I Sistemi Esperti

L'approccio probabilistico

SISTEMI ESPERTI (o Sistemi Basati Sulla Conoscenza) (1980)

Un sistema **basato sulla conoscenza** è un sistema in grado di risolvere problemi in un **dominio limitato** ma con prestazioni **simili** a quelle di un **esperto** umano del dominio stesso.

Generalmente esamina un largo numero di possibilità e costruisce dinamicamente una soluzione.

*"La potenza di un programma intelligente nel risolvere un problema dipende primariamente dalla **quantità e qualità** di conoscenza che possiede su tale problema". (Feigenbaum)*

Il programma non è un insieme di **istruzioni immutabili** che rappresentano la soluzione del problema, ma un ambiente in cui:

- rappresentare;
- utilizzare;
- modificare;

una **base di conoscenza.**

Caratterizzato dalle seguenti **proprietà:**

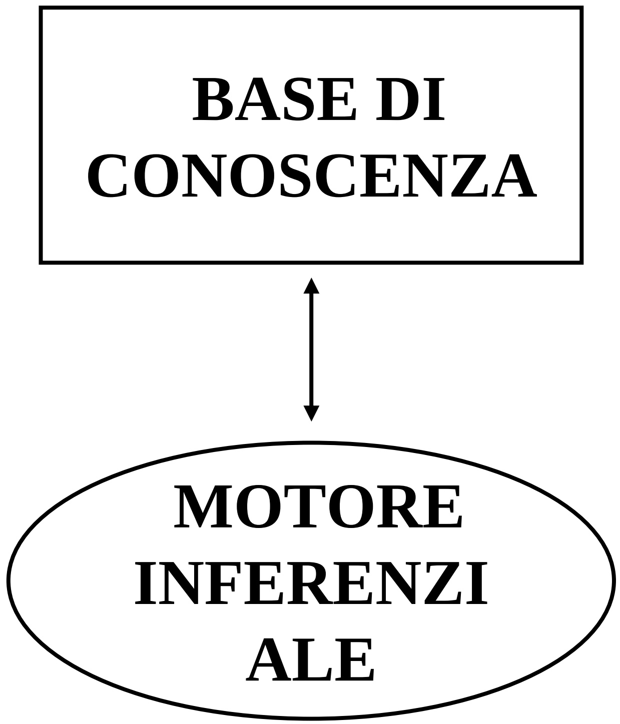
- *Generalità.*
- *Rappresentazione esplicita della conoscenza.*
- *Meccanismi di ragionamento.*
- *Capacità di spiegazione.*
- *Capacità di operare in domini mal strutturati.*

PRINCIPI ARCHITETTURALI

Ogni sistema basato sulla conoscenza deveriuscire ad esprimere **due** tipi di conoscenza in modo **separato** e **modulare:**

Conoscenza sul dominio dell’applica-zione **(COSA);**

Conoscenza su **COME** utilizzare la conoscenza sul dominio per risolvere problemi **(CONTROLLO).**



Problemi:

•Come esprimere la conoscenza sul problema?

•Quale strategia di controllo utilizzare?

SISTEMI DI PRODUZIONE

Un sistema a regole di produzione (production system) è costituito da tre componenti fondamentali:

- **Base di conoscenza a regole** (che prende spesso il nome di “memoria a lungo termine”) in cui sono contenute le regole di produzione;

- **Memoria di lavoro** (memoria a breve termine) in cui sono contenuti i dati e in cui vengono mantenute le conclusioni raggiunte dal sistema;

- **Motore inferenziale.**

  Ogni regola di produzione ha la seguente forma:

```
if <condizione> then
   <conclusione/azione>
```

STRATEGIE DI CONTROLLO

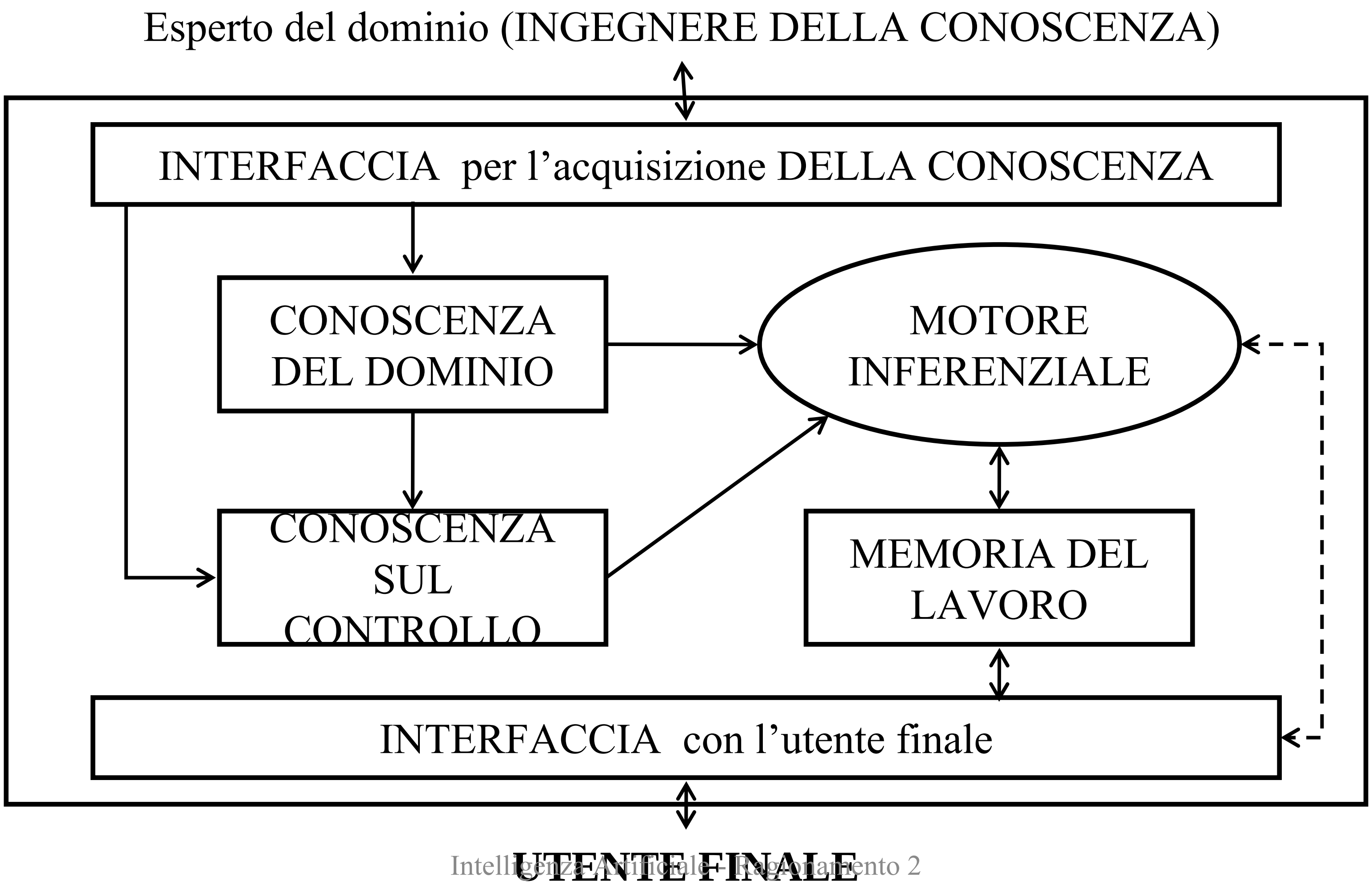
- Strategia ***forward*** (in avanti) o controllo guidato dai dati;

```
while <obiettivo non raggiunto> do
begin
<MATCH:determina l'insieme delle regole
  applicabili (cioè le regole il cui
  antecedente è soddisfatto dai fatti
  contenuti nella memoria di lavoro)>;
<CONFLICT_RESOLUTION: seleziona la regola
  da applicare>;
<FIRE: esegui l'azione associata alla
  regola>
end.
```

- Strategia *"**backward**"* (all'indietro) o controllo guidato dal goal G.

```
if <G è un fatto nella memoria di lavoro>
then <G è dimostrato>
else
begin
<MATCH:seleziona le regole il cui
  conseguente può essere unificato con G>;
<CONFLICT_RESOLUTION: seleziona la regola
  da applicare>;
<l'antecedente della regola selezionata
  diventa il nuovo obiettivo (congiunzione
  di obiettivi) da verificare>
end.
```

SISTEMI BASATI SULLA CONOSCENZA ARCHITETTURA



Rappresentazione della Conoscenza:

- Regole;
- Frames;
- Proposizioni Logiche;
- Vincoli;
- Procedure;
- Demoni;
- Oggetti;
- Fattori di Certezza, Variabili Fuzzy

Modalità di Inferenza:

- Ragionamento *forward;*
- Ragionamento *backward;*
- Risoluzione;
- Propagazione di vincoli;
- Strategie di ricerca euristiche;
- Ragionamento Ipotetico ed Abduttivo....

SVILUPPO DI UN SISTEMA ESPERTO (ingegneria della conoscenza):

AMBIENTI

a) **Skeletal systems** (SHELL)

Ottenuti togliendo da Sistemi Esperti già costruiti la conoscenza propria del dominio e lasciando solo il motore inferenziale e le *facilities* di supporto.

EMYCIN (Empty MYCIN) deriva da MYCIN;
KAS deriva da PROSPECTOR;
EXPERT deriva da CASNET.

b) **Sistemi general-purpose** (TOOLS);

KEE, ART, Knowledge-Craft, Nexpert, KAPPA

non sono strettamente legati a una particolare classe di problemi: essi permettono una più ampia varietà di rappresentazione della conoscenza e strutture di controllo.

c) **Linguaggi simbolici** (Prolog, Lisp) **e non** (C, C++).

