Bayesian Networks – Representation

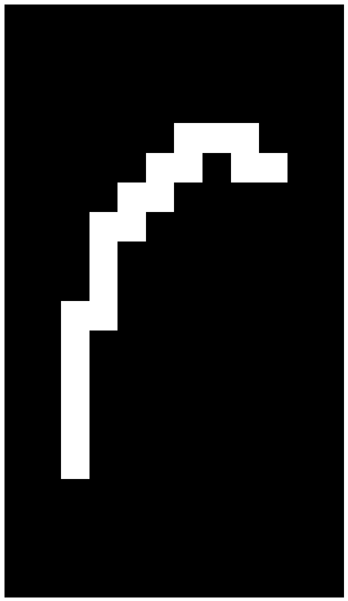
Machine Learning – 10701/15781

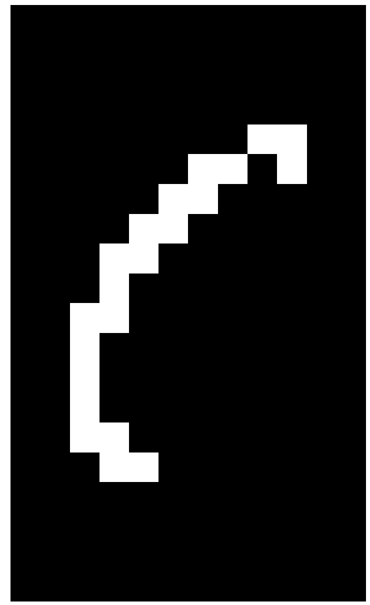
Carlos Guestrin

Carnegie Mellon University

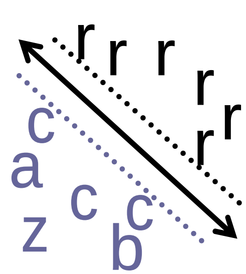
March 19th, 2007

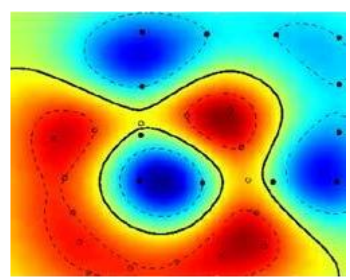
Handwriting recognition





Character recognition, e.g., kernel SVMs





Webpage classification

Company home page

vs

Personal home page

vs

University home page

vs

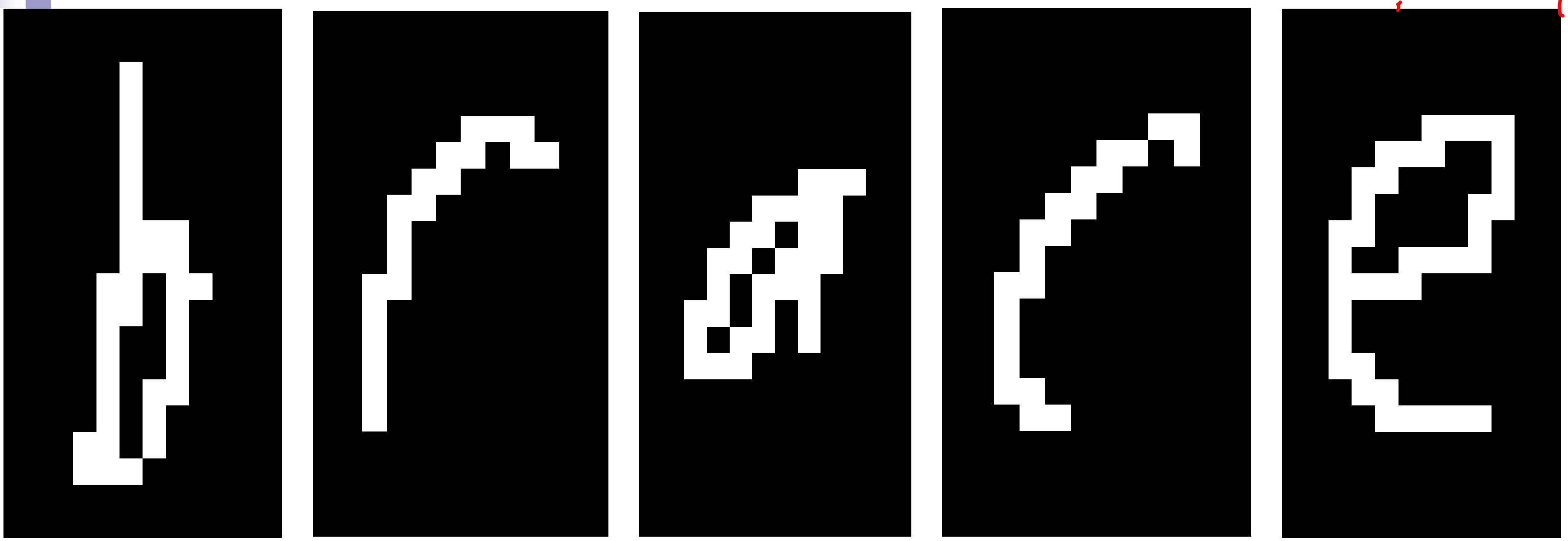
…

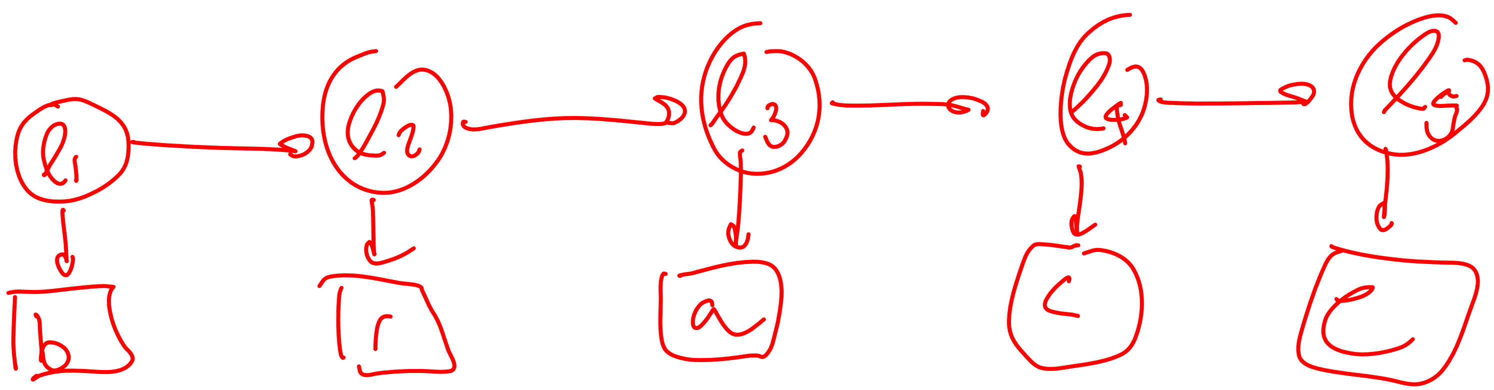
Handwriting recognition 2

