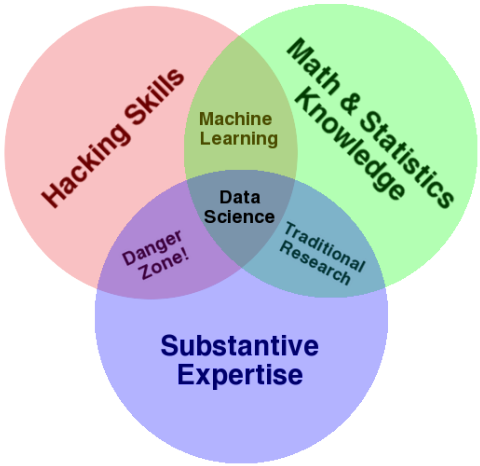
# Probabilitistic Graphical Models for Fraud and Anomaly Detection in Insurance

Mick Cooney
michael.cooney@applied.ai

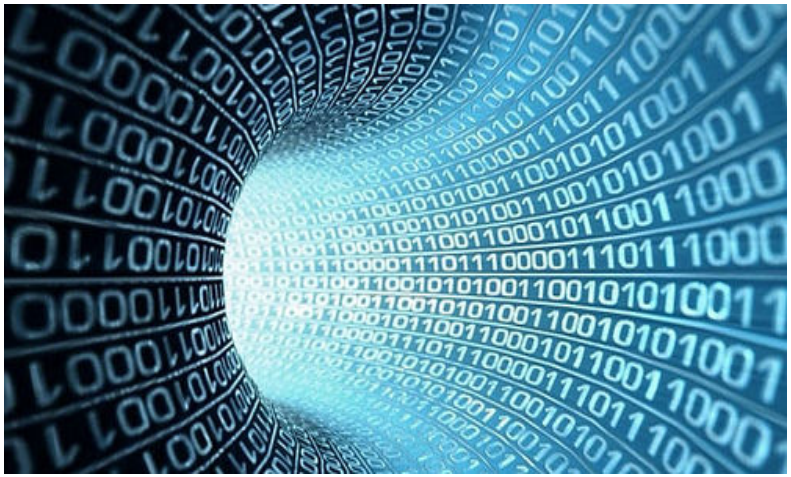
6 July 2016

# How to Build a Model with No Data and No Domain Knowledge...

# Obligatory Cliches




*"All models are wrong, but some are useful."*
- George Box



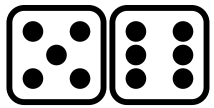
*"Data is the new gold."*
- Some Marketing Chancer

# Structure of Talk

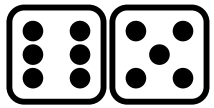
- Conditional Dependence, Independence and Bayesian Networks
- The Sprinkler Network
- Medical Non-disclosure
- Building a Model
- Expanding the Model
- Beyond Bayesian Networks
- Summary

# Conditional Probability

Probability of 2D6 totalling 11?



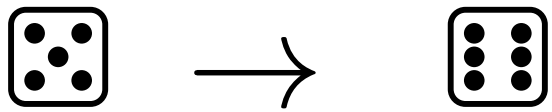
or



$$P(T = 11) = \frac{2}{36} = 0.05556$$

# Conditional Probability

Probability of 2D6 totalling 11 if first dice is 5?



$$P(T = 11|D_1 = 5) = \frac{1}{6} = 0.1667$$

# Conditional Dependence and Independence

Three variables, $A$, $B$, $C$:

$A$, $B$ independent
$C$ depends on $A$
$C$ depends on $B$

Learn information for $C$?

$A$ and $B$ *conditionally dependent* given $C$.

# 2D6 Example

Define variables $D_1$, $D_2$ and $T$:

$$T = D_1 + D_2$$

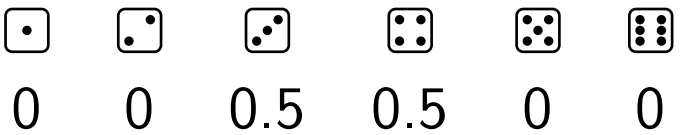
$D_1$ and $D_2$ independent, $T$ depends on both.

