Bayesian Networks – Representation

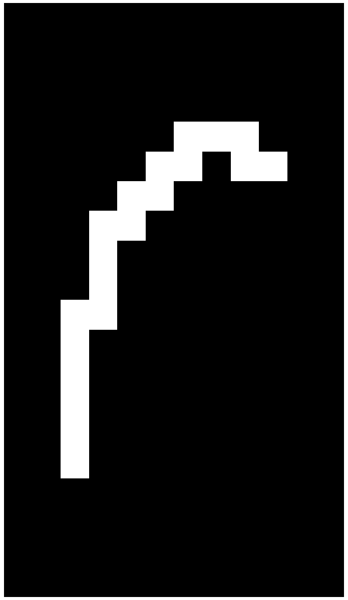
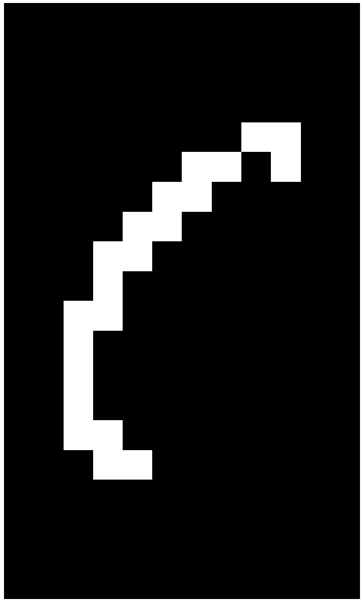
Machine Learning – 10701/15781