PASSI DI PROGETTAZIONE

- **IDENTIFICAZIONE** delle caratteristiche del problema.
- **CONCETTUALIZZAZIONE.**
- **FORMALIZZAZIONE :** progetto della struttura in cui organizzare la conoscenza.
- **IMPLEMENTAZIONE :** scrittura effettiva della conoscenza sul dominio.
- **VALIDAZIONE.**

SCELTA DI UN TOOL:

1. non bisogna utilizzare un Tool più generale del necessario;
2. bisogna testare il Tool costruendo un piccolo prototipo del sistema;
3. bisogna scegliere un Tool che sia affidabile ed mantenuto da chi lo ha sviluppato;
4. quando il tempo di sviluppo è critico, è bene scegliere un Tool con *facilities* di spiegazioni/interazione incorporate;
5. bisogna considerare le caratteristiche del problema per determinare le caratteristiche che deve avere il Tool.

APPLICAZIONI

Migliaia di Sistemi Esperti nei settori più svariati.

Interpretazione: Si analizzano dati complessi e potenzialmente rumorosi per la determinazione del loro significato (Dendral, Hearsay-II).

Diagnosi: Si analizzano dati potenzialmente rumorosi per la determinazione di malattie o errori (Mycin, ...).

Monitoring: I dati si interpretano continuamente per la generazione di allarmi in situazioni critiche. Al sistema è richiesta una risposta in tempo reale soddisfacente (VM).

Planning e Scheduling: Si determina una sequenza intelligente di azioni per raggiungere un determinato obiettivo (Molgen).

OBIETTIVO DELLE CATEGORIE GENERALI DI APPLICAZIONI

Interpretazione: Inferire descrizioni di situazioni da dati rilevati da sensori.

Predizione: Inferire conseguenze future da una data situazione.

Diagnosi: Inferire malfunzionamenti da osservazioni.

Progetto: Configurare oggetti rispettando vincoli.

Pianificazione: Progettare sequenze di azioni.

Monitoring: Confrontare osservazioni in tempo reale per identificare situazioni di allarme.

Debugging: Prescrivere rimedi per malfunzionamenti.

Riparazione: Eseguire un piano per ottenere il rimedio necessario

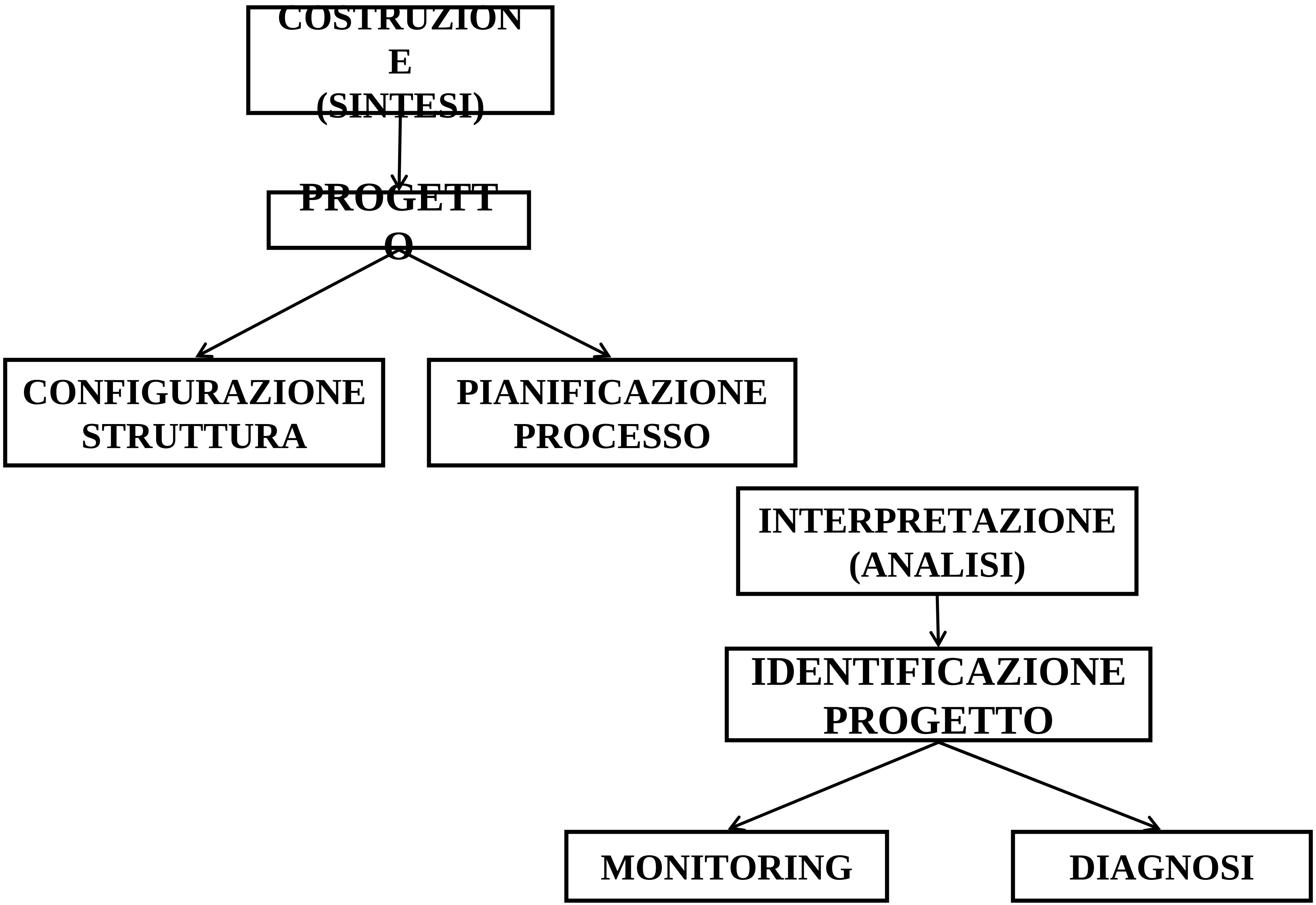
Insegnamento: Diagnosi, Debugging e Riparazione del comportamento di uno studente.

Controllo: Interpretare, Predire, Riparare, a Monitorare il comportamento di un sistema.

Alcune applicazioni ne includono altre.

CLASSIFICAZIONE: SISTEMI DI SINTESI E DI ANALISI

Si possono raggruppare le operazioni in termini di quelli che costruiscono un sistema (sintesi) e quelli che interpretano un sistema (analisi).


COSTRUZIONE (SINTESI)
PROGETTO
CONFIGURAZIONE STRUTTURA
PIANIFICAZIONE PROCESSO
INTERPRETAZIONE (ANALISI)
IDENTIFICAZIONE PROGETTO
MONITORING
DIAGNOSI

CLASSIFICAZIONE E DIAGNOSI (1)

Il più semplice tipo di classificazione consiste nell'identificare alcuni oggetti sconosciuti o fenomeni come appartenenti a una classe conosciuta di oggetti, eventi o processi.

Tipicamente queste classi sono tipi organizzati gerarchicamente e il procedimento di identificazione corrisponde al ***matching*** delle osservazioni di entità sconosciute con caratteristiche note delle classi.

Un esempio: Mycin

Sviluppato da E.M. Shortliffe a partire dal 1972;

Obiettivi:

- decidere se il paziente ha un’infezione che deve essere curata;
- determinare, se sì, quale è probabilmente l'organismo infettivo;
- scegliere fra le medicine adatte per combattere l’infezione quella più appropriata in rapporto alle condizioni del paziente.

Mycin risolve il problema di identificare un oggetto sconosciuto dalle culture di laboratorio, mediante un *matching* dei risultati di laboratorio con la gerarchia di batteri.

([1]) W. J. Clancey, ”Heuristic Classification”, Artificial Intelligence 27, North Holland, 1985 pp 289-350.

CARATTERISTICA ESSENZIALE DELLA CLASSIFICAZIONE

Il motore di inferenza seleziona partendo da un insieme di soluzioni pre-enumerate. Quindi **non costruisce** una nuova soluzione.

Le evidenze (osservazioni) possono essere incerte per cui il sistema può dare come risposta una lista di possibili ipotesi numerate in ordine di plausibilità.

Esistono anche regole di inferenza oltre al *matching*.

In molti problemi le caratteristiche di una soluzione non sono date direttamente come dati, ma sono inferite mediante regole di inferenza che esprimono:

1. **Astrazione definizionale:** basata sulle caratteristiche necessarie di un concetto: “Se è un mammifero allora allatta i figli”.

2. **Astrazione qualitativa:** coinvolge dati quantitativi e li confronta con valori normali: “Se il paziente è adulto e il valore dei globuli bianchi è minore di 2500 allora il valore è basso”

3. **Generalizzazione in una gerarchia di sottotipi**
   “Se il cliente è un giudice, allora è una persona educata”

Classificazione

Classificare un oggetto significa riconoscerlo come appartenente ad una determinata classe.

Una caratteristica essenziale della classificazione è che seleziona da un insieme predefinito di soluzioni.

Le classi identificano delle regolarità e tutti i membri di una classe le condividono (condizioni necessarie).

