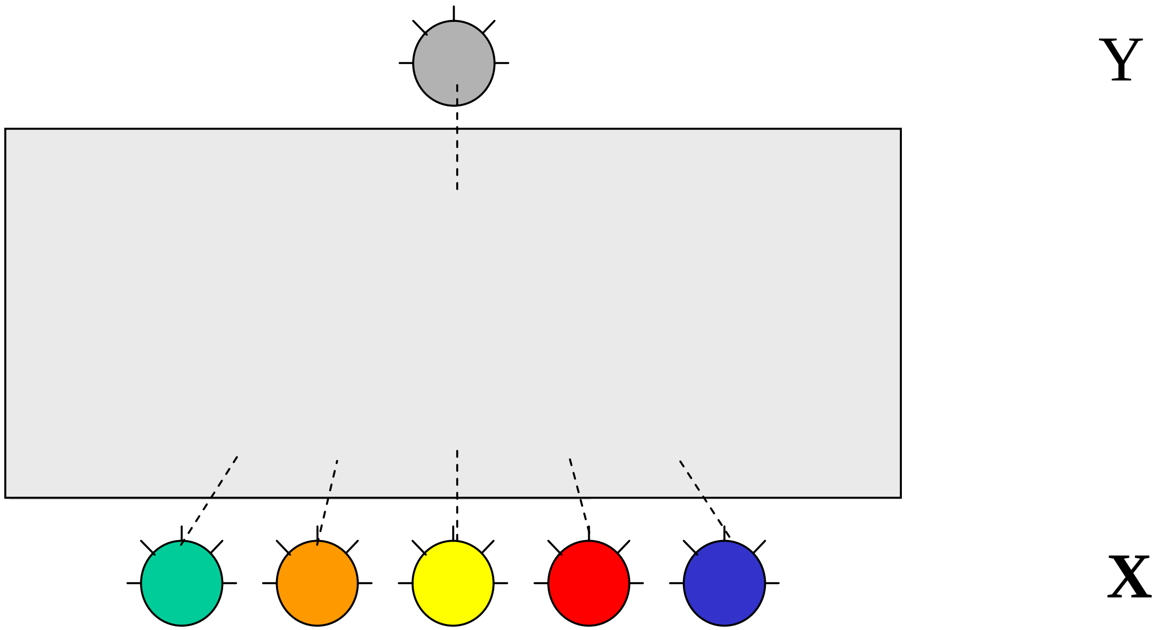
Lecture 5: Causality and Feature Selection

Isabelle Guyon

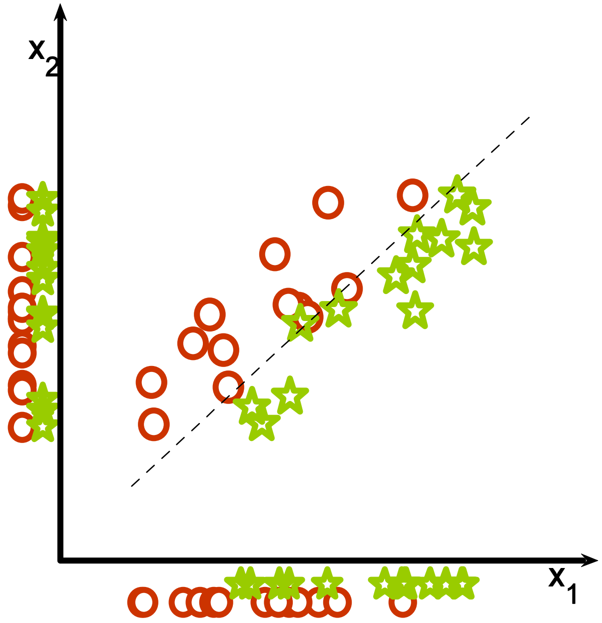
isabelle@clopinet.com

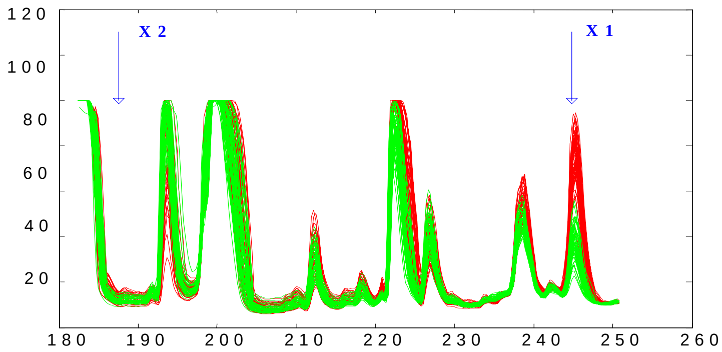
Variable/feature selection



Remove features X_i to improve (or least degrade) prediction of Y.

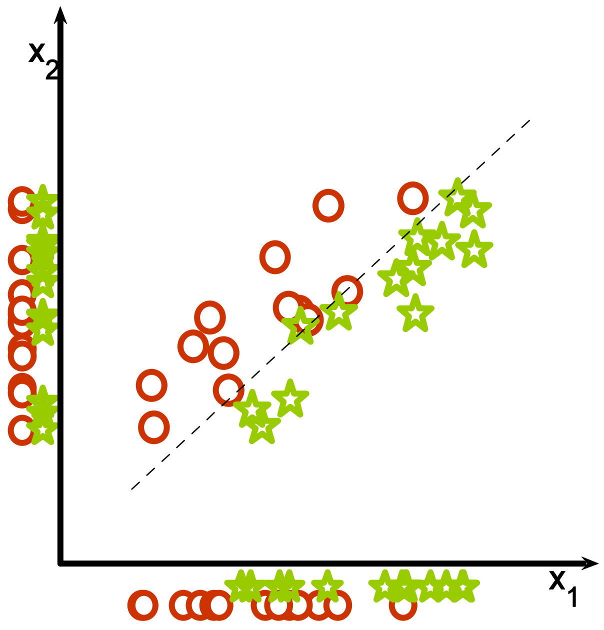
What can go wrong?

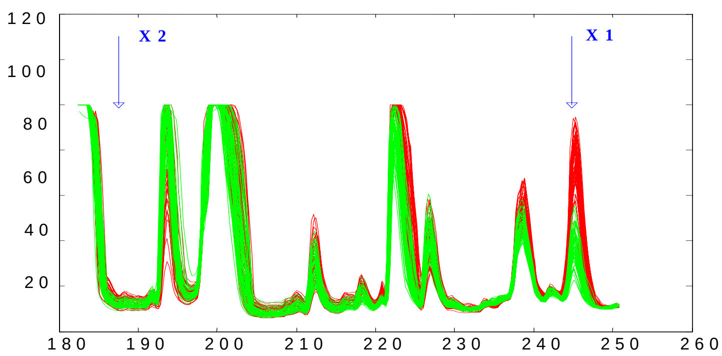


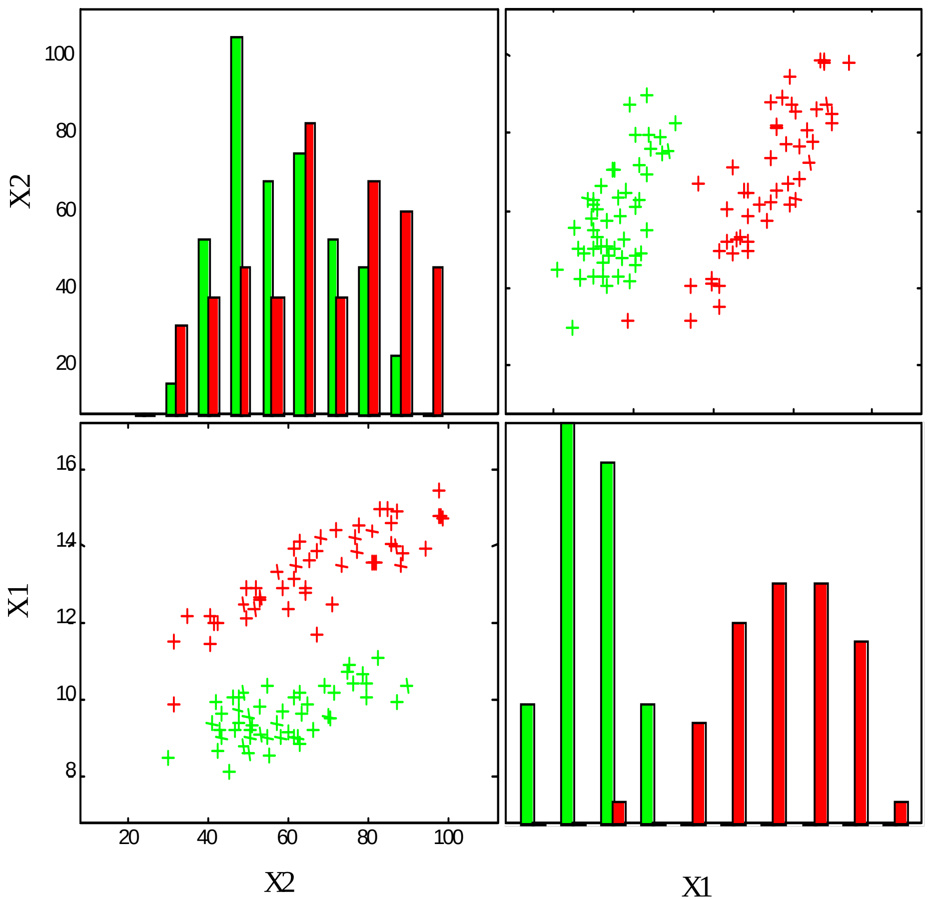


Guyon-Aliferis-Elisseeff, 2007

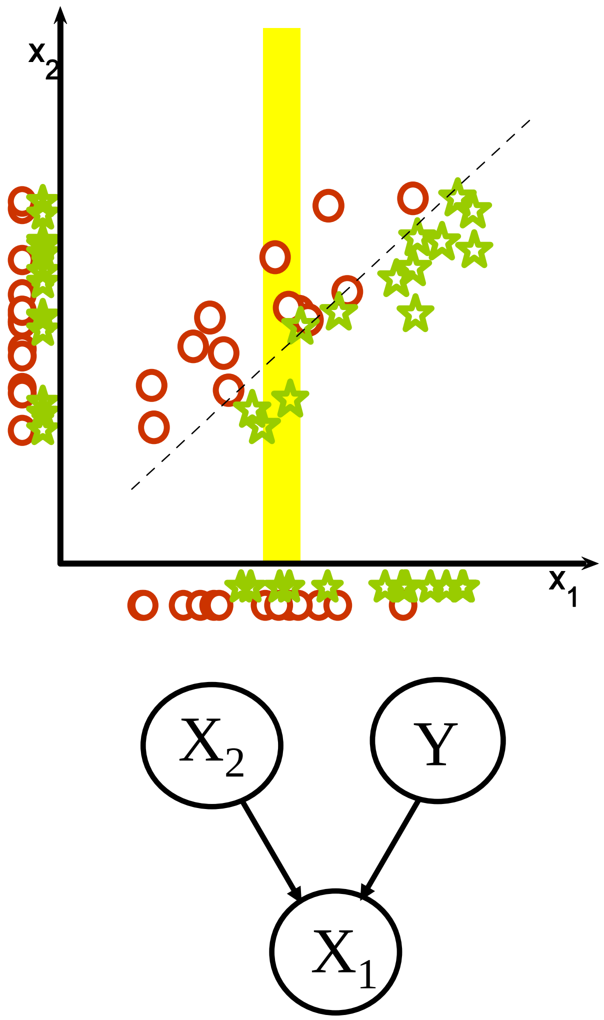
What can go wrong?

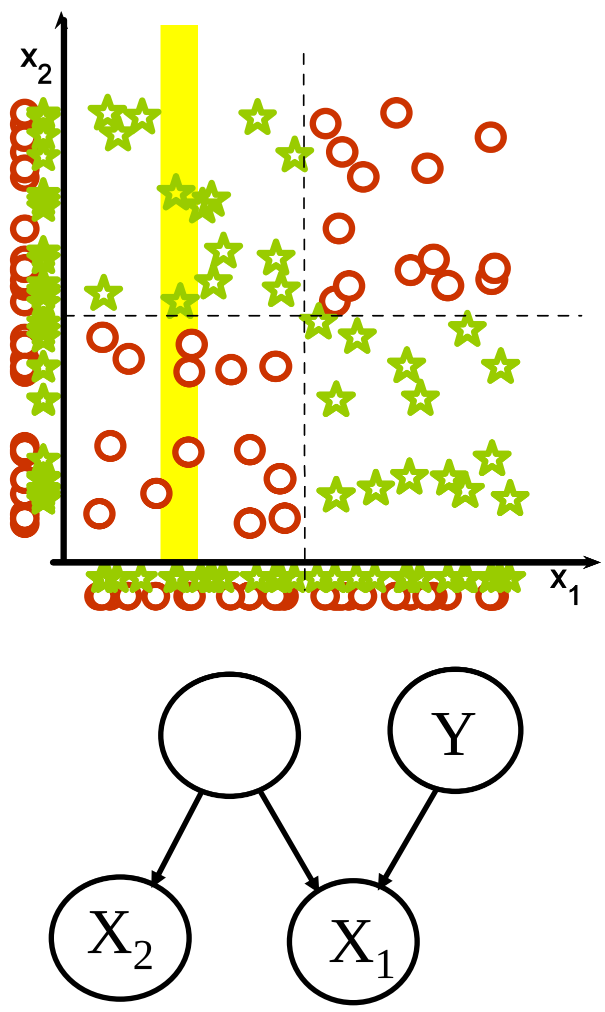






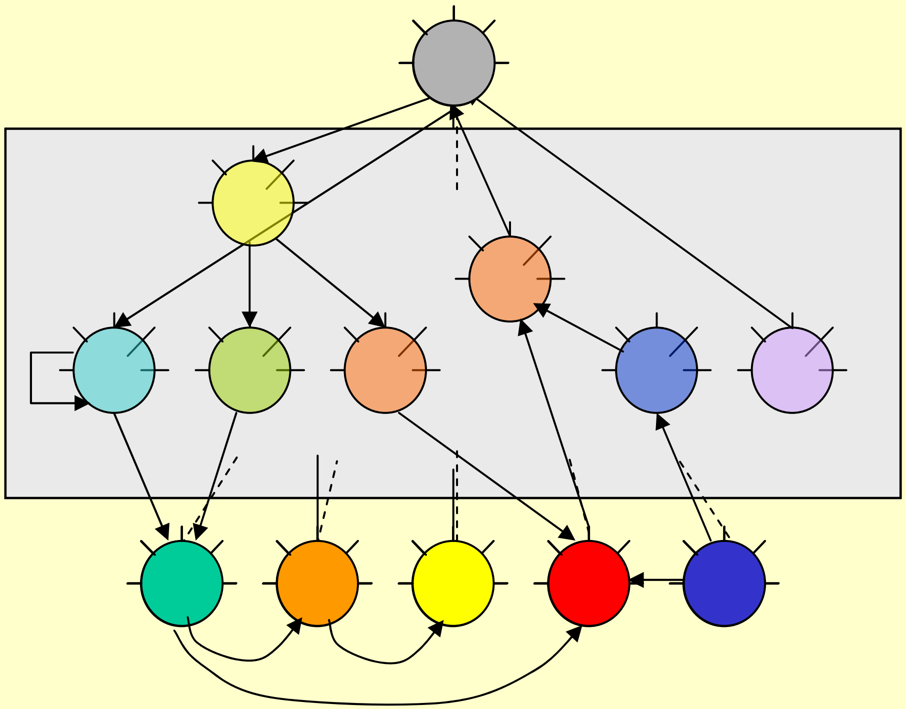
What can go wrong?





