Undirected Graphical Models

Representation

Professor Ajoodha

Lecture 4

School of Computer Science and Applied Mathematics
The University of the Witwatersrand, Johannesburg

Problem Statement

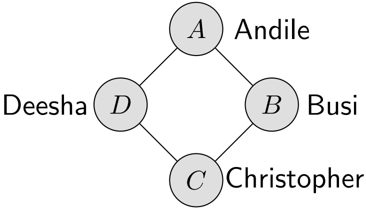
Problem

What do we do when we cannot naturally ascribe a directionality to the interaction between variables?

1. We use undirected graphical models (Markov networks)
2. Markov networks offer a simpler perspective on directed models in terms of:
   - Independence structure
   - Inference tasks
3. We can also use a combined framework (directed and undirected)
4. In Markov networks, sometimes we can only use discrete spaces

Problem Statement: Misconception Example

Suppose we have 4 students who are working in pairs to solve a class exercise

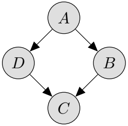


In this problem we would like to model $(A \perp C \mid D, B)$ and $(B \perp D \mid A, C)$.

Can these independence assertions be modelled in a Bayesian network?

Problem Statement: Misconception Example

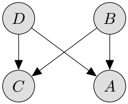
We need: $(A \perp C \mid D, B)$ and $(B \perp D \mid A, C)$.



1. We could do it this way:

   But this also implies:

   1. $(D \perp B \mid A)$
   2. $(D \not\perp B \mid A, C)$ (v-structure activated)

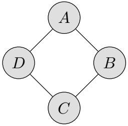


2. We could also try it a second way:
   1. Although we get $(A \perp C \mid B, D)$
   2. but then B and D are **marginally independent** (v-structure only activated if C or A is given).

This cannot be done in a Bayesian network.

Problem Statement: Summary

- In summary, it is not possible to capture **both of the below** two assumptions in a Bayesian Network.
  1. $(A \perp C \mid D, B)$
  2. $(B \perp D \mid A, C)$
- Any Bayesian network I-map would have extraneous edges and will not capture at least one of the above assumptions.
- We would like a model which can represent correlations without forcing a direction of influence.
- We can however use, a Markov network (Undirected graphical Model) to capture the above two assumptions.
- We build the Markov network naturally as follows:



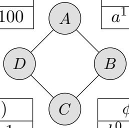
Markov Networks

The components of a Markov network correspond to:

- Nodes: Variables;
- Edges: direct probabilistic interaction between neighbouring variables - an interaction that is not mediated by any other variables in the network.
- We do not need CPDs, so we use **general purpose functions** (also called factors). E.g. $\phi_1(A,B) \mapsto \mathbb{R}^+$

$\phi_4(D,A)$		
d^0	a^0	100
d^0	a^1	1
d^1	a^0	1
d^1	a^1	100

$\phi_1(A,B)$		
a^0	b^0	30
a^0	b^1	5
a^1	b^0	1
a^1	b^1	10



$\phi_3(C,D)$		
c^0	d^0	1
c^0	d^1	100
c^1	d^0	100
c^1	d^1	1

More likely to disagree

$\phi_2(B,C)$		
b^0	c^0	100
b^0	c^1	1
b^1	c^0	1
b^1	c^1	100

More likely to agree

Joint distribution and partition function

Just as in Bayesian networks the joint distribution can be expressed as:

$$P(A,B,C,D) = \frac{1}{Z}\phi_1(A,B)\cdot\phi_2(B,C)\cdot\phi_3(C,D)\cdot\phi_4(D,A)$$

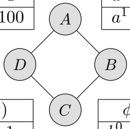
where Z is an unnormalised constant called the **partition function**:

$$Z = \sum_{a,b,c,d} \phi_1(A,B)\cdot\phi_2(B,C)\cdot\phi_3(C,D)\cdot\phi_4(D,A)$$

