Part II
Preliminaries:
Probability, a bit Statistics, and Graphical Models

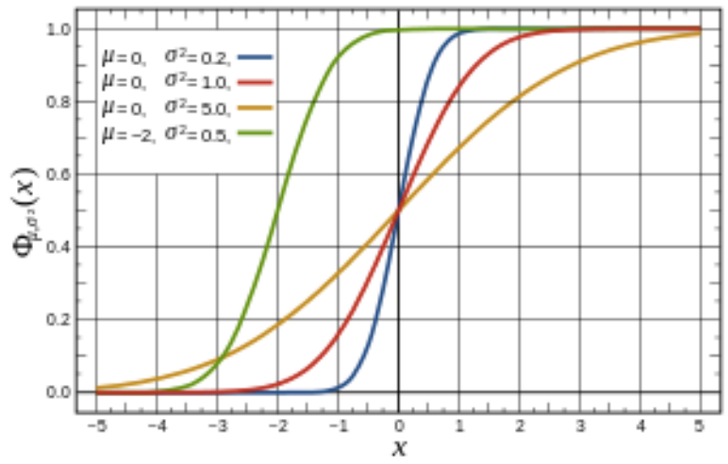
- Probability, (conditional) independence, and a bit statistics

- Graphical models, Markov condition, and representation of causal models

How to Specify Prob. Measures of Random Variables

- Cumulative distribution function (CDF):
  A function $F_X : \mathbb{R} \rightarrow [0,1]$ which specifies a probability measure as

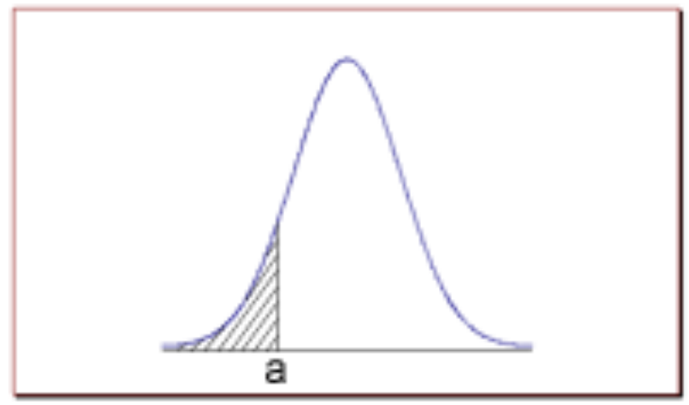
$$F_X(x) \triangleq P(X \leq x)$$



- Probability Mass Functions (PMFs) for discrete variables

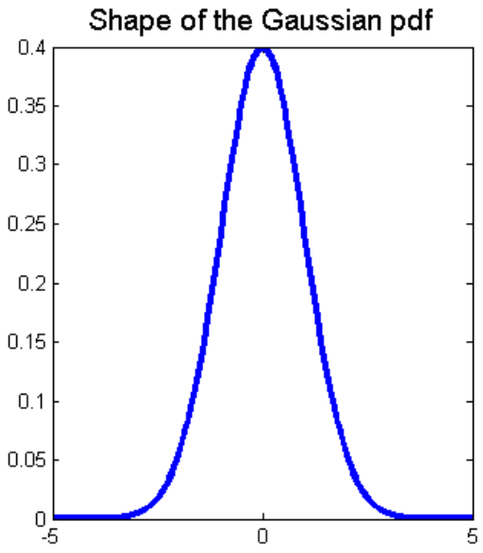
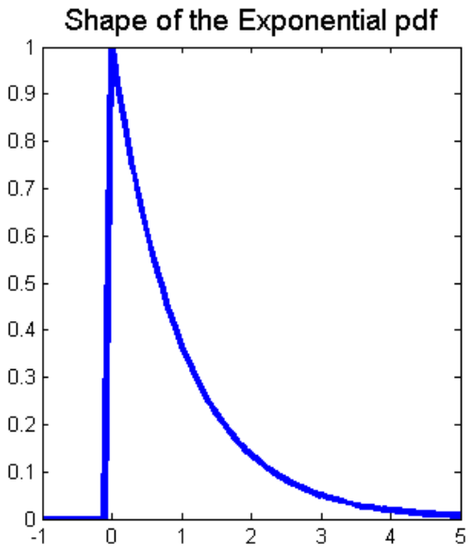
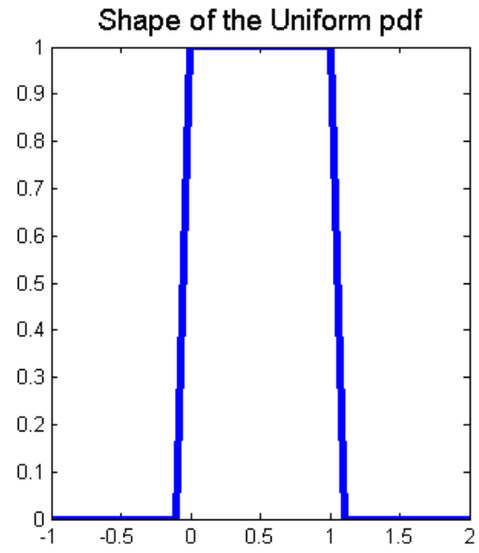
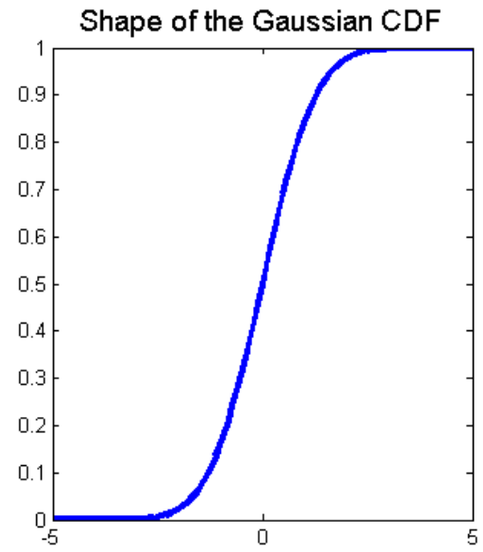
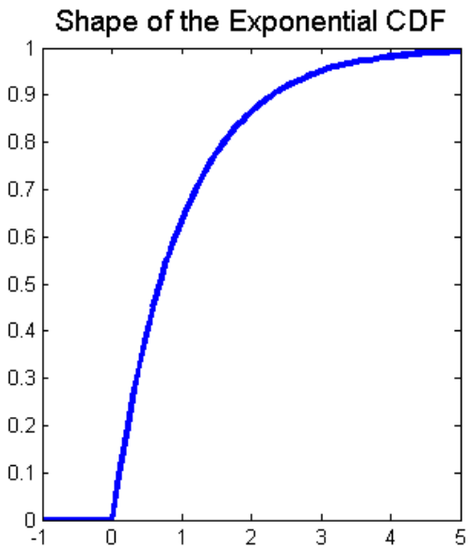
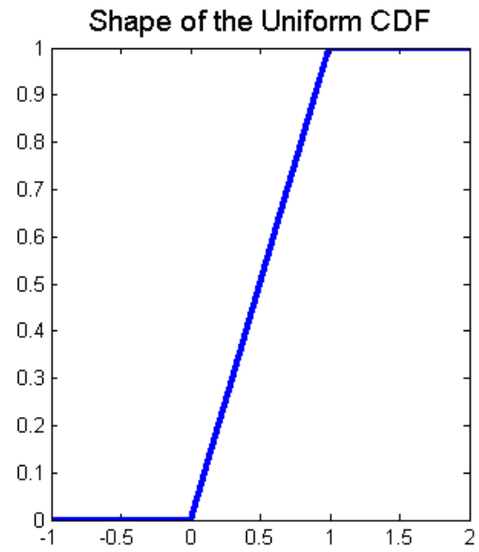
- Probability density function (PDF): derivative of the CDF for variables whose CDFs are differentiable everywhere

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

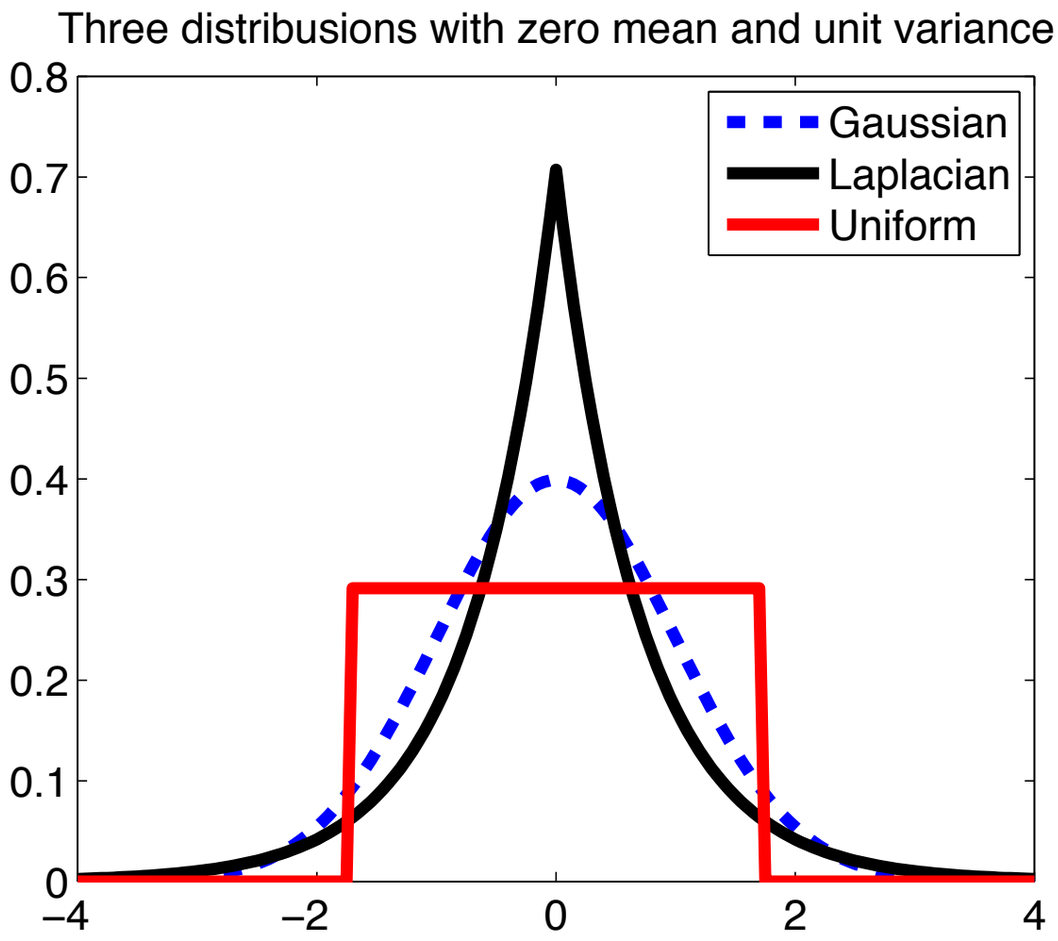


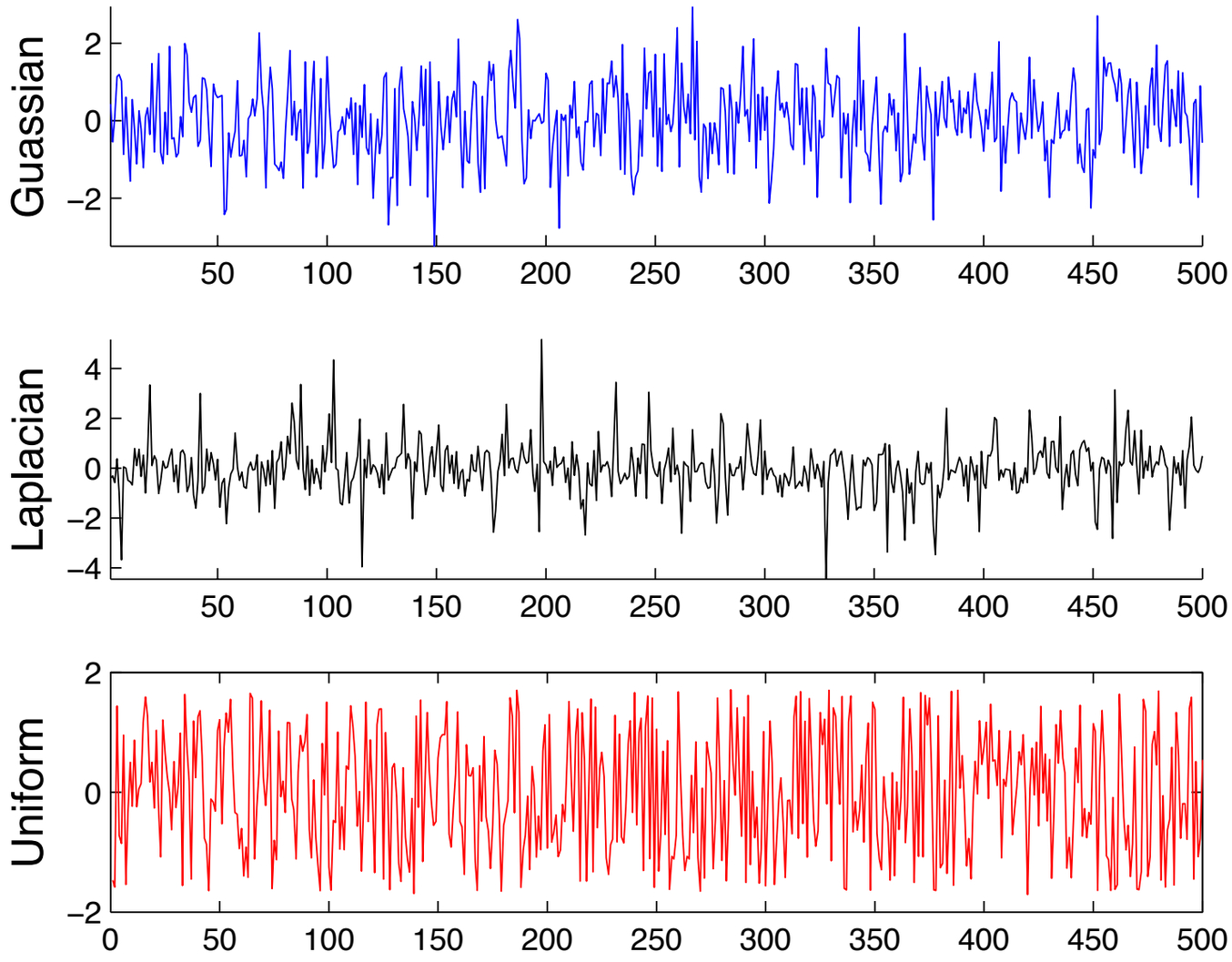
Probability: Examples

- Discrete variables: *Bernoulli*(*p*), *Binomial*(*n*,*p*)...
- Continuous variables:

	Gaussian	Exponential	Uniform
PDF			
CDF			

Some Distributions





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 $P(O_1, O_2, \ldots, O_5)$?
- In practice we often need fewer parameters...
- Divide-and-conquer

Curse of Dimensionality

- More generally, suppose you want to find a mapping from $(x_1,x_2,\ldots,x_n)$ to y

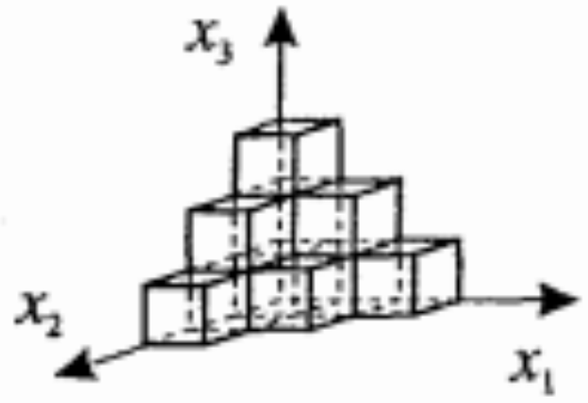


Figure 1.5. One way to specify a mapping from a d-dimensional space $x_1, \ldots, x_d$ to an output variable y is to divide the input space into a number of cells, as indicated here for the case of $d = 3$, and to specify the value of y for each of the cells. The major problem with this approach is that the number of cells, and hence the number of example data points required, grows exponentially with d, a phenomenon known as the 'curse of dimensionality'.

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 $p_{XY}(x,y) = p_X(x)p_Y(y)$, or $p_{X|Y}(x|y) = p_X(x)$ whenever $p_Y(y)\neq 0$

Can you show
if $p_{XY}(x,y) = g(x)h(y)$ for all x and y, they X and Y are independent?