Carlos Guestrin

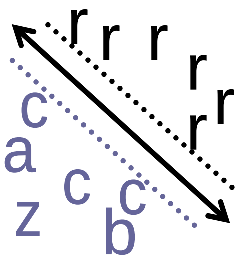
Carnegie Mellon University

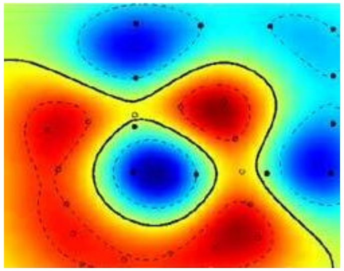
March 16th, 2005

Handwriting recognition

Character recognition, e.g., kernel SVMs





Webpage classification

→ Company home page

vs

Personal home page

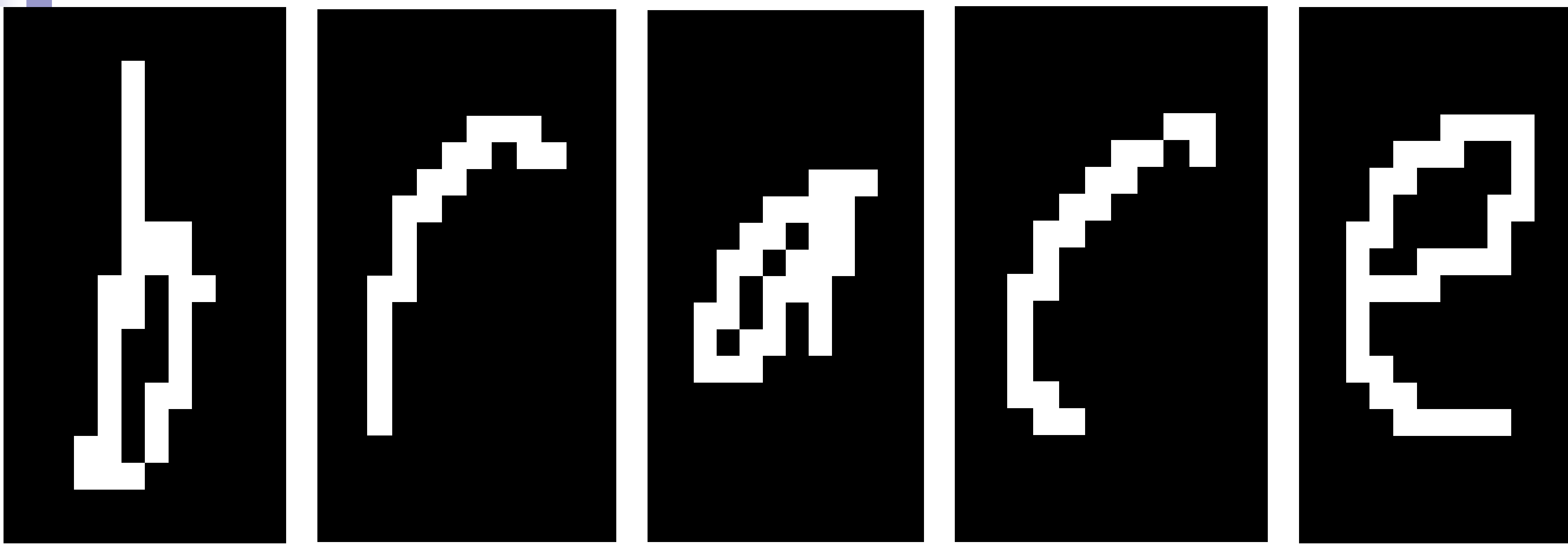
vs

Univeristy home page

vs

…

Handwriting recognition 2



A, B, ... Z

A, B, ... Z

one class per word: AAAAA

AAAAB