brace

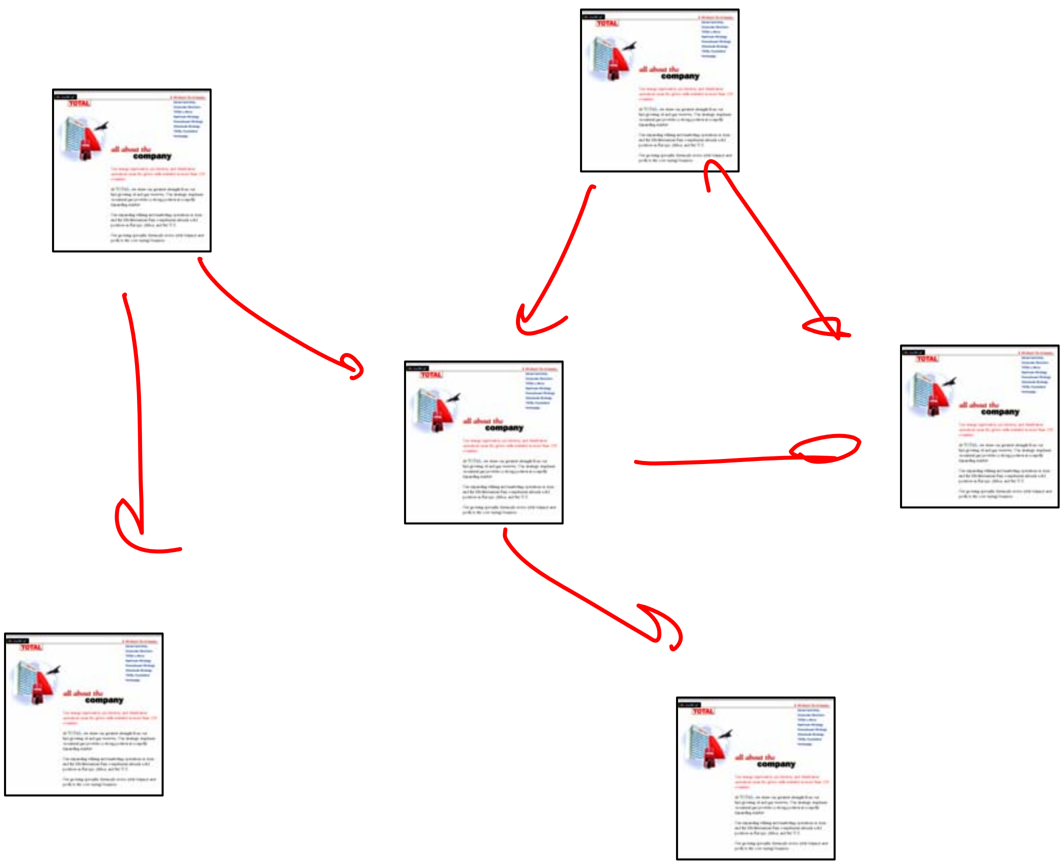
~~zzzzz~~ # classes $\doteq (26)^5$

brake





Webpage classification 2



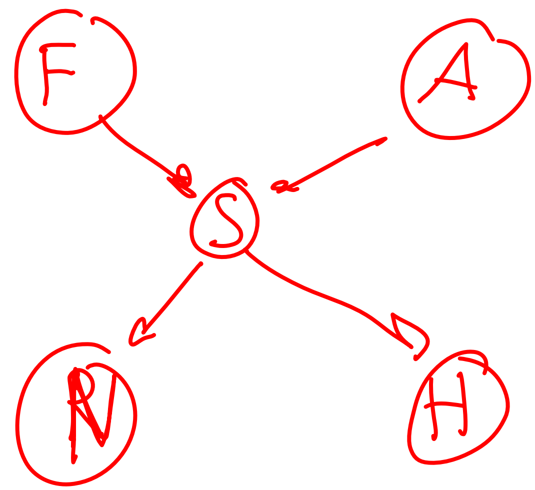
look at links between pages

Today – Bayesian networks

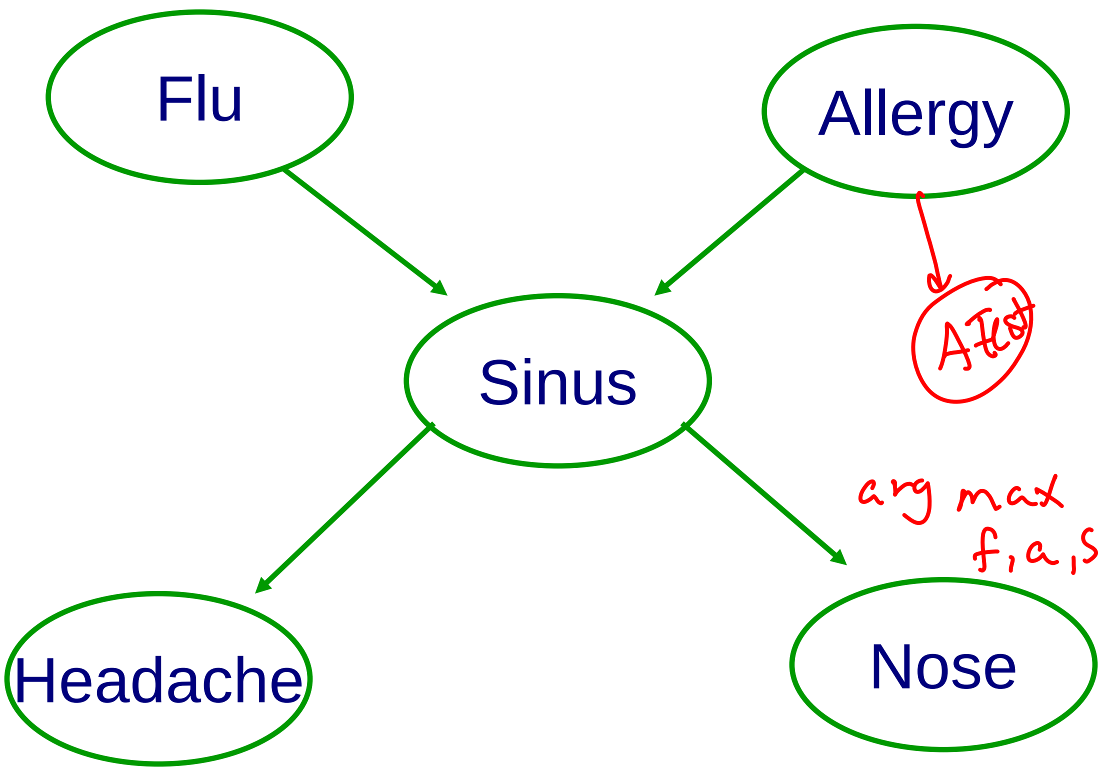
- One of the most exciting advancements in statistical AI in the last 10-15 years
- Generalizes naïve Bayes and logistic regression classifiers
- Compact representation for exponentially-large probability distributions
- Exploit conditional independencies

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?



