# Bayesian Networks – Representation

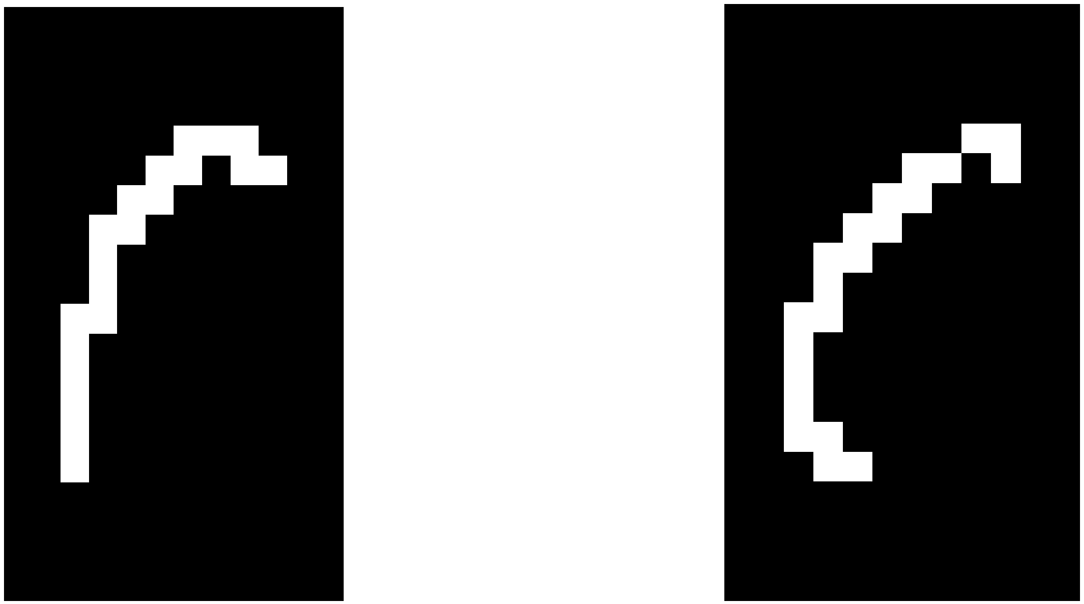
Machine Learning – 10701/15781
Carlos Guestrin
Carnegie Mellon University

March 20th, 2006

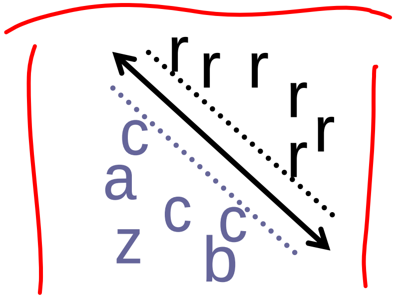
# Announcements

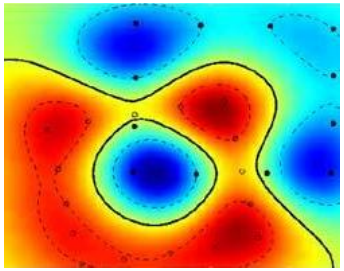
- Welcome back!
- One page project proposal due Wednesday
- We’ll go over midterm in this week’s recitation