Example

$\phi_4(D, A)$		
d^0	a^0	100
d^0	a^1	1
d^1	a^0	1
d^1	a^1	100

$\phi_1(A, B)$		
a^0	b^0	30
a^0	b^1	5
a^1	b^0	1
a^1	b^1	10



$\phi_3(C, D)$		
c^0	d^0	1
c^0	d^1	100
c^1	d^0	100
c^1	d^1	1

More likely to disagree

$\phi_2(B, C)$		
b^0	c^0	100
b^0	c^1	1
b^1	c^0	1
b^1	c^1	100

More likely to agree

$$P(a^0, b^0, c^1, d^0) = \frac{1}{Z}\phi_1(a^0, b^0) \cdot$$
$$\phi_2(b^0, c^1) \cdot \phi_3(c^1, d^0) \cdot \phi_4(d^0, a^0)$$
$$= \frac{1}{Z} 30 \cdot 1 \cdot 100 \cdot 100 = 300000$$
$$= \frac{1}{Z} 300000 = \frac{1}{7201840} 300000 = 0.04165$$

Assignment				Unr	Nor
a^0	b^0	c^0	d^0	3E5	0.04
a^0	b^0	c^0	d^1	3E5	0.04
a^0	b^0	c^1	d^0	3E5	0.04
a^0	b^0	c^1	d^1	30	4.1E-6
a^0	b^1	c^0	d^0	500	6.9E-5
a^0	b^1	c^0	d^1	500	6.9E-5
a^0	b^1	c^1	d^0	5E6	0.69
a^0	b^1	c^1	d^1	500	0.69E-5
a^1	b^0	c^0	d^0	100	1.4E-5
a^1	b^0	c^0	d^1	1E6	0.14
a^1	b^0	c^1	d^0	100	1.4E-5
a^1	b^0	c^1	d^1	100	1.4E-5
a^1	b^1	c^0	d^0	10	1.4E-6
a^1	b^1	c^0	d^1	1E5	0.014
a^1	b^1	c^1	d^0	1E5	0.014
a^1	b^1	c^1	d^1	1E5	0.014
				7201840	

Factors

Since the edges are not directed, then we can not have "directed" parameterization.

Factors subsumes both the joint distribution and CPD:

- The joint distribution is a factor over variable set **D**.
- A CPD $P(X|\,\mathbf{U})$ is a factor over variable set $\{X\} \cup \mathbf{U}$.

However, factors do not have to be normalised.

In Markov networks we use a **general representation of factors** over multiple subsets of variables.

E.g. $\phi_1(\mathbf{X}, \mathbf{Y})$ or $\phi_2(\mathbf{Y}, \mathbf{Z})$

There are several operations that we can perform on factors.

1. Factor Product
2. and Factor Reduction

Factor Product with Example

A factor product $(\phi_1 \times \phi_2)$ is defined as a factor
$\psi : Val(\mathbf{X,Y,Z}) \mapsto \mathbb{R}$ as $\psi(\mathbf{X,Y,Z}) = \phi_1(\mathbf{X},\mathbf{Y}) \cdot \phi_2(\mathbf{Y},\mathbf{Z})$

a^1	b^1	0.5
a^1	b^2	0.8
a^2	b^1	0.1
a^2	b^2	0
a^3	b^1	0.3
a^3	b^2	0.9

$\phi(A,B)$

$\times$

b^1	c^1	0.5
b^1	c^2	0.7
b^2	c^1	0.1
b^2	c^2	0.2

$\phi(B,C)$

$=$

a^1	b^1	c^1	$0.5 \cdot 0.5$
a^1	b^1	c^2	$0.5 \cdot 0.7$
a^1	b^2	c^1	$0.8 \cdot 0.1$
a^1	b^2	c^2	$0.8 \cdot 0.2$
a^2	b^1	c^1	$0.1 \cdot 0.5$
a^2	b^1	c^2	$0.1 \cdot 0.7$
a^2	b^2	c^1	$0 \cdot 0.1$
a^2	b^2	c^2	$0 \cdot 0.2$
a^3	b^1	c^1	$0.3 \cdot 0.5$
a^3	b^1	c^2	$0.3 \cdot 0.7$
a^3	b^2	c^1	$0.9 \cdot 0.1$
a^3	b^2	c^2	$0.9 \cdot 0.2$