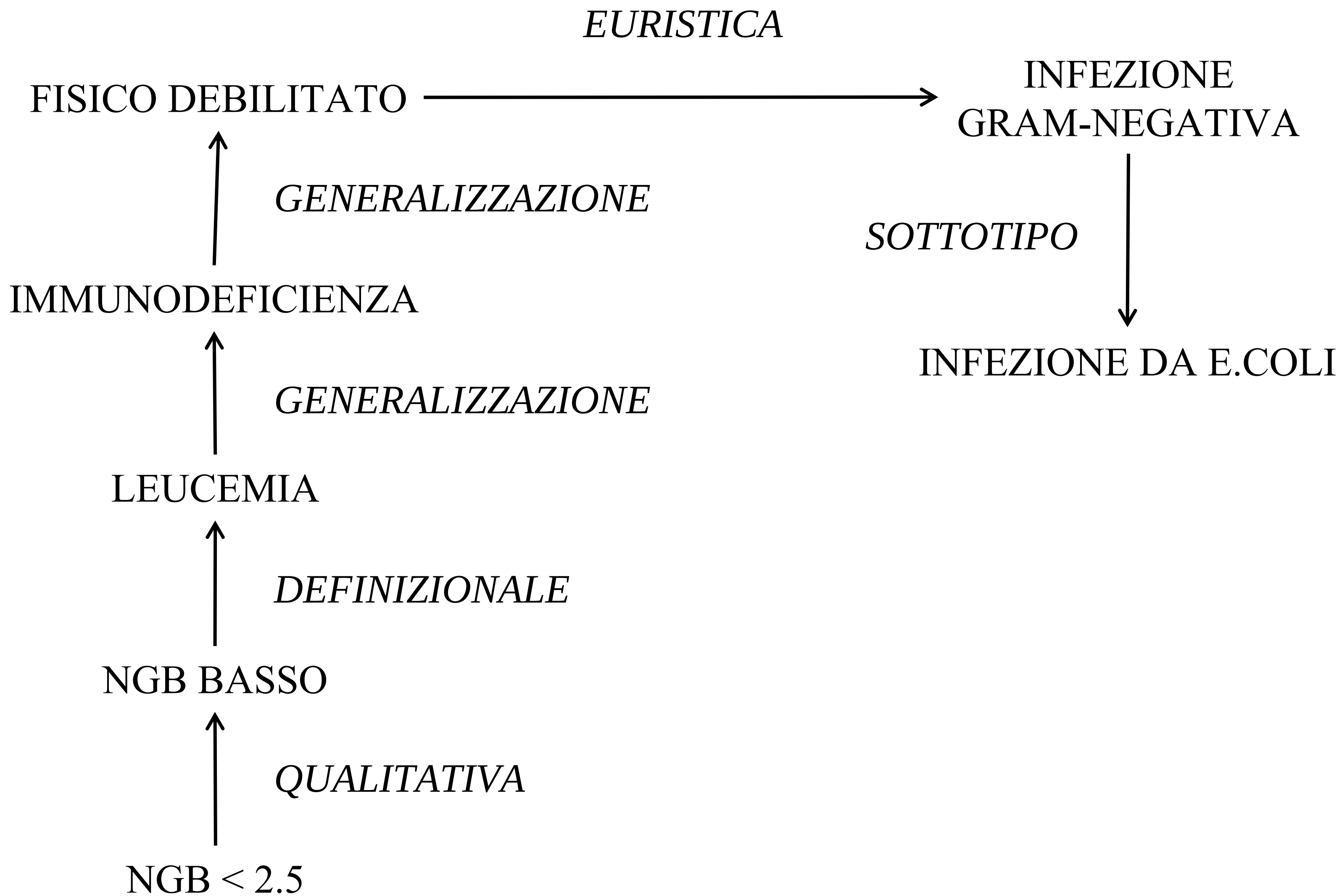
Di solito escludiamo il caso, denominato configurazione, in cui le soluzioni sono un insieme finito, ma molto ampio e quindi determinato dinamicamente come insieme potenza di elementi più semplici (componenti).

CLASSIFICAZIONE SEMPLICE ED EURISTICA

Nella classificazione semplice i dati hanno un ***match diretto*** con le caratteristiche delle soluzioni (eventualmente dopo un passo di astrazione).

Nella classificazione euristica le soluzioni possono anche essere trovate usando un ***matching euristico*** mediante un’associazione diretta e euristica con la gerarchia delle soluzioni.

Normalmente l’associazione euristica è di tipo empirico.


EURISTICA
FISICO DEBILITATO
INFEZIONE
GRAM-NEGATIVA
GENERALIZZAZIONE
SOTTOTIPO
IMMUNODEFICIENZA
INFEZIONE DA E.COLI
GENERALIZZAZIONE
LEUCEMIA
DEFINIZIONALE
NGB BASSO
QUALITATIVA
NGB < 2.5

In pratica, nella classificazione euristica i dati del problema opportunamente astratti sono associati con classi di soluzioni di problemi.

Esempi: Mycin, Grundy che seleziona i libri che una persona può preferire mediante una classificazione della personalità e poi del libro che può essere più adatto, Sophie che classifica un circuito elettronico in termini dei componenti che causano un malfunzionamento.

FATTORI DI CERTEZZA

Compromesso rispetto a un sistema bayesiano puro. Introdotti per la prima volta nel Sistema Esperto Mycin.

Regole non certe (l'implicazione corretta sarebbe invertita) mediate da un **fattore di certezza.**

Un fattore di certezza è un numero reale "ad hoc" che varia fra +1 e -1.

Permette l'inserimento di fatti apparentemente contradditori, ambedue plausibili con differenti valori di certezza.

Fatti:

```
(<attributo> <entità> <valore>
  <fattore di certezza>.
```

Esempi di fatti espressi in Mycin sono (Lisp):

```
(SITE CULTURE-1 BLOOD 1.0)
(IDENT ORGANISM-2 KLEBSIELLA .25)
(IDENT ORGANISM-2 E.COLI 0.73)
(SENSITIVS ORGANISM-1 PENICILLIN -
  1.0)
```

REGOLE:

```
PREMISE <premessa> ACTION <azione>

PREMISE
($AND
(SAME CNTXT INFECT PRIMARY-
  BACTEREMIA)
(MEMBF CNTXT SITE STERILESITES)
(SAME CNTXT PORTAL G1))
ACTION
(CNTXT IDENT BACTEROIDES 0.7)
```

Significato della regola

<u>IF</u>

1)the infection is primary-bacteremia,

2)the site of the culture is one of sterilesites, and

3)the suspected portal of entry of the organism is the gastro-intestinal tract,

<u>THEN</u> there is a suggestive evidence (0.7) that the identity of the organism is bacteroides.

OSSERVAZIONI

I fattori di certezza iniziali sono forniti dagli esperti.

Ogni CF in una regola di Mycin rappresenta il contributo della regola al fattore di confidenza di un’ipotesi.

Rappresenta in un certo senso una probabilità condizionale **P(H|E)**.

In un sistema Bayesiano puro però si deve assumereche la sola evidenza rilevante per **H** sia **E**, altrimenti dobbiamo tenere conto delle probabilità congiunte.

Dunque Mycin assume che tutte le regole siano indipendenti ed è colui che scrive le regole che deve garantire ciò.

R1:

Sviluppato all'Università Carnegie-Mellon da John Mc Dermott per conto della Digital a partire dal 1978.

COMPITO PRINCIPALE: configurare il calcolatore VAX-11 /780 automaticamente.

In base all’ordine del cliente, R1 è in grado di:

- assicurare che l'ordine sia completo;
- determinare le relazioni spaziali fra le componenti.

Un tipico sistema ha più di 100 componenti con varie possibilità di interazione.

Esperti: technical editors

* Nato con un nucleo di 250 regole ora ne possiede circa 2800.

Dal 1980 è un prodotto funzionante Digital.

R1 è implementato in OPS-5.

TIPI DI CONOSCENZA IN R1

a) **Informazione sui componenti:**
in memoria di massa sono raccolte informazioni su circa 400 componenti della Digital che il Sistema va a recuperare quando necessario.

```
(RK711-EA
  !class bundle,
  !type disk drive
  !supported yes
  !component-list 1 070-12292-25
                  1 RK07-EA*
                  1 RK611)
```

b) Conoscenza sui vincoli:
è la conoscenza di come associare determinati componenti per formare configurazioni parziali corrette e di come associare fra di loro le configurazioni parziali

Distributed-mb-devices-3

```
IF: current active context is distributing
    massbus devices
AND there is a single port disk drive that
    has not been assigned to a massbus
AND there are no unassigned dual port disk
    drives
AND …
THEN: assign the disk drive to the massbus
```

c) **MEMORIA DI LAVORO:**
tiene traccia della conoscenza che viene accumulata dinamicamente durante il processo di configurazione.

CONTROLLO IN R1

Forward

Strategia Irrevocabile

Tasks non interagenti → sistema di produzione decomponibile

Nessuna strategia di controllo esplicita

Ambiente povero

Regole per il cambio di contesto

Check-voltage-and-frequency-1

IF: the MOST CURRENT ACTIVE CONTEXT is checking voltage and frequency

AND there is a component that requires one voltage or frequency

AND there is another component that requires a different voltage or frequency

THEN: ENTER THE CONTEXT of fixing voltage or frequency mismatches.

PROSPETTIVE

Verso gruppi di agenti intelligenti cooperanti (ingegneria del software).

Cooperazione con un essere umano

- interfacce grafiche.

Integrazione con vari sistemi informatici (interfacce grafiche, database ecc…), tecnologie tradizionali e non (reti neurali, sistemi fuzzy)

- sistemi ibridi.

Da costose workstations special-purpose (LISP-Machine) verso sistemi Unix o PC con l'utilizzo di Tools scritti in linguaggi più tradizionali (C, C++).

Sistemi basati sulla conoscenza in aree più "semplici": costruzione di libri e manuali non più passivi ma **attivi.**

In fondo, i linguaggi Object-Oriented sono nati nel 1967 e i Database Relazionali negli anni 70…

L’approccio probabilistico
Teoria della probabilità (richiami)

P(*A*): probabilità **incondizionata** (o probabilità **a priori**): quella che si ha *in mancanza di qualsiasi altra informazione*.