# Handwriting recognition



Character recognition, e.g., kernel SVMs



# Webpage classification


→ Company home page

vs

Personal home page

vs

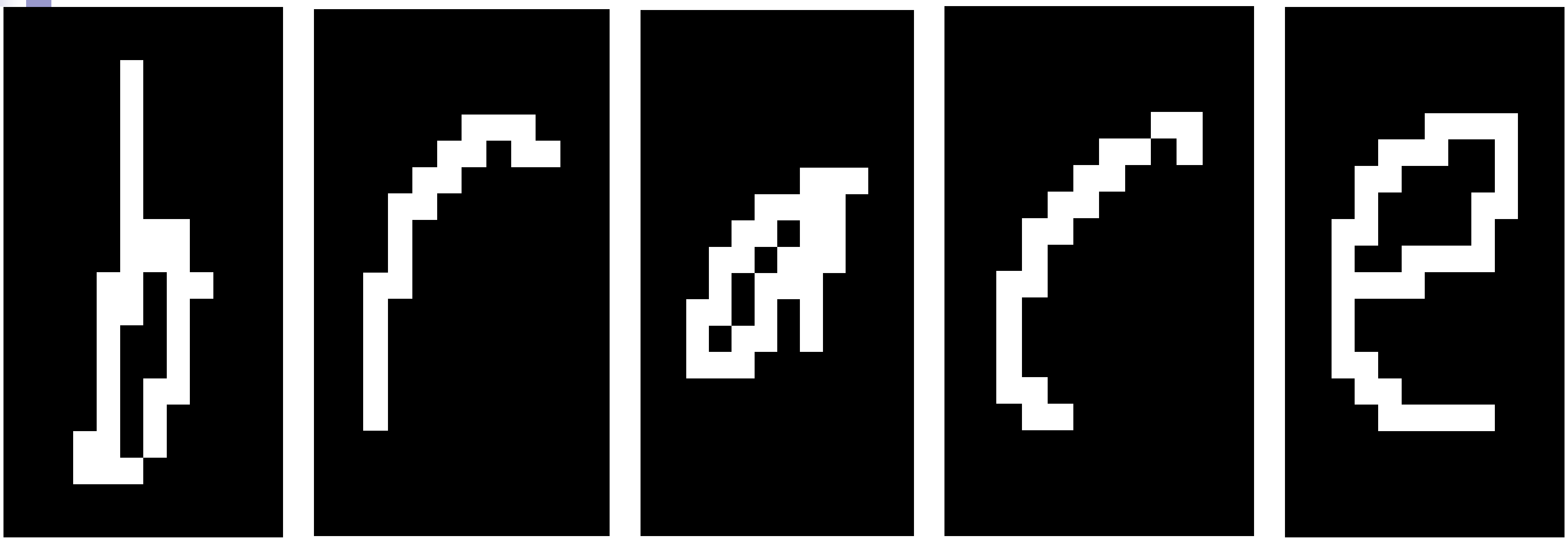
cannot spell

Univeristy home page

vs

…

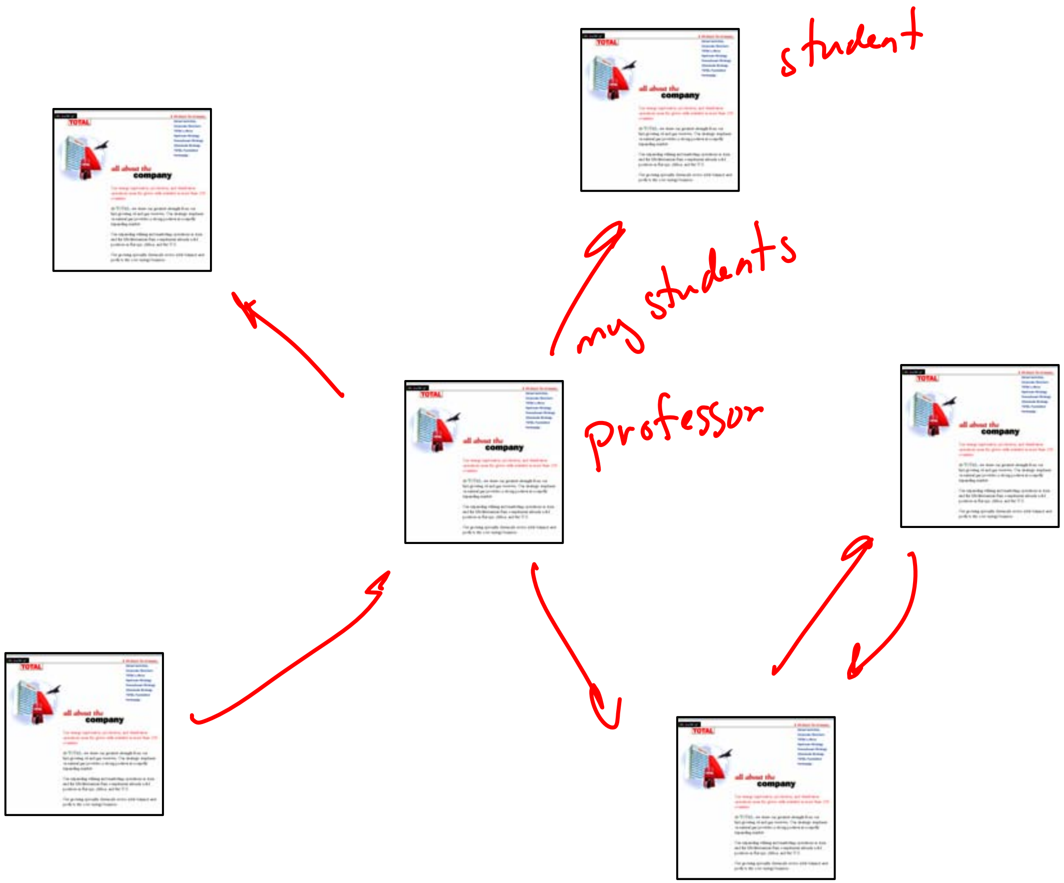
# Handwritting recognition 2



- Context
- examples not i.i.d.

+ correlations between labels!!

# Webpage classification 2



# Today – Bayesian networks

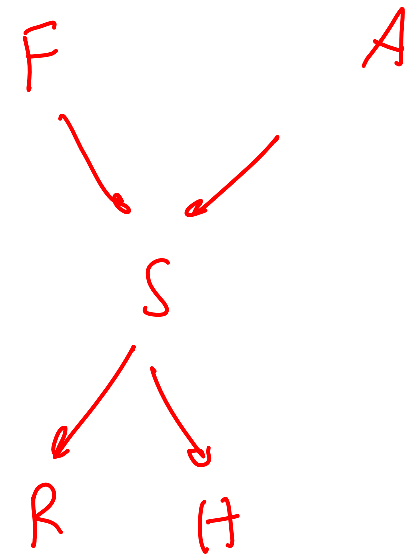
- One of the most exciting advancements in statistical AI in the last 10-15 years
- Generalizes naïve Bayes and logistic regression classifiers

  $p(y|x)$

  $p(y,x)$
- Compact representation for exponentially-large probability distributions
- Exploit conditional independencies

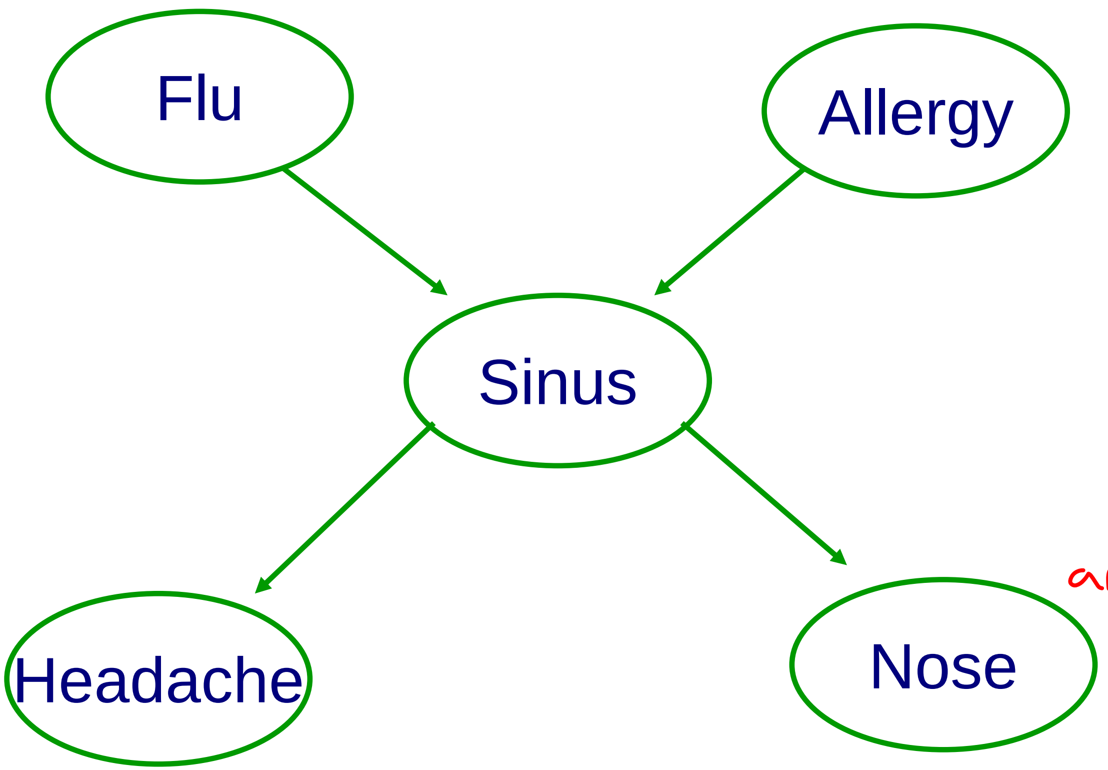
# Causal structure

- Suppose we know the following:
  - The flu causes sinus inflammation
  - Allergies cause sinus inflammation
  - Sinus inflammation causes a runny nose
  - Sinus inflammation causes headaches
- How are these connected?




$$P(F=t \mid H=t, R=f)$$

# Possible queries




- Inference

  $P(F=t \mid N=t)$

- Most probable explanation

  $H=t, N=t$

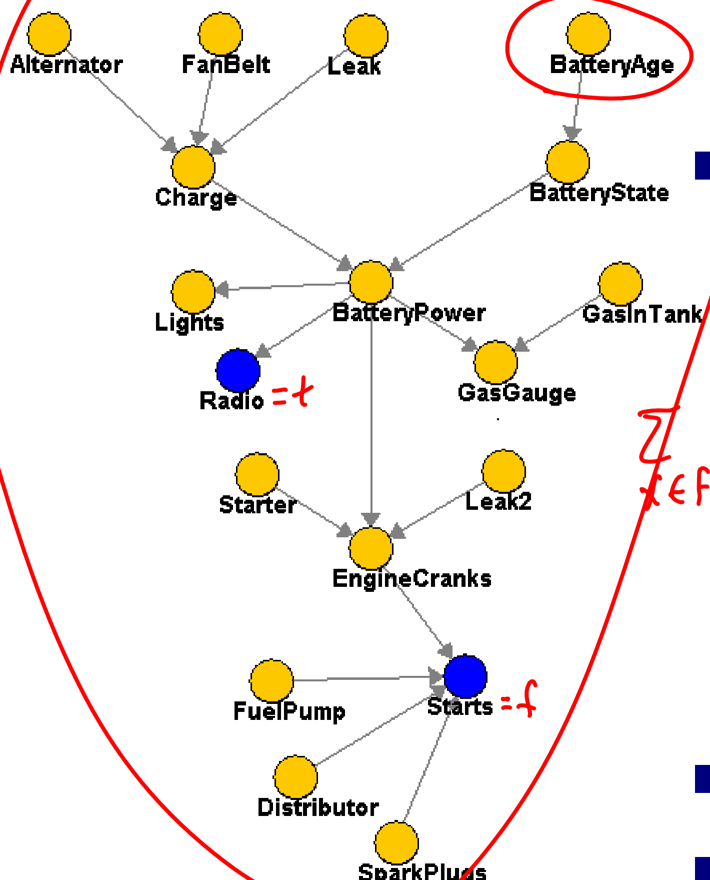
  $\underset{f,a,s}{\arg\max}\; P(f,a,s \mid H=t, N=t)$

- Active data collection

  $H=t$

  → Running Nose?