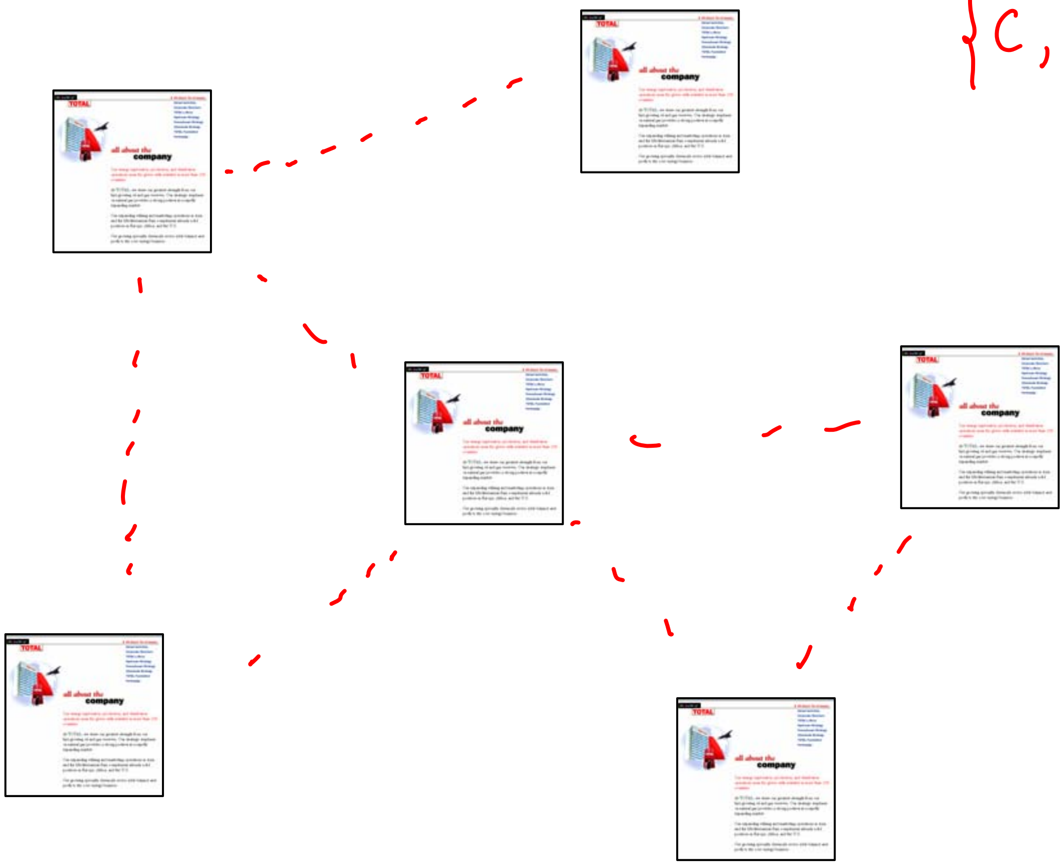
⋮

ZZZZZ

26^5

Webpage classification 2

$\{C, P, U, S\}^{\#pages}$

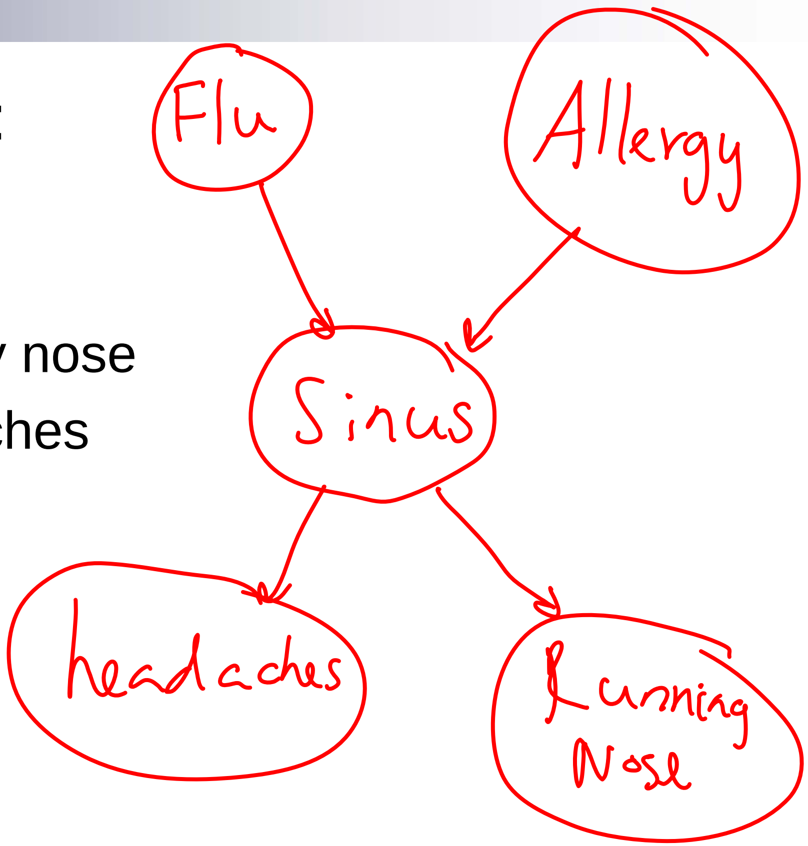


Today – Bayesian networks

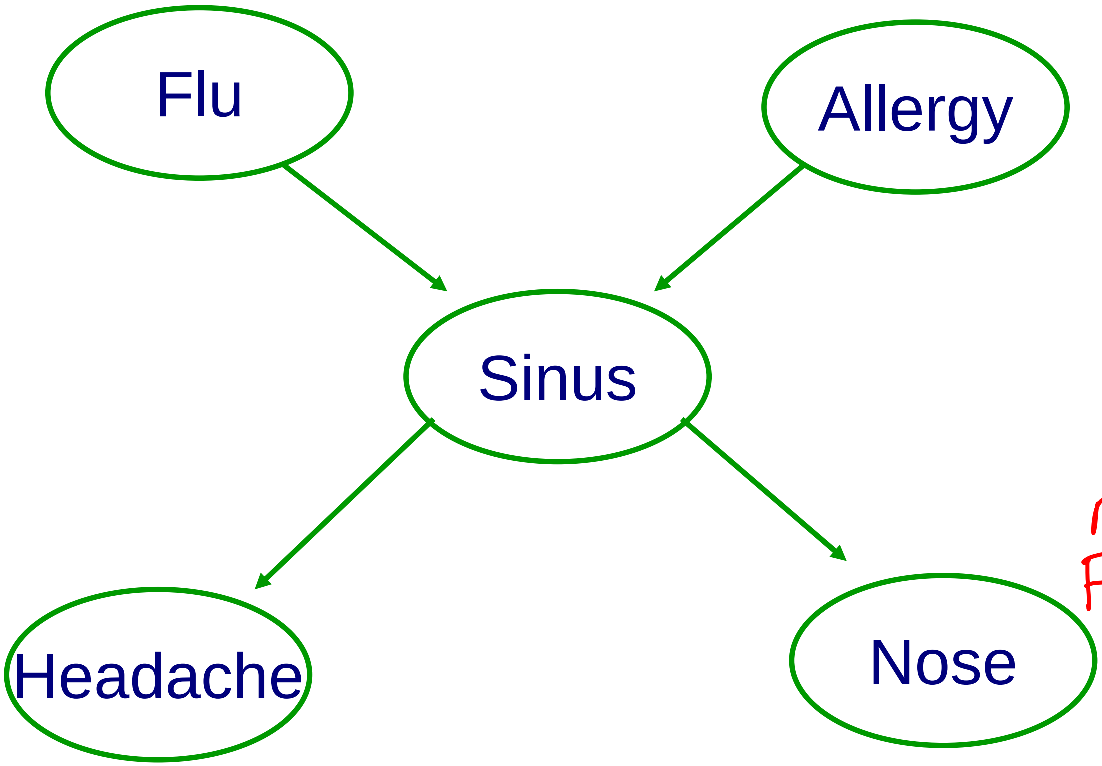
- One of the most exciting advancements in statistical AI in the last 10-15 years
- Generalizes naïve Bayes and logistic regression classifiers
- Compact representation for exponentially-large probability distributions
- Exploit conditional independencies

