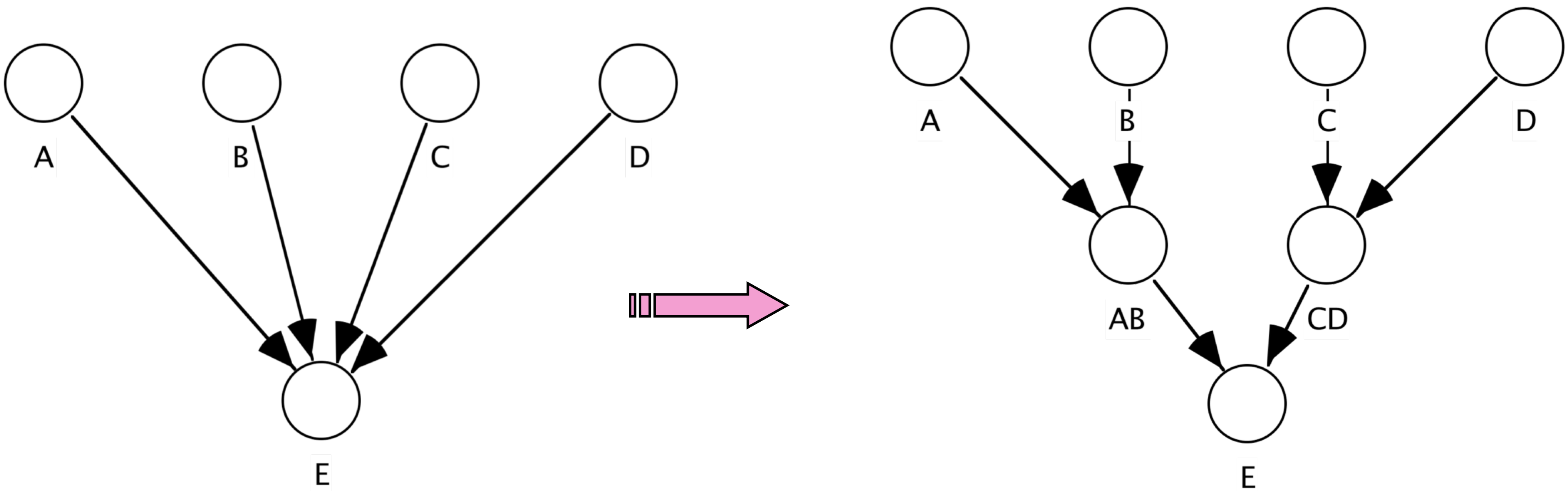
- Partition state set of parents into subsets
  - Set of subsets is called a partition
  - Each subset is a partition element
- Elicit one probability distribution per partition element
- Child is independent of parent given partition element
- Examples:
  - P(reported location | location, sensor-type, weather) independent of sensor type given weather=sunny
  - P(fever=high | disease) is the same for disease $\in$ {flu, measles}
- Knowledge elicitation using partitions
  - For each child variable ask expert to group states of parent variable having same probability distributions
  - Sometimes several children induce same partition
  - Partitions may form basis of type hierarchy
    - » When many indicators induce a common partition element we may name that partition element as a subtype of the parent variable
    - » e.g., acute infectious disease “isa” infectious disease “isa” disease

# Independence of Causal Influence

- Assumption: causal influences operate independently of each other in producing effect
  - Probability that C1 causes effect does not depend on whether C2 is operating
  - Excludes synergy or inhibition
- Examples
  - Noisy logic gates (Noisy-OR, Noisy-AND, Noisy-XOR)
  - Noisy adder
  - Noisy MAX
  - General noisy deterministic function
- Elicitation:
  - ICI structure:
    - » "Does the presence of C2, C3, … increase or decrease [strengthen or weaken] the impact of C1 on E?"
    - » "Does the presence of C2, C3, … increase or decrease the probability C1 will cause E to occur?"
  - Parameters:
    - » ICI structure allows calculation of entire probability table from single-cause distributions

# Divorcing

- Divorcing generalizes partitions and ICI
- An intermediate variable summarizes the effect of a subset of parents on the child



# Exploiting Context-Specific Indepdendence

- Context-specific independence can simplify elicitation
- Example:
  - Government supporters and apolitical people rarely criticize the government. Dissidents often do, as do government agents (because they are trying to lure Rechtian agents into thinking they are dissidents)
  - Need to specify only one distribution given {Supporter, Apolitical} and another probability given {Agent, Dissident}
- Eliciting partitions from expert:
  - Ask expert: “Which variables help to distinguish between supporters and agents?”
  - Expert answers: “Criticizing the government”
  - Ask: “Does this variable give any information to help distinguish between agents and dissidents? Between supporters and apolitical people?”
  - Expert answers: “No”
  - Result: Distribution for criticism is same for agents and dissidents, and same for supporters and apolitical people

# Specialists and Generalists

- In many disciplines experts tend to partition problems into sub-categories exhibiting context-specific independence.
- If we know which sub-category to focus on, we can ignore cues relevant for other categories.
  - Specialists focus on difficult cases in one sub-category
  - Generalists focus on
    - Solving easy cases in any category
    - Diagnosing when to call in a specialist and which specialist to call
- Context-specific independence justifies and supports this strategy
  - Some variables are most useful for sorting cases into sub-categories, and may be (approximately) independent of which hypothesis is correct within sub-category
  - Other cues may be useful for discriminating within sub-categories
- Partitions can represent this type of reasoning

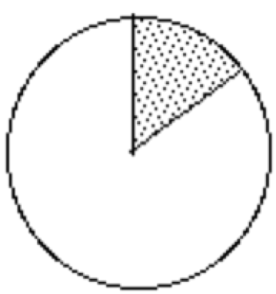
# Assessing Probability Distributions

- Theory
  - Parametric expression may be suggested on theoretical grounds
    - » P(Node|Parents) = f(parents,parameter)
  - Elicit parameter from expert or estimate from observations
  - Test theory against data
- Statistical estimation (when data are available)
  - Frequencies (problem with zero probabilities)
  - Posterior distribution given Dirichlet prior distribution
    - » Uniform prior
    - » Elicit prior frequencies and virtual prior sample size from expert
  - Regression model (discrete or continuous)
  - Independence of causal influence
- Direct elicitation
  - Probabilities
  - Odds (better for extreme probabilities)
  - Continuous distributions: percentiles, density function, parameters

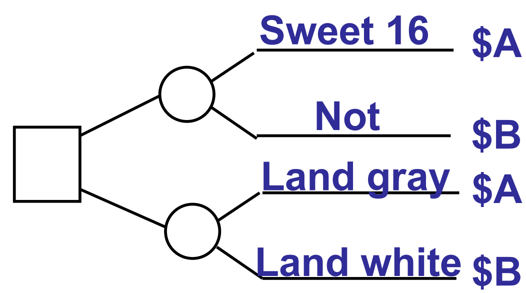
# Direct Elicitation

**Example: Will George Mason University make Sweet 16 in 2018?**

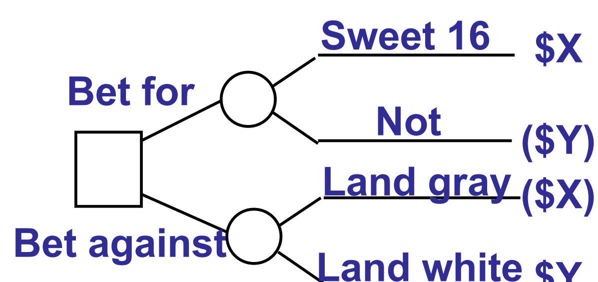
- Direct question: "What is the probability GMU will make Sweet 16 in 2009?"
- Visual representation
  - Adjust the size of the gray area so the probability that we will make Sweet 16 is the same as the probability that a spinner will land in the gray area



