



*CBMS Conference -- Foundations of Causal  
Graphical Models and Structure Discovery*

## *Lecture 2*

# Preliminaries: Probability Theory & Probabilistic Graphical Models

Instructor: Kun Zhang

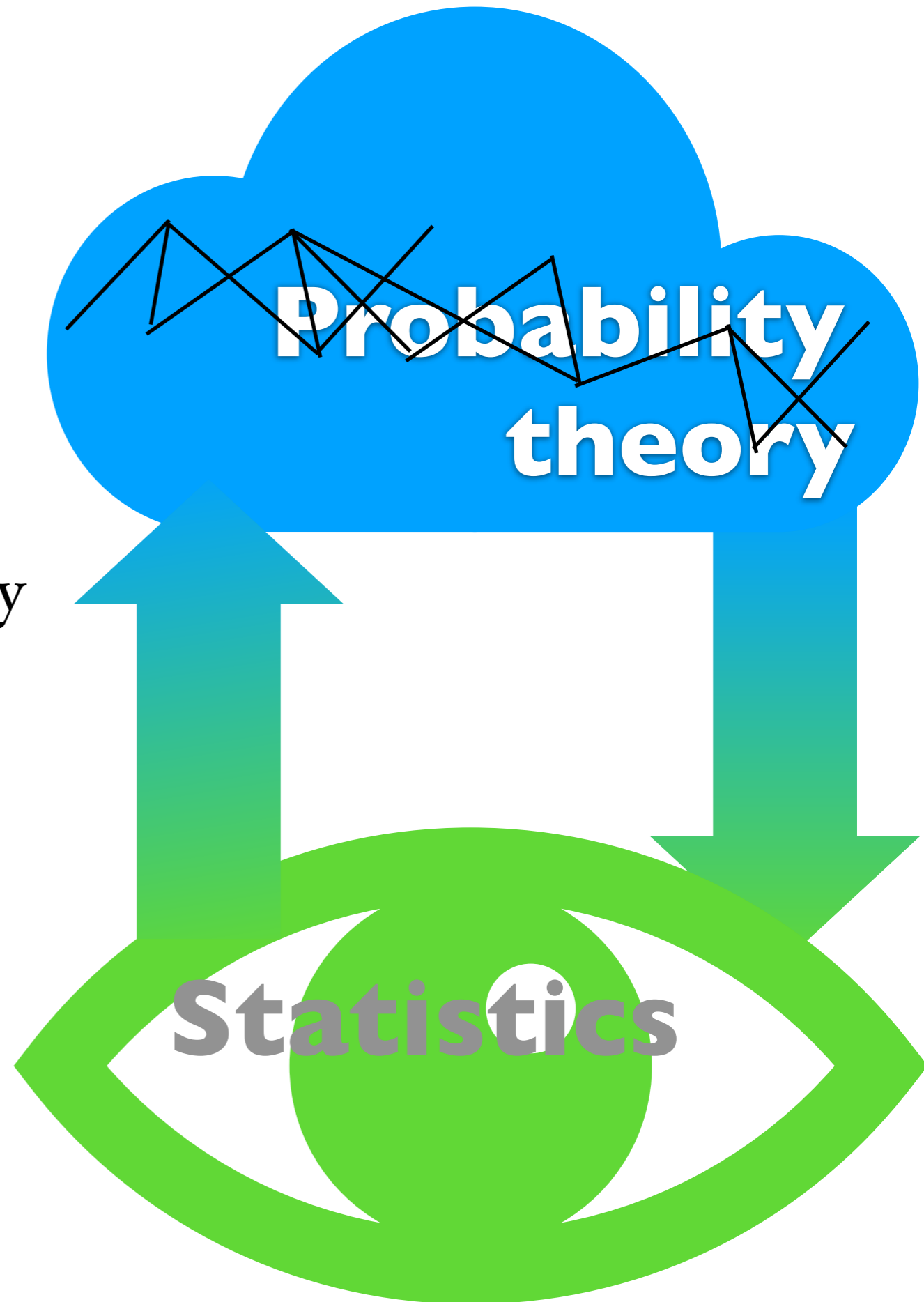
**Carnegie Mellon University**



MOHAMED BIN ZAYED  
UNIVERSITY OF  
ARTIFICIAL INTELLIGENCE

# Statistics...

- Relationship between probability theory & statistics



# Discrete vs. Continuous Random Variables



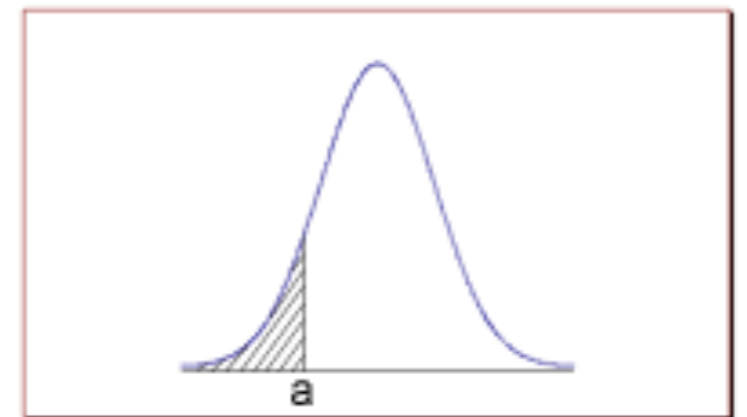
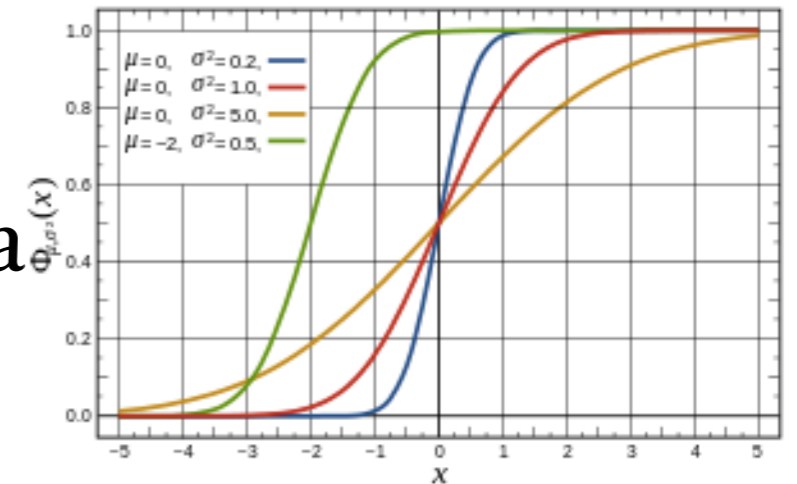
- A random variable is **discrete** if its range (the set of values that it can take) is finite or at most countably infinite
  - E.g., the sum of what I got on the two dice
  - $P(X=k) = P(\{\omega: X(\omega) = k\})$ ; tabular representation for the probability mass function (PMF)
- A random variable is **continuous** (not discrete) if its range (the set of values that it can take) is uncountably infinite
  - E.g., the height of a TAMU student
  - $P(a \leq X \leq b) = P(\{\omega: a \leq X(\omega) \leq b\})$

# How to Specify Prob. Measures of Random Variables

- PMFs for *discrete* variables
- Cumulative distribution function (CDF):  
A function  $F_X: \mathbb{R} \rightarrow [0,1]$  which specifies a probability measure as
- Probability density function (PDF):  
derivative of the CDF for *continuous* variables whose CDFs are differentiable everywhere

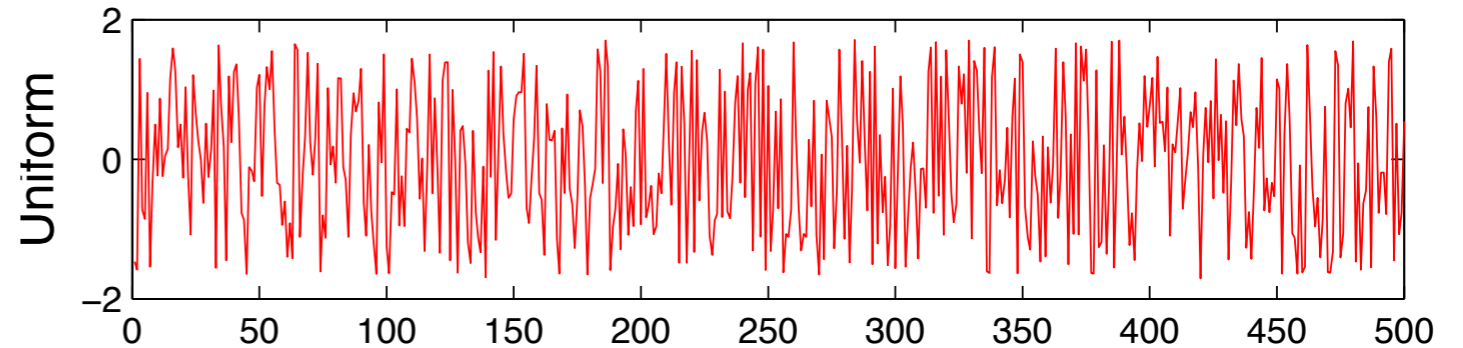
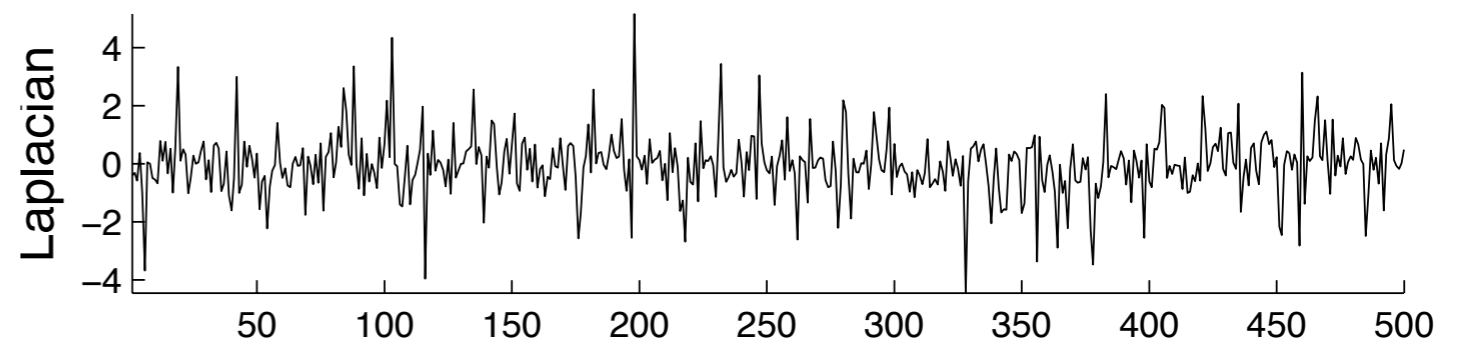
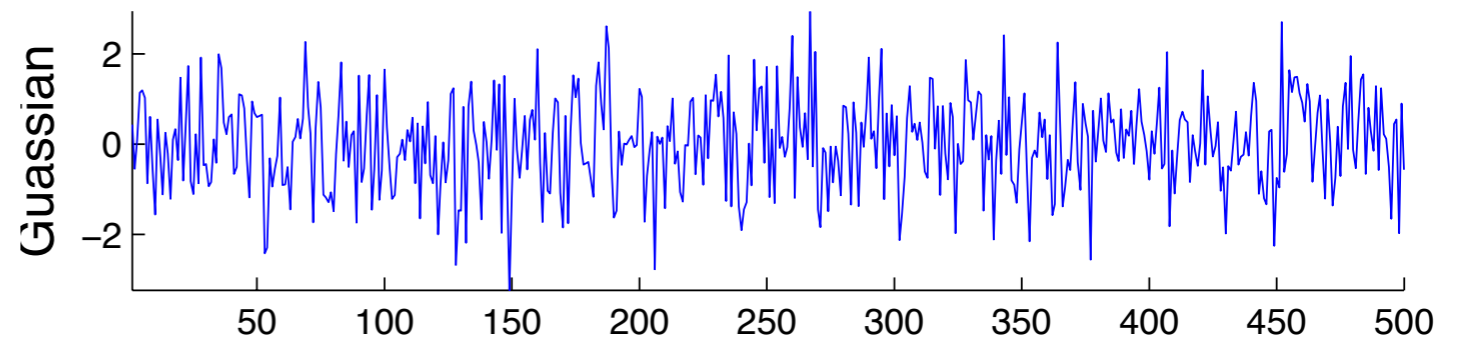
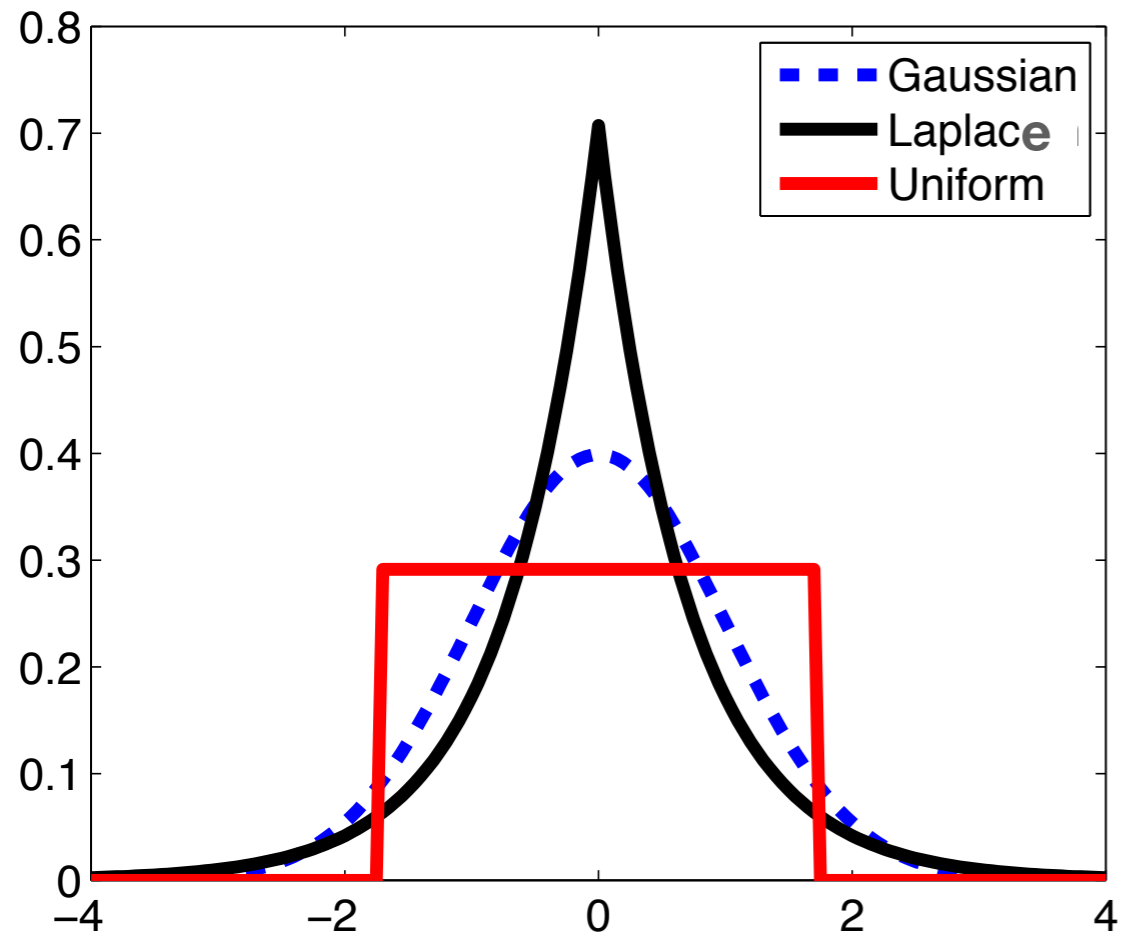
$$F_X(x) \triangleq P(X \leq x)$$

$$p_X(x) \triangleq \frac{dF_X(x)}{dx}$$



# Some Distributions

Three distributions with zero mean and unit variance



# Conditional Distributions

- Joint/marginal PMFs, CDFs, and PDFs:  
*straightforward*
- What is the probability distribution over  $X$ , when we know  $Y$  must take a certain value  $y$ ?

- Discrete case: Provided  $P_Y(y) \neq 0$ , conditional PMF of  $X$  given  $Y$  is

$$P_{X|Y} = \frac{P_{XY}(x, y)}{P_Y(y)}$$

- Continuous case: Provided  $p_Y(y) \neq 0$ , conditional PDF of  $X$  given  $Y$  is

$$p_{X|Y} = \frac{p_{XY}(x, y)}{p_Y(y)}$$

# A Question...

- With 5 coins which are not necessarily fair, how many parameters to represent the joint probability distribution  $P(O_1, O_2, \dots, O_5)$ ?
- In practice we often need fewer parameters...
- Divide-and-conquer



# Statistical Independence

- Two variables  $X$  and  $Y$  are independent if  $F_{XY}(x,y) = F_X(x) F_Y(y)$  for all values of  $x$  and  $y$ . Equivalently,
- For **discrete** variables,  $P_{XY}(x,y) = P_X(x)P_Y(y)$ , or  $P_{X|Y}(x|y) = P_X(x)$  whenever  $P_Y(y) \neq 0$
- For **continuous** variables:  $p$  instead of  $P$



# Pairwise Independence vs. Mutual Independence

- Pairwise independent: every pair of random variables is independent
- Mutually independent:  $F_{X_1 X_2 \dots X_n}(x, y) = F_{X_1}(x_1) F_{X_2}(x_2) \dots F_{X_n}(x_n)$
- Example of three coins:  $A \perp\!\!\!\perp B$ ; C is determined by A and B but  $C \perp\!\!\!\perp B$  and  $C \perp\!\!\!\perp A$

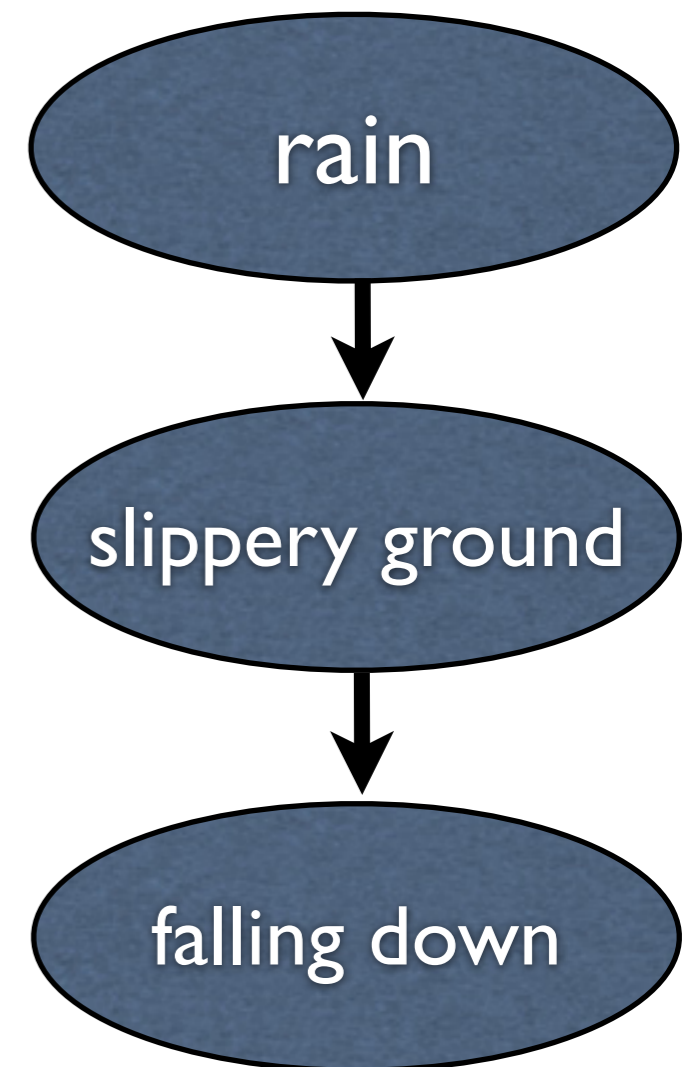
# Ways to Produce Dependence

- Common cause underlying them
- causal relations between them
- Selection (conditioning on the effect)!