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



- Inference

$P(Flu=t \mid H=t, N=f)$

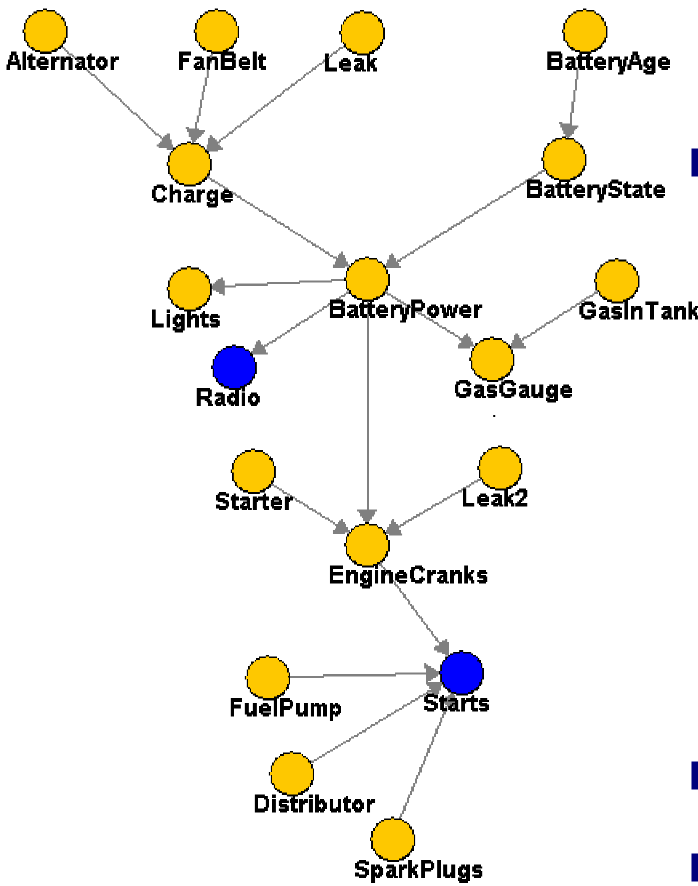
- Most probable explanation

$\max_{Flu, Allergy} P(Flu, Allergy \mid H=t, N=f)$

- Active data collection

what's best question to ask

Car starts BN

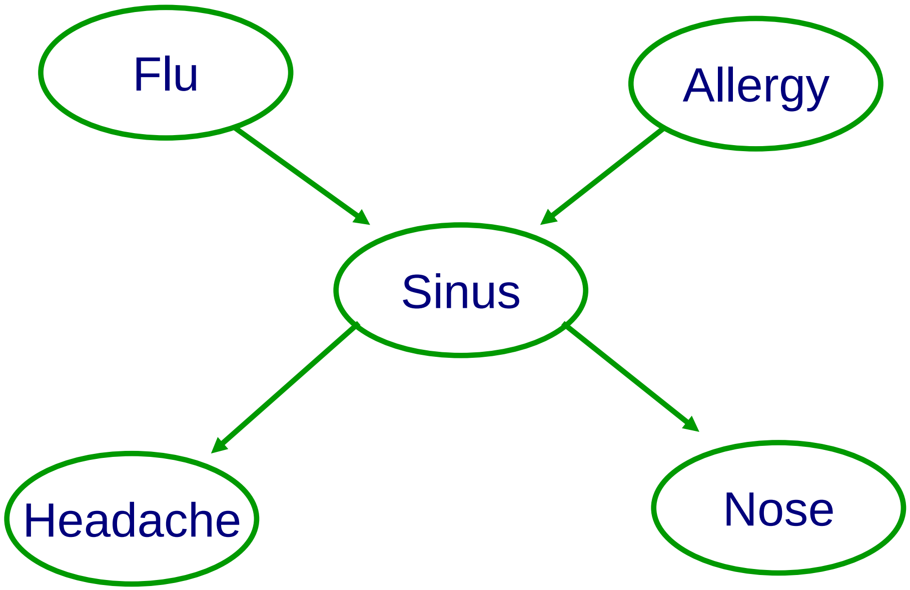


- 18 binary attributes
- Inference $P(B,S) = \sum_{A,F,L,\ldots} P(A,F,L,\ldots B,S)$
  - P(BatteryAge|Starts=f)

$$P(B{=}old \mid S{=}f) = \frac{P(B{=}old, S{=}f)}{P(S{=}f)}$$

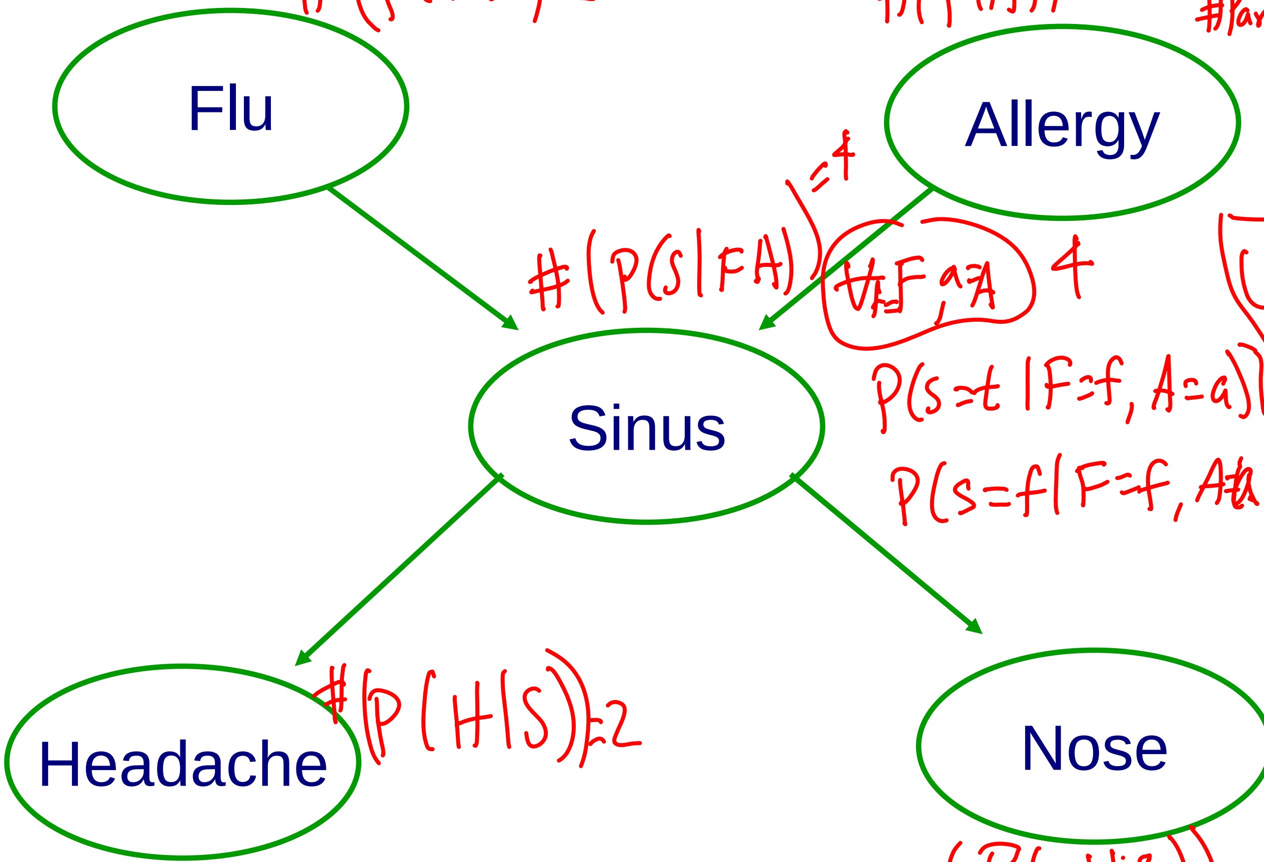
- 2^{18} terms, why so fast?
- Not impressed?
  - HailFinder BN – more than 3^{54} = 58149737003040059690390169 terms