Possible queries



Probabilistic

- Inference $P(X \mid e)$

  $e = H{=}t \wedge N{=}t$

  $P(A{=}t \mid H{=}t, N{=}t)$

- Most probable explanation

  $P(F{=}f, A{=}a, S{=}s \mid H{=}t, N{=}t)$

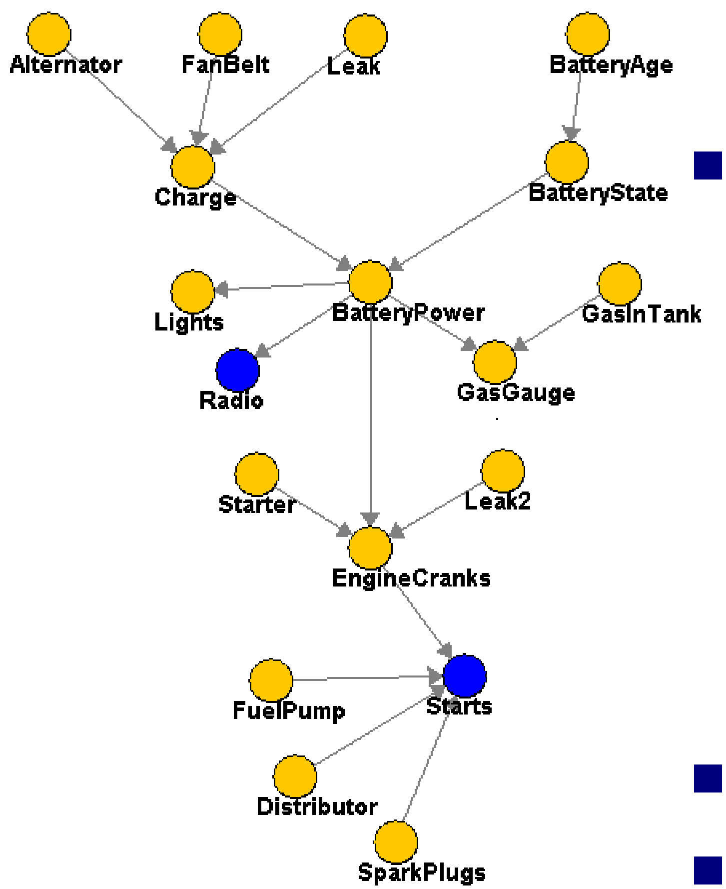
  e.g., $F{=}t$, $A{=}f$, $S{=}t$

- Active data collection

what test should I do?

go get Allergy test ...

Car starts BN



- 18 binary attributes
- Inference
  - P(BatteryAge|Starts=f) =old

go through all joint instantiations
add up probability
$\approx 2^{17}$ terms

- 2^{16} terms, why so fast?
- Not impressed?
  - HailFinder BN – more than 3^{54} = 58149737003040059690390169 terms

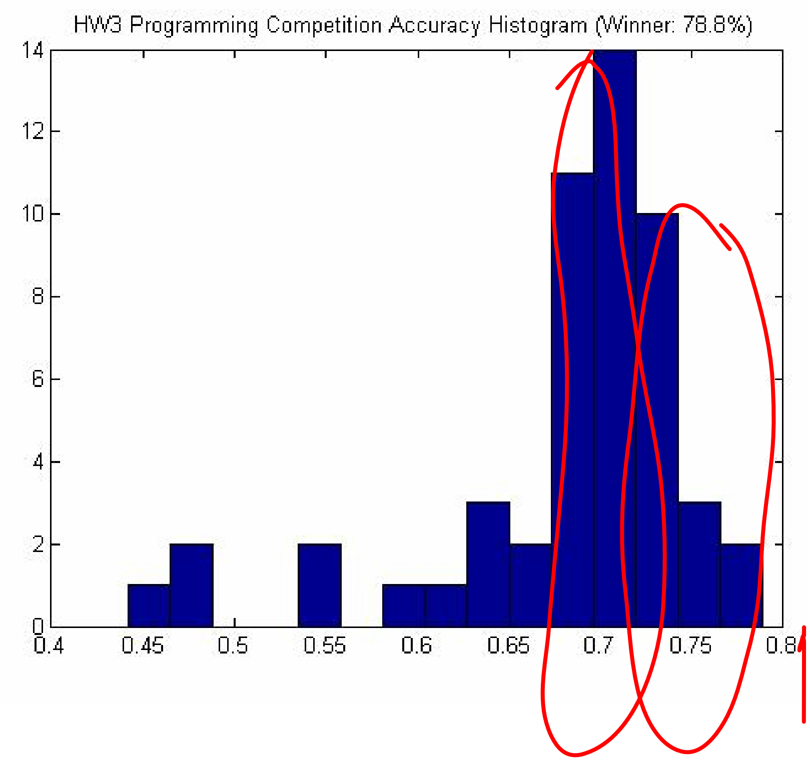
Announcements

- Welcome back!

- One page project proposal due Wednesday
  - Individual or groups of two
  - Must be something related to ML! ☺
  - It will be great if it's related to your research → it must be something you started this semester
- Midway progress report
  - 5 pages NIPS format
  - April 16th
  - Worth 20%
- Poster presentation
  - May 4, 2-5pm in the NSH Atrium
  - Worth 20%
- Final report
  - May 10th
  - 8 pages NIPS format
  - Worth 60%