Esempio:

P(*Carie*) = 0.1

(In una popolazione, il 10% degli individui ha la carie).

Le proposizioni possono includere uguaglianze che comprendono le cosiddette **variabili casuali**.

Esempio:

P(*Tempo* = *Soleggiato*) = 0.7

P(*Tempo* = *Piovoso*) = 0.2

P(*Tempo* = *Nuvoloso*) = 0.08

P(*Tempo* = *Nevoso*) = 0.02

Ogni variabile casuale X ha un **dominio** di possibili valori ($x_{1, \ldots ,} x_n$).

Si possono vedere i simboli proposizionali come variabili causali se si assume che abbiano dominio (*vero*, *falso*).

Quindi P(*Carie*) è un'abbreviazione di P(*Carie* = *vero*) e P(¬*Carie*) sta per P(*Carie* = *falso*).

Per denotare tutti i valori che può assumere una variabile (discreta) si utilizza una forma vettoriale $\mathbf{P}(A)$. Per l'esempio precedente:

$$\mathbf{P}(Tempo) = \langle 0.7, 0.2, 0.08, 0.02 \rangle$$

Questo enunciato definisce una **distribuzione di probabilità** della variabile *Tempo*.

L’espressione **P**(*Tempo*, *Carie*) denota le probabilità di tutte le combinazioni di valori dell’insieme di variabili casuali. Nell’esempio, è una tavola di 4 x 2 probabilità.

P(*A*|*B*): probabilità **condizionata** o **a posteriori**: probabilità di *A* dato che *tutto ciò che sappiamo* è *B*.

Esempio:

P(*Carie*|*MalDiDenti*) = 0.8

(se il paziente ha mal di denti e non c’è altra informazione, la probabilità che abbia la carie è dell’80%).

La notazione **P**(*X*|*Y*) corrisponde ad una tabella bidimensionale con i valori di P(X=x_i|Y=y_j) per ogni possibile *i*, *j*.

Vale la **regola del prodotto**:

$$\mathrm{P}(A \wedge B) = \mathrm{P}(A|B)\ \mathrm{P}(B)$$

(affinché A e B siano veri, occorre che sia vero B e quindi che A sia vero dato B).

Dalla precedente:

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

Vale anche:

$$\mathrm{P}(A \wedge B) = \mathrm{P}(B|A)\ \mathrm{P}(A)$$

Assiomi della teoria della probabilità

1) $0 \leq P(A) \leq 1$

2) Le proposizioni necessariamente vere (cioè valide) hanno probabilità 1, quelle necessariamente false (cioè insoddisfacibili) hanno probabilità 0; in formule: $P(Vero) = 1$ e $P(Falso) = 0$.

3) $P(A \lor B) = P(A) + P(B) - P(A \land B)$
(si desume dal diagramma di Venn).

Dagli assiomi si può dedurre, ad esempio, che

$$P(A \lor \neg A) = P(A) + P(\neg A) - P(A \land \neg A)$$

$$P(Vero) = P(A) + P(\neg A) - P(Falso)$$

$$1 = P(A) + P(\neg A)$$

$$P(\neg A) = 1 - P(A)$$

NOTA: la probabilità delle proposizioni può esprimere, più che una mera statistica, il grado di *credenza* su un fatto da parte di un agente: deve comunque conformarsi agli assiomi della probabilità (tesi di Bruno De Finetti, gioco delle scommesse tra due agenti).

Probabilità soggettiva

Esempio: *Qual è la probabilità che l'Inter vinca il campionato?*

- Approccio classico:

  P(I) = 1/18 (insoddisfacente)

- Approccio frequentista:
  impraticabile: non si può ripetere il campionato N volte.

De Finetti propone questo approccio:

"La probabilità P(E) di un evento E è la frazione di una somma m che un agente ritiene equo sia pagato per la promessa di ricevere in cambio la somma m nel caso E si verifichi" (**definizione**).

In pratica:

- Lo scommettitore (agente) entra in gioco pagando $m \cdot P(E)$ euro.
- Il bookmaker pagherà m euro se E si verifica, altrimenti niente.

La probabilità così definita corrisponde al grado soggettivo di aspettativa (*degree of belief*) del verificarsi dell'evento incerto *E*.

Si possono avere due situazioni:

- esiste una modalità di scommettere tale che l'agente perde sicuramente (*Dutch book*);
- non esiste tale modalità: in questo caso l'insieme di *beliefs* è **coerente**.

De Finetti dimostra che se un insieme di *beliefs* è coerente (ossia evita un *Dutch book*) allora tali *beliefs* devono necessariamente soddisfare gli assiomi della probabilità.

Distribuzione di probabilità congiunta

Siano $X_1 \ldots X_n$ variabili casuali che assumono valori determinati in un dominio, ciascuna con una certa probabilità.

Un **evento atomico** è un'assegnazione di valori particolari a tutte le variabili.

La distribuzione di probabilità congiunta $\mathbf{P}(X_1 \ldots X_n)$ assegna probabilità a tutti gli eventi atomici (è una tabella *n*-dimensionale). Esempio:

	MalDiDenti	¬*MalDiDenti*
Carie	0.04	0.06
¬*Carie*	0.01	0.89

Dalla distribuzione di probabilità si desume, ad esempio:

$P(Carie \vee MalDiDenti) = 0.04 + 0.01 + 0.06 = 0.11$

oppure

$$P(Carie \mid MalDiDenti) =$$

$$\frac{P(Carie \wedge MalDiDenti)}{P(MalDiDenti)} =$$

$$\frac{0.04}{0.04 + 0.01} = 0.80$$

In generale, per *n* variabili booleane, occorre specificare 2^n valori per la distribuzione di probabilità.

Regola di Bayes

Da:

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

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

uguagliando le parti a destra, si ha:

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)}$$

che è detta *regola di Bayes*.

Per variabili multivalori si ha:

$\mathbf{P}(Y|X) = \mathbf{P}(X|Y)\ \mathbf{P}(Y)\ /\ \mathbf{P}(X)$

(da intendere come insieme di equazioni che legano elementi corrispondenti delle tabelle).

Vale anche:

$\mathbf{P}(Y|X,E) = \mathbf{P}(X|Y,E)\ \mathbf{P}(Y|E)\ /\ \mathbf{P}(X|E)$

(dove E rappresentano le prove di fondo).

Esempio:

Un dottore sa che la meningite induce nel paziente, nel 50% dei casi, un irrigidimento del collo.

Sa anche che la probabilità a priori che un paziente abbia la meningite è 1/50.000, quella che un paziente abbia un irrigidimento del collo è 1/20.

Sia

S = paziente ha irrigidimento del collo,
M = paziente ha meningite.

Allora:

P(S|M) = 0.5

P(M) = 1/50000

P(S) = 1/20

Pertanto:

$$P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)} = \frac{0.5 \times \frac{1}{50000}}{\frac{1}{20}} = \frac{1}{5000}$$

Si osservi che P(S|M) riflette il funzionamento della meningite, mentre P(M|S) può variare, se ad esempio c'è un'epidemia di meningite (in qual caso P(M) cresce).

Normalizzazione

Si è visto che:

$$P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)}$$

La possibilità che un paziente abbia un irrigidimento del collo per un colpo di frusta è:

$$P(W \mid S) = \frac{P(S \mid W)P(W)}{P(S)}$$

La **verosimiglianza relativa** della meningite e del colpo di frusta, dato un irrigidimento del collo, non richiede la probabilità a priori di un irrigidimento del collo:

$$\frac{P(M \mid S)}{P(W \mid S)} = \frac{P(S \mid M)P(M)}{P(S \mid W)P(W)}$$

Un altro caso in cui non si deve valutare direttamente la probabilità a priori dei "sintomi" si ha quando si considera un insieme esaustivo di casi. Nell'esempio precedente, si considera il caso M e il caso $\neg M$:

$$P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)}$$

$$P(\neg M \mid S) = \frac{P(S \mid \neg M)P(\neg M)}{P(S)}$$

Ora è anche $P(M|S) + P(\neg M|S) = 1$, quindi:

$P(S) = P(S|M)\ P(M) + P(S|\neg M)\ P(\neg M)$

e sostituendo sopra:

$$P(M \mid S) = \frac{P(S \mid M) P(M)}{P(S \mid M) P(M) + P(S \mid \neg M) P(\neg M)}$$

Questo processo è detto **normalizzazione**, poiché tratta 1/P(S) come costante di normalizzazione che permette ai termini condizionali di dare 1 come somma (in altri termini, si preferisce calcolare P($S|\neg M$) piuttosto che P(S)).

Nel caso generale multivalore, la regola di Bayes diventa

$$\mathbf{P}(Y|X) = \alpha\mathbf{P}(X|Y)\,\mathbf{P}(Y)$$

dove α è la costante di normalizzazione necessaria per rendere la somma dei valori della tabella $\mathbf{P}(Y|X)$ pari a 1.

La combinazione delle prove

Si suppongano note le due probabilità condizionate:

$P(Carie|MalDiDenti) = 0.8$

$P(Carie|Colpire) = 0.95$

e si voglia concludere la probabilità che il sondino colpisca il dente dolorante del paziente.

Per calcolare

P($Carie|MalDiDenti \wedge Colpire$)

occorerebbe conoscere la probabilità congiunta.

Anche l'applicazione diretta della regola di Bayes:

$$P(Carie \mid MalDiDenti \wedge Colpire =$$
$$P(Carie) \cdot \frac{P(MalDiDenti \wedge Colpire \mid Carie)}{P(MalDiDenti \wedge Carie)}$$

non aiuta, poiché richiede la conoscenza della probabilità condizionata

(in generale, date n variabili individuali differenti, occorre calcolare n^2 valori differenti, e una diagnosi potrebbe dipendere da dozzine di variabili, non solo due!).