Factored joint distribution - Preview



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

Number of parameters



$\#(P(F)) = 1$

$\#(P(A)) = 1$

$\#Par(P(F, A, S, H, N)) = 2^5 - 1 = 31$

$\#(P(S|FA)) = 4$

$\forall F, A$ 4

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

$P(S=f \mid F=f, A=a)$ } 1

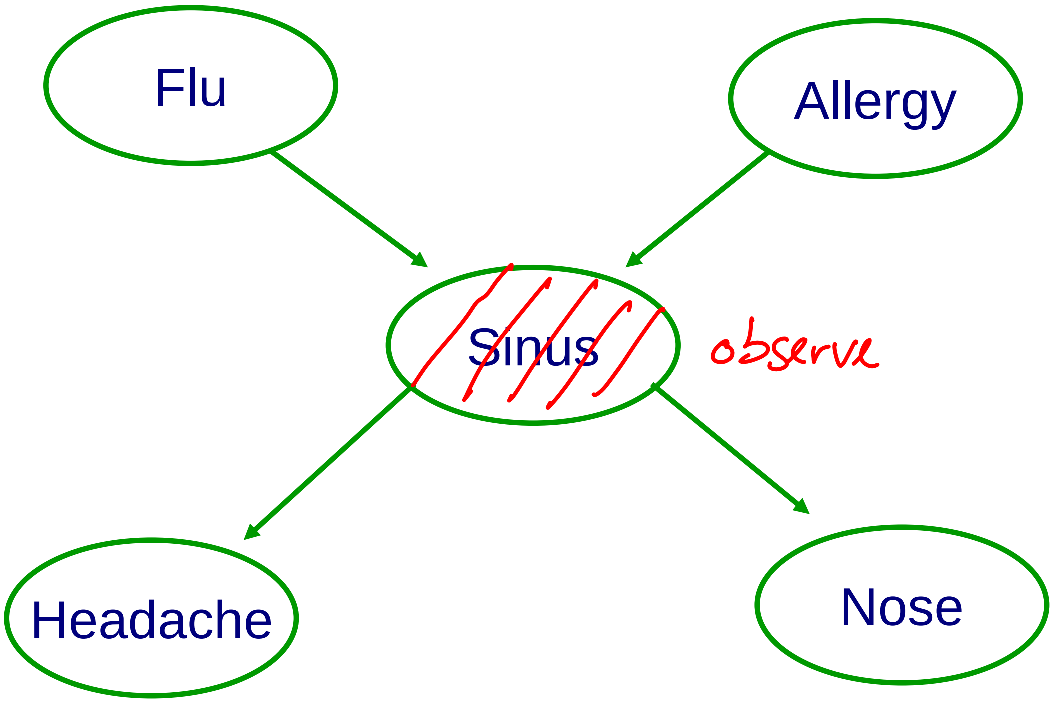
$\#(P(H|S)) = 2$

$\#(P(N|S)) = 2$

Using BN:

$\#pars = 10$

Key: Independence assumptions



F, A independent a priori

F, N are "dependent"

F, N are independent given S

Knowing sinus separates the variables from each other

(Marginal) Independence

- Flu and Allergy are (marginally) independent

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

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

Flu = t	0.1
Flu = f	0.9

Allergy = t	0.2
Allergy = f	0.8

- More Generally:

Independence:

A, B independent: $(A \perp B)$

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

$\Updownarrow$

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

$\Updownarrow$

$P(AB) = P(A).P(B)$

$P(F,A)$	Flu = t	Flu = f
Allergy = t	0.1×0.2	0.9×0.2
Allergy = f	0.1×0.8	0.9×0.8

Conditional independence

- Flu and Headache are not (marginally) independent

  $P(F) \neq P(F|H)$

- Flu and Headache are independent given Sinus infection

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

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

- More Generally:

  $(A \perp B \mid S)$ A,B independent given S

  $P(A|S) = P(A|SB)$

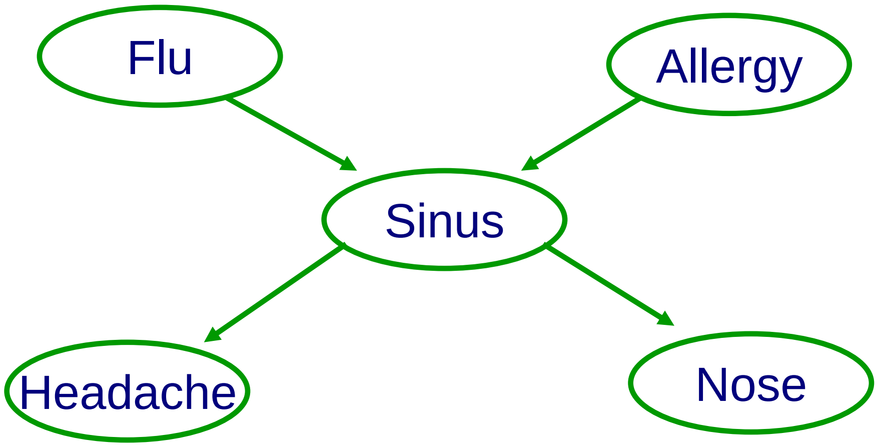
  $\Updownarrow$

  $P(B|S) = P(B|SA)$

  $\Updownarrow$

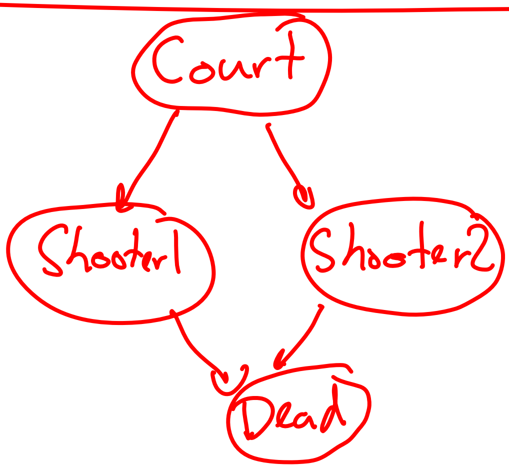
  $P(AB|S) = P(A|S) \cdot P(B|S)$

The independence assumption



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

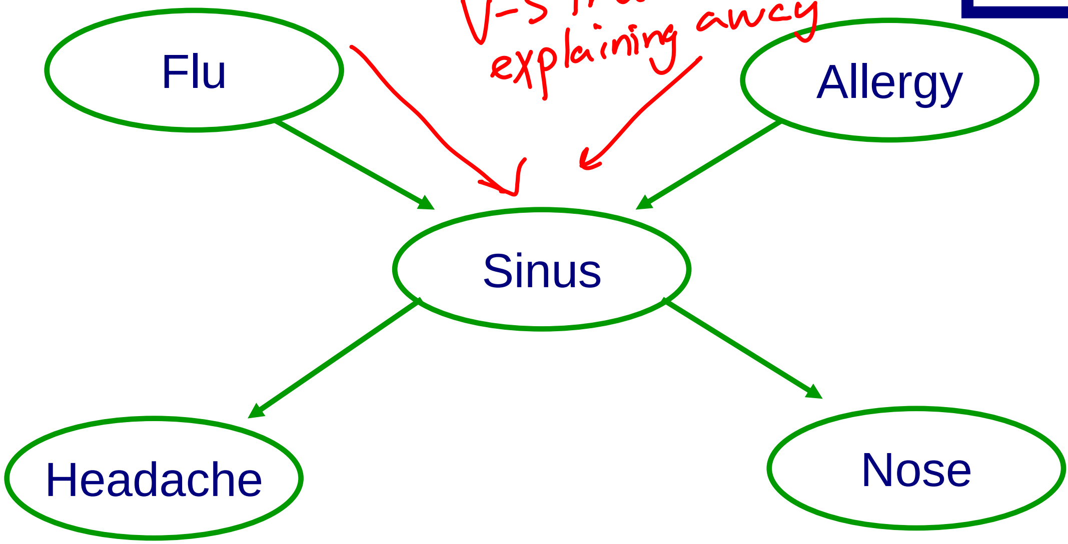
Example with Cycle



Explaining away

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

V-structure
explaining away



F, A independent

Give you S = t

F, A not independent:

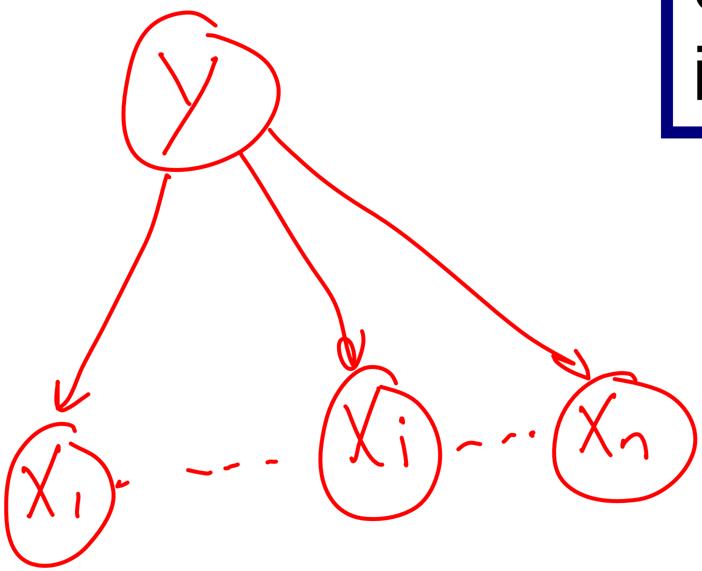
$\cancel{(F \perp A \mid S)}$

observe S = t → increase $P(F \mid S=t)$
$P(A \mid S=t)$

but S=t, A=t → decrease prob. F=t

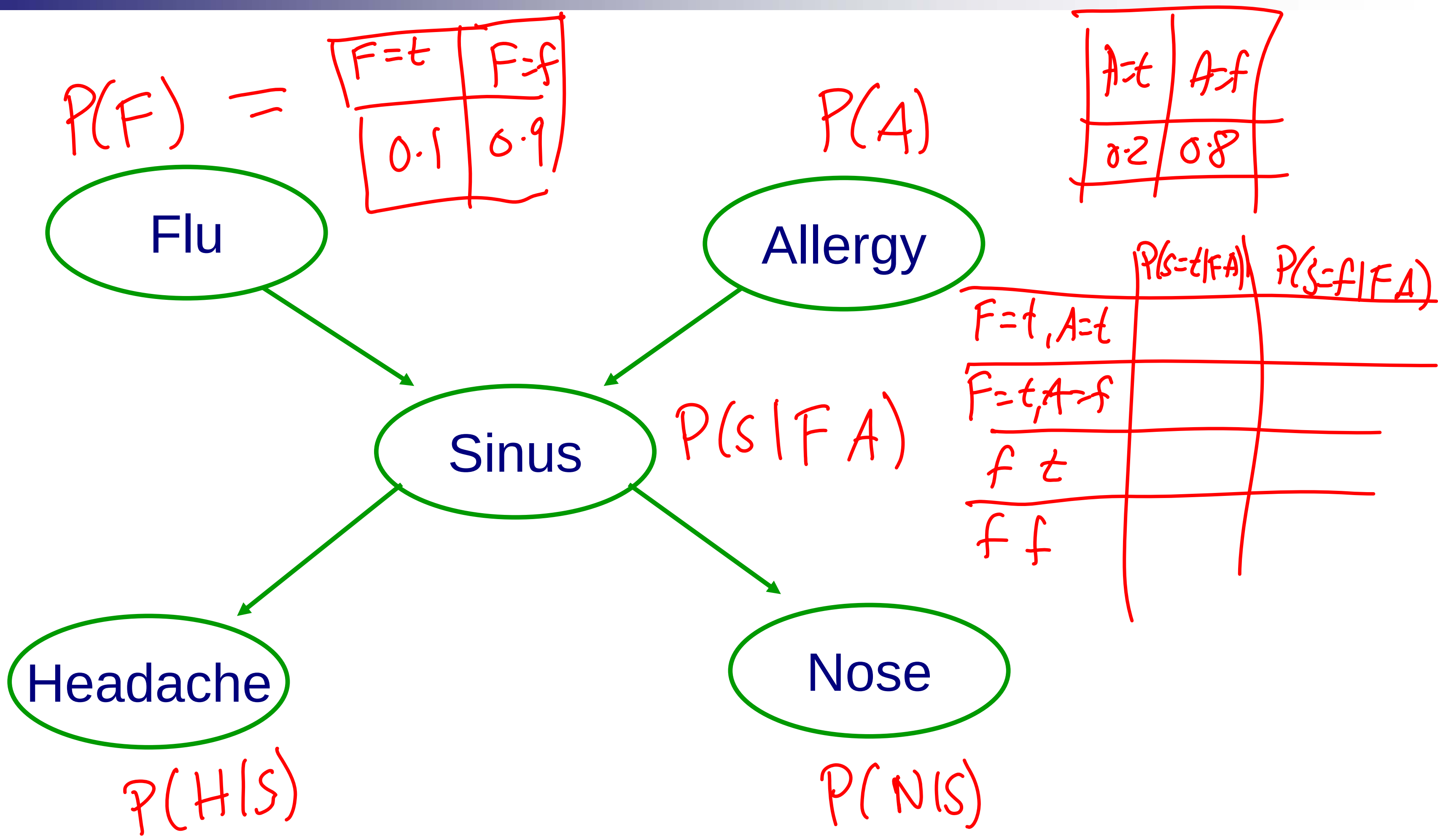
Naïve Bayes revisited

Y – class
X_i – features

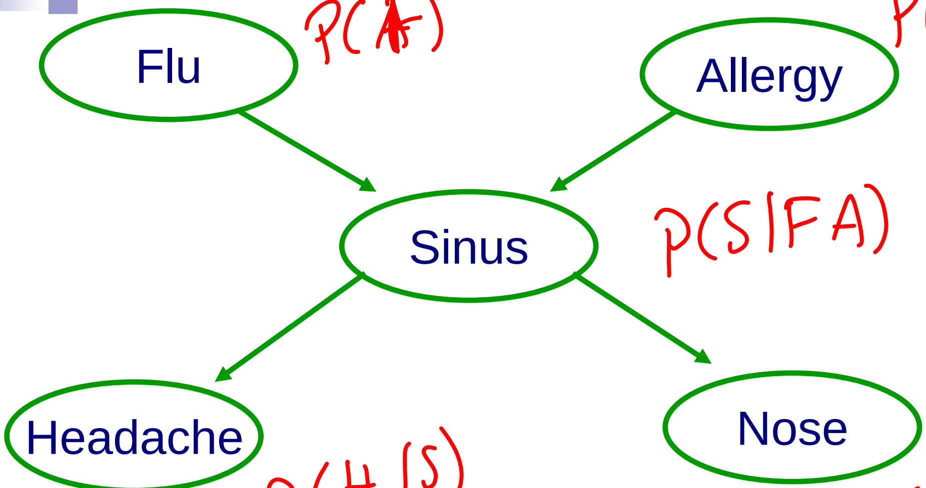


Local Markov Assumption: A variable X is independent of its non-descendants given its parents

What about probabilities?
Conditional probability tables (CPTs)