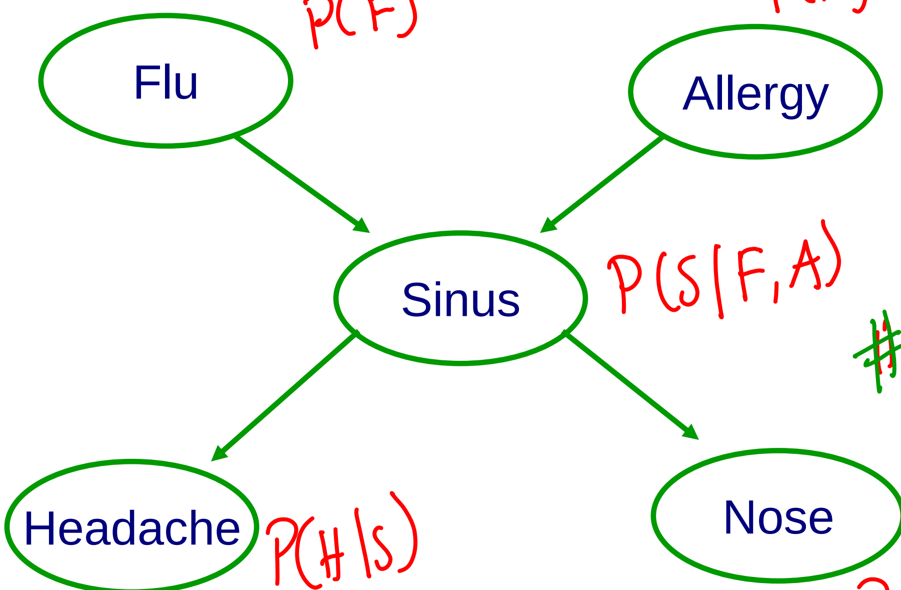
- It will be fun!!! ☺

And the winner is…



Factored joint distribution - Preview

params? $2^5 - 1$
$= 31$



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

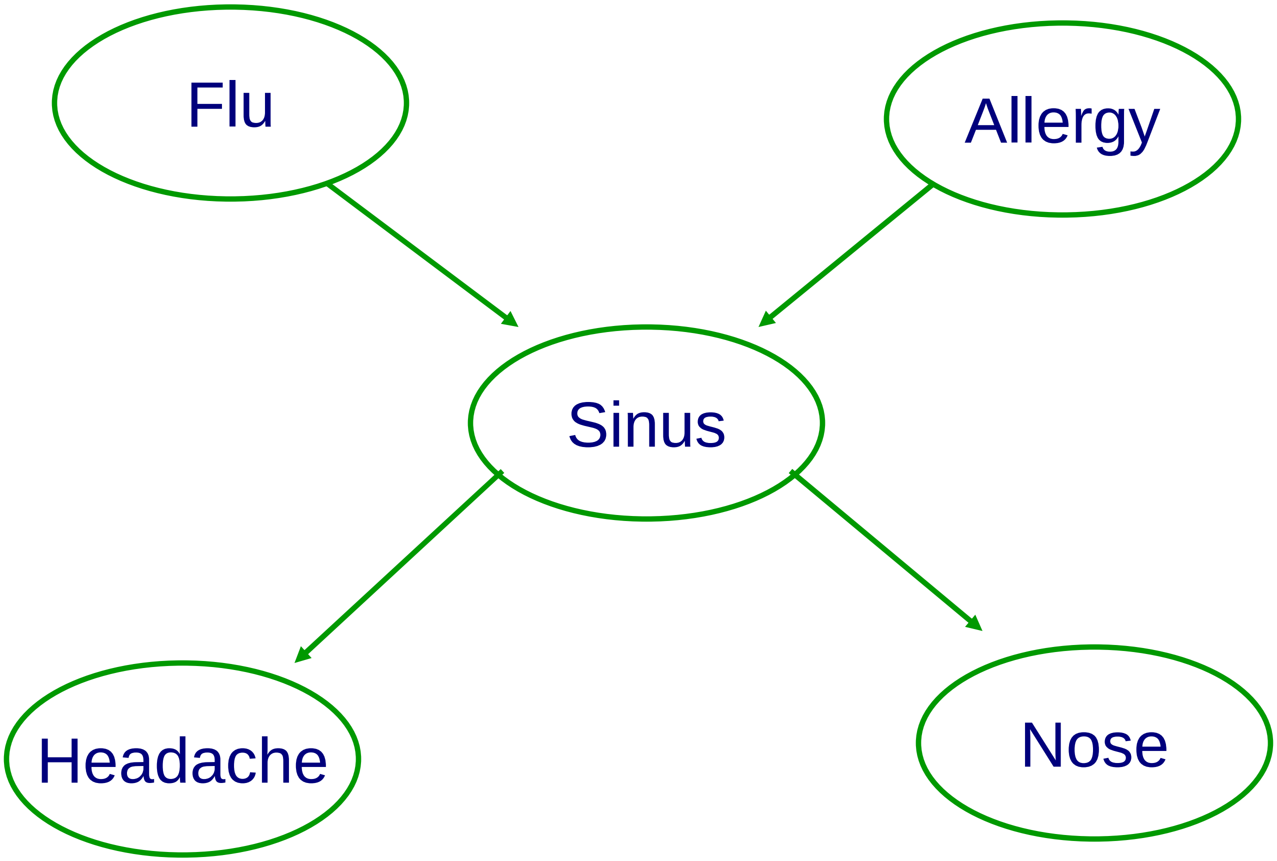
# params	1	1	4	2	2

$= 10$

$P(S|FA) =$

A \ F	t	f
t	P(S=t|Par)=1 P(S=f|Par)=0	.5 .5
f	.7 .3	.1 .9

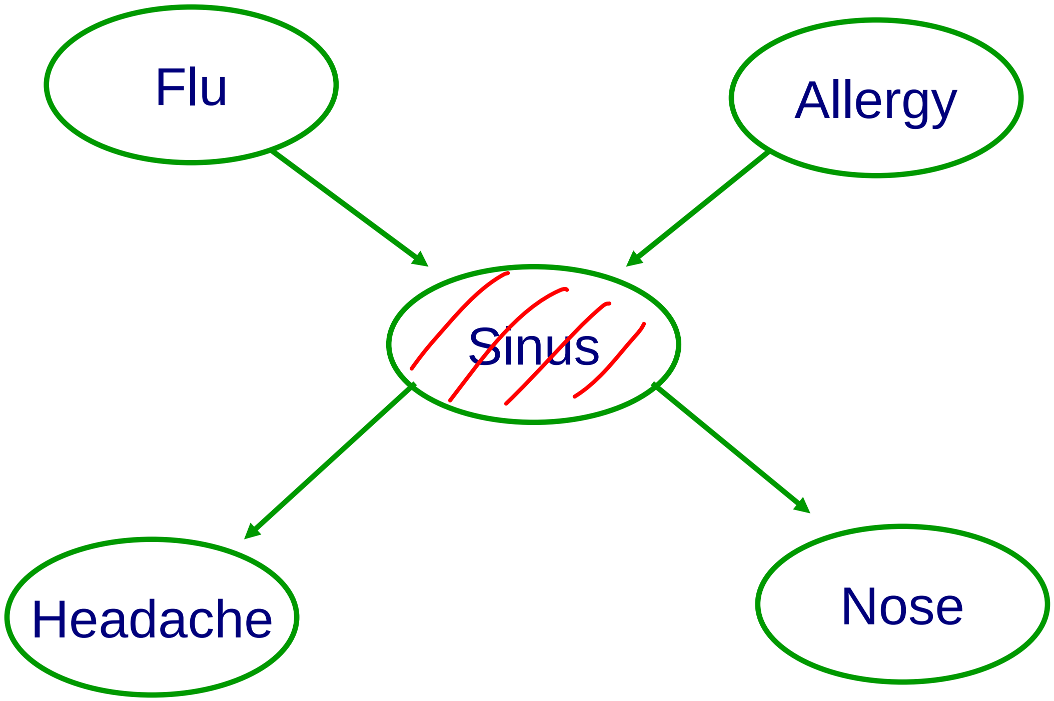
Number of parameters



Key: Independence assumptions

$A \perp B \equiv A$ indep of B

not $N \perp F$



$F \perp N \mid S$

$A \perp H \mid S$

$A \perp N \mid S$

$F \perp H \mid S$

$H \perp N \mid S$

Knowing sinus separates the variables from each other

(Marginal) Independence

- Flu and Allergy are (marginally) independent

$F \perp A$

$\Rightarrow P(F, A) = P(A) . P(F)$

$P(F) =$

Flu = t	.4
Flu = f	.6

- More Generally:

$X_i \perp X_j$

$\Rightarrow P(X_i, X_j) = P(X_i) . P(X_j)$

$P(A)$

Allergy = t	.3
Allergy = f	.7

$P(F, A)$

	Flu = t	Flu = f
Allergy = t	$.3 \times .4 = 0.12$	$.3 \times .6$
Allergy = f	$.4 \times .7$	$.7 \times .6$

Marginally independent random variables

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

possibly assignments

$$P(X=x, Y=y) = P(X=x) \cdot P(\cancel{X} Y=y)$$

- Shorthand:
  - **Marginal independence:** $P \models (\mathbf{X} \perp \mathbf{Y})$

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

$$P(X|Y) = P(X)$$
$$P(Y|X) = P(Y)$$

Conditional independence

- Flu and Headache are not (marginally) independent

not $F \perp H$

- Flu and Headache are independent given Sinus infection $F \perp H \mid S$

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

$$P(F \mid H, S) = P(F \mid S)$$

- More Generally: $X_i \perp X_j \mid Y \Leftrightarrow P(X_i \mid X_j, Y) = P(X_i \mid Y)$

Conditionally independent random variables

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

$\Rightarrow P(X=x | Z=z) = P(X=x | Z=z, Y=y)$

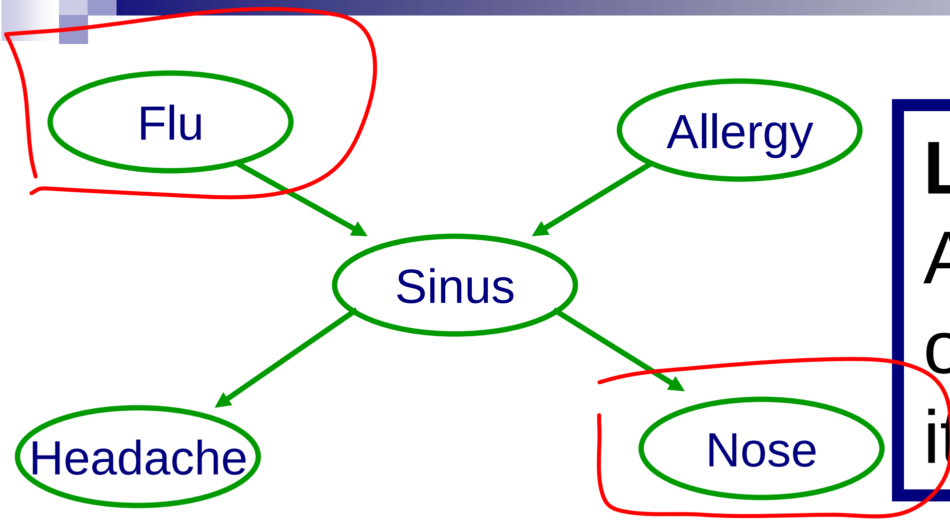
- Shorthand:
  - **Conditional independence:** $P \models (\mathbf{X} \perp \mathbf{Y} | \mathbf{Z})$
  - For $P \models (\mathbf{X} \perp \mathbf{Y} | \emptyset)$, write $P \models (\mathbf{X} \perp \mathbf{Y})$