Another Example

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

$$X_1 \to X_2 \to ... \to X_n$$

Conditional Independence

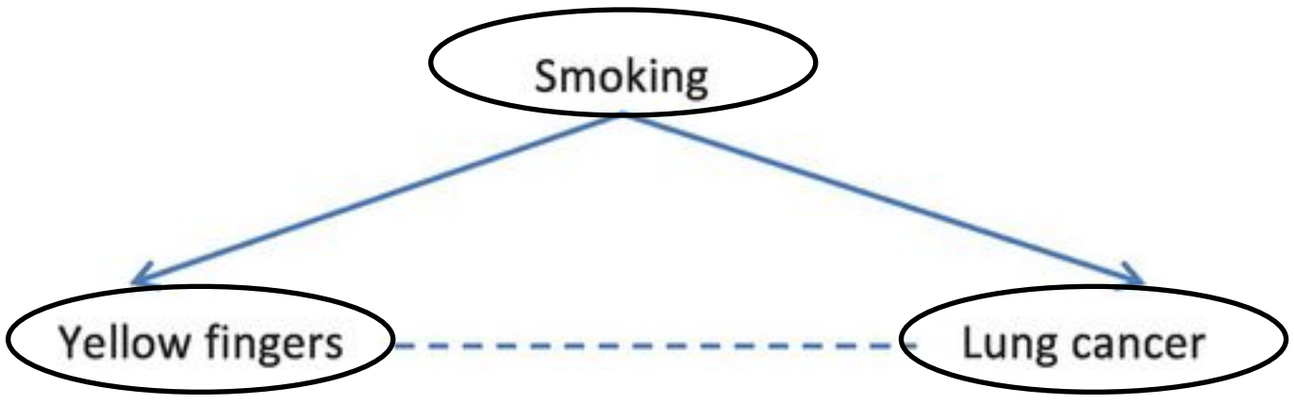
- Two variables X and Y are conditionally independent given Z if $F_{XY|Z}(x,y|z) = F_{X|Z}(x|z)\, F_{Y|Z}(y|z)$ for all values of x, y and z. Equivalently,
  - For discrete variables, $P_{XY|Z}(x,y|z) = P_{X|Z}(x|z)P_{Y|Z}(y|z)$, or $P_{X|Y,Z}(x|y,z) = P_{X|Z}(x|z)$ whenever $P_{YZ}(y,z)\neq 0$
  -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

Some Properties of (Conditional) Independence

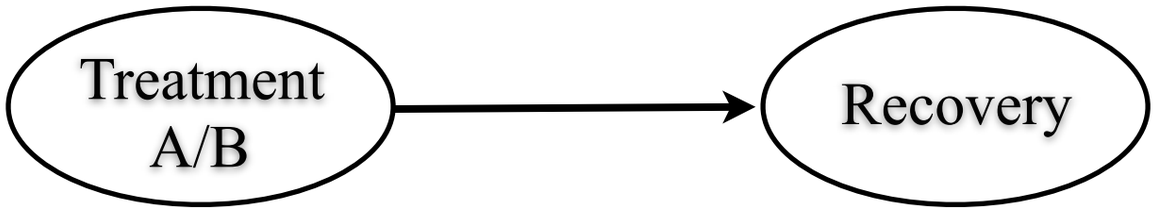
- Symmetry $X \perp\!\!\!\perp Y \quad \Rightarrow \quad Y \perp\!\!\!\perp X$
- Decomposition $X \perp\!\!\!\perp A, B \quad \Rightarrow \quad \text{and} \begin{cases} X \perp\!\!\!\perp A \\ X \perp\!\!\!\perp B \end{cases}$
- Weak union $X \perp\!\!\!\perp A, B \quad \Rightarrow \quad \text{and} \begin{cases} X \perp\!\!\!\perp A \mid B \\ X \perp\!\!\!\perp B \mid A \end{cases}$
- Contraction $\left.\begin{matrix} X \perp\!\!\!\perp A \mid B \\ X \perp\!\!\!\perp B \end{matrix}\right\} \text{and} \quad \Rightarrow \quad X \perp\!\!\!\perp A, B$

Relationship between independence & conditional independence?

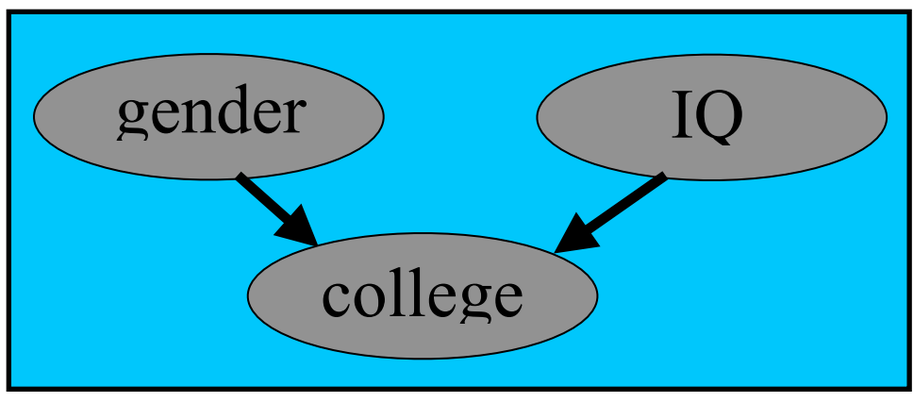
Ways to Produce Dependence



- Common cause



- Causal relation between them



- Conditional dependence given common effect

Expectation, Variance, and Standard Deviation

- Expectation:

$$E[g(X)] \triangleq \sum_x g(x) P_X(x) \quad \text{(for discrete varaibles) } \textit{or}$$

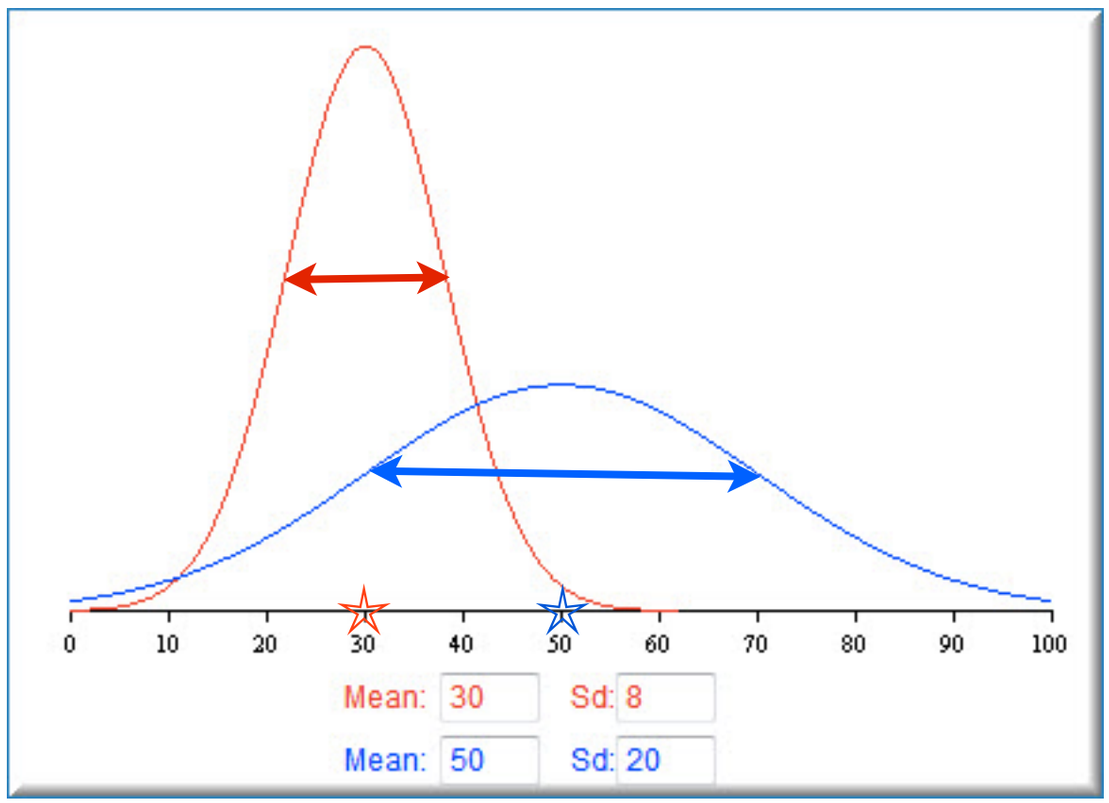
$$E[g(X)] \triangleq \int_{-\infty}^{+\infty} g(x) p_X(x) dx \quad \text{(for continuous varaibles)}$$

- Mean of X: E[X]
- Variance:

$$Var[X] \triangleq E\{[X - E(X)]^2\}$$

- Standard deviation:

$$Std[X] \triangleq \sqrt{Var[X]}$$



Covariance and Correlation

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

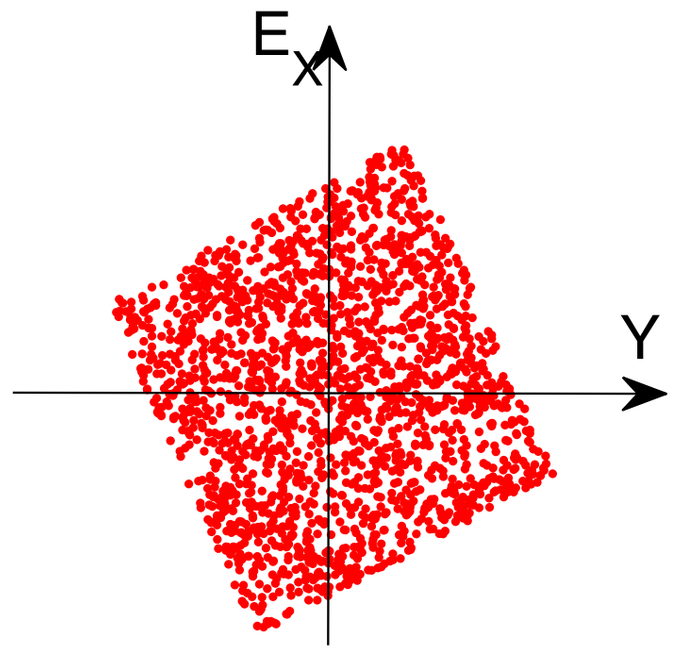
Independence and Uncorrelatedness

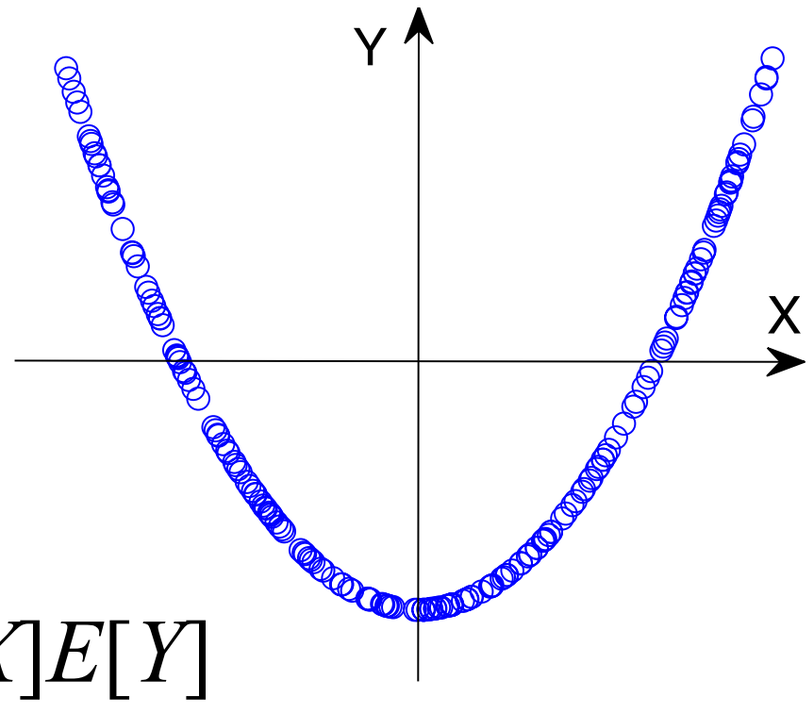
- Independence ⇒ uncorrelatedness
- How about the reverse direction?

Normal distribution !

(Conditional) Independence

- X⫫Y | Z : given *Z*, *Y* not informative to *X*
- Divide & conquer, remove irrelevant info...
- By construction, regression residual is uncorrelated (but not necessarily independent !) from the predictor

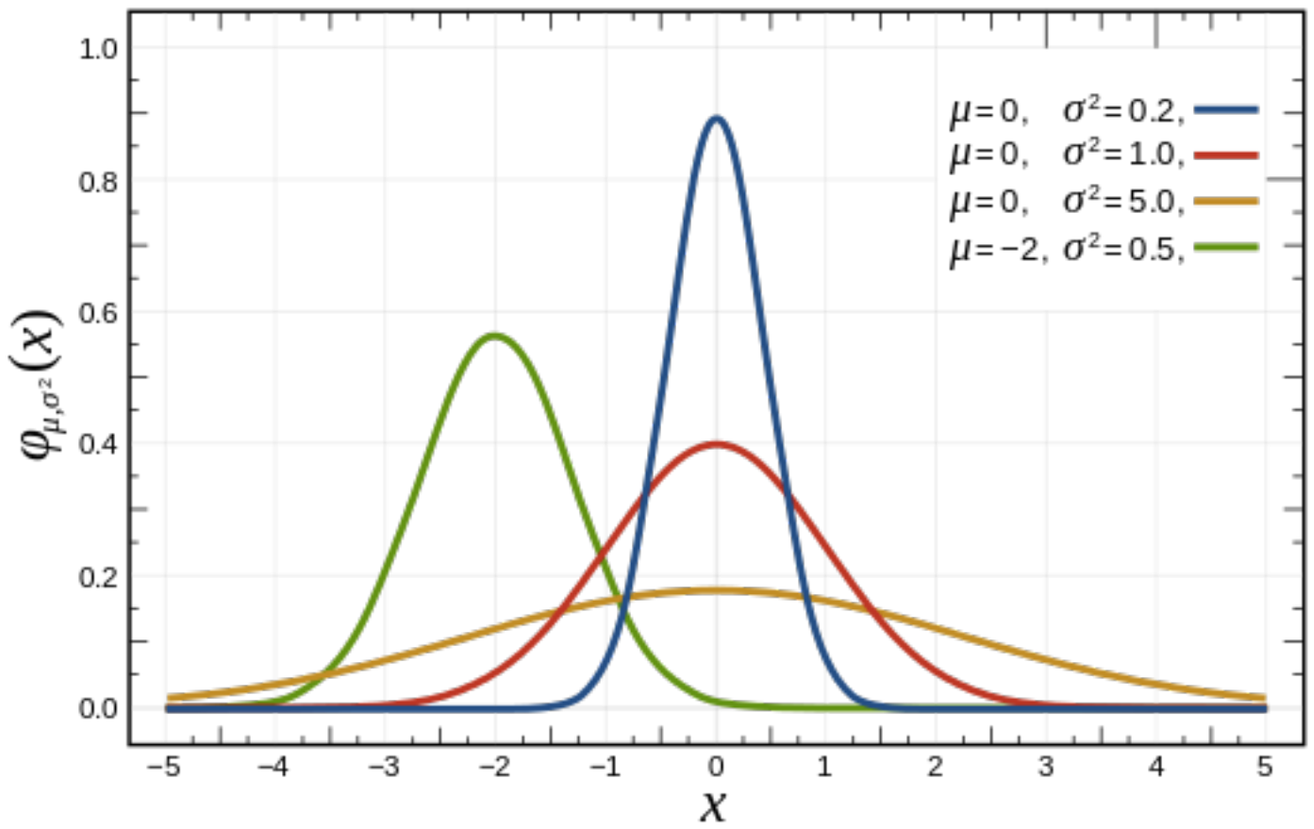


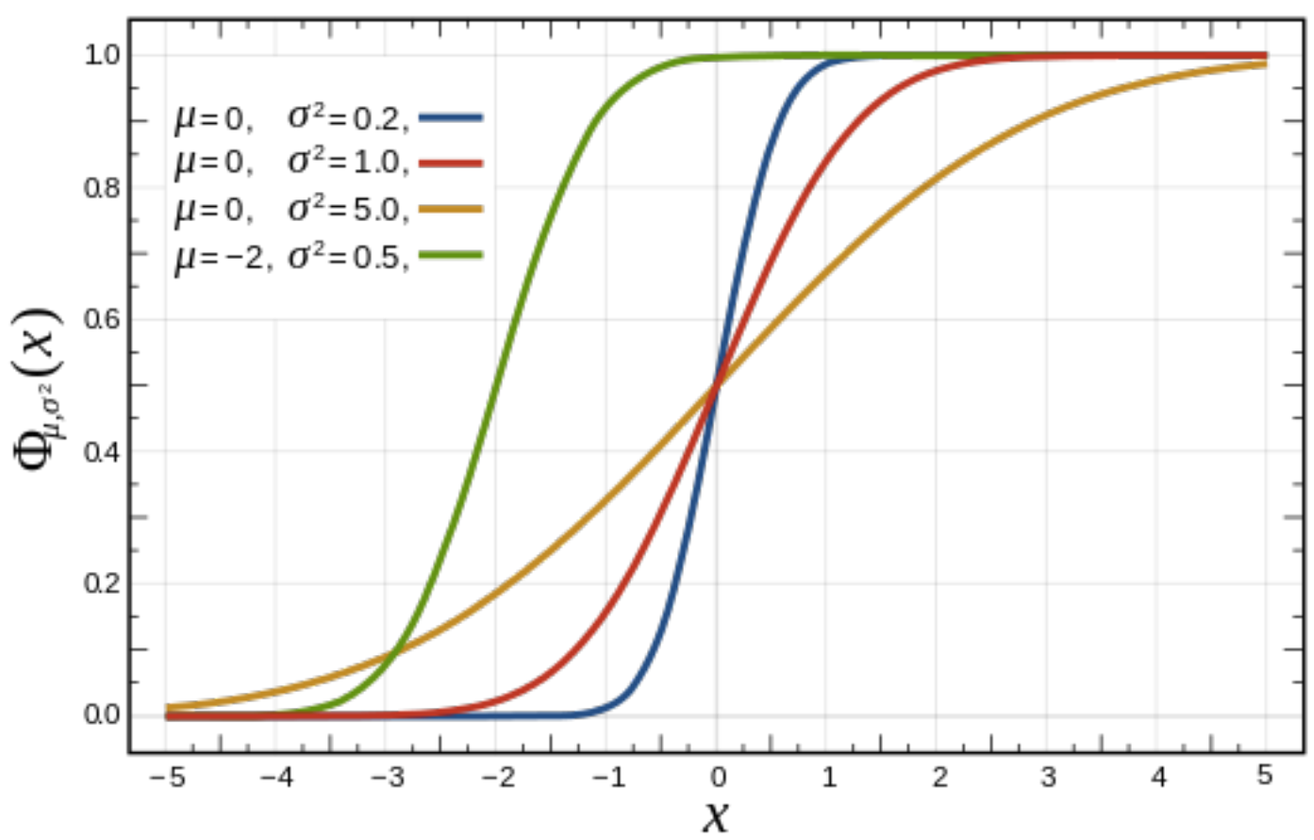


Uncorrelatedness: $E[XY] = E[X]E[Y]$

Normal Distribution

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





Normal Distribution

- Very common distribution (sometimes also informally known as bell curve)
- PDF specified by mean μ and standard deviation σ (or variance σ²):

