

# Missing Data

- You cannot just ignore missing data unless you know it is missing at random.
- Is the reason data is missing correlated with something of interest?

- You cannot just ignore missing data unless you know it is missing at random.
- Is the reason data is missing correlated with something of interest?
- **Example:** data in a clinical trial to test a drug may be missing because:

- You cannot just ignore missing data unless you know it is missing at random.
- Is the reason data is missing correlated with something of interest?
- **Example:** data in a clinical trial to test a drug may be missing because:
  - ▶ the patient dies

- You cannot just ignore missing data unless you know it is missing at random.
- Is the reason data is missing correlated with something of interest?
- **Example:** data in a clinical trial to test a drug may be missing because:
  - ▶ the patient dies
  - ▶ the patient had severe side effects

- You cannot just ignore missing data unless you know it is missing at random.
- Is the reason data is missing correlated with something of interest?
- **Example:** data in a clinical trial to test a drug may be missing because:
  - ▶ the patient dies
  - ▶ the patient had severe side effects
  - ▶ the patient was cured

- You cannot just ignore missing data unless you know it is missing at random.
- Is the reason data is missing correlated with something of interest?
- **Example:** data in a clinical trial to test a drug may be missing because:
  - ▶ the patient dies
  - ▶ the patient had severe side effects
  - ▶ the patient was cured
  - ▶ the patient had to visit a sick relative.

- You cannot just ignore missing data unless you know it is missing at random.
- Is the reason data is missing correlated with something of interest?
- **Example:** data in a clinical trial to test a drug may be missing because:
  - ▶ the patient dies
  - ▶ the patient had severe side effects
  - ▶ the patient was cured
  - ▶ the patient had to visit a sick relative.

— ignoring some of these may make the drug look better or worse than it is.

- You cannot just ignore missing data unless you know it is missing at random.
  - Is the reason data is missing correlated with something of interest?
  - **Example:** data in a clinical trial to test a drug may be missing because:
    - ▶ the patient dies
    - ▶ the patient had severe side effects
    - ▶ the patient was cured
    - ▶ the patient had to visit a sick relative.
- ignoring some of these may make the drug look better or worse than it is.
- In general you need to model why data is missing.



# Missing Data

- Suppose there is a drug claimed to treat a disease.
- The drug does not actually affect the disease or its symptom, but makes sick people sicker.
- Suppose patients were randomly assigned the drug or a placebo,

# Missing Data

- Suppose there is a drug claimed to treat a disease.
- The drug does not actually affect the disease or its symptom, but makes sick people sicker.
- Suppose patients were randomly assigned the drug or a placebo, but the sickest people dropped out of the study, because they become too sick to participate.
- What happens if the missing data (from patients who dropped out) is ignored?

# Missing Data

- Suppose there is a drug claimed to treat a disease.
- The drug does not actually affect the disease or its symptom, but makes sick people sicker.
- Suppose patients were randomly assigned the drug or a placebo, but the sickest people dropped out of the study, because they become too sick to participate.
- What happens if the missing data (from patients who dropped out) is ignored?
- It looks like the treatment works; there are fewer sick people among the people who took the treatment and remained in the study!

Handling missing data requires more than a probabilistic model that models correlation. It requires a causal model of how the data is missing.

# Missingness graph

A **missingness graph**, or ***m*-graph**, is a causal model of data where some values might be missing.

- Given a causal network, the *m*-graph contains all variables of the original graph with the same parents.

# Missingness graph

A **missingness graph**, or ***m*-graph**, is a causal model of data where some values might be missing.

- Given a causal network, the *m*-graph contains all variables of the original graph with the same parents.
- For each variable  $V$  that could be observed with some values missing, the *m*-graph contains two extra variables:

# Missingness graph

A **missingness graph**, or ***m*-graph**, is a causal model of data where some values might be missing.

- Given a causal network, the *m*-graph contains all variables of the original graph with the same parents.
- For each variable  $V$  that could be observed with some values missing, the *m*-graph contains two extra variables:
  - ▶ Boolean variable  $M_V$  is true when  $V$ 's value is missing.  
The parents of  $M_V$  are the variables missingness depends on.

# Missingness graph

A **missingness graph**, or ***m*-graph**, is a causal model of data where some values might be missing.