Guyon-Aliferis-Elisseeff, 2007

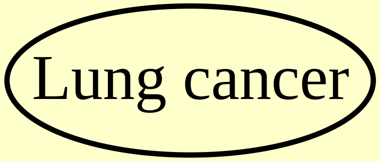
Causal feature selection



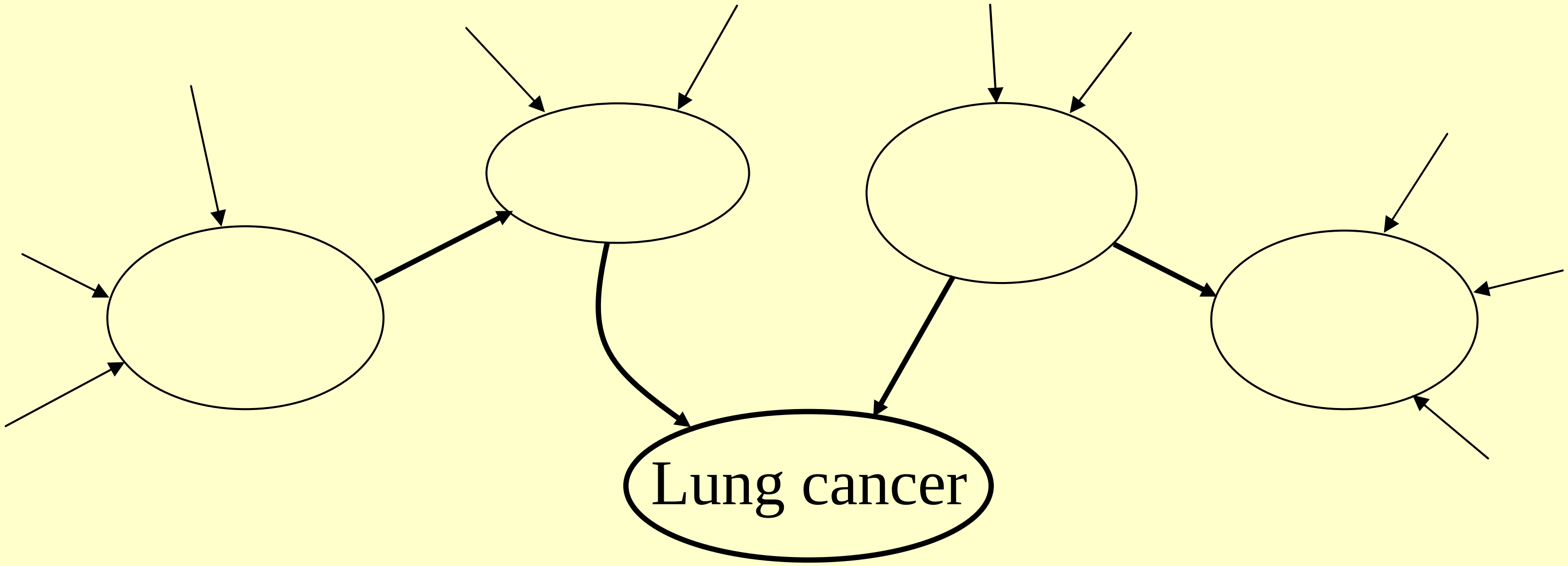
X

Uncover causal relationships between X_i and Y.

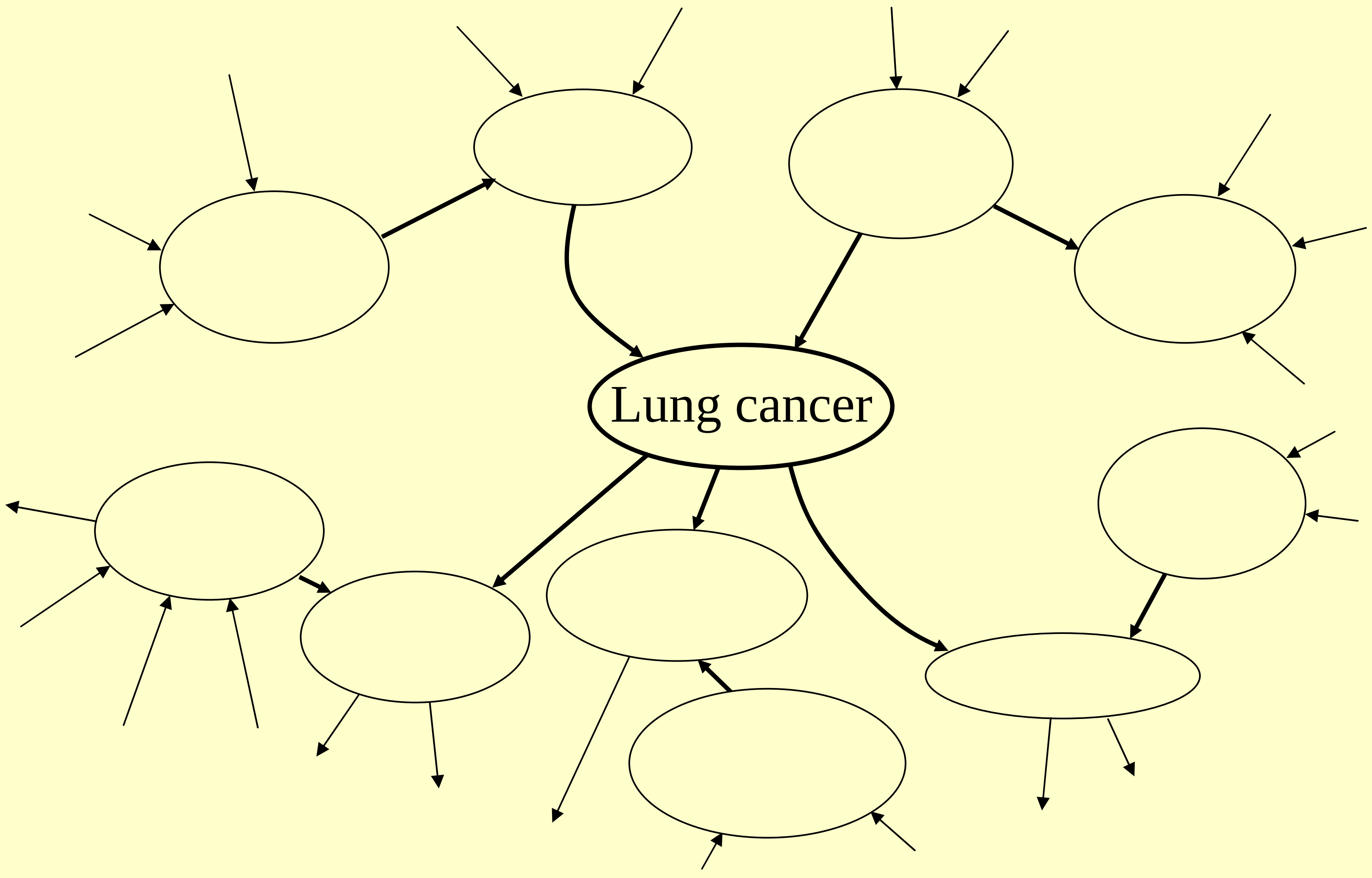
Causal feature relevance



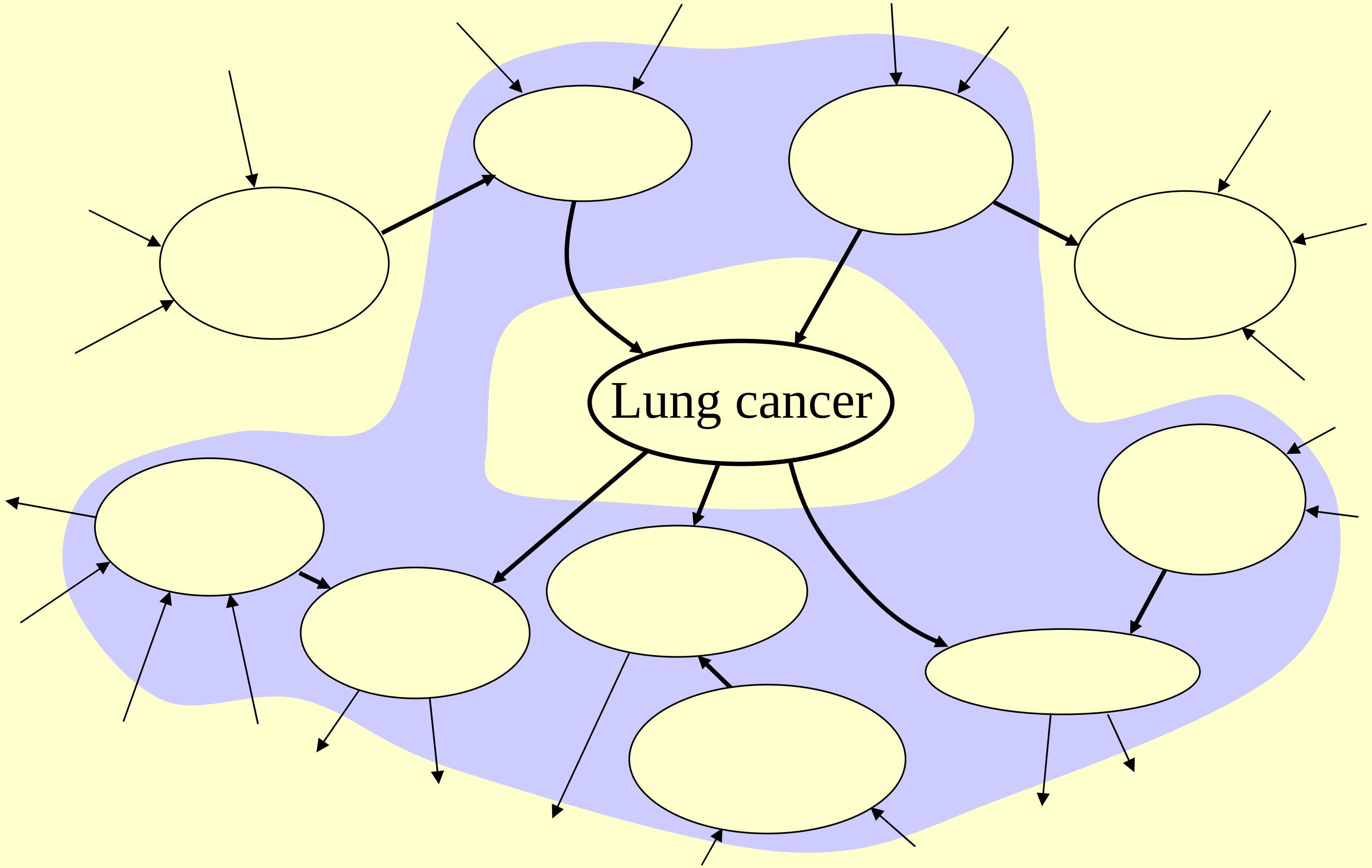
Causal feature relevance



Causal feature relevance



Markov Blanket



Strongly relevant features *(Kohavi-John, 1997)* ⇔ **Markov Blanket** *(Tsamardinos-Aliferis, 2003)*

Feature relevance

- **Surely irrelevant** feature X_i:

  $$P(X_i, Y \mid \mathbf{S}^{\setminus i}) = P(X_i \mid \mathbf{S}^{\setminus i}) P(Y \mid \mathbf{S}^{\setminus i})$$

  for all $\mathbf{S}^{\setminus i} \subseteq \mathbf{X}^{\setminus i}$ and all assignment of values to $\mathbf{S}^{\setminus i}$

- **Strongly relevant** feature X_i:

  $$P(X_i, Y \mid \mathbf{X}^{\setminus i}) \neq P(X_i \mid \mathbf{X}^{\setminus i}) P(Y \mid \mathbf{X}^{\setminus i})$$

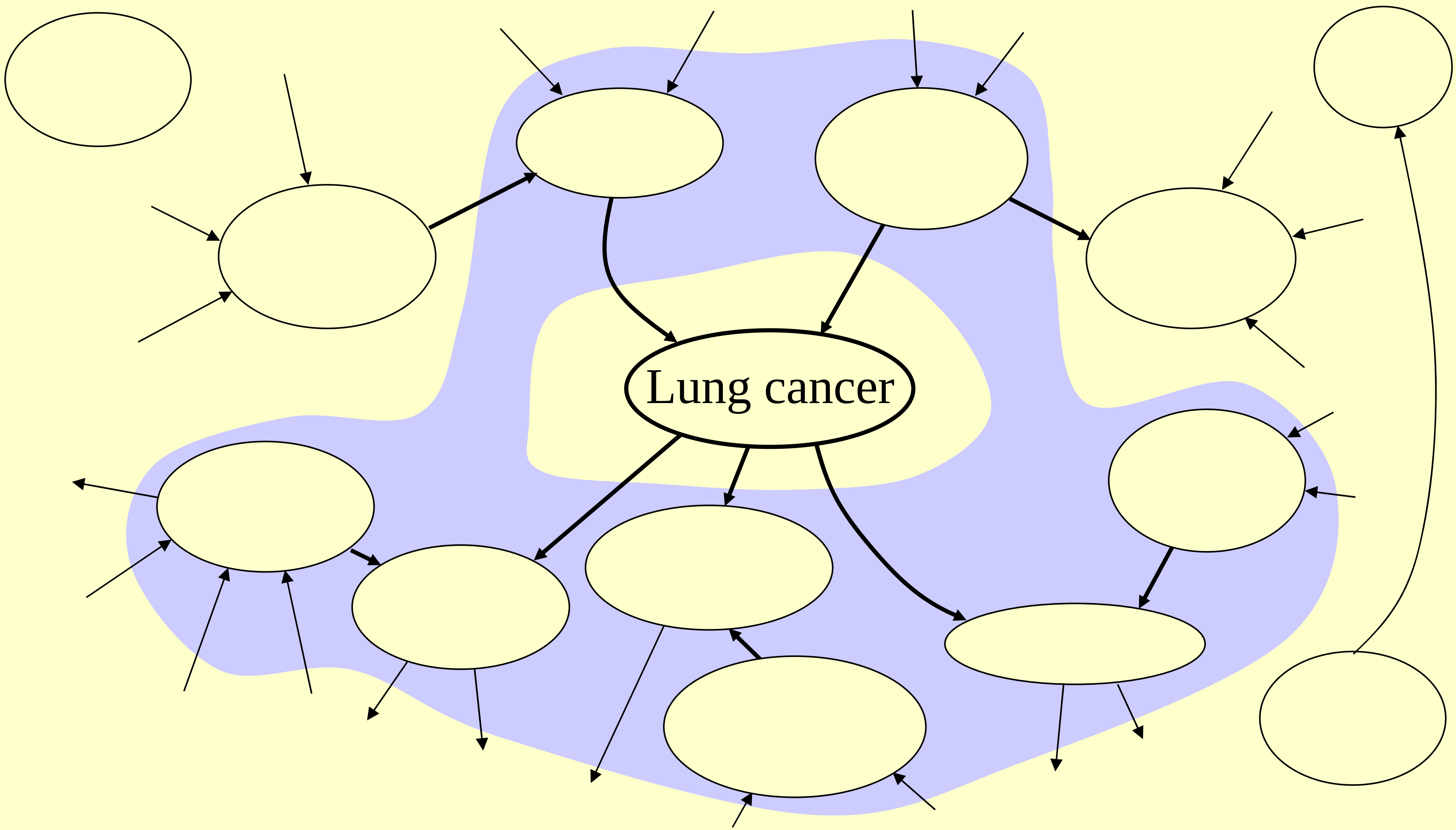
  for some assignment of values to $\mathbf{X}^{\setminus i}$

- **Weakly relevant** feature X_i:

  $$P(X_i, Y \mid \mathbf{S}^{\setminus i}) \neq P(X_i \mid \mathbf{S}^{\setminus i}) P(Y \mid \mathbf{S}^{\setminus i})$$