- Ask for probability or area on wheel so that decision maker is indifferent between 2 lotteries:




- Ask decision maker about betting odds and solve for probability




  - For small prizes, p=Y/(X+Y)

# Comments

- Many decision makers are uncomfortable with numerical probabilities
  - Decision makers may prefer qualitative terms such as "fairly likely" or "improbable"
  - These phrases have ambiguous meaning and can cause miscommunication unless they are calibrated to agreed-upon numerical values
  - Sometimes visual devices are a good compromise. If manipulated on a computer screen they can be translated directly into numerical probabilities
  - Betting odds are often less useful in practice than asking directly about probabilities - betting odds come from probabilities and not vice versa
  - People find frequencies (3 out of every 100) easier than probabilities (0.03)
  - For very small probabilities orders of magnitude must be used
    - "State a1 is 100 times more likely than State a2"
    - "We will generally see about 100 cases of State a1 for every case of State a2"
  - Assessing very small probabilities is difficult
- Probability assessors should be aware of systematic distortions of probability judgment
  - Treating low-probability events as impossible
  - Overconfidence and other anchoring effects
  - Neglect of base rates
  - Overweighting salient events
- If time permits it is good to phrase questions in multiple ways and to feed back consequences of judgments to decision maker

# Probability Scales: an Empirical Study

- Twenty-three NATO intelligence analysts were given statements about likelihood of events, e.g.:
  - ”It is highly likely that the Soviets will invade Czechoslovakia”
- The basic sentence structure was fixed but qualifiers varied, e.g.:
  - ”It is almost certain that the Soviets will invade Czechoslovakia”
- Analysts indicated percentage they would assign to each statement if it appeared in an intelligence assessment

Source: Barclay, et al (1977)

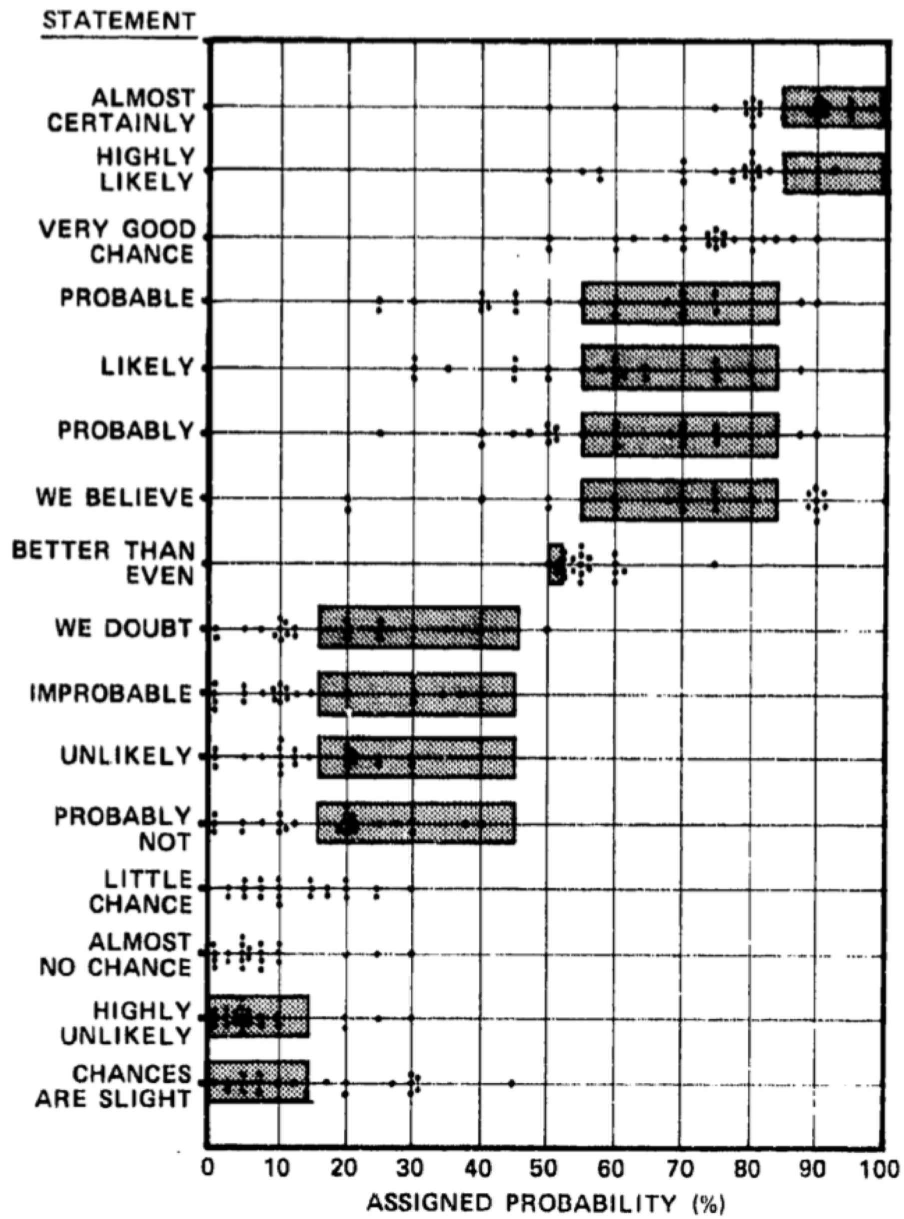



Figure 3·5
WHAT UNCERTAINTY STATEMENTS MEAN TO DIFFERENT READERS

# Assessing Odds

- Direct assessment of probabilities can yield very poor results on extreme probabilities
  - Direct assessment focuses on absolute magnitude
  - 0.01 seems "not much different" from 0.001
  - Orders of magnitude are important in Bayes Rule

$$\frac{P(H_1 | E)}{P(H_2 | E)} = \frac{P(E | H_1)P(H_1)}{P(E | H_2)P(H_2)}$$

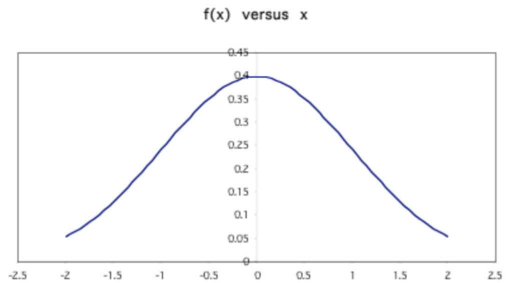
- Assessing by odds
  - "State a1 of variable A is 3 times more likely than state a2"
  - Yields equation P(a1) = 3P(a2)
  - Solve for P(a1) and P(a2)

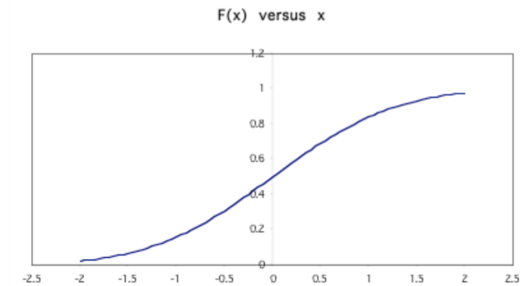
# Continuous Distributions

- Continuous random variables can take on values on a continuum
  - Parametric models (e.g., Normal, Gamma, Chi-square)
  - Nonparametric models
  - "Semi-parametric" models (kernel density functions)
- The cumulative distribution function (cdf)
  - $F(x) = P(X \leq x)$
  - Value of cdf at x is the probability that the random variable is less than or equal to the number x
  - F(x) is a step function for discrete variables, and a smoothly increasing function for continuous variables
- Probability density function (pdf)
  - The pdf measures the relative probability of different values of the continuous random variable. The value $f(x)\Delta x$ is approximately equal to the probability that X lies in the small interval $[x-\Delta x, x+\Delta x]$
  - The pdf is the derivative of the cdf:
    - » $f(x) = \frac{dF(x)}{dx}$
  - The cdf is the integral of the pdf:
    - » $F(x) = \int_{u \leq x} f(u)du$
  - The probability that X lies in the interval [a, b] is equal to the area under the curve defined by the cdf:
  - $P(a \leq X \leq b) = F(b) - F(a) = \int_a^b f(u)du$

