Probabilistic Machine Learning
— Introduction —

Michael U. Gutmann

Probabilistic Machine Learning (INFR11298)
School of Informatics, The University of Edinburgh

Autumn Semester 2026

Course Administration

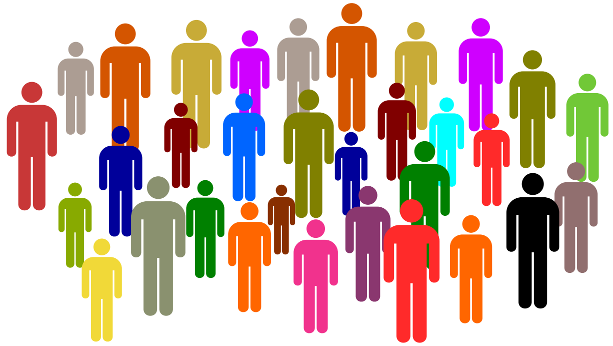
- Website: `https://opencourse.inf.ed.ac.uk/pml`
  - Course overview and key dates
  - Slides and tutorial materials
  - Optional materials
  - Information on what to expect from the course
- Via Learn (see link on course website):
  - Lecture recordings
  - Initial login to EdStem (Q&A forum)
  - Gradescope for three formative quizzes
- No coursework or assignment.
- EdStem forum to ask questions anonymously (towards your class mates).

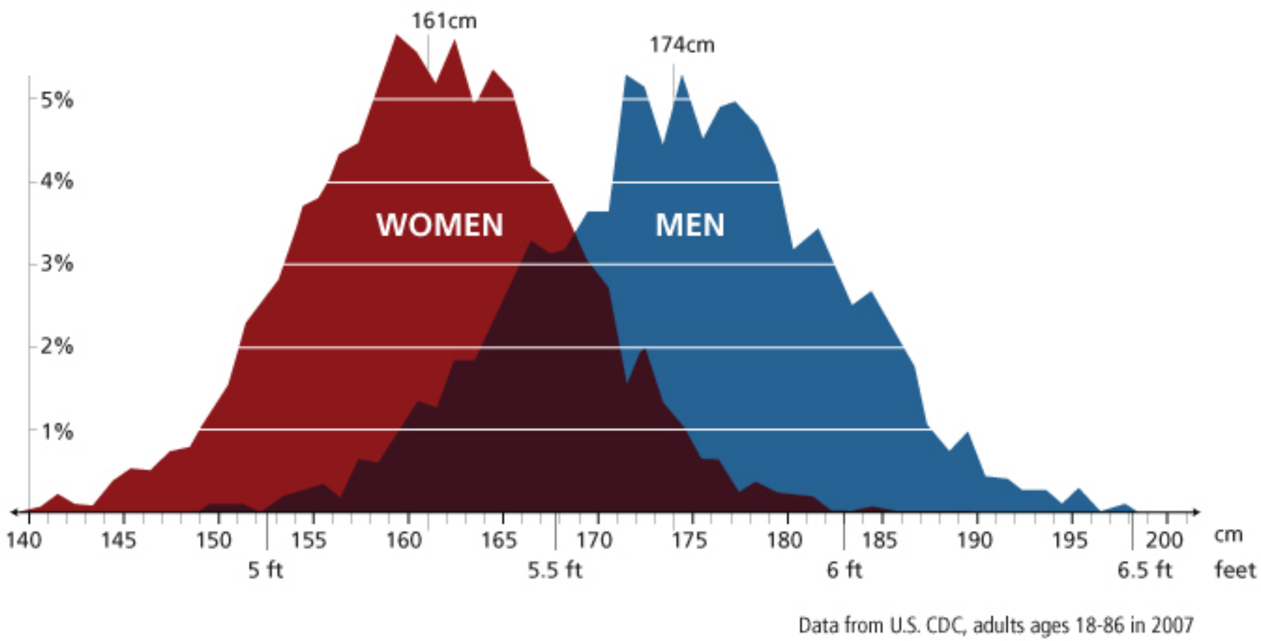
Course Administration

- Tutorial format:
  - 8 tutorials starting in week 3
  - Attempt to solve the exercises prior to the tutorials
  - It's ok to get stuck, but you need to have tried!
  - Tutorials focus on discussing solutions and difficulties encountered.
  - Detailed solutions will be provided.
- Time per week for independent study during the teaching weeks (rough guide):
  - Learning / getting the job done and not time spent is the most important thing.
  - Required time will depend on your background and learning techniques.
  - Lectures: 2h per lecture: 6h
  - Tutorials: 1h prep, 1h study of solutions: 2h
  - Self-study exercises: 2h
  - Total: about 10 hours on average

Variability

- Variability is part of nature
- Human heights vary
- Men are typically taller than women but height varies a lot