$D_1$ if $T = 7$, $D_1 = 4$?

$$P(D_2 = 3) = 1$$

# 2D6 Example

Fix $T = 9$:

$P(D_1) =$ 

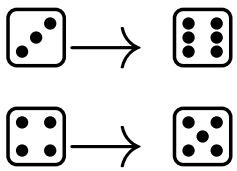
| 0 | 0 | 0.5 | 0.5 | 0 | 0 |
|---|---|---|---|---|---|

induces

$P(D_2) =$ 

| 0 | 0 | 0 | 0 | 0.5 | 0.5 |
|---|---|---|---|---|---|

because



# Conditional Independence

Still have $T = D_1 + D_2$

Define new variables:

$$X_1 = \begin{cases} 1 \text{ iff } T \text{ even} \\ 0 \text{ otherwise} \end{cases} \qquad X_2 = \begin{cases} 1 \text{ iff } T >= 9 \\ 0 \text{ otherwise} \end{cases}$$

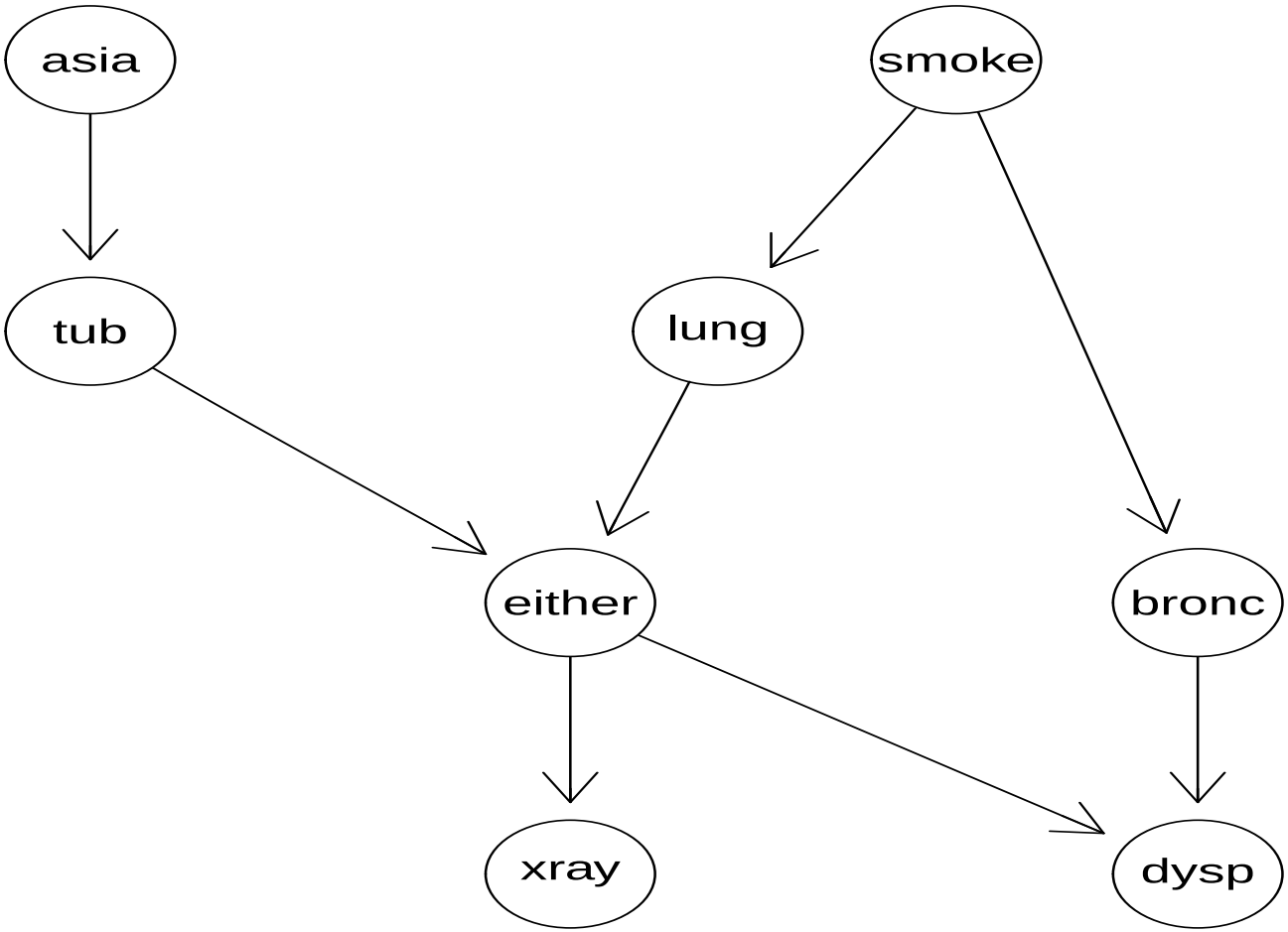
$$\begin{aligned} T \text{ not known} &\Rightarrow X_1, X_2 \text{ dependent} \\ T \text{ known} &\Rightarrow X_1, X_2 \text{ independent} \end{aligned}$$