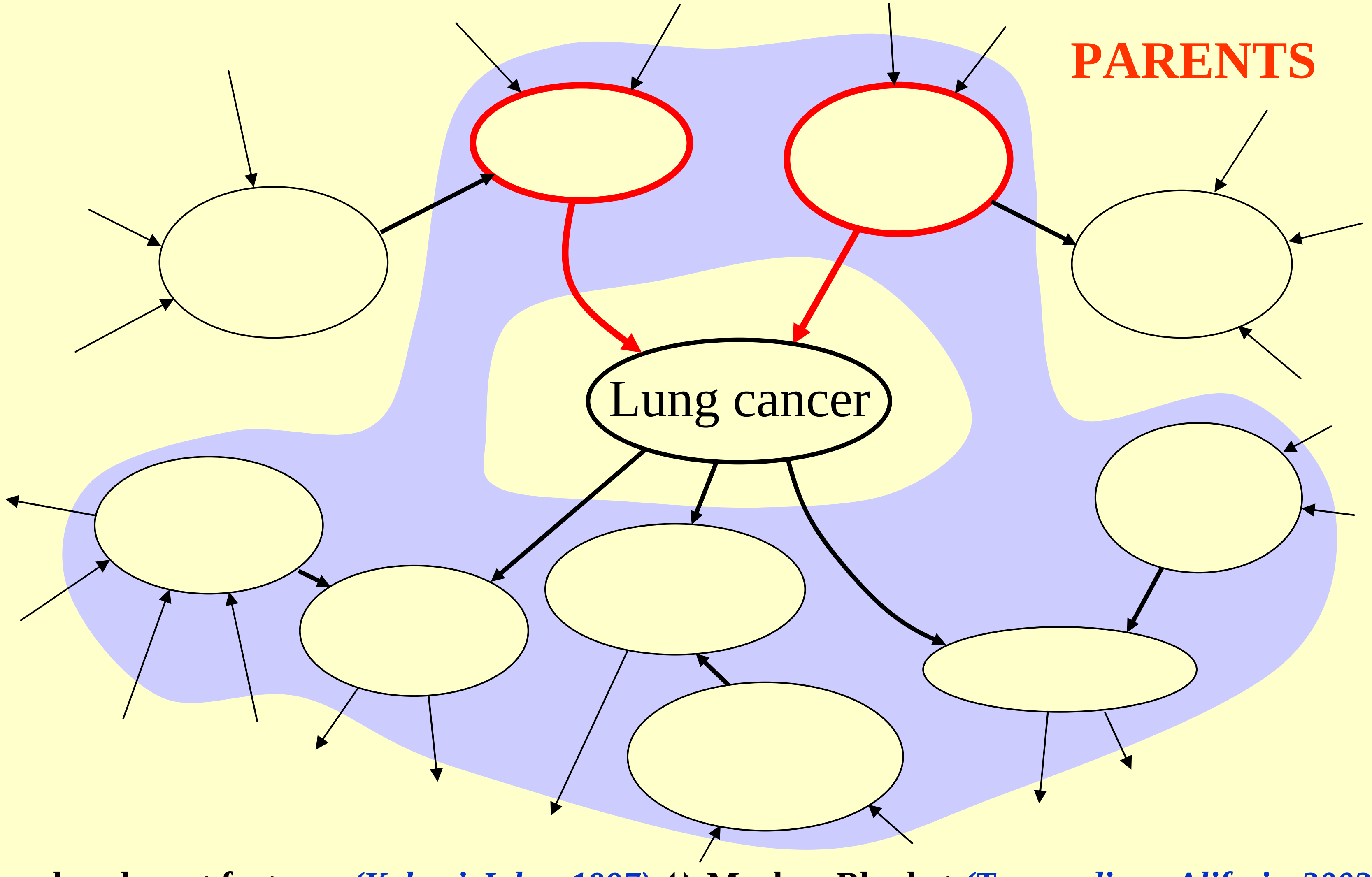
  for some assignment of values to $\mathbf{S}^{\setminus i} \subset \mathbf{X}^{\setminus i}$

Markov Blanket



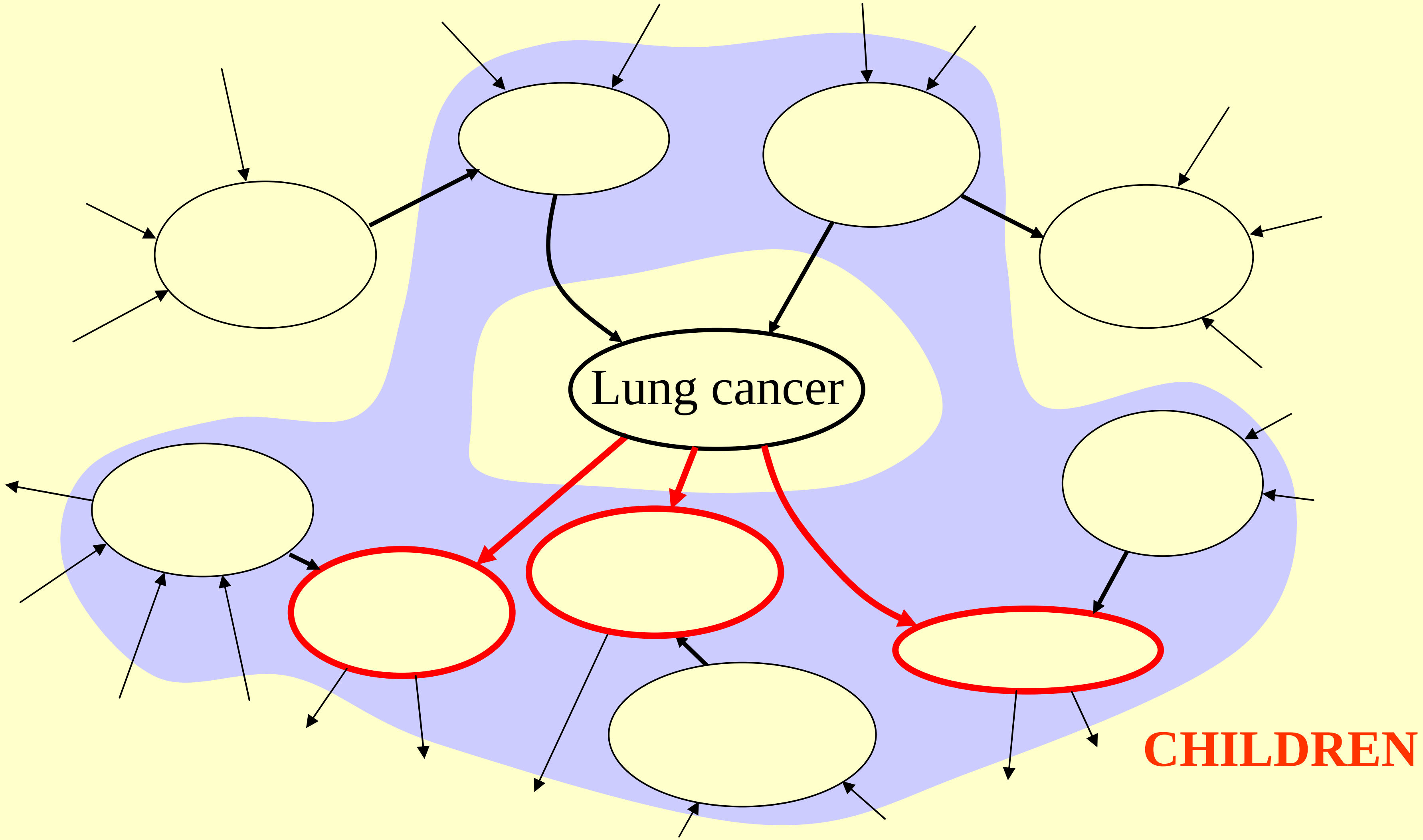
Strongly relevant features *(Kohavi-John, 1997)* ⇔ **Markov Blanket** *(Tsamardinos-Aliferis, 2003)*

Markov Blanket



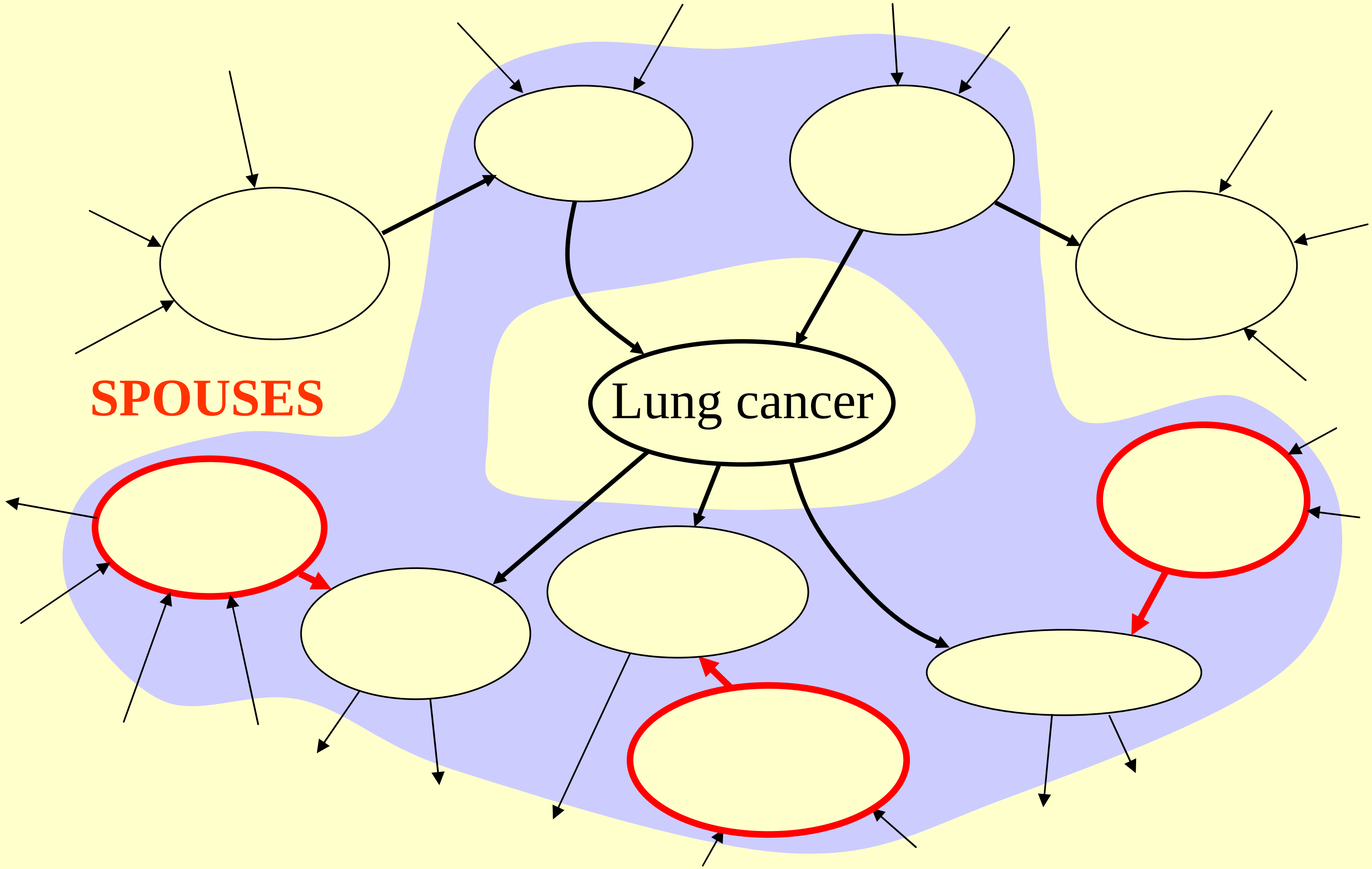
Strongly relevant features ***(Kohavi-John, 1997)*** **⇔ Markov Blanket** ***(Tsamardinos-Aliferis, 2003)***

Markov Blanket



Strongly relevant features *(Kohavi-John, 1997)* ⇔ **Markov Blanket** *(Tsamardinos-Aliferis, 2003)*

Markov Blanket



Strongly relevant features *(Kohavi-John, 1997)* ⇔ **Markov Blanket** *(Tsamardinos-Aliferis, 2003)*

Causal relevance

- **Surely irrelevant** feature X_i:

  $$P(X_i, Y \mid \mathbf{S}^{\backslash i}) = P(X_i \mid \mathbf{S}^{\backslash i}) P(Y \mid \mathbf{S}^{\backslash i})$$

  for all $\mathbf{S}^{\backslash i} \subseteq \mathbf{X}^{\backslash i}$ and all assignment of values to $\mathbf{S}^{\backslash i}$

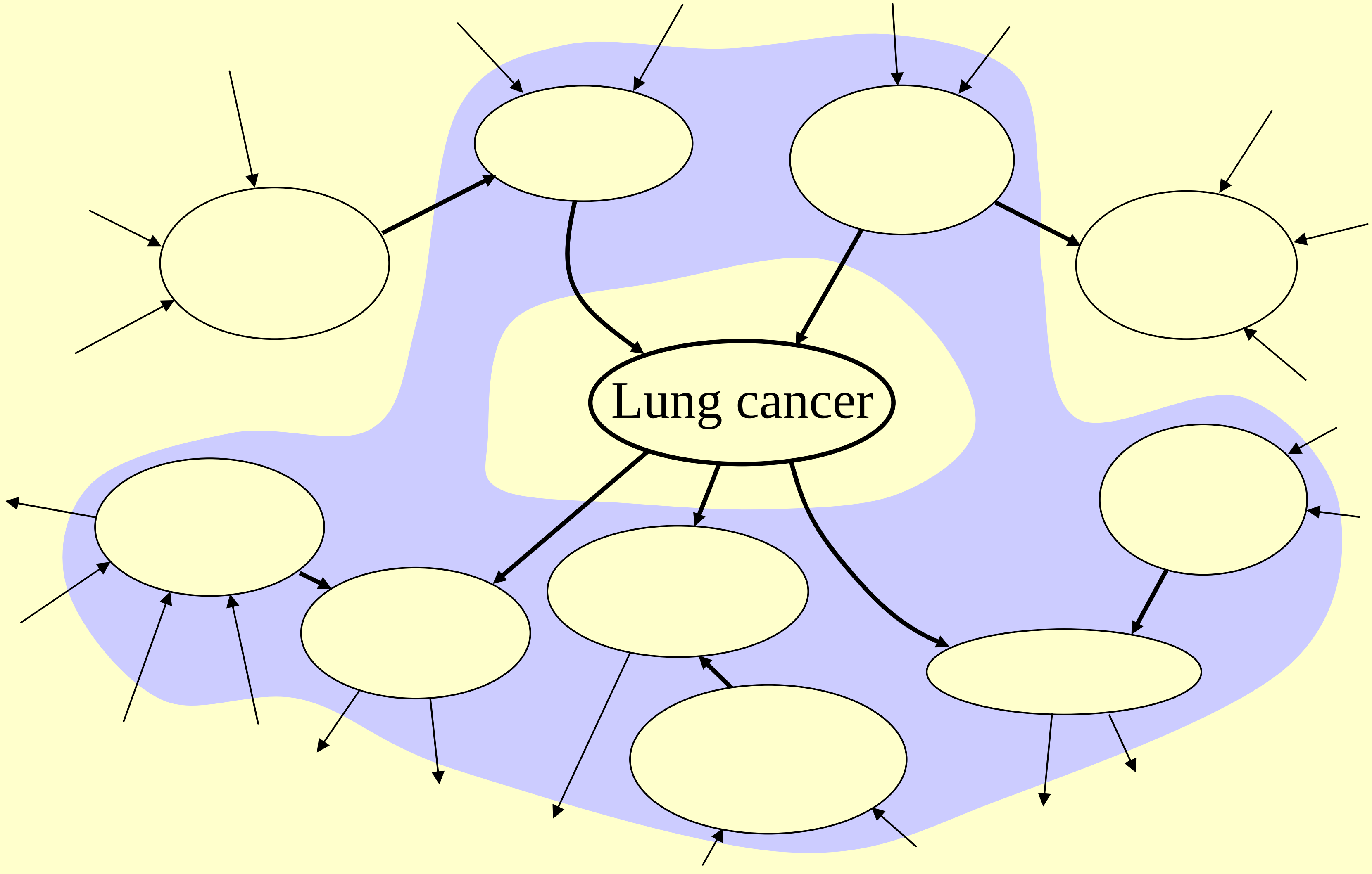
- **Causally relevant** feature X_i:

  $$P(X_i, Y \mid \mathbf{do}(\mathbf{S}^{\backslash i})) \neq P(X_i \mid \mathbf{do}(\mathbf{S}^{\backslash i})) P(Y \mid \mathbf{do}(\mathbf{S}^{\backslash i}))$$