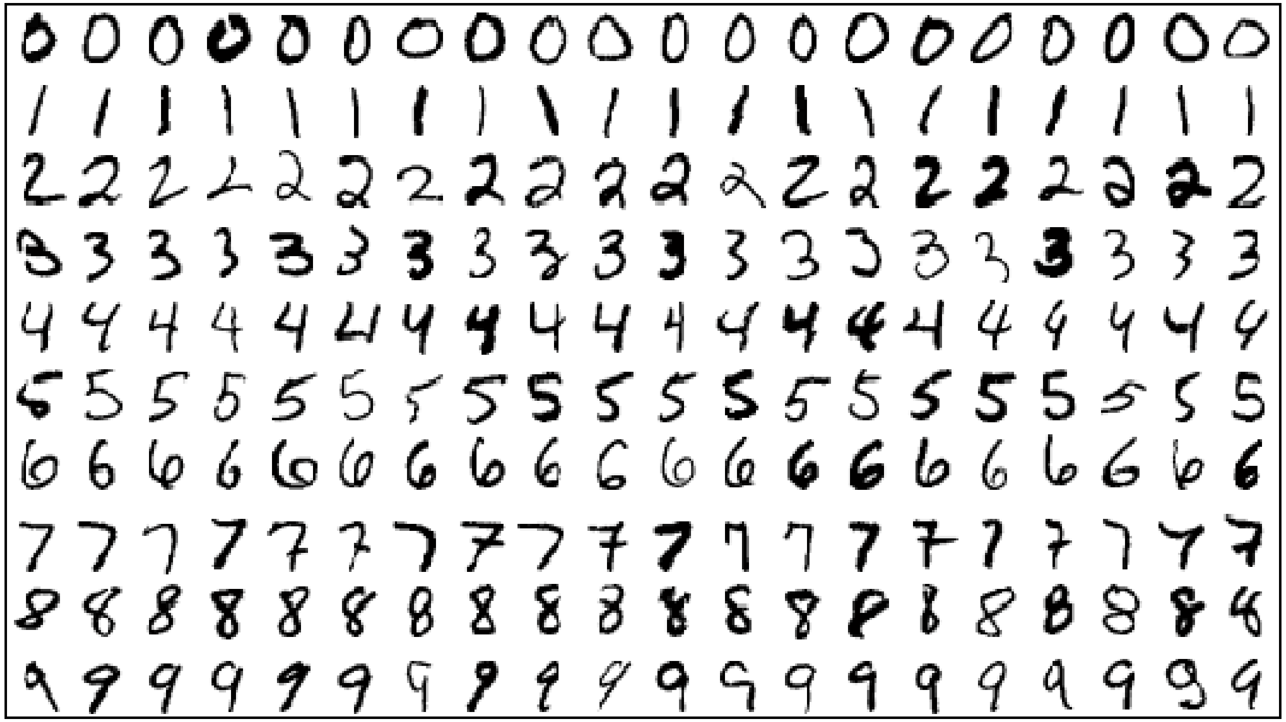
Variability

- Variability can be nice and desirable
- Picture your favourite ice-cream/gelato place!

(from https://italybest.com/10-best-gelateria-in-rome-italy/)

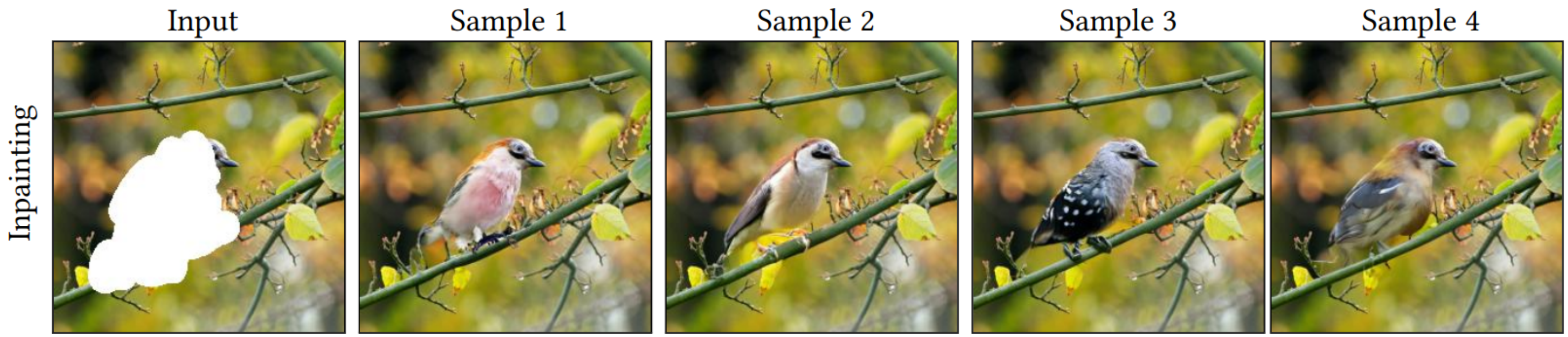
Variability

- Variability leads to uncertainty
- Which ice-cream/gelato flavour is best?
- Is it a 1 or a 7?



Need to capture uncertainty

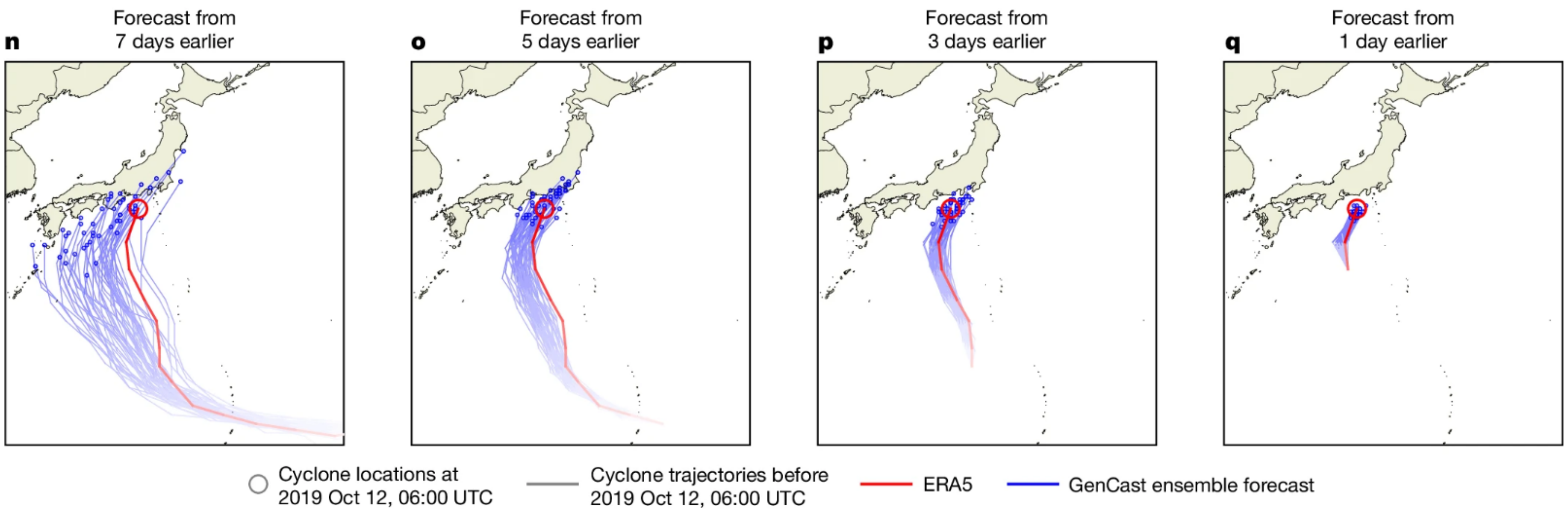
- Capturing uncertainty is important in many domains
- Example of image generation (inpainting)