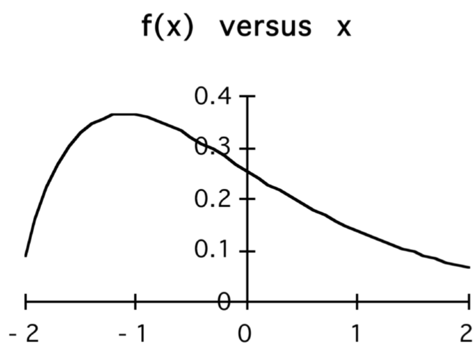
# Example Types of Density Function

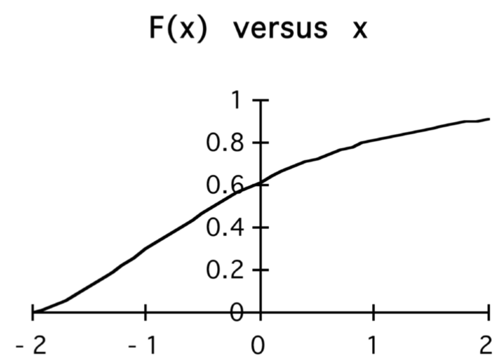
- Symmetric and unimodal



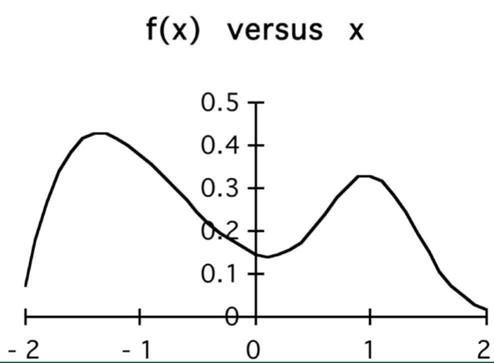



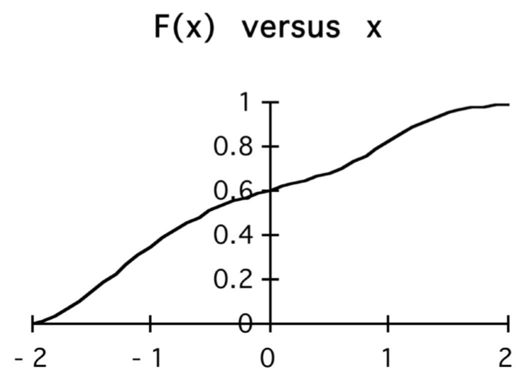

- Symmetric and skewed distribution




- Bimodal




We can show the expert different shapes and ask which best fits judgments

It may help to show density function and report percentiles ("With this density function, 23% of the cases will have value less than 1.5.")

Bimodal distribution may mean a missing parent

# Assessing Continuous Distributions

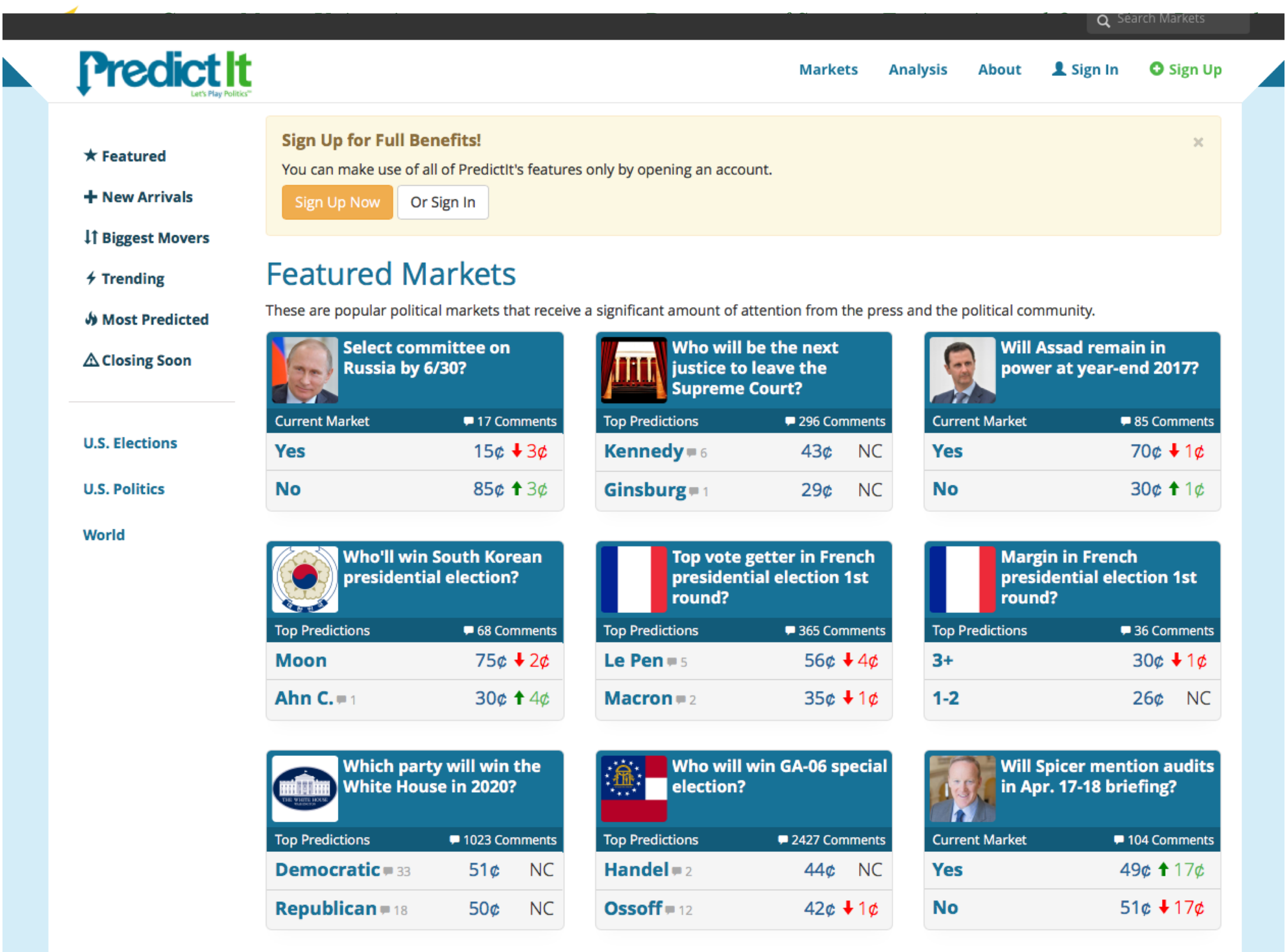
- Continuous distributions are often assessed by asking about the cdf
  - Pick values and assess probability (this method can also be used for discrete random variables) :
    - » P(sales ≤ \$10,000) = ?
    - » P(sales ≤ \$20,000) = ?
    - » P(sales ≤ \$40,000) = ?
    - » P(sales ≤ \$65,000) = ?
    - » P(sales ≤ \$100,000) = ?
  - Pick probabilities and assess values:
    - » P(sales ≤ ?) = .10
    - » P(sales ≤ ?) = .30
    - » P(sales ≤ ?) = .50
    - » P(sales ≤ ?) = .70
    - » P(sales ≤ ?) = .90
- Depending on how the judgments will be used, we may interpolate between these points or we may fit a parameterized probability distribution to the expert's judgments
- This method may be problematic for parameters whose meaning is not straightforward to the decision maker
  - e.g., What is your cdf for the mean number of transmission errors per hour?
- Another method is to ask about shape of density function
- These methods can yield different results
  - Suggestion: use both and resolve inconsistencies