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

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

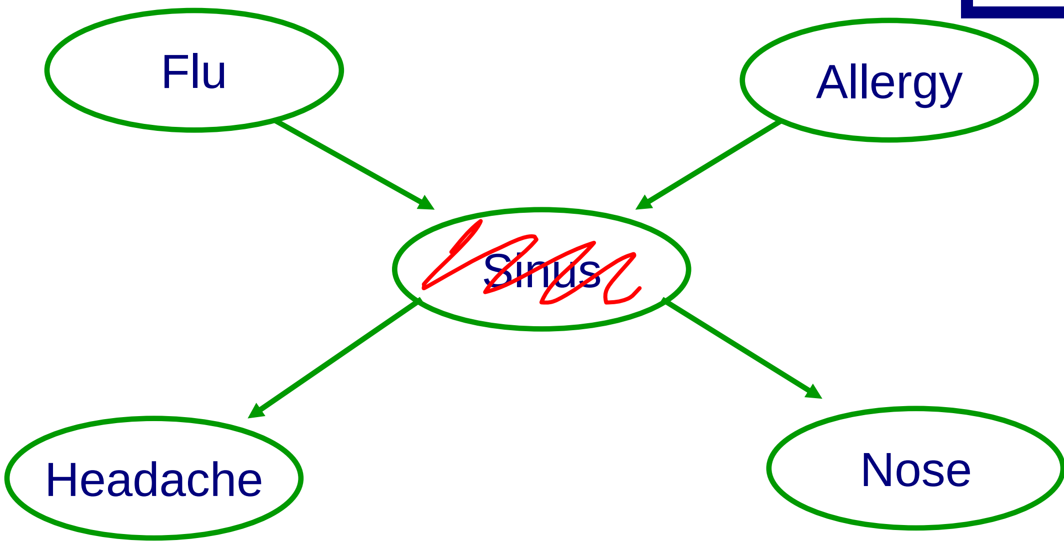
$N \perp \{F, A, H\} \mid S$

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

$F \perp A \mid \emptyset$

Explaining away

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



$\boxed{F \perp A}$

Suppose $S=t$

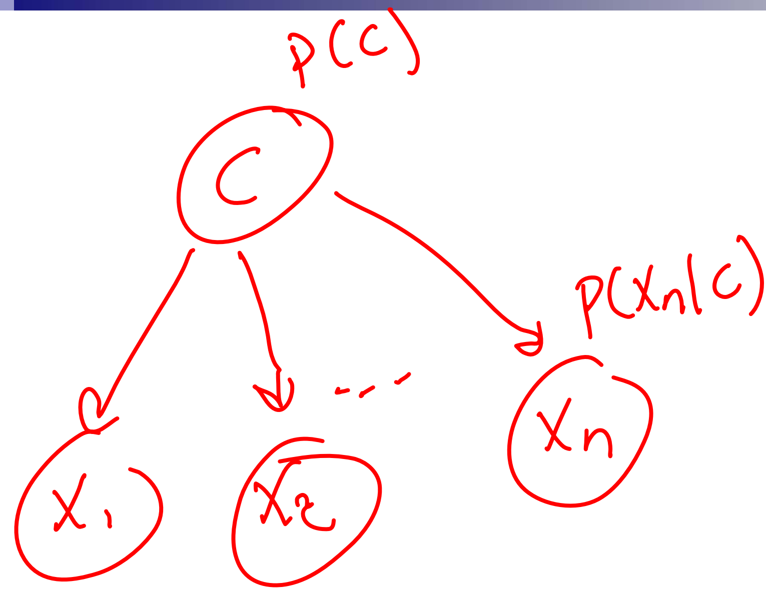
$P(A=t \mid S=t) > P(A=t)$

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

$< P(A=t \mid S=t)$

not $F \perp A \mid S$

Naïve Bayes revisited



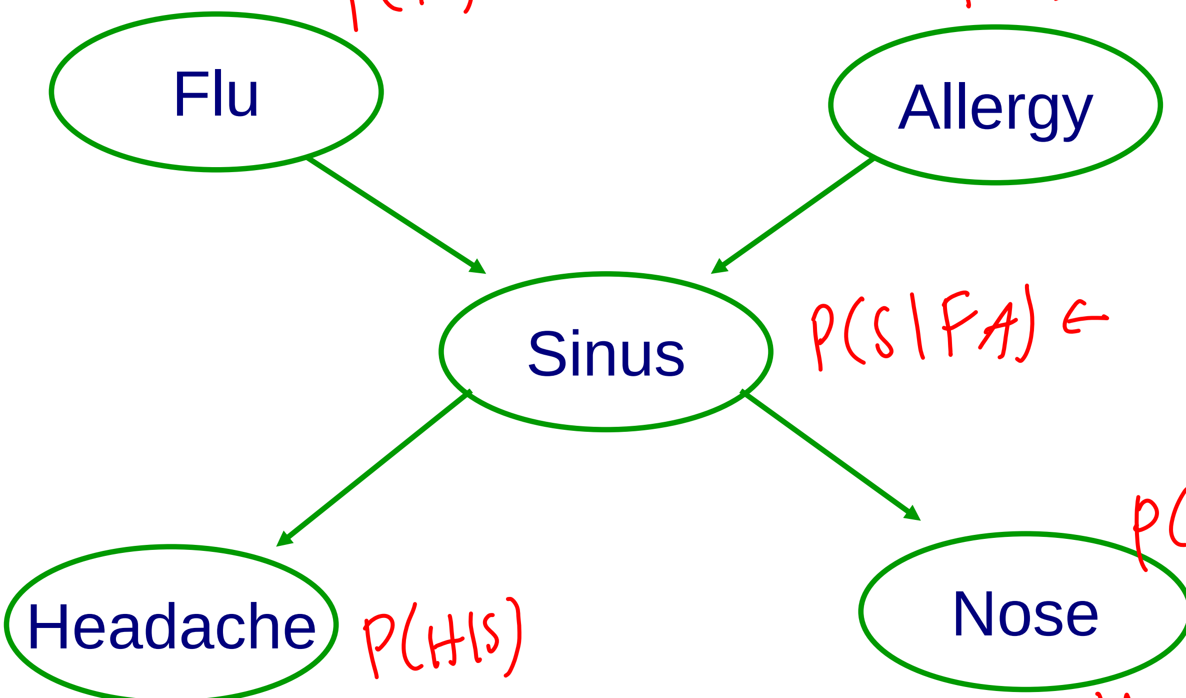
Local Markov Assumption: A variable X is independent of its non-descendants given its parents

$$\boxed{X_i \perp X_j \mid C}$$

$$X_1 \perp \{X_2 \ldots X_n\} \mid C$$

$$P(C, X_1 \ldots X_n) = P(C) \cdot \prod_{j=1}^{n} P(X_i \mid C)$$

What about probabilities?
Conditional probability tables (CPTs)



$P(F)$

$P(A)$

A	
t	.3
f	.7

$P(S|FA) \leftarrow$

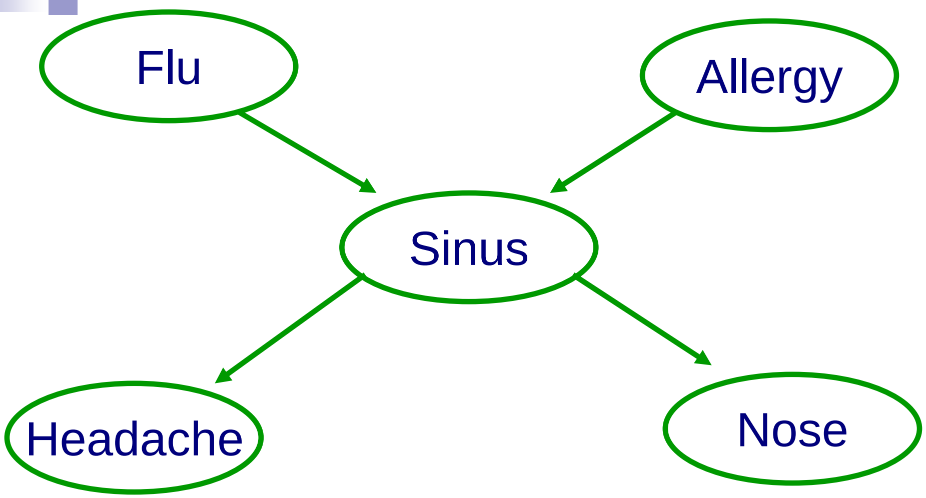
$P(N|S)$

$P(H|S)$

$P(H|S) =$

S \ H	t	f
t	.9	.1
f	.2	.8

Joint distribution



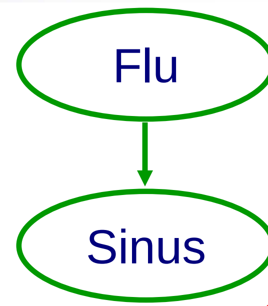
$$P(F, A, S, H, N)$$
$$= P(F) \cdot P(A) \cdot P(S \mid F A) \cdot P(H \mid S) \cdot P(N \mid S)$$

Why can we decompose? Markov Assumption!