# Car starts BN



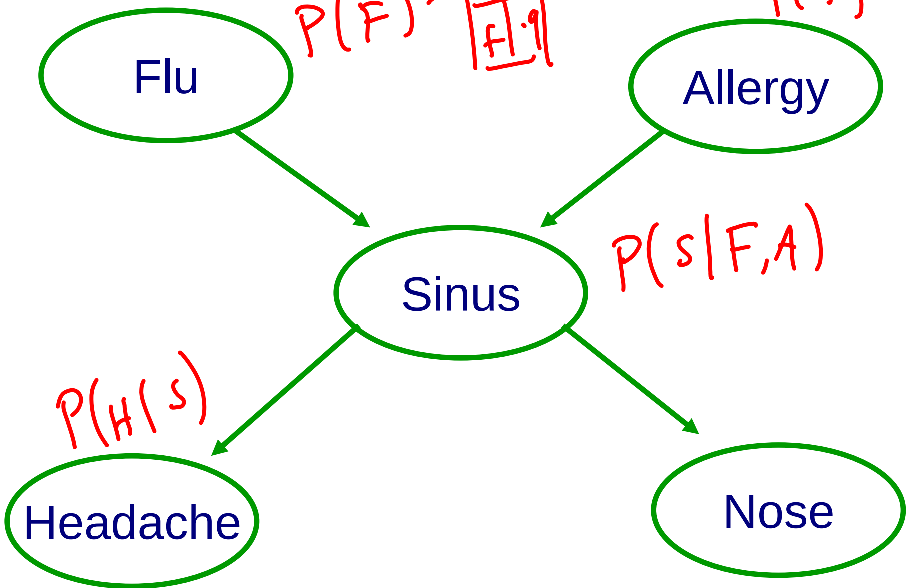

- 18 binary attributes
- Inference
  - P(BatteryAge|Starts=f) $\propto P(BA, S=f)$

Marginalization

$$\sum_{x \in F} P(BA, F=x, S=f) = P(BA, S=t)$$

- $2^{18}$ terms, why so fast?
- Not impressed?
  - HailFinder BN – more than $3^{54}$ = 58149737003040059690390169 terms

# Factored joint distribution - Preview




$P(F) =$

| | |
|---|---|
| t | .1 |
| f | .9 |

$P(A)$

$P(S|F,A)$

$P(H|S)$

$P(N|S)$

Indep: $(F \perp A)$

$(F \perp N \mid S)$

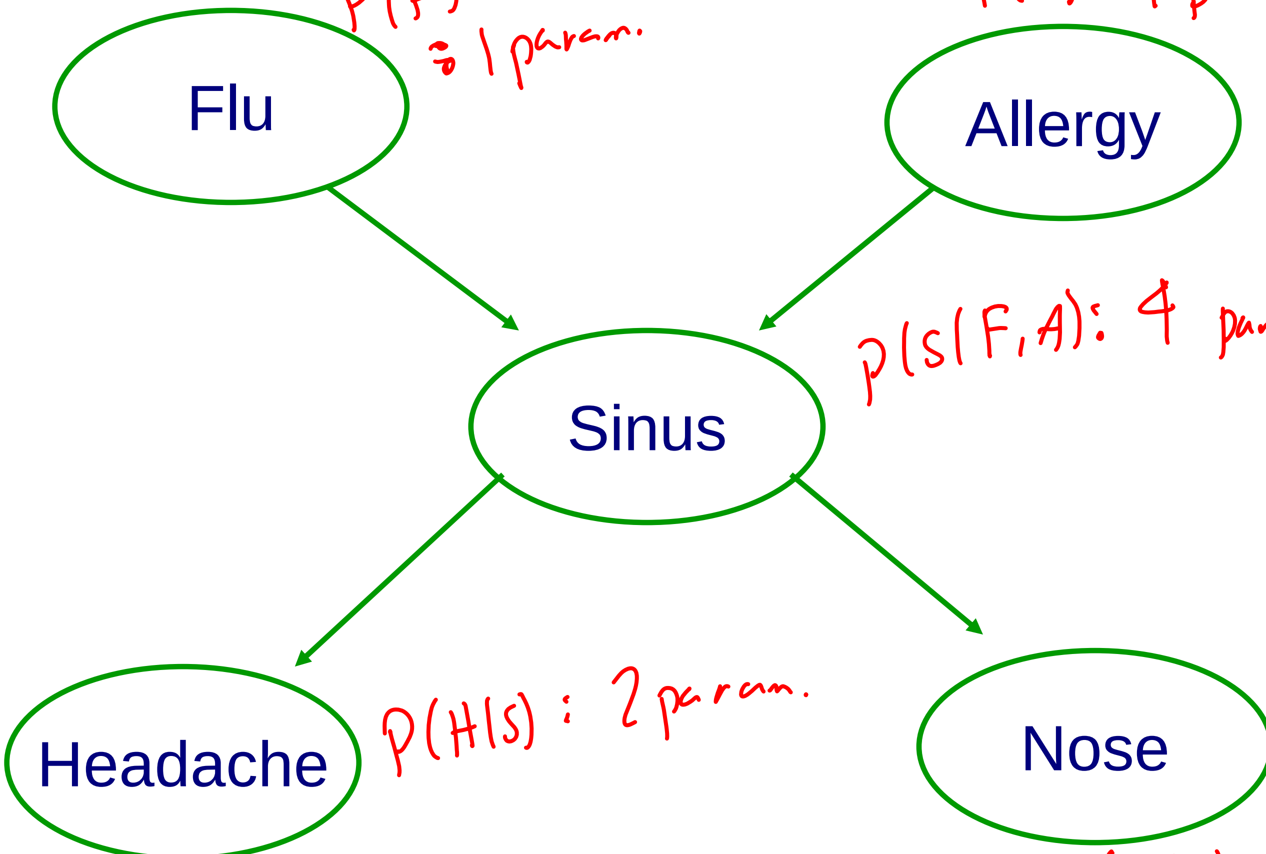
$$P(F,A,S,H,N)$$
$$= P(F) \cdot P(A) \cdot P(S|FA) \cdot P(H|S) \cdot P(N|S)$$

$P(N|S) =$

| N \ S | t | f |
|---|---|---|
| t | .8 | .3 |
| f | 1-.8 = .2 | 1-.3 = .7 |

2 parameters

# Number of parameters




$P(F)$ : 1 param.

$P(A)$ : 1 param.

$P(S|F,A)$ : 4 param.

$P(H|S)$ : 2 param.

$P(N|S)$ : 2 param.