- Given a causal network, the *m*-graph contains all variables of the original graph with the same parents.
- For each variable  $V$  that could be observed with some values missing, the *m*-graph contains two extra variables:
  - ▶ Boolean variable  $M_V$  is true when  $V$ 's value is missing. The parents of  $M_V$  are the variables missingness depends on.
  - ▶ Variable  $V^*$ , with domain  $dom(V) \cup \{missing\}$ .  $missing$  is a new value (not in the domain of  $V$ )  
 $V$  and  $M_V$  and the parents of  $V^*$ , with:

$$P(V^*=missing \mid M_V=true) = 1$$

$$P(V^*=v \mid M_V=false \wedge V=v) = 1.$$

# Missingness graph

A **missingness graph**, or ***m*-graph**, is a causal model of data where some values might be missing.

- Given a causal network, the *m*-graph contains all variables of the original graph with the same parents.
- For each variable  $V$  that could be observed with some values missing, the *m*-graph contains two extra variables:
  - ▶ Boolean variable  $M_V$  is true when  $V$ 's value is missing. The parents of  $M_V$  are the variables missingness depends on.
  - ▶ Variable  $V^*$ , with domain  $dom(V) \cup \{missing\}$ .  $missing$  is a new value (not in the domain of  $V$ )  
 $V$  and  $M_V$  and the parents of  $V^*$ , with:

$$P(V^*=missing \mid M_V=true) = 1$$

$$P(V^*=v \mid M_V=false \wedge V=v) = 1.$$

- If  $V$  is observed to be  $v$ ,  $V^*=v$  is conditioned on.  
If the value for  $V$  is missing,  $V^*=missing$  is conditioned on.



# Missingness graph

A **missingness graph**, or ***m*-graph**, is a causal model of data where some values might be missing.

- Given a causal network, the *m*-graph contains all variables of the original graph with the same parents.
- For each variable  $V$  that could be observed with some values missing, the *m*-graph contains two extra variables:
  - ▶ Boolean variable  $M_V$  is true when  $V$ 's value is missing. The parents of  $M_V$  are the variables missingness depends on.
  - ▶ Variable  $V^*$ , with domain  $dom(V) \cup \{missing\}$ .  $missing$  is a new value (not in the domain of  $V$ )  
 $V$  and  $M_V$  and the parents of  $V^*$ , with:

$$P(V^*=missing \mid M_V=true) = 1$$

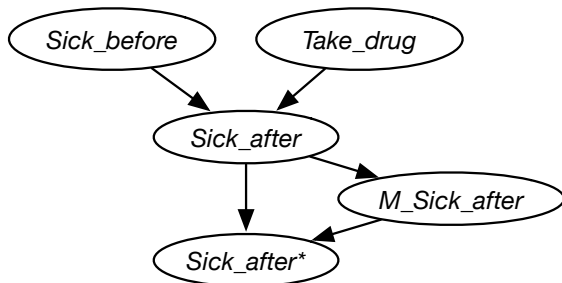
$$P(V^*=v \mid M_V=false \wedge V=v) = 1.$$

- If  $V$  is observed to be  $v$ ,  $V^*=v$  is conditioned on.  
If the value for  $V$  is missing,  $V^*=missing$  is conditioned on.
- $V^*$  is always observed.

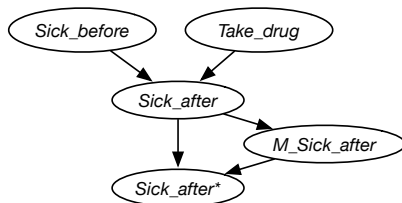
## Example $m$ -graph

A drug that just makes people sicker and so drop out, giving missing data.

Missingness depends on whether they are sick after:

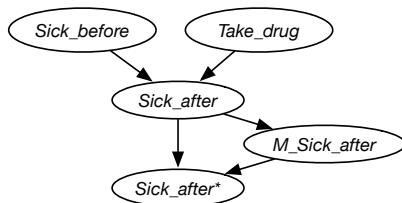


# Training with Expectation Maximization



This could be trained using expectation maximization (EM) with *Sick\_after* unobserved, *however*.

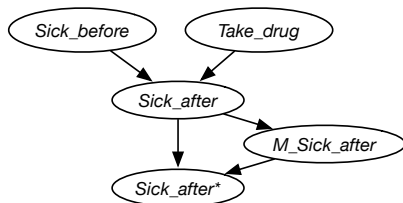
# Training with Expectation Maximization



This could be trained using expectation maximization (EM) with *Sick\_after* unobserved, *however*:

- There are many distributions consistent with the data:
  - all of the unobserved could be very sick after
  - none could be sick after taking the drug.