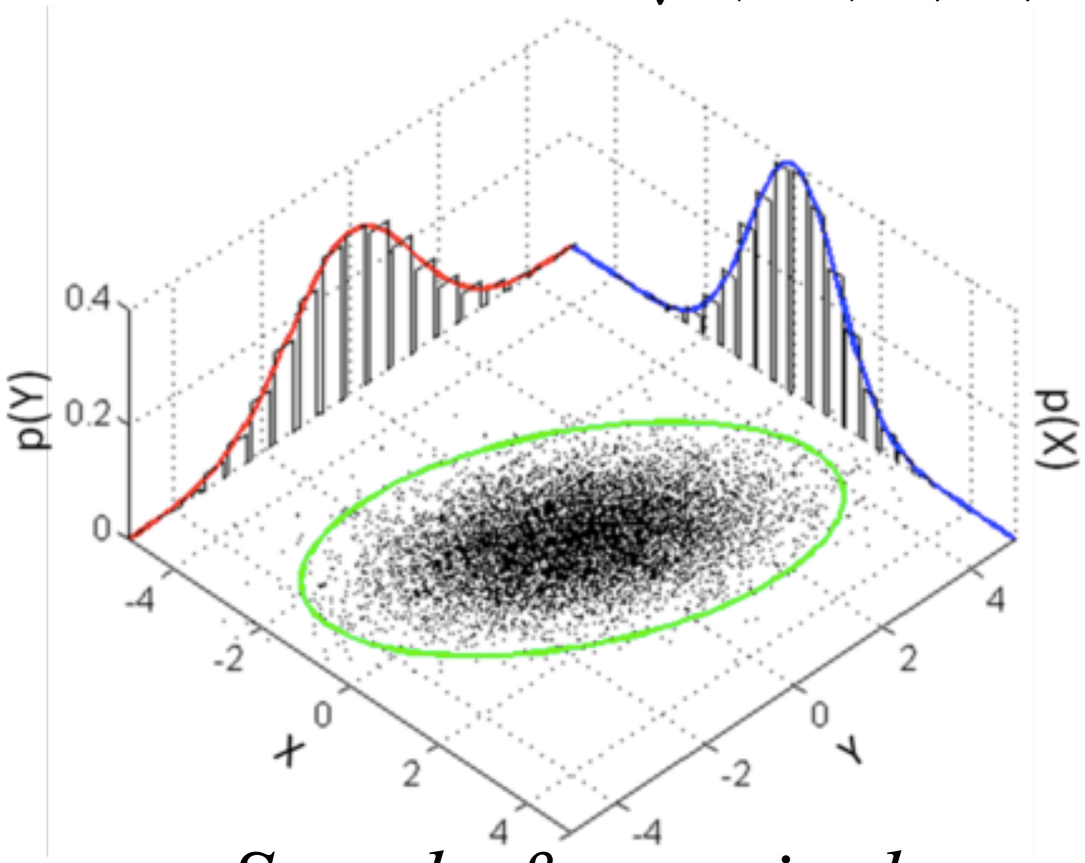
$$p_X(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

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

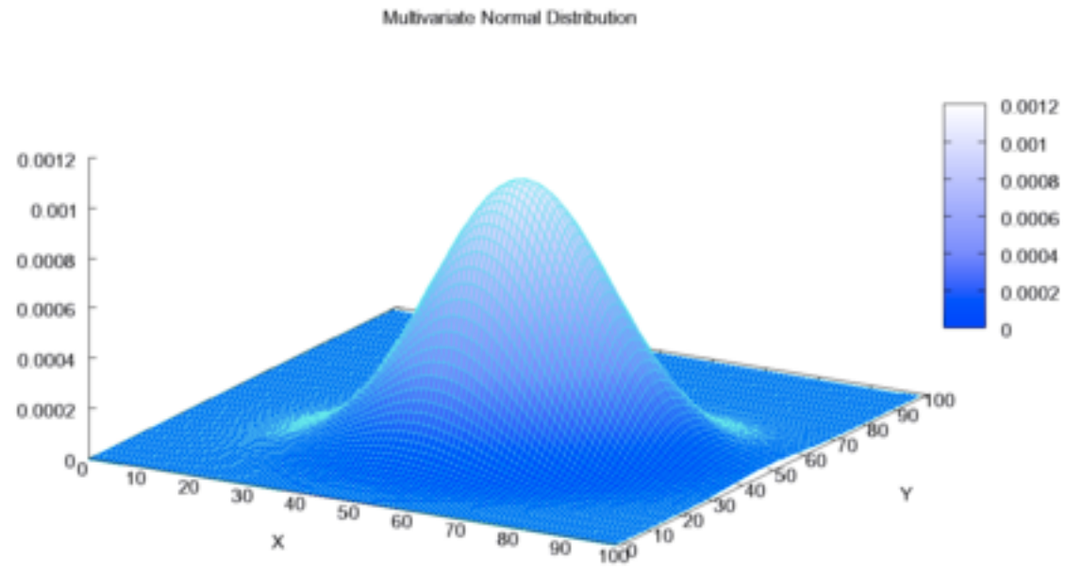
Multivariate Normal Distribution

- PDF for point $\mathbf{x} = (\mathbf{x}_1,..., \mathbf{x}_k)$, specified by mean $\boldsymbol{\mu}$ and covariance matrix :

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



Sample & marginal



pdf

Some Properties

- 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!
- Relation to χ^2:

Let $Q = \sum_{i=1}^{k} Z_i^2$ with independent *standard normal* variables Z_i, $Q \sim \chi_k^2$

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:

$$D_{\mathrm{KL}}(P\|Q) = E_{X\sim P}\Big[\log\frac{P(X)}{Q(X)}\Big] = \sum_i P(i)\,\log\frac{P(i)}{Q(i)}.$$

$$D_{\mathrm{KL}}(p(x)\|q(x)) = E_{X\sim p}\Big[\log\frac{p(X)}{q(X)}\Big] = \int_{-\infty}^{\infty} p(x)\,\log\frac{p(x)}{q(x)}\,\mathrm{d}x.$$

- Non-negative; asymmetric; zero iff identical

Are Two Variables Independent?

- Natural measure of statistical dependence: mutual information

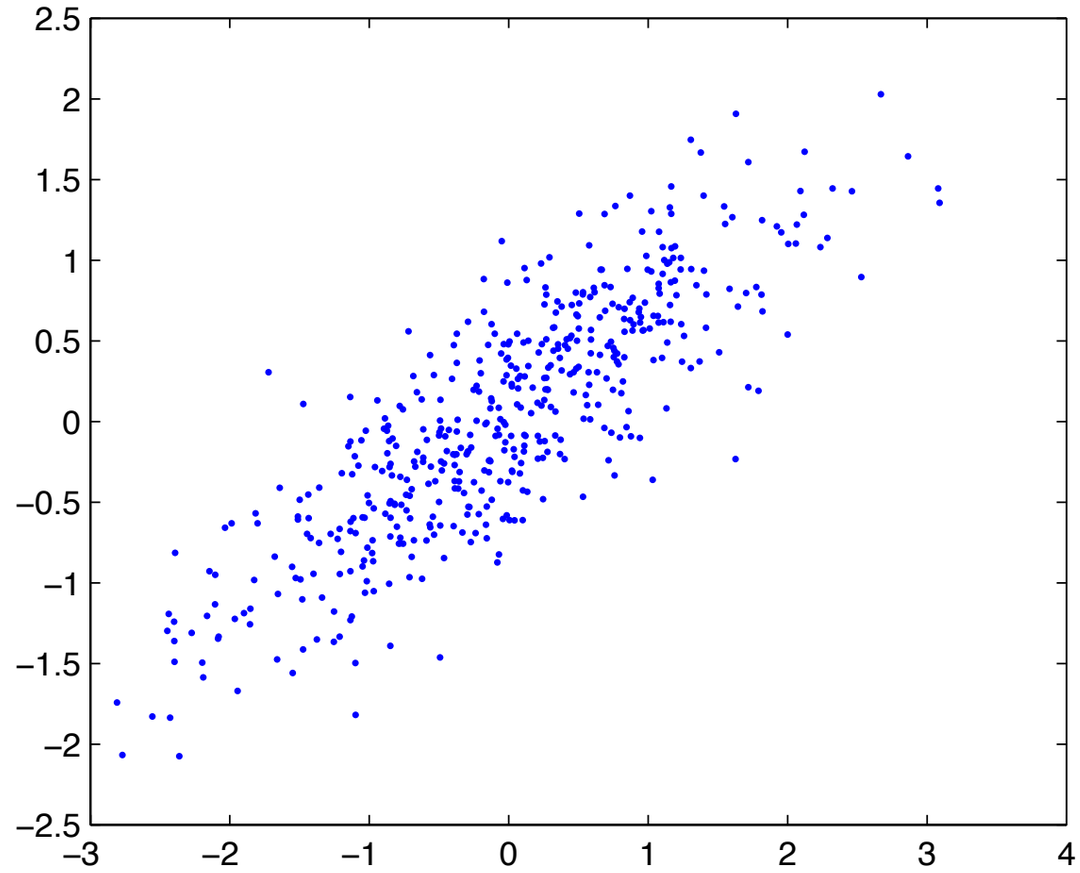
$$I(X;Y) = \sum_y \sum_x P(x,y) \log \left(\frac{P(x,y)}{P(x)\,P(y)} \right),$$

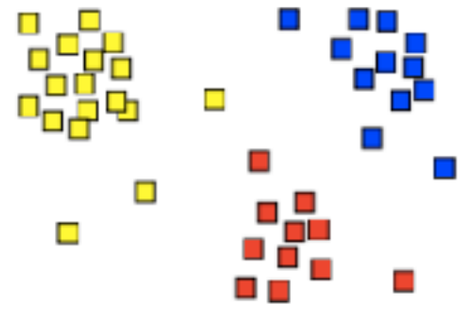
$$I(X;Y) = \int \int p(x,y) \log \left(\frac{p(x,y)}{p(x)\,p(y)} \right) dx\,dy,$$

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

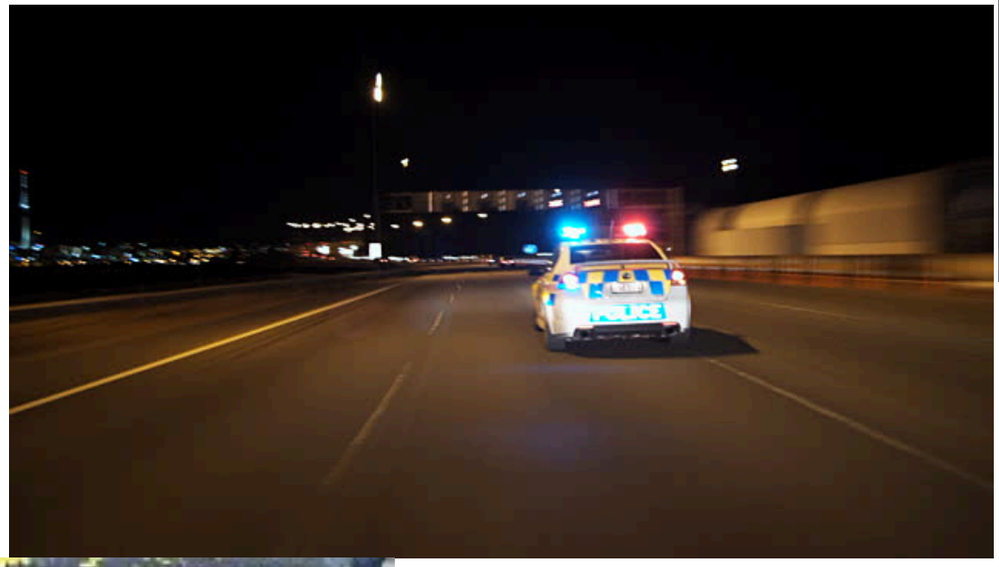
Let's Come Closer to Reality...

- Find knowledge from data, which has randomness. E.g.,
- Bayesian inference
- parameter estimation and hypothesis test
- learning
  - supervised learning
  - unsupervised learning
  - causal discovery





How to Update Our Belief?



Bayes' Rule

Thomas Bayes (1701-1761)

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

- Use evidence to update probabilities
- How to find $P(B)$?