$\psi(A,B,C)$

Factor Reduction with Example

Let $\phi(\mathbf{Y})$ be a factor, and $\mathbf{U} = \mathbf{u}$ an assignment for $\mathbf{U} \subseteq \mathbf{Y}$. The reduction of a factor ϕ to the context $\mathbf{U} = \mathbf{u}$, denoted $\phi[\mathbf{U} = \mathbf{u}]$ is a factor over scope $\mathbf{Y'} = \mathbf{Y} - \mathbf{U}$, such that $\phi[\mathbf{u}](\mathbf{y}') = \phi(\mathbf{y}', \mathbf{u})$.

a^1	b^1	c^1	0.25
a^1	b^1	c^2	0.35
a^1	b^2	c^1	0.08
a^1	b^2	c^2	0.16
a^2	b^1	c^1	0.05
a^2	b^1	c^2	0.07
a^2	b^2	c^1	0
a^2	b^2	c^2	0
a^3	b^1	c^1	0.15
a^3	b^1	c^2	0.21
a^3	b^2	c^1	0.09
a^3	b^2	c^2	0.18

$\phi(A, B, C)$

$\rightarrow$

a^1	b^1	c^1	0.25
a^1	b^2	c^1	0.08
a^2	b^1	c^1	0.05
a^2	b^2	c^1	0
a^3	b^1	c^1	0.15
a^3	b^2	c^1	0.09

$\phi[C = c^1](A, B) = \phi(A, B, c^1)$

Gibbs Distributions

We can describe the joint distribution as a Gibbs distribution.

A distribution P_Φ is a Gibbs distribution parameterised by a set of factors $\Phi = \{\phi_1(\mathbf{D_1}), \ldots, \phi_K(\mathbf{D_k})\}$ if it is defined as:

$$P_\Phi(X_1, \ldots, X_n) = \frac{1}{Z}\tilde{P}_\Phi(X_1, \ldots, X_n),$$

where

$$\tilde{P}_\Phi(X_1, \ldots, X_n) = \phi_1(\mathbf{D_1}) \times \ldots \times \phi_K(\mathbf{D_k})$$

is an **unnormalised measure** and

$$Z = \sum_{X_1, \ldots, X_n} \tilde{P}_\Phi(X_1, \ldots, X_n)$$

Factorization

A factor in a Markov network is only one contribution to the overall distribution.

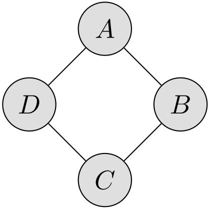
The full joint distribution comes from all of the contributing factors.

How does the Gibbs distribution relate to the graph structure?

If a single factor (clique potential) contains the variables X and Y, then there is a direct interaction between them.

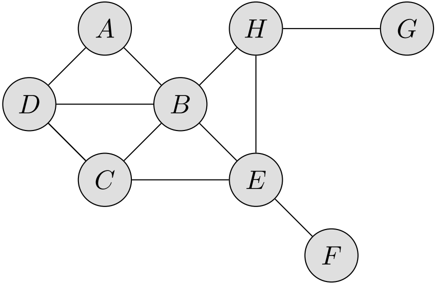
This connection must appear in the graph structure.

Maximal Clique Potentials: Example 1



$$\Phi = \{\phi_1(A,B),$$
$$\phi_2(B,C),$$
$$\phi_3(C,D),$$
$$\phi_4(D,A)\}$$

Maximal Clique Potentials: Example 2



$$\Phi = \{\phi_1(A, D, B),$$
$$\phi_2(D, B, C),$$
$$\phi_3(B, C, E),$$
$$\phi_4(B, H, E),$$
$$\phi_5(H, G),$$
$$\phi_6(E, F)\}$$

Basic Independencies

Influence in Markov networks freely flows via edges between variables

We call a path where influence can flow an "active path".

However, conditioning on a variables blocks an active path

We say that two sets of nodes, $\mathbf{X}$ and $\mathbf{Y}$, are seperated in $\mathcal{H}$ by $\mathbf{Z}$, denoted $\texttt{sep}_{\mathcal{H}}(\mathbf{X};\mathbf{Y} \mid \mathbf{Z})$, if there is no active trail between $\mathbf{X}$ and $\mathbf{Y}$ given $\mathbf{Z}$.

