

# Probabilistic Modelling and Reasoning — Introduction —

Chris Williams  
(based on slides by Michael U. Gutmann)

Probabilistic Modelling and Reasoning (INFR11134)  
School of Informatics, The University of Edinburgh

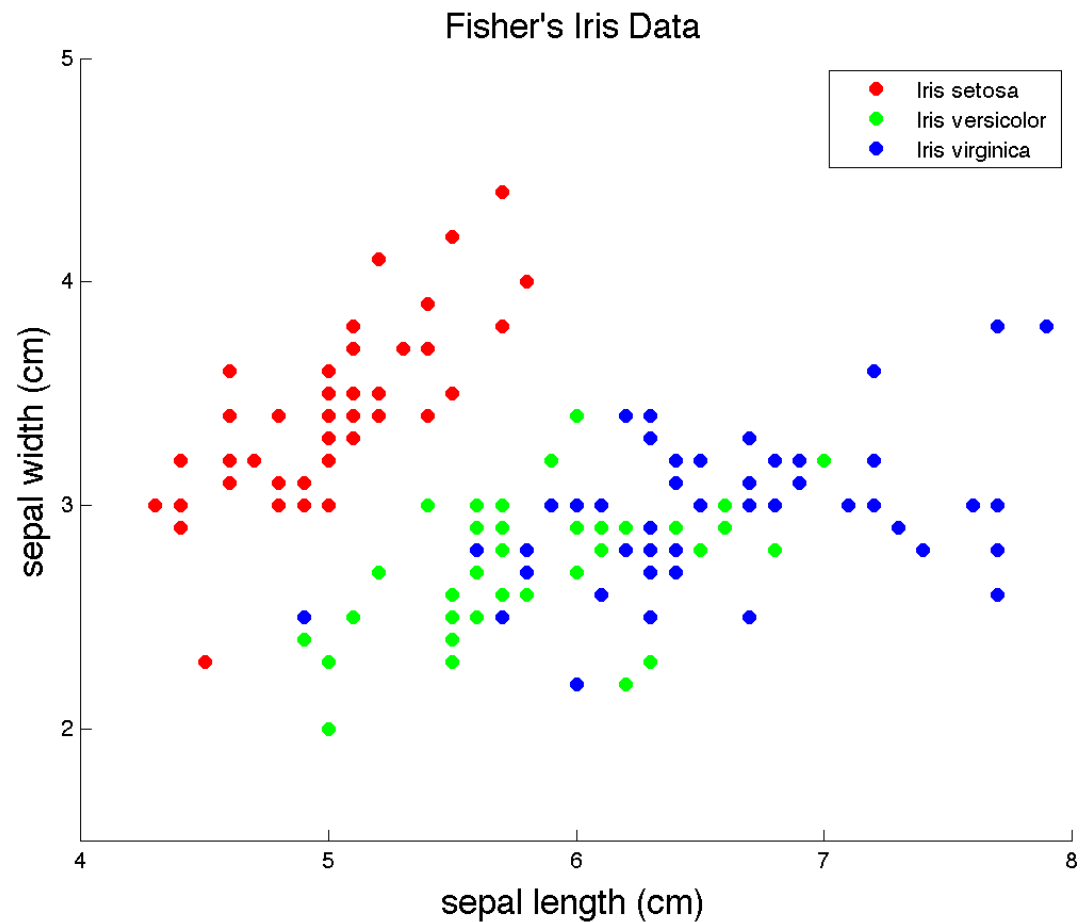
Spring Semester 2023

# Course Administration

- ▶ Welcome!
- ▶ Course website: `https://www.inf.ed.ac.uk/teaching/courses/pmr/22-23/`
- ▶ Lecture recordings via Learn
- ▶ Lectures
- ▶ Tutorials
- ▶ Labs
- ▶ Quizzes
- ▶ No assignments
- ▶ Resources
- ▶ Maths background (MLPR resources)

# Variability

- ▶ Variability is part of nature
- ▶ Data for 3 species of iris, from Ronald Fisher (1936)



# Variability

- ▶ Our handwriting is unique
- ▶ Variability leads to uncertainty: e.g. 1 vs 7 or 4 vs 9



Josef Stepan - Own work, CC BY-SA 4.0, <https://commons.wikimedia.org/w/index.php?curid=64810040>

# Variability

- ▶ Variability leads to uncertainty
- ▶ Reading handwritten text in a foreign language

**啤酒冷たいです**  
 毎週火曜日の中生200円！

・サッポロ生ビール (中) 390 (409)  
 (大) 750 (787)  
 ・爽比壽! 中ビシ 500 (526)  
 爽比壽! 黒小ビシ 480 (500)  
 ・ゴロチニニマズ/SP/ALL  
 ・バビワササ/CP/SP/ALL  
 ・ホッピーあります! 380 (399)  
 中4 200 (210)  
 ・ピルピル (1.5リットル) 480 (500)

**サウ~290~**

チョコアイス 400円  
 トロトロアイス 300円  
 (おまけ) 300円  
 グラス 290 (306) 生ビール 590 (617)

梅干し 200円  
 スライス 200円  
 トマト 200円  
 生卵 200円  
 生クリーム 200円

**カクテル**

・カクテル 430 (451)  
 ・カクテル 480 (500)  
 ・カクテル 980 (1009)

**デザート**

・デザート 580 (609)  
 ・デザート 600 (621)

**梅酒**

・梅酒 500 (526)  
 ・梅酒 580 (609)  
 ・梅酒 680 (706)

**その他**

・その他 250 (266)  
 ・その他 300 (316)

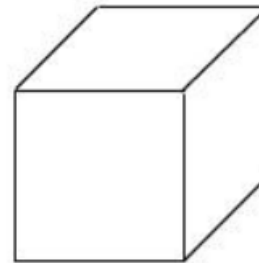


# Example: Screening and diagnostic tests

- ▶ Early warning test for Alzheimer's disease (Scharre, 2010, 2014)
- ▶ Detects “mild cognitive impairment”

- ▶ Takes 10–15 minutes
- ▶ Freely available
- ▶ Assume a 70 year old man tests positive.
- ▶ Should he be concerned?