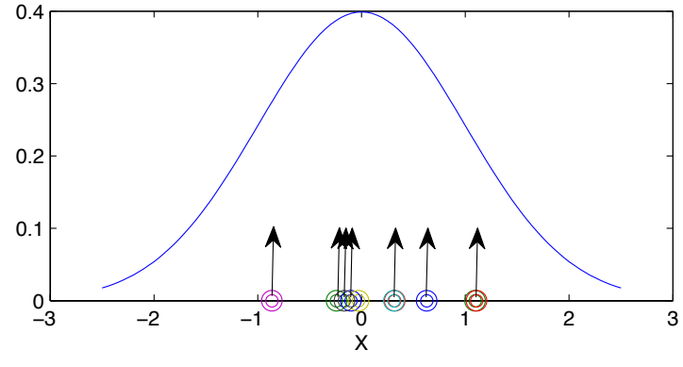
$$P(A_i \mid B) = \frac{P(B|A_i)P(A_i)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_{k=1}^{m} P(B|A_k)P(A_k)}$$

Bayes' Rule: Example

$$P(A_i \mid B) = \frac{P(B|A_i)P(A_i)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_{k=1}^{m} P(B|A_k)P(A_k)}$$

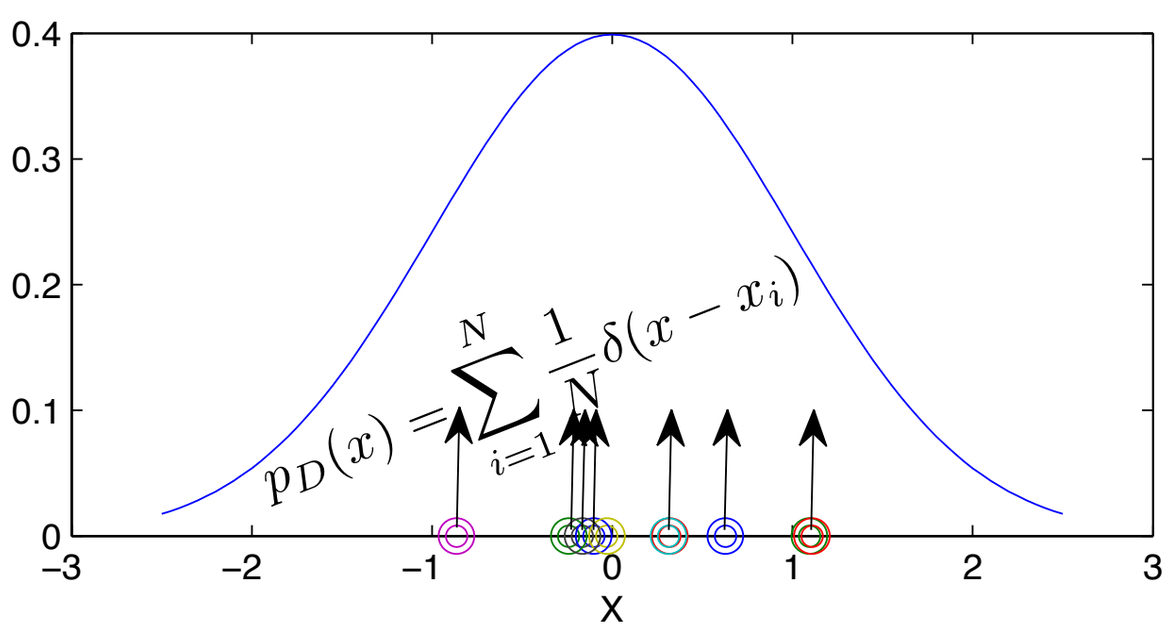
- Suppose a drug test is 99% sensitive and 99% specific. That is, the test will produce 99% true positive results for drug users and 99% true negative results for non-drug users.
- Suppose that 0.1% of people are users of the drug.
- If a randomly selected individual tests positive, what is the probability he or she is a user?
- P(User | +) ≈ ? *A. 0.1,* *B. 0.4,* *C. 0.9*

Maximum Likelihood Estimation



- Estimate characteristics of the model distribution from the sample
  - so that the distribution underlying the sample is close to the model distribution
- Suppose we have functional form of the pdf/pmf $f(x;\theta)$ with unknown parameters $\theta \in \Theta$
- Aim to find a point estimator of θ, i.e., a member of $\{f(x;\theta) \mid \theta \in \Theta\}$ as the most likely pmf/pdf

How to Find the Best Parameter



- $f(x;\theta)$ should be as close as possible to $p_D(x) = \sum_{i=1}^{N} \frac{1}{N} \delta(x - x_i)$
- How? Kullback-Leiber divergence...
- **Maximum likelihood estimator**:

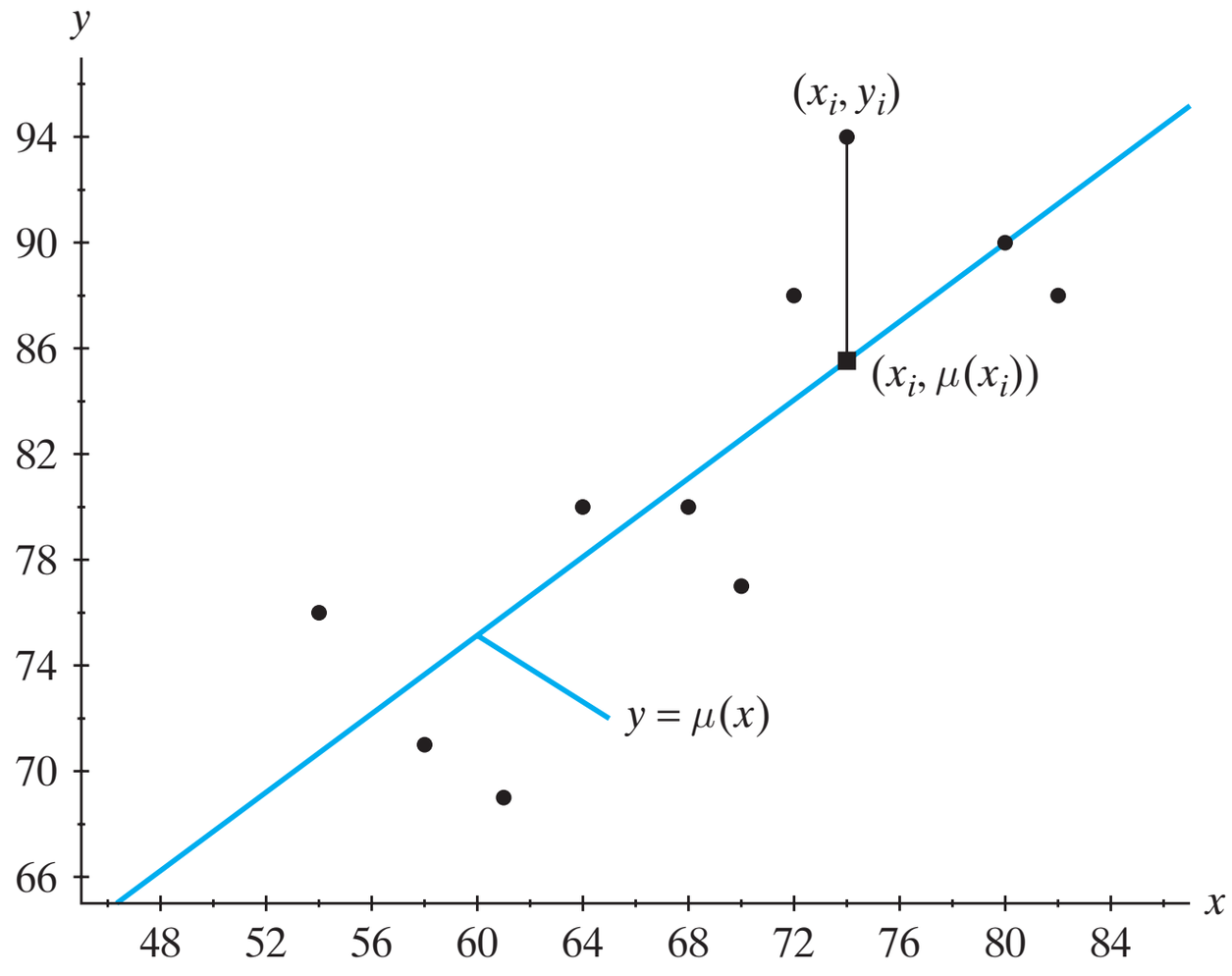
$$\hat{\theta}_{ML} = \arg\max_{\theta} f(x_1;\theta) f(x_2;\theta) \ldots f(x_N;\theta) = \arg\max_{\theta} \sum_{i=1}^{N} \log f(x_i;\theta)$$

Likelihood function $L(\theta;x_1,x_2,\ldots,x_N)$ — *Log likelihood*

- Crucial assumption: I.I.D.
- Closed-form solution or numerical optimization

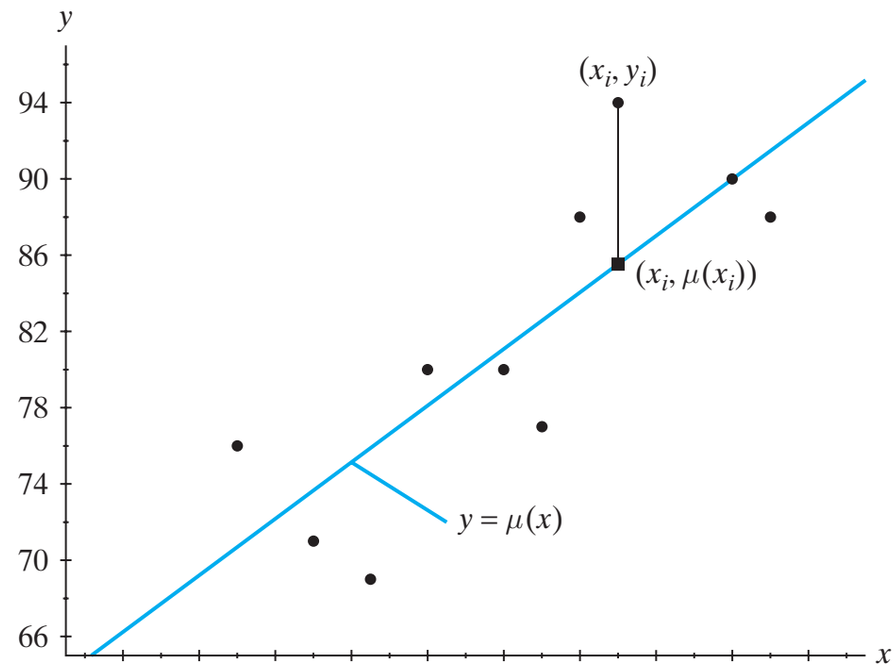
Linear Regression

- To explain/predict Y with X
- $Y = aX + u + \varepsilon$
- Y: dependent variable; X: explanatory / independent variable.
- How to find the regression line $\hat{y} = \alpha x + c$ from data points $(x_1, y_1), \ldots, (x_n, y_n)$?



Linear Regression: MLE

- $Y = aX + u + \varepsilon$, *where* $\varepsilon \sim N(0,\sigma^2)$
- MLE:



$$L(a, u, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y_i - ax_i - u)^2}{2\sigma^2} \right]$$

$$\log L(a, u, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{\sum_{i=1}^{n} (y_i - ax_i - u)^2}{2\sigma^2}$$

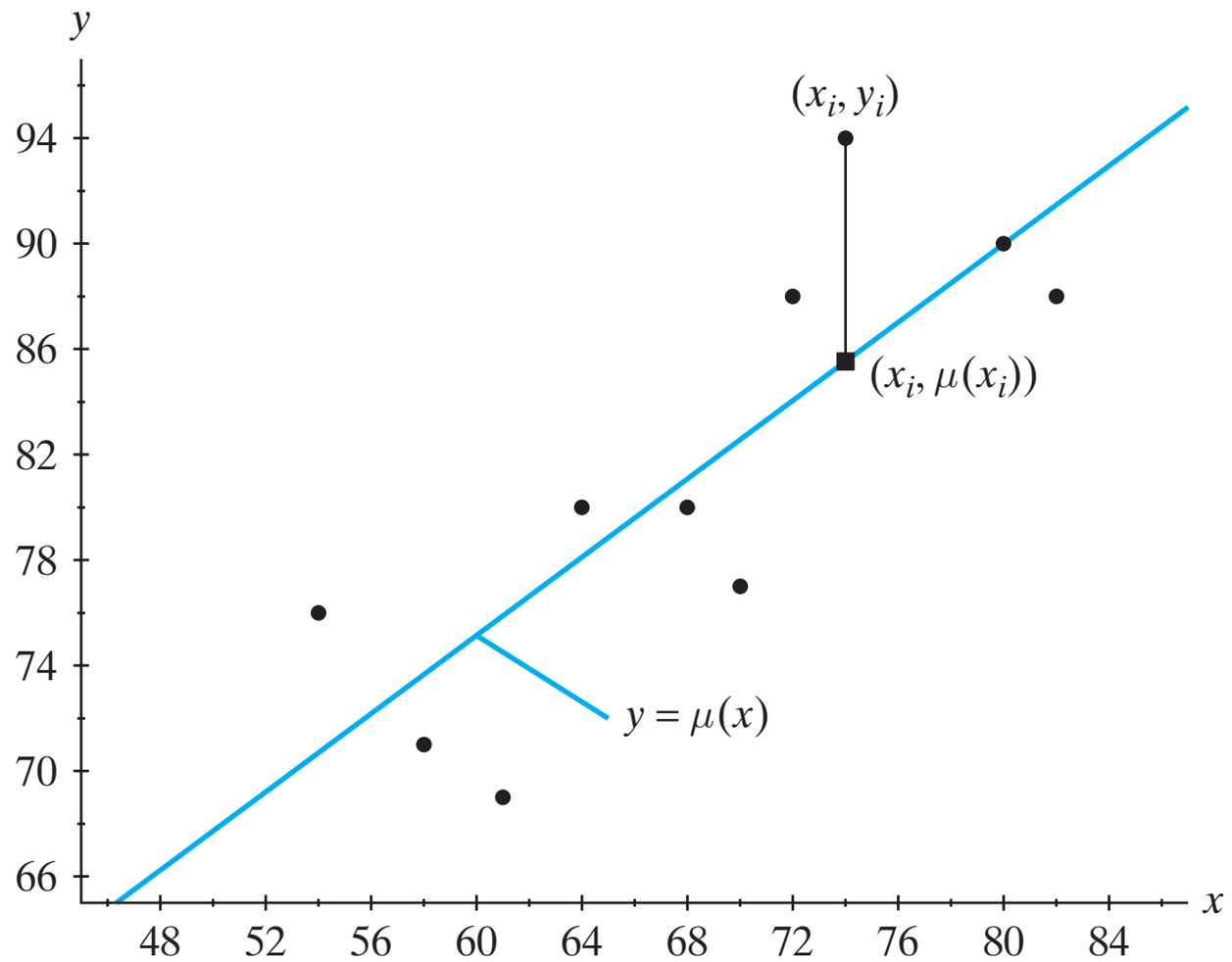
$$\hat{a} = \frac{s_{XY}}{s_X^2}, \quad \hat{u} = \bar{y} - \hat{a}\bar{x}, \quad \hat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)2$$

Linear Regression: Least Squares

- regression line $\hat{y}=\alpha x + c$
- Method of least squares:

$$\text{Minimize } \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

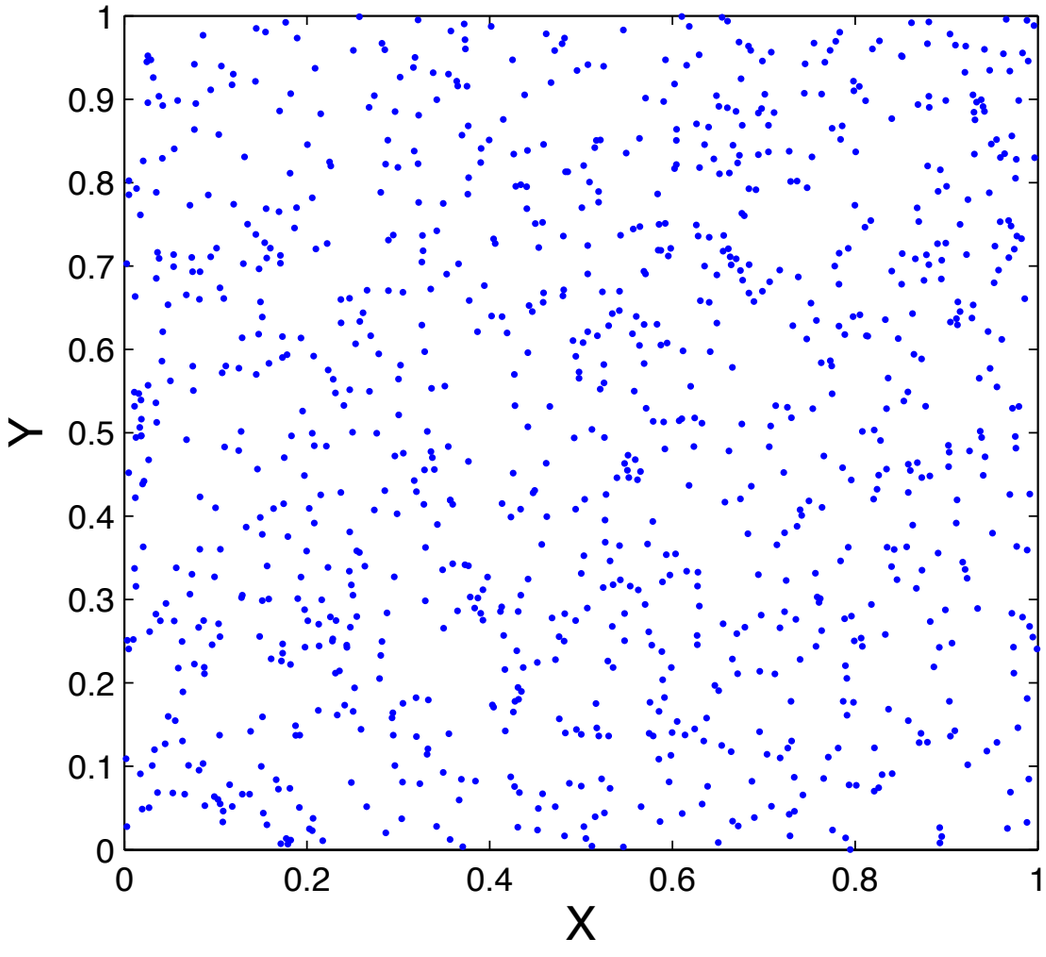
$$= \sum_{i=1}^{n}(y_i - \alpha x_i - c)^2$$