# Another Example

- What if  $X_i$ 's are not mutually independent but we know they were generated the following way?

$$X_1 \rightarrow X_2 \rightarrow \dots \rightarrow X_n$$



# Conditional Independence

- Two variables  $X$  and  $Y$  are **conditionally independent given  $Z$**  if  $F_{XY|Z}(x,y|z) = F_{X|Z}(x|z) F_{Y|Z}(y|z)$  for all values of  $x$ ,  $y$  and  $z$ . Equivalently,
  - For discrete variables,  $P_{XY|Z}(x,y|z) = P_{X|Z}(x|z)P_{Y|Z}(y|z)$ , or  $P_{X|Y,Z}(x|y,z) = P_{X|Z}(x|z)$  whenever  $P_{YZ}(y,z) \neq 0$
  - For continuous variables...
- $X \perp\!\!\!\perp Y \mid Z$ : If  $Z$  is known,  $Y$  is not useful when modeling/predicting  $X$

# Some Properties of (Conditional) Independence

- Symmetry

$$X \perp\!\!\!\perp Y \Rightarrow Y \perp\!\!\!\perp X$$

- Decomposition

$$X \perp\!\!\!\perp (A, B) \Rightarrow \text{and } \begin{cases} X \perp\!\!\!\perp A \\ X \perp\!\!\!\perp B \end{cases}$$

- Weak union

$$X \perp\!\!\!\perp (A, B) \Rightarrow \text{and } \begin{cases} X \perp\!\!\!\perp A \mid B \\ X \perp\!\!\!\perp B \mid A \end{cases}$$

- Contraction

$$\left. \begin{array}{l} X \perp\!\!\!\perp A \mid B \\ X \perp\!\!\!\perp B \end{array} \right\} \text{and } \Rightarrow X \perp\!\!\!\perp (A, B)$$

*Relationship between independence & conditional independence?*

# Some Properties of (Conditional) Independence

$$P(A, B | X) = P(A, B)$$

$\Rightarrow P(A | X) = P(A)$  (by marginalizing  $B$  out)

- Symmetry

$$X \perp\!\!\!\perp Y \Rightarrow Y \perp\!\!\!\perp X$$

- Decomposition

$$X \perp\!\!\!\perp (A, B) \Rightarrow \text{and } \begin{cases} X \perp\!\!\!\perp A \\ X \perp\!\!\!\perp B \end{cases}$$

- Weak union

$$X \perp\!\!\!\perp (A, B) \Rightarrow \text{and } \begin{cases} X \perp\!\!\!\perp A | B \\ X \perp\!\!\!\perp B | A \end{cases}$$

- Contraction

$$\left. \begin{array}{l} X \perp\!\!\!\perp A | B \\ X \perp\!\!\!\perp B \end{array} \right\} \text{and } \Rightarrow X \perp\!\!\!\perp (A, B)$$

*Relationship between independence & conditional independence?*

# Some Properties of (Conditional) Independence

$$P(X|A,B) = P(X);$$

$$P(X|A) = P(X).$$

$$\Rightarrow P(X|A,B) = P(X|A), \text{ i.e., } X \perp\!\!\!\perp B|A$$

- Symmetry

$$X \perp\!\!\!\perp Y \Rightarrow Y \perp\!\!\!\perp X$$

- Decomposition

$$X \perp\!\!\!\perp (A, B) \Rightarrow \text{and } \begin{cases} X \perp\!\!\!\perp A \\ X \perp\!\!\!\perp B \end{cases}$$

- Weak union

$$X \perp\!\!\!\perp (A, B) \Rightarrow \text{and } \begin{cases} X \perp\!\!\!\perp A | B \\ X \perp\!\!\!\perp B | A \end{cases}$$

- Contraction

$$\left. \begin{array}{l} X \perp\!\!\!\perp A | B \\ X \perp\!\!\!\perp B \end{array} \right\} \text{and } \Rightarrow X \perp\!\!\!\perp (A, B)$$

*Relationship between independence & conditional independence?*

# Covariance and Correlation

- Covariance:  $Cov[X, Y] \triangleq E[(X - E[X])(Y - E[Y])]$
- Uncorrelated if  $Cov[X, Y] = 0$
- Correlation:  $Corr[X, Y] \triangleq \frac{Cov[X, Y]}{\sqrt{Var[X]Var[Y]}}$



# Independence and Uncorrelatedness

- Independence  $\Rightarrow$  uncorrelatedness
- How about the reverse direction?

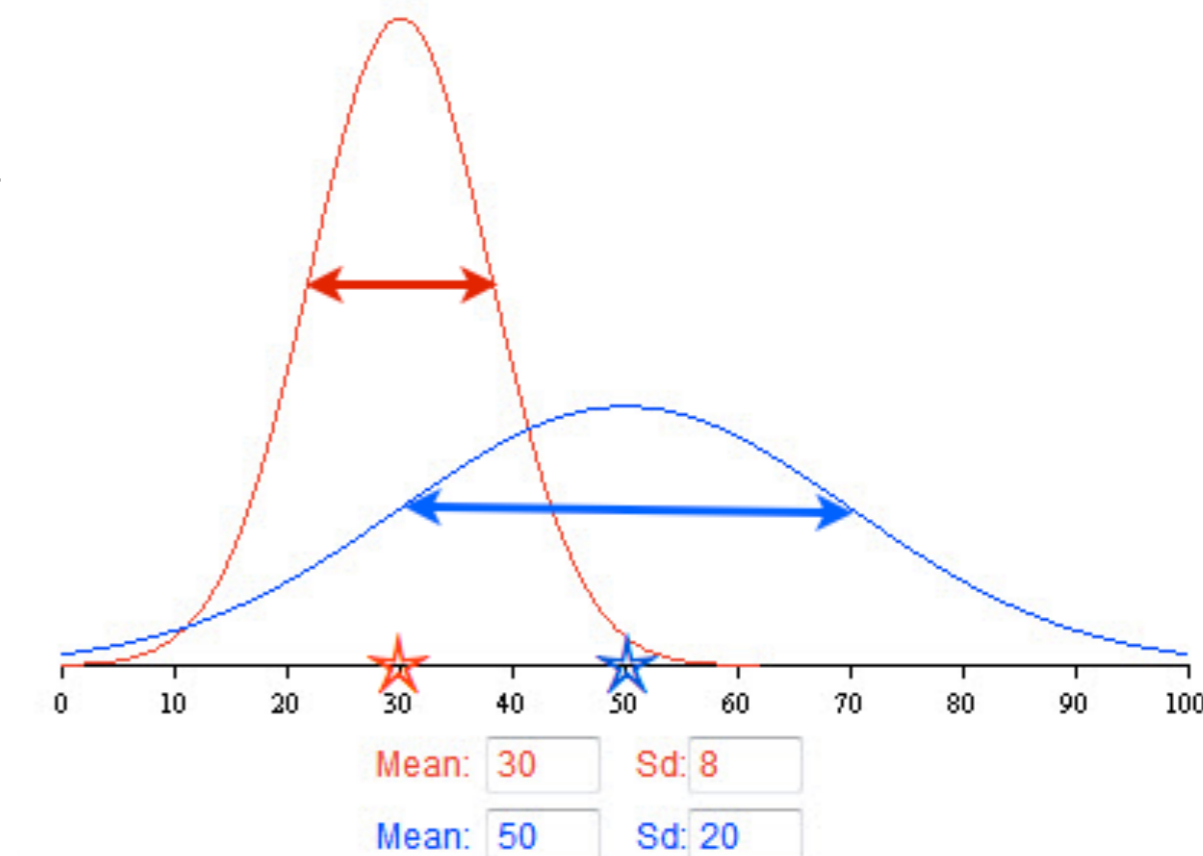
*Normal distribution !*

# Normal Distribution

- Very common distribution (sometimes also informally known as bell curve)
- PDF specified by mean  $\mu$  and standard deviation  $\sigma$  (or variance  $\sigma^2$ ):

$$p_X(x | \mu, \sigma) = \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Often denoted by  $N(\mu, \sigma^2)$

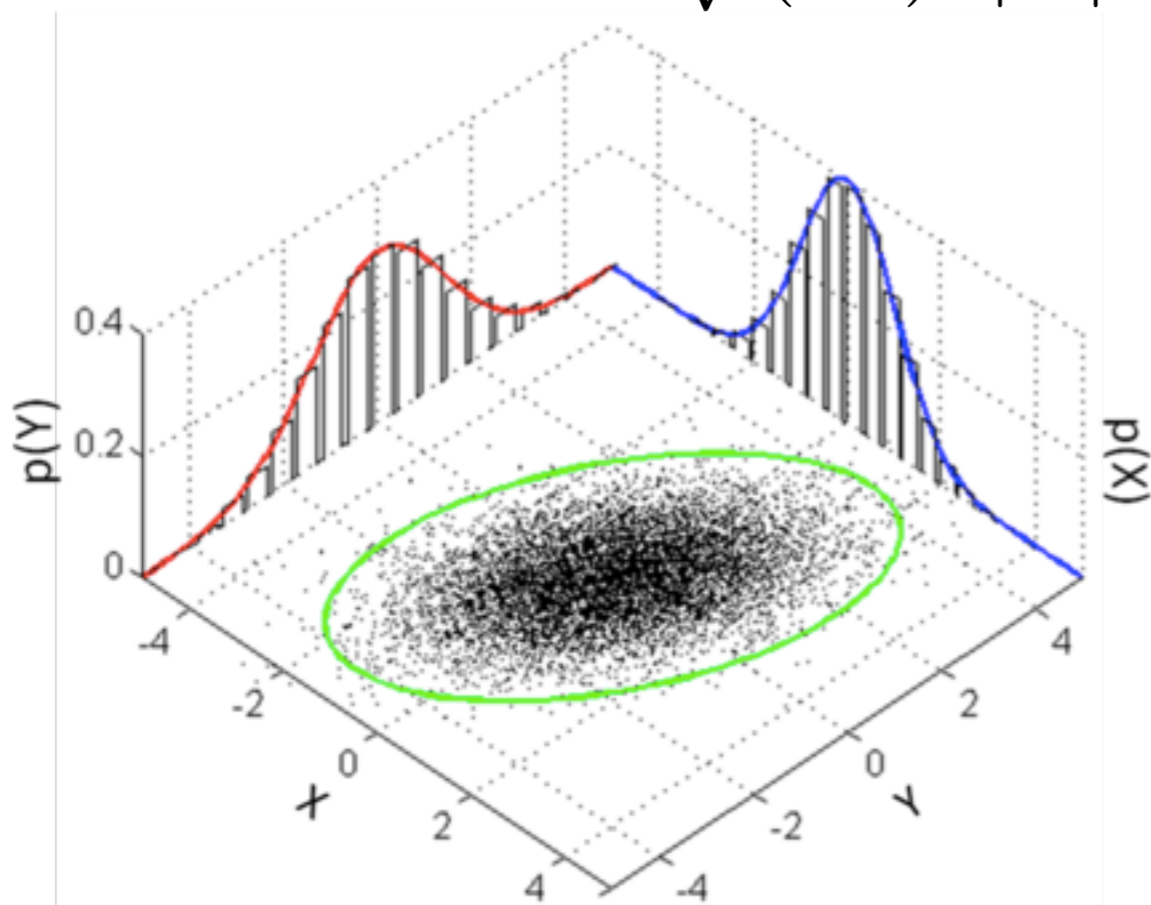


# Multivariate Normal Distribution

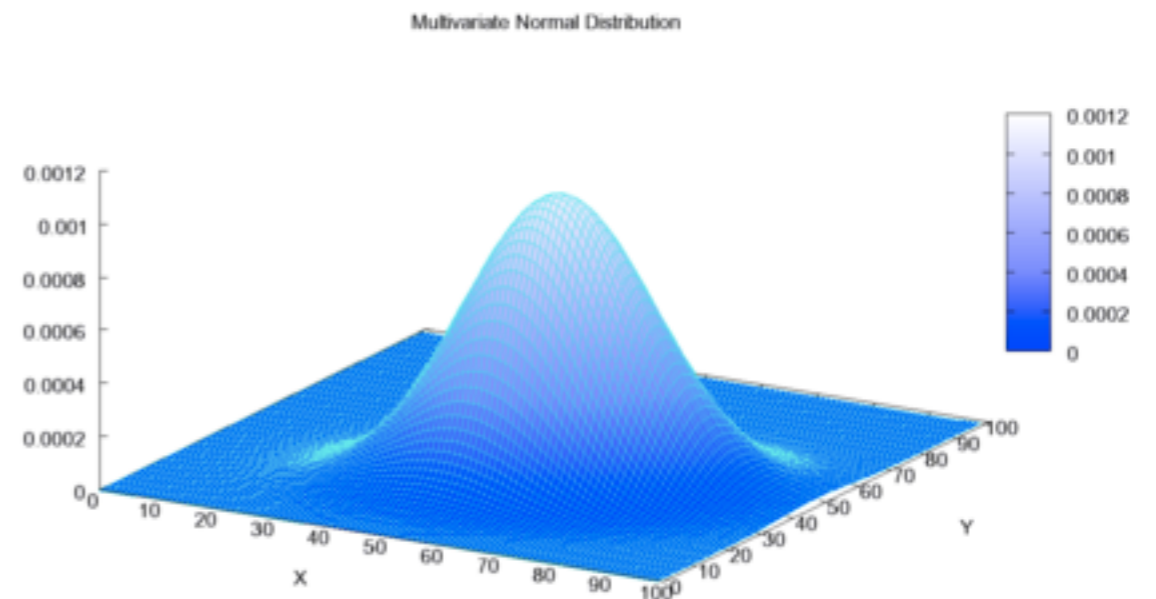
$$\Sigma = \begin{bmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) \\ \text{Cov}(X_2, X_1) & \text{Var}(X_2) \end{bmatrix}$$

- PDF for point  $\mathbf{x} = (x_1, \dots, x_k)$ , specified by mean  $\boldsymbol{\mu}$  and covariance matrix :

$$p_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$$



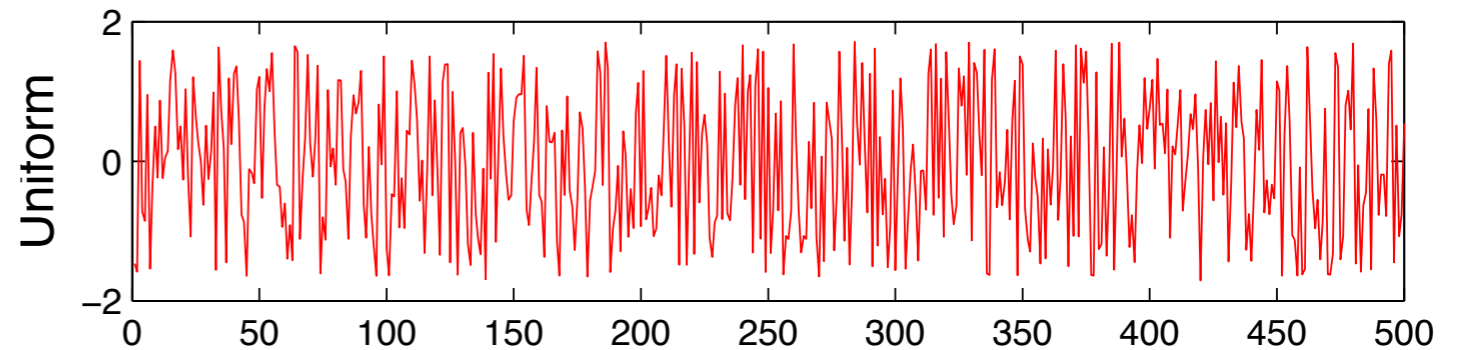
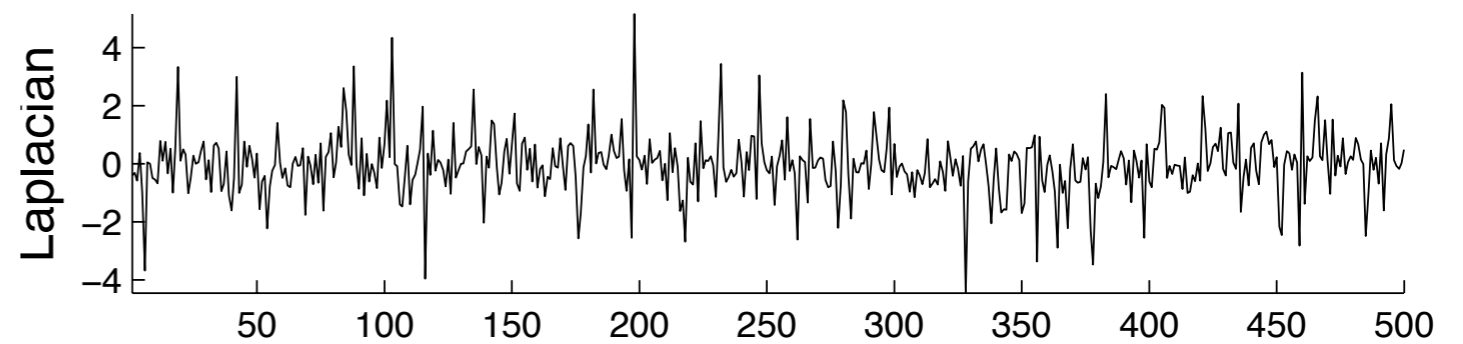
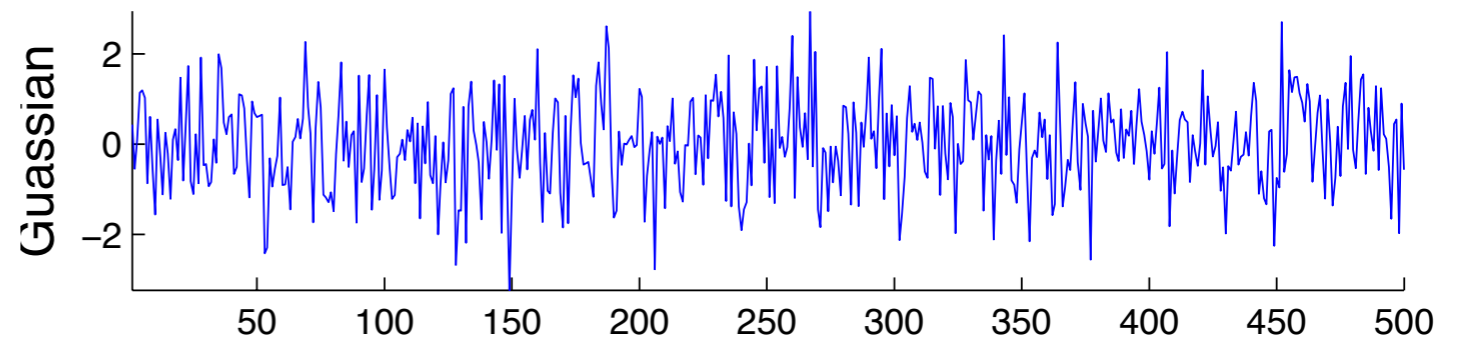
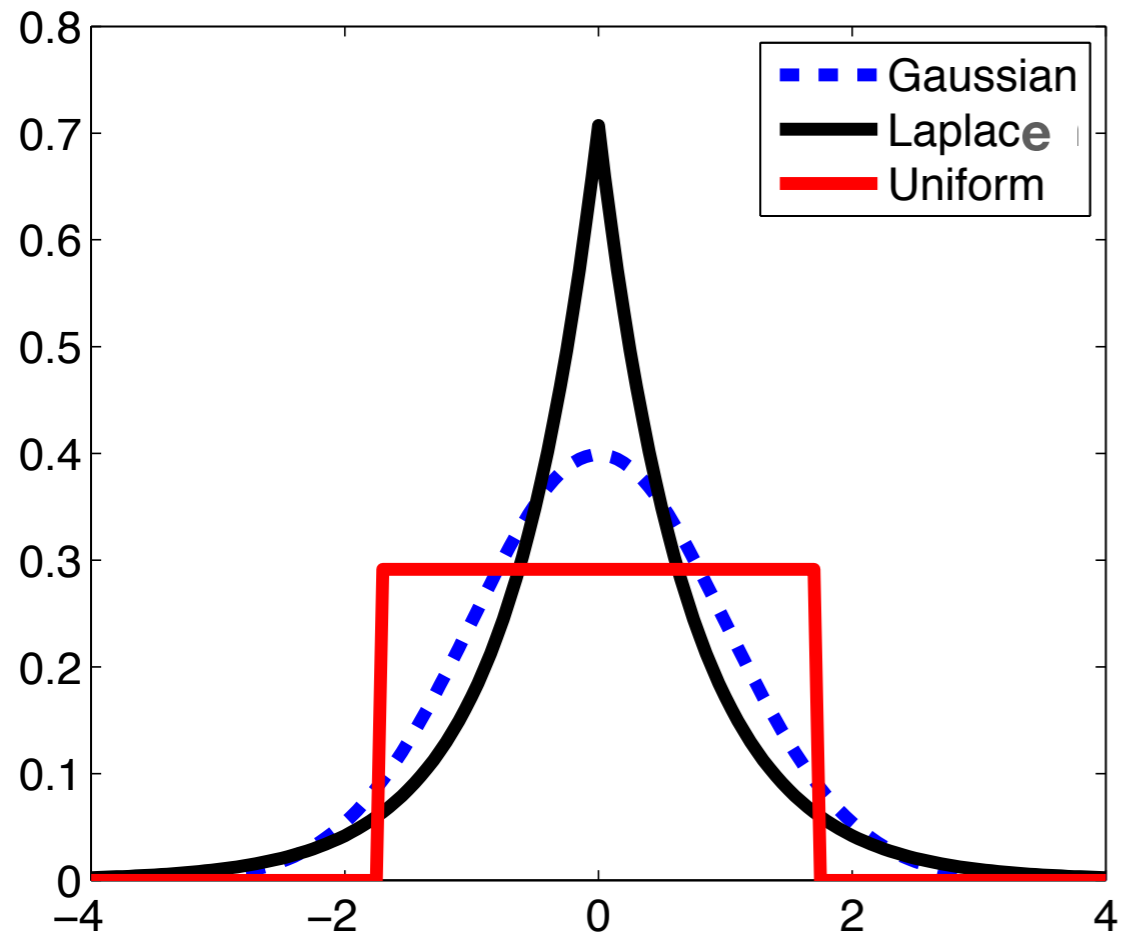
*Sample & marginal*