In molti domini è possibile semplificare l'applicazione della regola di Bayes mediante un processo detto **aggionamento bayesiano**, il quale *incorpora prove un pezzo per volta*, *modificando la credenza precedente nella variabile ignota*.

Nel caso in esame, si parte dal considerare il solo *MalDiDenti*:

$$P(Carie \mid MalDiDenti) = P(Carie) \cdot \frac{P(MalDiDenti \mid Carie)}{P(MalDiDenti)}$$

Quando si considera *Colpire*, si applica la regola di Bayes con *MalDiDenti* come "contesto costante condizionante":

$$P(Carie \mid MalDiDenti \wedge Colpire) =$$

$$P(Carie \mid MalDiDenti) \cdot \frac{P(Colpire \mid MalDiDenti \wedge Carie)}{P(Colpire \mid MalDiDenti)} =$$

$$P(Carie) \cdot \frac{P(MalDiDenti \mid Carie)}{P(MalDiDenti)} \cdot \frac{P(Colpire \mid MalDiDenti \wedge Carie)}{P(Colpire \mid MalDiDenti)}$$

Per calcolare

$P(Colpire|MalDiDenti \wedge Carie)$

occorre un'assunzione sostanziale, che parte dalla considerazione che la carie è la *causa diretta* sia del mal di denti che del colpire la carie con il sondino:

pertanto, se si sa che il paziente ha la carie, la probabilità del sondino di colpire la carie non dipende dalla presenza del mal di denti;

così, il sondino che finisce nella carie non cambia la probabilità che la carie sia causa del mal di denti.

In formule:

$P(\mathit{Colpire}|\mathit{Carie} \wedge \mathit{MalDiDenti}) =$
$P(\mathit{Colpire}|\mathit{Carie})$

e

$P(\mathit{MalDiDenti}|\mathit{Carie} \wedge \mathit{Colpire}) =$
$P(\mathit{MalDiDenti}|\mathit{Carie})$

equazioni che esprimono l'**indipen-denza condizionale** di *MalDiDenti* e *Colpire* dato *Carie*.

L’espressione di prima diventa:

$$P(Carie \mid MalDiDenti \wedge Colpire) =$$

$$P(Carie) \cdot \frac{P(MalDiDenti \mid Carie)}{P(MalDiDenti)} \cdot \frac{P(Colpire \mid Carie)}{P(Colpire \mid MalDiDenti)}$$

Si noti che il termine

P(*Colpire*|*MalDiDenti*)

che coinvolge le coppie (in generale, triple, ecc.) di sintomi, può essere eliminato:

infatti al denominatore c’è il prodotto

P(*Colpire*|*MalDiDenti*) P(*MalDiDenti*) che è uguale a

$\mathrm{P}(\mathit{MalDiDenti} \wedge \mathit{Colpire})$

e, come prima, può essere eliminato con la normalizzazione, dato che si può disporre di

$\mathrm{P}(\mathit{MalDiDenti} | \neg \mathit{Carie})$ e
$\mathrm{P}(\mathit{Colpire} | \neg \mathit{Carie})$.

Concludendo, occorre solo valutare la probabilità a priori per la causa e le probabilità condizionali per ognuno degli effetti.

Nel caso multivalore, per dire che X e Y sono indipendenti dato Z, si scrive:

$$\mathbf{P}(X|Y,Z) = \mathbf{P}(X|Z)$$

che rappresenta un insieme di enunciati condizionali indipendenti individuali.

La corrispondente semplificazione della regola di Bayes per le prove multiple è

$$\mathbf{P}(Z|X,Y) = \alpha\ \mathbf{P}(Z)\ \mathbf{P}(X|Z)\ \mathbf{P}(Y|Z)$$

dove α è una costante di normalizzazione tale che i valori in $\mathbf{P}(Z|X,Y)$ sommati diano 1.

In pratica, se $\boldsymbol{e}_n = e^1, e^2, \ldots, e^n$ è una sequenza di dati osservati in passato (contesto) e se e è una nuova informazione, il risultato precedente ci dice che:

$$P(H \mid e, e_n) \propto P(e \mid H, e_n) P(H \mid e_n)$$

cioè il nuovo belief è proporzionale ($\propto$) al prodotto del vecchio belief per la verosimiglianza che la nuova informazione possa materializzarsi data l'ipotesi e l'esperienza passata.

Se, come succede spesso, l'esperienza passata è irrilevante a determinare la verosimiglianza, ovverosia si verifica che $e \perp e_n \mid H$, allora la formula precedente si semplifica:

$$P(H \mid e, e_n) \propto P(e \mid H) P(H \mid e_n) \ .$$

Esempio:

- L'epidemia E può aver causato il morbillo M nel paziente;
- Il morbillo può aver causato macchie rosse R o febbre F.

Effetti di una stessa causa diventano indipendenti una volta nota la causa. Dunque $F \perp E \mid M$.

Inoltre se E causa M e M causa R, allora conoscere E diventa irrilevante per R una volta che M è noto. Dunque $E \perp R \mid M$ e $E \perp F \mid M$.

L'approccio naturale è modellare gli effetti *date* le cause, non viceversa: P(e|H) piuttosto che P(H|e).

Nel nostro esempio:

- $P(M|E)$
- $P(R|M)$
- $P(F|M)$
- $P(E)$ (a priori, dato che non ha cause)

che somigliano a regole del tipo:
if *M* **then** *R*, con una certa confidenza

(nota: il significato è "**if** *M* **and** *tutto il resto è irrilevante* **then** *R*, con una certa confidenza")

Ora:

- Il teorema di Bayes ci serve per invertire i legami condizionali nella direzione di interesse.
- L'indipendenza condizionale ci serve a semplificare eliminando l'informazione irrilevante:

$$P(M \mid R, F, E) \propto$$
$$P(R \mid M, F, E)P(M \mid F, E) =$$
$$P(R \mid M)P(M \mid F, E)$$

e, a sua volta,

$$P(M \mid F, E) \propto P(F \mid M, E)P(M \mid E) =$$
$$P(F \mid M)P(M \mid E)$$

quindi

$$P(M \mid R, F, E) \propto$$
$$P(R \mid M)P(F \mid M)P(M \mid E)$$

che usa le "regole" probabilistiche codificate nel modello.

Si osservi che il procedimento precedente poteva essere sostituito da un approccio di "bruta forza":

$$P(M \mid R, F, E) = \frac{P(M, R, F, E)}{P(R, F, E)} =$$

$$\frac{P(M, R, F, E)}{\sum_M P(M, R, F, E)}$$

L'utilizzo dell'indipendenza condi-zionale riduce la complessità dell'inferenza.

Rete di credenze o Rete bayesiana

Una rete di credenze è un grafo per cui valgono le seguenti proprietà:

1. Un insieme di variabili casuali costituiscono i nodi della rete.

2. Un insieme di archi con un verso connette le coppie di nodi. Il significato intuitivo di una freccia dal nodo *X* al nodo *Y* è che *X* ha *un'influenza diretta* su *Y*.

3. Ogni nodo ha una tabella di probabilità condizionate che quantifica gli effetti che i genitori hanno sul nodo. I genitori di un nodo sono tutti quei nodi che hanno frecce che puntano al nodo.

4. Il grafo non ha cicli diretti (dunque si tratta di un grafo diretto aciclico, DAG, Directed Acyclic Graph).

Esempio:

In una casa di Los Angeles è installato un antifurto abbastanza affidabile nello scoprire gli intrusi, ma che si attiva anche in caso di piccolo terremoti.

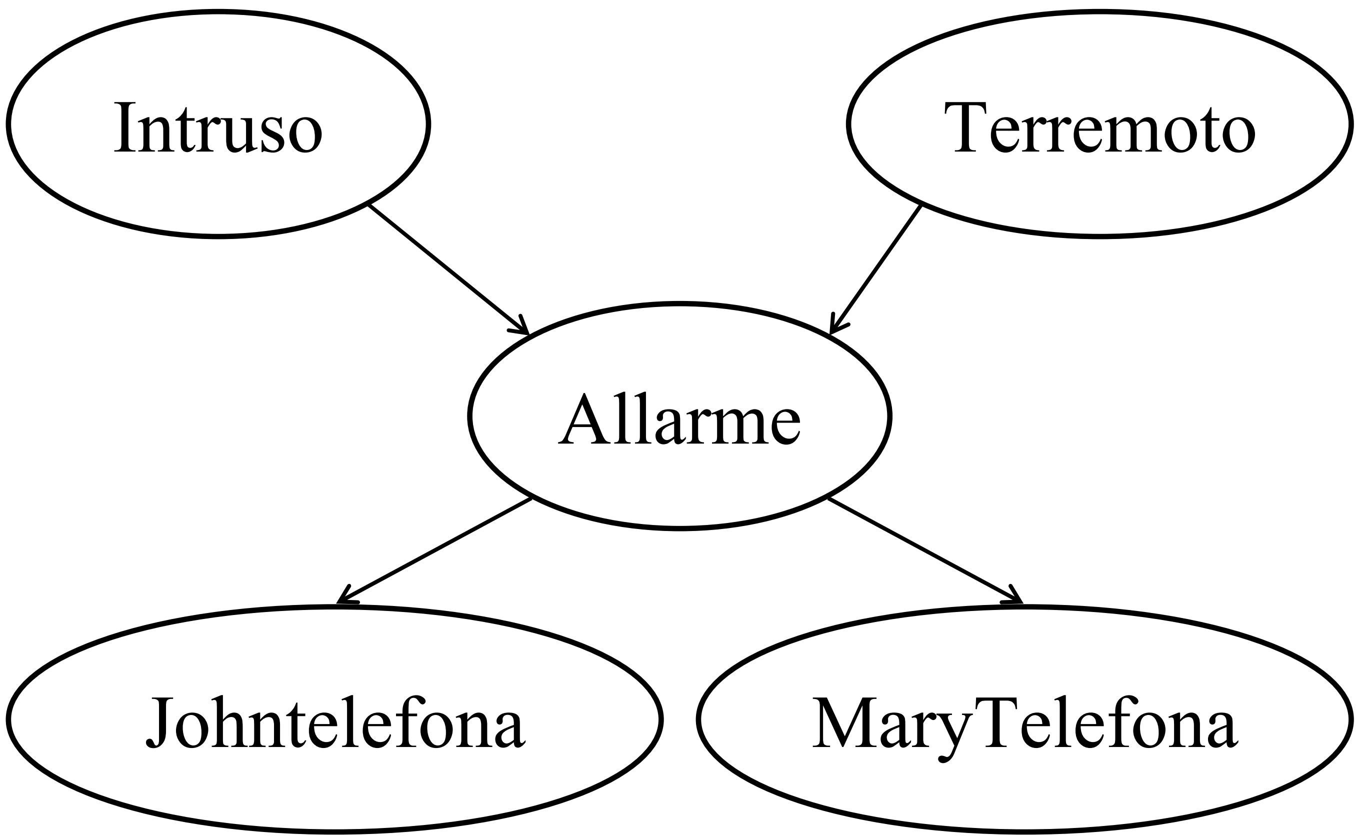
Due vicini, John e Mary, hanno promesso di avvertire in ufficio se sentono l’allarme.