$X_1$ and $X_2$ are *conditionally independent* given $T$

# Conditional Independence

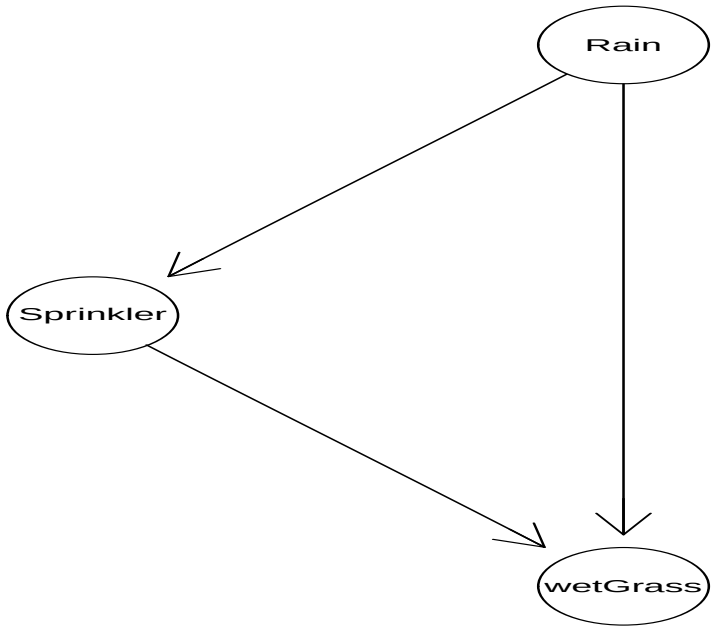
Probabilistic Graphical Models represent structural dependence amongst variables

# Conditional Independence



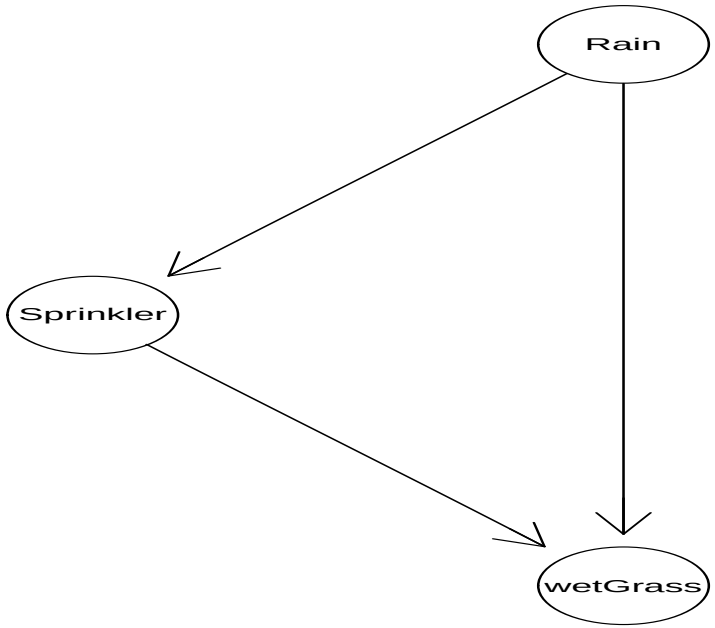
# Bayesian Networks

PGM with *directed, acyclic graph* (DAG):