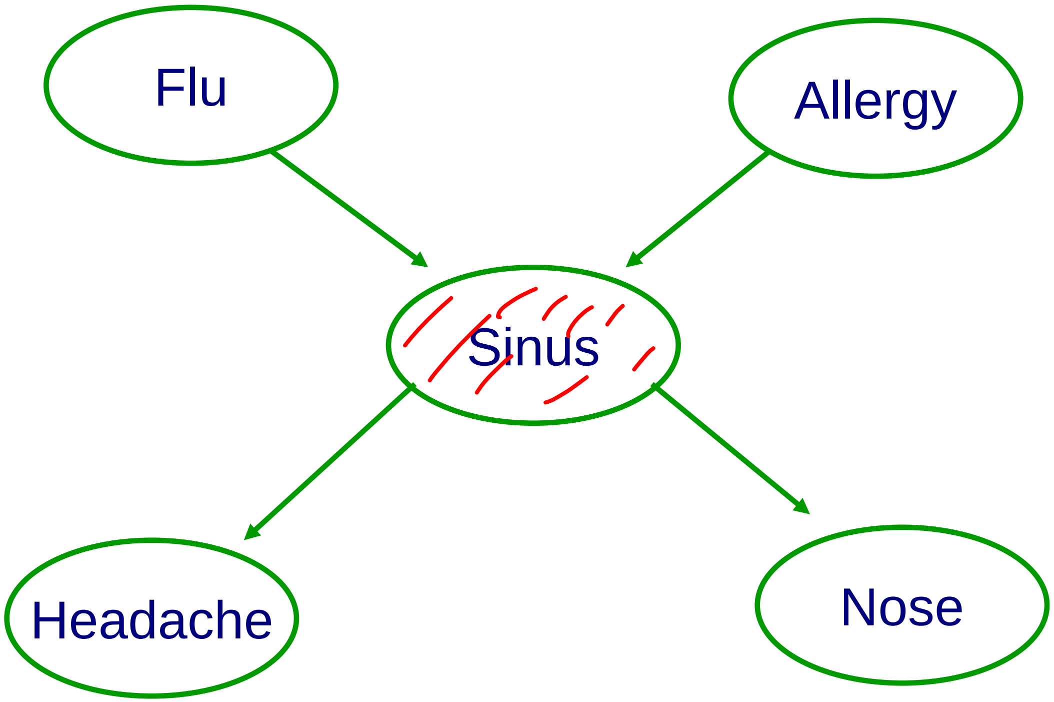
$P(A,F,S,H,N)$

→ naively : 31 param.

$P(A) \cdot P(F) \cdot P(S|F,A) \cdot P(H|S) \cdot P(N|S)$

10 param.

# Key: Independence assumptions




$(F \perp A)$

$(F \perp N \mid S)$

$(F \perp H \mid S)$

$(A \perp N \mid S)$

$(A \perp H \mid S)$

$(H \perp N \mid S)$

Knowing sinus separates the variables from each other

# (Marginal) Independence

- Flu and Allergy are (marginally) independent

$P(F,A) = P(F) \cdot P(A)$

| | |
|---|---|
| Flu = t | .1 |
| Flu = f | .9 |

- More Generally:

$P(A|F) = P(A)$

| | |
|---|---|
| Allergy = t | .2 |
| Allergy = f | .8 |

$P(F,A) =$

| | Flu = t | Flu = f |
|---|---|---|
| Allergy = t | | $.2 \times .9$ $= .18$ |
| Allergy = f | | |

# Marginally independent random variables

$\perp \leftarrow$ indep.

- **Sets** of variables **X**, **Y**
- X is independent of Y if
  - $P \vDash (\mathbf{X}=\mathbf{x} | \mathbf{Y}=\mathbf{y})$, $\forall$ $\mathbf{x} \in Val(\mathbf{X})$, $\mathbf{y} \in Val(\mathbf{Y})$

entails

- Shorthand:
  - **Marginal independence:** $P \vDash (\mathbf{X} \perp \mathbf{Y})$
- **Proposition:** *P* statisfies $(\mathbf{X} \perp \mathbf{Y})$ if and only if
  - $P(\mathbf{X},\mathbf{Y}) = P(\mathbf{X})\, P(\mathbf{Y})$

# Conditional independence

- Flu and Headache are not (marginally) independent

$P(F|H) \neq P(F)$ , e.g., $P(F=t|H=t) \neq P(F=t)$

- Flu and Headache are independent given Sinus infection

$P(F|S,H) = P(F|S)$ e.g., $P(F=t|S=t, H=t) = P(F=t|S=t)$

- More Generally:

$P(F|S,H) = P(F|S)$

or

$P(F,H|S) = P(F|S) \cdot P(H|S)$

# Conditionally independent random variables

- **Sets** of variables **X**, **Y**, **Z**
- X is independent of Y given Z if
  - $P \models (\mathbf{X}=\mathbf{x} \perp \mathbf{Y}=\mathbf{y} \mid \mathbf{Z}=\mathbf{z})$, $\forall\ \mathbf{x} \in Val(\mathbf{X})$, $\mathbf{y} \in Val(\mathbf{Y})$, $\mathbf{z} \in Val(\mathbf{Z})$

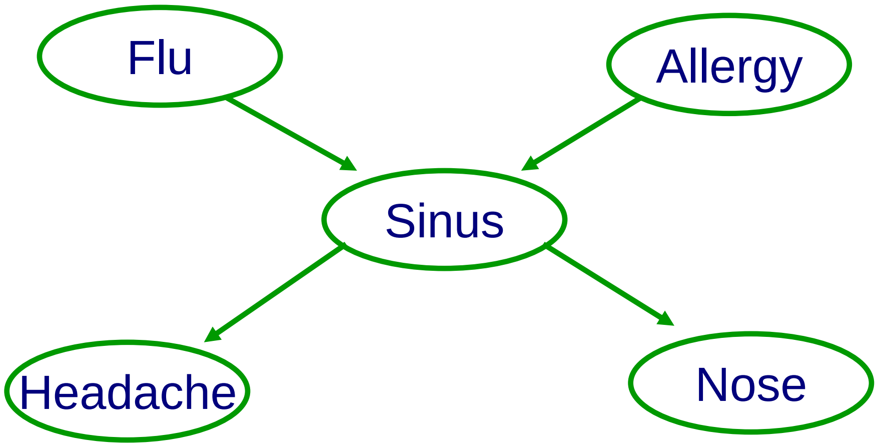
- Shorthand:
  - **Conditional independence:** $P \models (\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})$
  - For $P \models (\mathbf{X} \perp \mathbf{Y} \mid \emptyset)$, write $P \models (\mathbf{X} \perp \mathbf{Y})$

- **Proposition:** $P$ statisfies $(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})$ if and only if
  - $P(\mathbf{X},\mathbf{Y} \mid \mathbf{Z}) = P(\mathbf{X} \mid \mathbf{Z})\, P(\mathbf{Y} \mid \mathbf{Z})$

# Properties of independence

- **Symmetry:**
  - $(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}) \Rightarrow (\mathbf{Y} \perp \mathbf{X} \mid \mathbf{Z})$
- **Decomposition:**
  - $(\mathbf{X} \perp \mathbf{Y},\mathbf{W} \mid \mathbf{Z}) \Rightarrow (\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})$
- **Weak union:**
  - $(\mathbf{X} \perp \mathbf{Y},\mathbf{W} \mid \mathbf{Z}) \Rightarrow (\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z},\mathbf{W})$
- **Contraction:**
  - $(\mathbf{X} \perp \mathbf{W} \mid \mathbf{Y},\mathbf{Z})$ & $(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}) \Rightarrow (\mathbf{X} \perp \mathbf{Y},\mathbf{W} \mid \mathbf{Z})$
- **Intersection:**
  - $(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{W},\mathbf{Z})$ & $(\mathbf{X} \perp \mathbf{W} \mid \mathbf{Y},\mathbf{Z}) \Rightarrow (\mathbf{X} \perp \mathbf{Y},\mathbf{W} \mid \mathbf{Z})$
  - Only for positive distributions!
  - $P(\alpha)>0, \forall\alpha, \alpha\neq\emptyset$