*pdf*

# Some Distributions

Three distributions with zero mean and unit variance



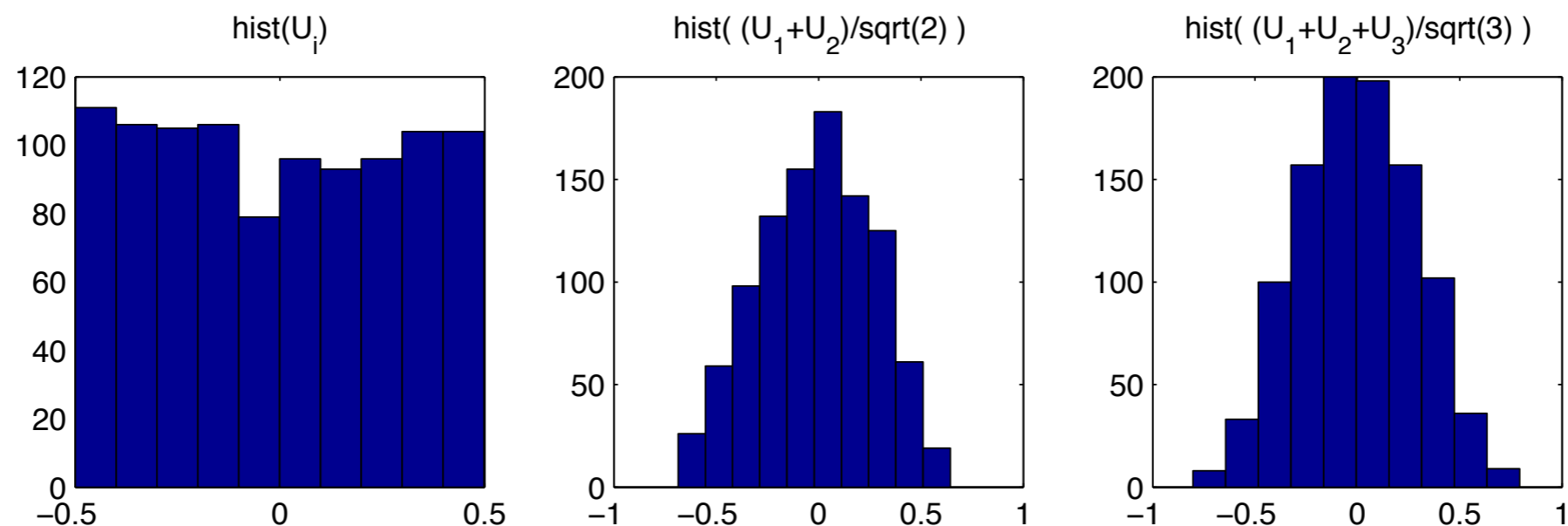
# Some Properties of Normal Distributions

- Simplicity
- Uncorrelatedness implies independence
- Approximately holds in many cases because of *central limit theorem* (CLT)
- CLT: Under some conditions,  $S = \frac{1}{n} \sum_{i=1}^n X_i$  converges to a normal distribution for independent  $X_i$  with finite mean and variance
- Are they really normal? Cramer's decomposition theorem!

*Interested students may refer to Chapter 7 of  
"Probability theory: The logic of science"*

# Central Limit Theorem: An Illustration

- CLT: Under some conditions,  $S = \frac{1}{n} \sum_{i=1}^n X_i$  converges to a normal distribution for independent  $X_i$  with finite mean and variance



- Are they really normal? Cramer's decomposition theorem!  
*E. T. Jaynes. Probability Theory: The Logic of Science. 1994. Chapter 7.*



## THE CENTRAL GAUSSIAN, OR NORMAL, DISTRIBUTION

*“My own impression . . . is that the mathematical results have outrun their interpretation and that some simple explanation of the force and meaning of the celebrated integral . . . will one day be found . . . which will at once render useless all the works hitherto written.”* - - - Augustus de Morgan (1838)

Here, de Morgan was expressing his bewilderment at the “curiously ubiquitous” success of methods of inference based on the gaussian, or normal, “error law” (sampling distribution), even in cases where the law is not at all plausible as a statement of the actual frequencies of the errors. But the explanation was not forthcoming as quickly as he expected.

In the middle 1950’s the writer heard an after-dinner speech by Professor Willy Feller, in which he roundly denounced the practice of using gaussian *probability* distributions for errors, on the grounds that the *frequency* distributions of real errors are almost never gaussian. Yet in spite of Feller’s disapproval, we continued to use them, and their ubiquitous success in parameter estimation continued. So 145 years after de Morgan’s remark the situation was still unchanged, and the same surprise was expressed by George Barnard (1983): *“Why have we for so long managed with normality assumptions?”*

Today we believe that we can, at last, explain (1) the inevitably ubiquitous use and (2) the ubiquitous success, of the gaussian error law. Once seen, the explanation is indeed yet to the best of our knowledge it is not recognized in any of the previous literature because of the universal tendency to think of probability distributions in terms of frequencies. We cannot understand what is happening until we learn to think of probability distributions in terms of their demonstrable *information content* instead of their imagined (and irrelevant) frequency connections.

A simple explanation of these properties – stripped of past irrelevancies – has been achieved only very recently, and this development changed our plans for the present work. We decided that it is so important that it should be inserted at this somewhat early point in the narrative, even though we must then appeal to some results that are established only later. In the present Chapter, then, we survey the historical basis of gaussian distributions and get a quick preliminary understanding of their functional role in inference. This understanding will then guide us directly – without the usual false starts and blind alleys – to the computational procedures which yield the great majority of the useful applications of probability theory.

*Interested students may refer to Chapter 7 of “Probability theory: The logic of science”*



# Three Ways to Derive Gaussian PDFs

- Found by de Moivre (1733), without realizing its importance
- Independence + isotropy (Herschel 1785)
- Maximum likelihood estimate = arithmetic mean (Gauss, 1809)
- Stability in its form under small perturbation (Landon, 1941)

*Interested students may refer to Chapter 7 of  
“Probability theory: The logic of science”*



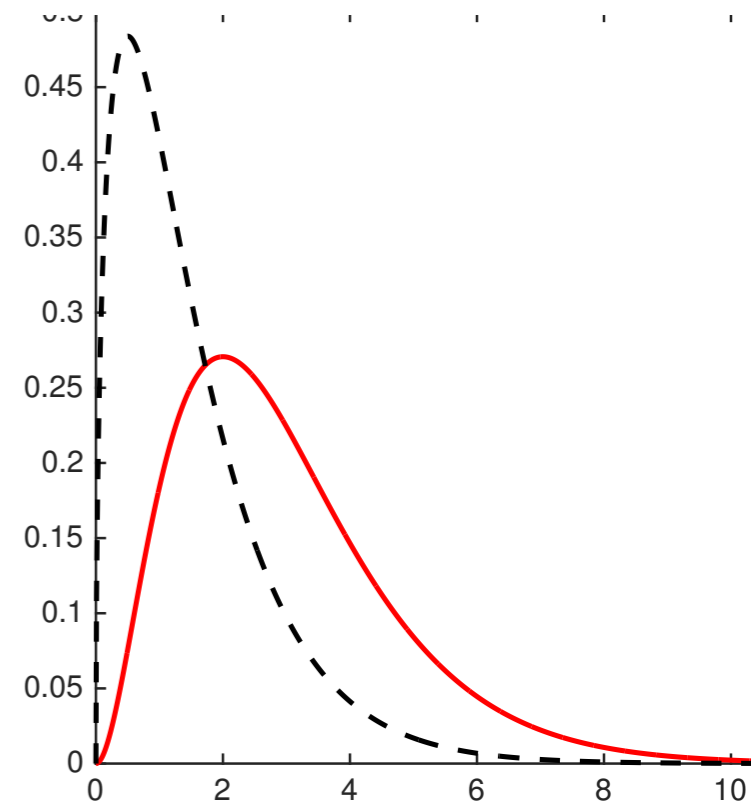
# Distance Between Distributions: Are Two Distributions the Same?

- Kullback-Leibler divergence:

$$D_{\text{KL}}(P\|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}.$$

$$D_{\text{KL}}(p(x)\|q(x)) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx.$$

- Non-negative; asymmetric; zero iff identical



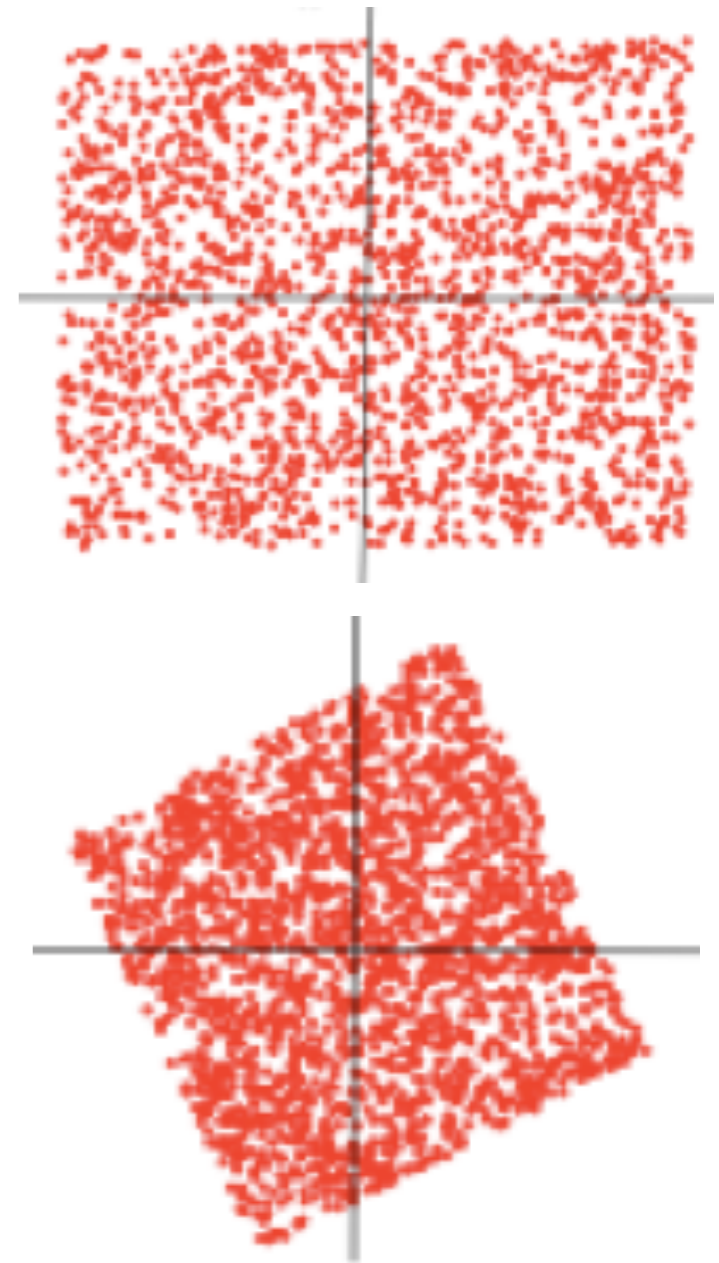
# Are Two Variables Independent?

- Natural measure of statistical dependence: mutual information

$$I(X; Y) = \sum_y \sum_x P(x, y) \log \left( \frac{P(x, y)}{P(x) P(y)} \right),$$

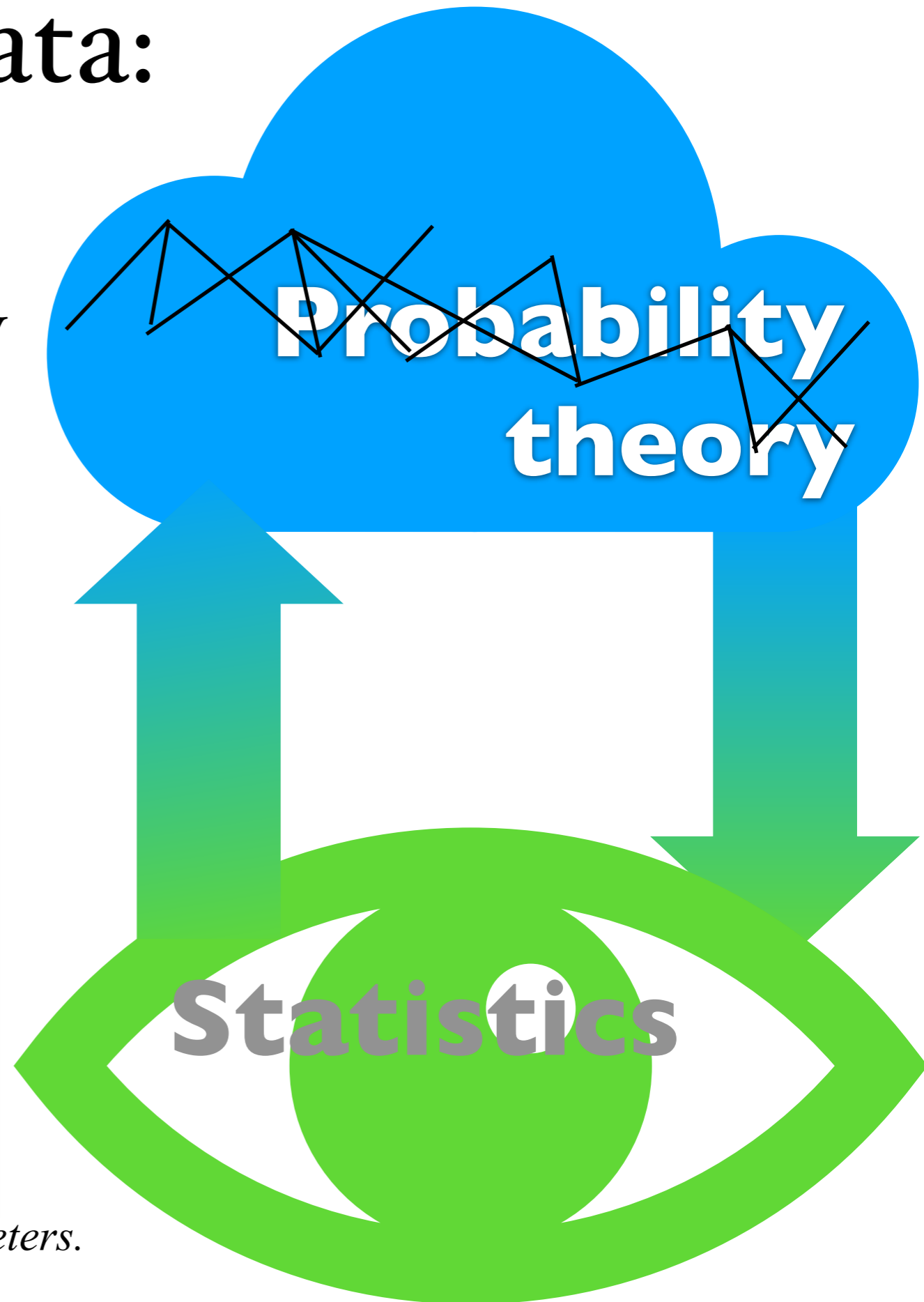
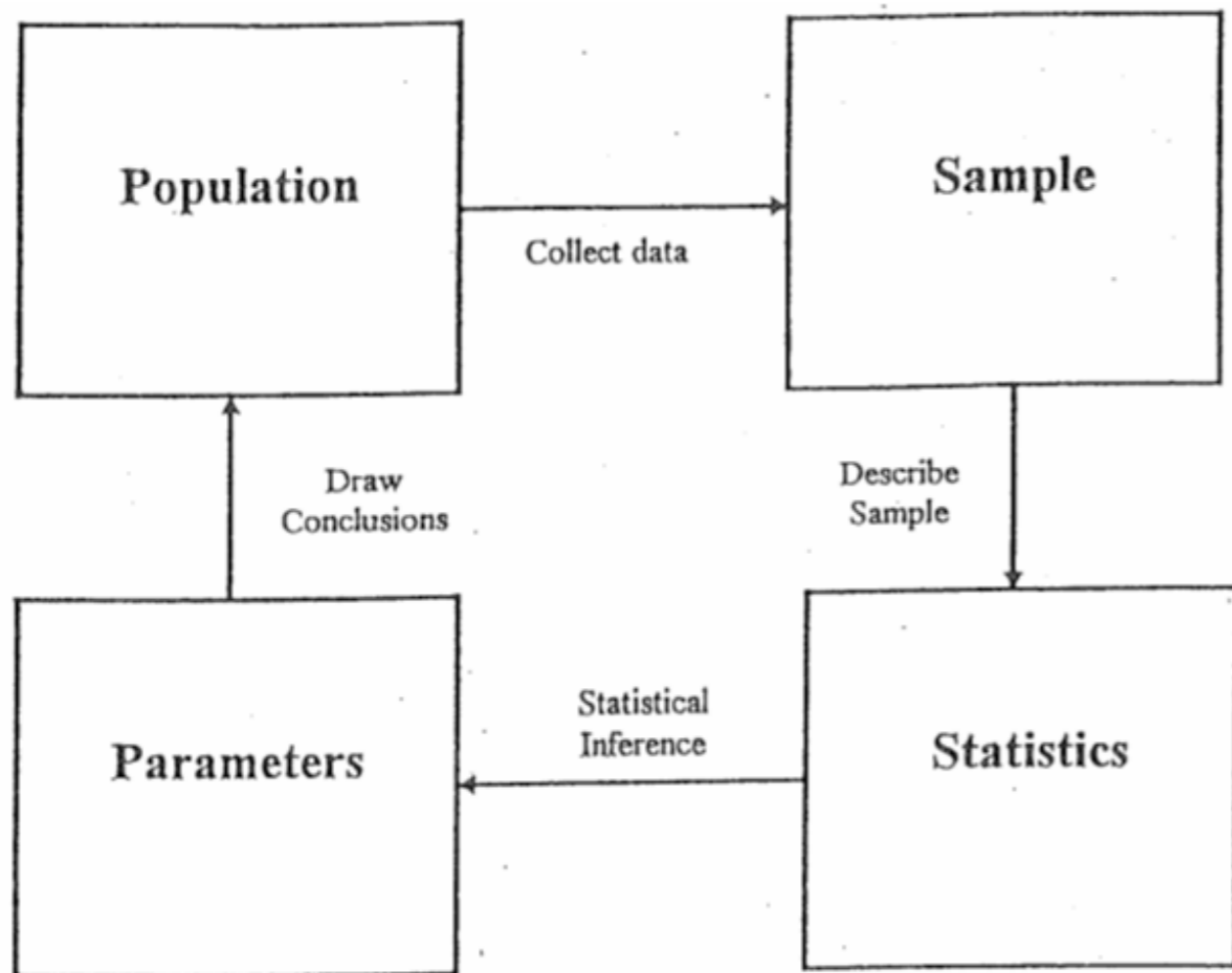
$$I(X; Y) = \int \int p(x, y) \log \left( \frac{p(x, y)}{p(x) p(y)} \right) dx dy,$$