# Parameterized Continuous Distributions

- Use of standard parameterized distributions may greatly economize on elicitation, e.g.:
  - Normal
  - Log-normal
  - Gamma
  - Exponential
- Assess several percentiles and select parameters to fit the percentiles (this may take some work)
- Check that shape of density function is acceptable
- Ask about "sufficient statistics"
  - What do you think is the average number of transmission errors per 8-hour period when averaged over many periods?
- When data are available they can be used to augment expert judgment
  - Use expert judgment to specify a prior distribution
  - Learn posterior distribution for parameters
  - Use structure learning method to check structural assumptions
    - » Partitions, presence/absence of arcs, functional form of distribution

# Combining Models from Multiple Sources

- Problem decomposition
  - Some elements learned from data; others elicited from experts
  - Different experts specify different parts of model
- Aggregating different inputs on same model component
  - There is a large literature on combining probability estimates
  - We can combine estimates from multiple experts, multiple models, or both
  - Typically the combination rule is some kind of average
  - Many weighting schemes have been proposed; it has proven surprisingly difficult to beat simple averaging
- Prediction markets
  - Useful for forecasting well-defined events on which outcome feedback will become available


Search Markets
PredictIt
Let's Play Politics
Markets
Analysis
About
Sign In
Sign Up
Featured
New Arrivals
Biggest Movers
Trending
Most Predicted
Closing Soon
U.S. Elections
U.S. Politics
World
Sign Up for Full Benefits!
You can make use of all of PredictIt's features only by opening an account.
Sign Up Now
Or Sign In
Featured Markets
These are popular political markets that receive a significant amount of attention from the press and the political community.
Select committee on Russia by 6/30?
Current Market 17 Comments
Yes 15¢ 3¢
No 85¢ 3¢
Who will be the next justice to leave the Supreme Court?
Top Predictions 296 Comments
Kennedy 6 43¢ NC
Ginsburg 1 29¢ NC
Will Assad remain in power at year-end 2017?
Current Market 85 Comments
Yes 70¢ 1¢
No 30¢ 1¢
Who'll win South Korean presidential election?
Top Predictions 68 Comments
Moon 75¢ 2¢
Ahn C. 1 30¢ 4¢
Top vote getter in French presidential election 1st round?
Top Predictions 365 Comments
Le Pen 5 56¢ 4¢
Macron 2 35¢ 1¢
Margin in French presidential election 1st round?
Top Predictions 36 Comments
3+ 30¢ 1¢
1-2 26¢ NC
Which party will win the White House in 2020?
Top Predictions 1023 Comments
Democratic 33 51¢ NC
Republican 18 50¢ NC
Who will win GA-06 special election?
Top Predictions 2427 Comments
Handel 2 44¢ NC
Ossoff 12 42¢ 1¢
Will Spicer mention audits in Apr. 17-18 briefing?
Current Market 104 Comments
Yes 49¢ 17¢
No 51¢ 17¢

# Unit 7 Outline

- The Knowledge Acquisition Lifecycle
- Building the Model
- Managing and Evaluating the Model
- Knowledge Engineering for Relational Graphical Models

# Managing Knowledge Acquisition

- Record rationale for modeling decisions
- Develop “style guide” to maintain consistency across multiple subproblems
  - Naming conventions
  - Variable definitions
  - Modeling conventions
- Enforce configuration management
  - History of model versions
  - Protocols for making and logging changes to current model
  - Rationale for changes
- Develop protocol for testing models
  - Record of test results traced to model changes and rationale

# Configuration Management

- Formal process is required for managing evolution of complex models
- Configuration management includes:
  - Archiving history of evolving versions
  - Protocols for making and logging changes to current knowledge base
  - Protocols for documenting changes and rationale
  - Automated comparison of similarities and differences between different versions of knowledge base

# Model Agility

- Requirement: rapid adaptation of model to a new situation
- Support for model agility
  - Libraries of reusable model fragments
  - Documentation of stable and changeable aspects of model
  - Development of data sources for inputs to changeable model components
    - Protocols for data collection and maintenance
    - Protocols for importing data into knowledge base
  - Automated support for propagating impact of changes

# Model Evaluation

- Model walk-through
  - Present completed model to "fresh" experts and/or modelers
  - Evaluate all components of model
- Sensitivity analysis
  - Measures effect of one variable on another
  - Compare with expert intuition to evaluate model
  - Evaluate whether additional modeling is needed
- Case-based evaluation
  - Run model on set of test cases
    - » Cases to check local model fragments (component testing)
    - » Cases to test behavior of global model (whole-model testing)
  - Compare with expert judgment or “ground truth”
  - Important issue: selection of test cases