John chiama tutte le volte che sente l’allarme, ma a volte confonde il suono del telefono con l’allarme e chiama lo stesso.

Mary invece ascolta spesso la musica a volume alto e a volte non sente l’allarme.

Data la prova di chi ha chiamato si vorrebbe sapere la probabilità di un intruso a casa.

Questo dominio può essere descritto dalla seguente rete di credenze:



Specificata la topologia, occorre specificare, per ogni nodo, la **tabella delle probabilità condizionate**. Ogni riga della tabella contiene la probabilità condizionata del valore di ogni nodo per un **caso condizionante**.

Un caso condizionante è una possibile combinazione dei valori per i nodi genitori (un evento atomico).

Ad esempio, per la variabile *Allarme* potrebbe essere:

Intruso	Terremoto	$\mathbf{P}(Allarme \mid Intruso, Terremoto)$	
		Vero	*Falso*
Vero	*Vero*	0.950	0.050
Vero	*Falso*	0.940	0.060
Falso	*Vero*	0.290	0.710
Falso	*Falso*	0.001	0.999

La somma di ogni riga dà 1, perché gli elementi rappresentano un insieme esaustivo di casi delle variabili (quindi è sufficiente rappresentare una sola colonna per P(*Allarme*)).

La rete completa per l'esempio precedente può essere il seguente:

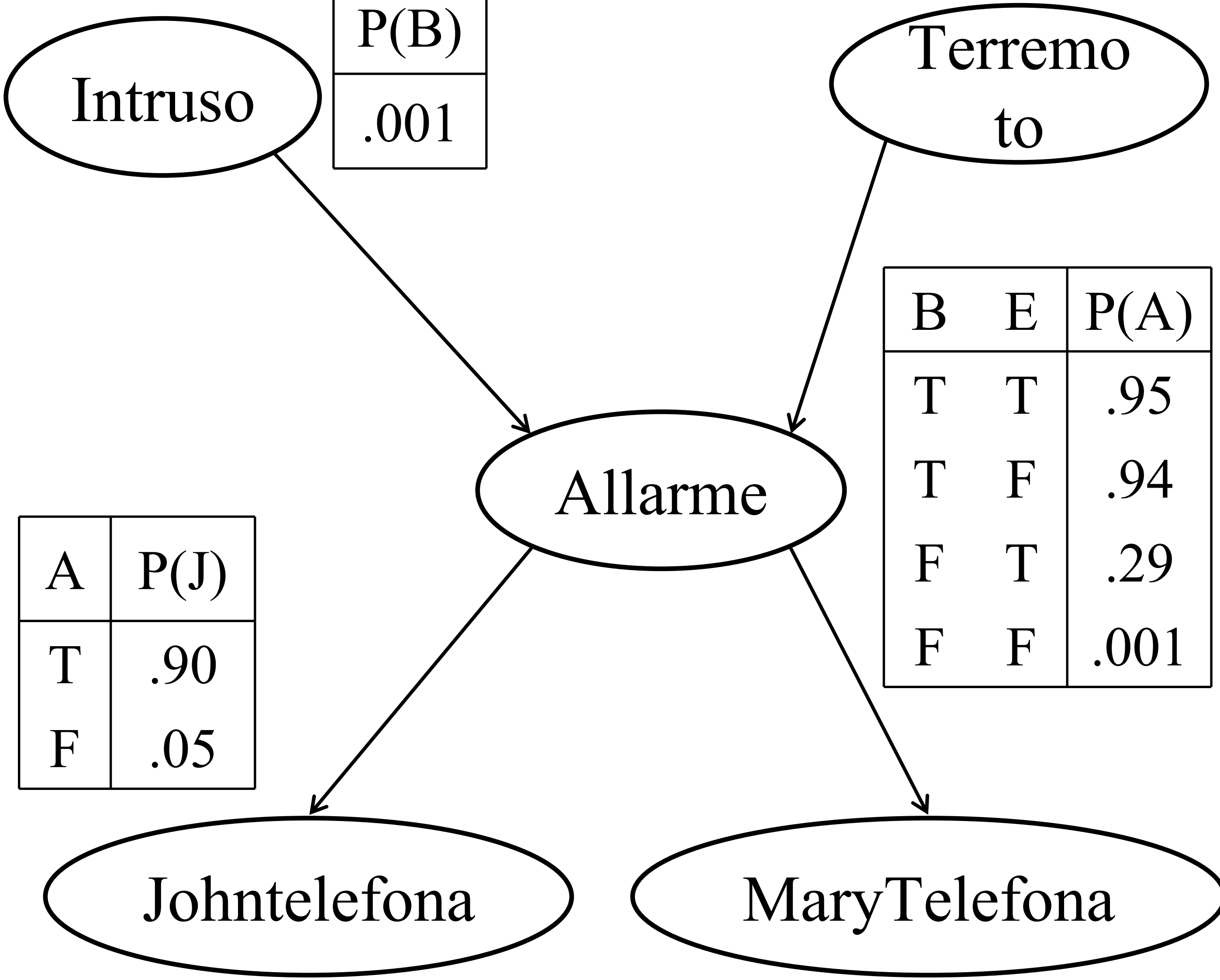
$B = Intruso$

$E = Terremoto$

$A = Allarme$

$J = JohnTelefona$

$M = MaryTelefona$



P(B)
.001

P(E)
.002

B	E	P(A)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(J)
T	.90
F	.05

A	P(M)
T	.70
F	.01

Tutte le variabili sono booleane: in qualsiasi riga vale $P(\neg X) = 1\text{-}P(X)$.

Vi sono due modi d'intendere la *semantica* di una rete di credenze:

- si può vedere la rete come una rappresentazione della distribuzione della probabilità congiunta (utile per comprendere come *costruire* la rete);
- si può vedere come una codifica di un insieme di enunciati di indipendenza condizionale (utile per il progetto delle procedure di inferenza).

Una rete di credenze fornisce una descrizione completa del dominio.

Ogni valore della tabella della probabilità congiunta può essere calcolata dall'informazione nella rete.

Un elemento generico nella tabella congiunta è la probabilità di una congiunzione di assegnazioni particolari ad ogni variabile:

$$P(x_1,...,x_n) = P(X_1 = x_1 \wedge ... \wedge X_n = x_n)$$

Può essere calcolata con la seguente formula:

$$P(x_1,\ldots,x_n) = \prod_{i=1}^{n} P(x_i \mid Genitori(X_i))$$

Dunque ogni elemento nella tabella congiunta è rappresentato dal prodotto degli elementi appropriati delle tabelle di probabilità condizionate nella rete di credenze.

Tali tabelle pertanto forniscono una rappresentazione decomposta della distribuzione congiunta.

Esempio di uso: calcolo della probabilità dell'evento che l'allarme sia partito ma non è stato né a causa di un intruso né di un terremoto e hanno telefonato sia Mary che John.

$$P(J \wedge M \wedge A \wedge \neg B \wedge \neg E) =$$
$$P(J \mid A)P(M \mid A)P(A \mid \neg B \wedge \neg E)P(\neg B)P(\neg E) =$$
$$0.90 \times 0.70 \times 0.001 \times 0.999 \times 0.998 = 0.00062$$

Metodo per la costruzione delle reti di credenze

Si riscriva la distribuzione congiunta in termini di una probabilità condizionale usando la definizione di probabilità condizionata:

$$P(x_1,\ldots,x_n) = P(x_n \mid x_{n-1},\ldots,x_1)\cdot P(x_{n-1},\ldots,x_1)$$

Si ripeta questo processo, riducendo ogni probabilità congiuntiva ad una probabilità condizionale ed una piccola congiunzione. Si finisce con un unico prodotto:

$$P(x_1,\ldots,x_n) =$$

$$P(x_n \mid x_{n-1},\ldots,x_1)\cdot P(x_{n-1} \mid x_{n-2},\ldots,x_1)\cdot\ldots\cdot P(x_2 \mid x_1)\cdot P(x_1) =$$

$$\prod_{i=1}^{n} P(x_i \mid x_{i-1},\ldots,x_1)$$

Se si riprende l'equazione usata per la rete di credenze, che era

$$P(x_1,\ldots,x_n) = \prod_{i=1}^{n} P(x_i \mid Genitori(X_i))$$

e si pone

$$Genitori(X_i) \subseteq \{x_{i-1},\ldots,x_1\}$$

si nota che la specifica della congiunzione è equivalente.