Therefore, the global independences in a Markov network is simply:

$$\mathcal{I}(\mathcal{H}) = \{(\mathbf{X};\mathbf{Y} \mid \mathbf{Z}) : \texttt{sep}_{\mathcal{H}}(\mathbf{X};\mathbf{Y} \mid \mathbf{Z})\}$$

Theorem: A (Gibbs) distribution P factorises over a Markov network $\mathcal{H}$ if and only if $\mathcal{H}$ is an I-map of P. *(proof not examinable)*

Local and Global Independencies

Recall in Bayesian networks we had the following local and global independence assumptions respectively:

Local: each node is independent of its non-descendants given its parents.
Global: induced by d-separation.

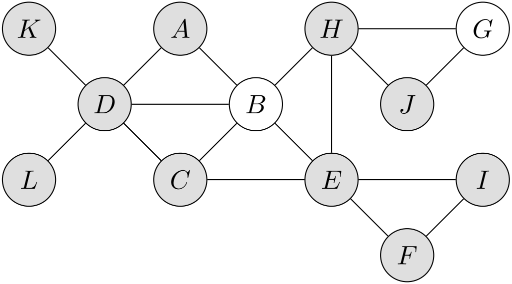
For Markov networks we can have 2 local assumptions and 1 global one:

Local 1: Two nodes are independent given all other nodes in the graph
Local 2: We can block all influence on a node by conditioning on its immediate neighbours.
Global (discussed previously):
$\mathcal{I}_{\texttt{global}}(\mathcal{H}) = \{(\mathbf{X};\mathbf{Y} \mid \mathbf{Z}) : \texttt{sep}_{\mathcal{H}}(\mathbf{X};\mathbf{Y} \mid \mathbf{Z})\}$

Markov Network: Local 1

Let $\mathcal{H}$ be a Markov network. We define the pairwise independencies associated with $\mathcal{H}$ to be:

$$\mathcal{I}_{\texttt{pairwise}}(\mathcal{H}) = \{(X \perp Y \mid \mathcal{X} - \{X, Y\}) : X - Y \notin \mathcal{H}\}$$

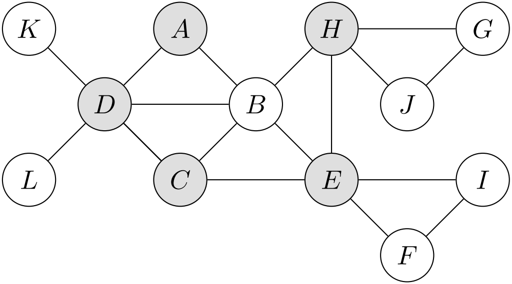


Two nodes are independent given all other nodes in the graph.

Markov Network: Local 2

For a given graph $\mathcal{H}$, we deine a Markov blanket of X in $\mathcal{H}$, denoted $\texttt{MB}_{\mathcal{H}}(X)$ to be the neighbours of X in $\mathcal{H}$. We define the local independencies of associated with $\mathcal{H}$ to be:

$$\mathcal{I}_{\ell\texttt{ocal}}(\mathcal{H}) = \{(X \perp \mathcal{X} - \{X\} - \texttt{MB}_{\mathcal{H}}(X) \mid \texttt{MB}_{\mathcal{H}}(X)) : X \in \mathcal{X}\}$$



We can block all influence on a node by conditioning on its immediate neighbours.

Corollary on Markov Model Independence

We can also make a claim about general positive distribution in terms of these assumptions.

The following three statements are equivalent for a **positive distribution** P.

1. $P \models \mathcal{I}_{\texttt{local}}(\mathcal{H})$
2. $P \models \mathcal{I}_{\texttt{pairwise}}(\mathcal{H})$
3. $P \models \mathcal{I}_{\texttt{global}}(\mathcal{H})$

Remember that a positive distribution is one where all events have a positively assigned probability assignment. i.e. $P(\alpha) > 0$

So how can we construct $\mathcal{H}$ from P?

We can use the same strategies we used in Bayesian networks using local independence assumptions to build minimal I-maps.