# **The** independence assumption



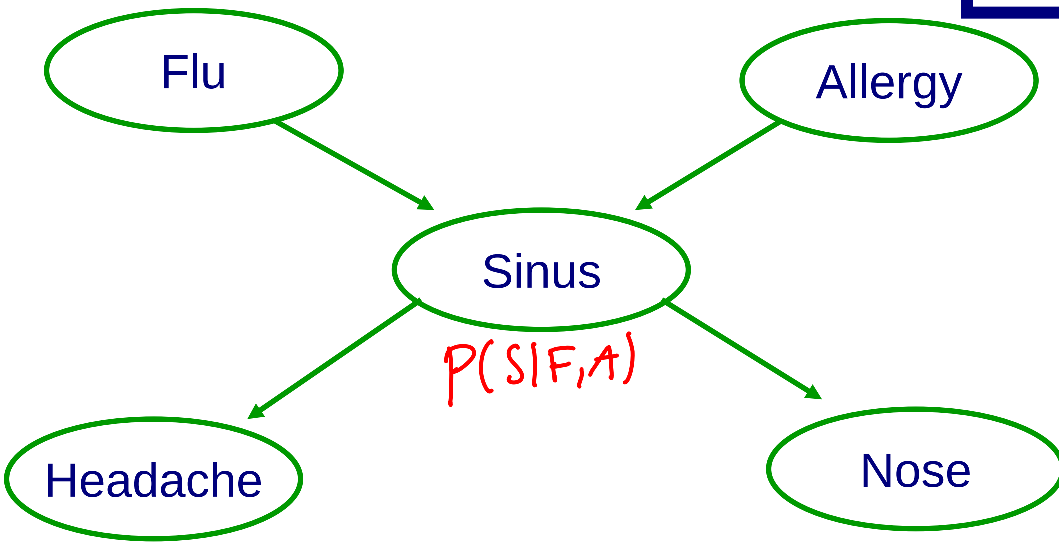

**Local Markov Assumption:**
A variable X is independent of its non-descendants given its parents

$(F \perp A)$

$(N \perp \{F, A, H\} \mid S)$

# Explaining away

**Local Markov Assumption:**
A variable X is independent of its non-descendants given its parents




$(F \perp A)$ marginally

what if $S=t$

$S=t$ $\quad P(F=t \mid S=t) > P(F=t)$

but

$S=t$ & $A=t$ :

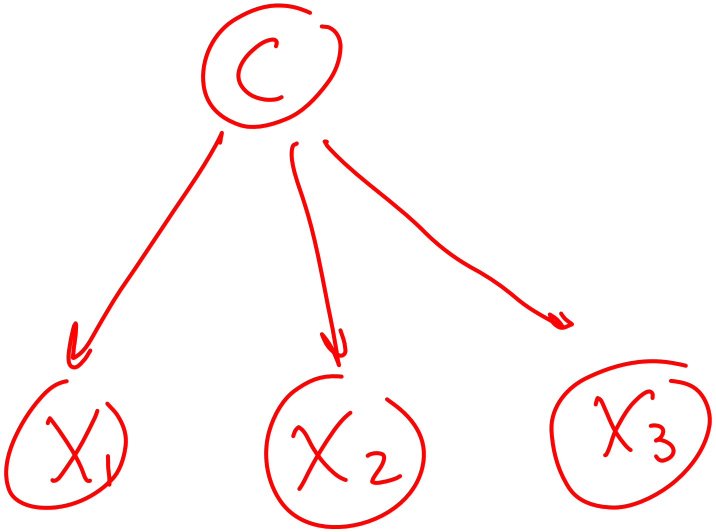
$$P(F=t \mid S=t) > P(F=t \mid S=t, A=t)$$
$$> P(F=t)$$

F not indep. A given S

what if $N=t$

Same explaining away!!

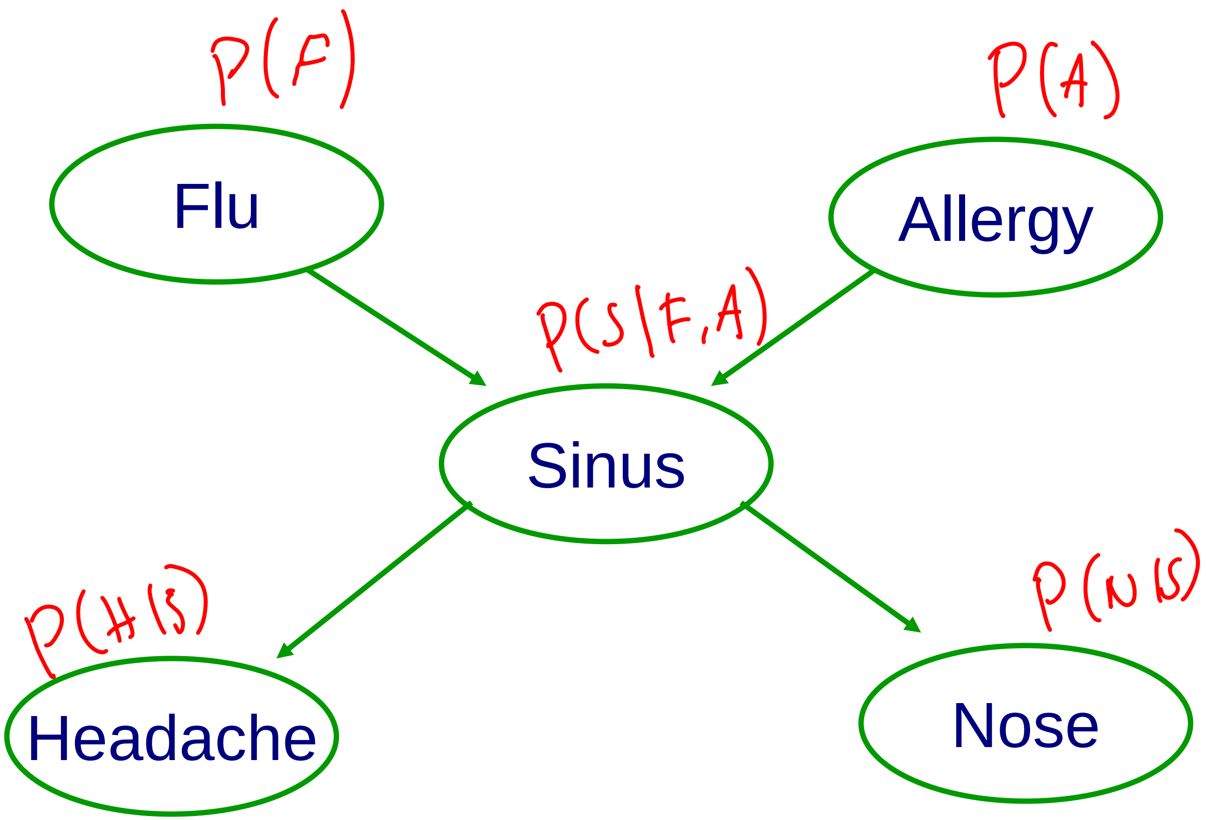
# Naïve Bayes revisited



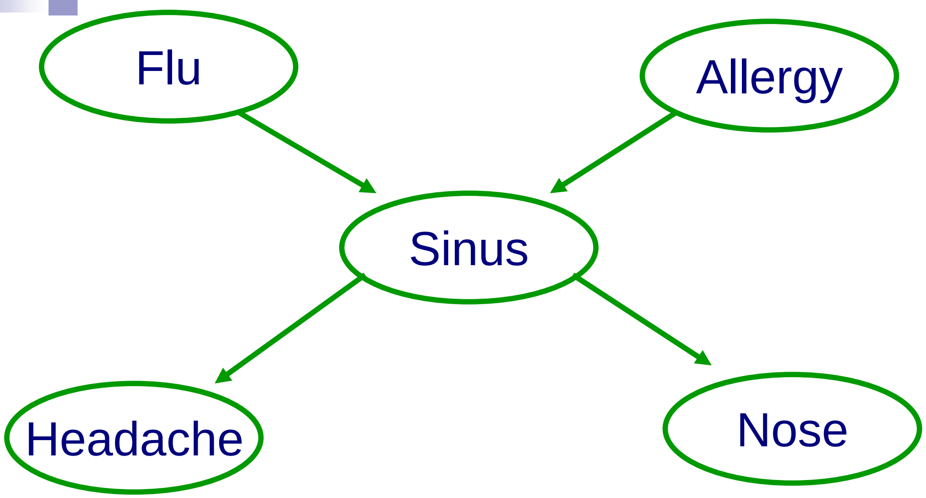

**Local Markov Assumption:**
A variable X is independent of its non-descendants given its parents

$(X_i \perp X_j \mid C)$

# What about probabilities? Conditional probability tables (CPTs)



# Joint distribution




$P(F, A, S, H, N)$
$= P(F) \cdot P(A) \cdot P(S|F,A) \cdot$
$P(H|S) \cdot P(N|S)$

More edges $\rightarrow$ fewer indep. assumptions!