(From Fig. 7 in Palette: Image-to-Image Diffusion Models, Saharia et al, *ACM Siggraph*, 2022)

Need to capture uncertainty

- Example of weather forecasting
- Uncertainty in the forecast is key for downstream decisions
  - Daily life: shall I wear the raincoat or not?
  - Renewable-energy planning: how much reserve should we allocate?
  - Flood warnings: do we close a road or not?
  - Disaster warnings: do we evacuate or not?



(From Fig. 2 in Probabilistic weather forecasting with machine learning, Price et al, *Nature*, 2025)

Probabilistic machine learning

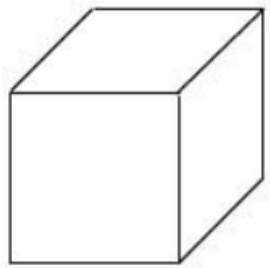
- Probabilistic machine learning embraces uncertainty.
- We use probabilities to describe our beliefs of how things usually behave and how much they vary.
  $\rightarrow$ Topic 1: Representing probabilistic models
- It allows us to compute probabilities for things that we have not or cannot observe from things that we can observe.
  $\rightarrow$ Topic 2: Exact inference
- We can combine beliefs with cost/utilities to choose actions that take uncertainty into account.
  $\rightarrow$ Topic 3: Actions and decision making
- It provides methods to learn/update beliefs from data.
  $\rightarrow$ Topic 4: Learning
- It shows us that tracking all eventualities is often impossible.
  $\rightarrow$ Topic 5: Approximate inference and learning

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:



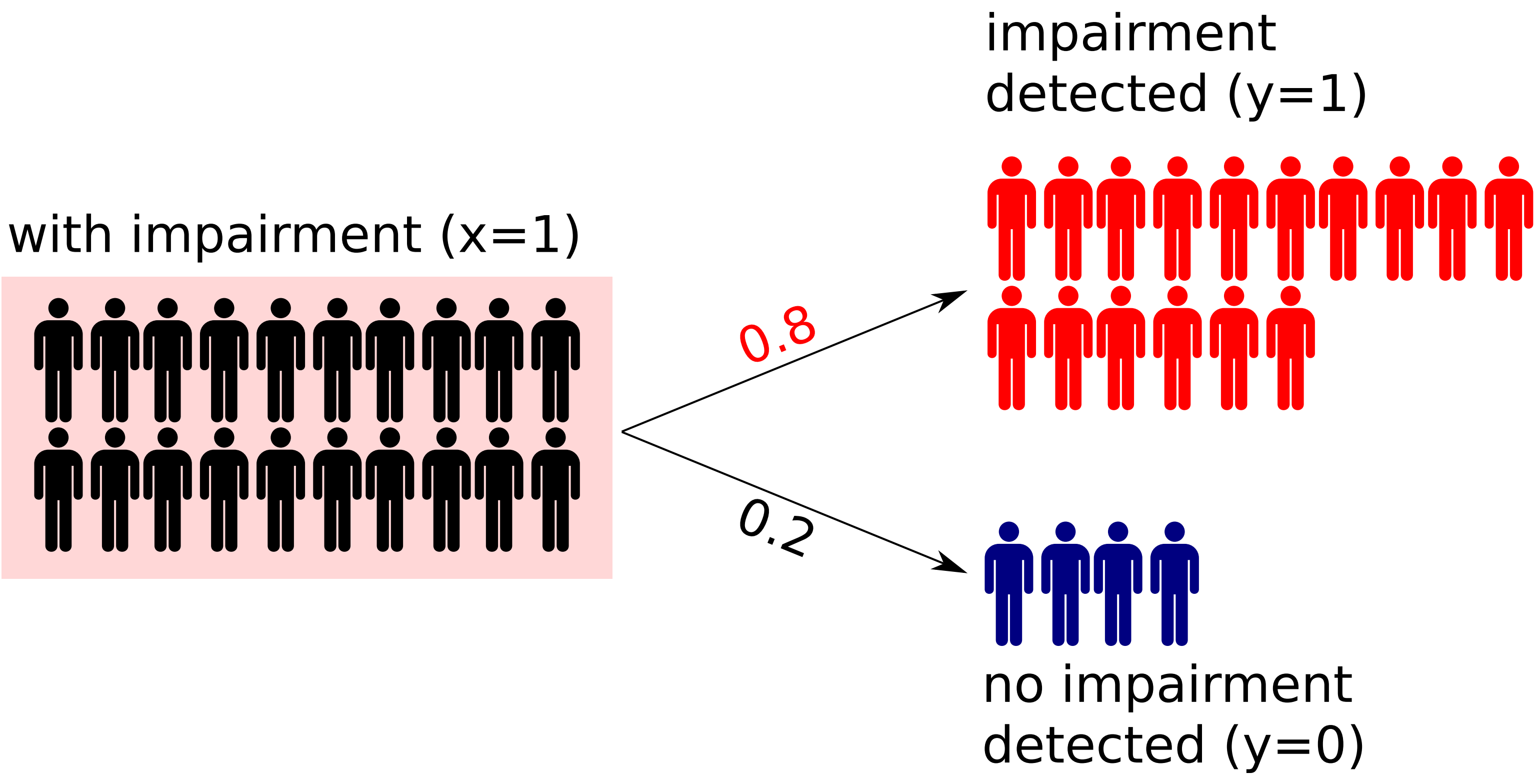
8. Drawing test

- 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)

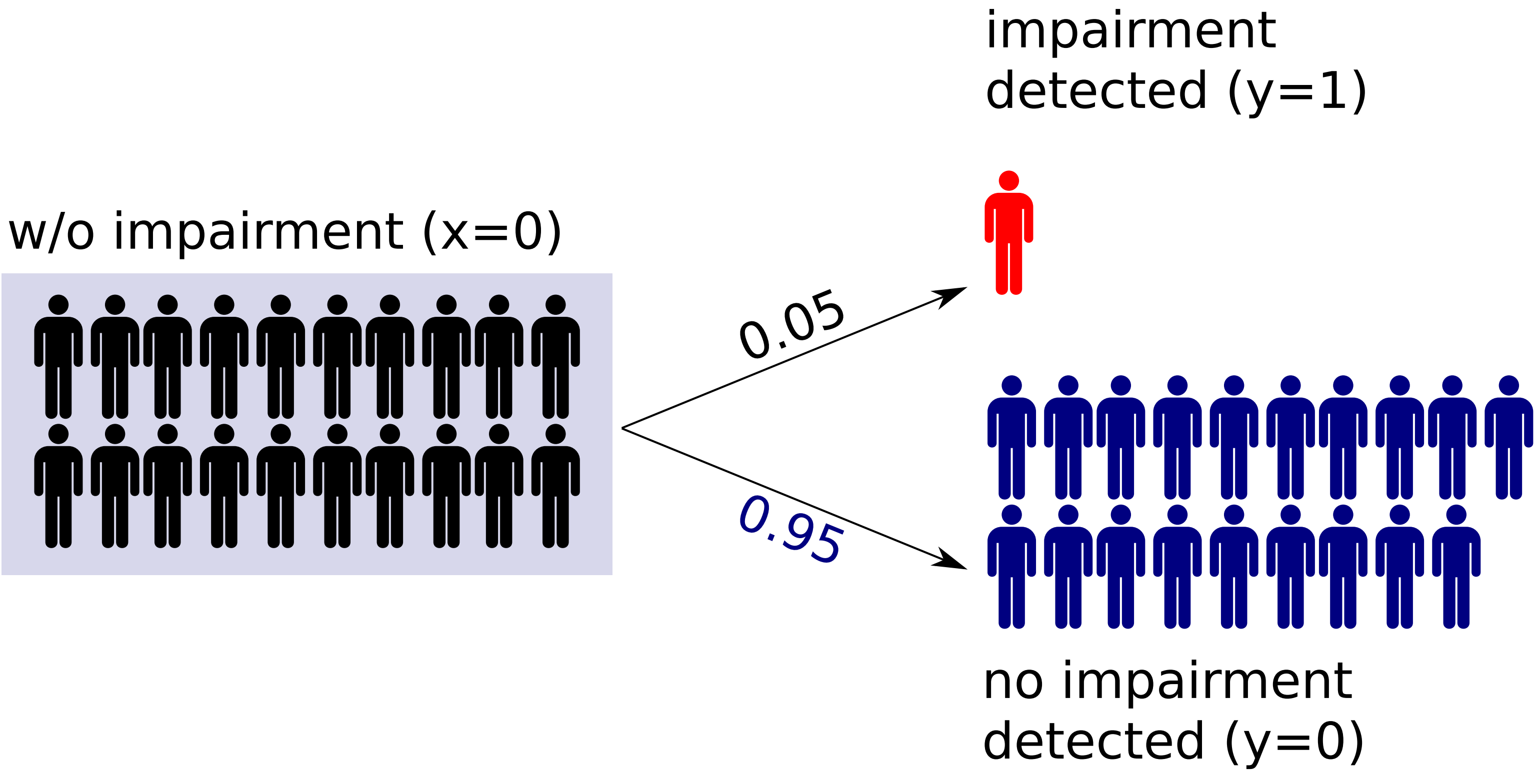
Accuracy of the test

- Sensitivity of 0.8 and specificity of 0.95 (Scharre, 2010)
- Values were determined from data.
- 80% correct for people with impairment



Accuracy of the test

- Sensitivity of 0.8 and specificity of 0.95 (Scharre, 2010)
- Values were determined from data.
- 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