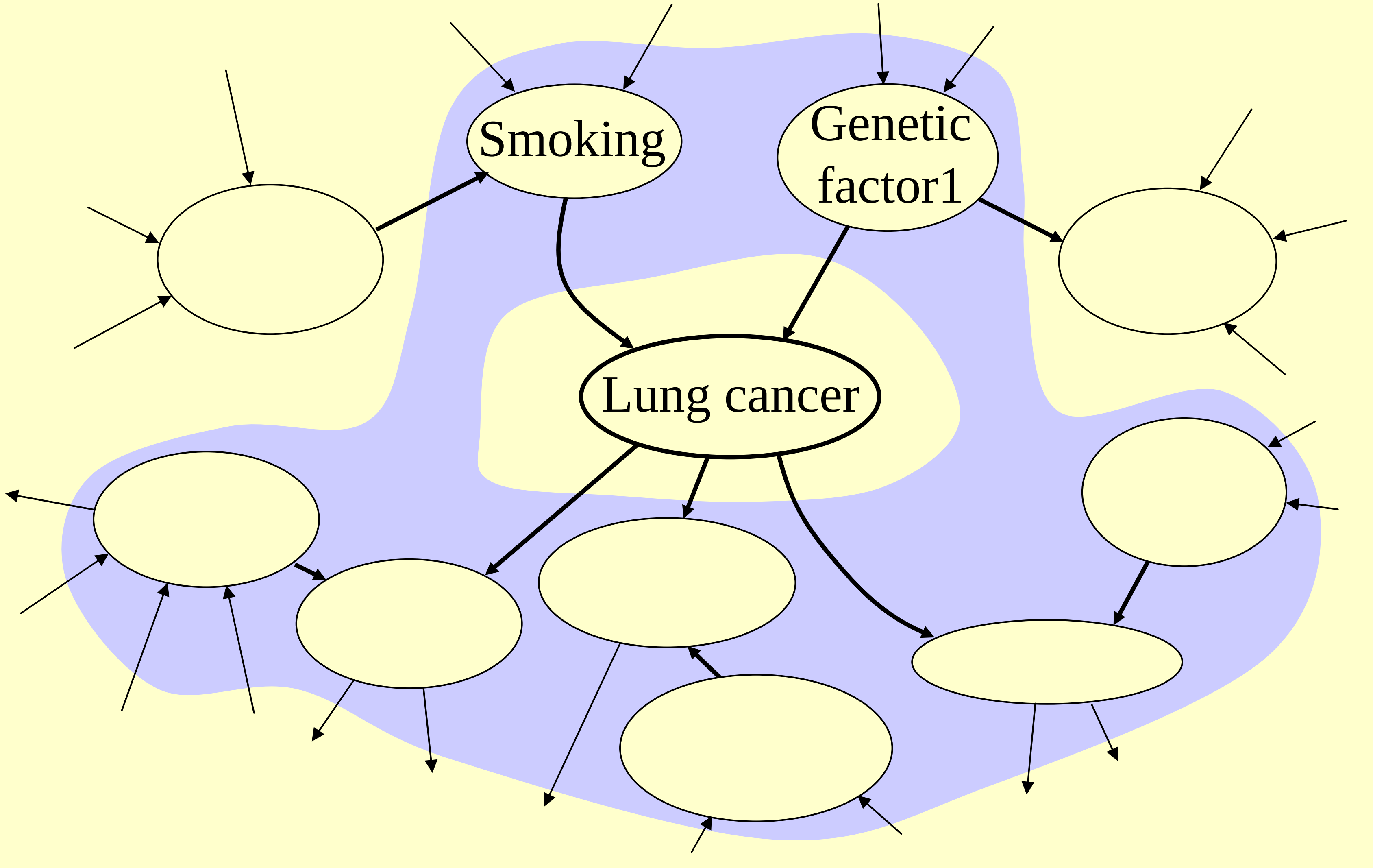
  for some assignment of values to $\mathbf{S}^{\backslash i}$

- **Weak/strong causal relevance:**
  - ***Weak=ancestors, indirect causes***
  - ***Strong=parents, direct causes.***

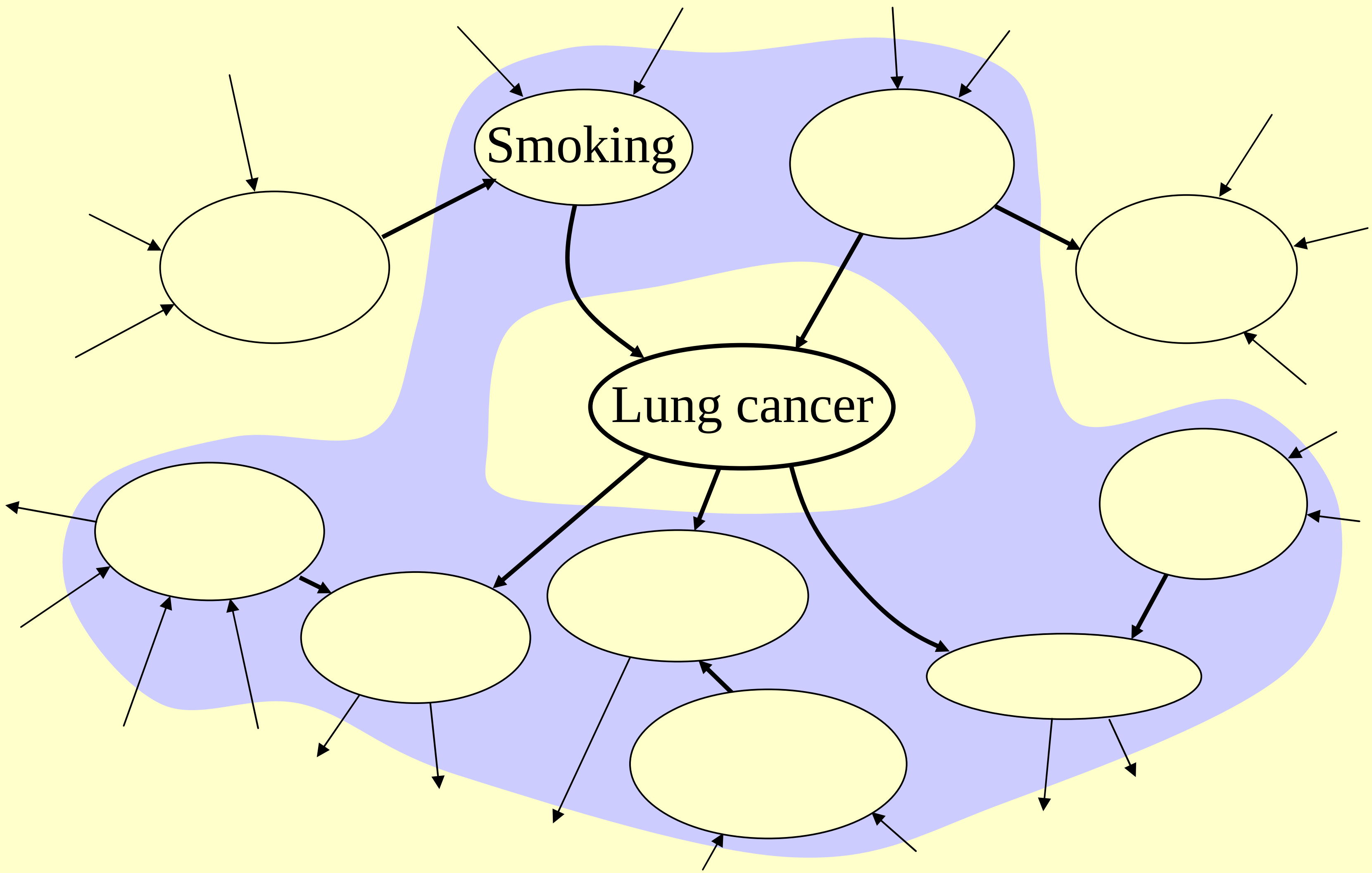
Examples



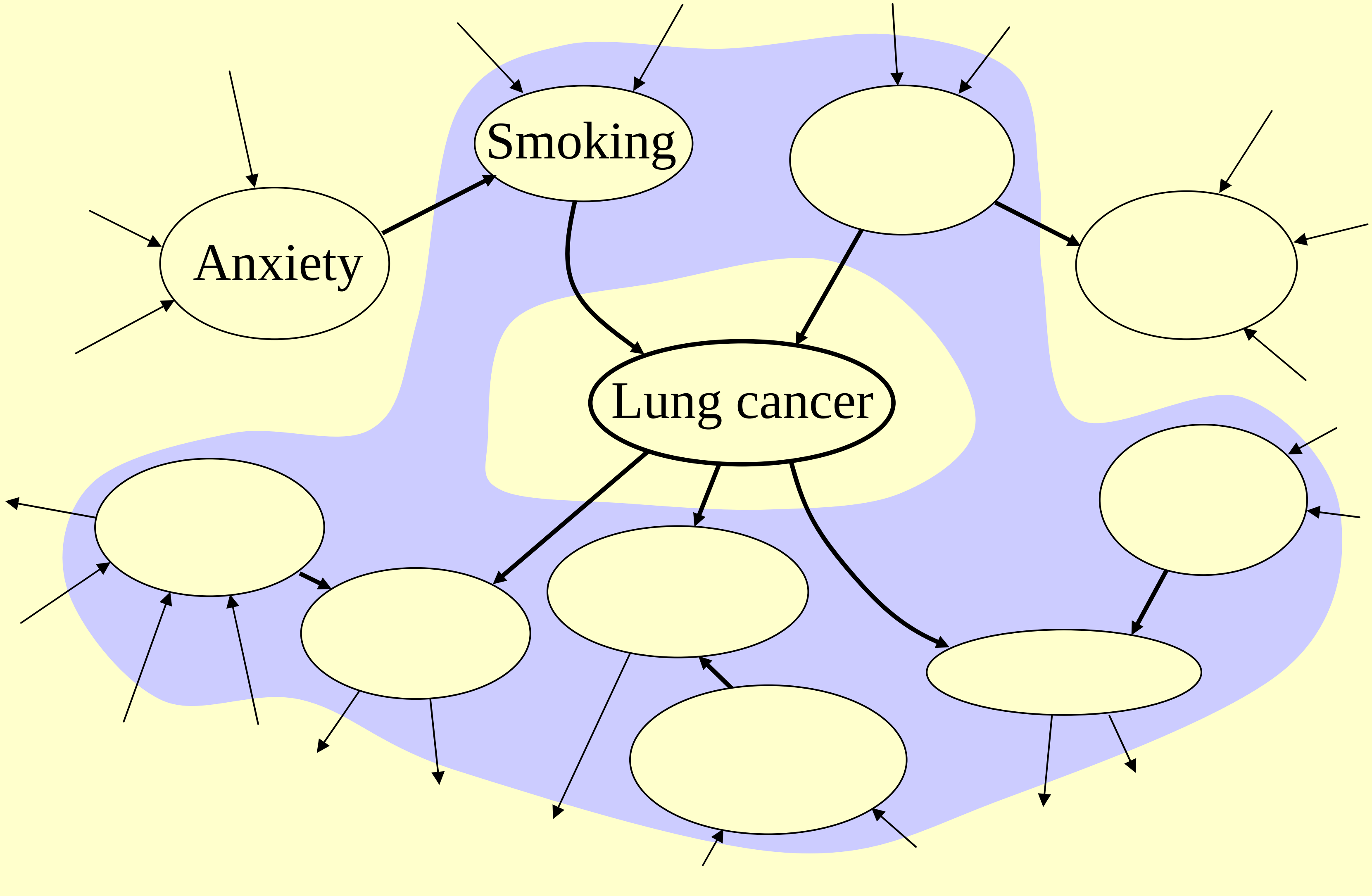
Immediate causes (parents)



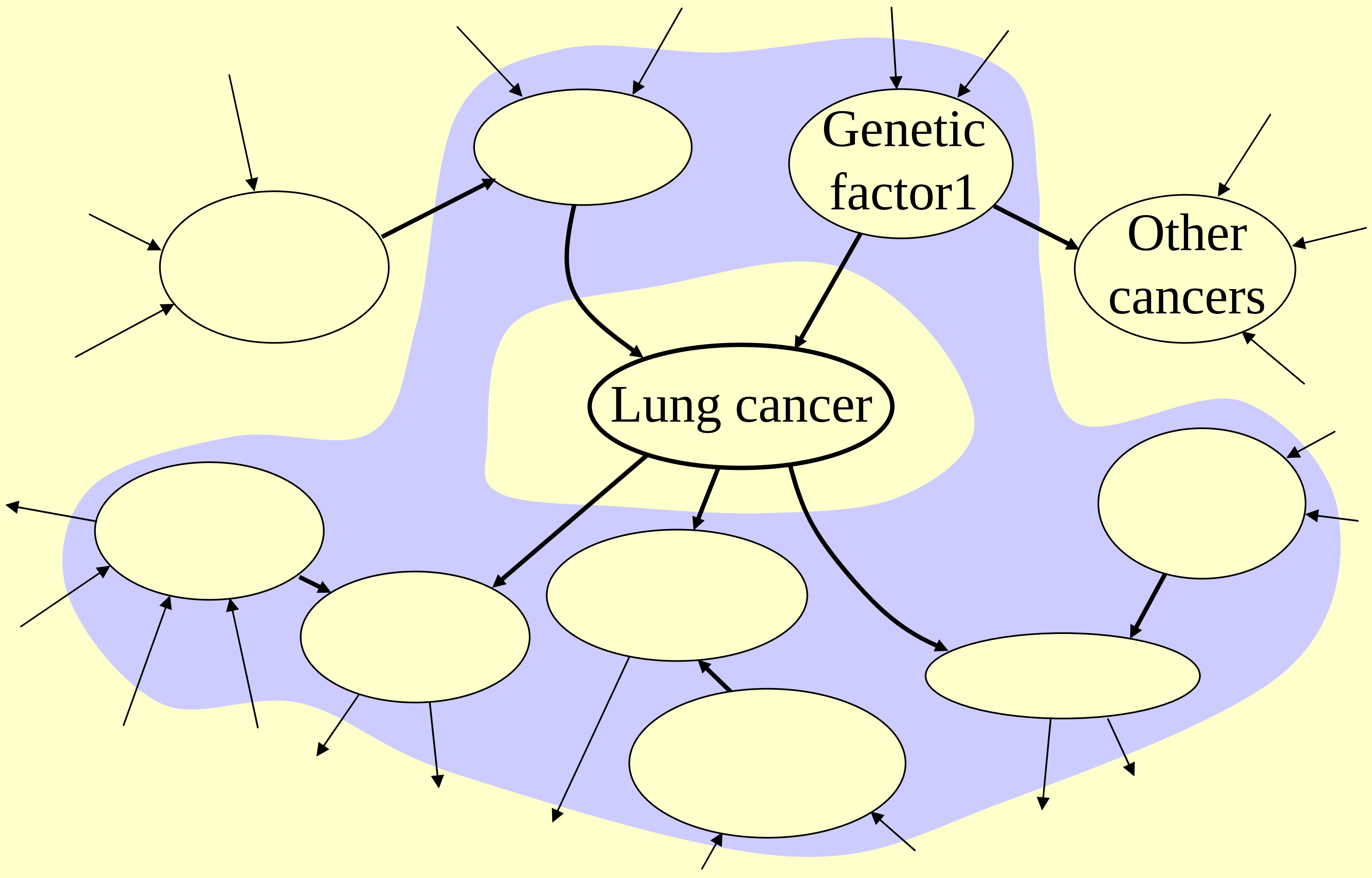
Immediate causes (parents)