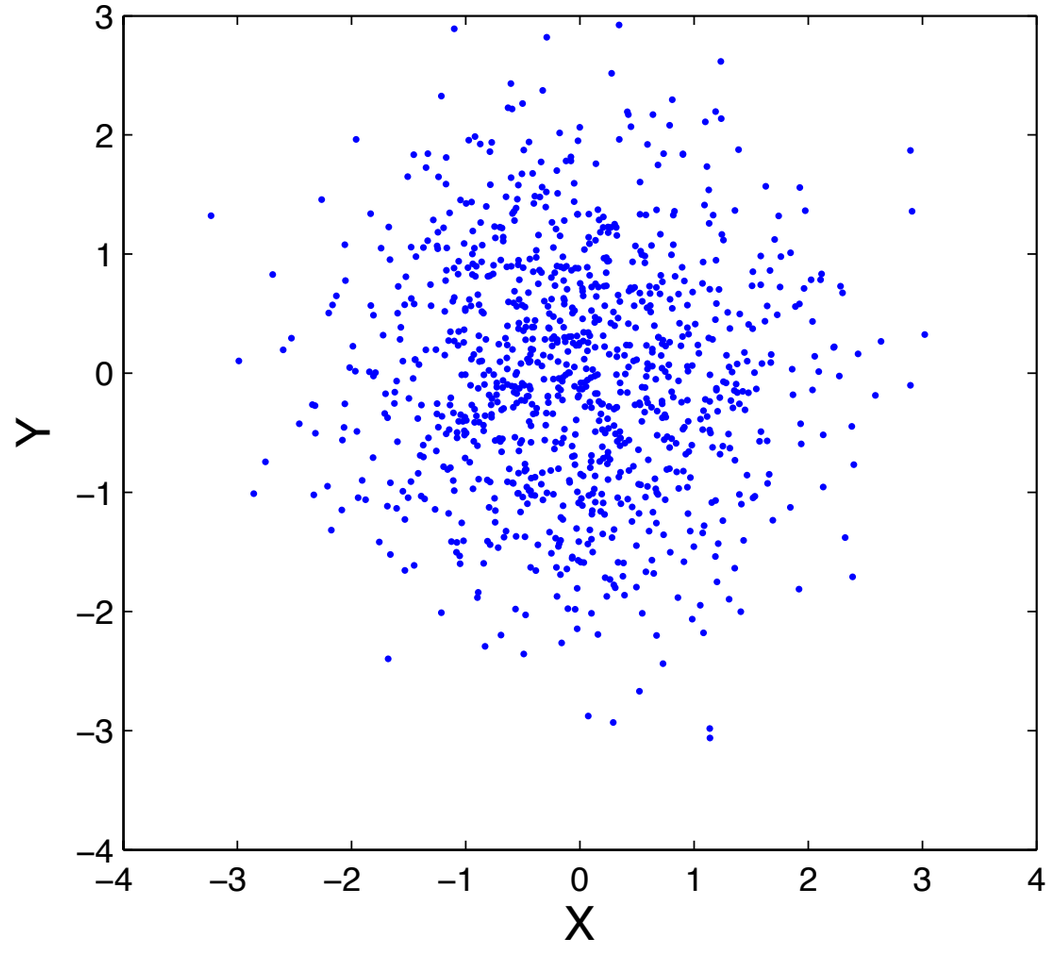


$$\alpha = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{s_{XY}}{s_X^2}$$

$$c = \bar{y} - \alpha\bar{x}$$

Can You See Whether They Are Independent?





- $p_{XY}(x,y)$ has the same shape for different values of y...

Independence Test: Discrete Case

- Set hypotheses:
- Formulate a plan:

r \ *c*	Have you taken an online course?		
	Yes	**No**	
Men	43	63	106
Women	95	113	208
	138	176	314

H_0: Variables are independent.
H_a: Variables are not independent.

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.

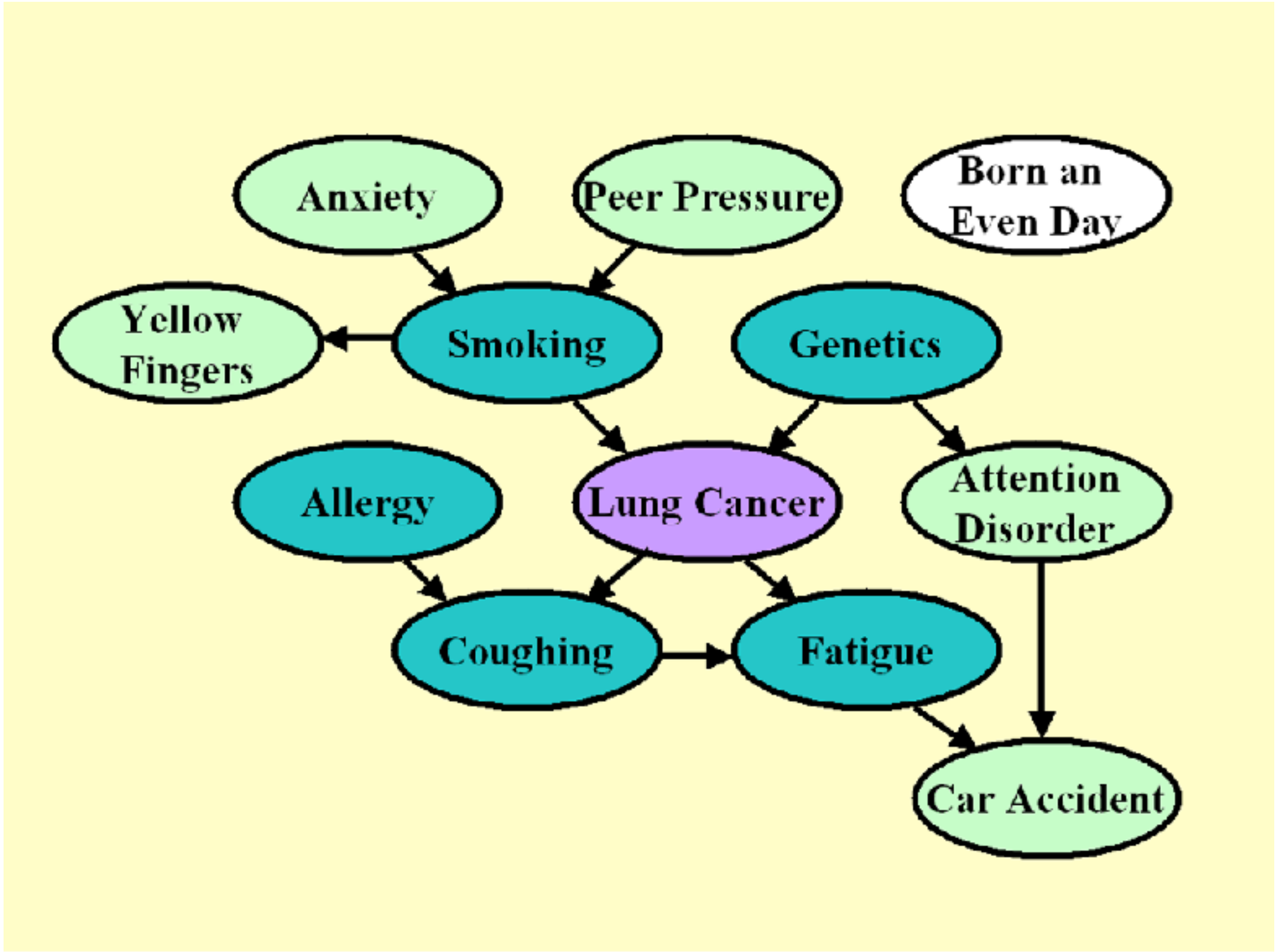
James Clerk Maxwell (1850)

- Why causal discovery?
  - understanding, manipulation, prediction, fusion...

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

−Democritus (460 B.C. – 370 B. C.)

Intuitive Way of Representing and Visualizing Relationships



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
- We mainly consider **directed acyclic** graphs (DAGs)

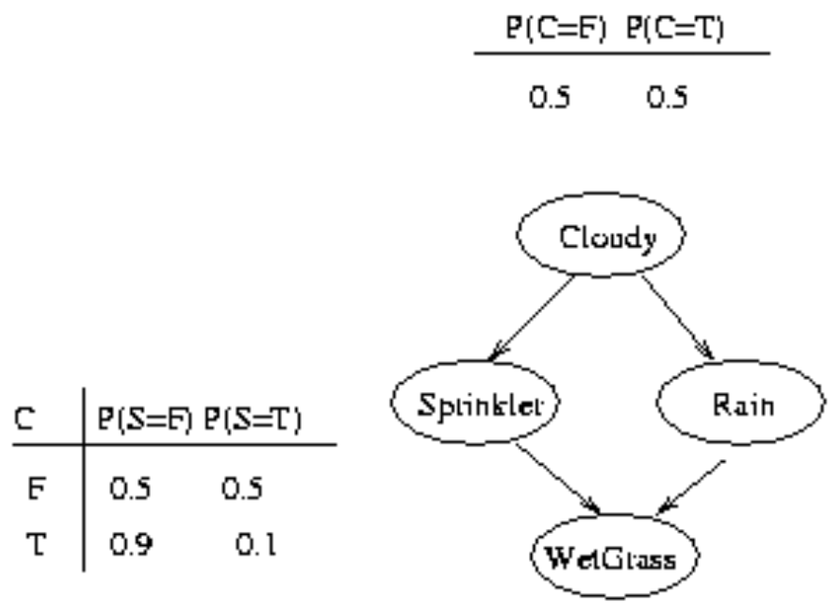
Probability theory + graph theory

represent uncertainty & interface models to data

intuitively appealing interface for humans

Directed Graphical Models

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

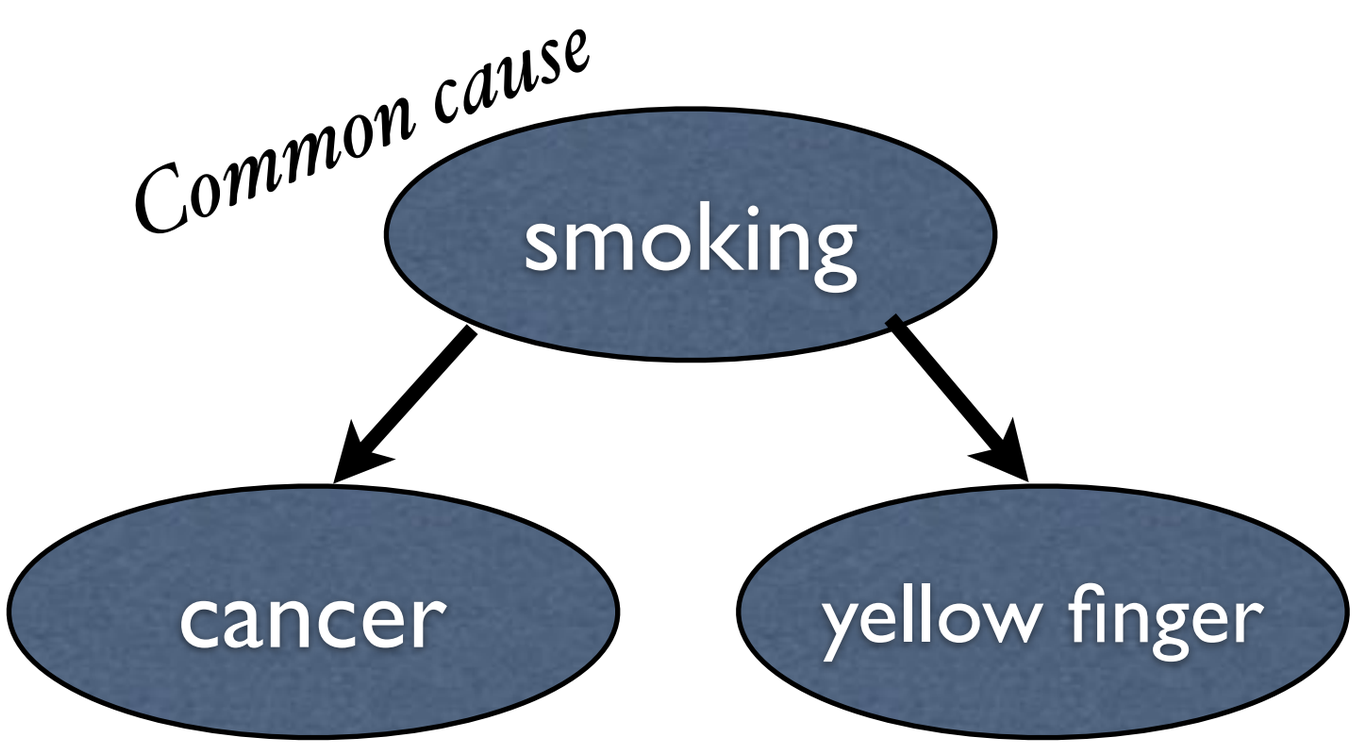


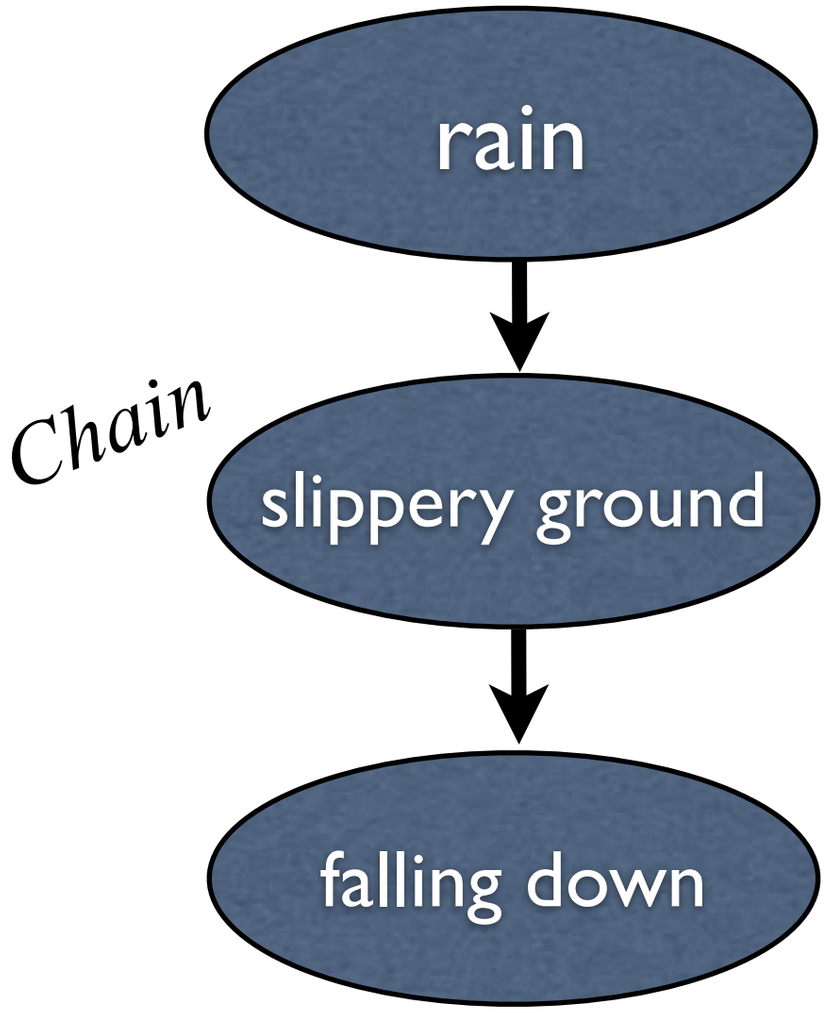
Terms:
nodes, edge, adjacent, path;
parents, children, spouses,
ancestors, descendants,
Markov blanket

Bayesian Networks: Story

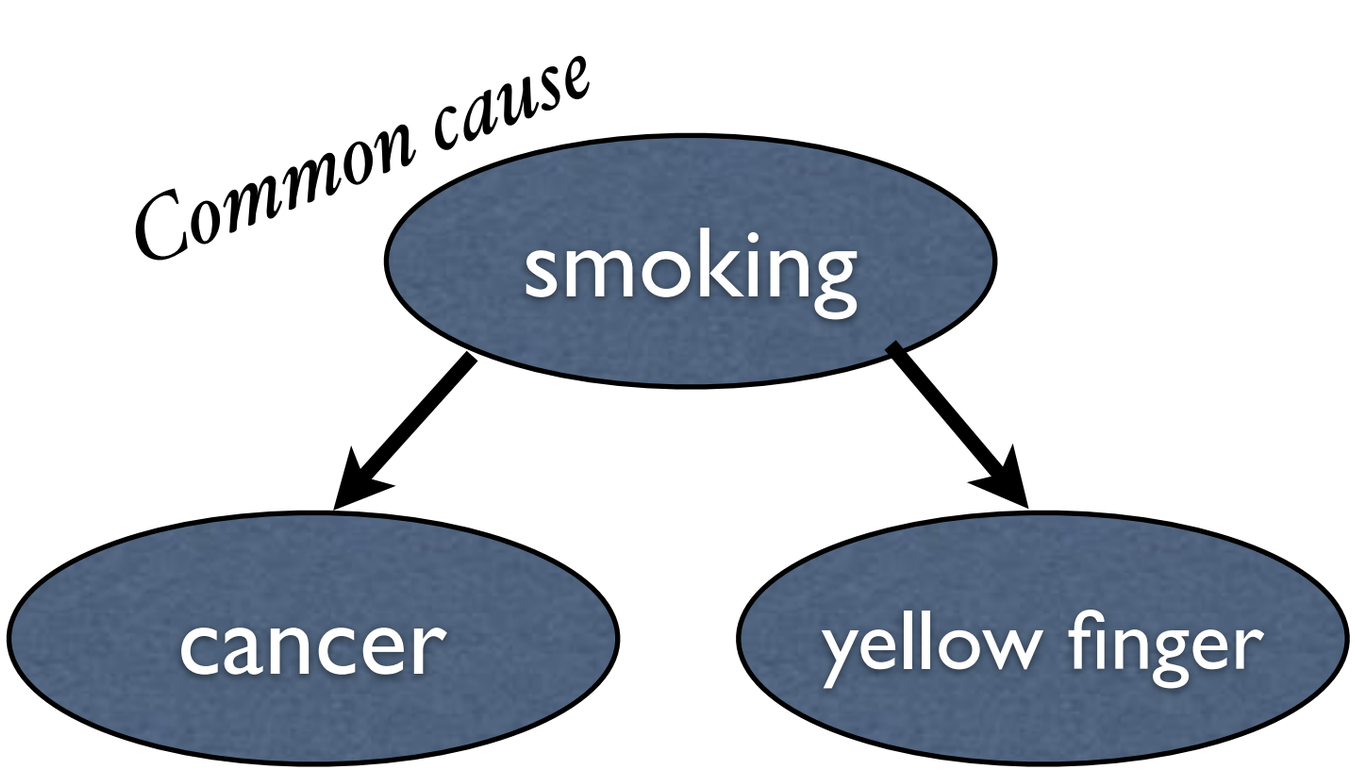
- Breakthrough in early 1980s (by Pearl et al.)
- In a joint probability distribution, every variable is, in general, related to all other variables.
- Pearl and others realized:
  - It is often reasonable to make the assumption that each variable is directly related to only a few other variables
  -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