*Conditional Probability Tables* (CPTs)

# The Sprinkler Network




Variables: (R)aining, (S)prinkler, wet(G)rass

# The Sprinkler Network

```
print(sprinkler_grain$cptlist$Rain)
```

```
## Rain
## yes  no
## 0.2 0.8
```

```
ftable(sprinkler_grain$cptlist$Sprinkler, row.vars = 'Rain')
```

```
##      Sprinkler  yes   no
## Rain
## yes            0.01 0.99
## no             0.40 0.60
```

```
ftable(sprinkler_grain$cptlist$wetGrass, row.vars = c('Rain', 'Sprinkler'))
```

```
##                wetGrass  yes   no
## Rain Sprinkler
## yes  yes                0.99 0.01
##      no                 0.90 0.10
## no   yes                0.80 0.20
##      no                 0.00 1.00
```

# Some Questions

What is the probability of the grass being wet?

```
querygrain(sprinkler_grain, nodes = 'wetGrass')$wetGrass
```

```
## wetGrass
##     yes      no
## 0.43618 0.56382
```

If the grass is wet, what is the probability that it is raining?

```
querygrain(sprinkler_grain, evidence = list(wetGrass = 'yes'), nodes = 'Rain')$Rain
```

```
## Rain
##      yes       no
## 0.413086 0.586914
```

# Medical Non-disclosure

SSQ Financial Group

**REQUEST FOR OPTIONAL LIFE INSURANCE**

PLEASE COMPLETE THIS FORM IN BLOCK LETTERS USING INK

**A. EMPLOYER INFORMATION**

| | |
|---|---|
| Policy Holder Name: | SSQ Group #: |
| Division Name: | Certificate #: |

**B. PARTICIPANT INFORMATION**

| | | |
|---|---|---|
| Last Name: | First Name: | S.I.N.: |
| Mailing Address: (including postal code) | | |
| Telephone: Home | Work | Language Preference: ☐ English ☐ French |
| Gender: ☐ M ☐ F | Date of Birth: D M Y | Salary: $ |

**C. REQUEST FOR OPTIONAL LIFE INSURANCE COVERAGE**

**IMPORTANT: Optional Life Insurance units of $10,000 are only available to plans that currently offer this benefit.**

| Participant: *(Please check N/A if request is only for spouse)* | | Spouse: | |
|---|---|---|---|
| Current amount of coverage (in force) | Additional amount of coverage (requested) | Current amount of coverage (in force) | Additional amount of coverage (requested) |
| ☐ None ☐ 1x salary ☐ 2x salary ☐ 3x salary _____ units of $10,000 | ☐ N/A ☐ 1x salary ☐ 2x salary ☐ 3x salary _____ units of $10,000 | ☐ None ☐ 25% _____ units of $10,000 | ☐ 25% ☐ 50% _____ units of $10,000 |

| | | |
|---|---|---|
| Spouse: | Last Name: | First Name: |
| | Gender: ☐ M ☐ F | Date of Birth: D M Y |

**D. SMOKING HABITS**

**Participant:** Non-Smoker ☐ Smoker ☐ **Spouse:** Non-Smoker ☐ Smoker ☐

"I declare that I do not smoke and have not smoked any tobacco products such as cigarettes, cigars, cigarillos or pipes, or any drugs during the past 12 months. This statement is an affirmative guarantee on my part." It is understood that the insurer may periodically require confirmation of non-smoker status. The participant must be in a position to meet the requirements then in force and return the confirmation within 30 days of the request, failing which the participant shall lose non-smoker status and the associated premium reduction shall cease to apply as of the date of the insurer's request. "I also acknowledge that a false or incomplete statement may cause the coverage to be null and void."

Participant: _____ Spouse: _____

# Problems

Data sparse / missing

Partially missing output variable

Low base-rate problem

Semi-supervised learning

# Fraud Detection

# Fraud Detection

# Fraud Detection

# Fraud Detection

Full automation difficult!