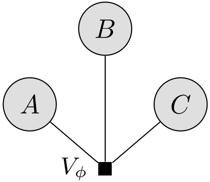
Factor Graphs

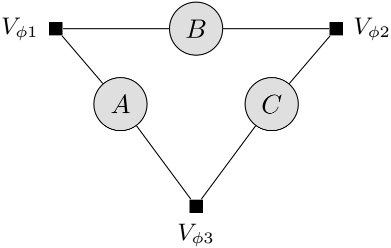
Recall the Markov graph $\mathcal{H}$ does not precisely give us the Gibbs parameterization.

That is, two Gibbs parameterization, Φ_A and Φ_B, can yield the same $\mathcal{H}$.

A factor graph makes the mapping between Φ and $\mathcal{H}$ explicit.

1. Circle nodes are variables
2. Square nodes are factors connected to their scope via edges





Problem Statement

Markov Networks

Factors

Gibbs Distributions

Independence

Factor Graphs

Knowledge Engineering

Applications

Picking Parameters

Knowledge Engineering

Application

Professor Ajoodha

School of Computer Science and Applied Mathematics
The University of the Witwatersrand, Johannesburg

Your new Machine Learning tookbox

So far we have looked at many ways to represent graphical models:

1. Bayesian networks
2. Markov networks

Ways of representing local probability models

1. Tabular CPDs
2. Deterministic CPDs
3. Context-Specific CPDs

The template settings of Graphical Models:

1. Dynamic Bayesian Networks
   - Factorial HMMs
   - Coupled HMMs
2. State-Observation Models
   - Hidden Markov Models
   - Linear Dynamical Systems

We learned about plate models:

1. Standard Plates
2. Overlapping Plates

What problem are you trying to solve?

- **Directed**

What problem are you trying to solve?

- ⋆ **Template**

- ⋆ **Directed**

What problem are you trying to solve?

- ⋆ **Template**
- ⋆ **Generative**
- ⋆ **Directed**

What problem are you trying to solve?

- ⋆ **Template**
- ⋆ **Generative**
- ⋆ **Directed**
- ⋆ **Undirected**

What problem are you trying to solve?

- ⋆ **Template**
- ⋆ **Generative**
- ⋆ **Directed**
- ⋆ **Undirected**
- ⋆ **Discriminative**

Knowledge Engineering
Professor Ajoodha
Problem Statement
Markov Networks
Factors
Gibbs Distributions
Independence
Factor Graphs
Knowledge Engineering
Applications
Picking Parameters

What problem are you trying to solve?

- ⋆ **Template**
- ⋆ **Generative**
- ⋆ **Directed**
- ⋆ **Undirected**
- ⋆ **Specific**
- ⋆ **Discriminative**

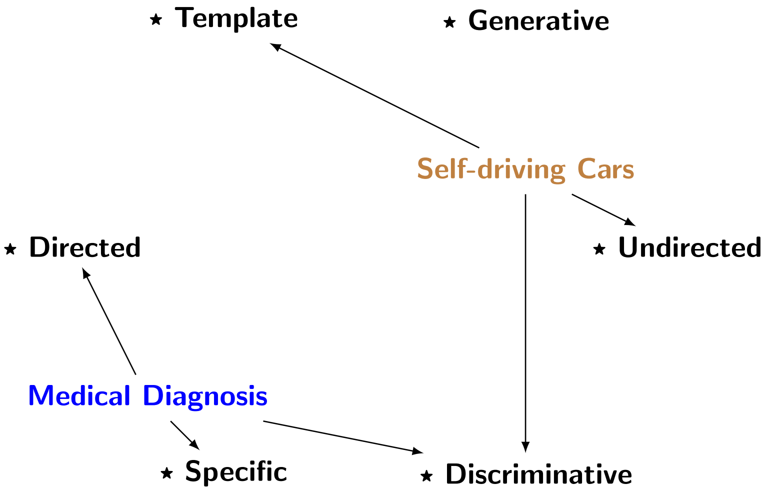
What problem are you trying to solve?