Si può quindi affermare che la rete di credenze è una rappresentazione corretta del dominio solo se ogni nodo è condizionalmente indipendente dai suoi predecessori nell'ordinamento dei nodi, dati i suoi genitori.

Pertanto per poter costruire una rete di credenze con la corretta struttura per il dominio, occorre scegliere i genitori di ogni nodo in maniera che questa proprietà valga.

Intuitivamente i genitori di un nodo X_i dovrebbe contenere quei nodi in $X_1,\ldots, X_{i-1}$ che influenzano *direttamente* X_i.

Nell'esempio precedente, si supponga di aver sistemato tutti gli altri nodi e dover scegliere i genitori di *MeryTelefona*.

Il comportamento di Mary è influenzato dalla presenza dell'intruso e dal terremoto, ma *non direttamente*.

È solo l'allarme ad influenzare *direttamente* il comportamento di Mary.

In modo formale, si può dire che vale il seguente enunciato di indipendenza condizionale:

$$\mathbf{P}(MaryTelefona|JohnTelefona, Allarme, Terremoto, Intruso) = \mathbf{P}(MaryTelefona|Allarme)$$

La procedura generale per la costruzione incrementale della rete è la seguente:

1. Si scelga un insieme di variabili rilevanti X_i che descrivono il dominio.

2. Si scelga un ordinamento per le variabili.

3. Finché sono rimaste variabili:

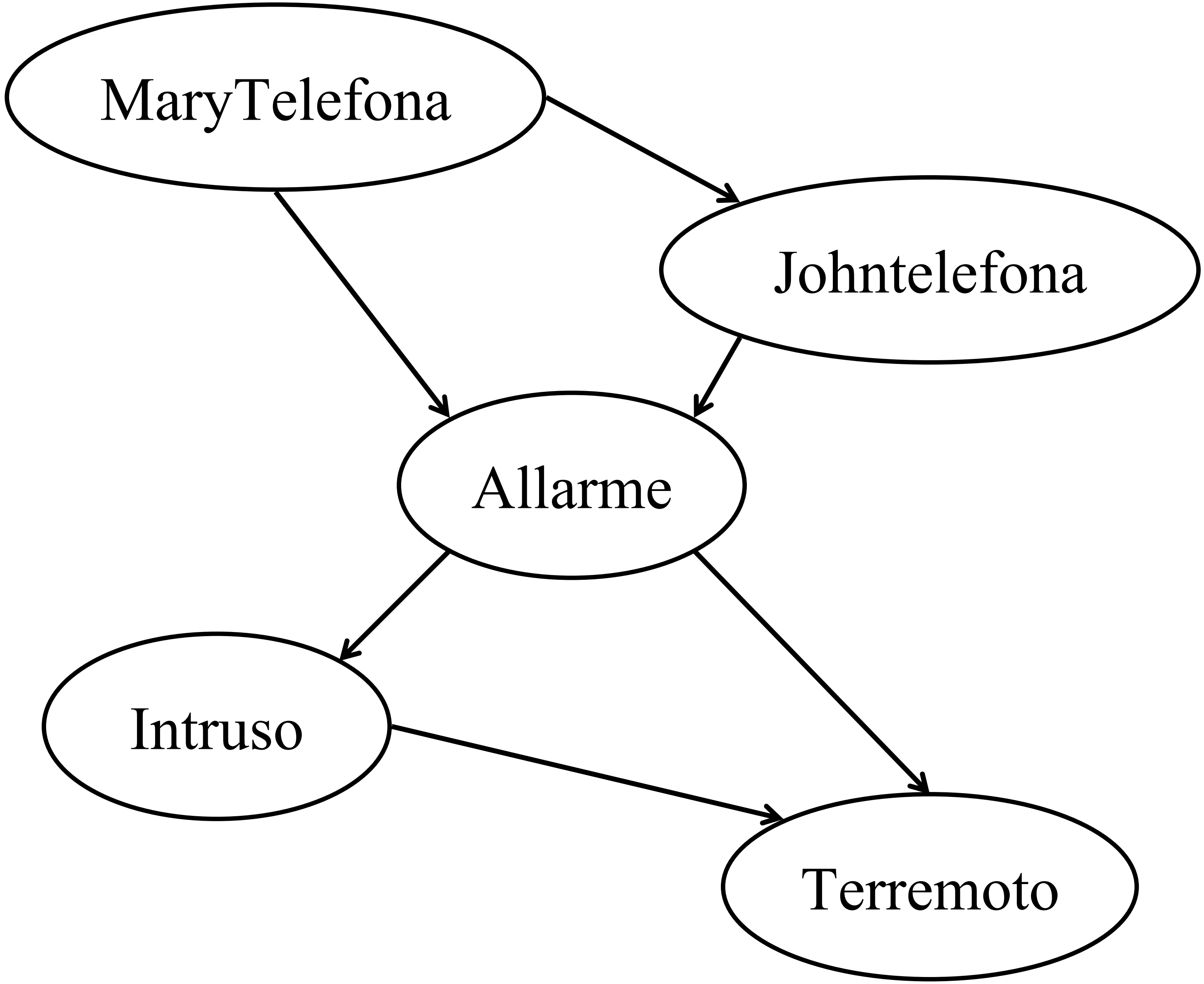
a) Si prenda una variabile X_i e si aggiunga un nodo alla rete.

b) Si ponga *Genitori*(X_i) come un qualche insieme minimo di nodi già presenti nella rete, tale che la proprietà di indipendenza condizionale sia soddisfatta.

c) Si definisca la tabella delle probabilità condizionate per X_i.

Siccome ogni nodo è connesso solo ai nodi precedenti, la rete è aciclica.

In base all'ordinamento dei nodi che è stato scelto, ci possono essere più rappresentazioni dello stesso dominio.


MaryTelefona
Johntelefona
Allarme
Intruso
Terremoto

Processo di creazione della rete:

- Aggiunta di *MaryTelefona*: nessun genitore.

- Aggiunta di *JohnTelefona*: se Mary telefona, probabilmente si è attivato l'allarme, quindi c'è dipendenza:

  $P(JohnTelefona|MaryTelefona)$
  $\neq P(JohnTelefona)$.

- Aggiunta di *Allarme*: sicuramente è un nodo correlato ai due precedenti.

- Aggiunta di *Intruso*: il nodo dipende direttamente solo dallo stato dell'allarme; la telefonata di John o di Mary non aggiunge informazione sull'intruso: $P(Intruso|Allarme, JohnTelefona, MaryTelefona) =$ $P(Intruso|Allarme)$.

- Aggiunta di *Terremoto*: sicuramente è in dipendenza diretta con *Allarme*, ma è anche influenzato dal nodo *Intruso*. Infatti se sappiamo che vi è stata un'intrusione, questo giustifica l'allarme e rende più bassa la probabilità di terremoto: $P(Terremoto|Intruso, Allarme, JohnTelefona, MeryTelefona) =$ $P(Terremoto|Intruso, Allarme)$

Questa rete ha due archi in più della precedente e richiede di specificare tre probabilità in più.

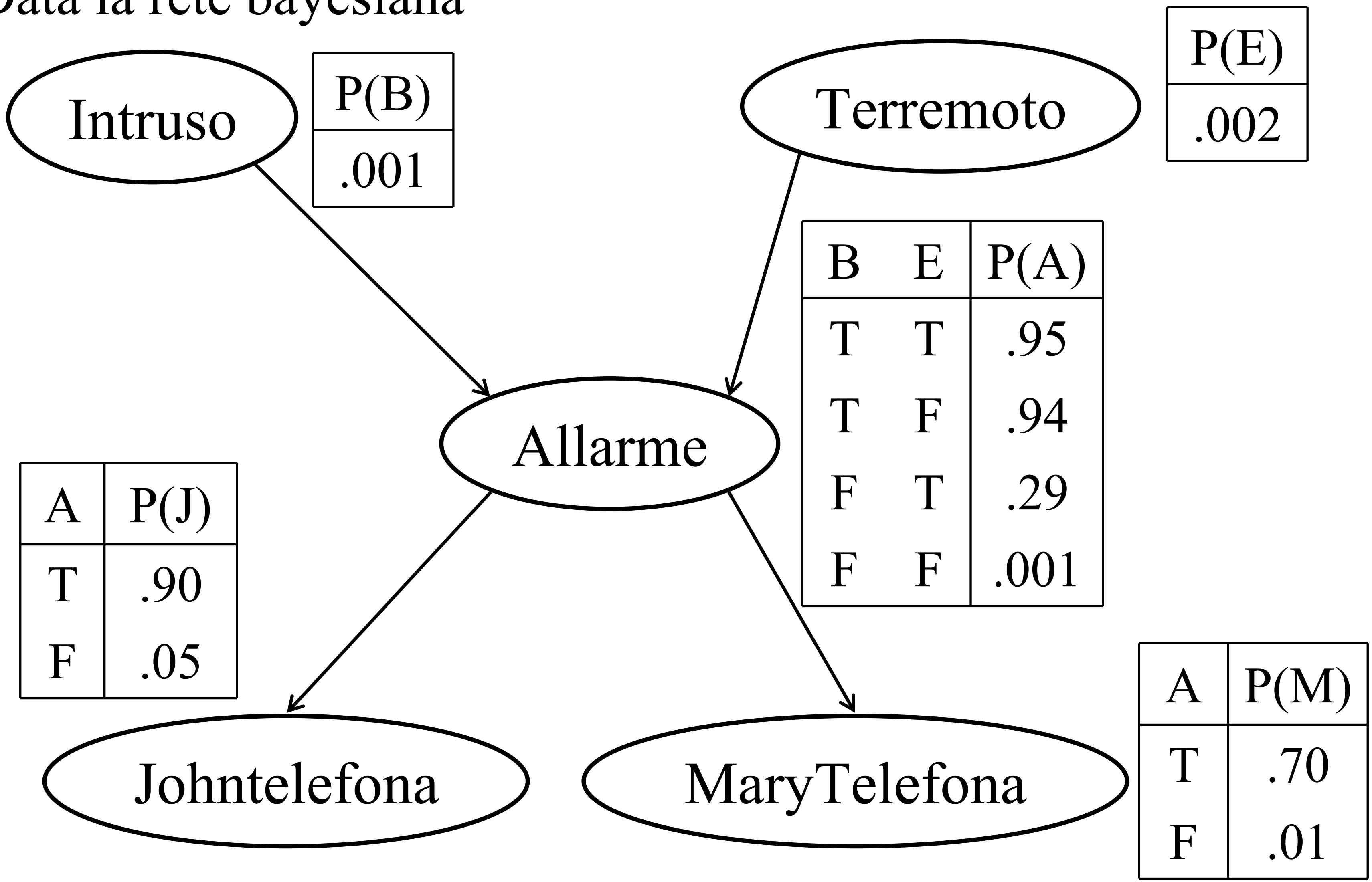
Si dimostra che conviene attenersi al *modello causale* (archi dalle cause ai sintomi, e non viceversa), poiché così occorre specificare meno numeri e i numeri stessi sono più significativi.