7. Copy this picture:



---

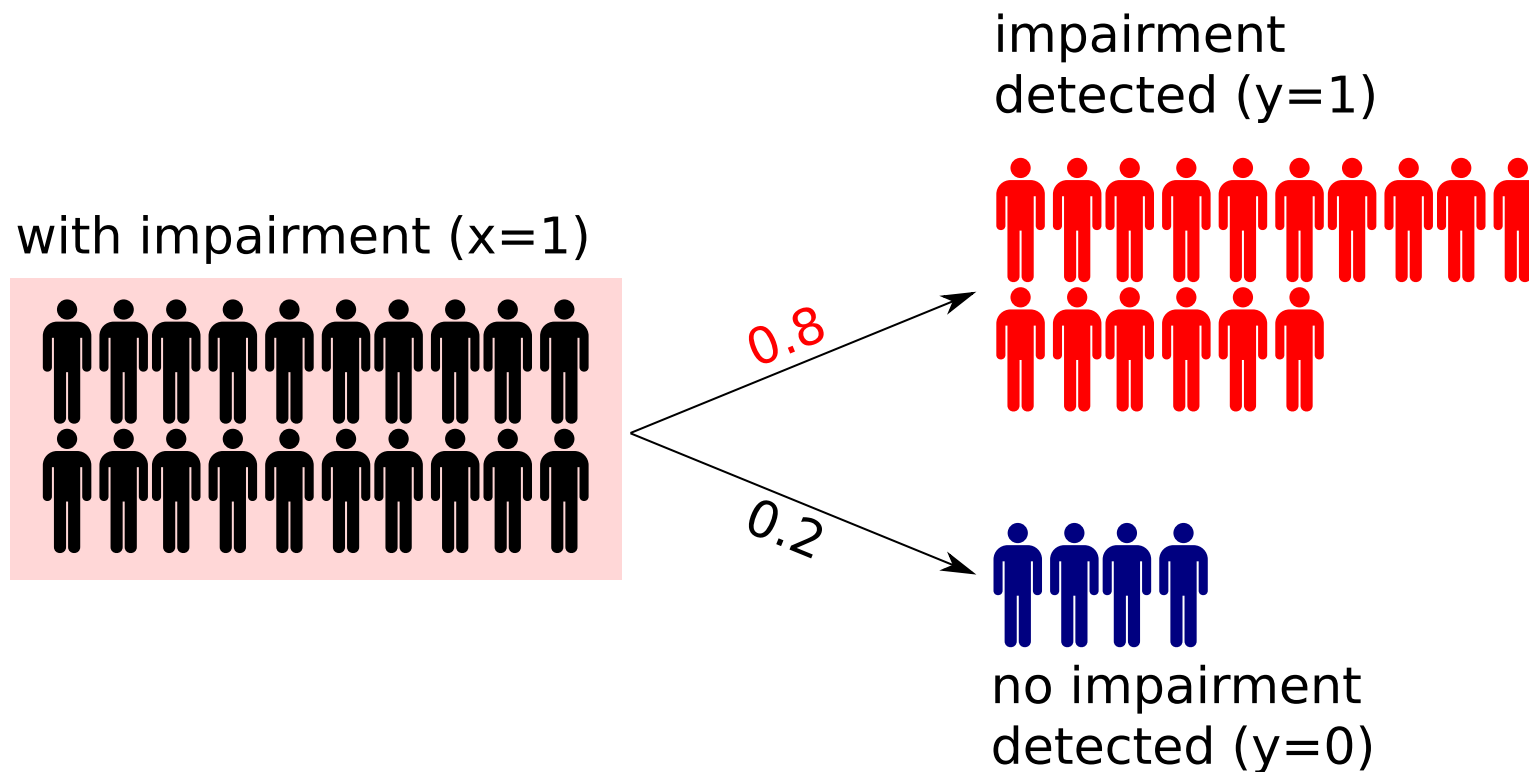
8. Drawing test

- Draw a large face of a clock and place in the numbers
- Position the hands for 5 minutes after 11 o'clock

(Example from [sagetest.osu.edu](http://sagetest.osu.edu))