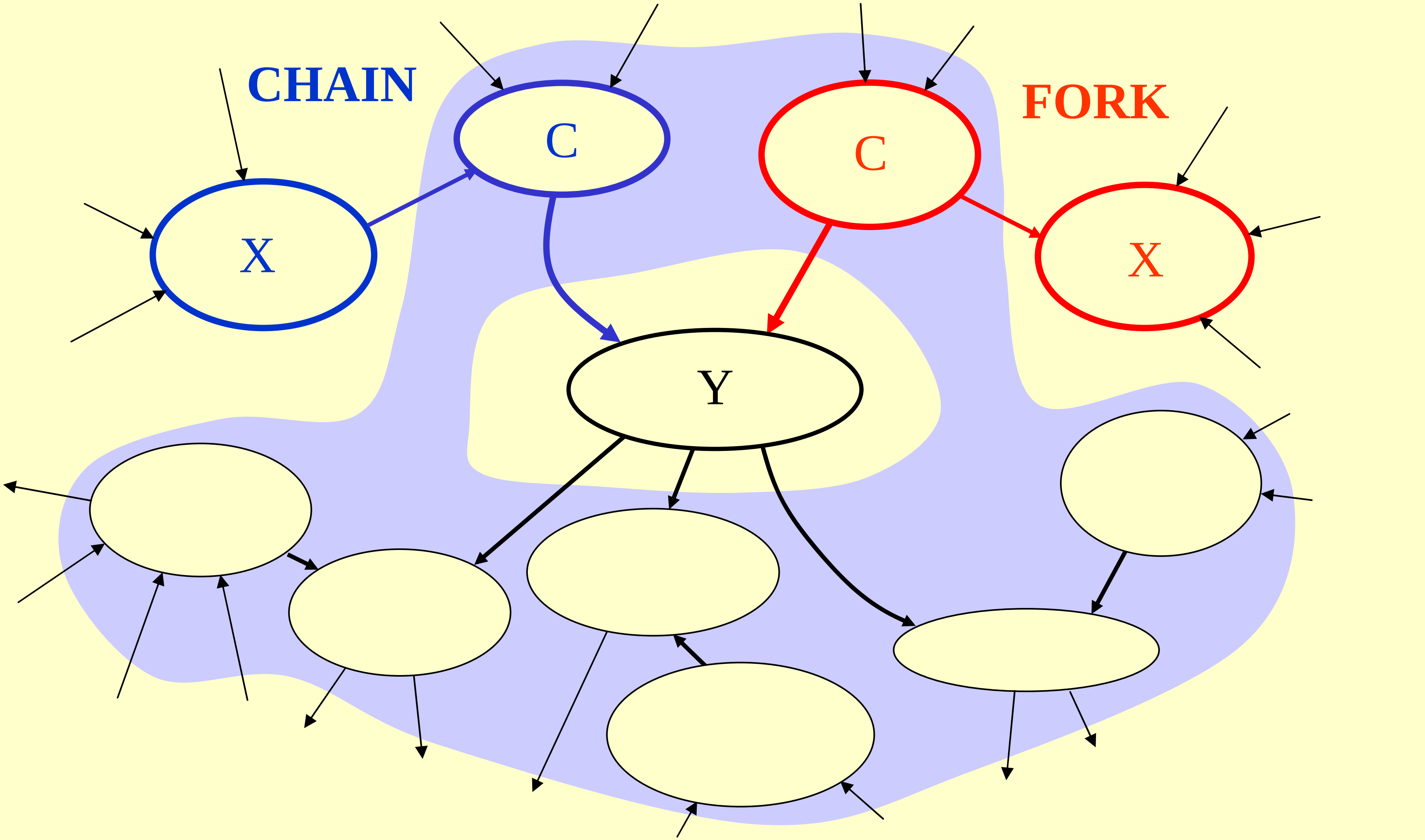
Non-immediate causes (other ancestors)



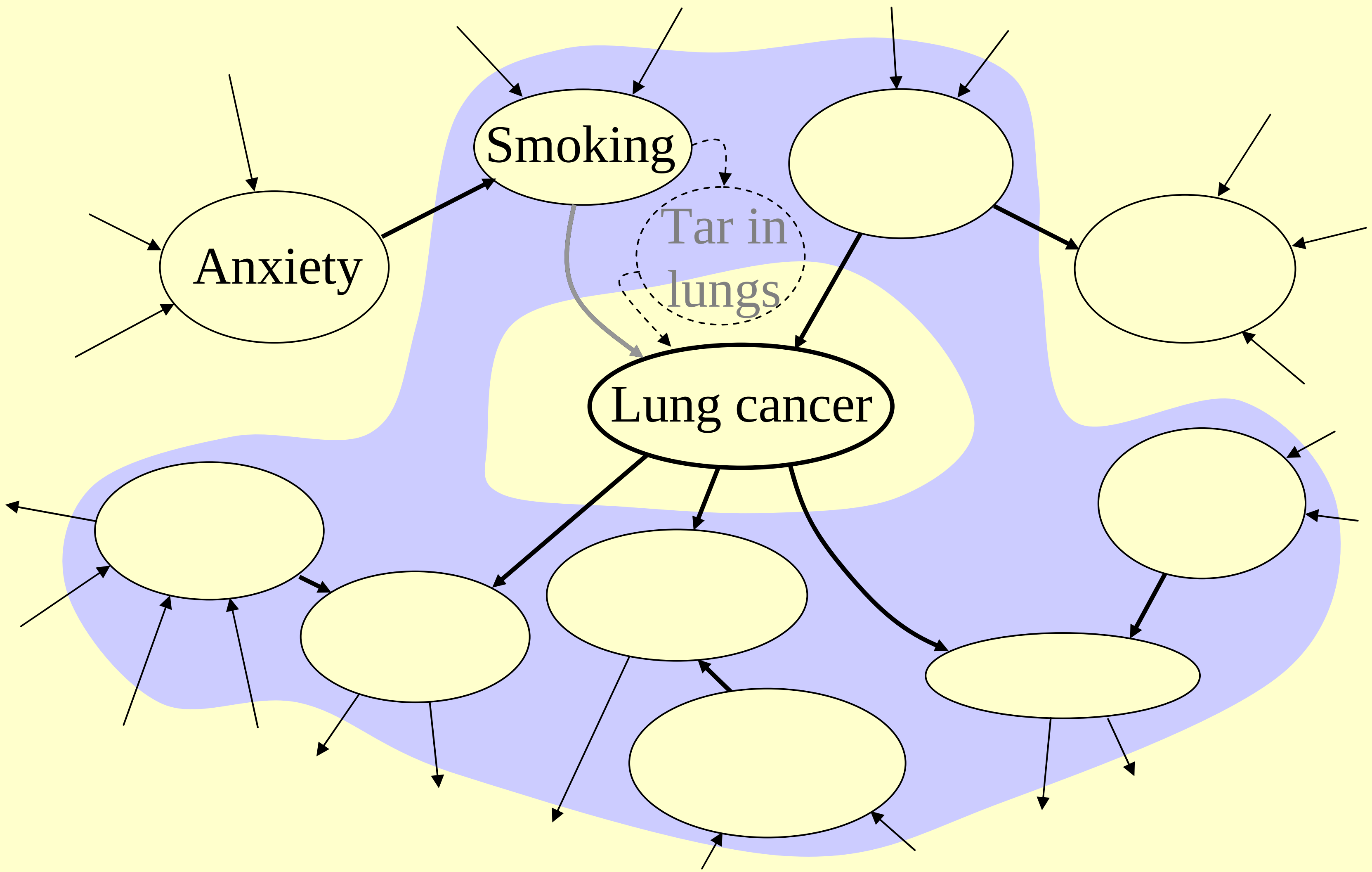
Non causes (e.g. siblings)