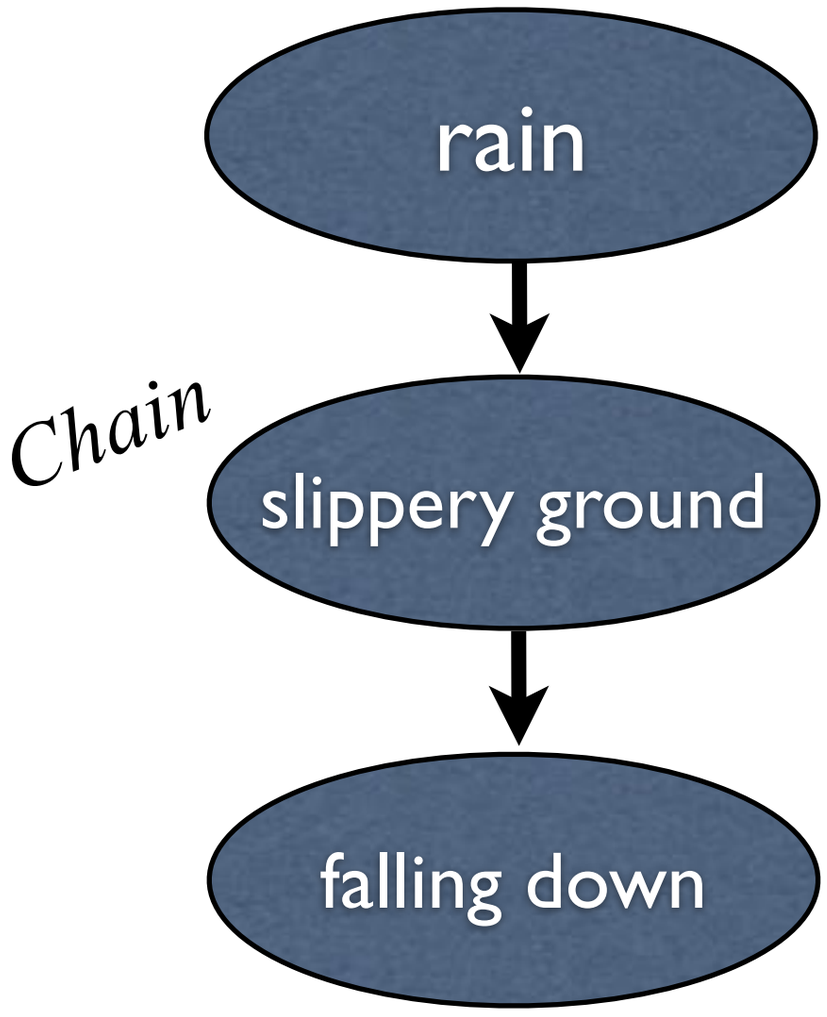
What Independence Relationships Can You See?





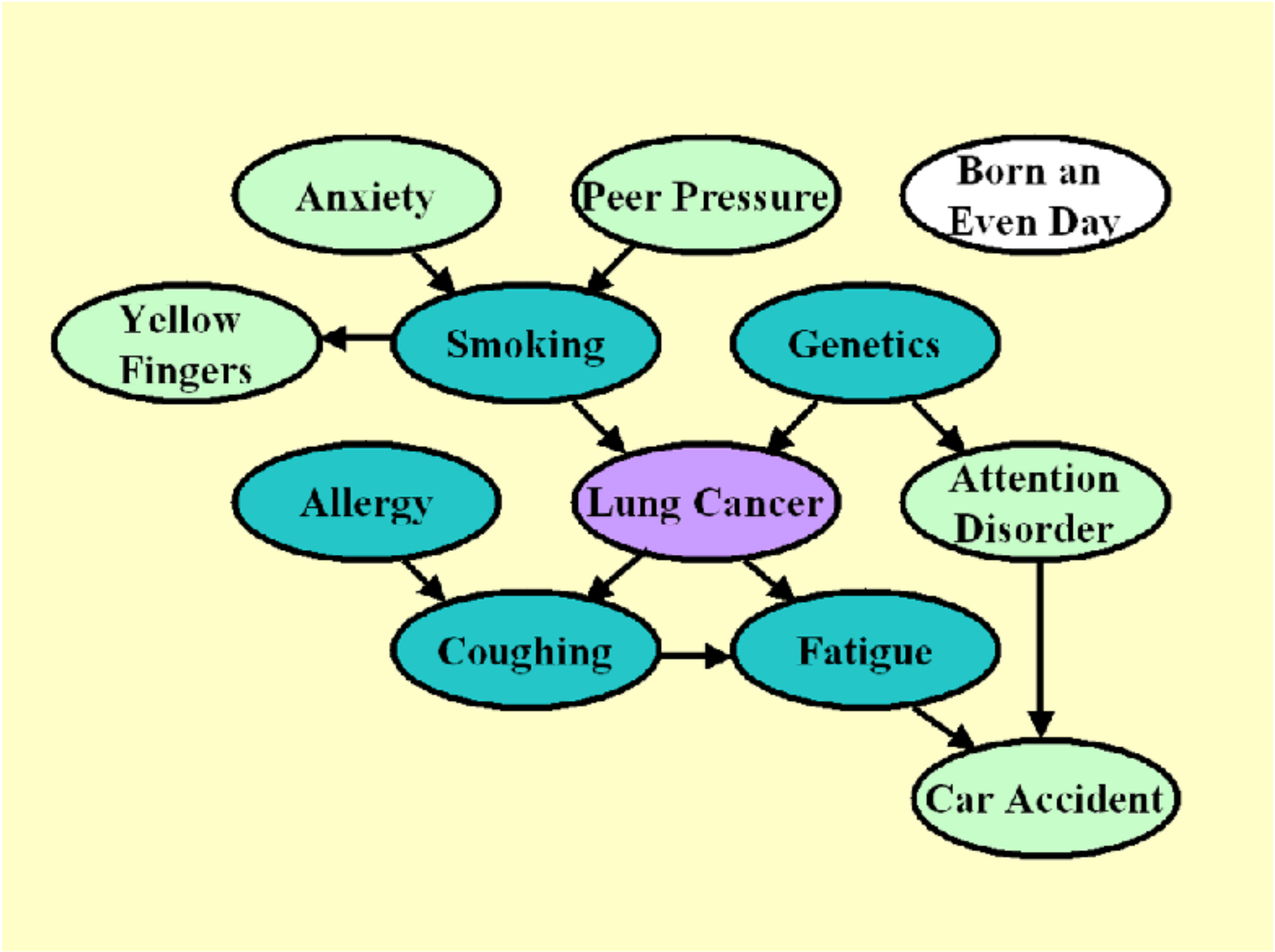
(Local) Markov Condition





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

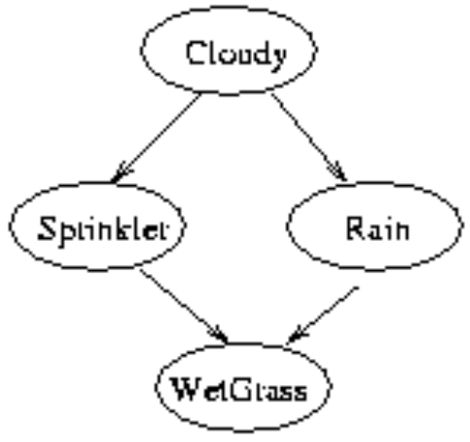
For Instance, What Independence Relations can You See?



Factorization According to Directed Graphs

- Chain rule of probability gives

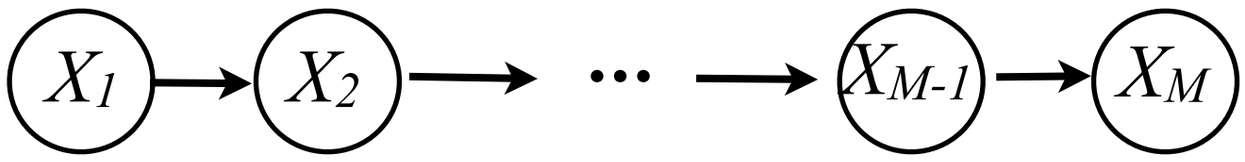
  P(C,S,R,W) = P(C) P(S|C) P(R|C,S) P(W|C,S,R)



- 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:
  - $P(X_1,\ldots,X_n) = \prod_i P(X_i \mid PA_i)$
- Remember this example?

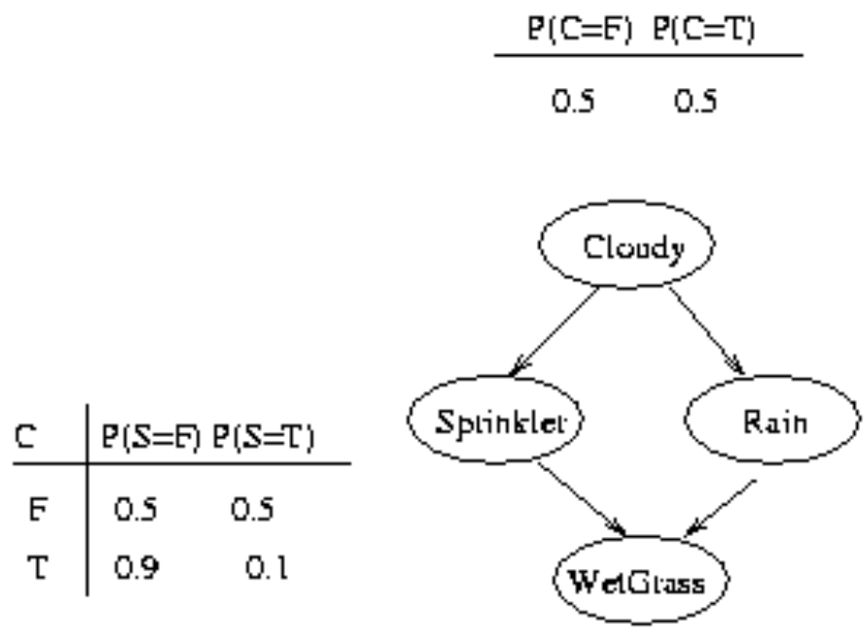


If we aim to represent causal info, is CI info enough?

X→Y or X←Y?

Tasks Related to Bayesian Networks

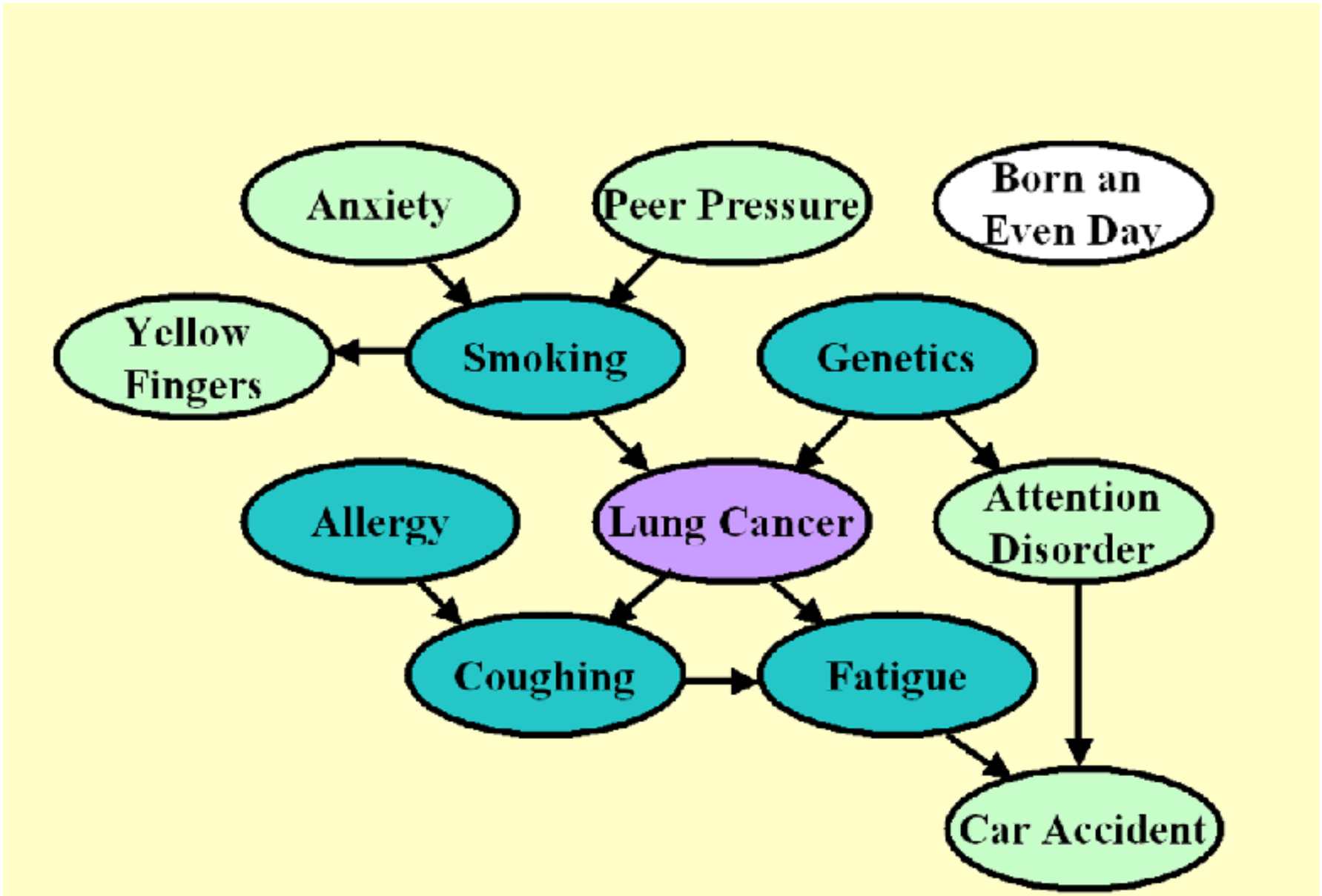
- **Probabilistic inference**: Calculate P(variables of interest | observed variables)
  - Most common task where we want to use Bayesian networks
    - How to find $P(S=1|W=1)$? $P(R=1|W=1)$?
- **Parameter learning**
- **Structure learning**: Learning the structure of the graphical model from observations