- ⋆ **Template**
- ⋆ **Generative**
- ⋆ **Directed**
- ⋆ **Undirected**

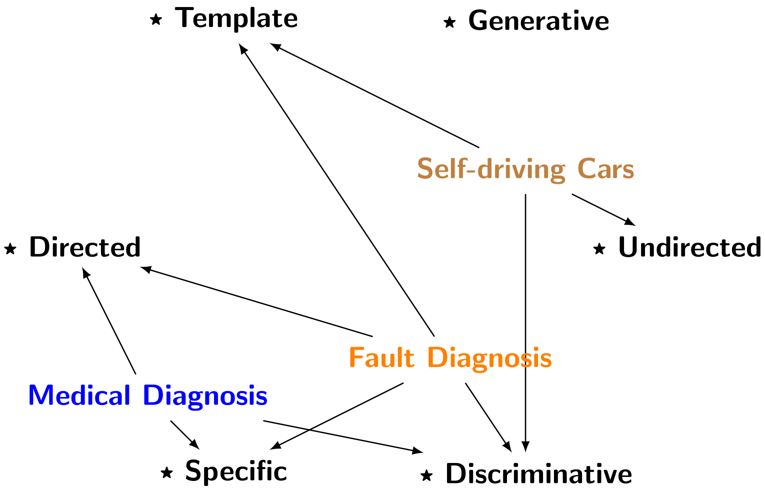
Medical Diagnosis

- ⋆ **Specific**
- ⋆ **Discriminative**

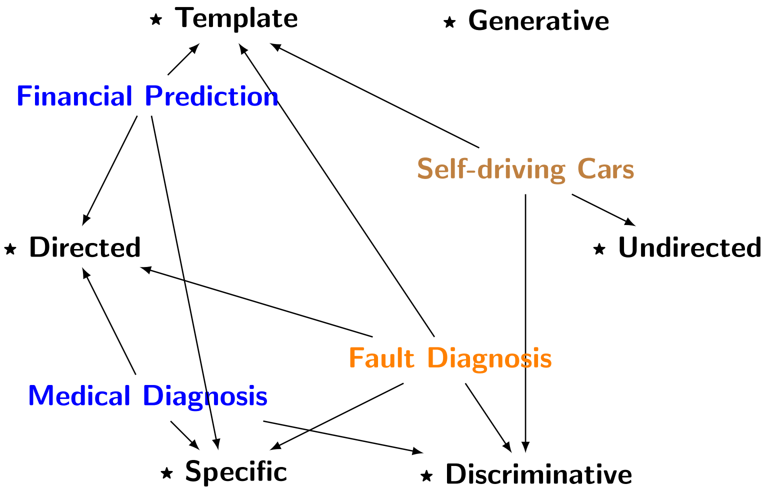
What problem are you trying to solve?



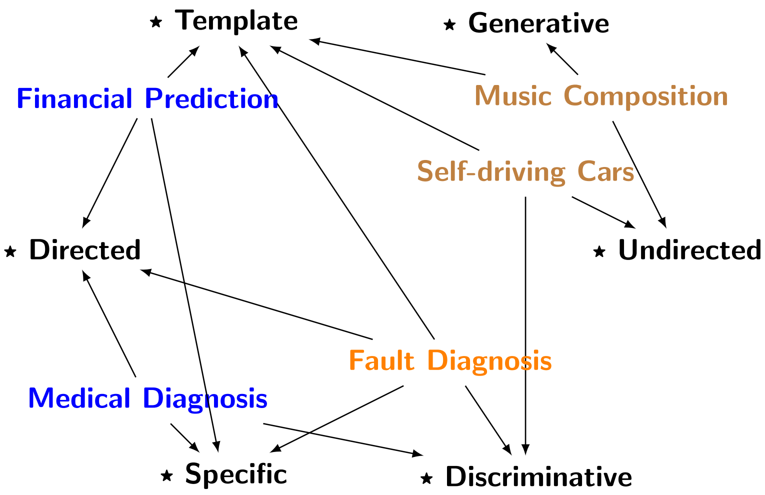
What problem are you trying to solve?