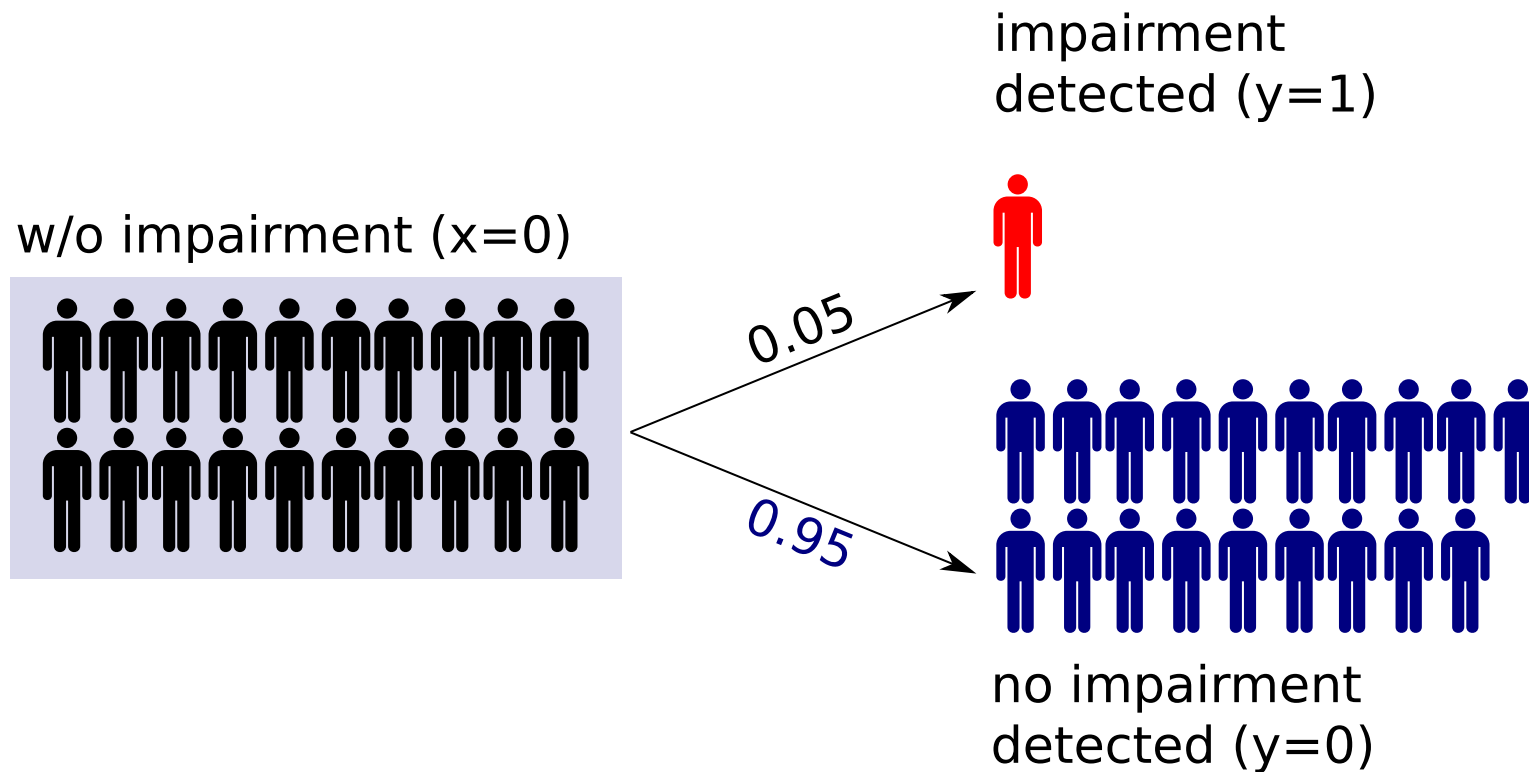
# Accuracy of the test

- ▶ Sensitivity of **0.8** and specificity of **0.95** (Scharre, 2010)
- ▶ **80%** correct for people with impairment



# Accuracy of the test

- ▶ Sensitivity of **0.8** and specificity of **0.95** (Scharre, 2010)
- ▶ **95%** correct for people w/o impairment





# Variability implies uncertainty

- ▶ People of the same group do not have the same test results
  - ▶ Test outcome is subject to variability
  - ▶ The data are noisy
- ▶ Variability leads to uncertainty
  - ▶ Positive test  $\equiv$  true positive ?
  - ▶ Positive test  $\equiv$  false positive ?
- ▶ What can we safely conclude from a positive test result?
- ▶ How should we analyse such kind of ambiguous data?

# Probabilistic approach

- ▶ The test outcomes  $y$  can be described with probabilities

$$\text{sensitivity} = 0.8 \quad \Leftrightarrow \quad \mathbb{P}(y = 1|x = 1) = 0.8$$

$$\Leftrightarrow \quad \mathbb{P}(y = 0|x = 1) = 0.2$$

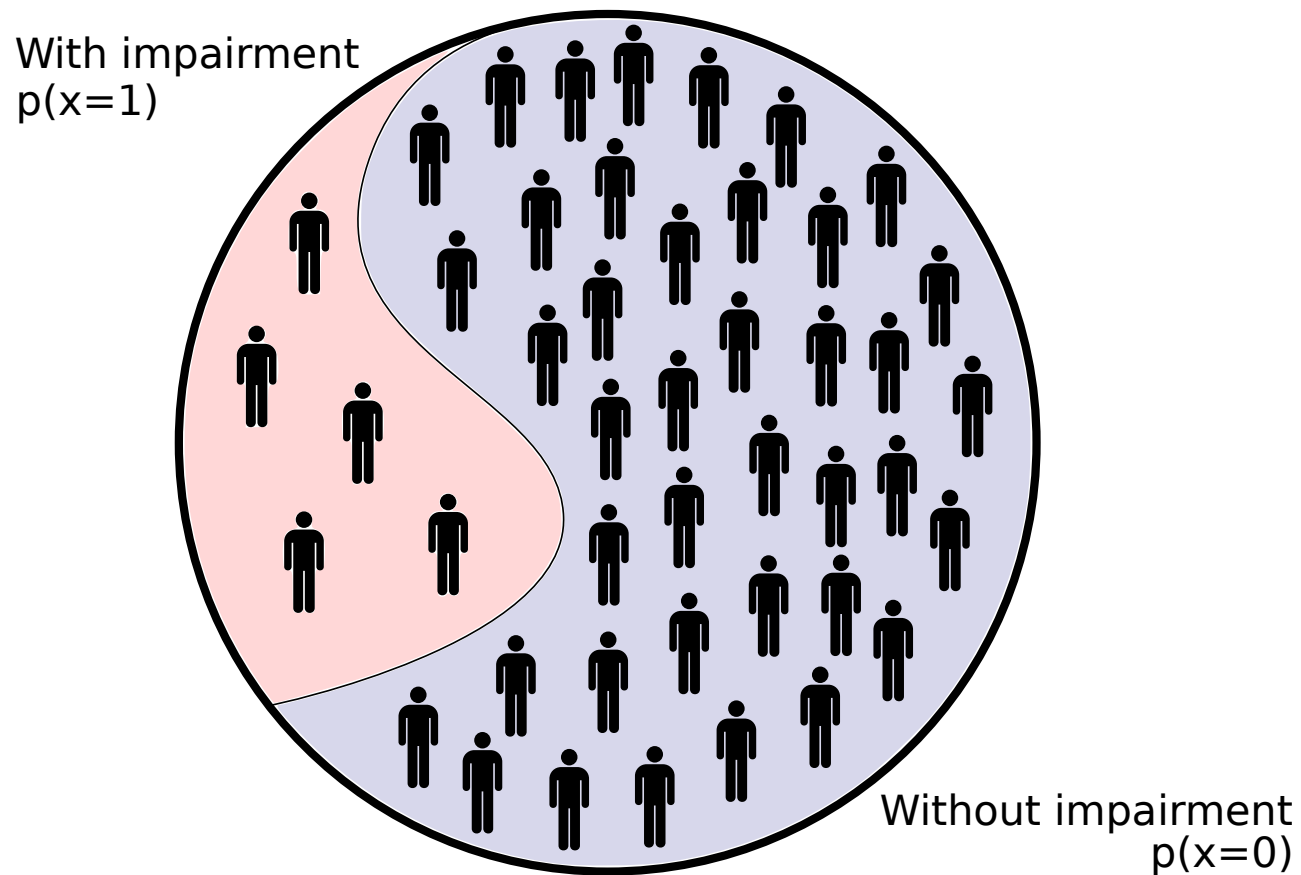
$$\text{specificity} = 0.95 \quad \Leftrightarrow \quad \mathbb{P}(y = 0|x = 0) = 0.95$$