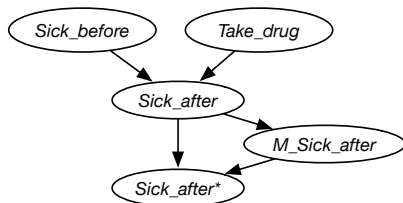
# Training with Expectation Maximization



This could be trained using expectation maximization (EM) with *Sick\_after* unobserved, *however*:

- There are many distributions consistent with the data: all of the unobserved could be very sick after none could be sick after taking the drug.
- EM could converge to any of these.

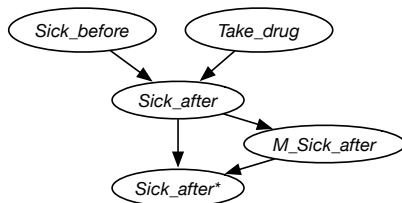
# Training with Expectation Maximization



This could be trained using expectation maximization (EM) with *Sick\_after* unobserved, *however*:

- There are many distributions consistent with the data: all of the unobserved could be very sick after none could be sick after taking the drug.
- EM could converge to any of these.
- EM makes up fiction about those with missing data.

# Training with Expectation Maximization



This could be trained using expectation maximization (EM) with *Sick\_after* unobserved, *however*:

- There are many distributions consistent with the data: all of the unobserved could be very sick after none could be sick after taking the drug.
- EM could converge to any of these.
- EM makes up fiction about those with missing data.
- We need to determine why the data is missing.

- A distribution is **recoverable** or **identifiable** from missing data if the distribution can be accurately measured from the data, even with parts of the data missing.



- A distribution is **recoverable** or **identifiable** from missing data if the distribution can be accurately measured from the data, even with parts of the data missing.
- Data for  $V$  is **missing completely at random (MCAR)** if  $V$  and  $M_V$  are independent. Missing data can be ignored.
- Variable  $Y$  is **missing at random (MAR)**, when  $Y$  is independent of  $M_Y$  given observed variables  $V_o$ .

$$P(Y | V_o, M_Y) = P(Y | V_o)$$

- A distribution is **recoverable** or **identifiable** from missing data if the distribution can be accurately measured from the data, even with parts of the data missing.
- Data for  $V$  is **missing completely at random (MCAR)** if  $V$  and  $M_V$  are independent. Missing data can be ignored.
- Variable  $Y$  is **missing at random (MAR)**, when  $Y$  is independent of  $M_Y$  given observed variables  $V_o$ .

$$P(Y | V_o, M_Y) = P(Y | V_o)$$

$$\text{then } P(Y, V_o) = P(Y | V_o, M_Y = \text{false})P(V_o)$$

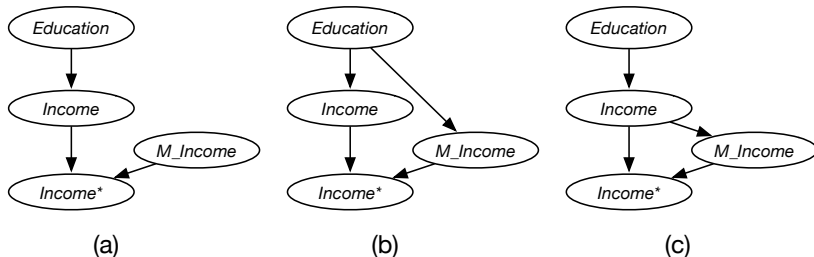
- A distribution is **recoverable** or **identifiable** from missing data if the distribution can be accurately measured from the data, even with parts of the data missing.
- Data for  $V$  is **missing completely at random (MCAR)** if  $V$  and  $M_V$  are independent. Missing data can be ignored.
- Variable  $Y$  is **missing at random (MAR)**, when  $Y$  is independent of  $M_Y$  given observed variables  $V_o$ .

$$P(Y | V_o, M_Y) = P(Y | V_o)$$

then  $P(Y, V_o) = P(Y | V_o, M_Y=false)P(V_o)$

- In other cases (e.g., previous case) the distribution may not be recoverable, depending on the graph structure.

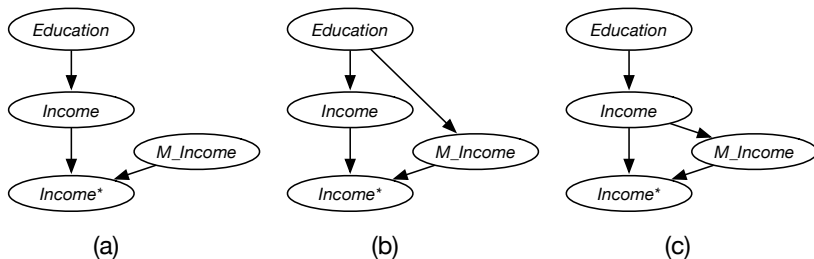
# Recoverability



*Education* is observed but *Income* might have missing values:

- (a) completely at random
- (b) missing at random
- (c) missing not at random

# Recoverability

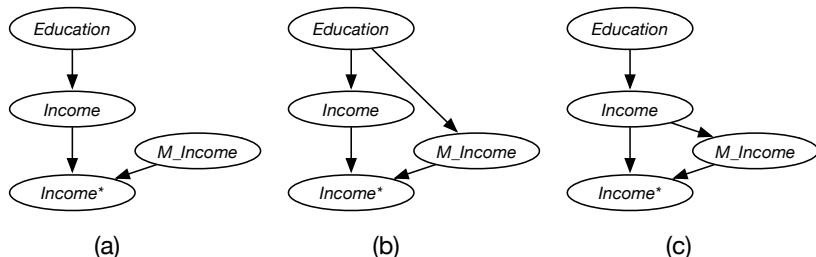


*Education* is observed but *Income* might have missing values:

(a) completely at random

$$\begin{aligned} P(\text{Income}, \text{Education}) \\ = P(\text{Income}^*, \text{Education} \mid M\_Income = \text{false}) \end{aligned}$$

# Recoverability



*Education* is observed but *Income* might have missing values:

(b) missing at random

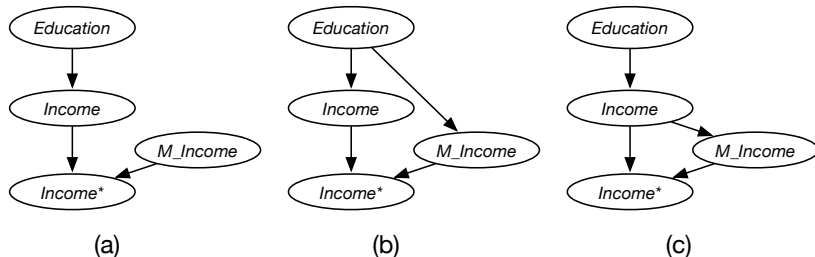
$$P(\text{Income}, \text{Education})$$

$$= P(\text{Income} \mid \text{Education}) * P(\text{Education})$$

$$= P(\text{Income} \mid \text{Education} \wedge M\_Income = \text{false}) * P(\text{Education})$$

$$= P(\text{Income}^* \mid \text{Education} \wedge M\_Income = \text{false}) * P(\text{Education})$$

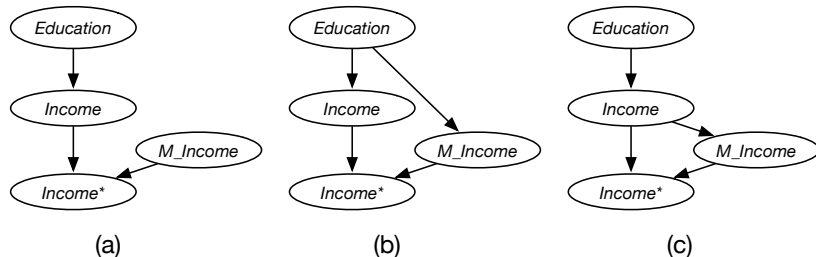
# Recoverability



*Education* is observed but *Income* might have missing values:

(c) missing not at random (MNAR).

- ▶ In this graph, the relationship between income and education cannot be estimated from data.

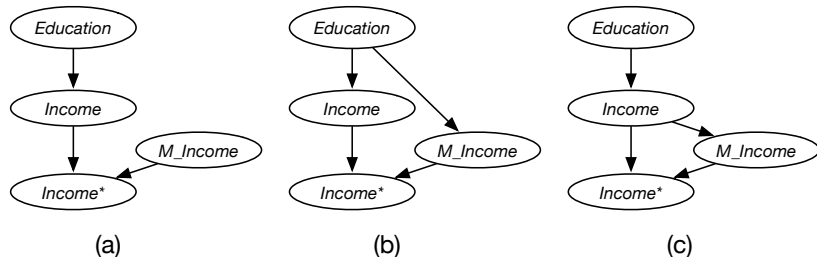


*Education* is observed but *Income* might have missing values:

(c) missing not at random (MNAR).

- ▶ In this graph, the relationship between income and education cannot be estimated from data.
- ▶ EM (and related algorithms) converge to fiction.





*Education* is observed but *Income* might have missing values:

(c) missing not at random (MNAR).

- ▶ In this graph, the relationship between income and education cannot be estimated from data.
- ▶ EM (and related algorithms) converge to fiction.
- ▶ In some cases of MNAR, probabilities can be computed, depending on the graph structure.