The chain rule of probabilities

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

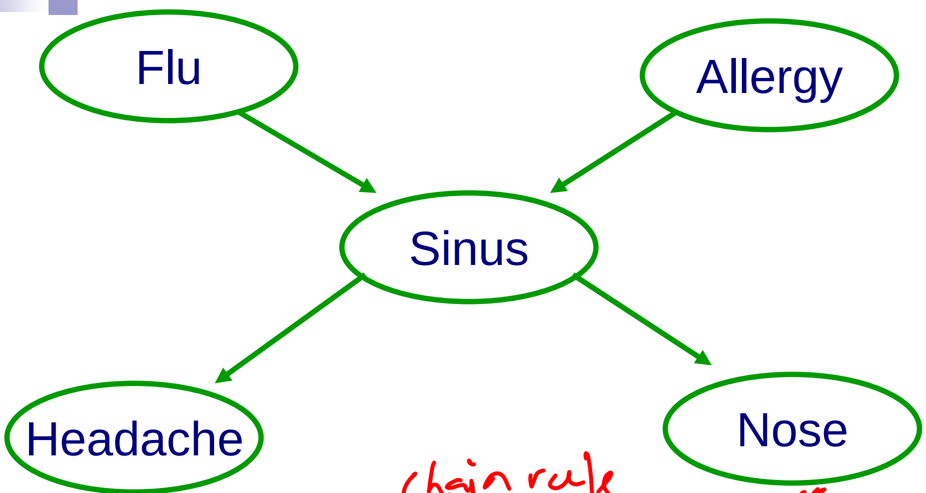
$P(F,S) = P(F) \cdot P(S|F)$



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

$P(FASHN) = P(F) \cdot P(A|F) \cdot P(S|AF) \cdot P(H|FAS) \cdot P(N|FASH)$

Chain rule & Joint distribution



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

$P(A \mid F) = P(A) \Leftarrow A \perp F$

$P(H \mid SFA) = P(H \mid S) \Leftarrow H \perp \{F A\} \mid S$

$P(N \mid SFAH) = P(N \mid S) \Leftarrow N \perp \{F A H\} \mid S$

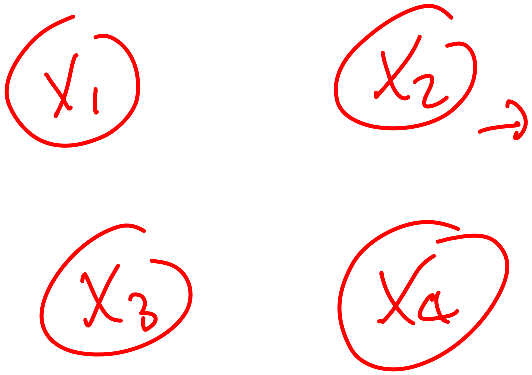
chain rule

$P(FASHN) = P(F)\, P(A \mid F)\, P(S \mid FA) \cdot P(H \mid SFA) \cdot P(N \mid SFAH)$

$P(F) \quad P(A) \quad P(S \mid FA) \quad P(H \mid S) \quad P(N \mid S)$

Two (trivial) special cases

Edgeless graph



$P(X_1\, X_2\, X_3\, X_4)$

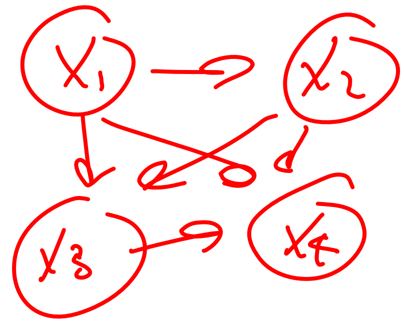
$= P(X_1) \cdot P(X_2)\, P(X_3)\, P(X_4)$

high bias

all nodes indep.

$X_1 \perp \{X_2 X_3 X_4\}$

Fully-connected graph



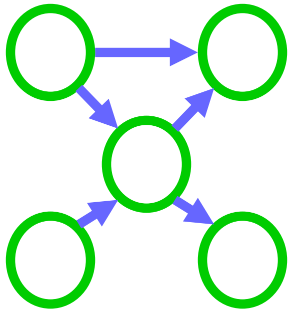
$P(X_1)\, P(X_2 | X_1)\, P(X_3 | X_1, X_2)$
$\cdot\, P(X_4 | X_1, X_2, X_3)$

high variance

no indeps.

The Representation Theorem – Joint Distribution to BN

BN:



Encodes independence assumptions

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

real world

Obtain

Joint probability distribution:

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

can represent the real world

Real Bayesian networks applications

- 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

- Directed acyclic graph
  - Encodes independence assumptions

- CPTs

- 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 – $P(X_i \mid \mathbf{Pa}_{X_i})$

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

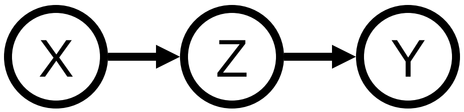
Independencies encoded in BN

- We said: All you need is the local Markov assumption
  - $(X_i \perp NonDescendants_{X_i} \mid \mathbf{Pa}_{X_i})$
- 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!!!

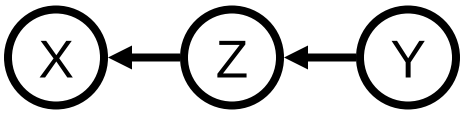
Understanding independencies in BNs – BNs with 3 nodes

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

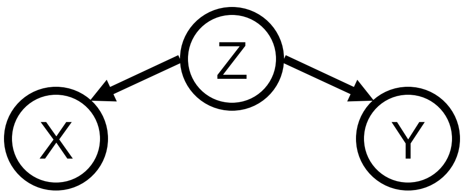
Indirect causal effect:



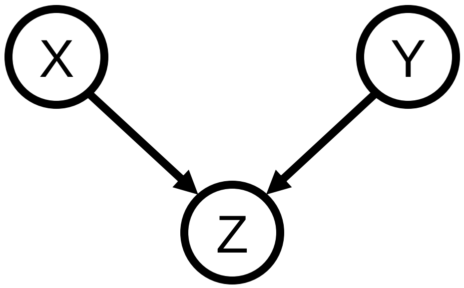
Indirect evidential effect:



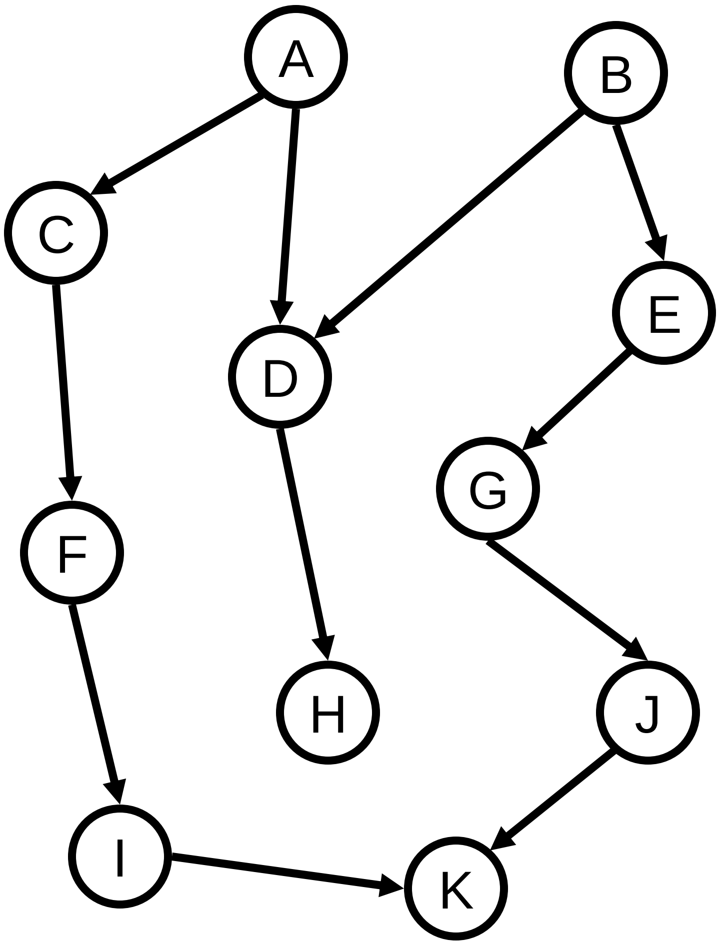
Common cause:



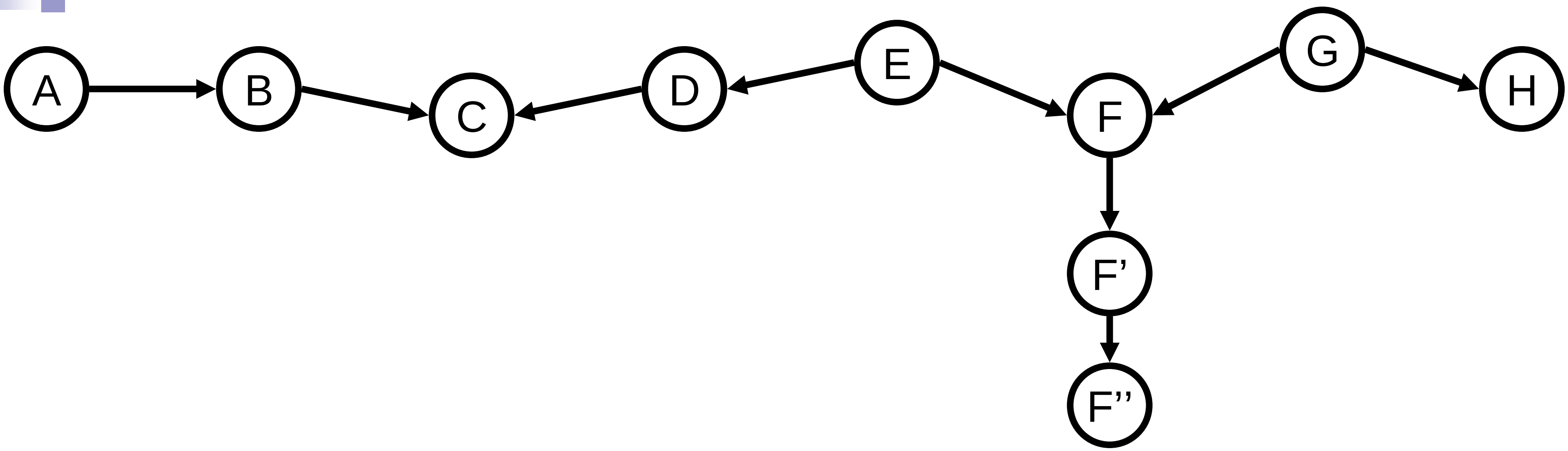
Common effect:



Understanding independencies in BNs – Some examples



An active trail – Example



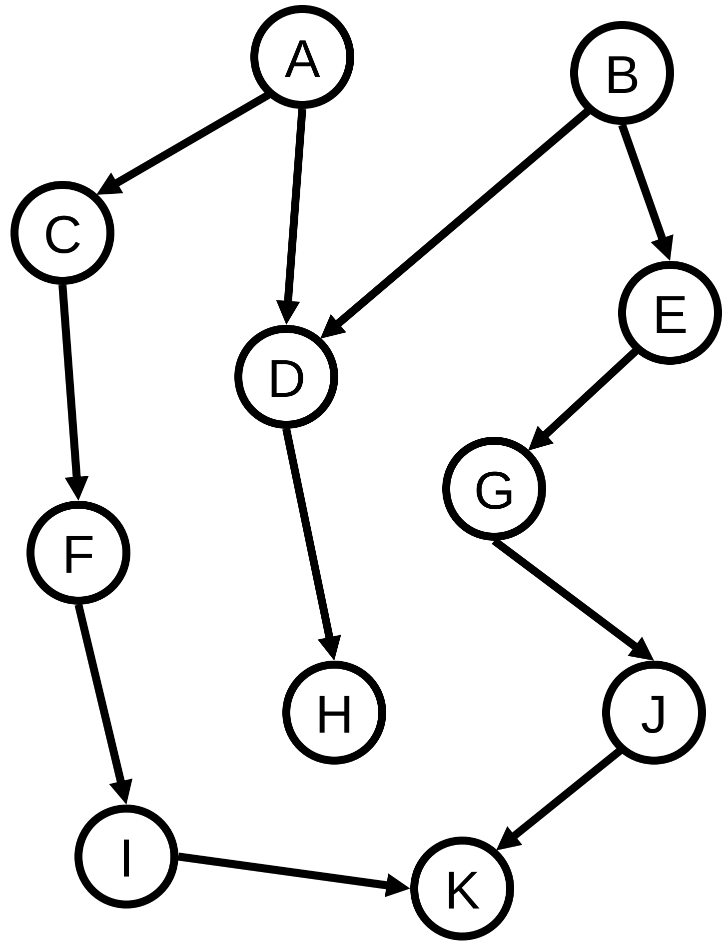
When are A and H independent?

Active trails formalized

- A path $X_1 - X_2 - \cdot\cdot\cdot - X_k$ 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*

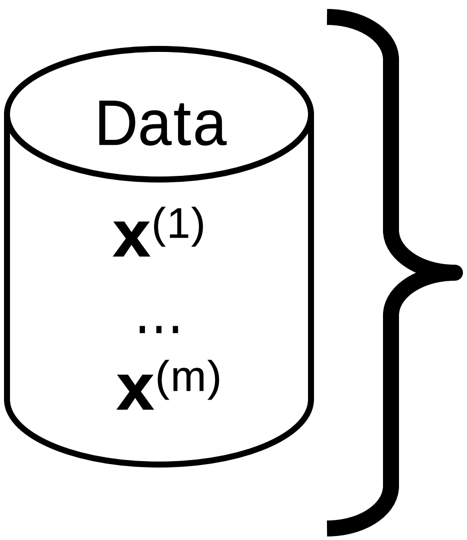
"Simpler" BNs

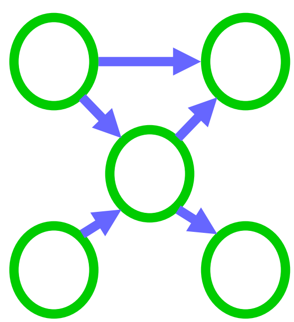
- A distribution can be represented by many BNs:

- Simpler BN, requires fewer parameters

Learning Bayes nets

	Known structure	Unknown structure
Fully observable data		
Missing data		





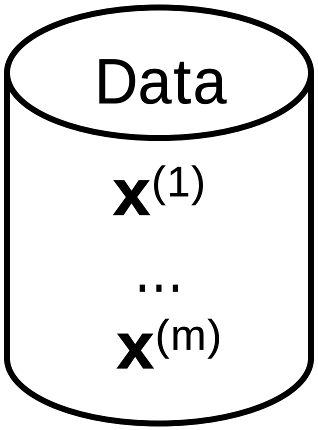
structure

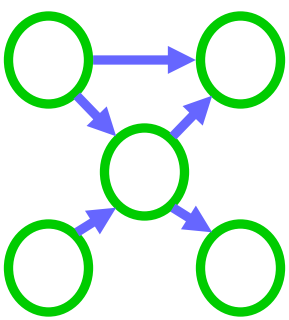
+

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

parameters

Learning the CPTs





For each discrete variable X_i

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

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