$$\Leftrightarrow \quad \mathbb{P}(y = 1|x = 0) = 0.05$$

- ▶  $\mathbb{P}(y|x)$ : model of the test specified in terms of (conditional) probabilities
- ▶  $x \in \{0, 1\}$ : quantity of interest (cognitive impairment or not)

# Prior information

Among people like the patient,  $\mathbb{P}(x = 1) = 5/45 \approx 11\%$  have a cognitive impairment (plausible range: 3% – 22%, Geda, 2014)

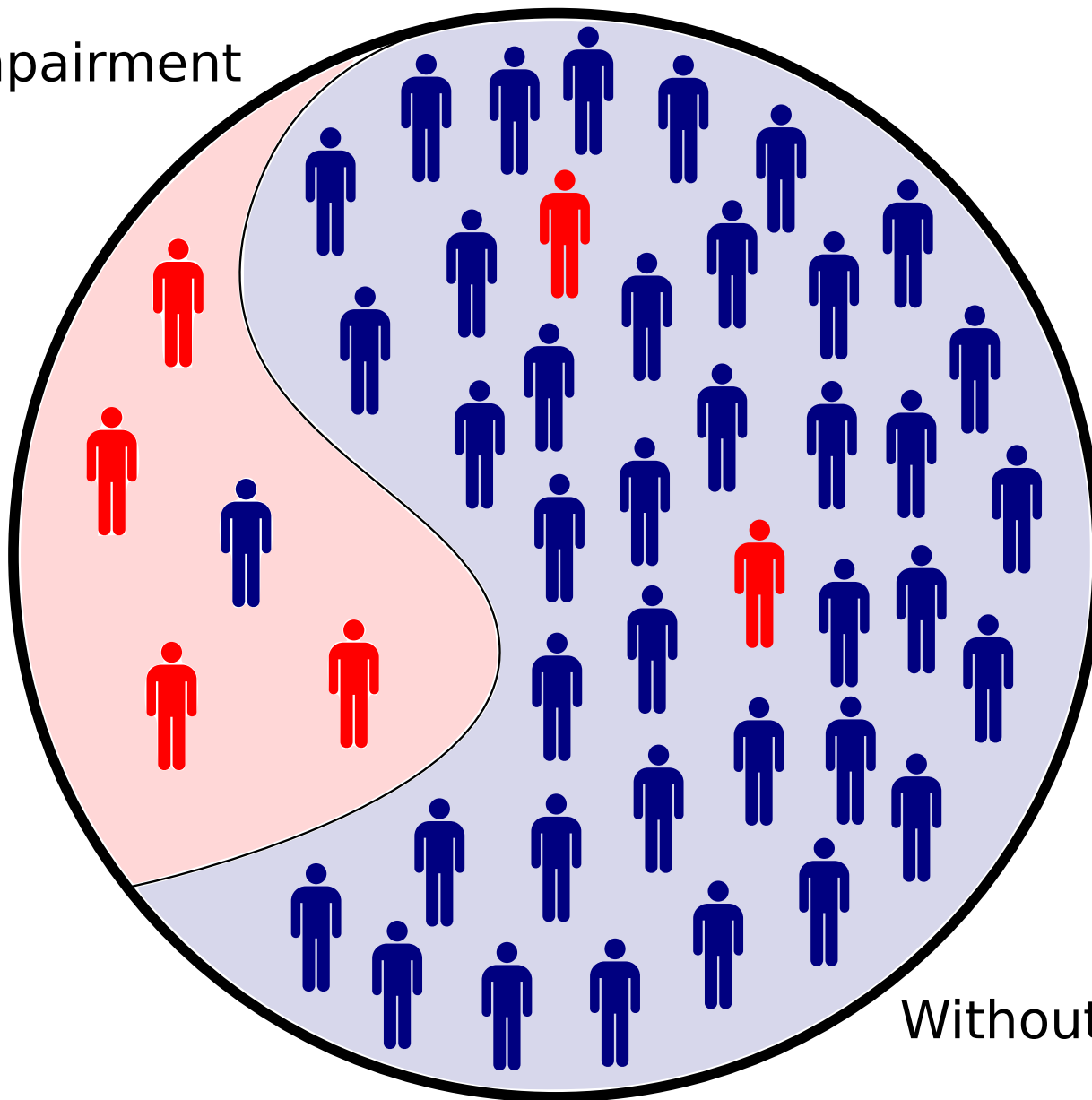


# Probabilistic model

- ▶ Reality:
    - ▶ properties/characteristics of the group of people like the patient
    - ▶ properties/characteristics of the test
  - ▶ Probabilistic model:
    - ▶  $\mathbb{P}(x = 1)$
    - ▶  $\mathbb{P}(y = 1|x = 1)$  or  $\mathbb{P}(y = 0|x = 1)$   
 $\mathbb{P}(y = 1|x = 0)$  or  $\mathbb{P}(y = 0|x = 0)$
- Fully specified by three numbers.
- ▶ A probabilistic model is an abstraction of reality that uses probability theory to quantify the chance of uncertain events.