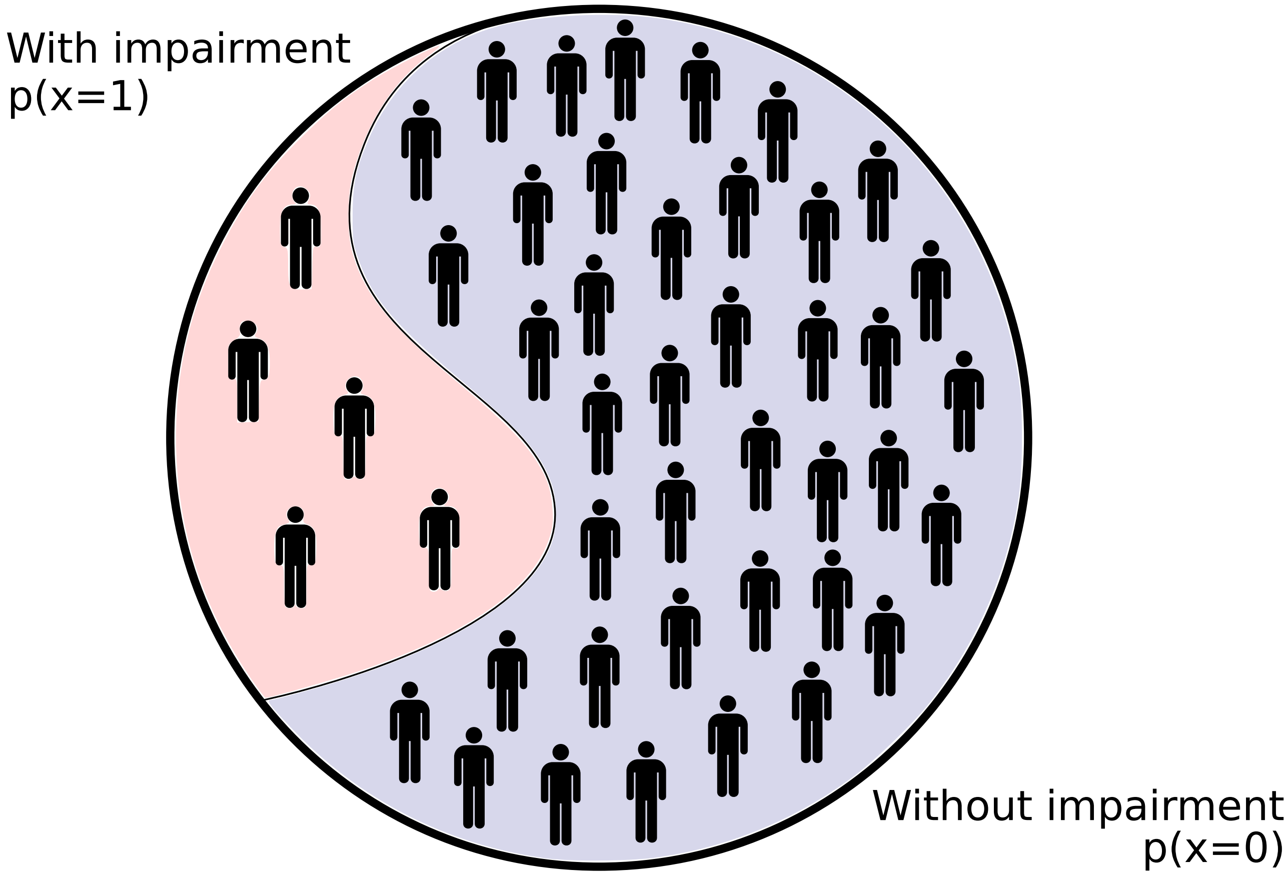
- The test outcomes y can be described with probabilities

$$
\begin{aligned}
\text{sensitivity} = 0.8 \quad &\Leftrightarrow \quad p(y = 1|x = 1) = 0.8 \\
&\Leftrightarrow \quad p(y = 0|x = 1) = 0.2 \\
\text{specificity} = 0.95 \quad &\Leftrightarrow \quad p(y = 0|x = 0) = 0.95 \\
&\Leftrightarrow \quad p(y = 1|x = 0) = 0.05
\end{aligned}
$$

- $p(y|x)$: 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, $p(x=1) = 5/45 \approx 11\%$ have a cognitive impairment (plausible range: 3% – 22%, Geda, 2014)