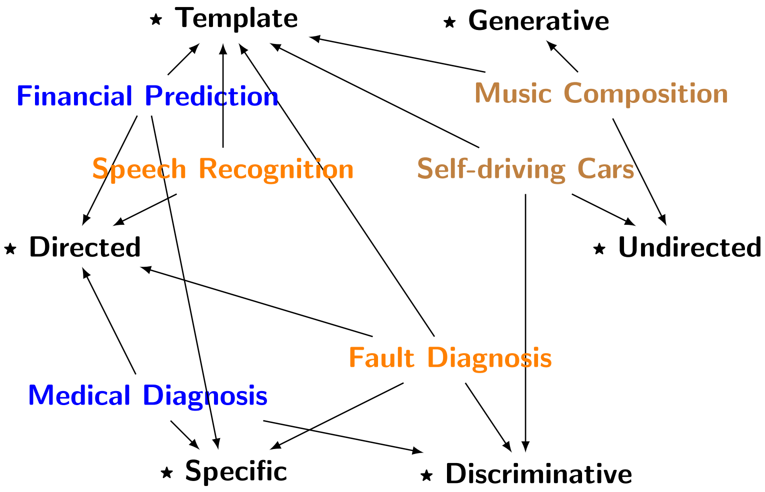
What problem are you trying to solve?



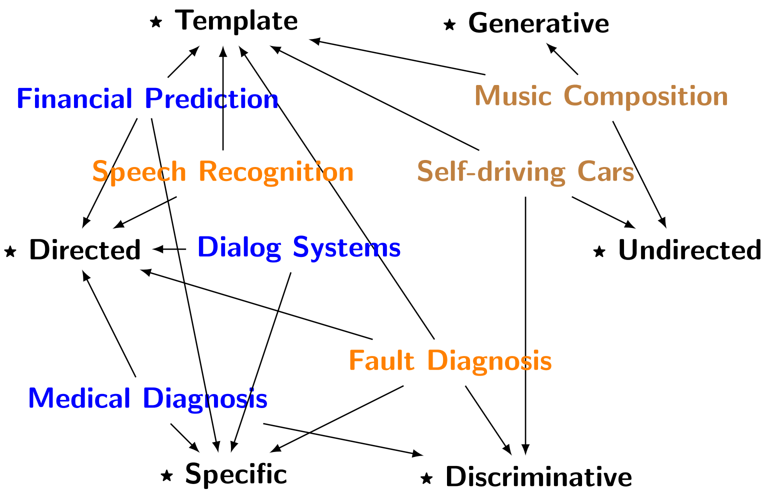
What problem are you trying to solve?



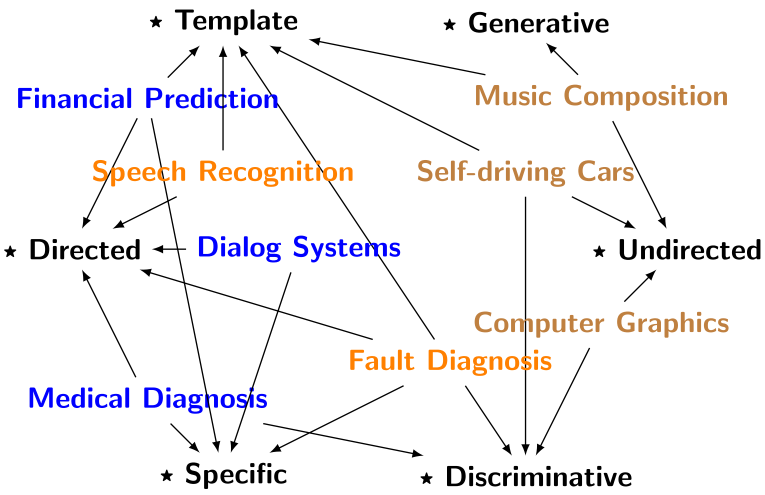
What problem are you trying to solve?



What problem are you trying to solve?



What problem are you trying to solve?



Building Graphical Models for Practical Applications

Picking Variables:

- We want **relevant variables** that we can observe.
- We want variables that we would like to **query**.
- Some abstractions as hidden variables?

Picking a Structure:

- Reflects the **causal order**: Causes are parents of effects
- Keep the network **sparse**: less expensive to use.

Picking Probabilities:

- Qualitatively: associate **rankings to a scale** ("common", "rare", "surprising")
- Don't assign **zero probabilities** to parameters (makes evidence negligible)
- Order of magnitude: Small differences (10E-5 vs 10E-6)
- Relative values: subtleties between probabilities with evidence
- Use **sensitivity analysis**: extent that a probability affects an outcome.