- Non-negative; is zero iff  $X$  and  $Y$  are independent



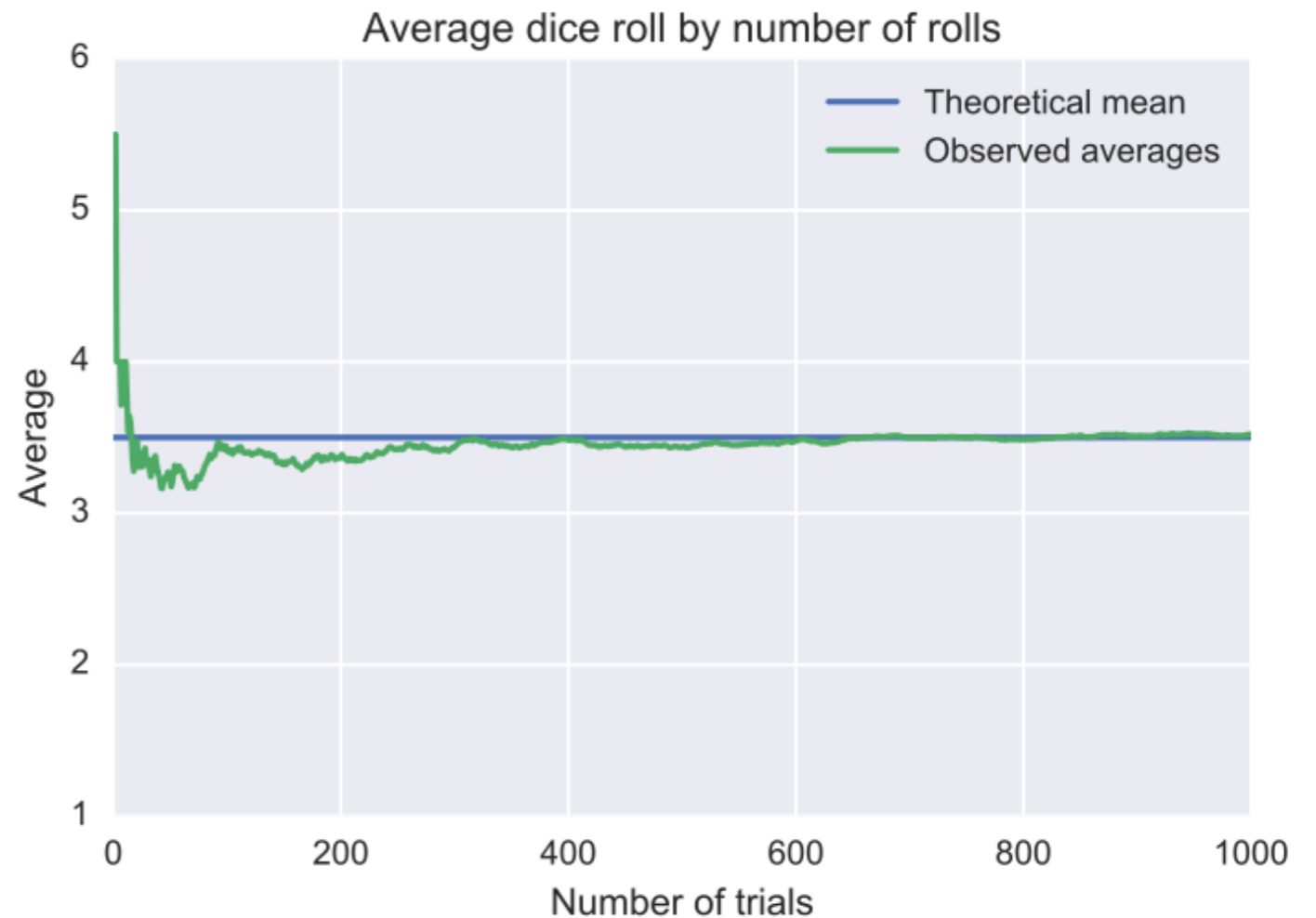
# Making Use of Data: Statistics...

- Relationship between probability theory & statistics



*Using sample statistics to estimate population parameters.*

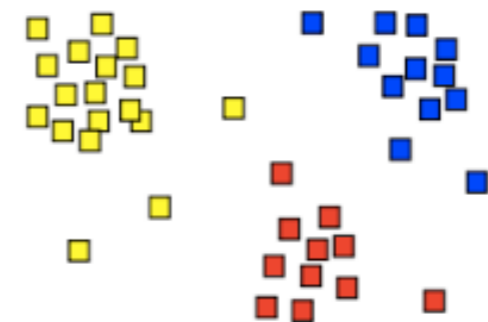
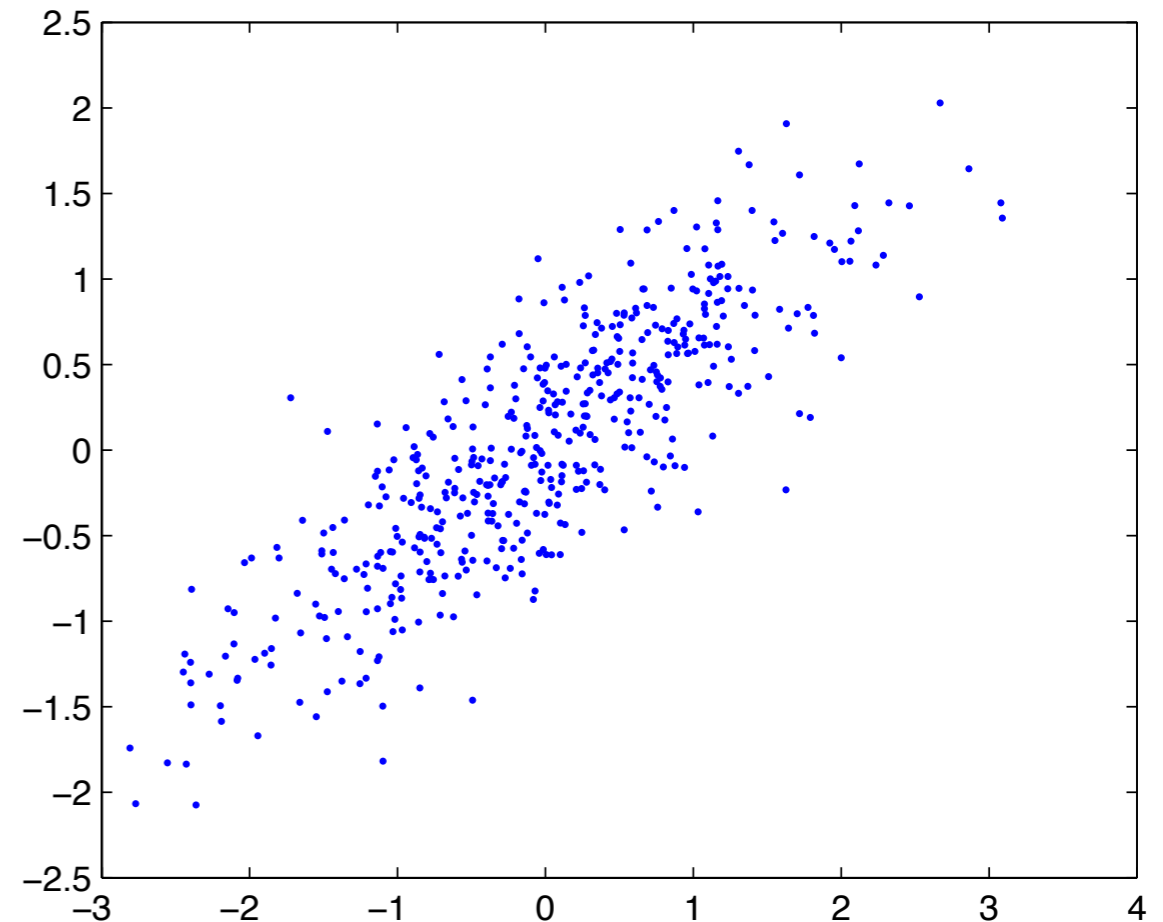
# Law of Large Numbers



- Law of large numbers (LLN) is a theorem that describes the result of performing the same experiment a large number of times: the average of the results obtained from a large number of trials should be close to the expected value, and will tend to become closer as more trials are performed.

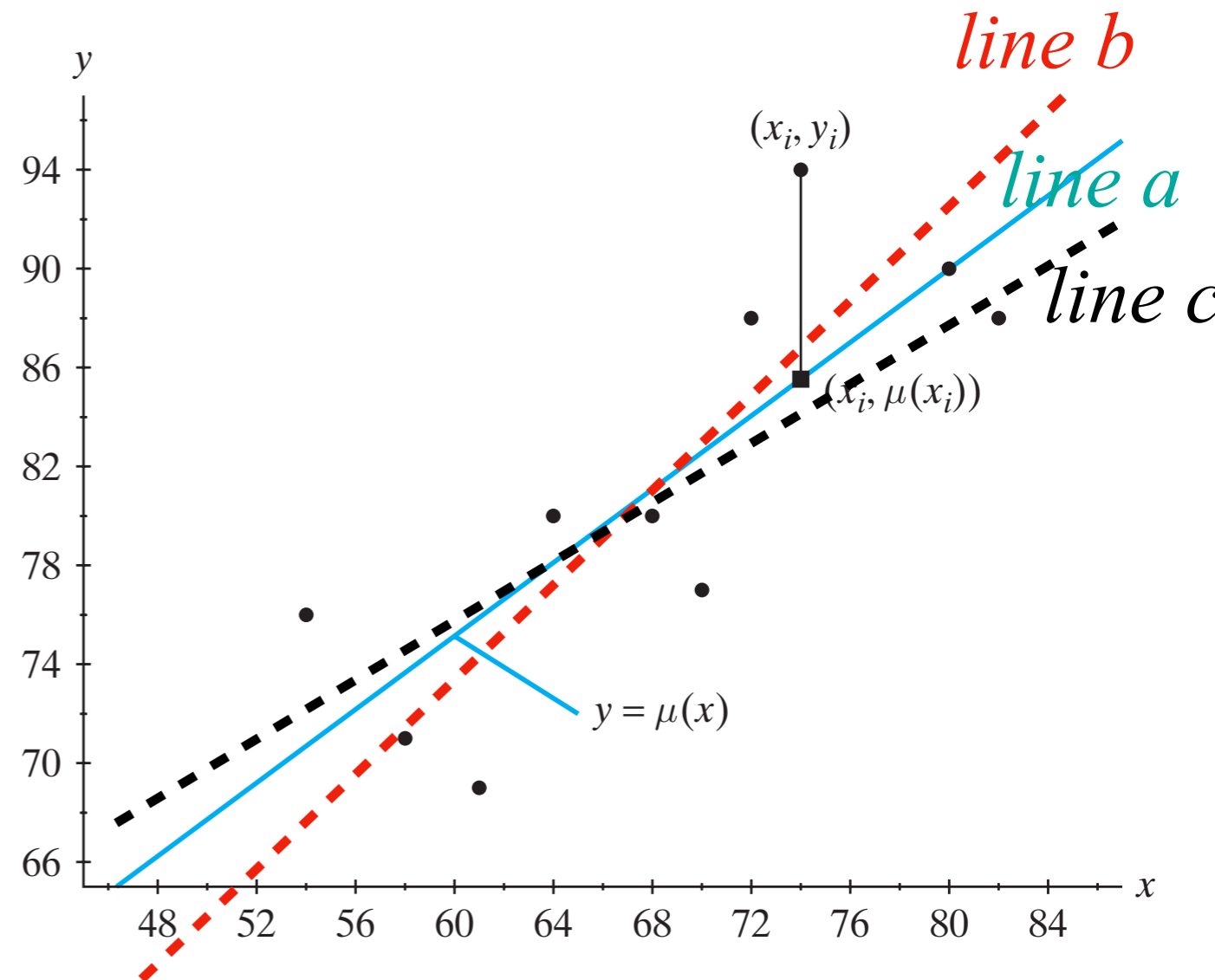
# Let's Come Closer to Reality...

- Find knowledge from data, which has randomness. E.g.,
- Bayesian inference
- Parameter estimation and hypothesis test
- Learning
  - Supervised learning
  - Unsupervised learning...
  - Causal discovery



# Linear Regression: The Two Directions

- Data generated by  $Y = aX + u + \varepsilon$ , where  $\varepsilon \sim N(0, \sigma^2)$
- Regression line in the reverse direction:  
$$\hat{x} = \beta y + c_2$$
- Consider different situations...

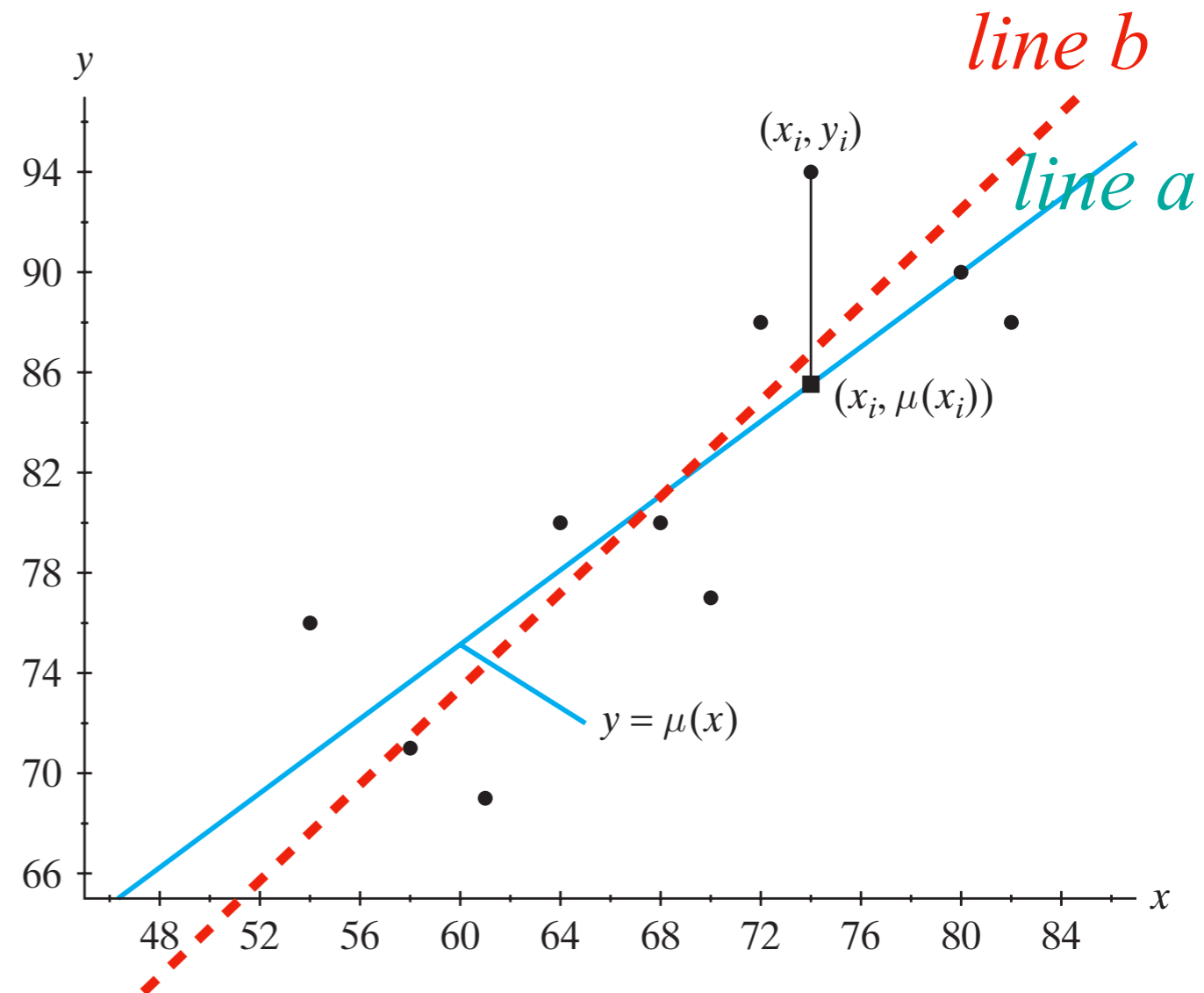


*Question: 1. Interpretation of the parameter.*

*2. Are the regression lines from X to Y and from Y to X identical?*

# Linear Regression: The Two Directions

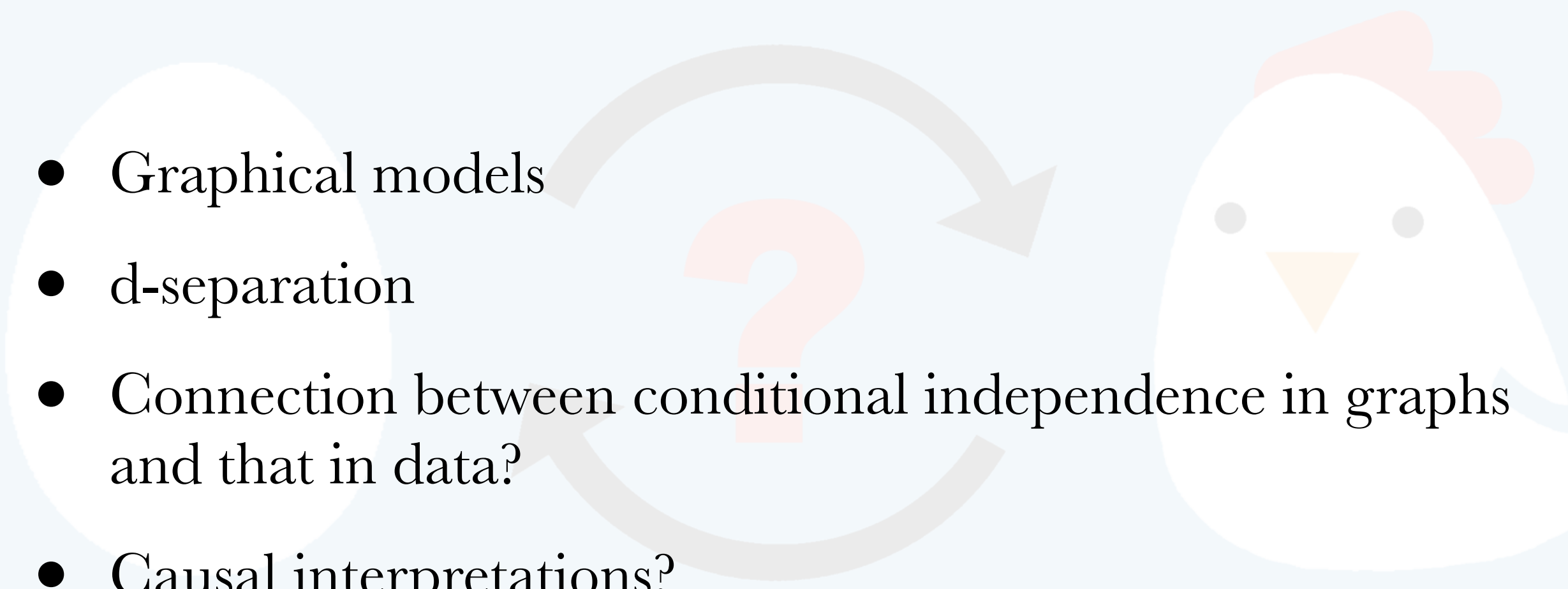
- Data generated by  $Y = aX + u + \varepsilon$ , where  $\varepsilon \sim N(0, \sigma^2)$
- Regression line in the reverse direction:  
$$\hat{x} = \beta y + c_2$$
- Consider different situations...