Si pensi, ad esempio, alla difficoltà di valutare statisticamente

$P(\mathit{Terremoto}|\mathit{Intruso}, \mathit{Allarme})$.

Esercizi

Data la rete bayesiana



P(B)
.001

P(E)
.002

B	E	P(A)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(J)
T	.90
F	.05

A	P(M)
T	.70
F	.01

$B = Intruso$
$E = Terremoto$
$A = Allarme$
$J = JohnTelefona$
$M = MaryTelefona$

calcolare:

- dato Intruso,
  P(JohnTelefona|Intruso)

  B = Intruso
  J = JohnTelefona

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

$P(J, B) =$

$P(J, B, \neg M, \neg A, \neg E) +$

$P(J, B, \neg M, \neg A, E) +$

$P(J, B, \neg M, A, \neg E) +$

$P(J, B, \neg M, A, E) +$

$P(J, B, M, \neg A, \neg E) +$

$P(J, B, M, \neg A, E) +$

$P(J, B, M, A, \neg E) +$

$P(J, B, M, A, E) =$

$$P(J|\neg A) \cdot P(B) \cdot P(\neg M|\neg A) \cdot P(\neg A|B \wedge \neg E) \cdot P(\neg E) +$$
$$P(J|\neg A) \cdot P(B) \cdot P(\neg M|\neg A) \cdot P(\neg A|B \wedge E) \cdot P(E) +$$
$$P(J|A) \cdot P(B) \cdot P(\neg M|A) \cdot P(A|B \wedge \neg E) \cdot P(\neg E) +$$
$$P(J|A) \cdot P(B) \cdot P(\neg M|A) \cdot P(A|B \wedge E) \cdot P(E) +$$
$$P(J|\neg A) \cdot P(B) \cdot P(M|\neg A) \cdot P(\neg A|B \wedge \neg E) \cdot P(\neg E) +$$
$$P(J|\neg A) \cdot P(B) \cdot P(M|\neg A) \cdot P(\neg A|B \wedge E) \cdot P(E) +$$
$$P(J|A) \cdot P(B) \cdot P(M|A) \cdot P(A|B \wedge \neg E) \cdot P(\neg E) +$$
$$P(J|A) \cdot P(B) \cdot P(M|A) \cdot P(A|B \wedge E) \cdot P(E)$$

Dividendo per P(B), si ottiene:

$P(J|B) =$

$0.05 \cdot 0.99 \cdot 0.06 \cdot 0.998 +$

$0.05 \cdot 0.99 \cdot 0.05 \cdot 0.002 +$

$0.9 \cdot 0.3 \cdot 0.94 \cdot 0.998 +$

$0.9 \cdot 0.3 \cdot 0.95 \cdot 0.002 +$

$0.05 \cdot 0.01 \cdot 0.06 \cdot 0.998 +$

$0.05 \cdot 0.01 \cdot 0.05 \cdot 0.002 +$

$0.9 \cdot 0.7 \cdot 0.94 \cdot 0.998 +$

$0.9 \cdot 0.7 \cdot 0.95 \cdot 0.002 =$

0.85

NB: procedendo in modo simbolico:

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

$= P(J|\neg A) \cdot P(\neg M|\neg A) \cdot [P(\neg A|B \wedge \neg E) \cdot P(\neg E) + P(\neg A|B \wedge E) \cdot P(E)] +$
$P(J|A) \cdot P(\neg M|A) \cdot [P(A|B \wedge \neg E) \cdot P(\neg E) + P(A|B \wedge E) \cdot P(E)] +$
$P(J|\neg A) \cdot P(M|\neg A) \cdot [P(\neg A|B \wedge \neg E) \cdot P(\neg E) + P(\neg A|B \wedge E) \cdot P(E)] +$
$P(J|A) \cdot P(M|A) \cdot [P(A|B \wedge \neg E) \cdot P(\neg E) + P(A|B \wedge E) \cdot P(E)] =$

$= P(J|\neg A) \cdot [P(\neg A|B \wedge \neg E) \cdot P(\neg E) +$
$P(\neg A|B \wedge E) \cdot P(E)] \cdot [P(\neg M|\neg A) + P(M|\neg A)] + P(J|A) \cdot [P(A|B \wedge \neg E)$
$\cdot P(\neg E) + P(A|B \wedge E) \cdot P(E)] \cdot [P(\neg M|A) + P(M|A)]$

e poiché

$$[P(\neg M|\neg A) + P(M|\neg A)] = 1$$
$$[P(\neg M|A) + P(M|A)] = 1$$

$P(J|B) =$

$$= P(J|\neg A) \cdot [P(\neg A|B \wedge \neg E) \cdot P(\neg E) + P(\neg A|B \wedge E) \cdot P(E)] +$$
$$P(J|A) \cdot [P(A|B \wedge \neg E) \cdot P(\neg E) + P(A|B \wedge E) \cdot P(E)]$$

da cui si vede che per il calcolo c'entra tutto quello che "sta sopra" JohnTelefona (linea causale), e non quello che sta "a fianco" (MaryTelefona).

- Dato JohnTelefona, $P(Intruso|JohnTelefona)$

$P(B|J) = P(J|B) \cdot P(B) / P(J)$

dove $P(J|B)$ è stato calcolato sopra, $P(B)$ è dato e $P(J)$ si calcola con modalità similari a quelli appena visti ($P(B|J) = 0.016$).

- Dati JohnTelefona e MaryTelefona, P(Allarme|JohnTelefona^MaryTelefona)

$P(A|J \wedge M) = P(A \wedge J \wedge M) / P(J \wedge M)$

$P(A \wedge J \wedge M) =$

$P(A \wedge J \wedge M \wedge \neg B \wedge \neg E) +$

$P(A \wedge J \wedge M \wedge \neg B \wedge E) +$

$P(A \wedge J \wedge M \wedge B \wedge \neg E) +$

$P(A \wedge J \wedge M \wedge B \wedge E) =$

$$P(J|A) \cdot P(M|A) \cdot P(A|\neg B \wedge \neg E) \cdot P(\neg B) \cdot P(\neg E) + P(J|A) \cdot P(M|A) \cdot P(A|\neg B \wedge E) \cdot P(\neg B) \cdot P(E) + P(J|A) \cdot P(M|A) \cdot P(A|B \wedge \neg E) \cdot P(B) \cdot P(\neg E) + P(J|A) \cdot P(M|A) \cdot P(A|B \wedge E) \cdot P(B) \cdot P(E) =$$

$$P(J|A) \cdot P(M|A) \cdot \{ P(\neg B) \cdot [P(A|\neg B \wedge \neg E) \cdot P(\neg E) + P(A|\neg B \wedge E) \cdot P(E)] + P(B) \cdot [P(A|B \wedge \neg E) \cdot P(\neg E) + P(A|B \wedge E) \cdot P(E)] \} = 0.0015852$$

$P(J \wedge M) = P(J \wedge M \wedge \neg A) + P(J \wedge M \wedge A) =$

$P(J \wedge M \wedge \neg A \wedge \neg B \wedge \neg E) +$
$P(J \wedge M \wedge \neg A \wedge \neg B \wedge E) +$
$P(J \wedge M \wedge \neg A \wedge B \wedge \neg E) +$
$P(J \wedge M \wedge \neg A \wedge B \wedge E) + P(J \wedge M \wedge A) =$

$P(J|\neg A) \cdot P(M|\neg A) \cdot \{ P(\neg B) \cdot [P(\neg A|\neg B \wedge \neg E) \cdot P(\neg E) + P(\neg A|\neg B \wedge E) \cdot P(E)] + P(B) \cdot [P(\neg A|B \wedge \neg E) \cdot P(\neg E) + P(\neg A|B \wedge E) \cdot P(E)] \} + P(J \wedge M \wedge A) =$
$= 0.0004987 + 0.0015852 = 0.0020839$

$P(A|J \wedge M) = P(A \wedge J \wedge M) / P(J \wedge M) =$
$= 0.0015852 / 0.0020839 = 0.76$