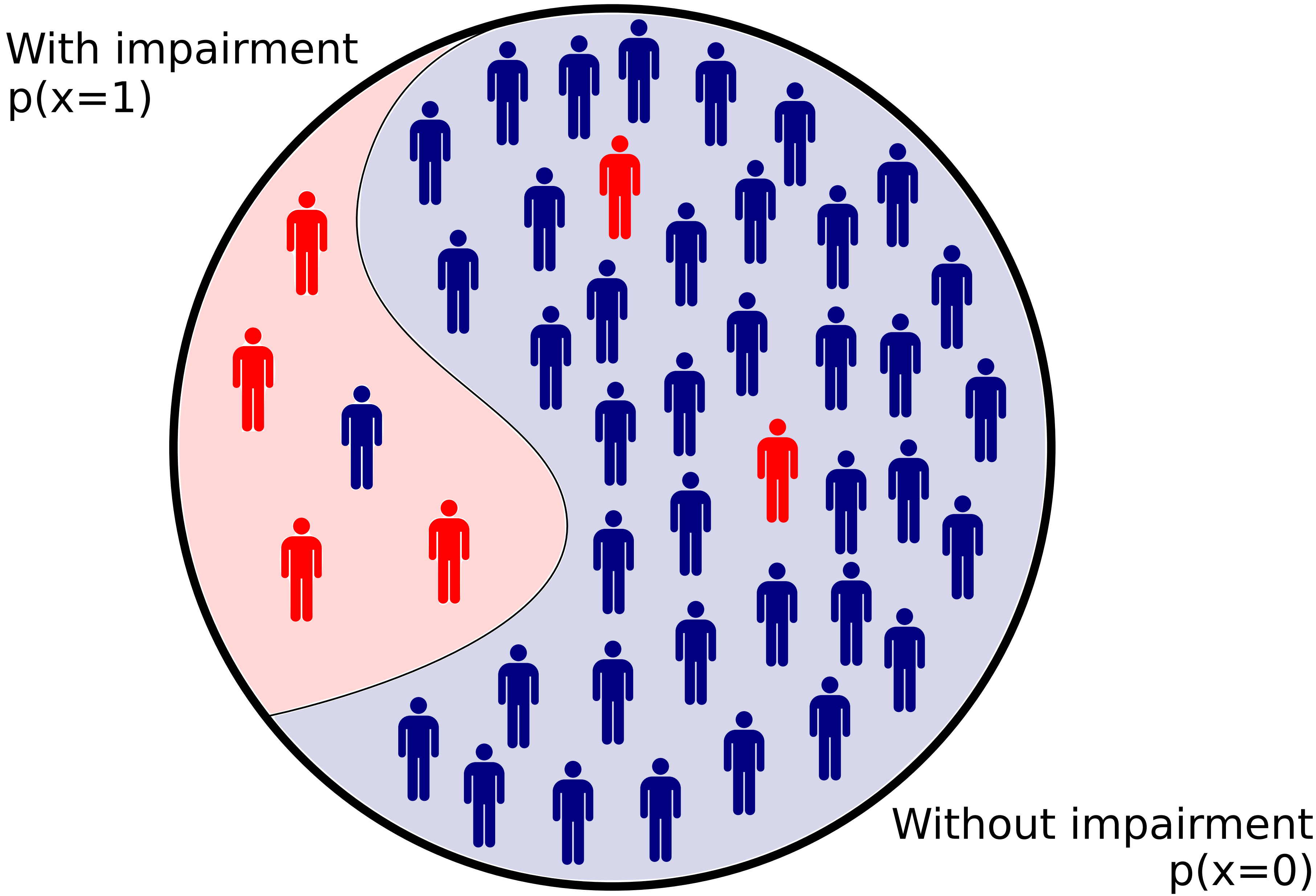


Probabilistic model

- Reality:
  - properties/characteristics of the group of people like the patient
  - properties/characteristics of the test
- Probabilistic model:
  - $p(x = 1)$
  - $p(y = 1|x = 1)$ or $p(y = 0|x = 1)$
    $p(y = 1|x = 0)$ or $p(y = 0|x = 0)$

  Fully specified by three numbers.
- A probabilistic model is an abstraction of reality that uses probability theory to quantify the chance of uncertain events.
- We will often learn probabilistic models from data, but some aspects may also be specified by hand.