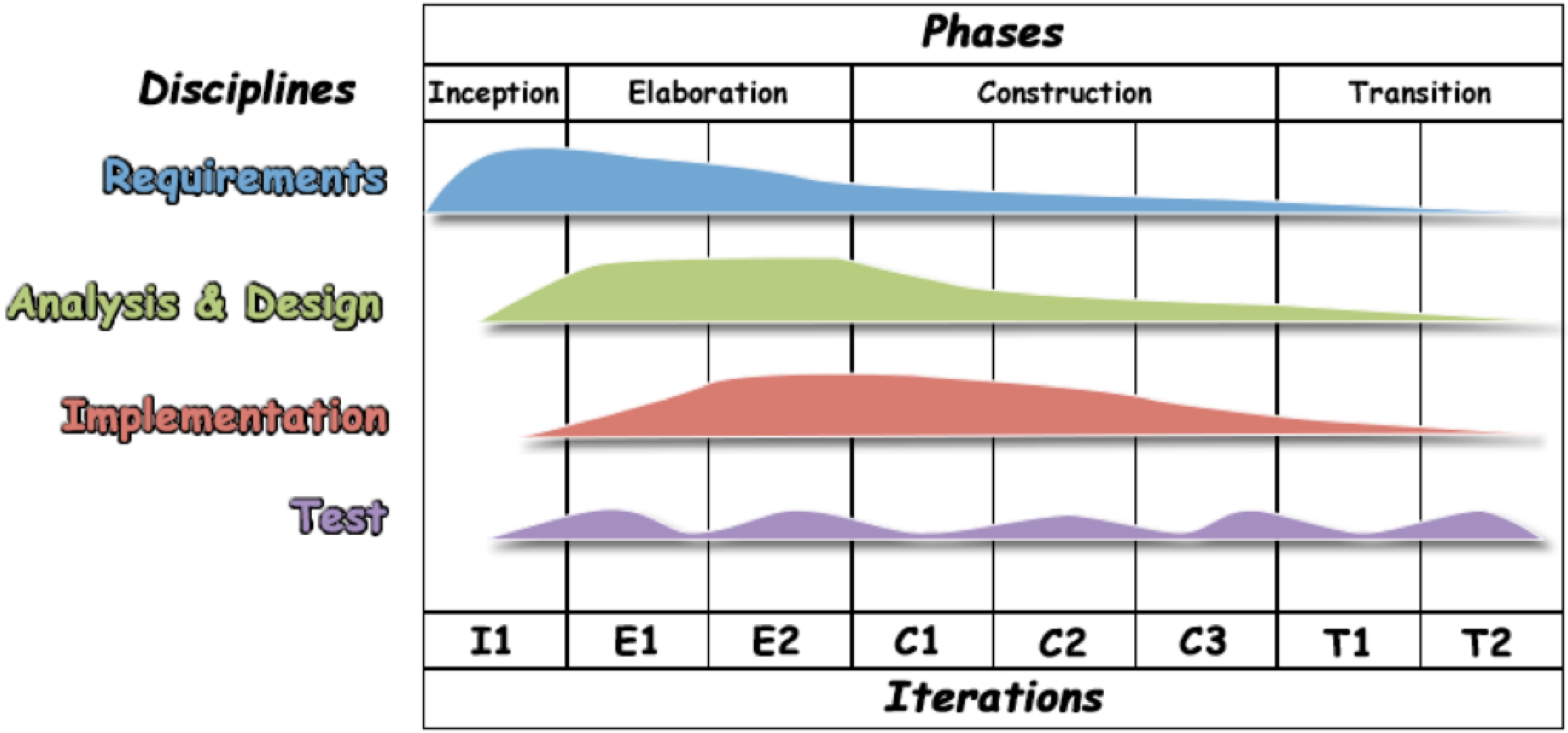
# Unit 7 Outline

- The Knowledge Acquisition Lifecycle
- Building the Model
- Managing and Evaluating the Model
- Knowledge Engineering for Relational Graphical Models

# Knowledge Engineering for Relational Models

- Relational representations (e.g., MEBN, PRM, OOBN) require knowledge about:
  - Entity types (e.g., patients, diseases, tests)
  - Attributes of entities (e.g., gender of patient, sensitivity of test)
  - Relationships among entities (e.g., patients have diseases; patients take tests)
- Entity-relationship model specification is needed for database schema design; object-oriented software design; ontological engineering
  - There is a literature on methodologies for entity-relationship modeling
- Literature in these areas usually does not treat uncertainty
- Growth area in probabilistic modeling
  - Software support is not widely available
  - Experience base is small of people knowledgeable in specifying relational models
  - Literature is still small

# Uncertainty Modeling Process for Semantic Technology (UMP-ST)

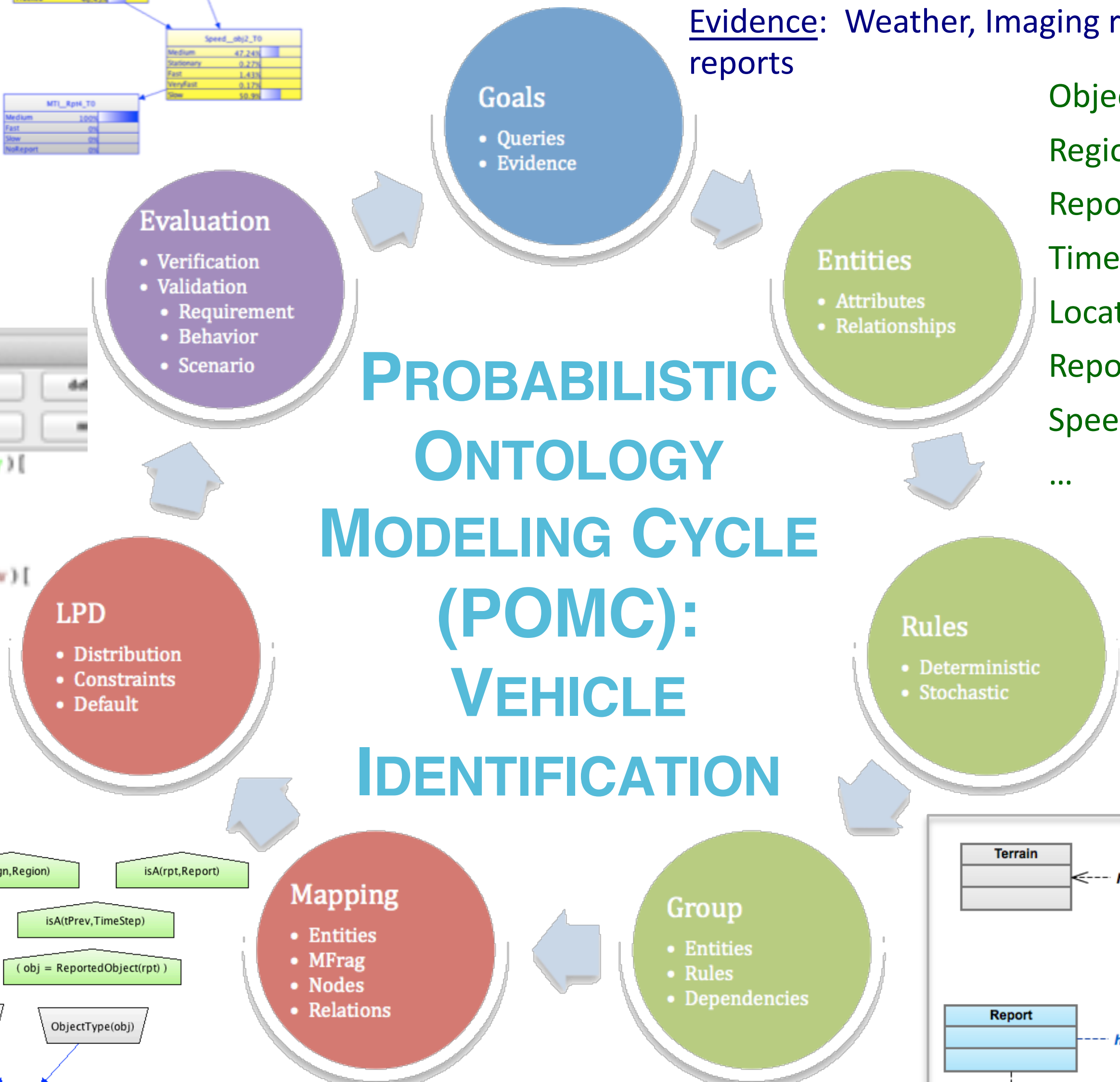



- Guideline for probabilistic ontology developers
- Describes main tasks involved in creating probabilistic ontologies
- Based on Unified Process (UP) for software engineering

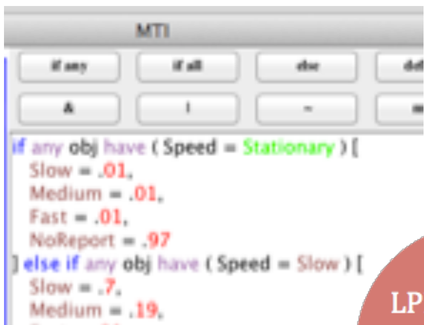
# Probabilistic Ontology Modeling Cycle (POMC): Vehicle Identification

Goal: Identify and type vehicles

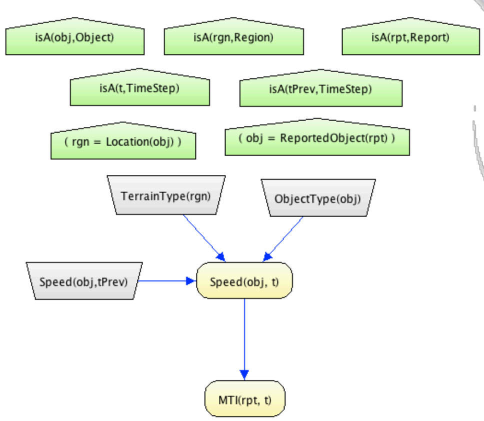
- Query: Is it a tracked vehicle? Wheeled vehicle? How fast is it going?
- Evidence: Weather, Imaging reports, MTI reports, GIS reports