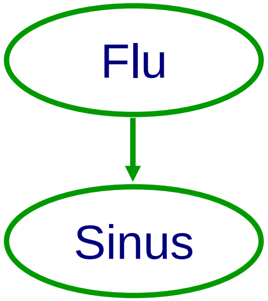
local

**Why can we decompose? Markov Assumption!**

# The chain rule of probabilities

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

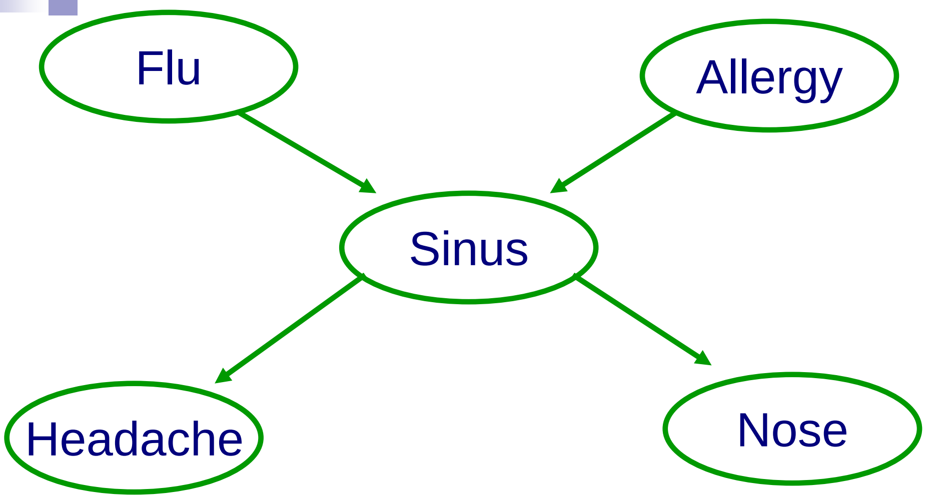
$P(F,S) = P(F) \cdot P(S|F)$




- More generally:
  - $P(X_1,\ldots,X_n) = P(X_1) \cdot P(X_2|X_1) \cdot \ldots \cdot P(X_n|X_1,\ldots,X_{n-1})$

$P(X_3|X_2,X_1) \cdot$

# Chain rule & Joint distribution




**Local Markov Assumption:**
A variable X is independent of its non-descendants given its parents

$(F \perp A) \Rightarrow P(A|F) = P(A)$

$(H \perp \{F,A\} | S) \Rightarrow P(H|SFA) = P(H|S)$

$(N \perp \{H,F,A\} | S) \Rightarrow P(N|FAHS) = P(N|S)$

chain rule
no assumptions

$P(F,A,S,H,N) =$

$P(F).\underbrace{P(A|F)}_{=P(A)} \cdot P(S|FA)\,\underbrace{P(H|SFA)}_{=P(H|S)}\,\underbrace{P(N|FASH)}_{=P(N|S)}$

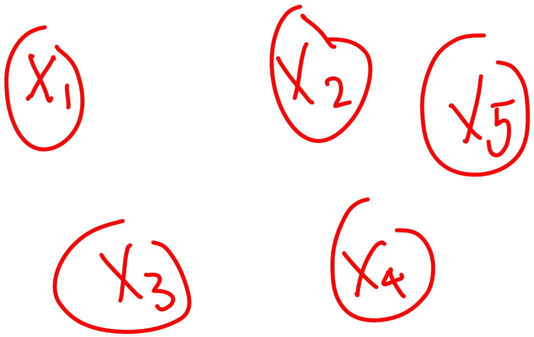
with local Markov Assumption:

$\Rightarrow P(F)\,P(A)\,P(S|FA)\;P(H|S)\;P(N|S)$

# Two (trivial) special cases

give you some P

**Edgeless graph**




only if all vars indep.

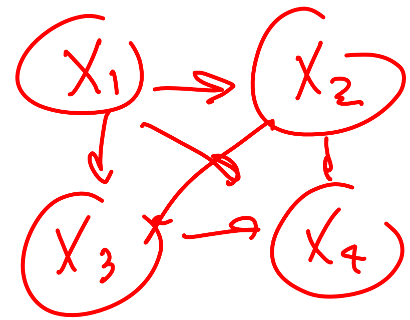
$(X_1 \perp X_4)$

$(X_2 \perp X_3 \mid X_5)$

$\vdots$

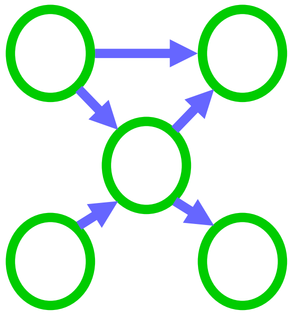
always!

**Fully-connected graph**




no indep. in graph

# The Representation Theorem – Joint Distribution to BN

**BN:** 

**Encodes independence assumptions**

**If conditional independencies in BN are subset of conditional independencies in *P***

Obtain

**Joint probability distribution:**

$$P(X_1,\ldots,X_n) = \prod_{i=1}^{n} P\left(X_i \mid \mathbf{Pa}_{X_i}\right)$$

P can be represented with BN

$\forall P$ exists at least one BN

# Real Bayesian networks applications

it's all about exploiting indep. (problem structure)

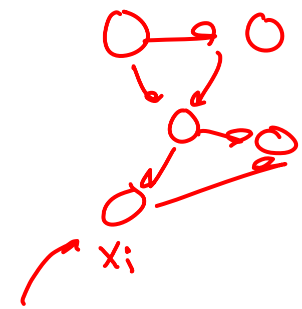
- Diagnosis of lymph node disease
- Speech recognition
- Microsoft office and Windows
  - http://www.research.microsoft.com/research/dtg/
- Study Human genome
- Robot mapping
- Robots to identify meteorites to study
- Modeling fMRI data
- Anomaly detection
- Fault dianosis
- Modeling sensor network data

# A general Bayes net

- Set of random variables

$X_1, X_2, X_3 \ldots, X_n$

- Directed acyclic graph
  - Encodes independence assumptions




- CPTs

$P(X_i \mid Pa_{X_i})$

- Joint distribution:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P\left(X_i \mid \mathbf{Pa}_{X_i}\right)$$

# How many parameters in a BN?

- Discrete variables $X_1, \ldots, X_n$
- Graph
  - Defines parents of $X_i$, $\mathbf{Pa}_{X_i}$
- CPTs – $P(X_i | \mathbf{Pa}_{X_i})$

one CPT $P(X_i \mid Pa_{X_i})$

$X_i$ takes on $|X_i|$ possible values

# param:

$$|Pa_{X_i}| \cdot (|X_i| - 1)$$

$$|Pa_{X_i}| = \prod_{X_j \in Pa_{X_i}} |X_j|$$

e.g. $\forall i: |X_i| = K$

no var has more than $d$ parents.