Create filter instead — triage cases

# Build a Model

*We want a model which, given the data observed in the policy application, allows us to estimate the probability of a subsequent medical exam changing the underwriting decision on the policy.*

*The model should incorporate our assumptions of the process and be as simple as possible.*

# Consequences

# Getting Started

Conditions:

(S)moker: Smoker, Quitter, Non-smoker

(B)MI: Normal, Overweight, Obese

Family (H)istory: None, HeartDisease

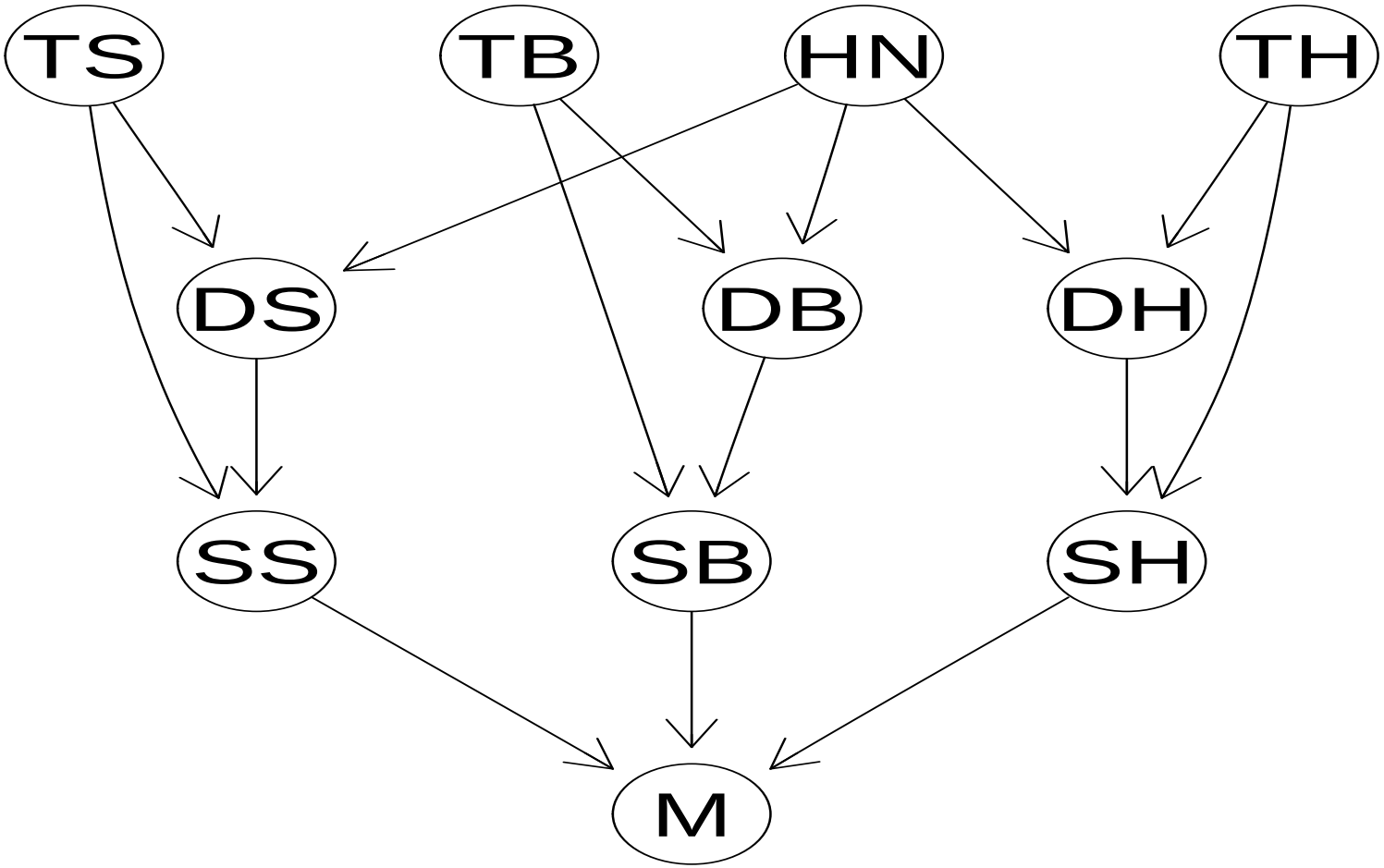
Aspects:

T True state

D Declared state

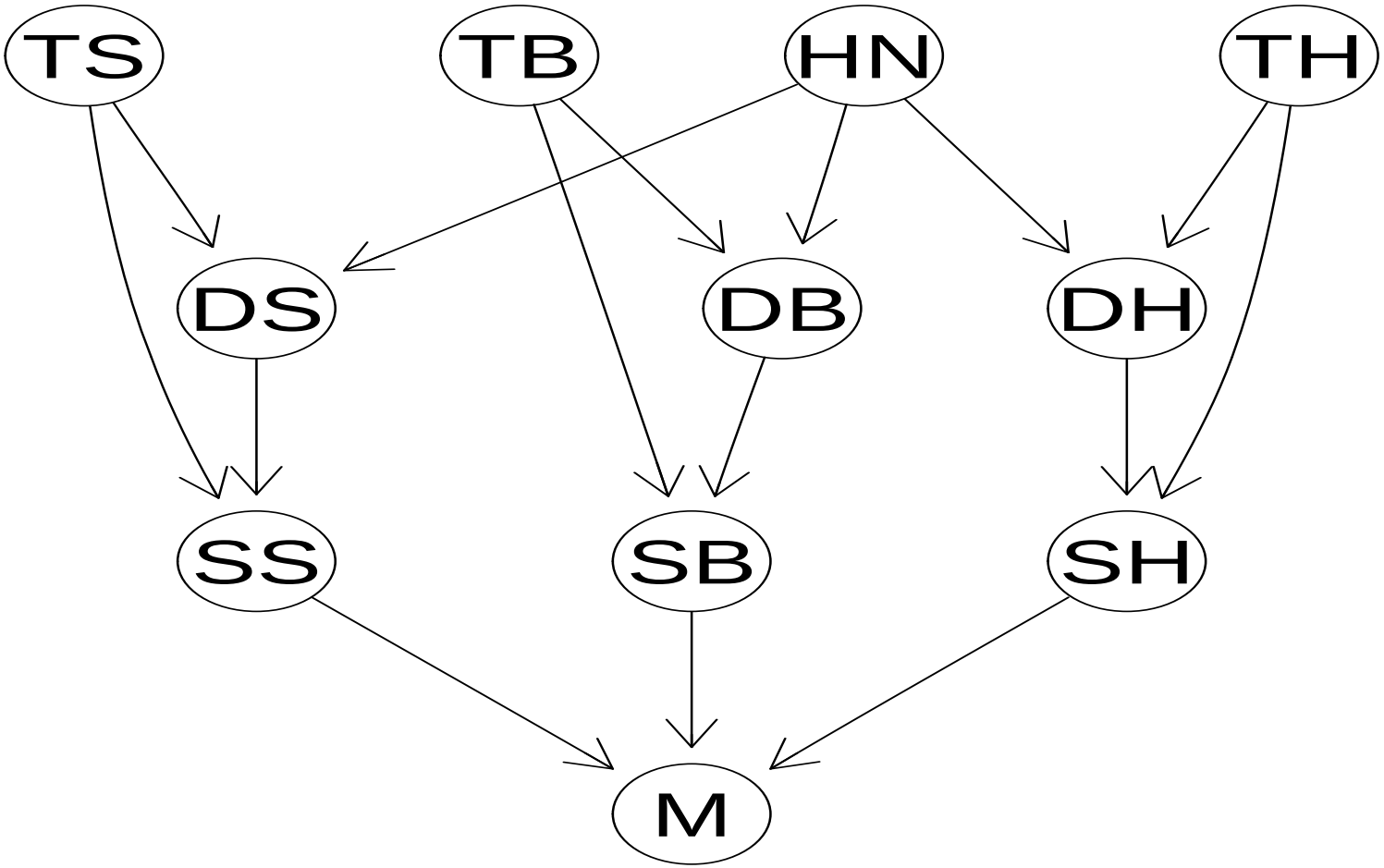
S Seriousness of condition's impact on decision