*Question: 1. Interpretation of the parameter.*

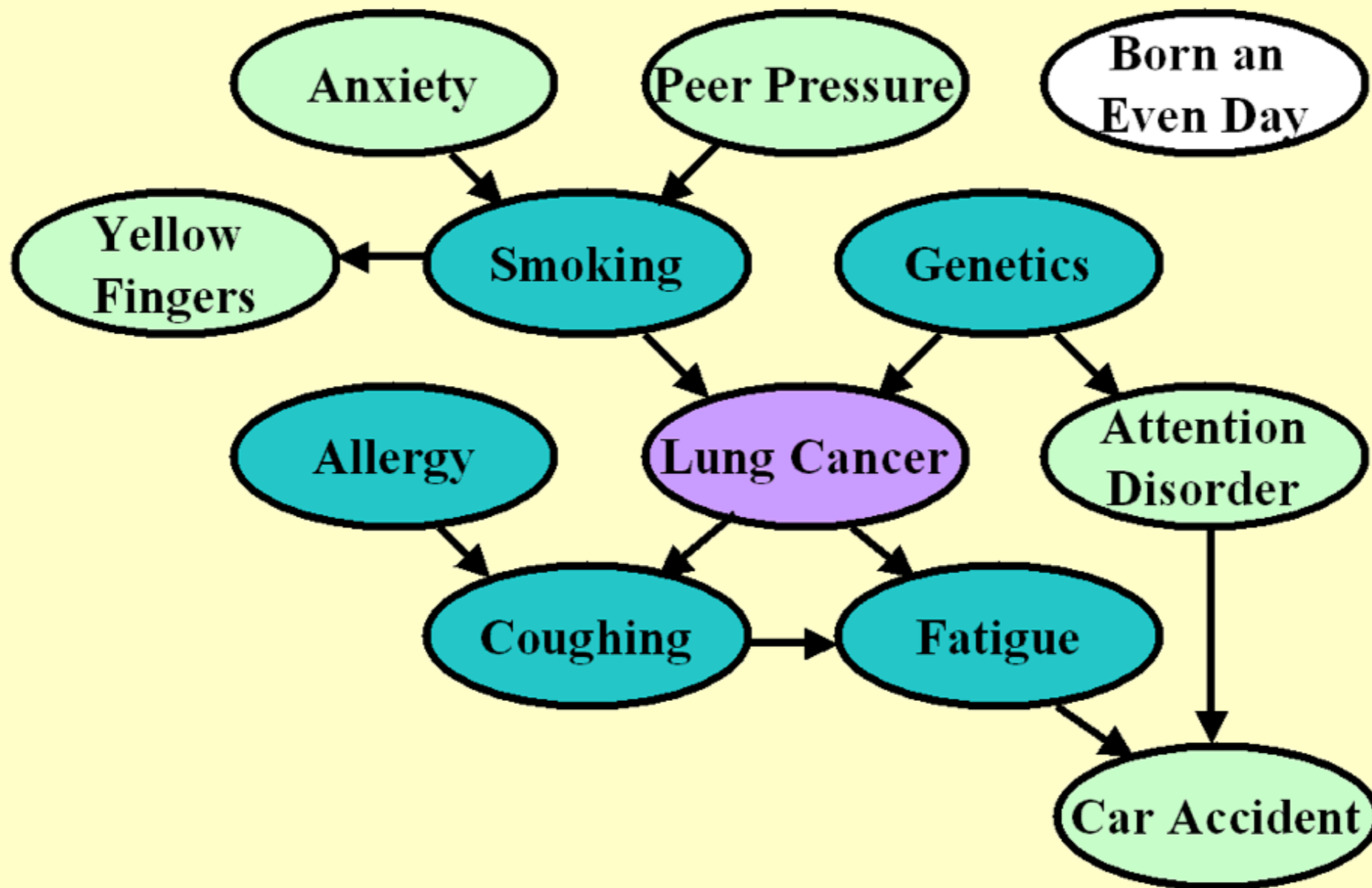
*2. Are the regression lines from X to Y and from Y to X identical?*

# Graphical Models

- Graphical models
  - d-separation
  - Connection between conditional independence in graphs and that in data?
  - Causal interpretations?
- 



# Intuitive Way of Representing and Visualizing Relationships



# Probabilities & Graphical Models

- Why graphical models?

- flexible, powerful and compact way to model relationships between random variables and do inference

- Why probabilities?

The actual science of logic is conversant at present only with things either certain, impossible, or entirely doubtful, none of which (fortunately) we have to reason on. Therefore the true logic for this world is the calculus of Probabilities, which takes account of the magnitude of the probability which is, or ought to be, in a reasonable man's mind.

*James Clerk Maxwell (1850)*

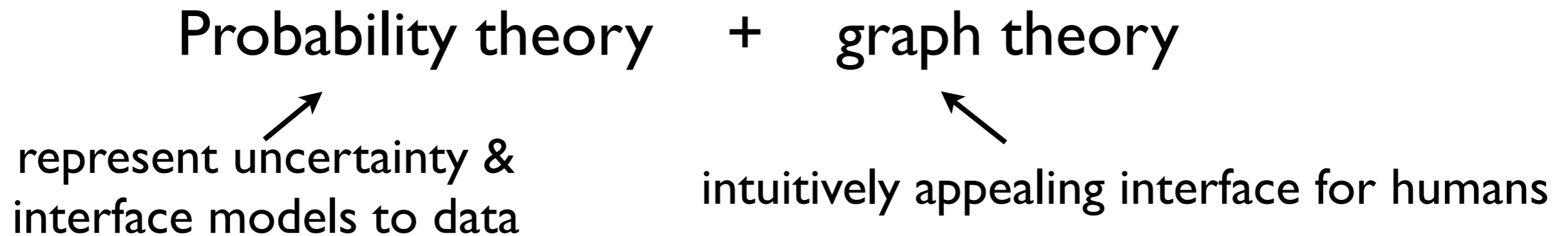
- Why causal discovery?

- understanding, manipulation, prediction, fusion...

*I would rather discover one true cause than gain the kingdom of Persia.*

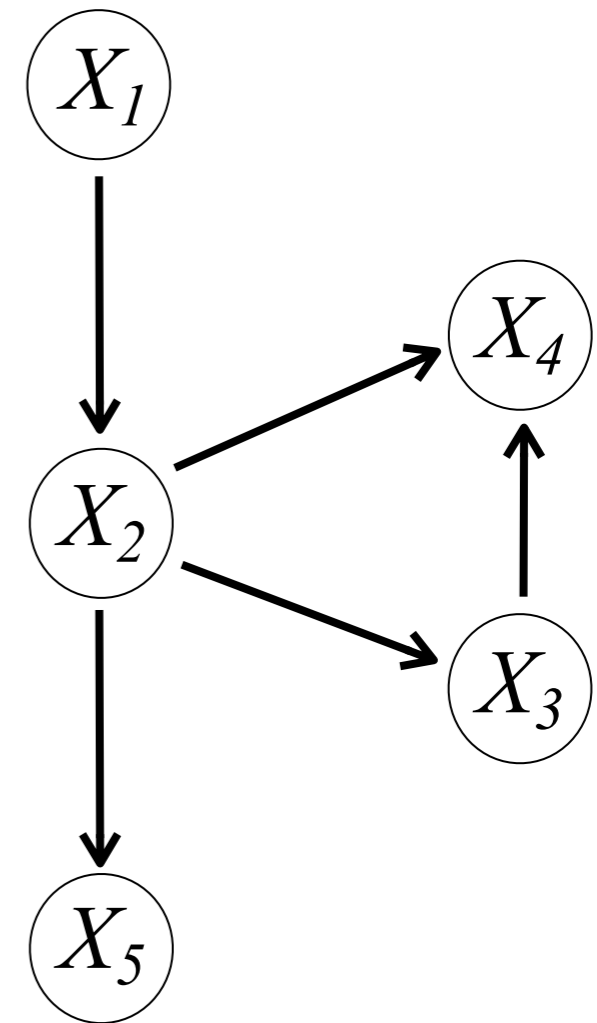
# Graphical Models

- A **graph** comprises nodes (also called vertices) connected by links (also known as edges or arcs)
- Probabilistic graphical models: **graph-based representation** as the basis for compactly encoding a complex distribution
  - **Node: a random variable** (or group of random variables)
  - **Links: direct probabilistic interactions** between them
- Categorization: Undirected graphs vs. **directed acyclic** graphs (DAGs)



# Directed Acyclic Graphs

- Let  $G(\mathcal{V}, \mathcal{E})$  be a directed acyclic graph, where  $\mathcal{V}$  are the nodes and  $\mathcal{E}$  are the edges of the graph.
- Let  $\{X_v : v \in \mathcal{V}\}$  be a collection of random variables indexed by the nodes of the graph.
- To each node  $v \in \mathcal{V}$ , let  $\pi_v$  denote the subset of indices of its parents.
- $X_{\pi_v}$  denotes the vector of random variables indexed by the parents of  $v$ ; sometimes written as  $\text{PA}_{X_v}$  or  $\text{PA}(X_v)$



## Terms:

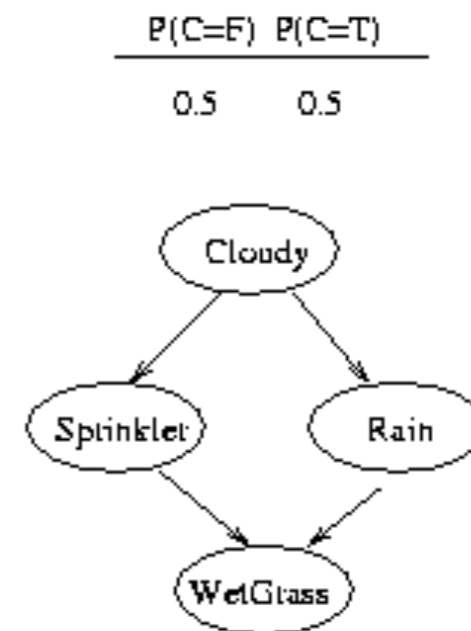
nodes, edge, adjacent, path;  
parents, children, spouses,  
ancestors, descendants,

Markov blanket

# Directed Acyclic Graphical Models

- Also known as Bayesian networks belief nets
- Two components
  - Graph structure (qualitative specification)
    - prior knowledge of causal/modular relationships, or expert knowledge
    - learned from data
  - Conditional probability distributions (CPDs)
    - discrete variables : conditional distribution tables (CPTs)
    - continuous variables: SEMs

C	P(S=F)	P(S=T)
F	0.5	0.5
T	0.9	0.1



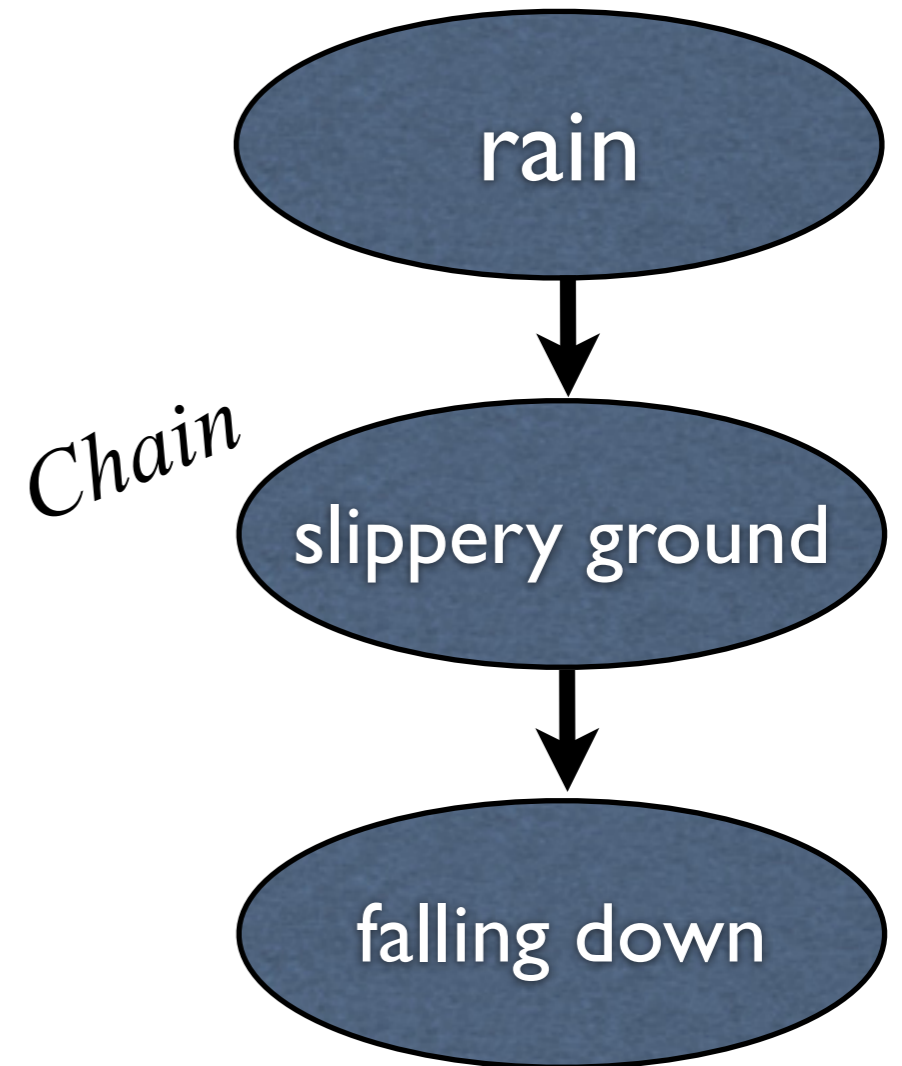
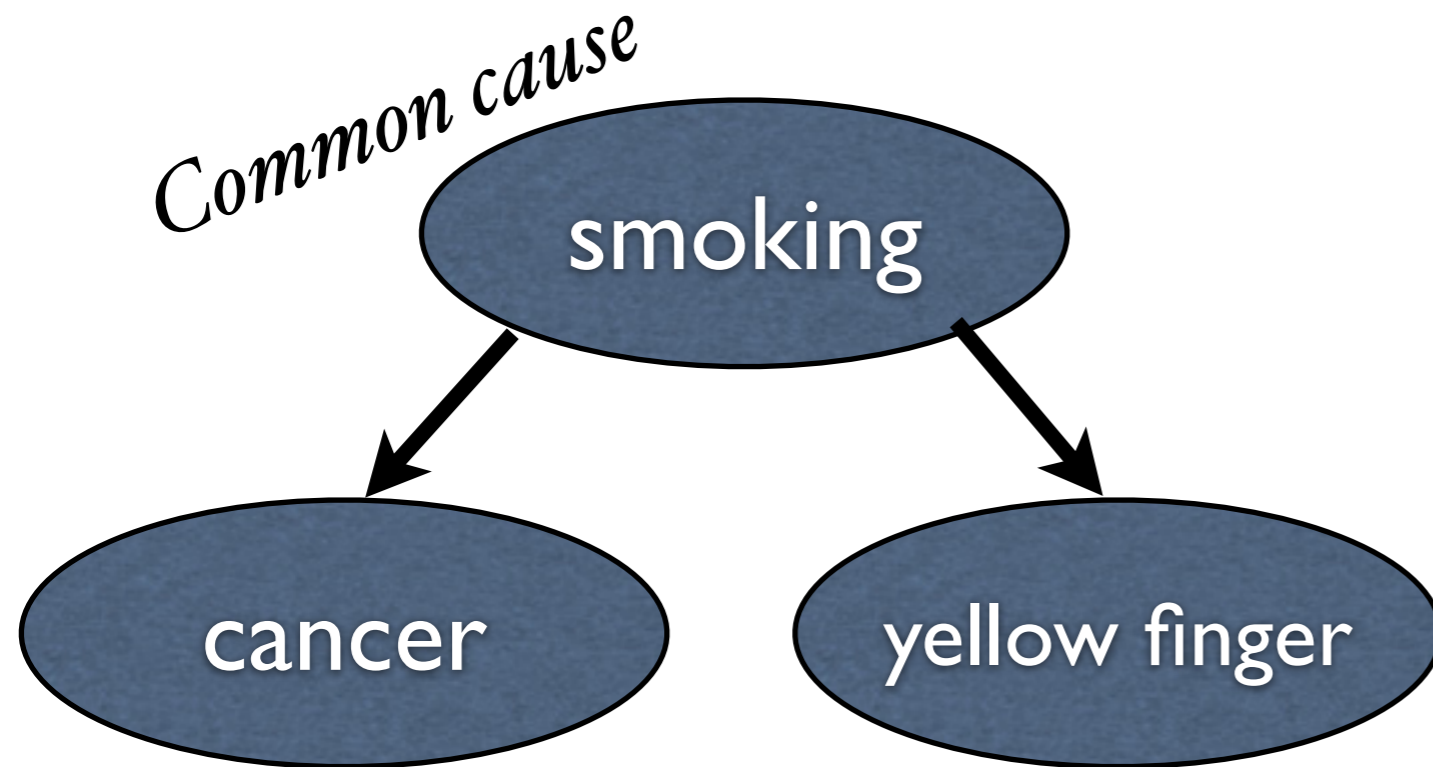
C	P(R=F)	P(R=T)
F	0.8	0.2
T	0.2	0.8