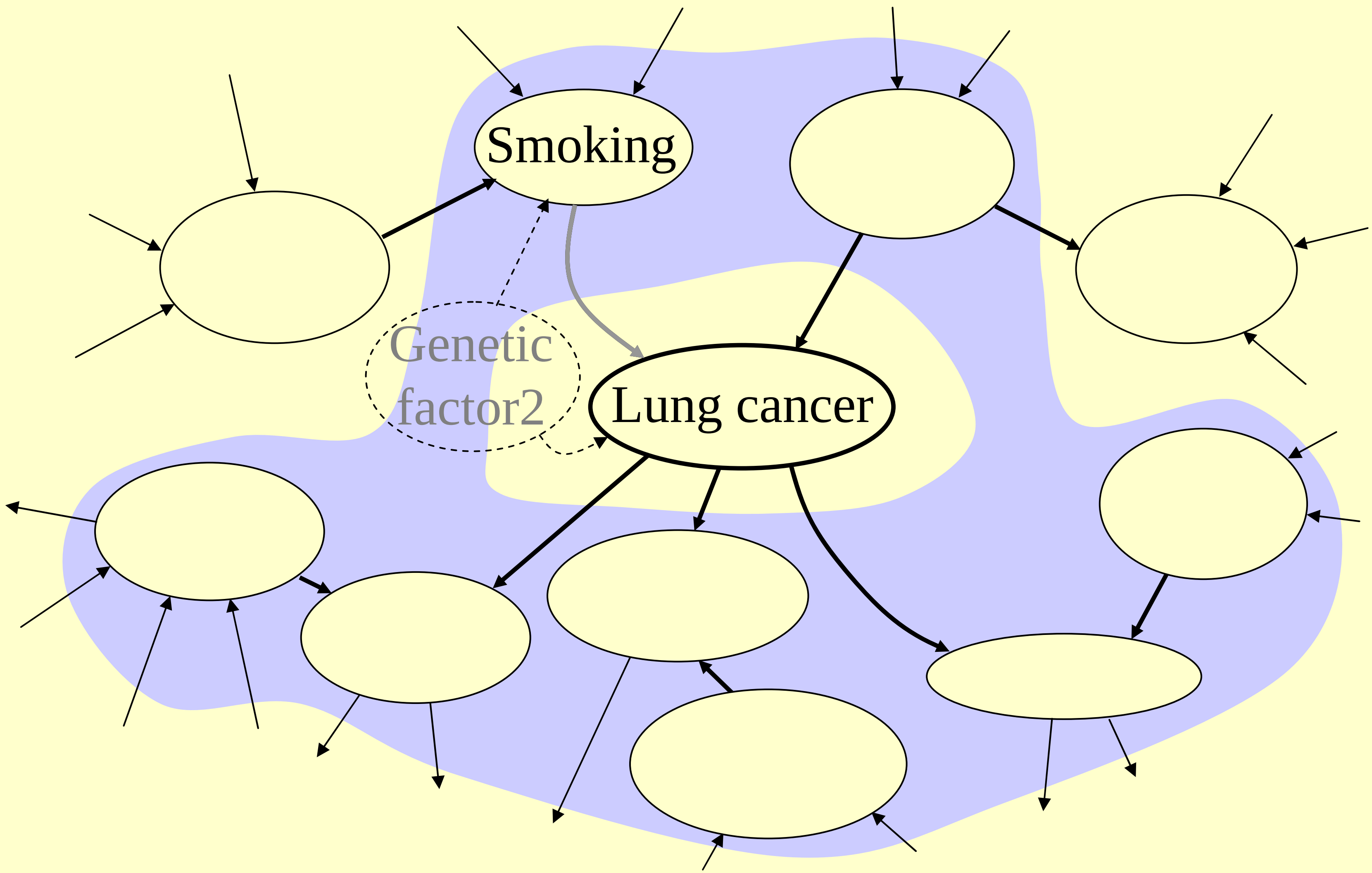
$$X \perp\!\!\!\perp Y \mid C$$



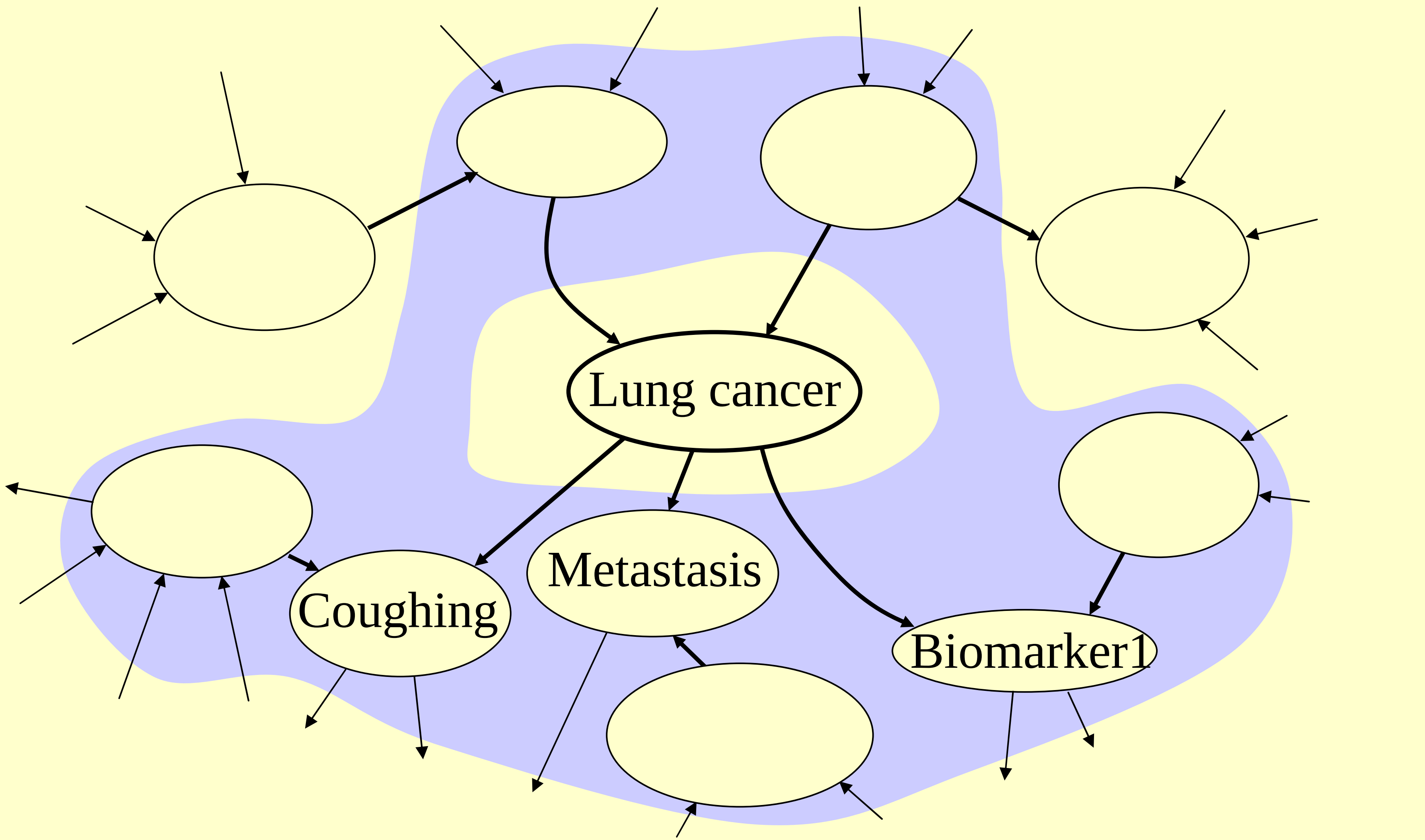
Hidden more direct cause



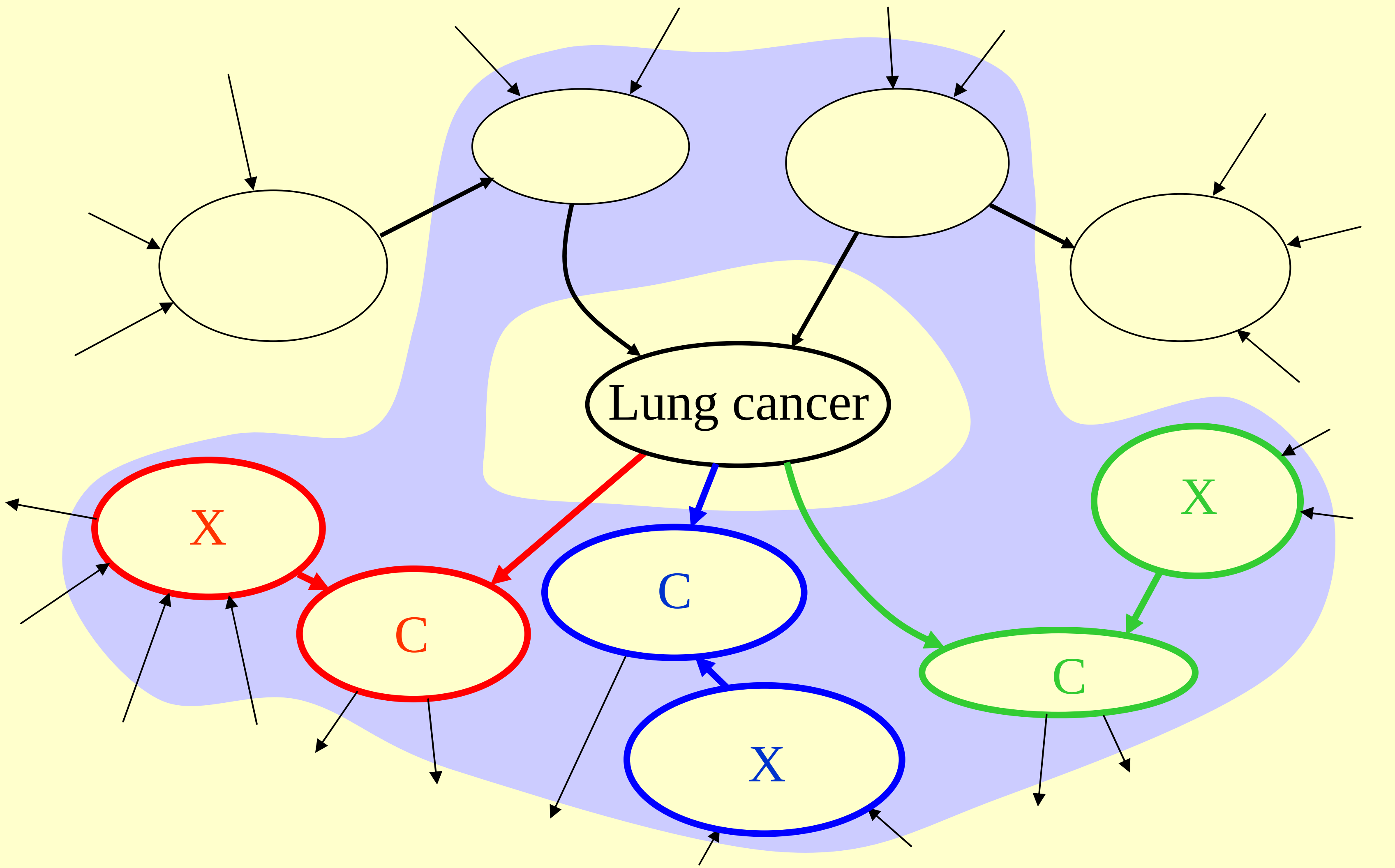
Confounder



Immediate consequences (children)

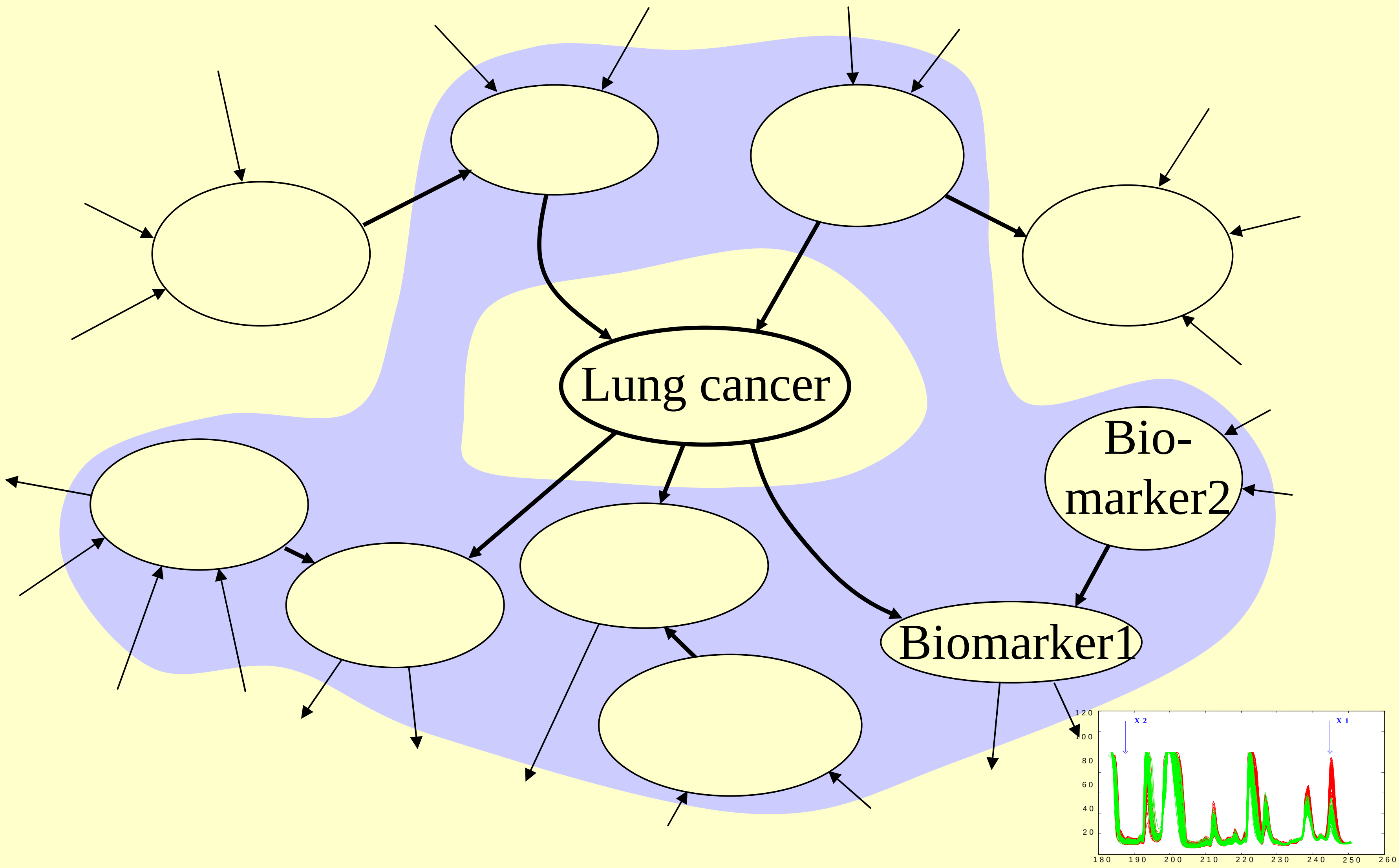


$$X \perp\!\!\!\perp Y \quad \textit{but} \quad \cancel{X \perp\!\!\!\perp} Y \mid C$$

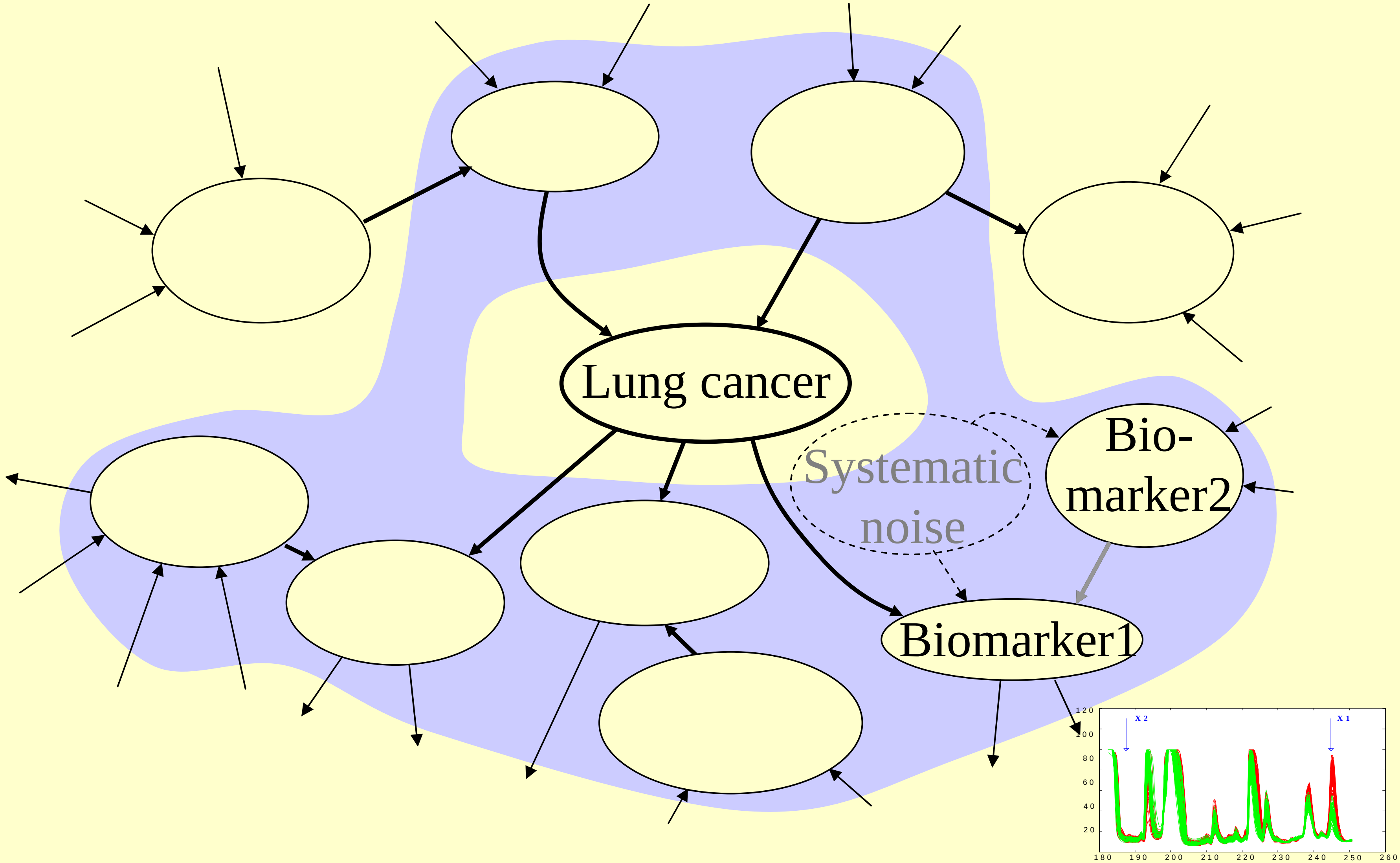


Strongly relevant features ***(Kohavi-John, 1997)*** ⬄ **Markov Blanket** ***(Tsamardinos-Aliferis, 2003)***

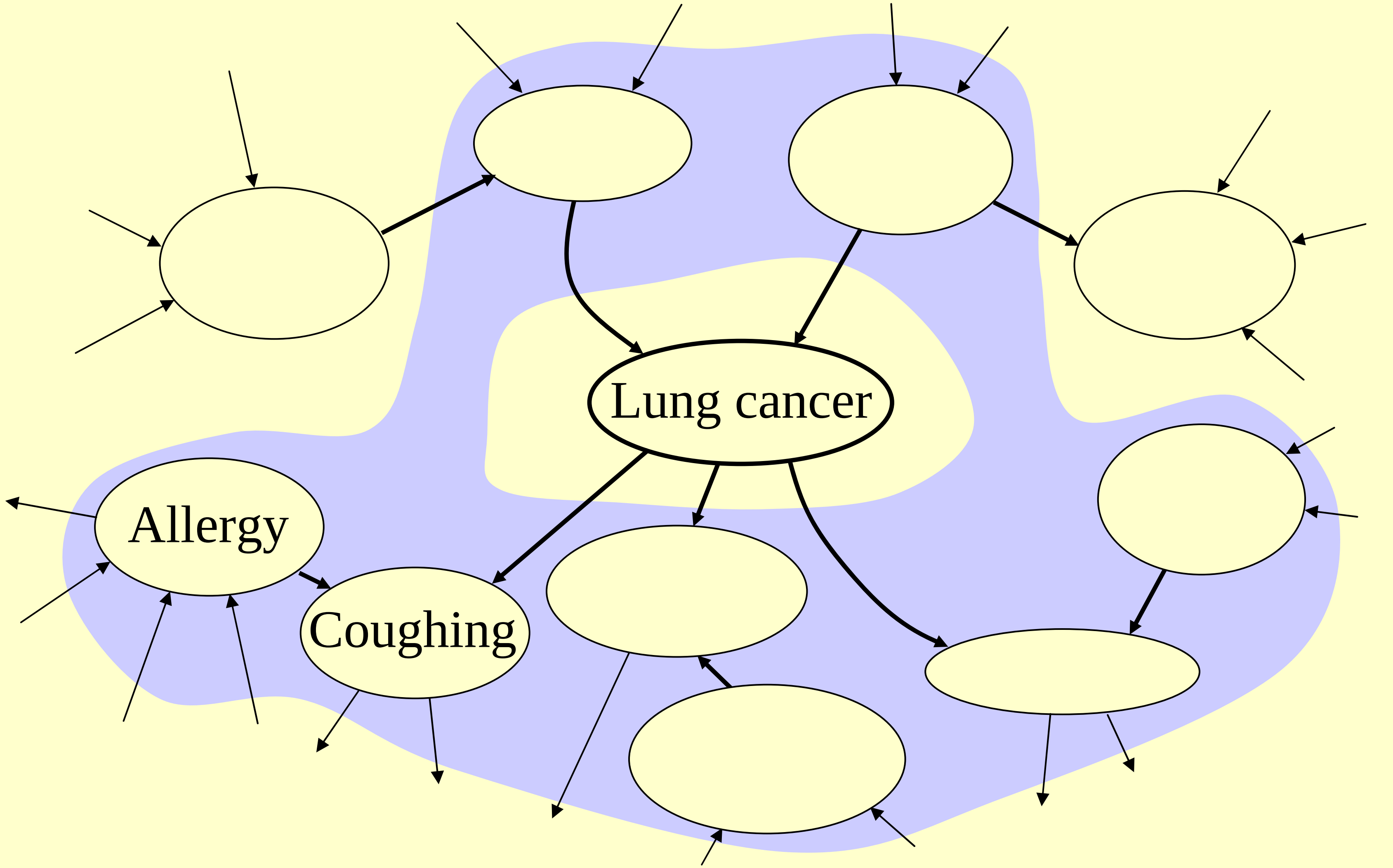
Non relevant spouse (artifact)