# If we tested the whole population

With impairment  
 $p(x=1)$

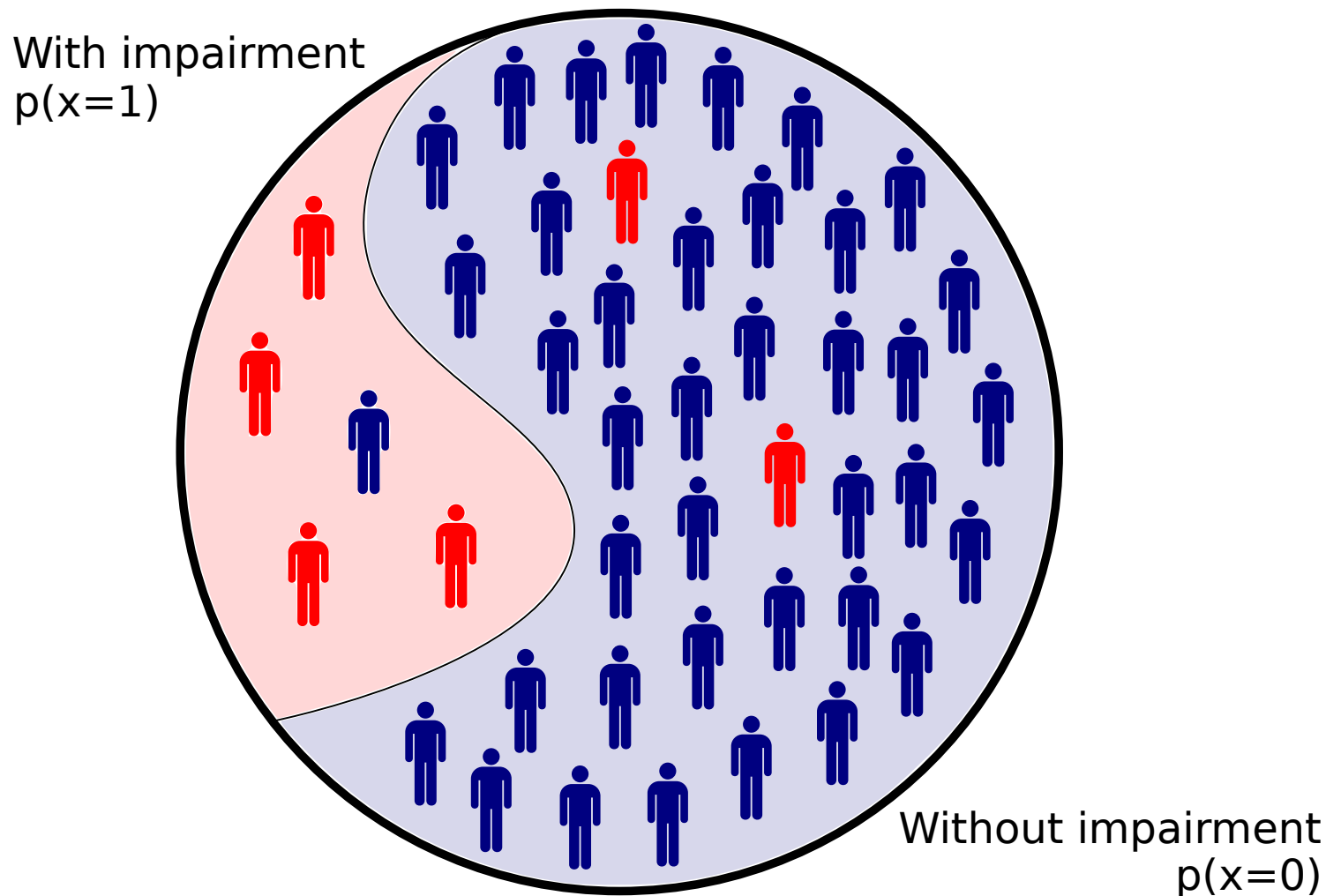


Without impairment  
 $p(x=0)$

# If we tested the whole population

Fraction of people who are impaired and have positive tests:

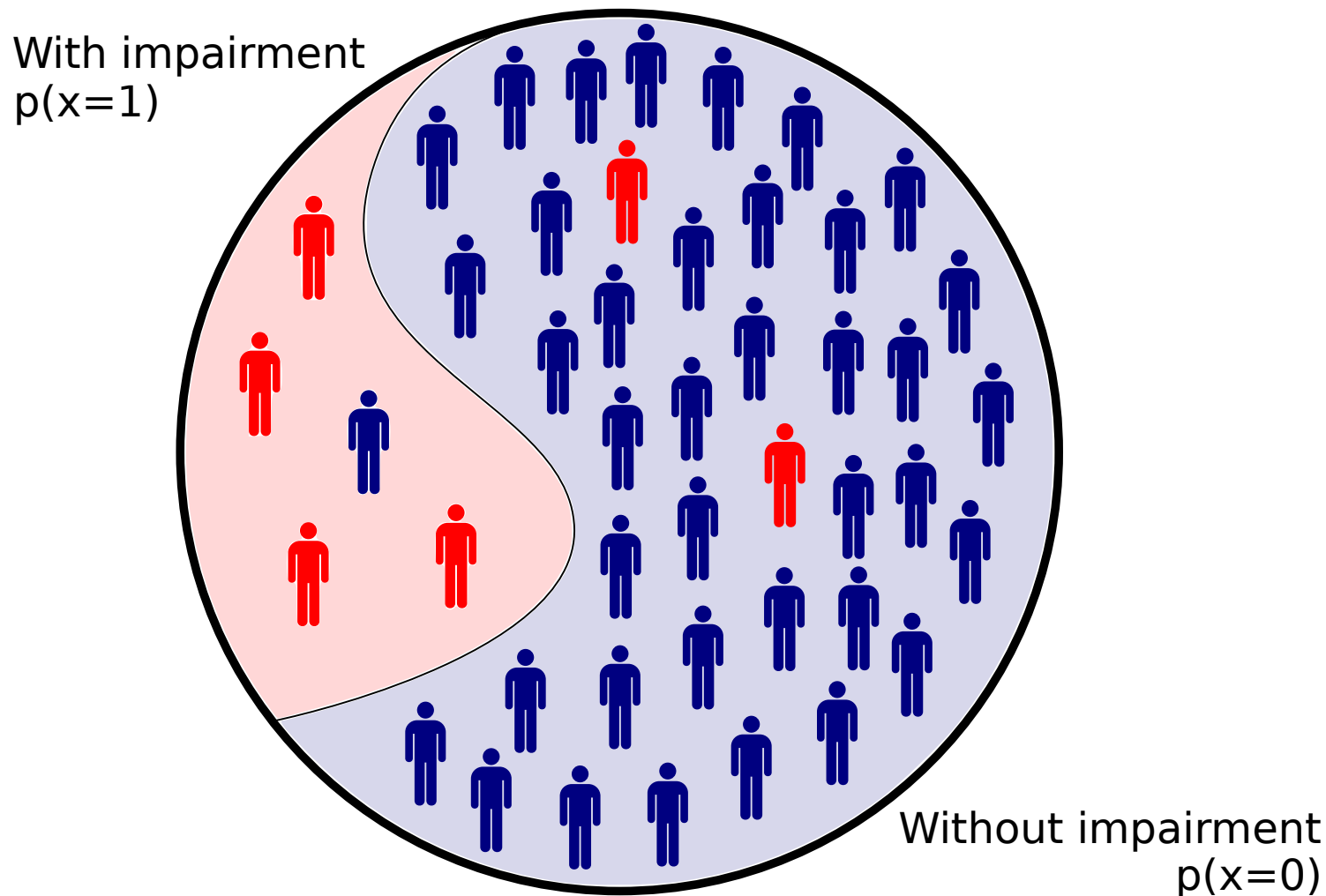
$$\mathbb{P}(x = 1, y = 1) = \mathbb{P}(y = 1|x = 1)\mathbb{P}(x = 1) = 4/45 \quad (\text{product rule})$$



# If we tested the whole population

Fraction of people who are not impaired but have positive tests:

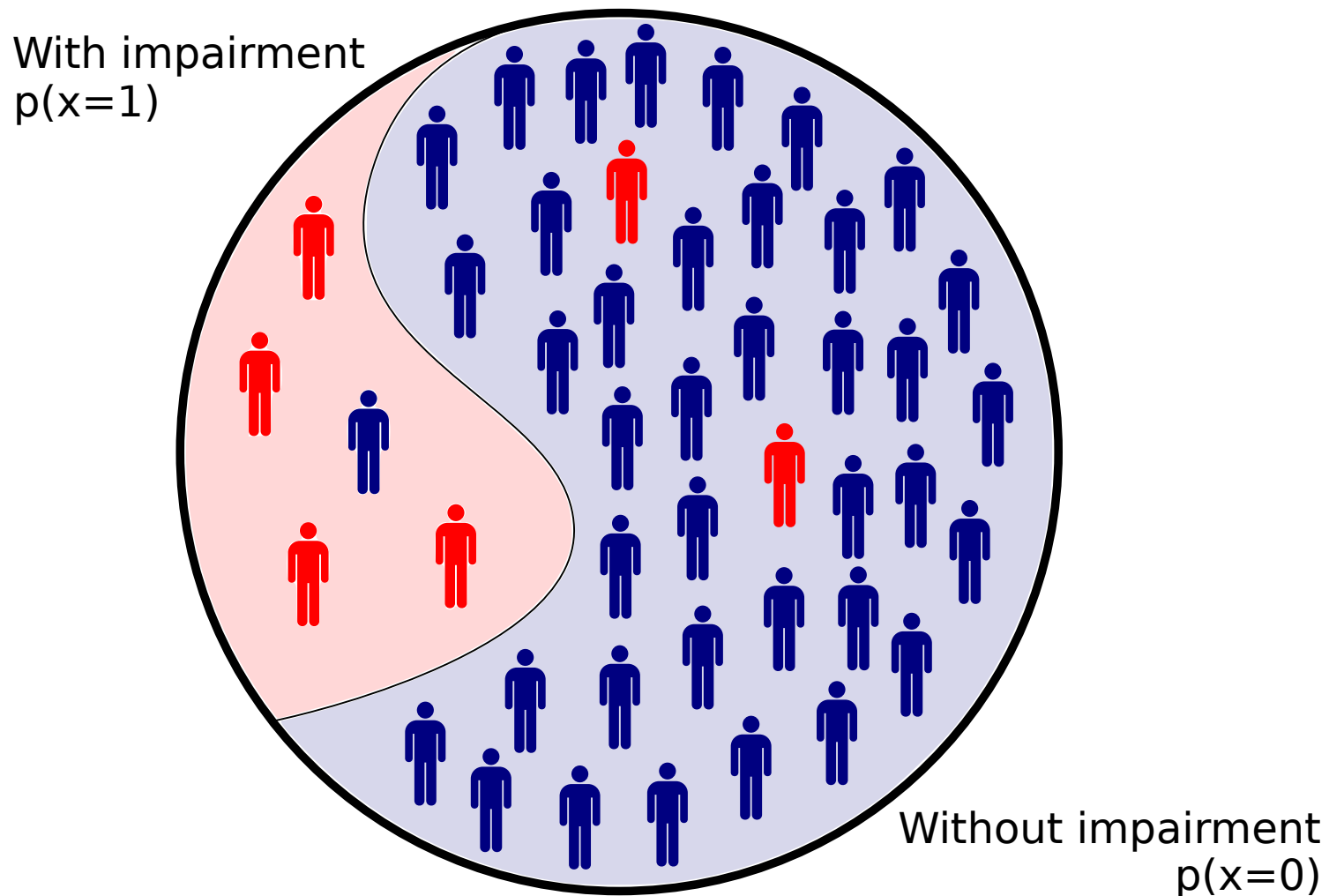
$$\mathbb{P}(x = 0, y = 1) = \mathbb{P}(y = 1|x = 0)\mathbb{P}(x = 0) = 2/45 \quad (\text{product rule})$$