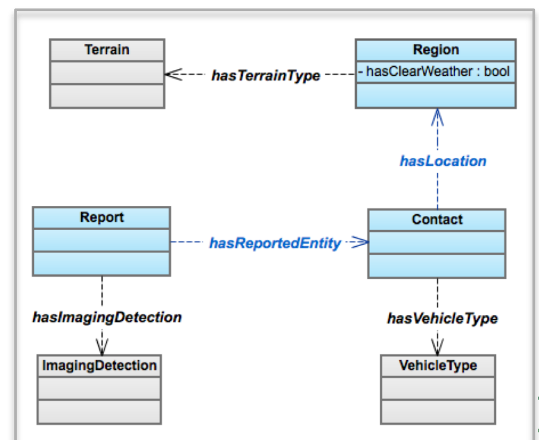



Object
Region
Report
Timestep
Location
ReportedObject
Speed
…



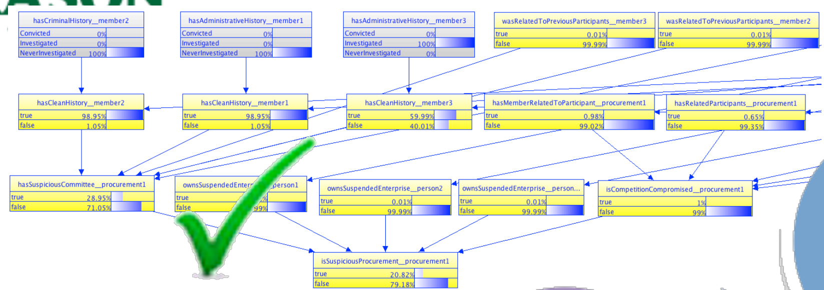

IF an object is moving fast and is on the road,
THEN it is more likely that it is a wheeled vehicle than a tracked vehicle.



# Use Case: Detecting Procurement Fraud

- The Comptroller's office is responsible for inspecting and auditing Brazilian Government projects and programs
  – Provides transparency and helps to prevent corruption
- All contracts with the private sector must follow the national Procurement law
  – Select the most advantageous proposal for a contract in its interest
  – Susceptible to many forms of corruption
- Use case purpose: support information fusion to detect possible fraud in procurements

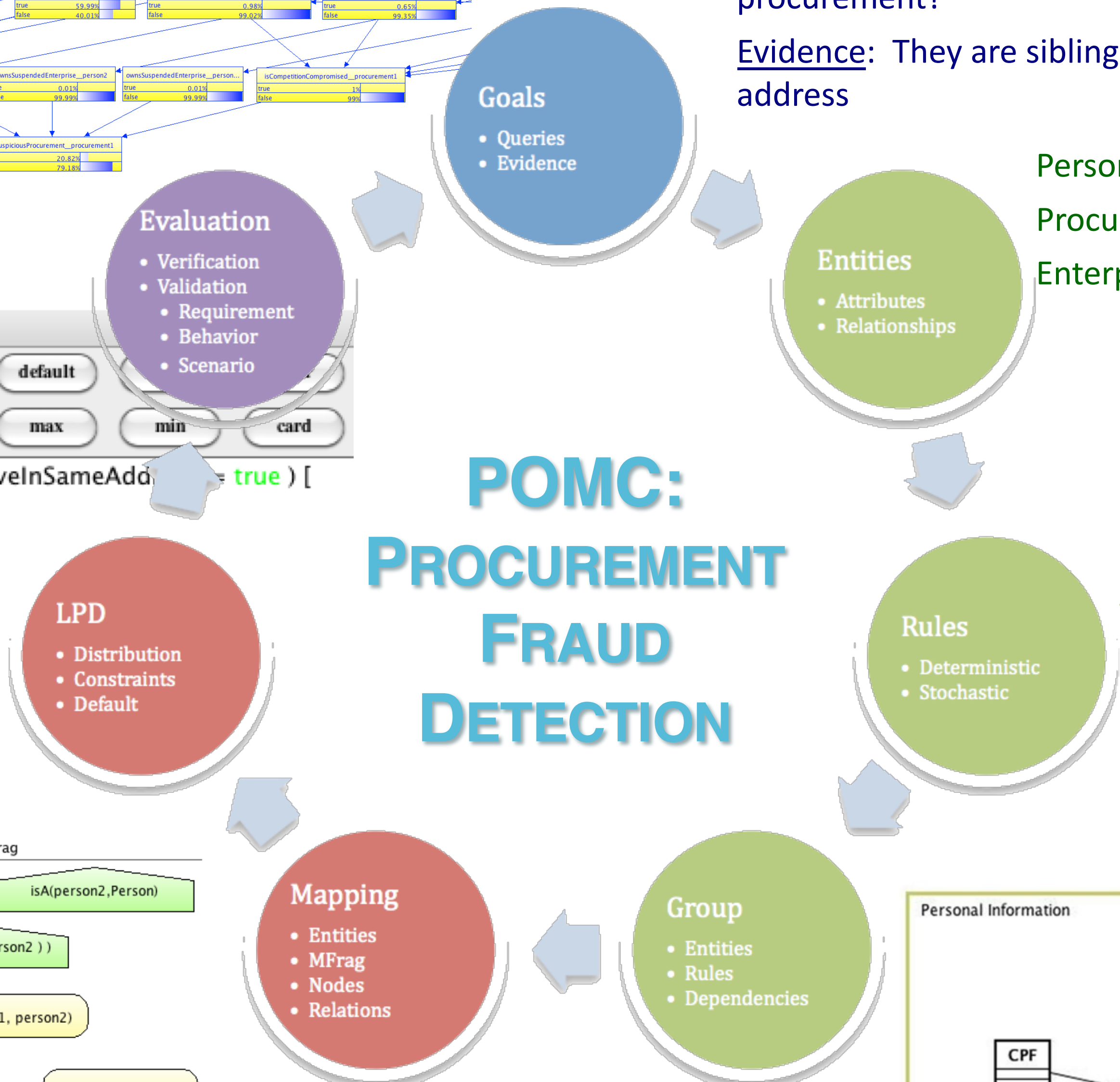


Goal: Find suspicious procurements

Query: Is there any relation between the committee and the enterprises that participated in the procurement?

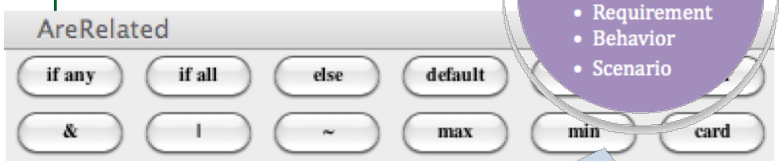
Evidence: They are siblings; They live at the same address

| | |
|---|---|
| Person | ownerOf |
| Procurement | participatesIn |
| Enterprise | livesAt |

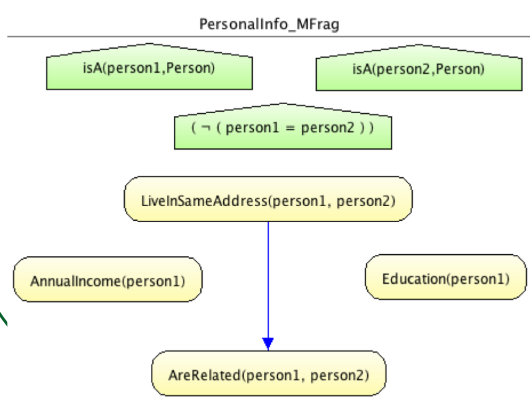


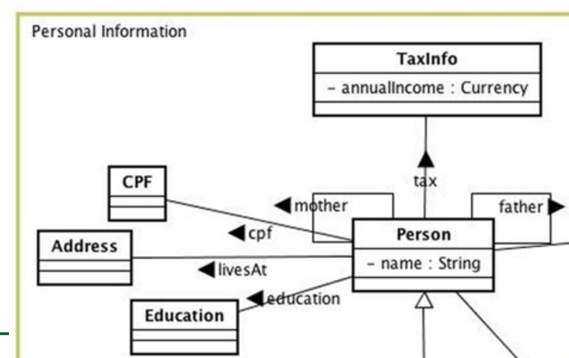

IF a member of the committee lives at the same address as a person responsible for a bidder in the procurement,

THEN it is more likely that a relation exists between the committee and the enterprises, which lowers competition.