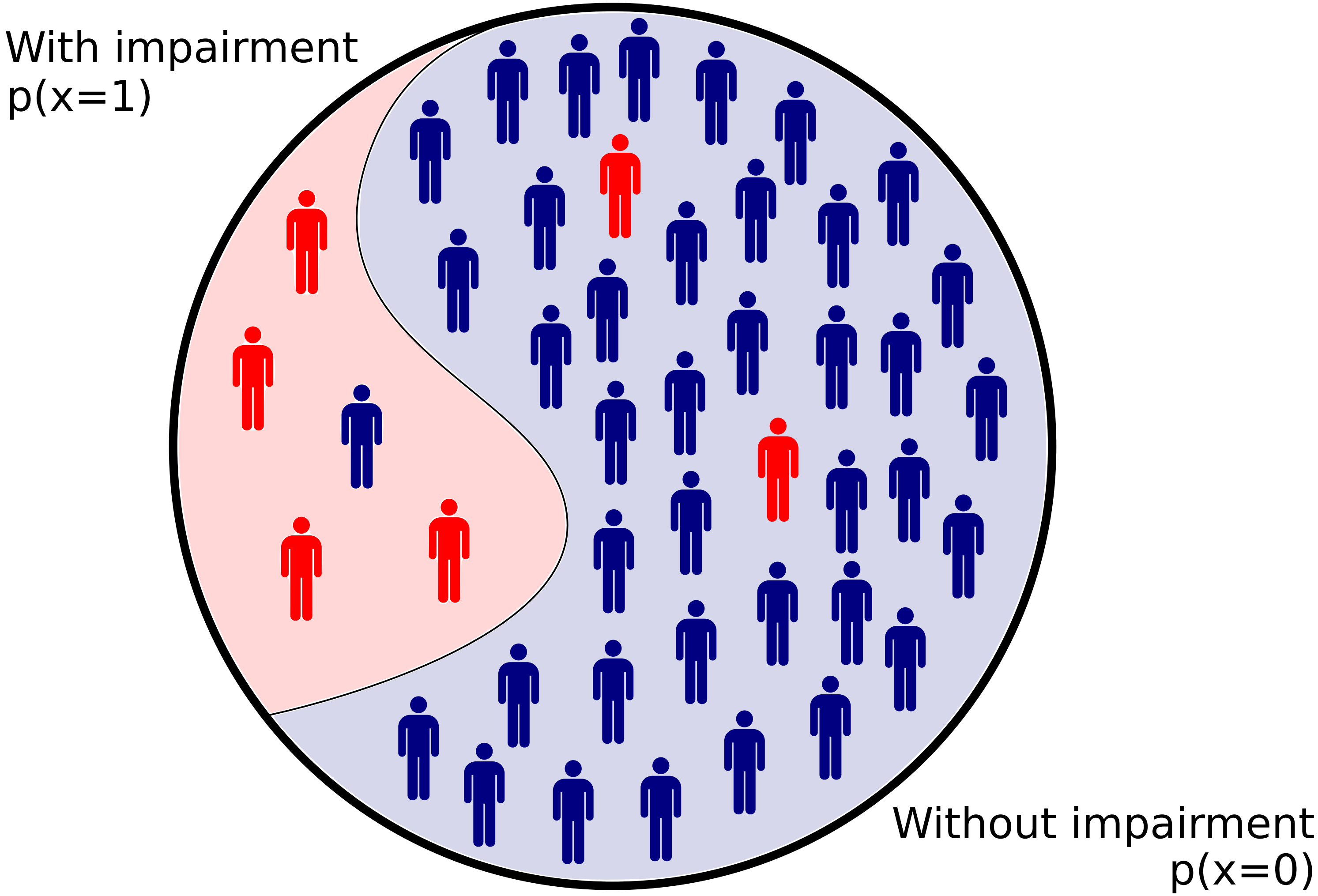
If we tested the whole population



If we tested the whole population

Fraction of people who are impaired and have positive tests:

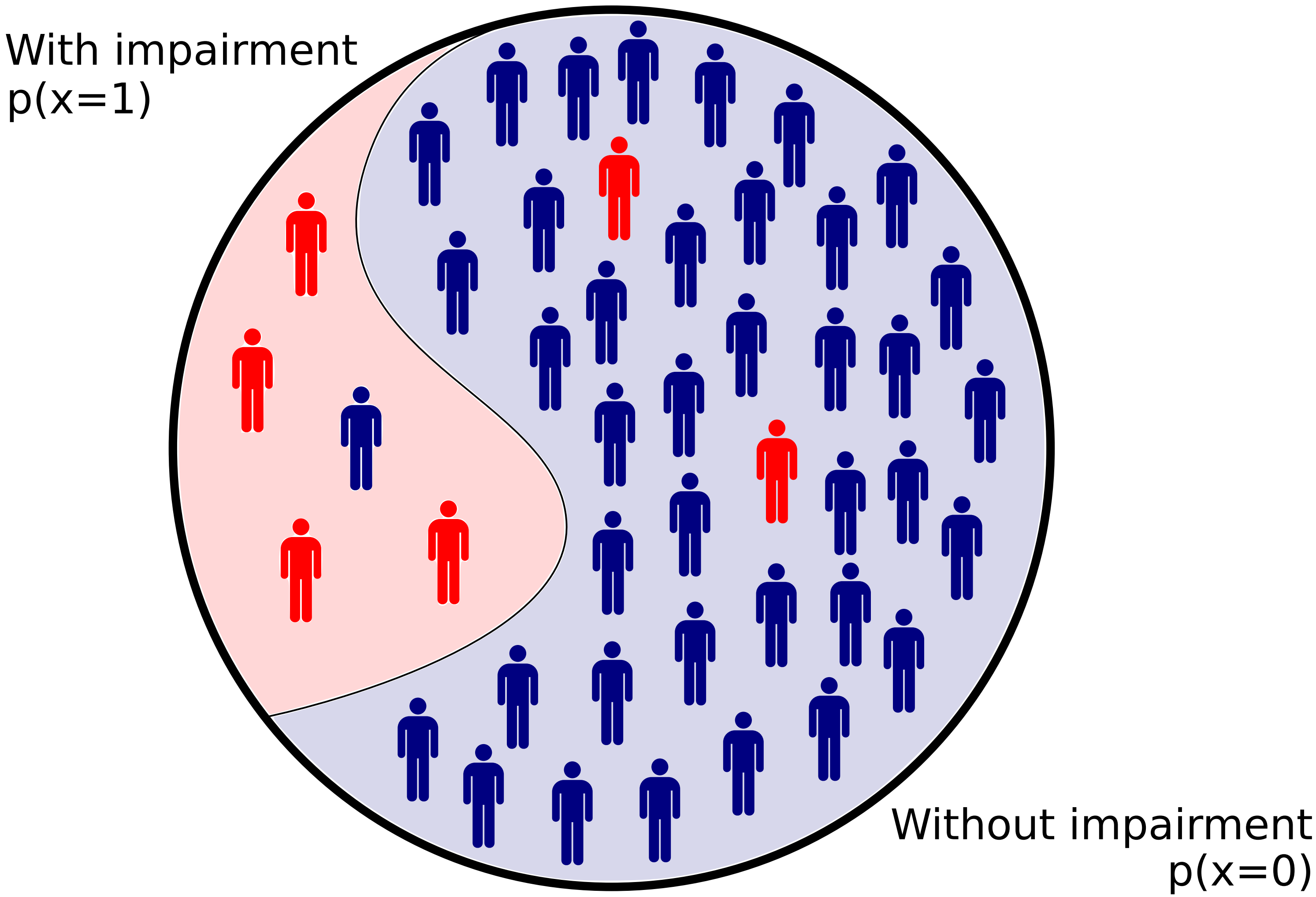
$$p(x=1, y=1) = p(y=1|x=1)p(x=1) = 4/45 \qquad \text{(product rule)}$$