# If we tested the whole population

Fraction of people where the test is positive:

$$\mathbb{P}(y = 1) = \mathbb{P}(x = 1, y = 1) + \mathbb{P}(x = 0, y = 1) = 6/45 \quad (\text{sum rule})$$





# Putting everything together

- ▶ Among those with a positive test, fraction with impairment:

$$\mathbb{P}(x = 1|y = 1) = \frac{\mathbb{P}(y = 1|x = 1)\mathbb{P}(x = 1)}{\mathbb{P}(y = 1)} = \frac{4}{6} = \frac{2}{3}$$

- ▶ Fraction without impairment:

$$\mathbb{P}(x = 0|y = 1) = \frac{\mathbb{P}(y = 1|x = 0)\mathbb{P}(x = 0)}{\mathbb{P}(y = 1)} = \frac{2}{6} = \frac{1}{3}$$

- ▶ Equations are examples of “Bayes’ rule”.
- ▶ Positive test increased probability of cognitive impairment from 11% (prior belief) to 67%, or from 6% to 51%.
- ▶ 51%  $\approx$  coin flip

# Probabilistic reasoning

- ▶ Probabilistic reasoning  $\equiv$  probabilistic inference:  
Computing the probability of an event that we have not or cannot observe from an event that we can observe
  - ▶ Unobserved/uncertain event, e.g. cognitive impairment  $x = 1$
  - ▶ Observed event  $\equiv$  evidence  $\equiv$  data, e.g. test result  $y = 1$
- ▶ “The prior”: probability for the uncertain event before having seen evidence, e.g.  $\mathbb{P}(x = 1)$
- ▶ “The posterior”: probability for the uncertain event after having seen evidence, e.g.  $\mathbb{P}(x = 1|y = 1)$
- ▶ The posterior is computed from the prior and the evidence via Bayes’ rule.

# Key rules of probability

## (1) Product rule:

$$\begin{aligned}\mathbb{P}(x = 1, y = 1) &= \mathbb{P}(y = 1|x = 1)\mathbb{P}(x = 1) \\ &= \mathbb{P}(x = 1|y = 1)\mathbb{P}(y = 1)\end{aligned}$$

## (2) Sum rule:

$$\mathbb{P}(y = 1) = \mathbb{P}(x = 1, y = 1) + \mathbb{P}(x = 0, y = 1)$$

Bayes' rule (conditioning) as consequence of the product rule

$$\mathbb{P}(x = 1|y = 1) = \frac{\mathbb{P}(x = 1, y = 1)}{\mathbb{P}(y = 1)} = \frac{\mathbb{P}(y = 1|x = 1)\mathbb{P}(x = 1)}{\mathbb{P}(y = 1)}$$

Denominator from sum rule, or sum rule and product rule

$$\mathbb{P}(y = 1) = \mathbb{P}(y = 1|x = 1)\mathbb{P}(x = 1) + \mathbb{P}(y = 1|x = 0)\mathbb{P}(x = 0)$$

# Key rules or probability

- ▶ The rules generalise to the case of multivariate random variables (discrete or continuous)
- ▶ Consider the conditional joint probability density function (pdf) or probability mass function (pmf) of  $\mathbf{x}, \mathbf{y}$ :  $p(\mathbf{x}, \mathbf{y})$

(1) Product rule:

$$\begin{aligned} p(\mathbf{x}, \mathbf{y}) &= p(\mathbf{x}|\mathbf{y})p(\mathbf{y}) \\ &= p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) \end{aligned}$$

(2) Sum rule:

$$p(\mathbf{y}) = \begin{cases} \sum_{\mathbf{x}} p(\mathbf{x}, \mathbf{y}) & \text{for discrete r.v.} \\ \int p(\mathbf{x}, \mathbf{y}) d\mathbf{x} & \text{for continuous r.v.} \end{cases}$$

# Probabilistic modelling and reasoning

- ▶ Probabilistic modelling:
  - ▶ Identify the quantities that relate to the aspects of reality that you wish to capture with your model.
  - ▶ Consider them to be random variables, e.g.  $\mathbf{x}, \mathbf{y}, \mathbf{z}$ , with a joint pdf (pmf)  $p(\mathbf{x}, \mathbf{y}, \mathbf{z})$ .
- ▶ Probabilistic reasoning:
  - ▶ Assume you know that  $\mathbf{y} \in \mathcal{E}$  (measurement, evidence)
  - ▶ Probabilistic reasoning about  $\mathbf{x}$  then consists in computing

$$p(\mathbf{x}|\mathbf{y} \in \mathcal{E})$$

or related quantities like  $\operatorname{argmax}_{\mathbf{x}} p(\mathbf{x}|\mathbf{y} \in \mathcal{E})$  or posterior expectations of some function  $g$  of  $\mathbf{x}$ , e.g.

$$\mathbb{E}[g(\mathbf{x}) | \mathbf{y} \in \mathcal{E}] = \int g(\mathbf{u})p(\mathbf{u}|\mathbf{y} \in \mathcal{E})d\mathbf{u}$$

# Solution via product and sum rule (discrete-valued variables)

Assume that all variables are discrete valued, that  $\mathcal{E} = \{\mathbf{y}_o\}$ , and that we know  $p(\mathbf{x}, \mathbf{y}, \mathbf{z})$ . We would like to know  $p(\mathbf{x}|\mathbf{y}_o)$ .