S	R	P(W=F)	P(W=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

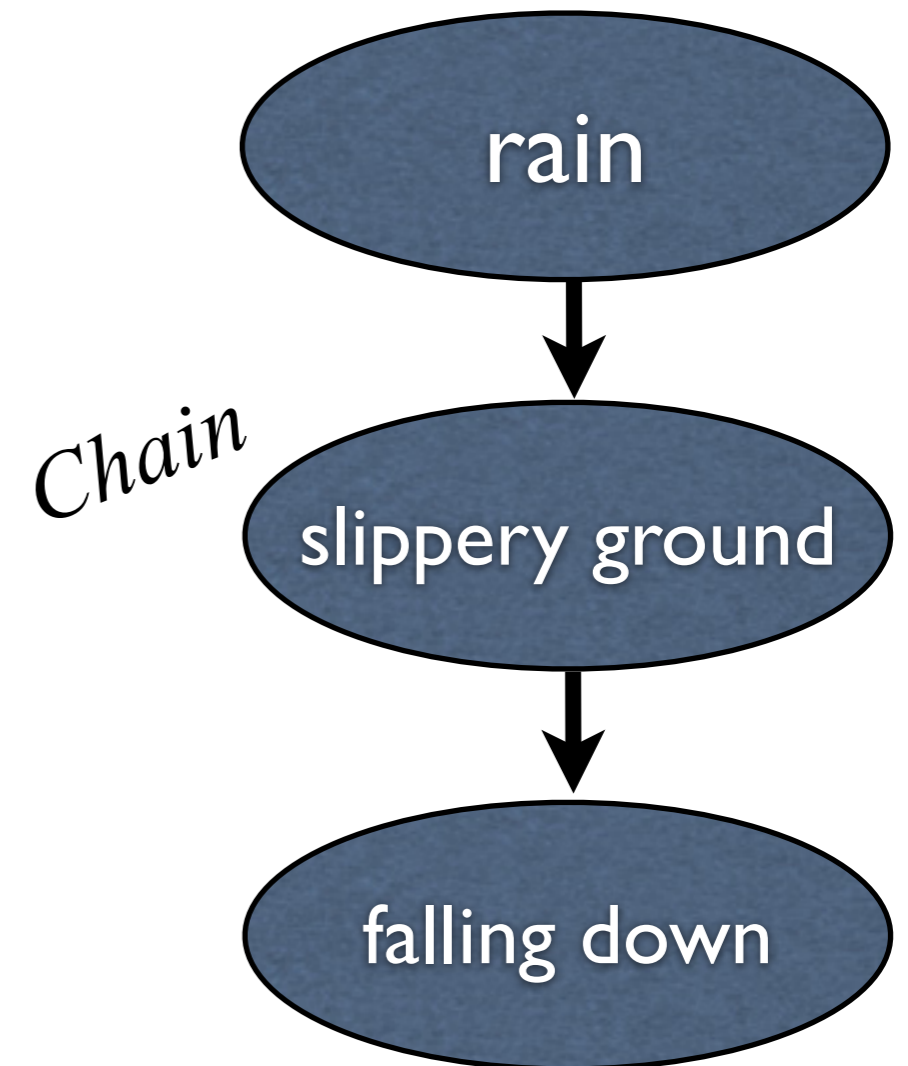
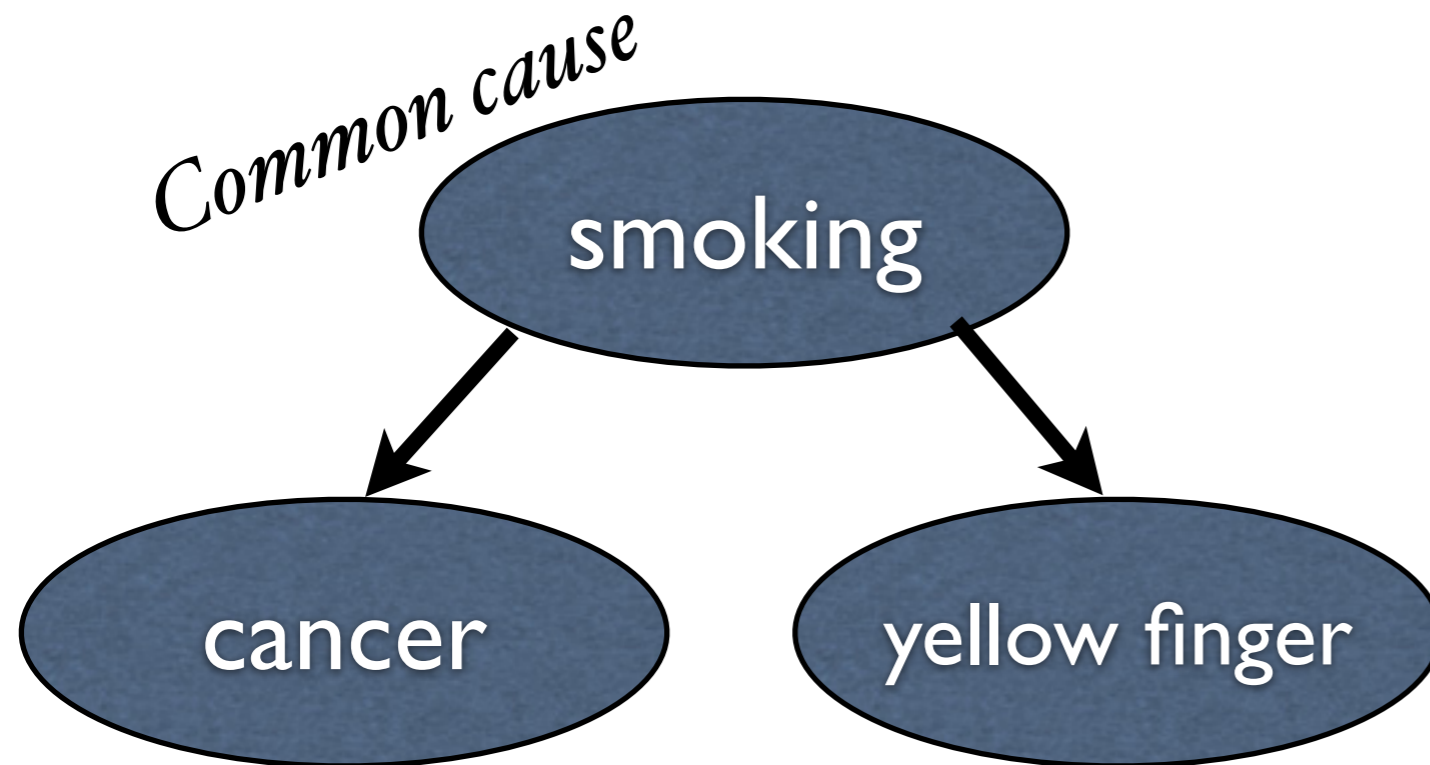
# Bayesian Networks: Story

- Breakthrough in early 1980s (by Pearl et al.)
- In a joint probability distribution, every variable is, in general, related to all other variables.
- Pearl and others realized:
  - It is often reasonable to make the assumption that each variable is directly related to only a few other variables
  - This leads to **modularity**: Allowing decomposing a complex model into small manageable pieces
  - Giving rise to **Bayesian networks**

# What Independence Relationships Can You See?



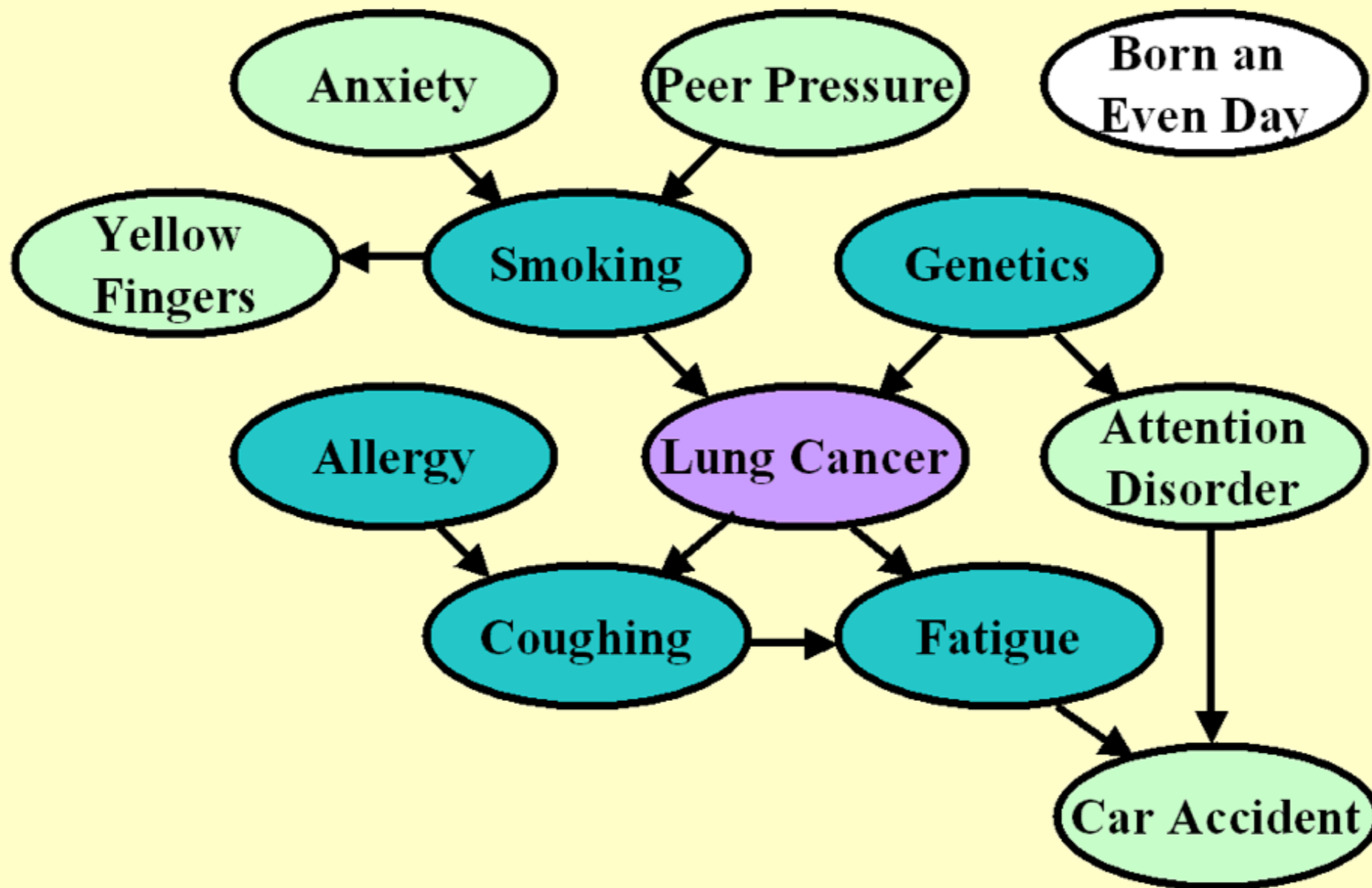
# (Local) Markov Condition



- Each variable is independent from its non-descendants given its parents



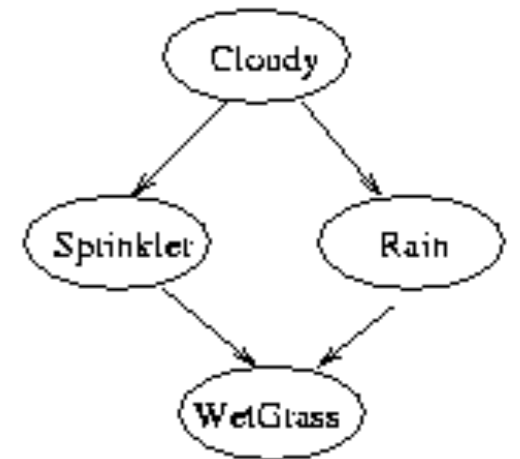
# For Instance, What Independence Relations can You See?



# Factorization According to Directed Graphs

- Chain rule of probability gives

$$P(C,S,R,W) = P(C) P(S|C) P(R|C,S) P(W|C,S,R)$$



- According to the CI relationships:

$$P(C,S,R,W) = P(C) P(S|C) P(R|C) P(W|S,R)$$

- The graph structure allows us to represent the joint distribution more compactly (*Markov factorization* or *Markov decomposition* of the joint distribution):

- $P(X_1, \dots, X_n) = \prod_i P(X_i | PA_i)$

- Remember this example?

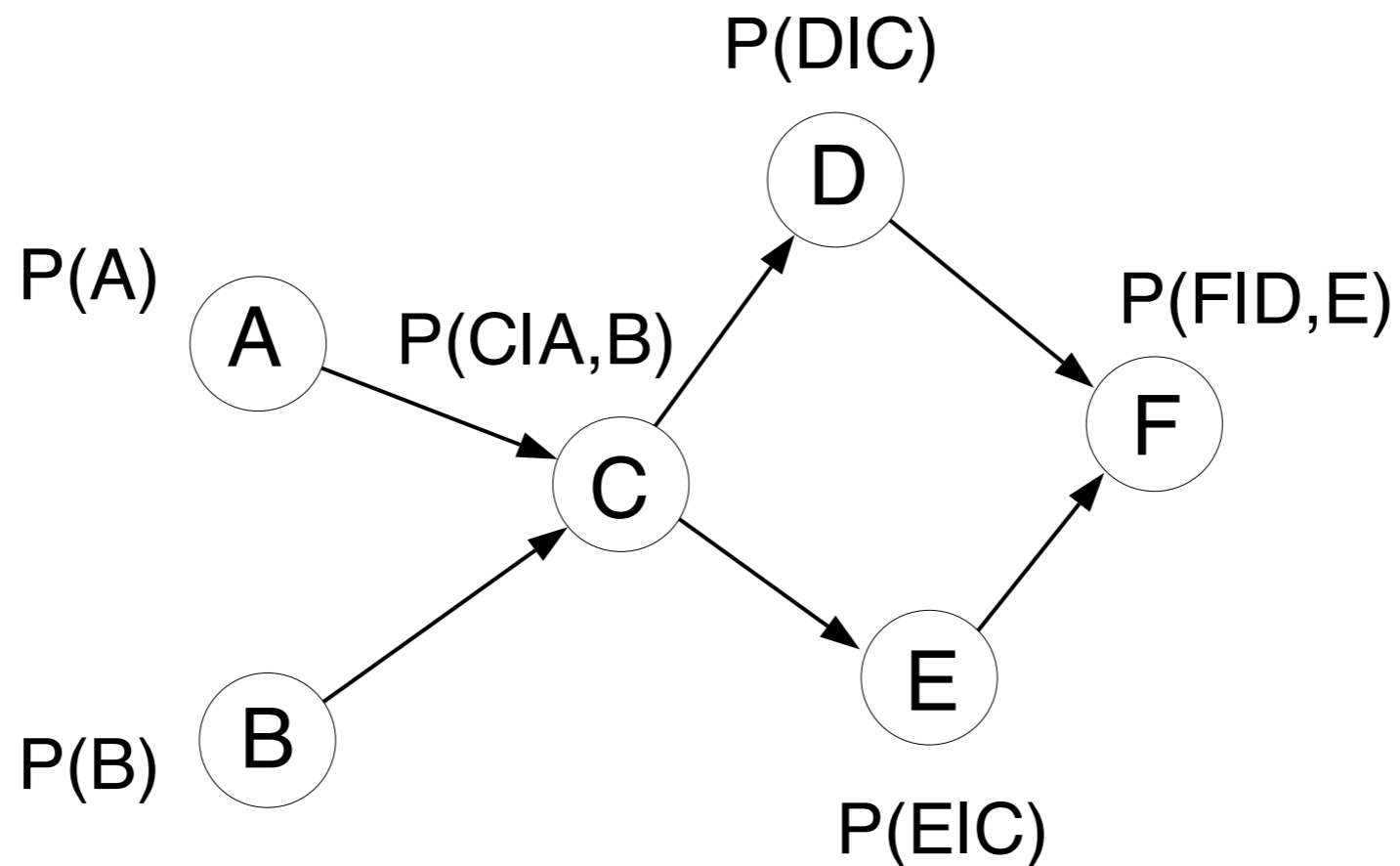


*If we aim to represent causal info, is CI info enough?*

*$X \rightarrow Y$  or  $X \leftarrow Y$ ?*

# Factorization According to Directed Graphs: Procedure

- Associate a conditional probability with each node
- Then take the product of the local probabilities to yield the global probabilities



# Tasks Related to Bayesian Networks

- **Probabilistic inference:**

Calculate  $P(\text{variables of interest} \mid \text{observed variables})$

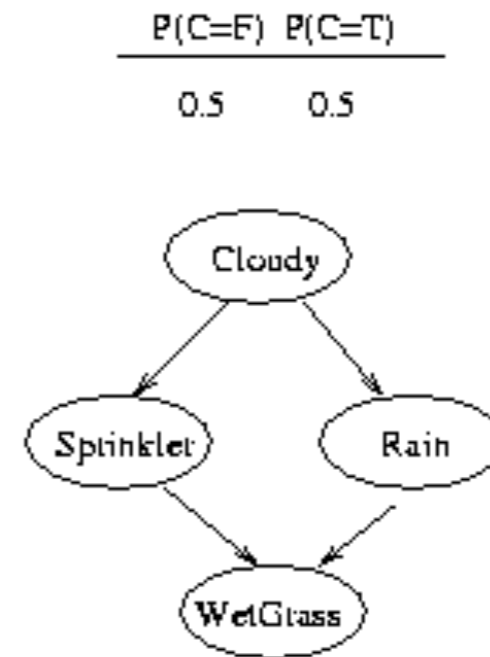
- Most common task where we want to use Bayesian networks

- How to find  $P(S=1 \mid W=1)$ ?  
 $P(R=1 \mid W=1)$ ?

- **Parameter learning**

- **Structure learning:** Learning the structure of the graphical model from observations