# Medical Exam Network



# Bad Teacher Syndrome

# Conditional Independence



```
print(underwriting_grain$cptlist$TH)
```

```
## TH
##         None HeartDisease
##         0.95         0.05
```

```
ftable(underwriting_grain$cptlist$DH, row.vars = c('HN', 'TH'))
```

```
##                         DH None HeartDisease
## HN        TH
## Dishonest None             0.9         0.1
##           HeartDisease     0.5         0.5
## Honest    None             0.9         0.1
##           HeartDisease     0.1         0.9
```

```
ftable(underwriting_grain$cptlist$SH, row.vars = c('TH', 'DH'))
```

```
##                            SH Serious NotSerious
## TH           DH
## None         None               0.01       0.99
##              HeartDisease       0.20       0.80
## HeartDisease None               0.60       0.40
##              HeartDisease       0.10       0.90
```

# Medical Exam

```
ftable(underwriting_grain$cptlist$M, row.vars = c('SS', 'SB', 'SH'))

##                                  M Medical NoMedical
## SS          SB          SH
## Serious     Serious     Serious         0.99      0.01
##                         NotSerious      0.85      0.15
##             NotSerious  Serious         0.95      0.05
##                         NotSerious      0.60      0.40
## NotSerious  Serious     Serious         0.90      0.10
##                         NotSerious      0.60      0.40
##             NotSerious  Serious         0.85      0.15
##                         NotSerious      0.10      0.90
```

What is the unconditional probability of a medical exam finding something?

```
querygrain(underwriting_grain, nodes = 'M')$M

## M
##    Medical NoMedical
##   0.177515  0.822485
```

Too high?

Probably flawed

# Assess the Model

Declares a clean bill of health ($DS$ = Nonsmoker, $DB$ = Normal, $DH$ = None)?

```
querygrain(underwriting_grain, nodes = 'M'
           ,evidence = list(DS = 'Nonsmoker'
                            ,DB = 'Normal'
                            ,DH = 'None'))$M
```

```
## M
##    Medical NoMedical
##   0.146951  0.853049
```

# Assess the Model

Declares history of heart disease? ($DH$ = HeartDisease)?

```
querygrain(underwriting_grain, nodes = 'M'
           ,evidence = list(DS = 'Nonsmoker'
                            ,DB = 'Normal'
                            ,DH = 'HeartDisease'))$M

## M
##    Medical NoMedical
##   0.257899  0.742101
```

# Assess the Model

Clean bill of health — are they dishonest?

```
querygrain(underwriting_grain, nodes = 'HN'
           ,evidence = list(DS = 'Nonsmoker'
                            ,DB = 'Normal'
                            ,DH = 'None'))$HN

## HN
## Dishonest    Honest
## 0.0138621 0.9861379
```

# Expanding the Model

Current model built by guessing CPTs

Use Data?

CPTs assist this - subsets of variables available

Bootstrap to assess calculation validity?

# Expanding Variable Levels

Add states/levels to variables – `HeartDisease`?

Limitations:

- Model separately?
- CPT specification complicated
- More data

# Add Variables

Add variables to model

- Family history?
- Work on honesty modelling
- Split exam types

Potential for bias

# Further Uses

- Underwriting fraud for car insurance
- Claims fraud
- Product recommendations
- Problematic customers
- Regulatory issues