# parameters $< K^d (K-1) \cdot n$

~~[illegible]~~ fully connected:

$$K^n - 1$$

# Another example

- Variables:
  - B – Burglar
  - E – Earthquake
  - A – Burglar alarm
  - N – Neighbor calls
  - R – Radio report
- Both burglars and earthquakes can set off the alarm
- If the alarm sounds, a neighbor may call
- An earthquake may be announced on the radio

# Another example – Building the BN

- B – Burglar
- E – Earthquake
- A – Burglar alarm
- N – Neighbor calls
- R – Radio report

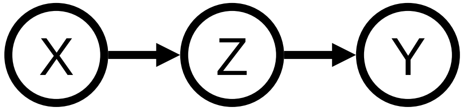
# Independencies encoded in BN

- We said: All you need is the local Markov assumption
  - $(X_i \perp \text{NonDescendants}_{Xi} \mid \mathbf{Pa}_{Xi})$
- But then we talked about other (in)dependencies
  - e.g., explaining away
- What are the independencies encoded by a BN?
  - Only assumption is local Markov
  - But many others can be derived using the algebra of conditional independencies!!!

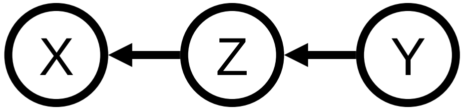
# Understanding independencies in BNs – BNs with 3 nodes

**Local Markov Assumption:**
A variable X is independent of its non-descendants given its parents

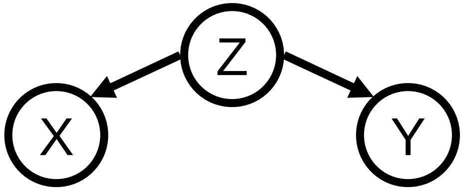
Indirect causal effect:



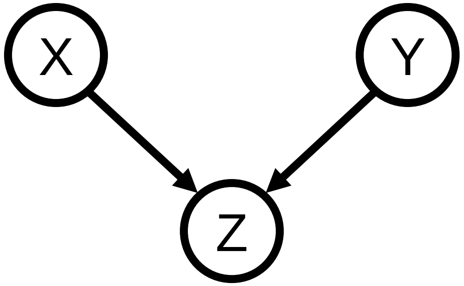

Indirect evidential effect:



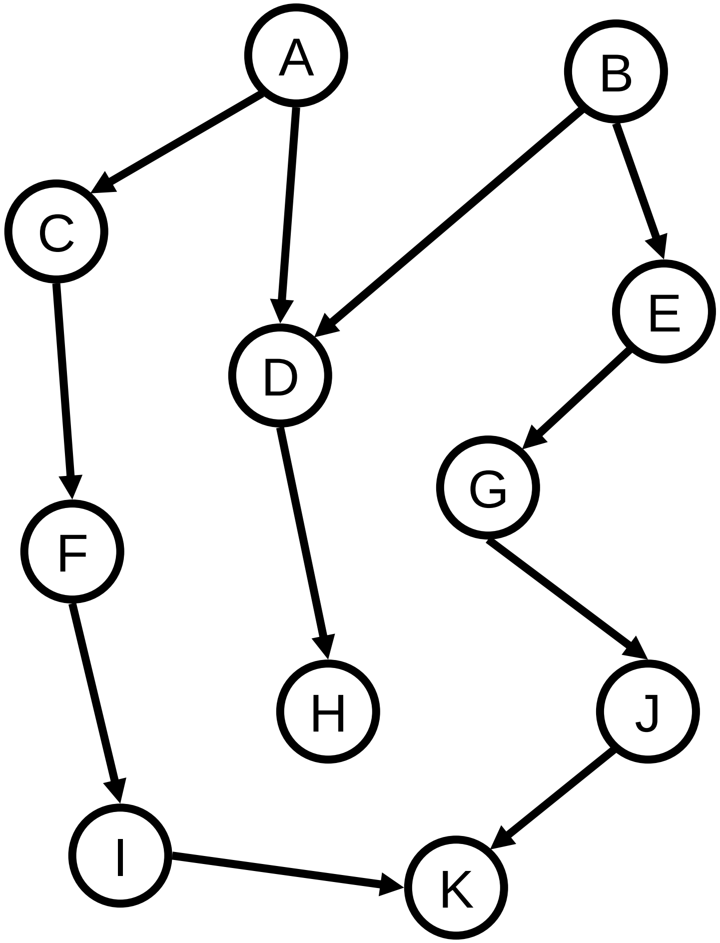

Common cause:



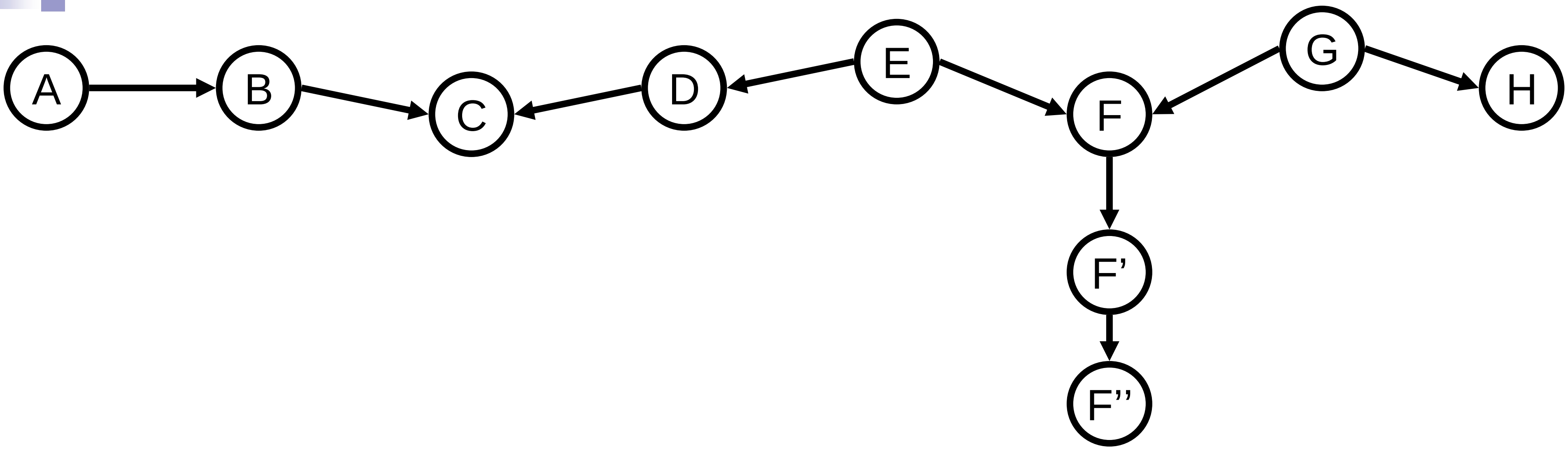

Common effect:



# Understanding independencies in BNs – Some examples



# An active trail – Example



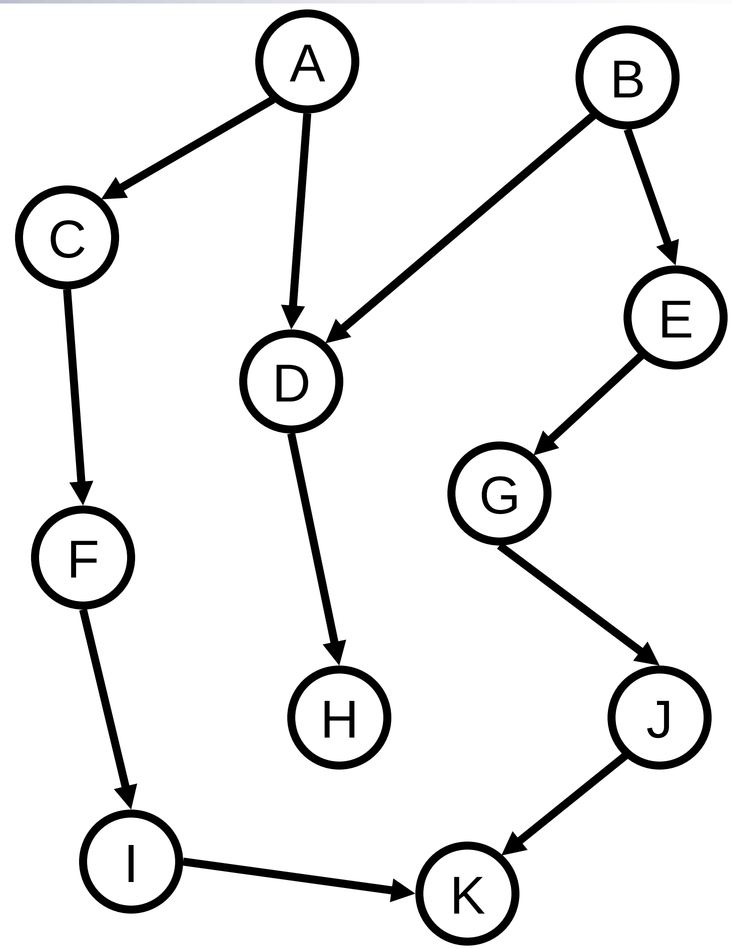

**When are A and H independent?**

# Active trails formalized

- A path $X_1 - X_2 - \cdots - X_k$ is an **active trail** when variables $\boldsymbol{O} \subseteq \{X_1,\ldots,X_n\}$ are observed if for each consecutive triplet in the trail:
  - $X_{i-1} \rightarrow X_i \rightarrow X_{i+1}$, and $X_i$ is **not observed** ($X_i \notin \boldsymbol{O}$)
  - $X_{i-1} \leftarrow X_i \leftarrow X_{i+1}$, and $X_i$ is **not observed** ($X_i \notin \boldsymbol{O}$)
  - $X_{i-1} \leftarrow X_i \rightarrow X_{i+1}$, and $X_i$ is **not observed** ($X_i \notin \boldsymbol{O}$)
  - $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$, and $X_i$ **is observed** ($X_i \in \boldsymbol{O}$), or **one of its descendents**

# Active trails and independence?

- **Theorem**: Variables $\mathbf{X_i}$ **and** $\mathbf{X_j}$ **are independent given** $\boldsymbol{Z}\subseteq\{X_1,\ldots,X_n\}$ if the is **no active trail** between $X_i$ and $X_j$ when variables $\boldsymbol{Z}\subseteq\{X_1,\ldots,X_n\}$ are observed



# The BN Representation Theorem

**If conditional independencies in BN are subset of conditional independencies in *P***

Obtain

**Joint probability distribution:**

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P\left(X_i \mid \mathbf{Pa}_{X_i}\right)$$

**Important because:**
**Every *P* has at least one BN structure *G***

**If joint probability distribution:**

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P\left(X_i \mid \mathbf{Pa}_{X_i}\right)$$

Obtain

**Then conditional independencies in BN are subset of conditional independencies in *P***

**Important because:**
**Read independencies of *P* from BN structure *G***

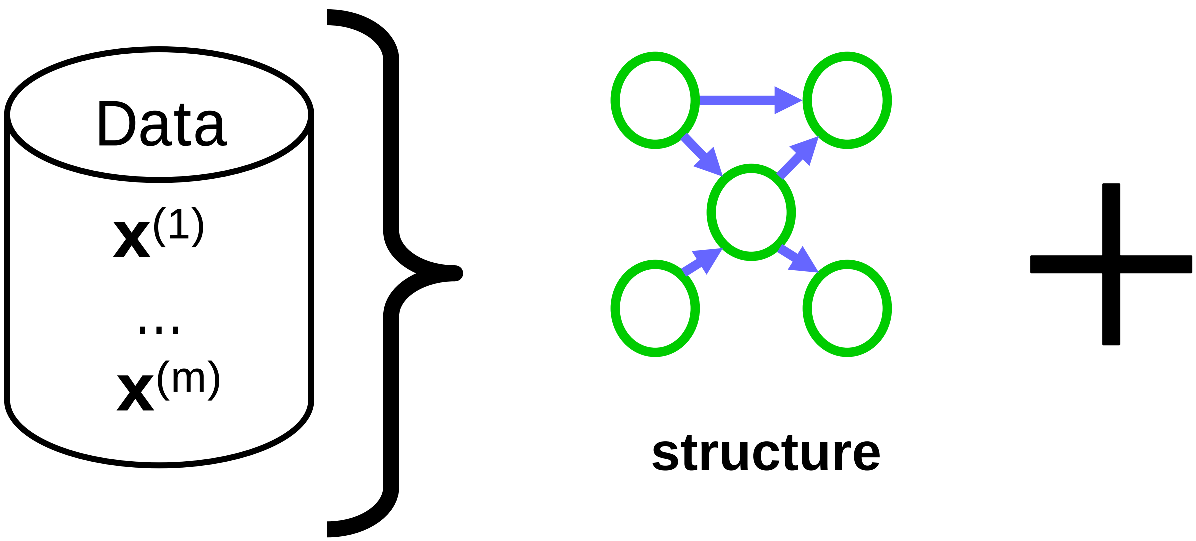
# “Simpler” BNs

- A distribution can be represented by many BNs:

- Simpler BN, requires fewer parameters

# Learning Bayes nets

| | Known structure | Unknown structure |
|---|---|---|
| Fully observable data | | |
| Missing data | | |

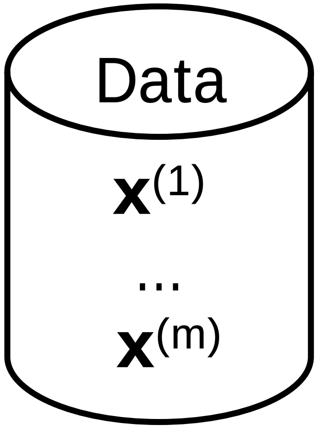


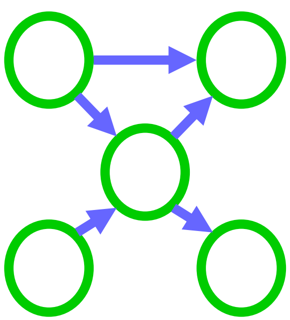

**structure**

CPTs – $P(X_i \mid \mathbf{Pa}_{X_i})$

**parameters**

# Learning the CPTs



For each discrete variable $X_i$

MLE: $P(X_i = x_i \mid X_j = x_j) = \dfrac{\text{Count}(X_i = x_i, X_j = x_j)}{\text{Count}(X_j = x_j)}$

# Queries in Bayes nets

- Given BN, find:
  - Probability of X given some evidence, $P(X|e)$
  - Most probable explanation, $\max_{x_1,\ldots,x_n} P(x_1,\ldots,x_n \mid e)$
  - Most informative query
- Learn more about these next class

# What you need to know

- Bayesian networks
  - A compact **representation** for large probability distributions
  - Not an algorithm
- Semantics of a BN
  - Conditional independence assumptions
- Representation
  - Variables
  - Graph
  - CPTs
- Why BNs are useful
- Learning CPTs from fully observable data
- Play with applet!!! ☺

# Acknowledgements

- JavaBayes applet
  - http://www.pmr.poli.usp.br/ltd/Software/javabayes/Home/index.html