- ▶ Product rule:  $p(\mathbf{x}|\mathbf{y}_o) = \frac{p(\mathbf{x}, \mathbf{y}_o)}{p(\mathbf{y}_o)}$
- ▶ Sum rule:  $p(\mathbf{x}, \mathbf{y}_o) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{y}_o, \mathbf{z})$
- ▶ Sum rule:  $p(\mathbf{y}_o) = \sum_{\mathbf{x}} p(\mathbf{x}, \mathbf{y}_o) = \sum_{\mathbf{x}, \mathbf{z}} p(\mathbf{x}, \mathbf{y}_o, \mathbf{z})$
- ▶ Result:

$$p(\mathbf{x}|\mathbf{y}_o) = \frac{\sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{y}_o, \mathbf{z})}{\sum_{\mathbf{x}, \mathbf{z}} p(\mathbf{x}, \mathbf{y}_o, \mathbf{z})}$$

# Roadmap for PMR

$$p(\mathbf{x}|\mathbf{y}_o) = \frac{\sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{y}_o, \mathbf{z})}{\sum_{\mathbf{x}, \mathbf{z}} p(\mathbf{x}, \mathbf{y}_o, \mathbf{z})}$$

Assume that  $\mathbf{x}, \mathbf{y}, \mathbf{z}$  each are  $d = 500$  dimensional, and that each element of the vectors can take  $K = 10$  values.

- ▶ **Issue 1:** To specify  $p(\mathbf{x}, \mathbf{y}, \mathbf{z})$ , we need to specify  $K^{3d} - 1 = 10^{1500} - 1$  non-negative numbers, which is impossible.

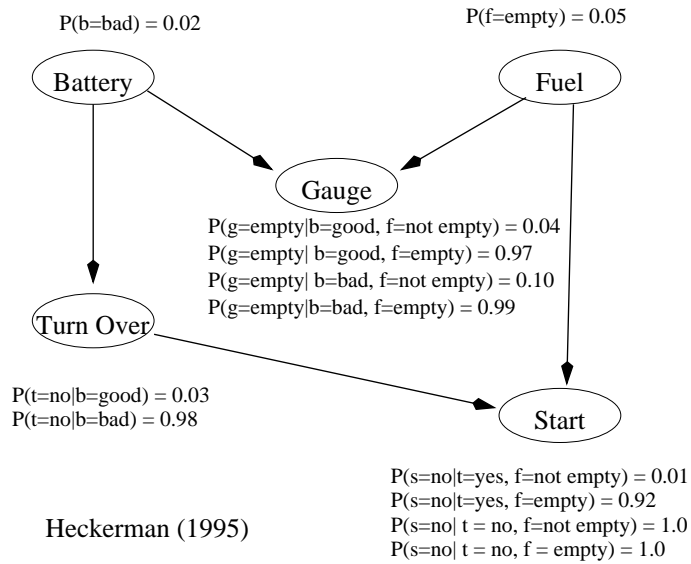
**Topic 1: Representation** What reasonably weak assumptions can we make to efficiently represent  $p(\mathbf{x}, \mathbf{y}, \mathbf{z})$ ?

# Roadmap for PMR

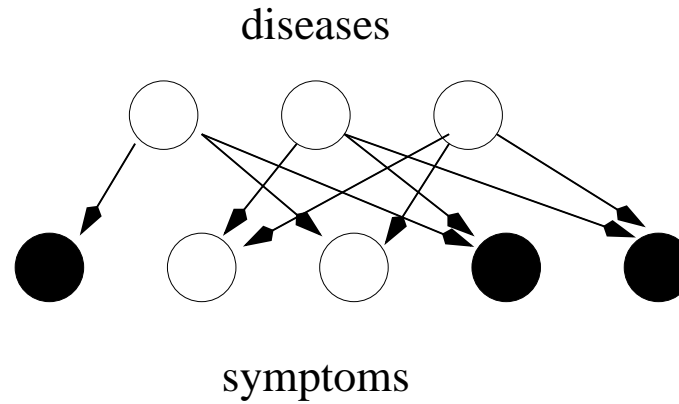
- ▶ **Issue 2:** The sum in the numerator goes over the order of  $K^d = 10^{500}$  non-negative numbers and the sum in the denominator over the order of  $K^{2d} = 10^{1000}$ , which is impossible to compute.

**Topic 2: Exact inference** Can we further exploit the assumptions on  $p(\mathbf{x}, \mathbf{y}, \mathbf{z})$  to efficiently compute the posterior probability or derived quantities?

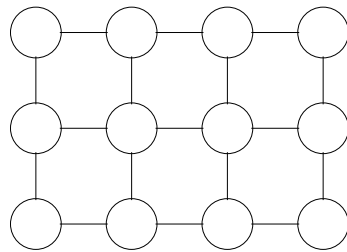




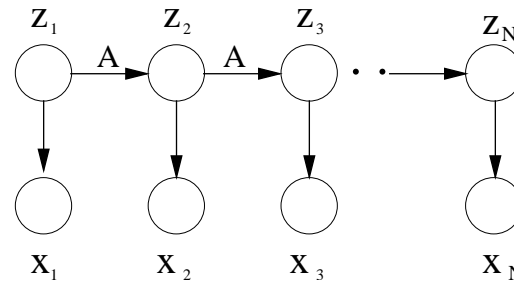
Directed graphical model



QMR-DT network



Ising model (statistical physics)  
(Undirected graphical model)



Hidden Markov model  
used for speech recognition etc.

# Roadmap for PMR

- ▶ **Issue 3:** Where do the non-negative numbers  $p(\mathbf{x}, \mathbf{y}, \mathbf{z})$  come from?

**Topic 3: Learning** How can we learn the numbers from data?

- ▶ **Issue 4:** For some models, exact inference and learning is too costly even after fully exploiting the assumptions made.

**Topic 4: Approximate inference and learning**

# Credits

These slides are modified from ones produced by Michael Gutmann, made available under Creative Commons licence CC BY 4.0.

©Michael Gutmann and Chris Williams, The University of Edinburgh 2018-2023 CC BY 4.0 .