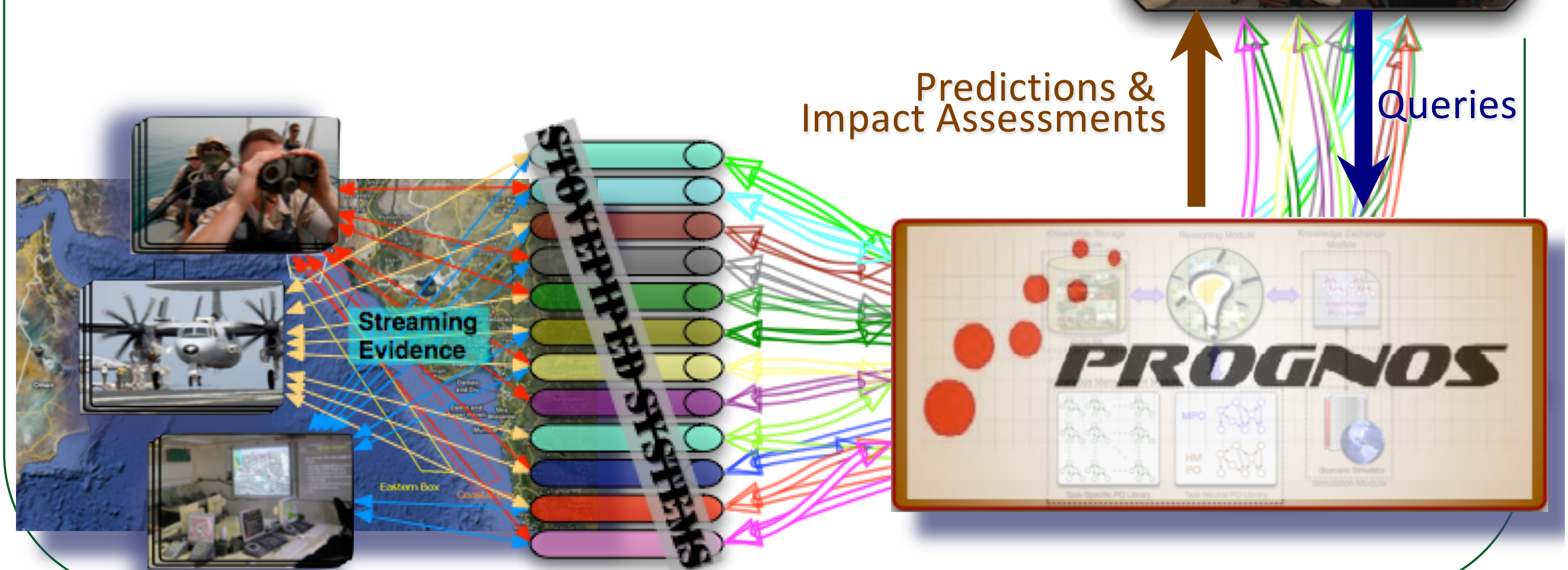

```
if any person1.person2 have ( LiveInSameAdd... = true ) [
  absurd = 0,
  false = .1,
  true = .9
] else [
  absurd = 0,
  false = .999,
  true = .001
]
```

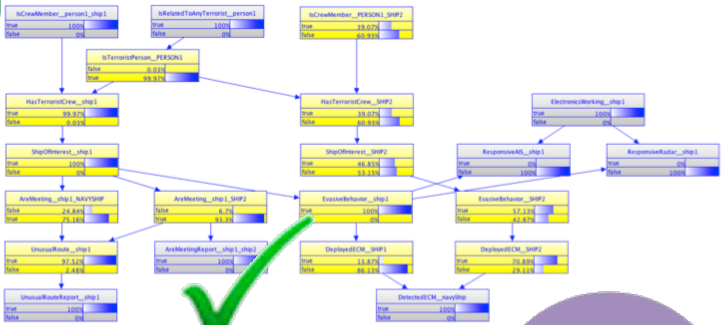


# Use Case: Maritime Security Operations

- Support maritime domain awareness (MDA)
- Enable predictive situational awareness in maritime operations
- Provide higher-level fusion to reason about threats in the maritime domain