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

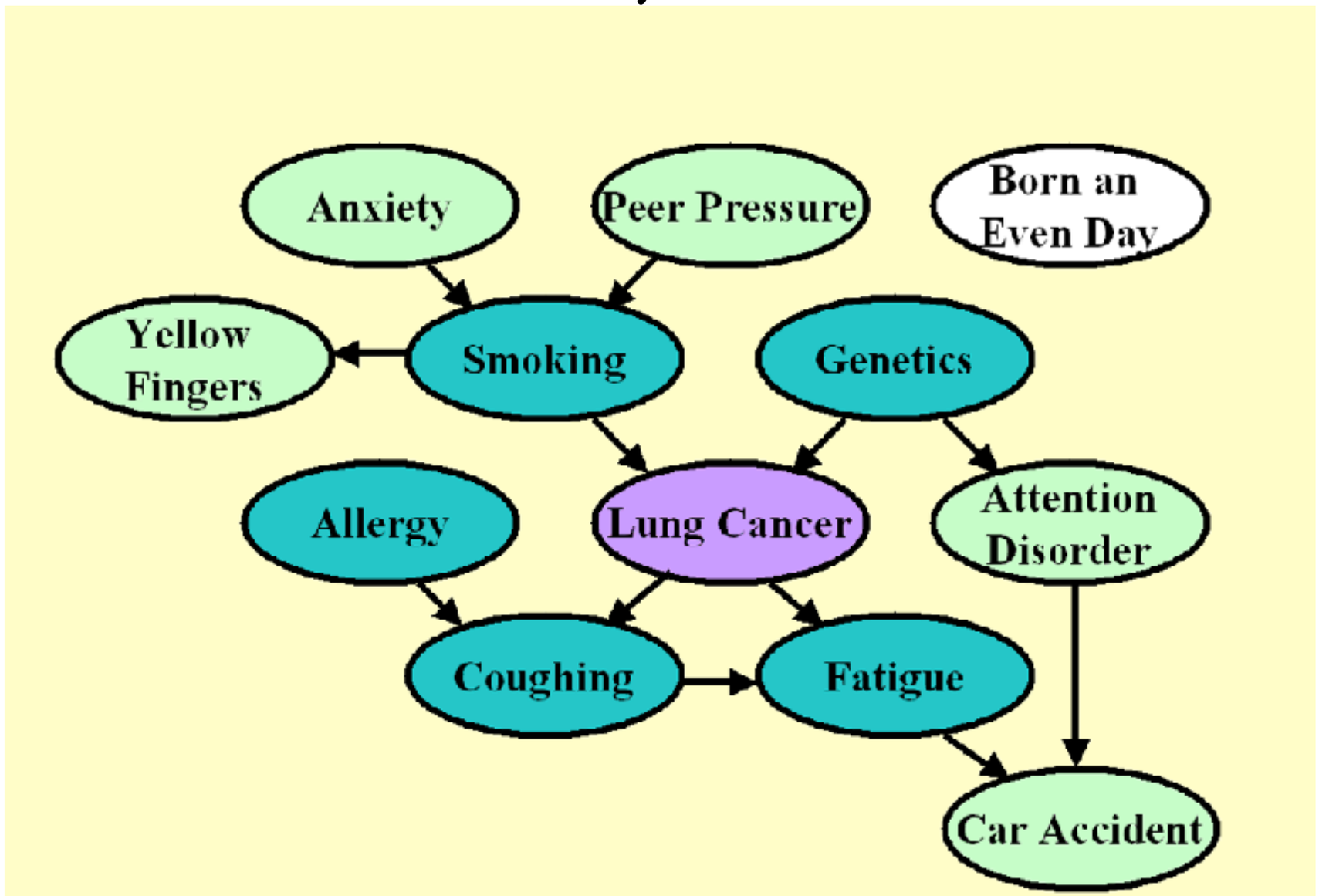
Is Local Markov Condition Enough?

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



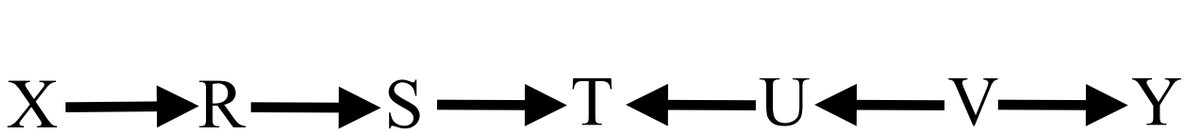
D-Separation Tells Conditional Independence

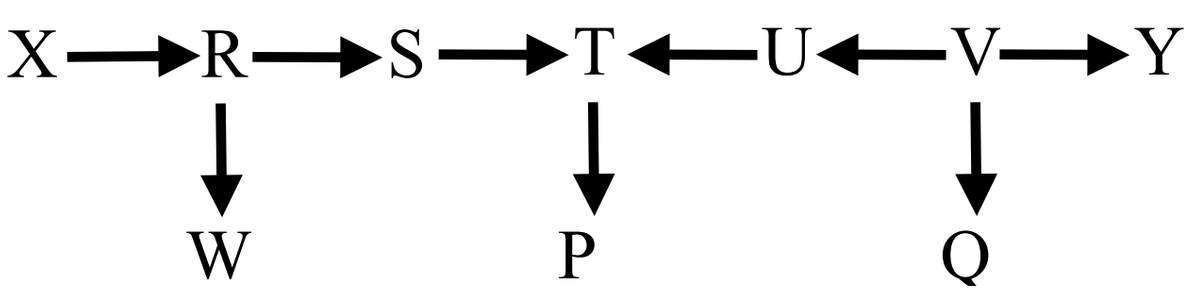
- If every path from a node in **X** to a node in **Y** is d-separated by **Z**, then **X** and **Y** are always conditionally independent given **Z**
- d: directional… You will see why



D-Separation: Story

- Developed by Pearl, Verma, Dan Geiger in the mid 1980s
- To make it possible for a robot to represent causal beliefs, incorporate uncertainty, and learn from its interactions with the world in an efficient way

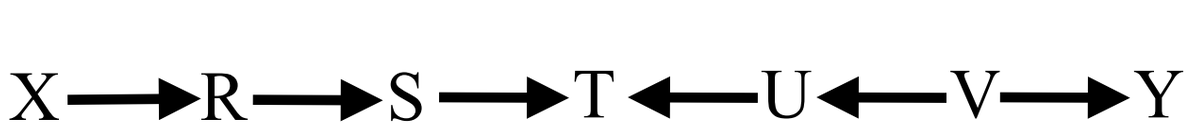




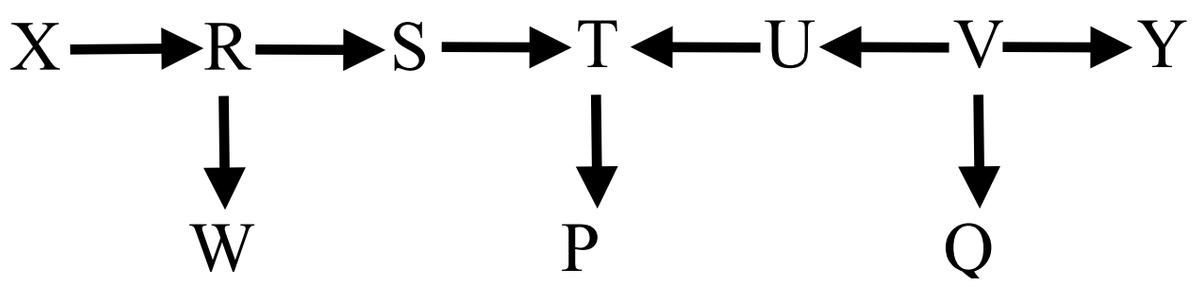
D-Separation

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

- A path p is blocked by a set of nodes **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 **Z**, **or**
  - p contains a collider $i \rightarrow m \leftarrow j$ such that the middle node m is in not **Z** and no descendant of m is in **Z**



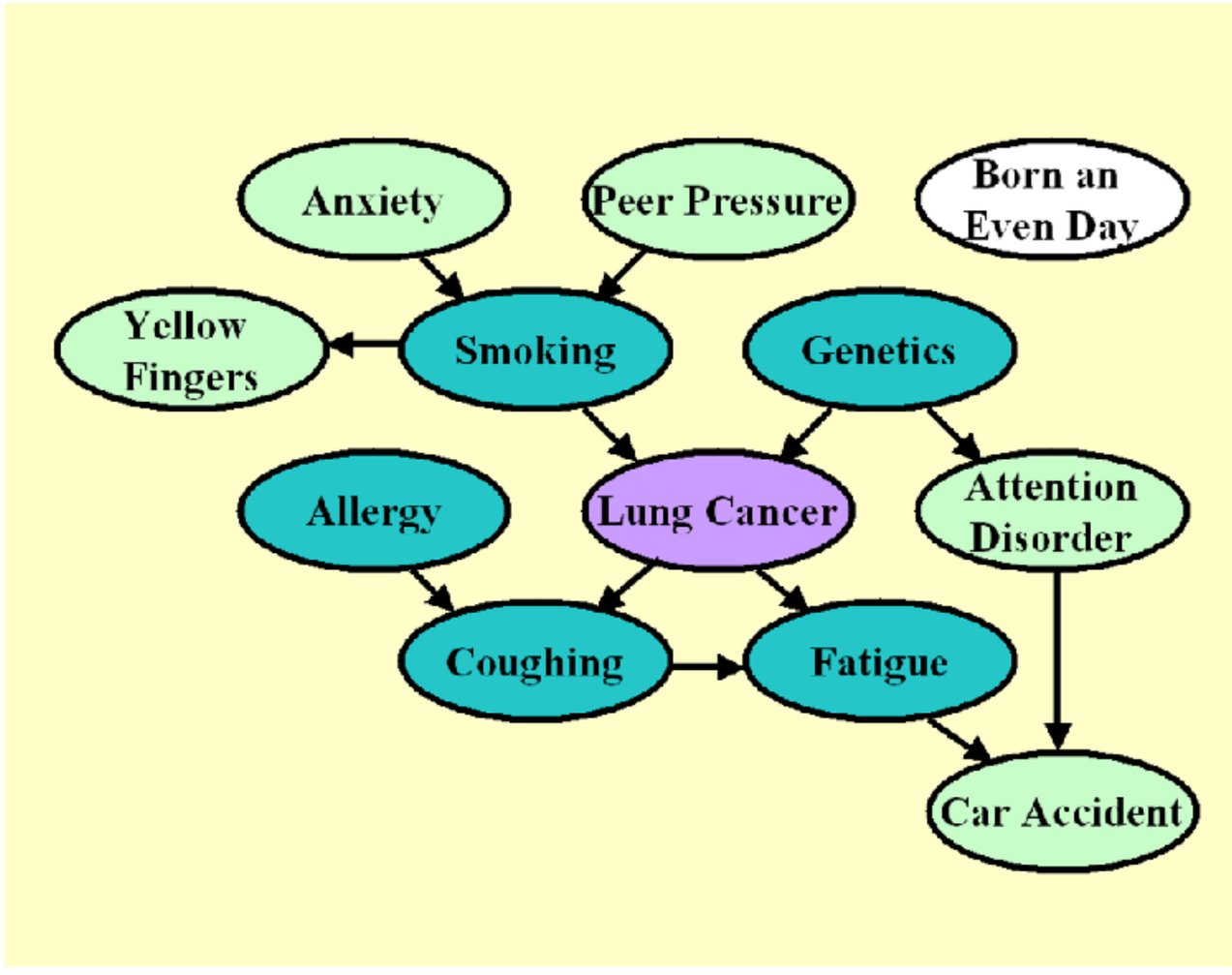
X and Y d-separated by {R,V}?
S and U d-separated by {R,V}?



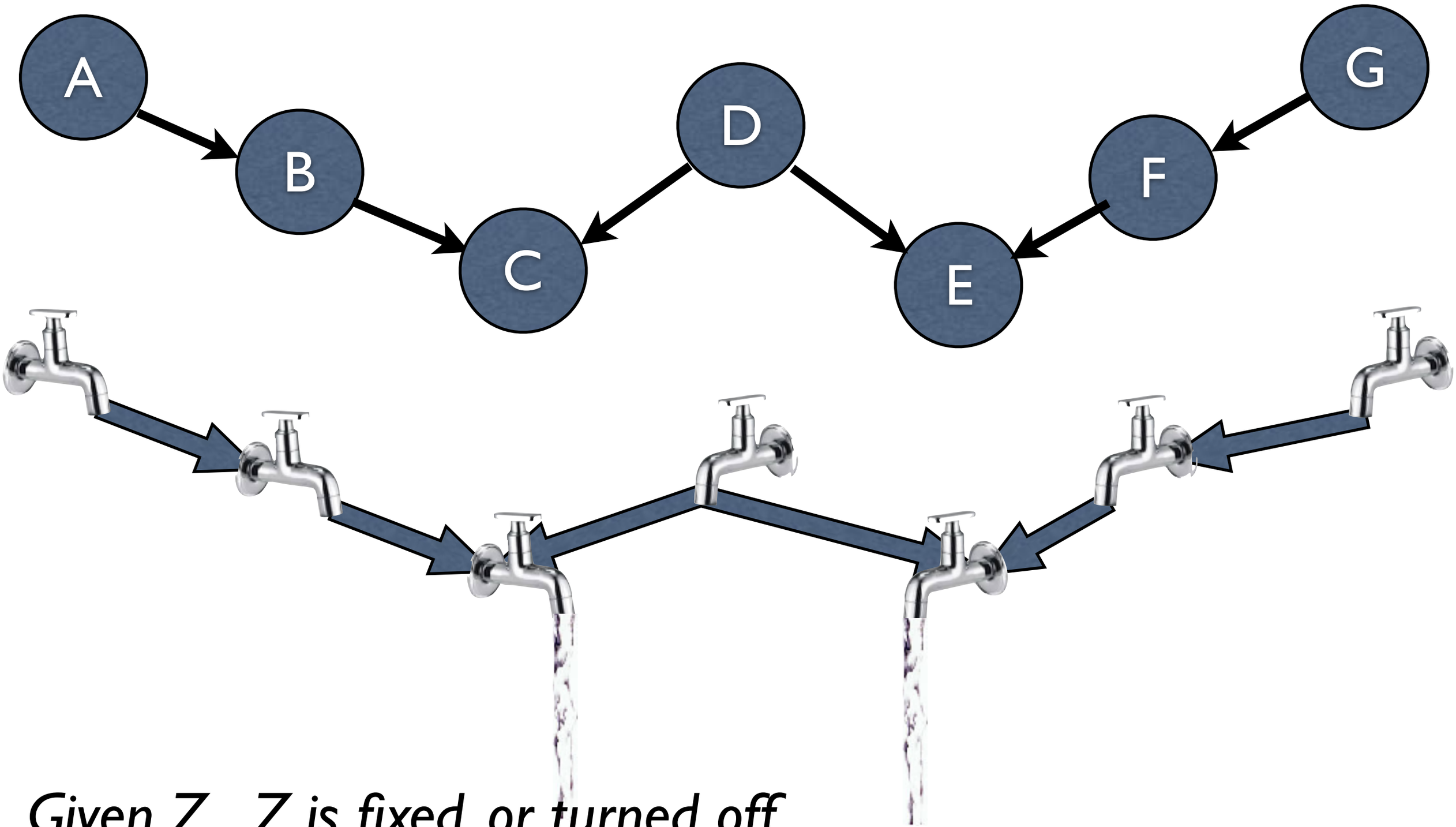
X and Y d-separated by {R, P}?

D-Separation: 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**)!