C	P(S=F)	P(S=T)
F	0.5	0.5
T	0.9	0.1

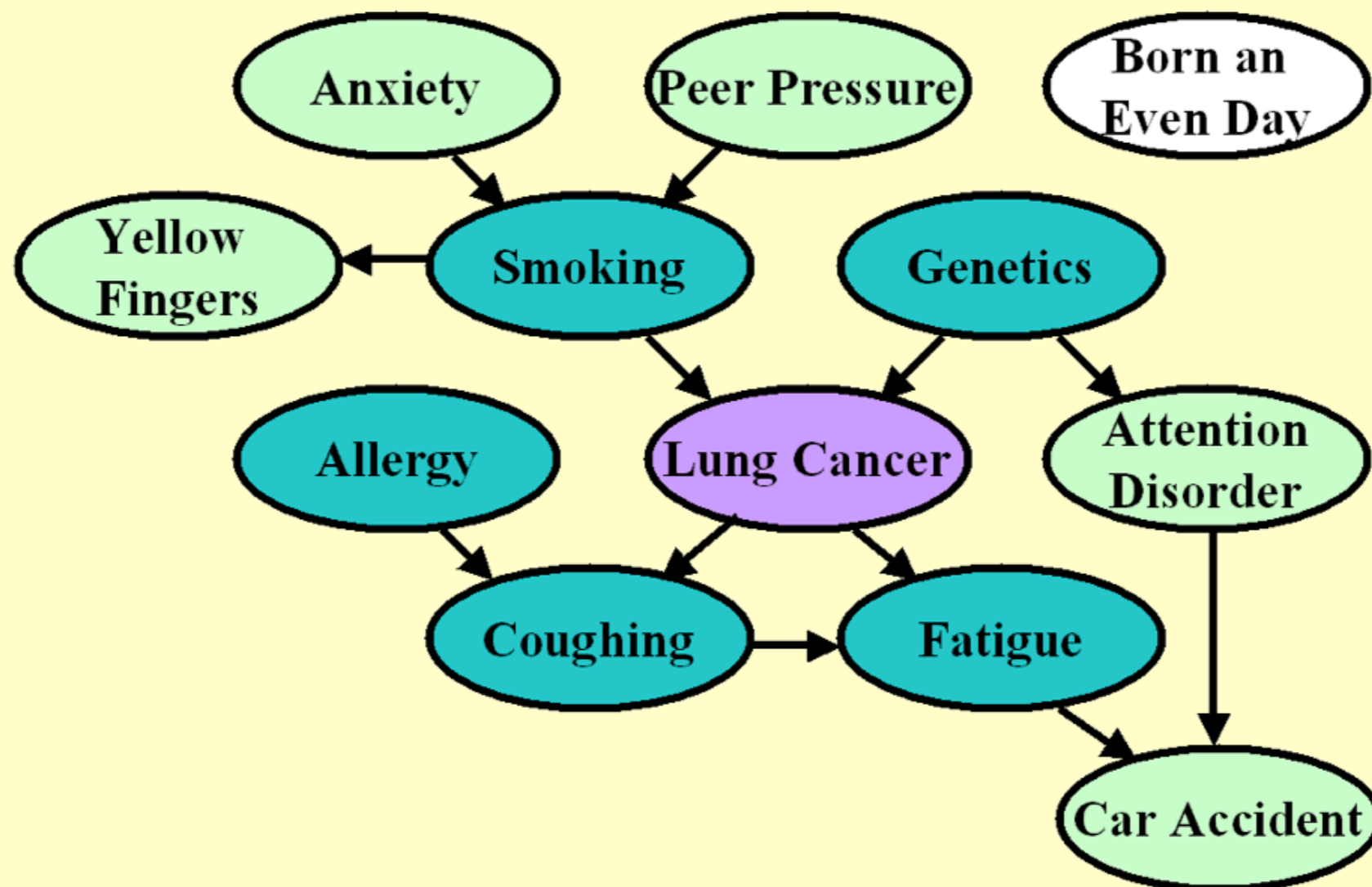


C	P(R=F)	P(R=T)
F	0.8	0.2
T	0.2	0.8

S	R	P(W=F)	P(W=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

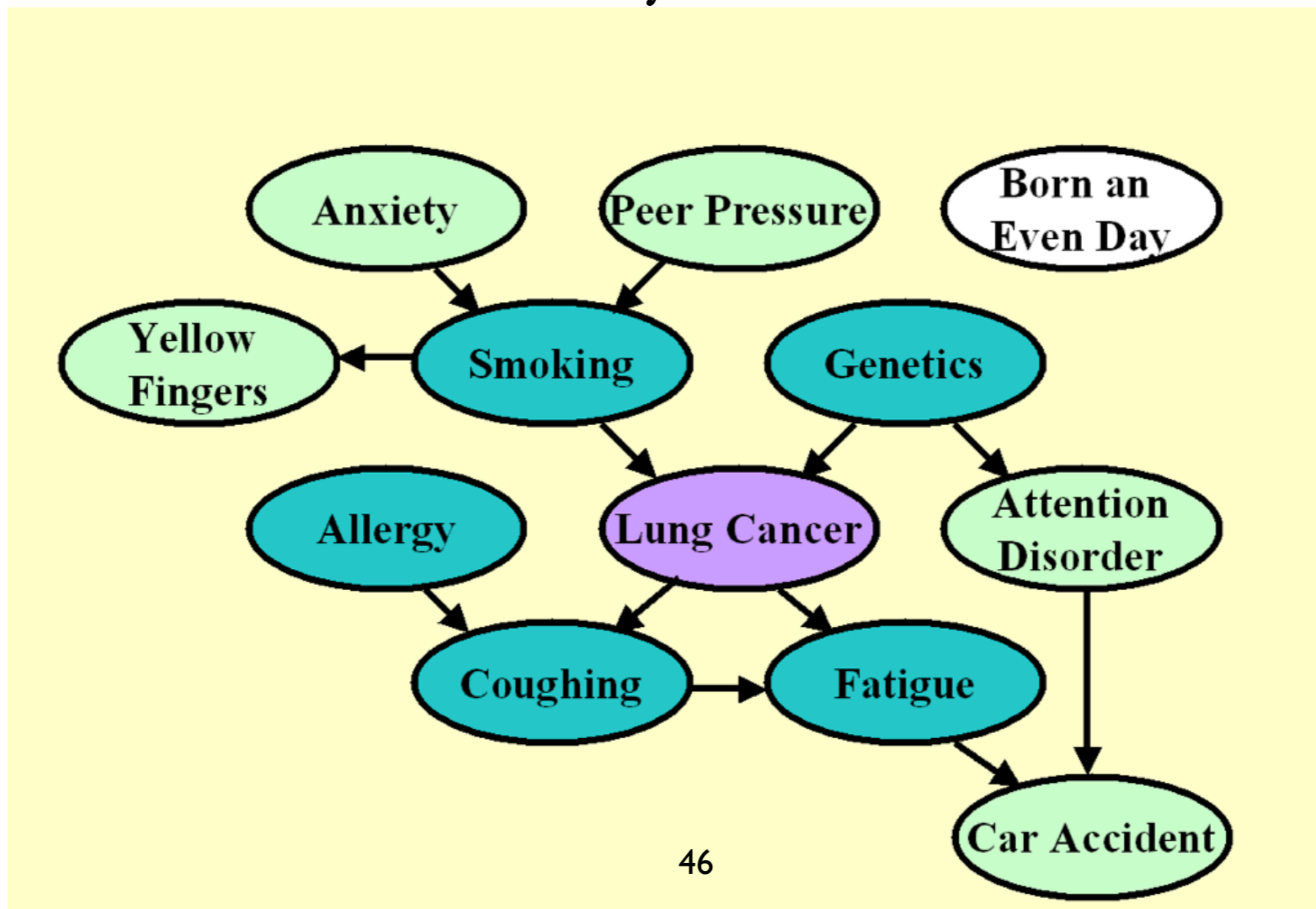
# Is Local Markov Condition Enough?

- Can we see whether **two arbitrary variables**,  $X$  and  $Y$ , are conditionally independent **given an arbitrary set of variables**,  $Z$ ?



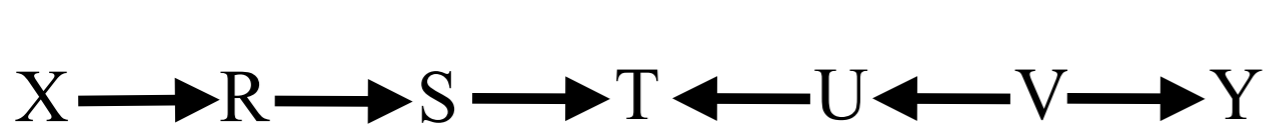
# D-Separation Tells Conditional Independence

- If every path from a node in **X** to a node in **Y** is **d-separated** by **Z**, then **X** and **Y** are **always conditionally independent** given **Z**
- d: directional... You will see why

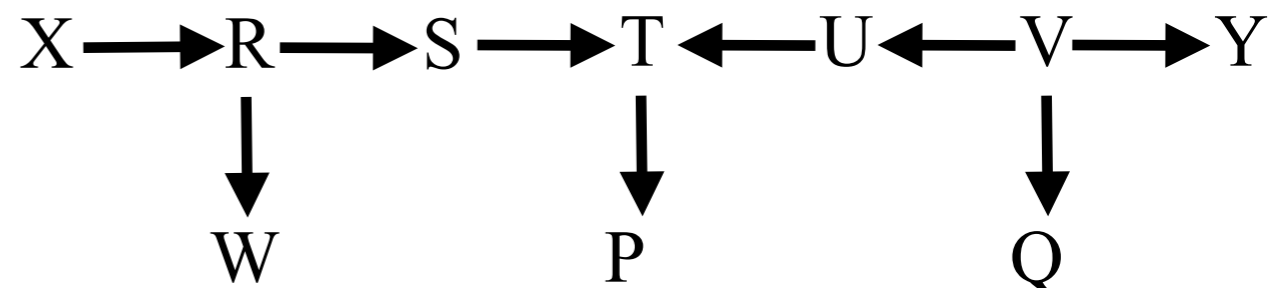


# D-Separation

- A set of nodes  $\mathbf{Z}$  d-separates two sets of nodes  $\mathbf{X}$  and  $\mathbf{Y}$  if every path from a node in  $\mathbf{X}$  to a node in  $\mathbf{Y}$  is blocked given  $\mathbf{Z}$ .
- A path  $p$  is blocked by a set of nodes  $\mathbf{Z}$  if
  - $p$  contains a chain  $i \rightarrow m \rightarrow j$  or a common cause  $i \leftarrow m \rightarrow j$  such that the middle node  $m$  is in  $\mathbf{Z}$ , or
  - $p$  contains a collider  $i \rightarrow m \leftarrow j$  such that the middle node  $m$  is in not  $\mathbf{Z}$  and no descendant of  $m$  is in  $\mathbf{Z}$



X and Y d-separated by  $\{R, V\}$ ?  
 S and U d-separated by  $\{R, V\}$ ?



X and Y d-separated by  $\{R, P\}$ ?

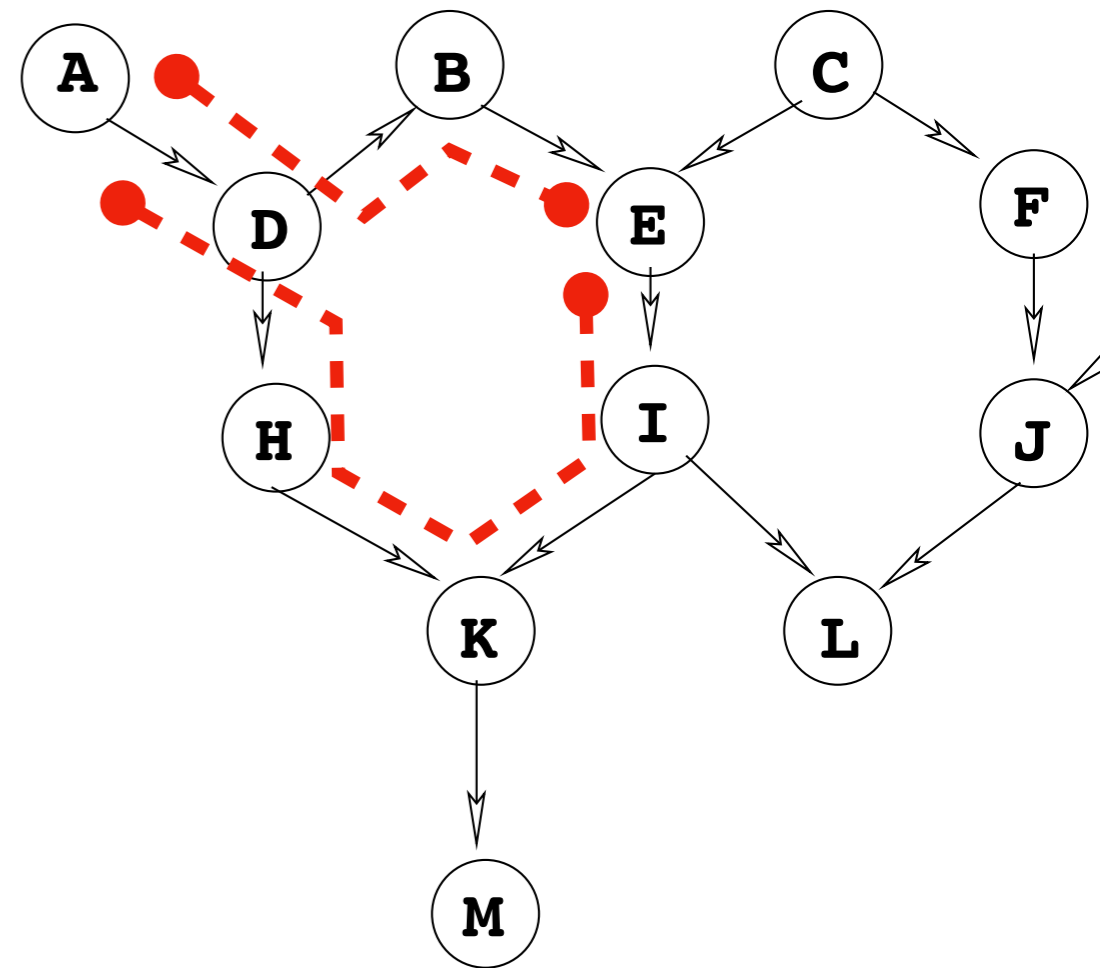
# D-Separation

- A set of nodes  $\mathbf{Z}$  d-separates two sets of nodes  $\mathbf{X}$  and  $\mathbf{Y}$  if every path from a node in  $\mathbf{X}$  to a node in  $\mathbf{Y}$  is blocked given  $\mathbf{Z}$ .

- A path  $p$  is blocked by a set of nodes  $\mathbf{Z}$  if

- $p$  contains a chain  $i \rightarrow m \rightarrow j$  or a common cause  $i \leftarrow m \rightarrow j$  such that the middle node  $m$  is in  $\mathbf{Z}$ , or

- $p$  contains a collider  $i \rightarrow m \leftarrow j$  such that the middle node  $m$  is not in  $\mathbf{Z}$  and no descendant of  $m$  is in  $\mathbf{Z}$



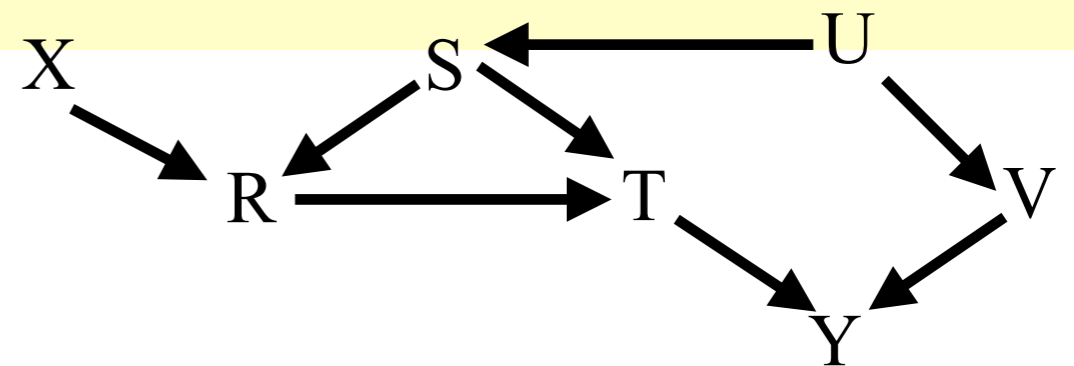
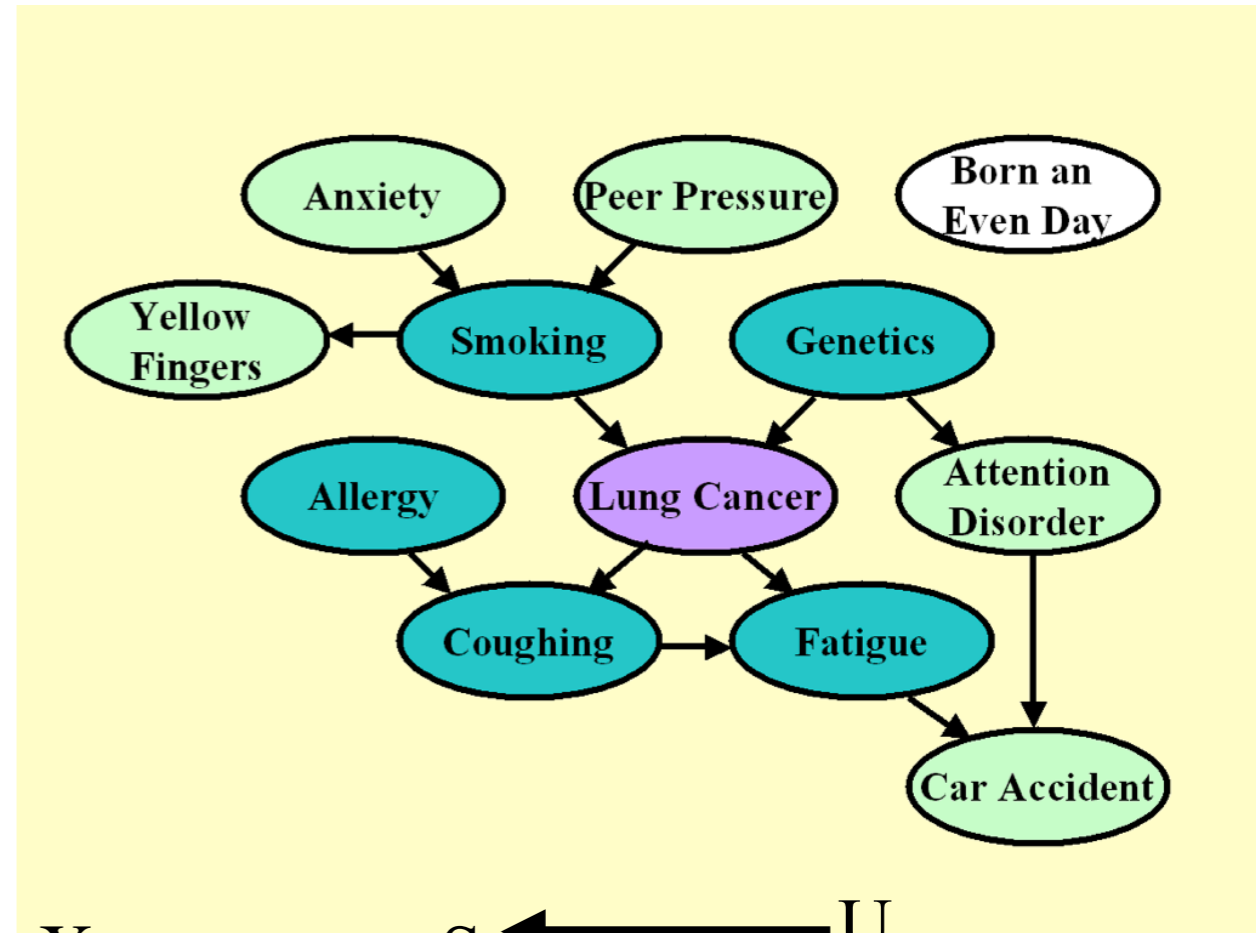
A and E d-separated by B ?

A and E d-separated by  $\{B, M\}$  ?



# D-Separation: Intuition

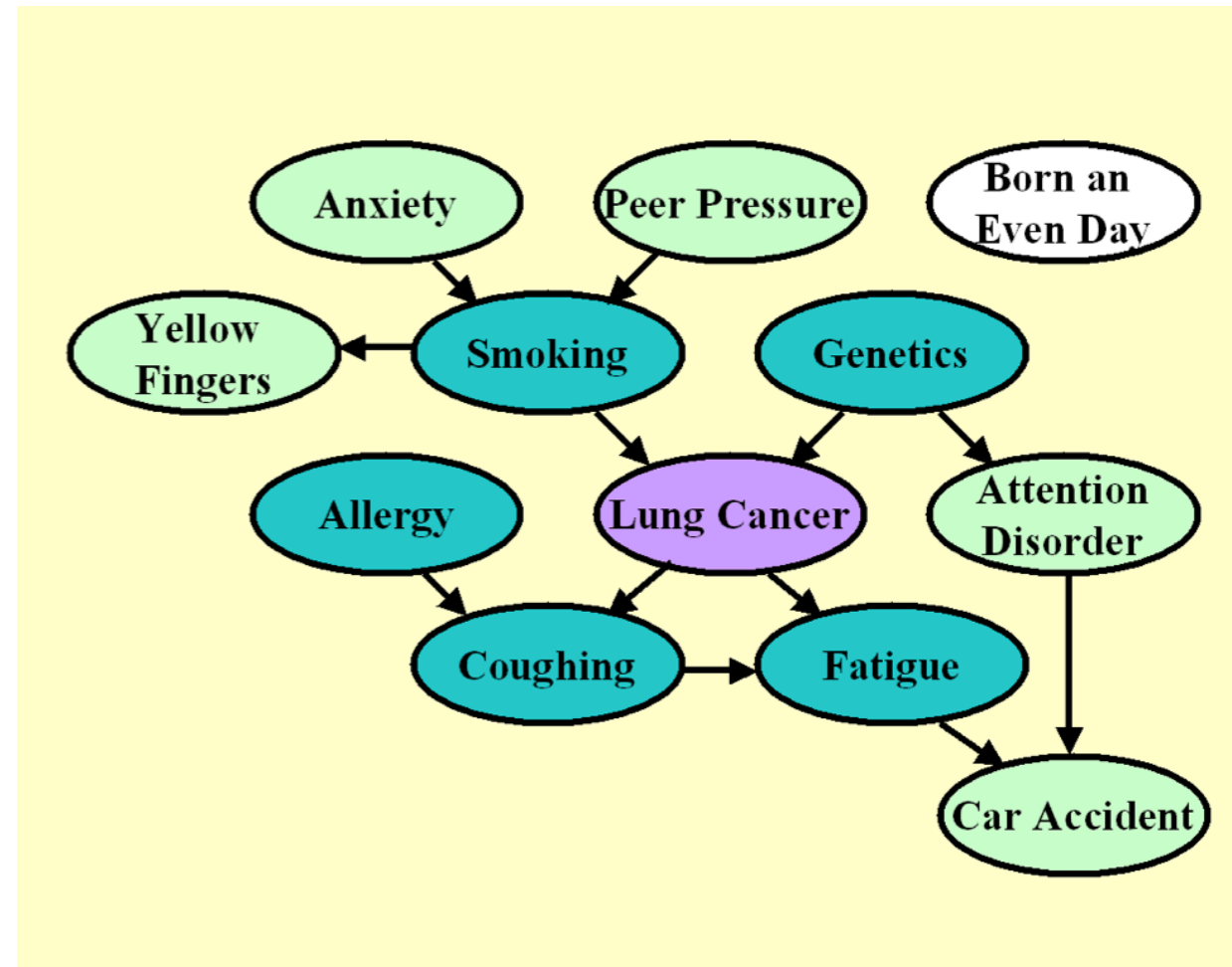
- Suppose  $X$  and  $Y$  are d-separated by  $Z$
- Then if you fix  $Z$ ,  $X$  and  $Y$ 
  - do not cause each other and
  - do not share a common cause
- $X$  and  $Y$  are independent (conditional on  $Z$ )!