# Beyond Bayesian Networks

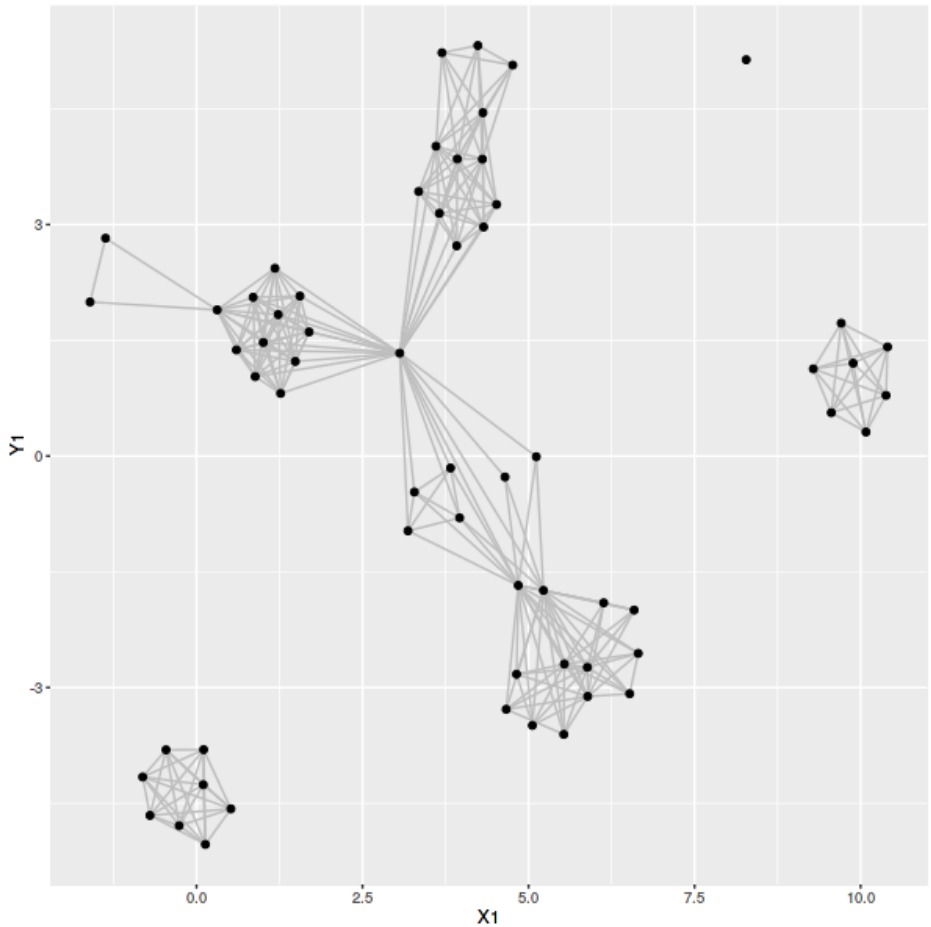
Require categorical variables

Binning continuous data loses information

- Markov Random Fields
- Chain graphs
- Conditional Random Fields

Semi-supervised Learning

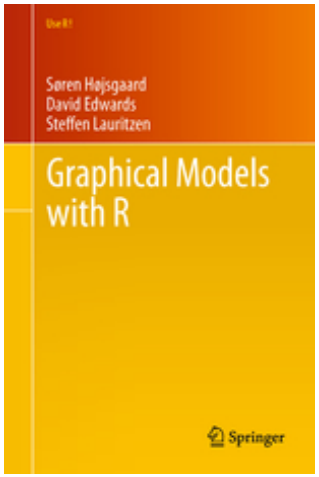
# One Last Thing...



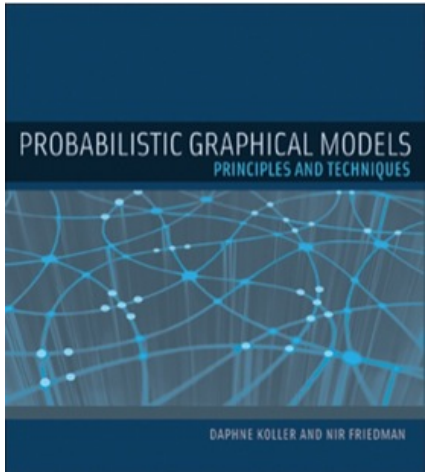
# Conclusions

- Classification very difficult
- Highly speculative – nowhere near production-ready
- Use as filter – no automation
- Outputs often counter-intuitive
- Work unfinished - lots more avenues to explore

# Further Resources




"Graphical Models with R"
Søren Højsgaard.




"Probabalistic Graphical Models: Principles and Techniques"
Koller and Friedman

Package Vignettes: `gRain` and `gRbase`

Coursera: Probabilistic Graphical Models `https://www.coursera.org/course/pgm`

# Get In Touch

Mick Cooney
michael.cooney@applied.ai

Slides and code available on GitHub:
`https://www.github.com/kaybenleroll/dublin_r_workshops`

Blogpost Series:
`http://blog.applied.ai/probabilistic-graphical-models-for-fraud-detection-part-1`
`http://blog.applied.ai/probabilistic-graphical-models-for-fraud-detection-part-2`
`http://blog.applied.ai/probabilistic-graphical-models-for-fraud-detection-part-3`