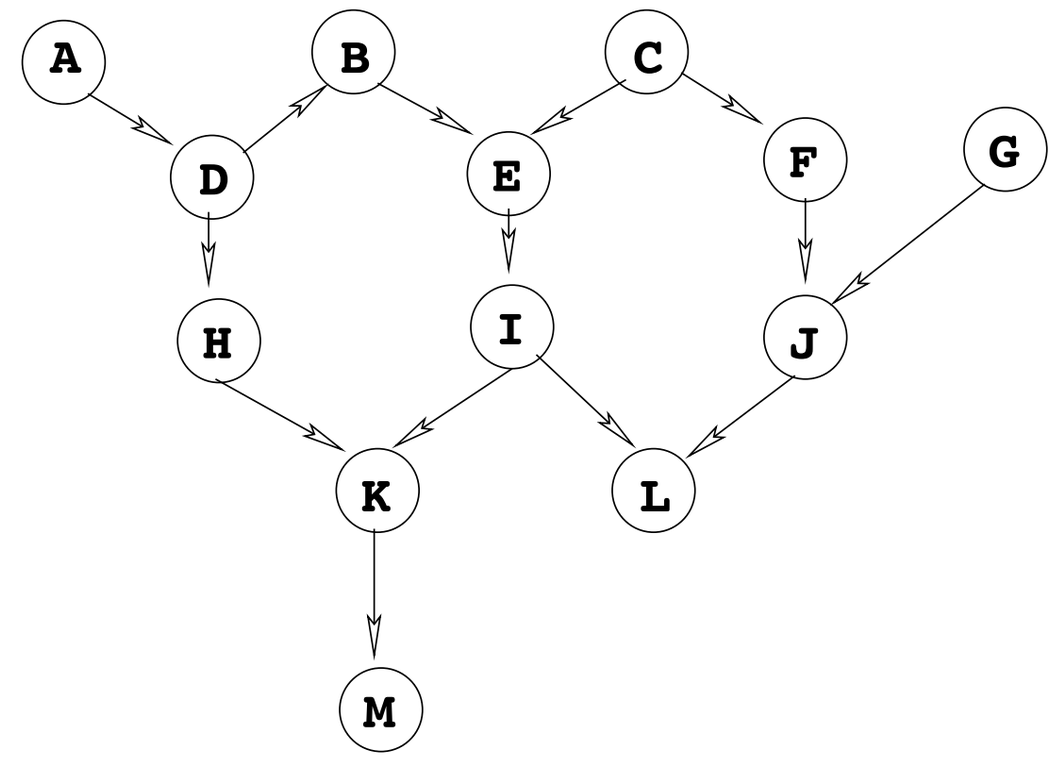
D-Separation: Intuition (2)



Given Z... Z is fixed, or turned off

Local & Global Markov Conditions

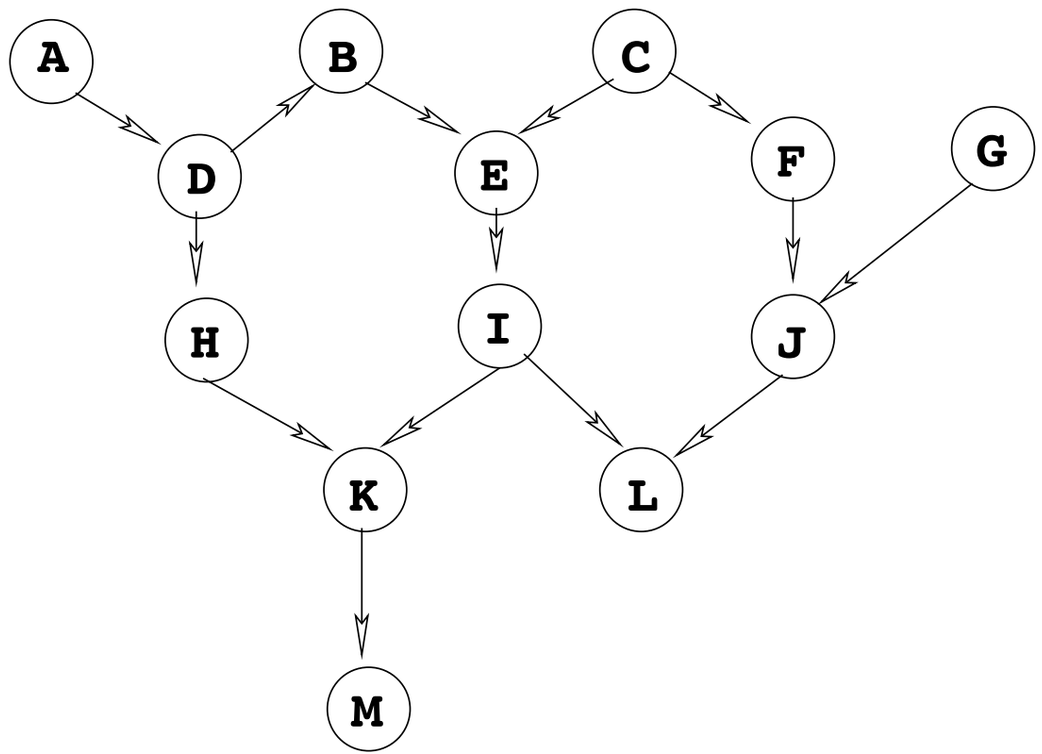
- **Local** Markov condition:
  - In a DAG, a variable X is independent of all its non-descendants given its parents



- **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!

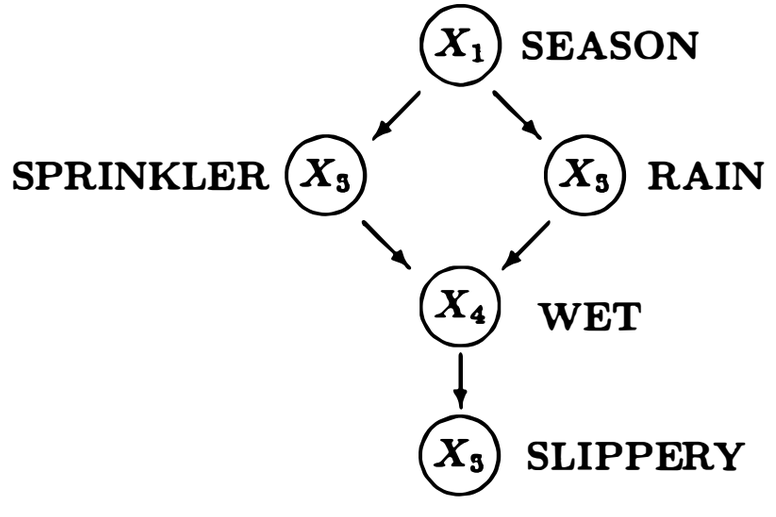
Markov Blanket

- 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 ?



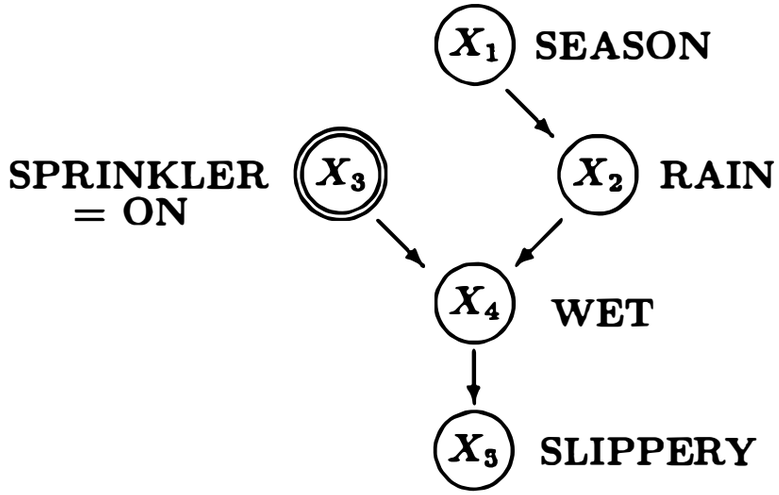
Causal Bayesian Networks (CBNs)

- 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. $P_x(V_i \mid PA_i) = P(V_i \mid PA_i)$ for all $V_i \notin$X, i.e., $P(V_i \mid PA_i)$ remains invariant to interventions not involving V_i.



What is $P_{X3=ON}(X_1,X_2,X_4,X_5)$?

Structural Causal Models

$$PA_i \longrightarrow X_i$$

- $X_i = f_i(PA_i, E_i)$, $i=1,\ldots,n$
  - E_i: exogenous variables / errors / disturbances
  - Each equation represents an *autonomous* mechanism
    - Describes how nature assigns values to variables of interest

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

- Distinction between structural equations & algebraic equations
- Associated with graphical causal models

CI from Data...

- We are able to see CI relationships from DAGs.
- How can we see that from data?
  - Useful to find information of the underlying DAG

Independence in Linear-Gaussian Case

- If X and Y are jointly normally distributed, their independence $\Leftrightarrow$ their zero correlation
- Zero correlation can be tested with, say, Fisher's z test

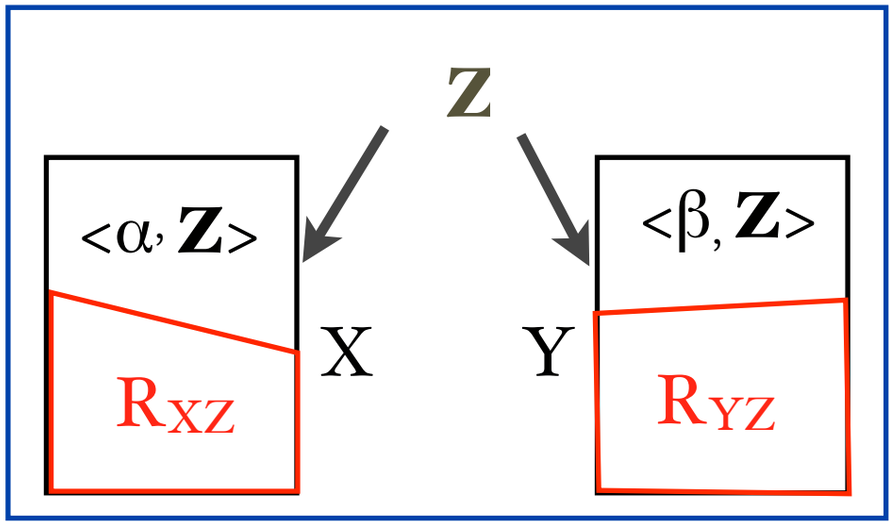
- Calculate sample correlation coefficient (statistic):

$$r = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}}.$$

- Under H_0 (zero correlation), $z \triangleq \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$ follows $\mathcal{N}(0, \frac{1}{N-3})$
- Given the statistic and its null distribution, we can find p value

Conditional Independence in Linear-Gaussian case: Partial Correlation

- Partial correlation: "Relationship" between X and Y while eliminating influence of $\mathbf{Z}$
  - Regress X and Y on $\mathbf{Z}$, respectively
    - Partial correlation $\rho_{XY \cdot \mathbf{Z}}$ is the correlation between residuals R_{XZ} and R_{YZ}



- If X, Y, and $\mathbf{Z}$ are jointly normally distribution, $X \perp\!\!\!\perp Y \mid \mathbf{Z} \Leftrightarrow \rho_{XY \cdot \mathbf{Z}} = 0$
- We can then test for zero partial correlation ('partialcorr' in MATLAB)

Another Causality Story

gs. One day, the scientist put the
e frog jumped four feet.

og with four feet, jumps four
the frogs legs. The scientist told
So the scientist wrote in his note
et."

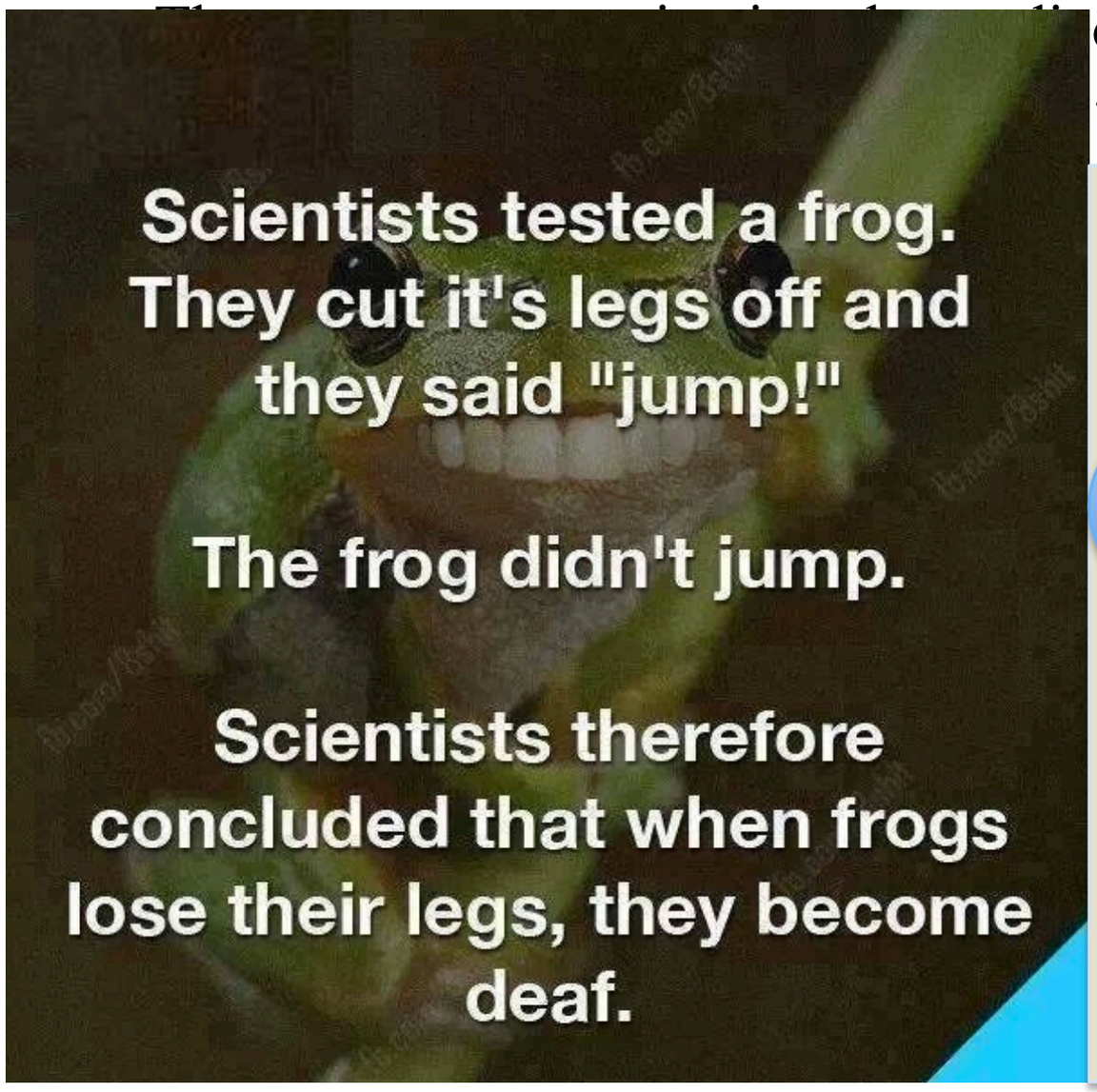
the frog to jump. The frog
his notebook "Frog with two feet,

the frog to jump. Frog jumped
book, "Frog with one foot, jumps

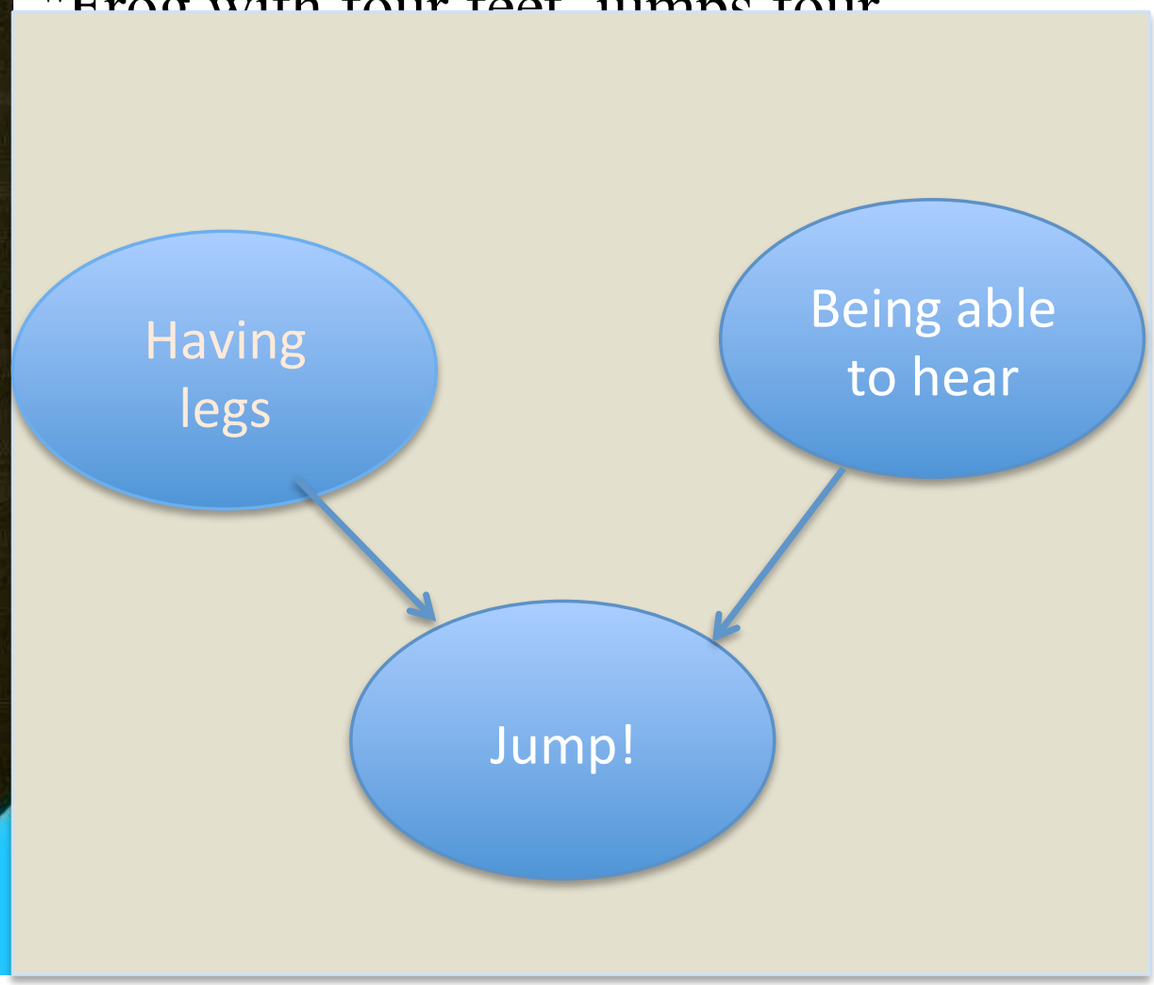
d, "Frog jump. Frog jump.

og with no feet, goes deaf."

Another Causality Story



d frogs. One day, the scientist put the
. The frog jumped four feet.
"Frog with four feet, jumps four



Large-scale...