If we tested the whole population

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

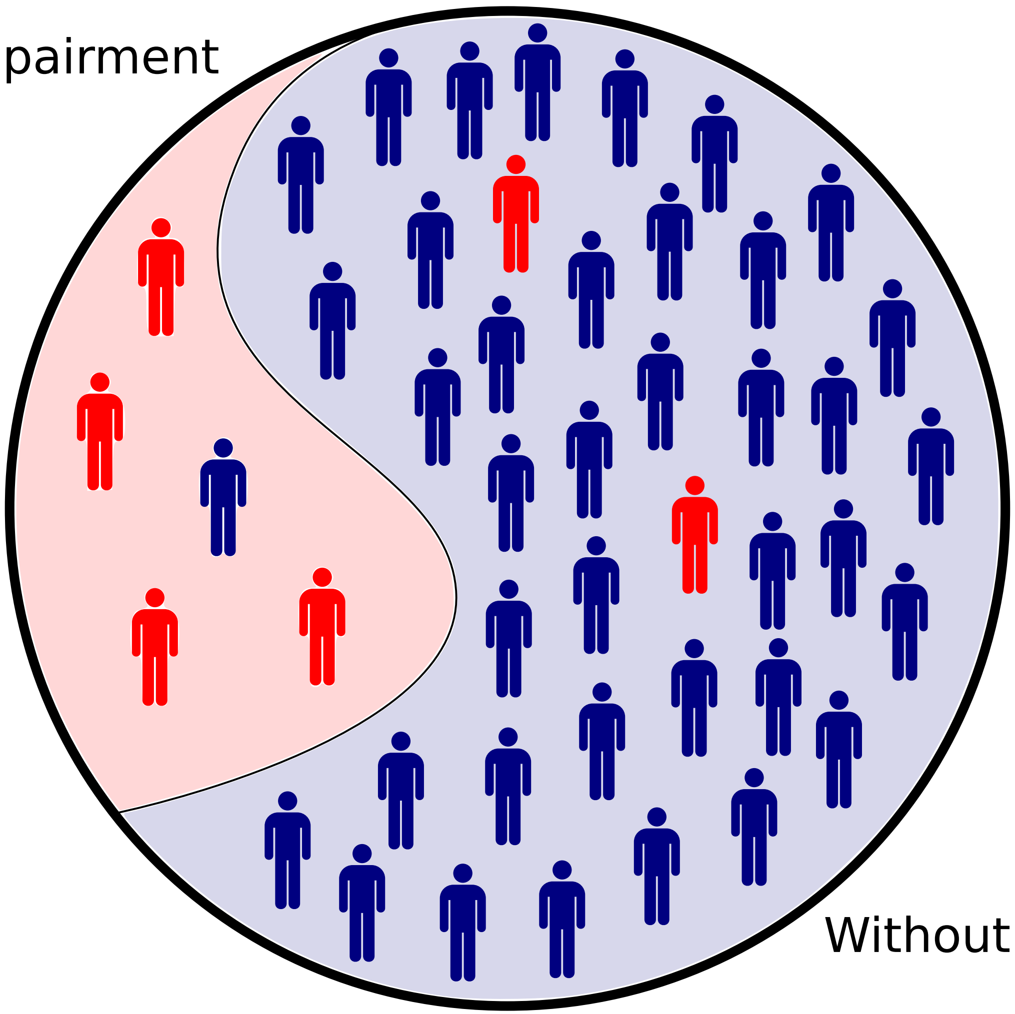
$$p(x=0, y=1) = p(y=1|x=0)p(x=0) = 2/45 \qquad \text{(product rule)}$$



If we tested the whole population

Fraction of people where the test is positive:

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



Putting everything together

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

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

- Fraction without impairment:

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

- Equations are examples of "Bayes' rule".
- Converts "prior" probabilities $p(x)$ into "posterior" probabilities $p(x|y = 1)$.
- Positive test increased probability of cognitive impairment from 11% (prior belief) to 67%, or from 6% to 51%.
- Given the probabilities, should the 70 year old man be concerned about the positive test? What should he best do?

Relating the example to the five PML topics

- The real-world test was represented in terms of a probabilistic model $\rightarrow$ topic 1.
- We used probabilistic inference to compute the probability of an event that we couldn't observe (the cognitive impairment $x = 1$) from an event that we could observe (the test result $y = 1$) $\rightarrow$ topic 2.
- We asked what the man should do given the probabilities $\rightarrow$ converting probabilities to actions will be part of topic 3.
- The numbers specifying the probabilistic model were learned from data $\rightarrow$ topic 4
- While the computations were tractable here, they become easily intractable and approximations will be neeed $\rightarrow$ topic 5.

Key rules of probability

(1) Product rule:

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

(2) Sum rule (marginalisation):

$$p(y=1) = p(x=1, y=1) + p(x=0, y=1)$$

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

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

Denominator from sum rule, or sum rule and product rule

$$p(y=1) = p(y=1|x=1)p(x=1) + p(y=1|x=0)p(x=0)$$

Key rules of 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 (marginalisation):

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

- Notation $\sum_{\mathbf{x}}$ means "sum over all values of $\mathbf{x}$".

Inference via the product and sum rule

Assume that all variables are discrete valued and that we know $p(\mathbf{x}, \mathbf{y}, \mathbf{z})$. Having observed $\mathbf{y} = \mathbf{y}_o$, 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:
$$p(\mathbf{x}|\mathbf{y}_o) = \frac{\sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{y}_o, \mathbf{z})}{\sum_{\mathbf{x},\mathbf{z}} p(\mathbf{x}, \mathbf{y}_o, \mathbf{z})}$$
- In case of continuous random variables, replace sum with integrals.
- Formally solves the inference task.

Roadmap for PML

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

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 PML

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

- 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?

- Issue 3: Thank you for the numbers. But what shall I best do?

  Topic 3: Actions and decision making How to predict the outcome of actions and choose optimal actions?

Roadmap for PML

- Issue 4: Where do the non-negative numbers $p(\mathbf{x}, \mathbf{y}, \mathbf{z})$ come from?

  Topic 4: Learning How can we learn the numbers from data?

- Issue 5: For some models, exact inference and learning is too costly even after fully exploiting the assumptions made.

  Topic 5: Approximate inference and learning