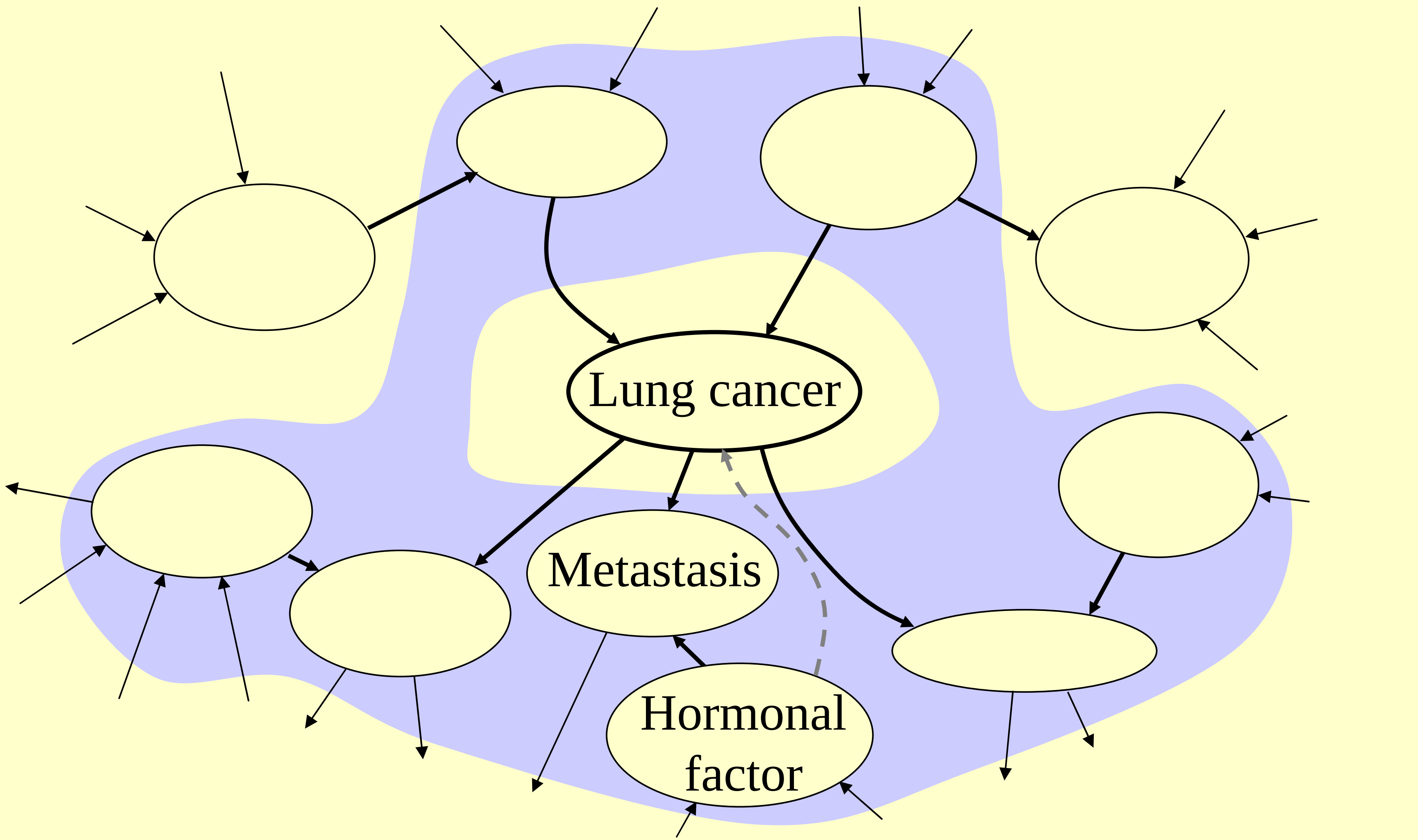
Another case of confounder



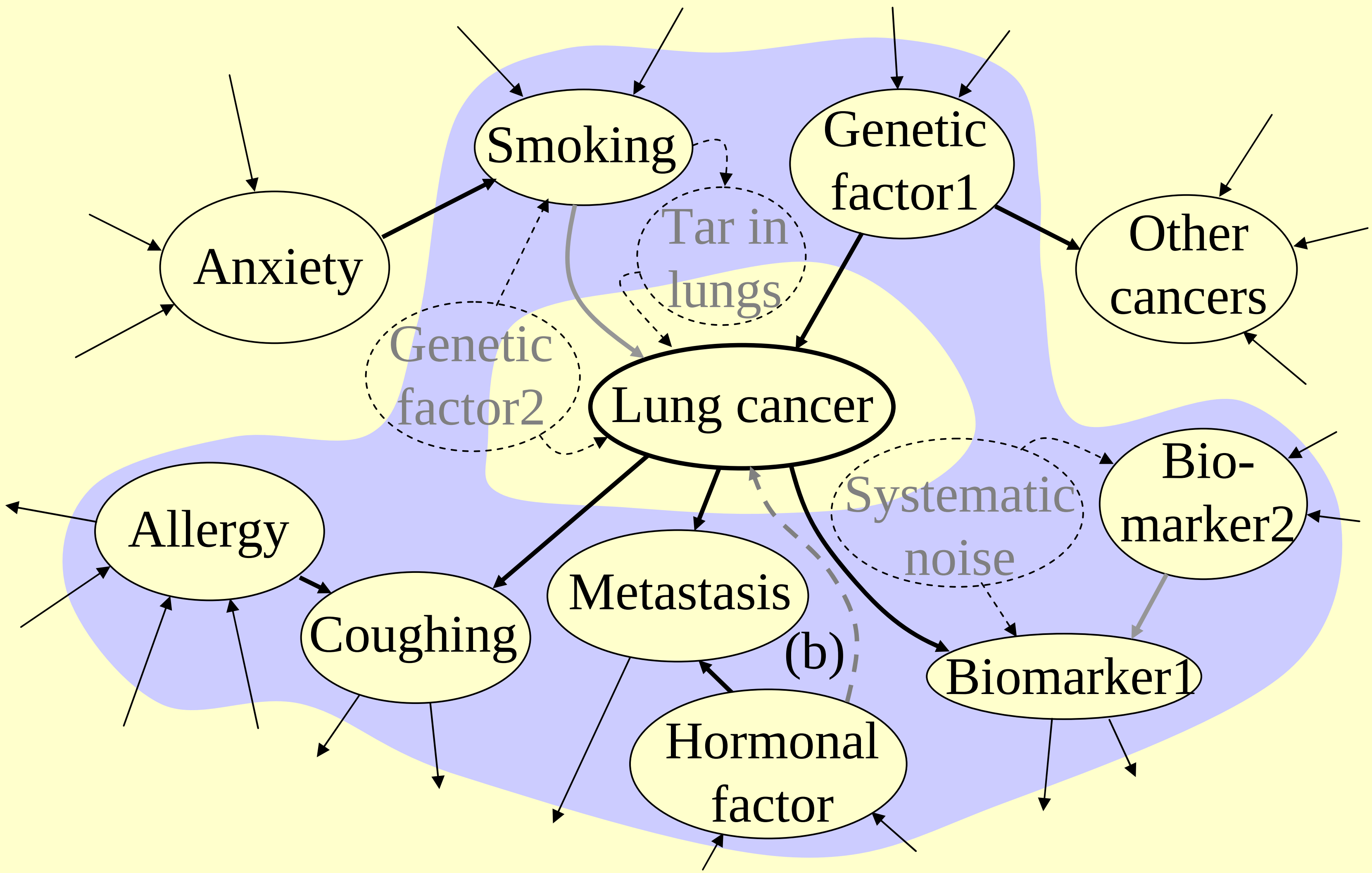
Truly relevant spouse



Sampling bias



Causal feature relevance



Formalism: Causal Bayesian networks

- **Bayesian network:**
  - Graph with random variables X_1, X_2, …X_n as nodes.
  - Dependencies represented by edges.
  - Allow us to compute $P(X_1, X_2, \ldots X_n)$ as

    $\prod_i P(X_i \mid Parents(X_i))$.
  - Edge directions have no meaning.
- **Causal Bayesian network:** egde directions indicate causality.

Example of Causal Discovery Algorithm

Algorithm: PC ***(Peter Spirtes and Clarck Glymour, 1999)***

Let A, B, C $\in$ **X** and **V** $\subset$ **X**.

Initialize with a fully connected un-oriented graph.

1. Find un-oriented edges by using the criterion that variable A shares a direct edge with variable B *iff* no subset of other variables **V** can render them conditionally independent (A $\perp$ B | **V**).
2. Orient edges in "collider" triplets (i.e., of the type: A $\rightarrow$ C $\leftarrow$ B) using the criterion that if there are direct edges between A, C and between C and B, but not between A and B, then A $\rightarrow$ C $\leftarrow$ B, *iff* there is no subset **V** containing C such that A $\perp$ B | **V**.
3. Further orient edges with a constraint-propagation method by adding orientations until no further orientation can be produced, using the two following criteria:

   (i) If A $\rightarrow$ B $\rightarrow$ … $\rightarrow$ C, and A — C (i.e. there is an undirected edge between A and C) then A $\rightarrow$ C.

   (ii) If A $\rightarrow$ B — C then B $\rightarrow$ C.

Computational and statistical complexity

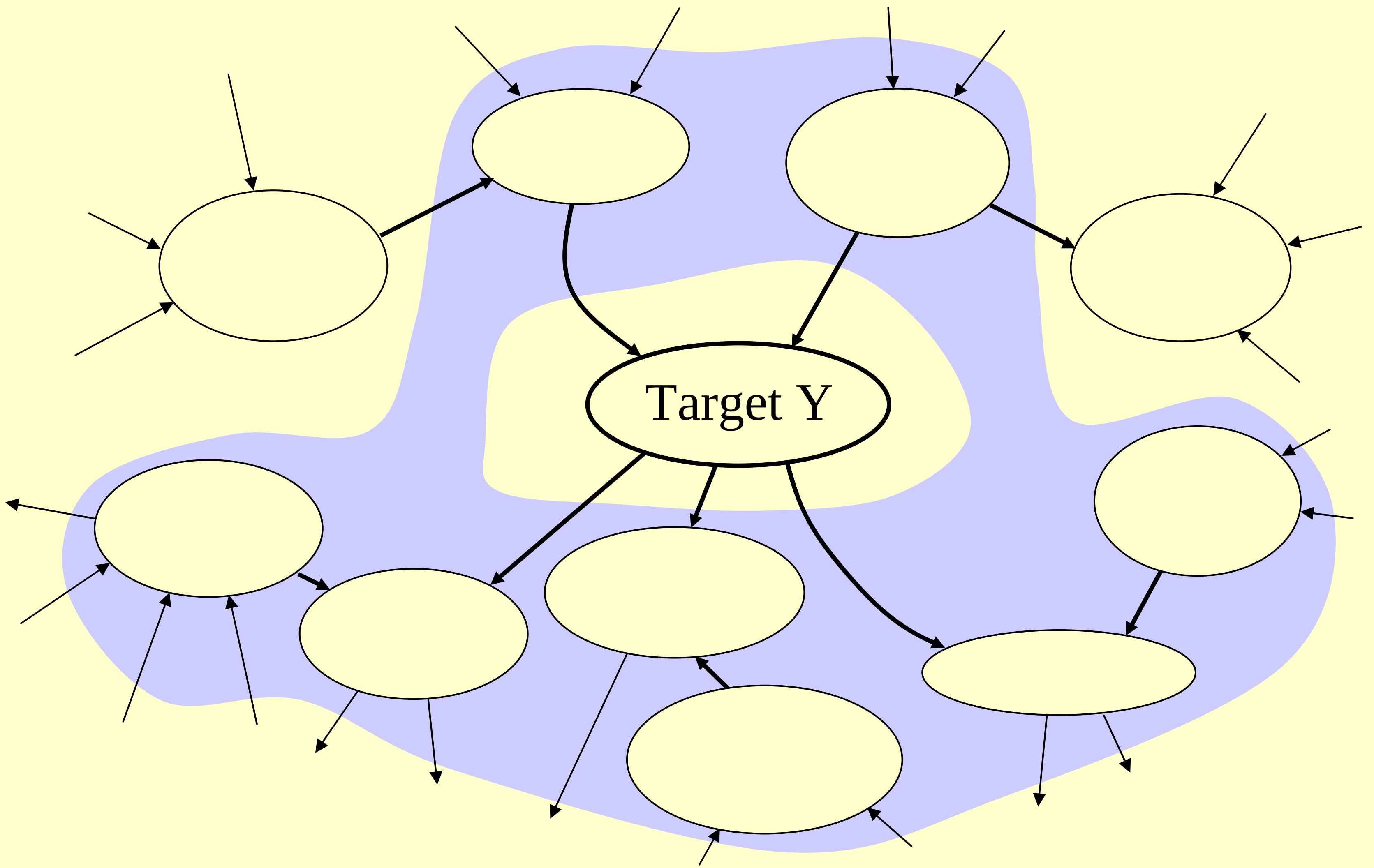
Computing the full causal graph poses:

- Computational challenges (intractable for large numbers of variables)
- Statistical challenges (difficulty of estimation of conditional probabilities for many var. w. few samples).