1.  $X$  and  $Y$  d-separated by  $\{R\}$ ?
2.  $X$  and  $Y$  d-separated by  $\{R, T\}$ ?
3.  $X$  and  $Y$  d-separated by  $\{T, V\}$ ?
4.  $X$  and  $V$  d-separated by  $\emptyset$ ?

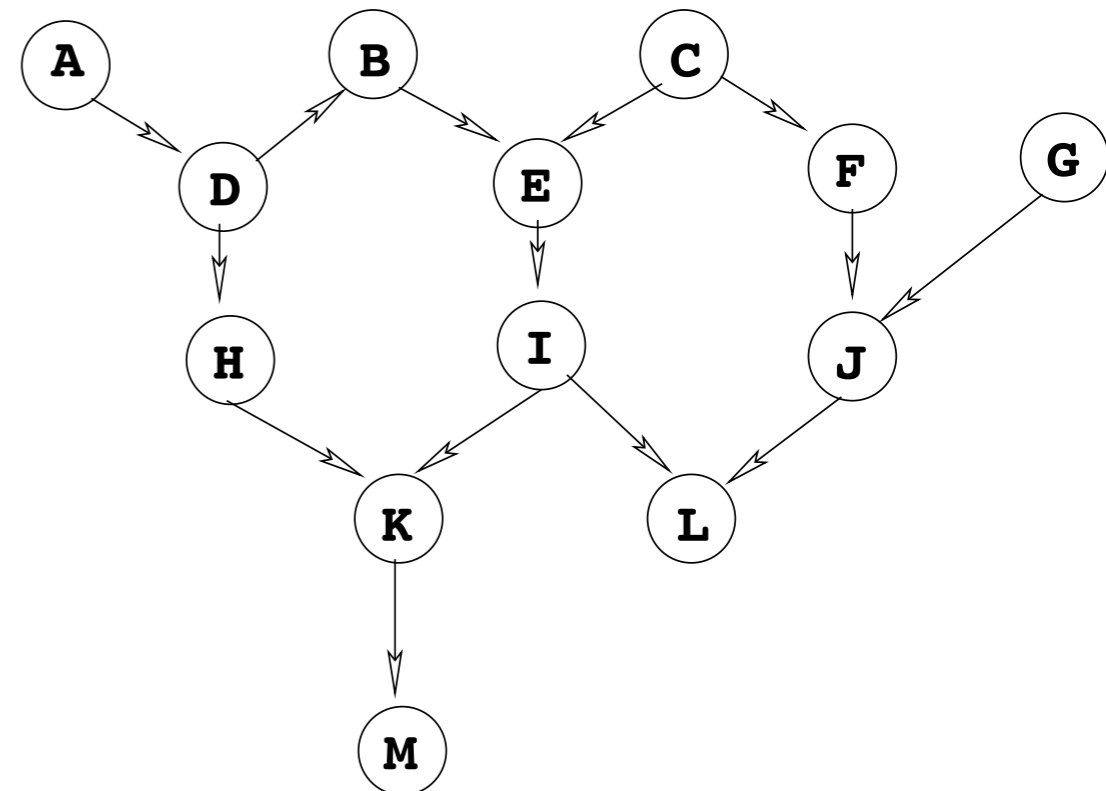
# Local & Global Markov Conditions

- **Local** Markov condition:
  - In a DAG, a variable  $X$  is independent of all its non-descendants given its parents
- **Global** Markov condition:
  - Given a DAG, let  $X$  and  $Y$  be two variables and  $\mathbf{Z}$  be a set of variables that does not contain  $X$  or  $Y$ . If  $\mathbf{Z}$  **d-separates**  $X$  and  $Y$ , then  $X \perp\!\!\!\perp Y \mid \mathbf{Z}$ .
- Actually equivalent on DAGs!



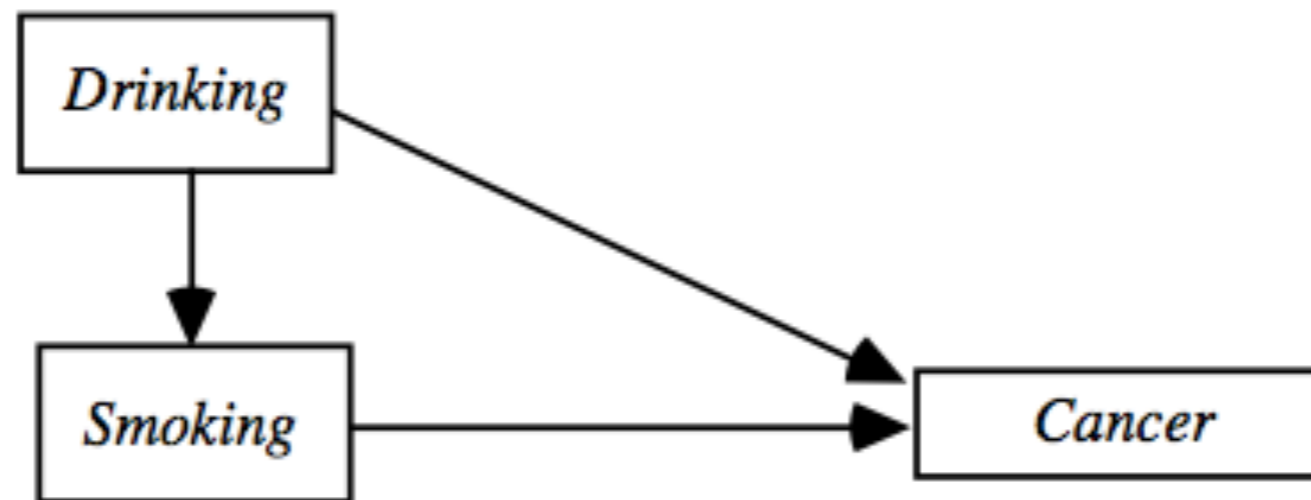
# Markov Blanket

- In a DAG, the Markov Blanket of a node  $X$  is the set consisting of
  - Parents of  $X$
  - Children of  $X$
  - Parents of children (i.e., spouses) of  $X$
- In a DAG, a variable  $X$  is conditionally independent from all other variables given its Markov Blanket
  - Implied by d-separation...
- The Markov blanket of  $I$ ?



# Representing Causal Relations with Directed Graphs

- A directed graph represents a causally sufficient causal structure



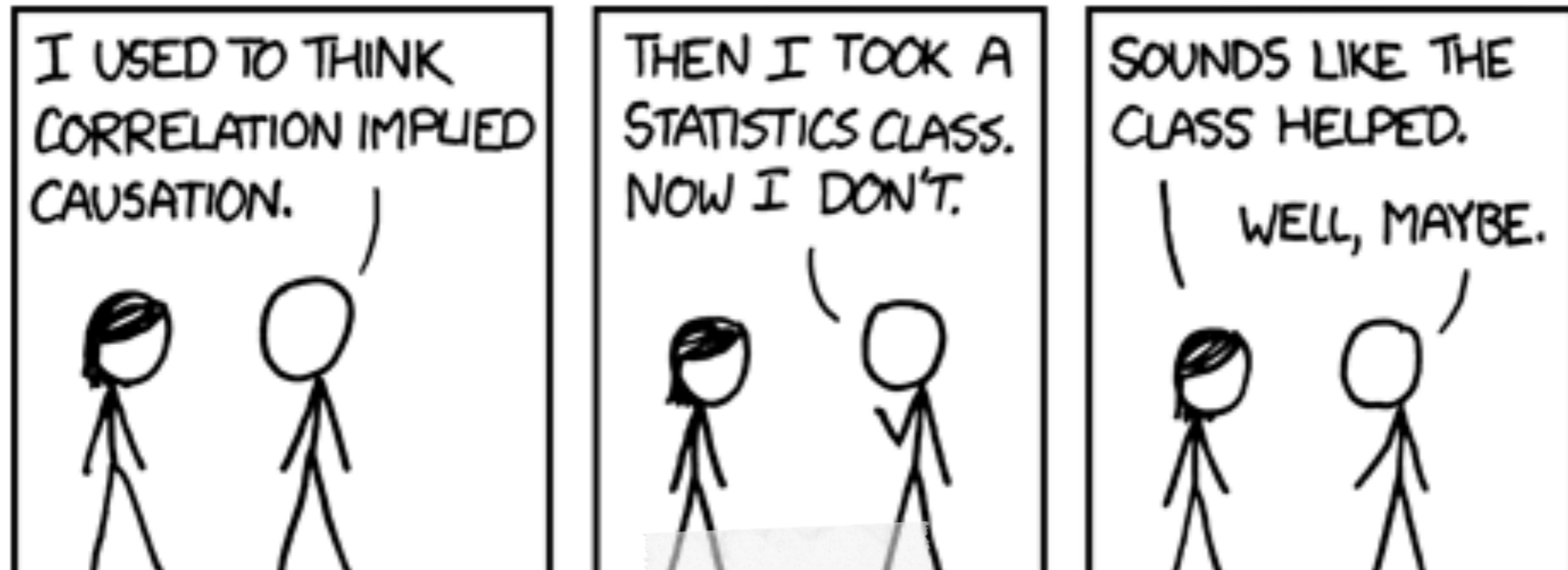
(adapted from “Causation, Prediction, and Search” by SGS, 1995)

- Directed edge from  $A$  to  $B$  means  $A$  is a direct cause of  $B$  relative to the given variable set  $V$

# Causality vs. Dependence



- Causality → dependence ! Dependence → causality



An intervention on  $X$  changes only the target variable  $X$ , leaving any other variable unchanged, at least for the moment.

$X$  and  $Y$  are **associated** iff

$$\exists x_1 \neq x_2 P(Y|X=x_1) \neq P(Y|X=x_2)$$

$X$  is a **cause** of  $Y$  iff

$$\exists x_1 \neq x_2 P(Y|\text{do } X=x_1) \neq P(Y|\text{do } X=x_2)$$

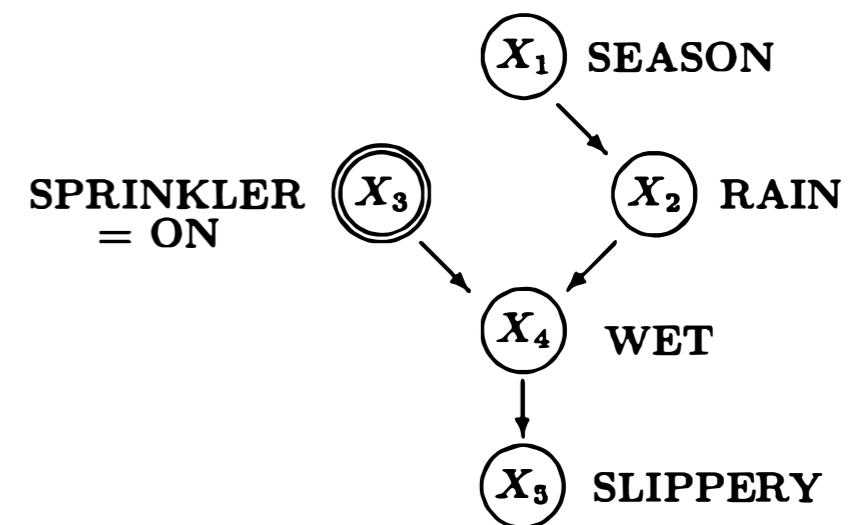
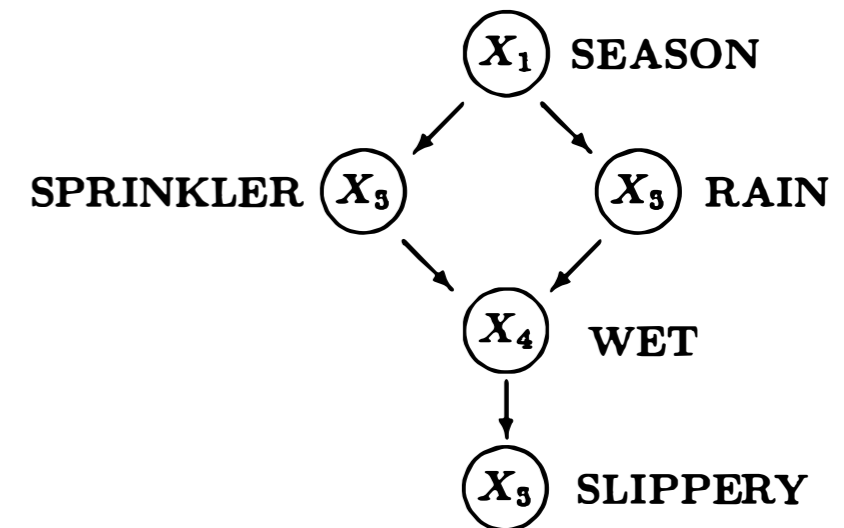
intervention

# Causal Bayesian Networks (CBNs)

- Bayesian networks: DAGs
- Causal Bayesian networks
  - More meaningful & able to **represent and respond to external or spontaneous changes**

Let  $P_x(V)$  be the distribution of  $V$  resulting from intervention  $do(X=x)$ . A DAG  $G$  is a CBN if

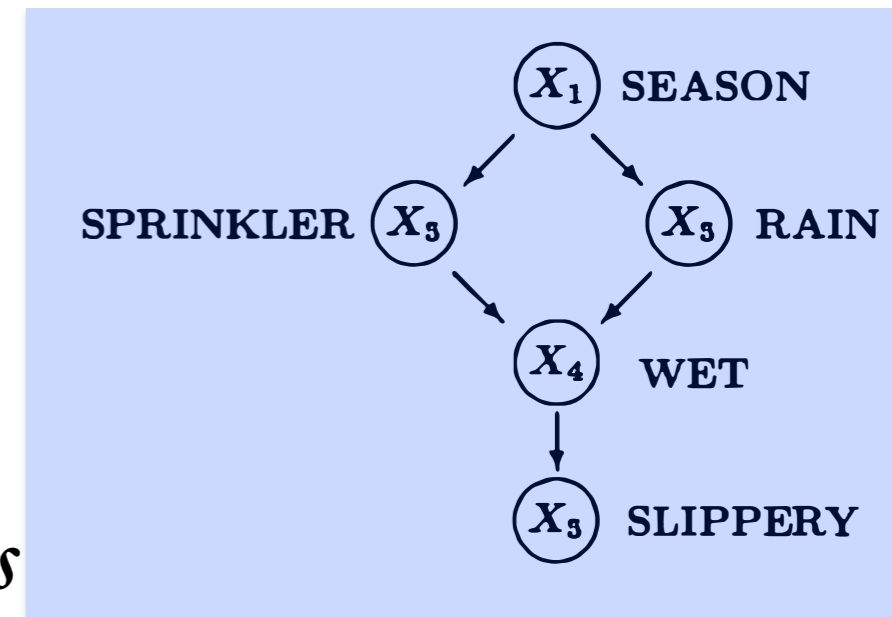
1.  $P_x(V)$  is Markov relative to  $G$ ;
2.  $P_x(V_i=v_i)=1$  for all  $V_i \in X$  and  $v_i$  consistent with  $X=x$ ;
3.  $P_x(V_i | PA_i) = P(V_i | PA_i)$  for all  $V_i \notin X$ , i.e.,  $P(V_i | PA_i)$  remains invariant to interventions not involving  $V_i$ .



What is  $P_{X_3=ON}(X_1, X_2, X_4, X_5)$ ?

# Structural Causal Models

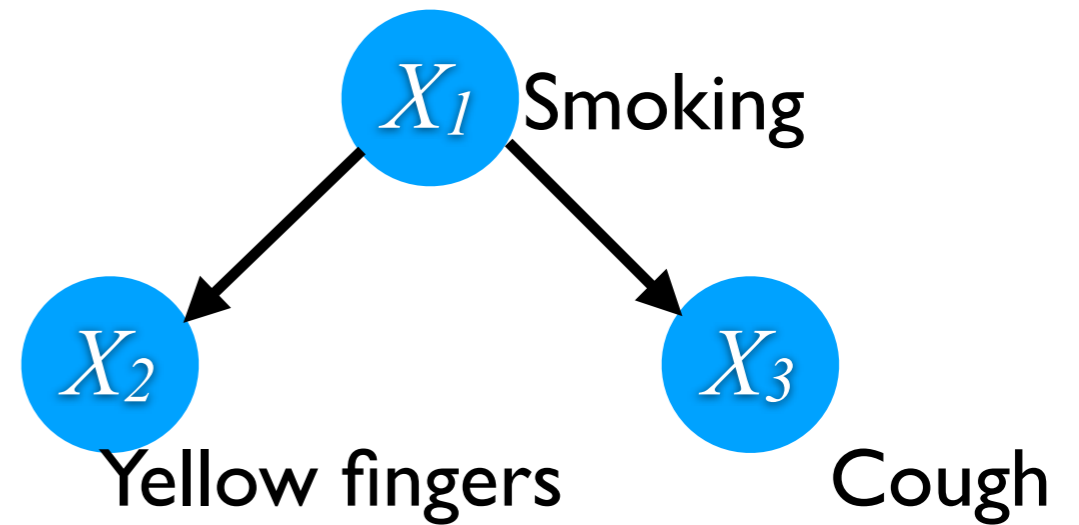
- $X_i = f_i(PA_i, E_i), i=1, \dots, n$
- $E_i$ : exogenous variables / errors / disturbances
- Each equation represents an *autonomous* mechanism
- Describes how nature assigns values to variables of interest
- Distinction between structural equations & algebraic equations
- Associated with graphical causal models



$$PA_i \longrightarrow X_i$$

$$\begin{aligned}
 X_1 &= E_1, \\
 X_2 &= f_2(X_1, E_2), \\
 X_3 &= f_3(X_1, E_3), \\
 X_4 &= f_4(X_2, X_3, E_4), \\
 X_5 &= f_5(X_4, E_5)
 \end{aligned}$$

# Three Types of Problems in Current AI



- Three questions:

$X_1$	$X_2$	$X_3$
1	0	0
0	0	1
0	1	1
1	1	1
0	0	0
0	1	0
1	1	1
1	1	1
0	0	0
1	0	0
...	...	...

- **Prediction:** Would the person cough if we *find* he/she has yellow fingers?

$$P(X_3 \mid X_2=1)$$

- **Intervention:** Would the person cough if we *make sure* that he/she has yellow fingers?

$$P(X_3 \mid \text{do}(X_2=1))$$

- **Counterfactual:** Would George cough *had* he had yellow fingers, *given that he does not have yellow fingers and coughs*?

$$P(X_3_{X_2=1} \mid X_2 = 0, X_3 = 1)$$



# Summary

- Probability theory & statistics: (conditional) independence, Gaussian distribution, regression...
- Directed acyclic graph
- d-separation
- Local and global Markov condition
- Causal graphical representation