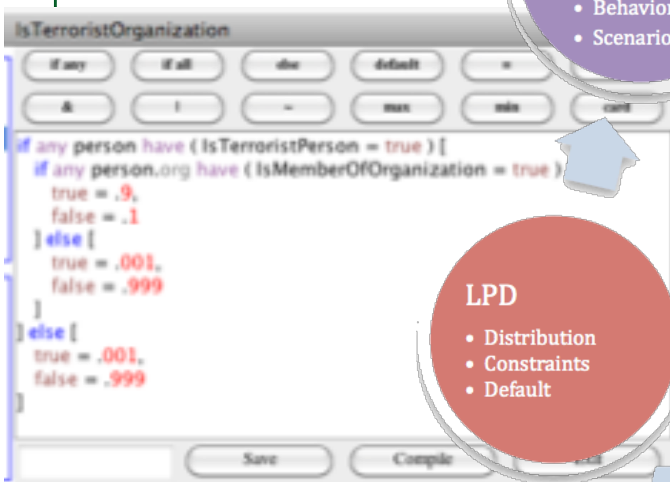


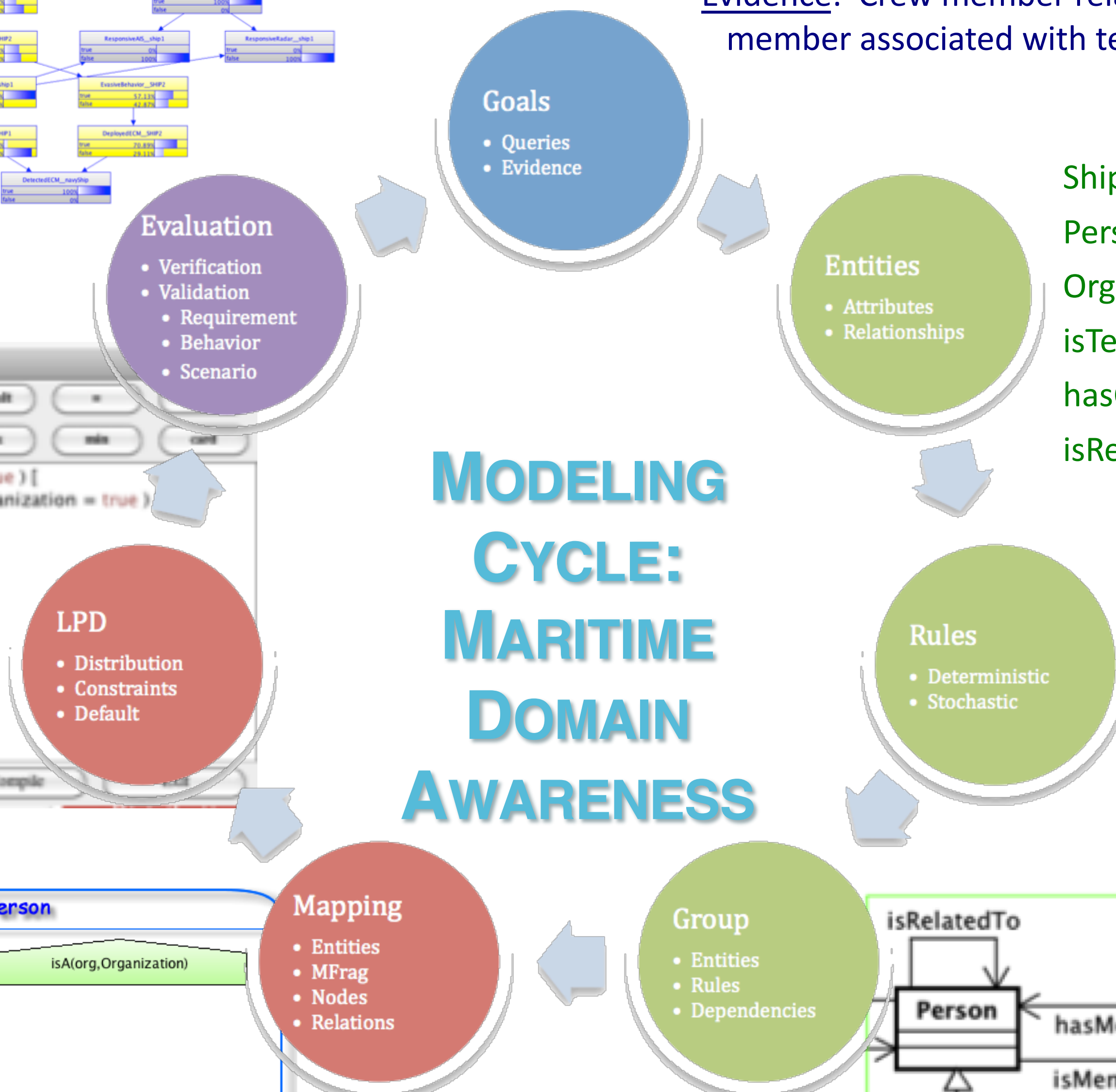


Goal: Identify whether a ship is a ship of interest

Query: Does the ship have a terrorist crew member?

Evidence: Crew member related to any terrorist; Crew member associated with terrorist organization




Ship

Person

Organization

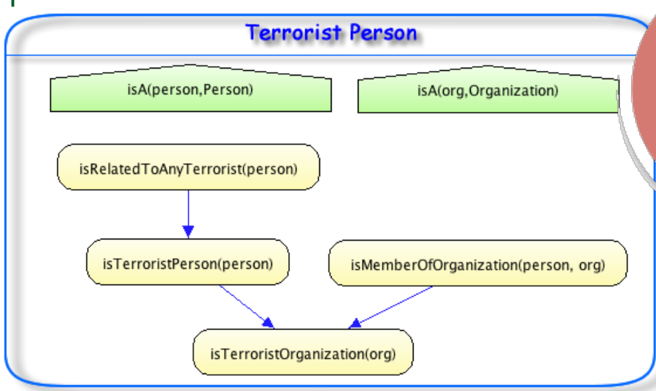
isTerroristPerson

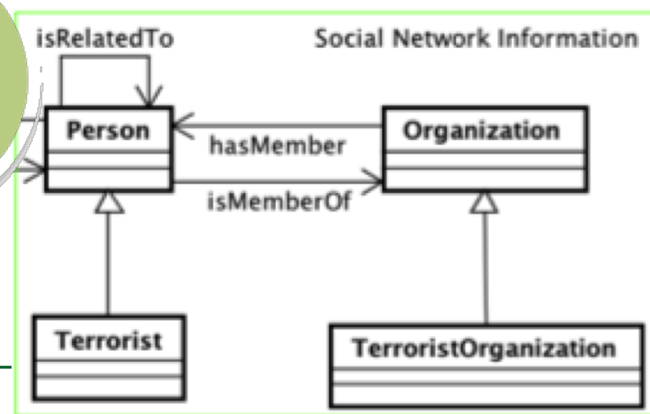
hasCrewMember

isRelatedTo

IF a crew member is related to a terrorist,

THEN it is more likely that he is also a terrorist



# UMP-ST Plugin to UnBBayes

- Provides support for probabilistic ontology creation, evolution and maintenance
- Addresses problems faced by PO designers:
  - Complexity of creating probabilistic ontologies
  - Difficulty of maintaining and evolving probabilistic ontologies
  - Lack of tool for documenting probabilistic ontologies
  - Need for tools to support requirements traceability
- Tool for documenting, maintaining and evolving probabilistic ontologies
  - Provides step by step guidance for creating probabilistic ontologies
  - Provides centralized tool for documenting probabilistic ontologies
  - Enables identification of impact of changes through requirements traceability

# Summary and Synthesis

- Graphical model development is a problem in system lifecycle engineering
  - Begin with a small subproblem
    - » Self-contained
    - » Can be completed in short time
    - » Interesting in its own right
    - » Reasonably representative of global problem
  - Iteratively expand and refine
  - Test and evaluate at each stage
    - » Elicitation review
    - » Sensitivity analysis – quantitative and qualitative
    - » Case-based testing – local and global
    - » Evaluation against empirical data
- Effective management of knowledge engineering process improves
  - Communication between domain experts and knowledge engineers
  - Quality of model
  - Reusability of results
- Software supports are needed for agile lifecycle engineering of probabilistic graphical models

# References for Unit 7


## *The knowledge elicitation process*

- Mahoney, S. and Laskey, K. Network Engineering for Complex Belief Networks, UAI 96
- Morgan, M.G., Henrion, M. and Small, M. Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis, Cambridge University Press, 1990
- Barclay, S., Brown, R., Kelly, C., Peterson, C., Phillips, L. and Selvidge, J., Handbook for Decision Analysis, Decisions and Designs, Inc., 1977 (a classic) https://www.researchgate.net/publication/235066943_Handbook_for_Decision_Analysis
- Pradhan, M., Provan, G., Middleton, B. and Henrion, M. Knowledge Engineering for Large Belief Networks, UAI-94
- Special issue of Journal of Machine Learning Research in 2003 on Knowledge and Data Fusion

## *Relational Knowledge Engineering*

- Carvalho RN, Laskey KB, Costa PCGD. (2016) Uncertainty modeling process for semantic technology. *PeerJ Computer Science* 2:e77 https://doi.org/10.7717/peerj-cs.77.

## *Probability elicitation*

- Special issue of IEEE TKDE: Probability Models: Where do the Numbers Come From?
- Morgan, Henrion, and Small, ibid.
- Barclay, et al., ibid.
- Spetzler, C.S. and Stael von Holstein, C-A.S., Probability Encoding in Decision Analysis, Management Science 22(3), 1975 (a classic)

## *Independence of causal influence*

- Pearl, J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann, 1988 (for noisy OR: not the original source, but easily available)
- Srinivas, S. A Generalization of the Noisy-OR Model, UAI-93

## *Elicitation by partitions*

- Heckerman, D. *Probabilistic Similarity Networks*, MIT Press, 1991
- Mahoney, S.M. and Laskey, K.B. Representing and Combining Partially Specified CPTs. In Laskey, K.B. and Prade, H. (eds.) Uncertainty in Artificial Intelligence: Proceedings of the Fifteenth Conference, San Mateo, CA: Morgan Kaufmann, 1991.
- N.Friedman,M.Goldszmidt, Learning Bayesian Networks with local structure, In Horvitz, E. and Jensen, F. (eds.) Uncertainty in Artificial Intelligence: Proceedings of the Fifteenth Conference, pp 252-262, San Mateo, CA: Morgan Kaufmann, 1996

## *Discretization*

- Clarke, E.J. and Barton, B.A. Entropy and MDL Discretization of Continuous Variables for Bayesian Belief Networks, International Journal of Intelligent Systems 15,1, 2000.