Compromise:

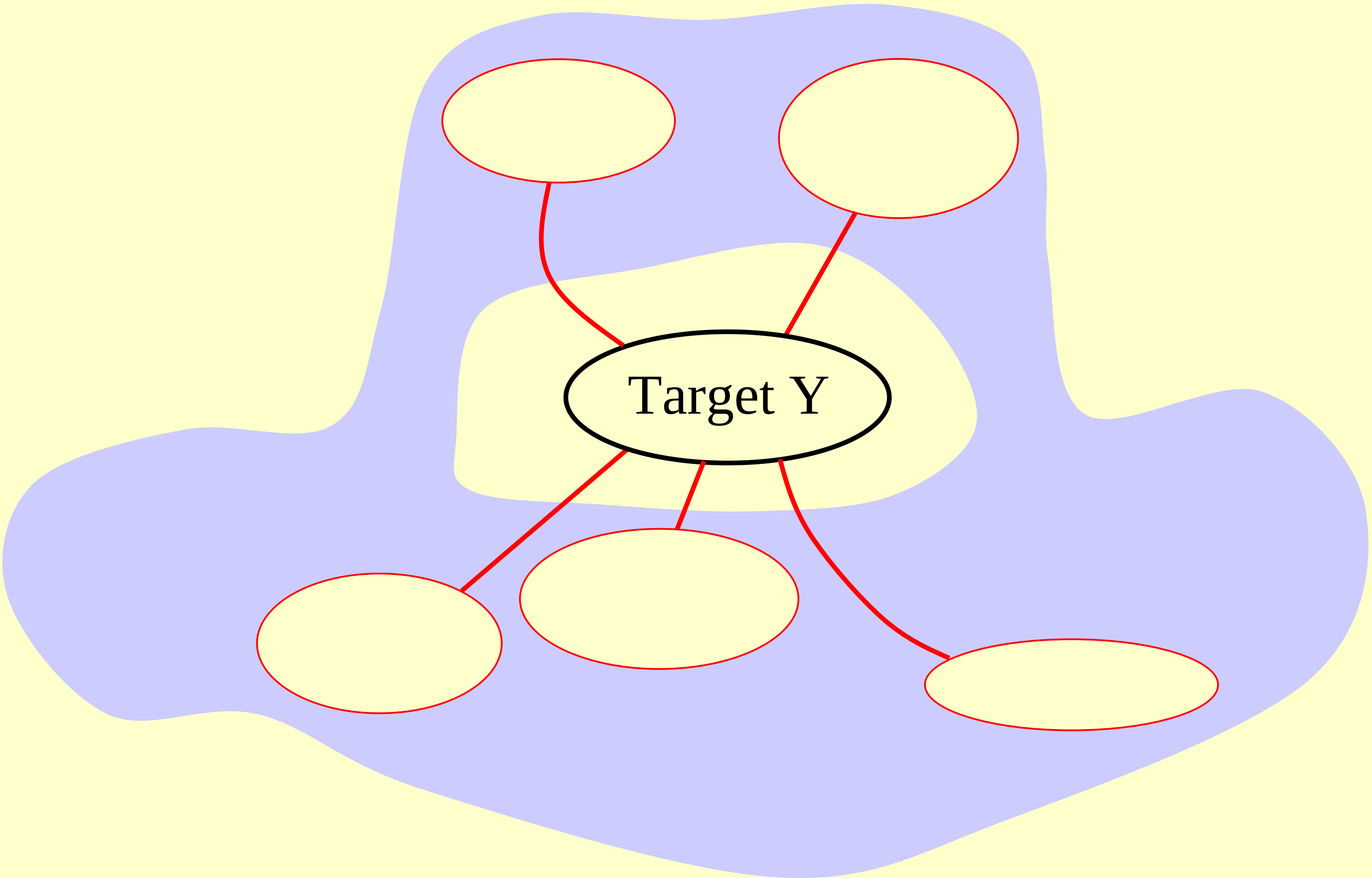
- Develop algorithms with good average- case performance, tractable for many real-life datasets.
- Abandon learning the full causal graph and instead develop methods that learn a local neighborhood.
- Abandon learning the fully oriented causal graph and instead develop methods that learn unoriented graphs.

A prototypical MB algo: HITON



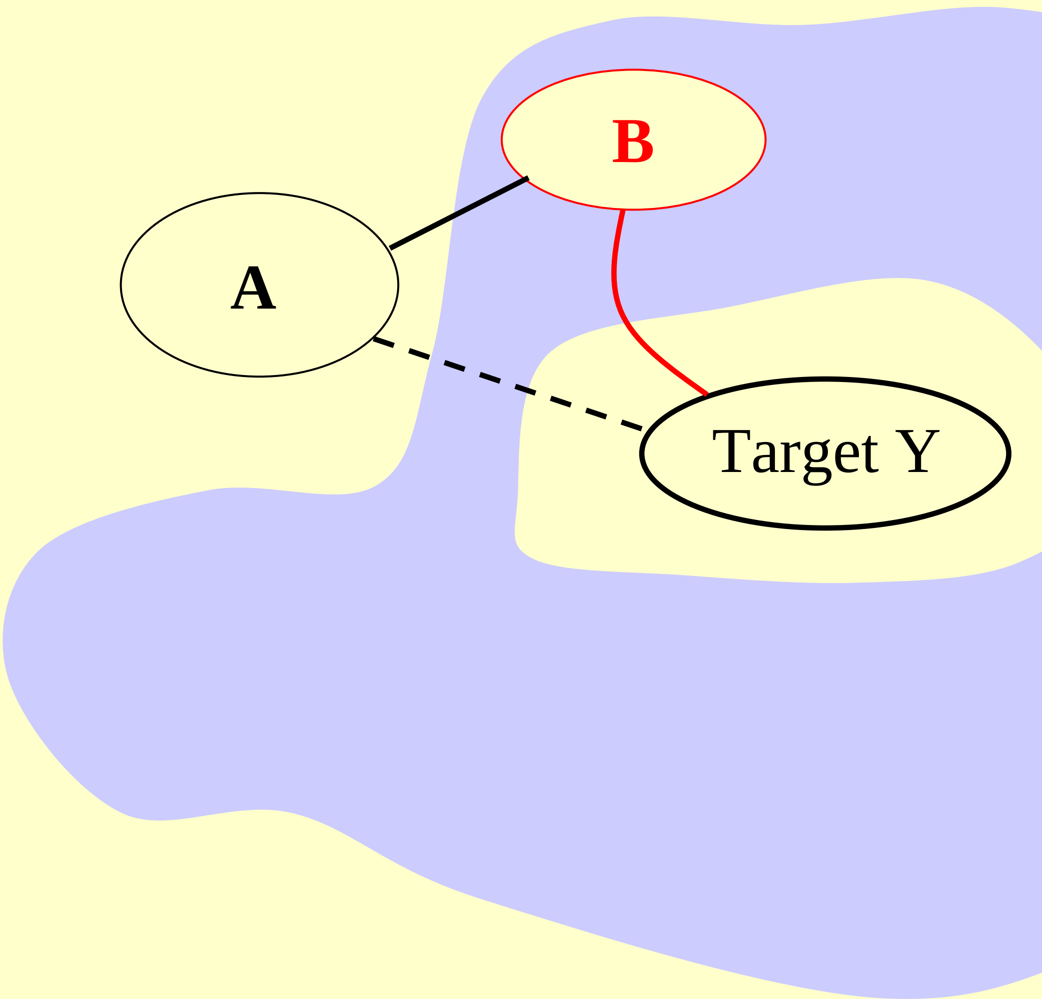
Aliferis-Tsamardinos-Statnikov, 2003)

1 – Identify variables with direct edges to the target (parent/children)



Aliferis-Tsamardinos-Statnikov, 2003)

1 – *Identify variables with direct edges to the target (parent/children)*



Iteration 1: add A

A

Iteration 2: add B

A **B**

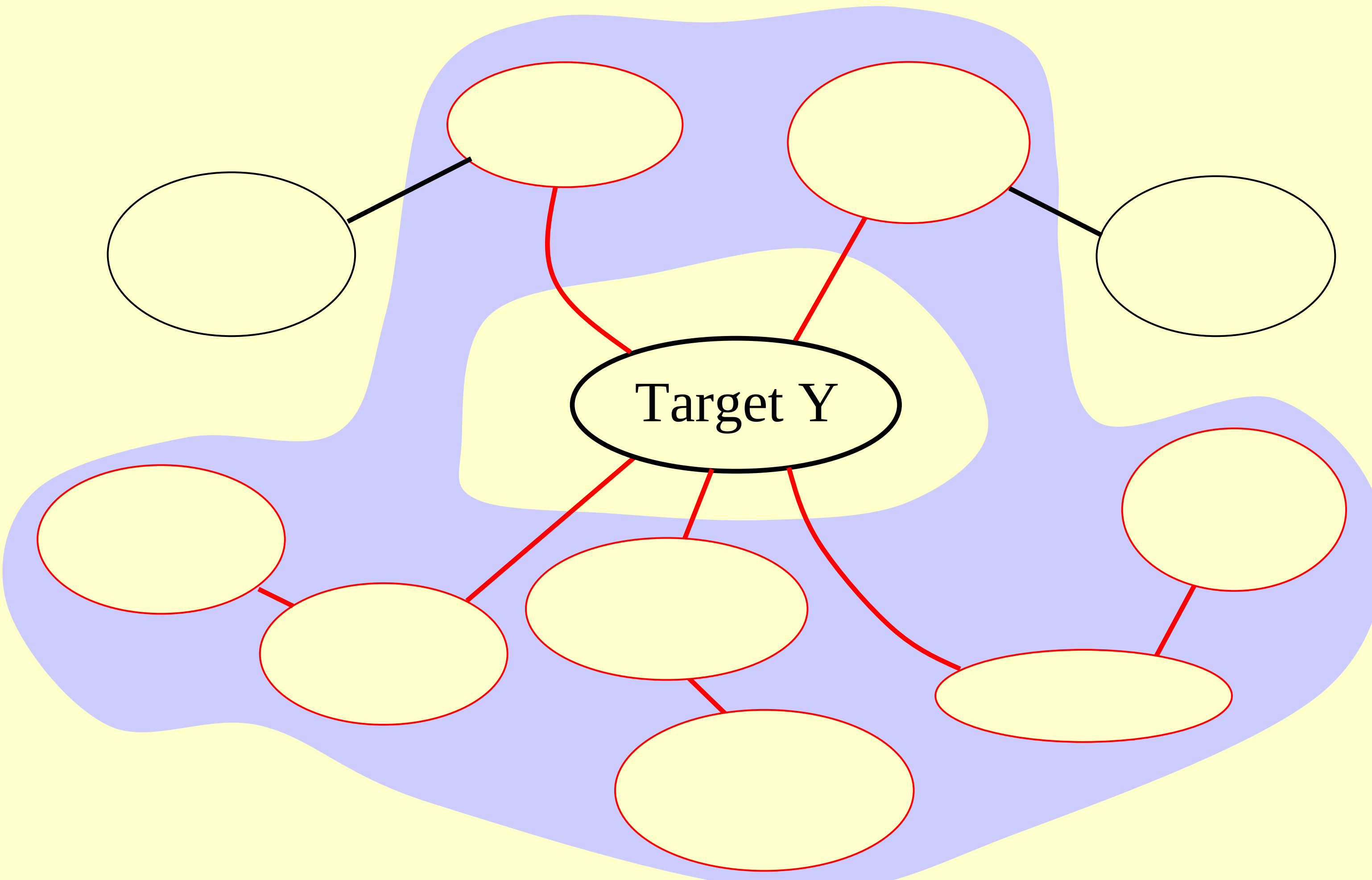
Iteration 3: remove B because $A \perp Y \mid B$

B

etc.

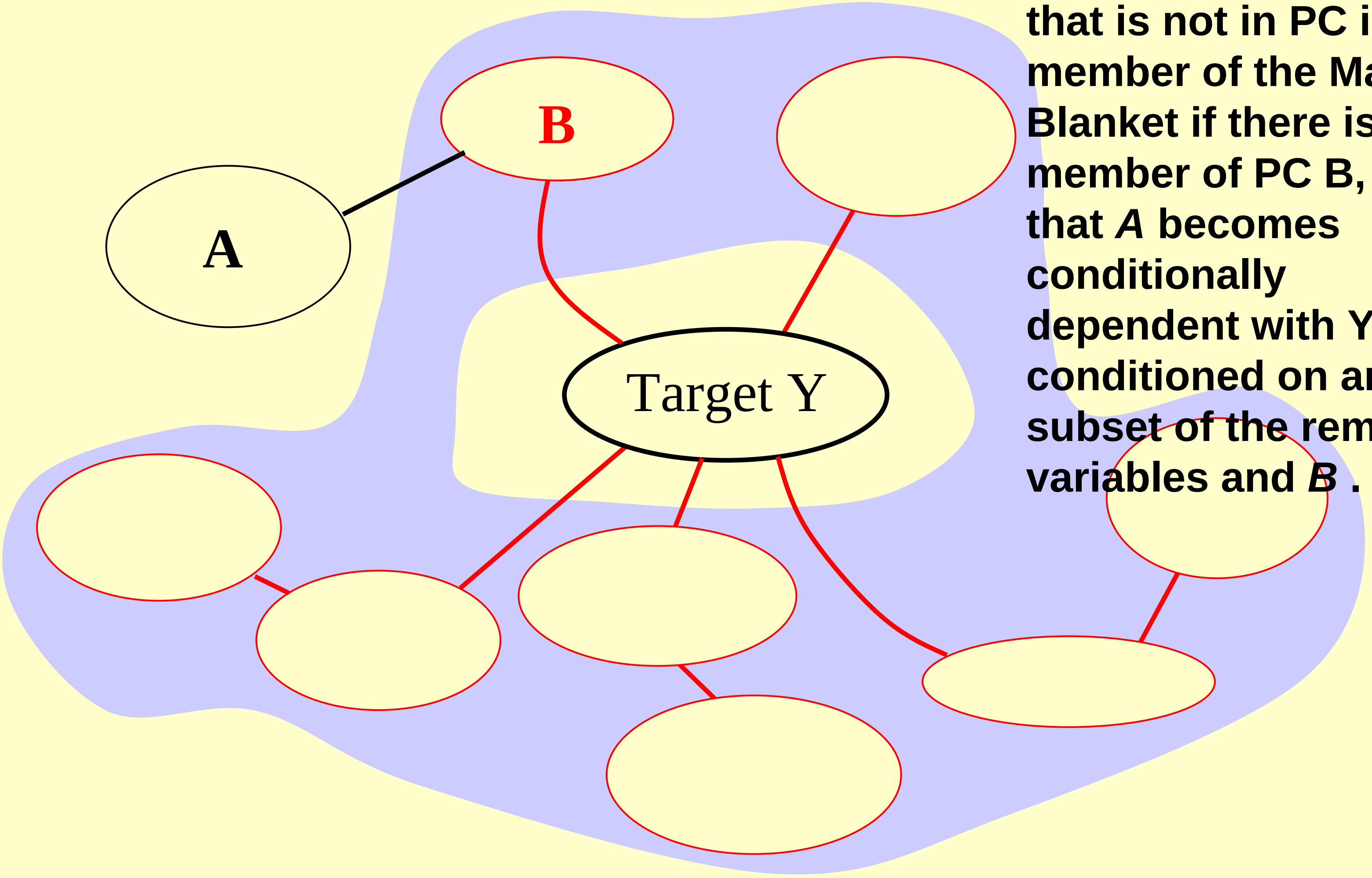
Aliferis-Tsamardinos-Statnikov, 2003)

2 – Repeat algorithm for parents and children of Y *(get depth two relatives)*



Aliferis-Tsamardinos-Statnikov, 2003)

3 – Remove non-members of the MB



A member *A* of PCPC that is not in PC is a member of the Markov Blanket if there is some member of PC B, such that *A* becomes conditionally dependent with Y conditioned on any subset of the remaining variables and *B* .

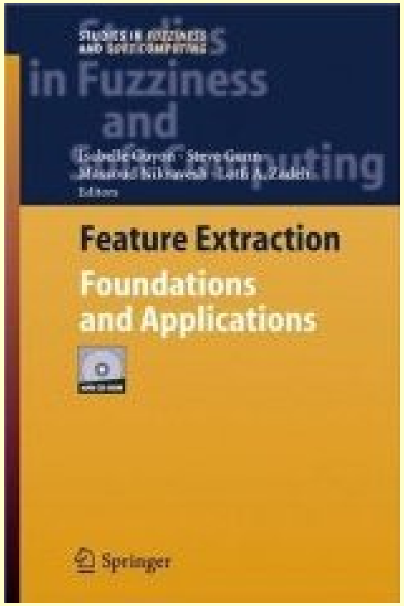
Aliferis-Tsamardinos-Statnikov, 2003)

Conclusion

- Feature selection focuses on uncovering subsets of variables X_1, $X_{2,}$ … predictive of the target Y.
- Multivariate feature selection is in principle more powerful than univariate feature selection, but not always in practice.
- Taking a closer look at the type of dependencies in terms of causal relationships may help refining the notion of variable relevance.

Acknowledgements and references

1) **Feature Extraction, Foundations and Applications**
   I. Guyon et al, Eds.
   Springer, 2006.
   [***http://clopinet.com/fextract-book***](http://clopinet.com/fextract-book)



2) **Causal feature selection**
   I. Guyon, C. Aliferis, A. Elisseeff
   To appear in "Computational Methods of Feature Selection",
   Huan Liu and Hiroshi Motoda Eds.,
   Chapman and Hall/CRC Press, 2007.
   [***http://clopinet.com/isabelle/Papers/causalFS.pdf***](http://clopinet.com/isabelle/Papers/causalFS.pdf)