Joint distribution



$P(F)$ $P(A)$ $P(S|FA)$ $P(H|S)$ $P(N|S)$

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

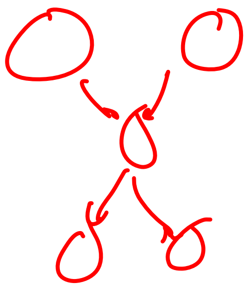
Why can we decompose? Markov Assumption!

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 F, A, S, H, N
- Directed acyclic graph
  - Encodes independence assumptions



- CPTs

Product of local tables

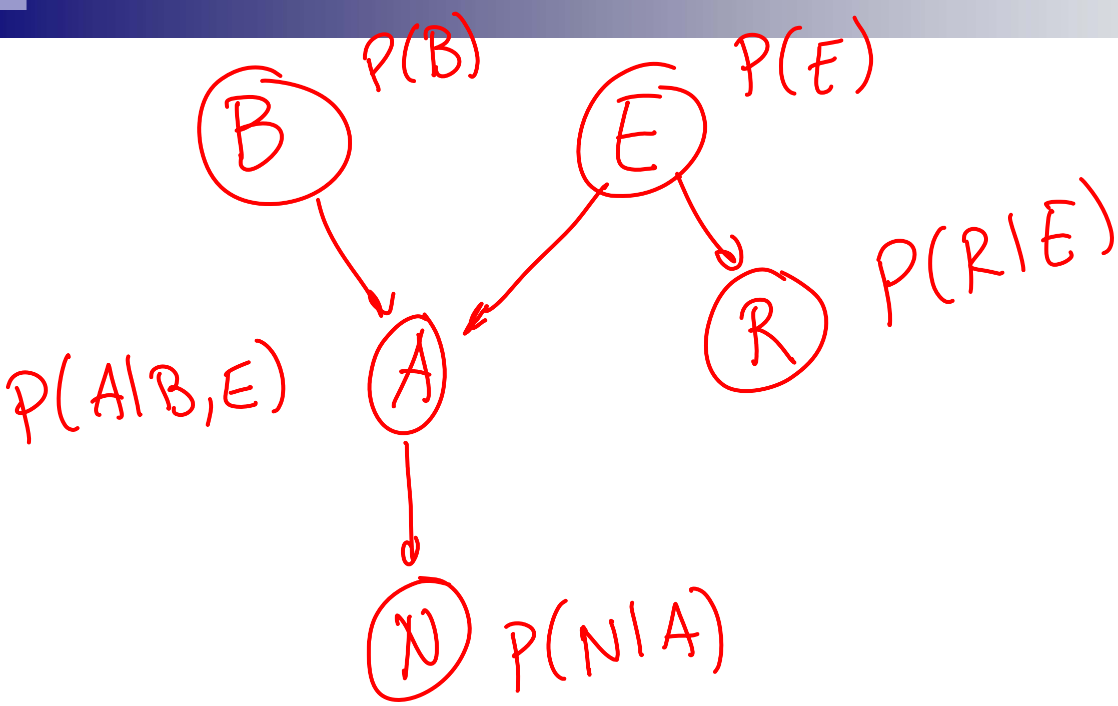
- Joint distribution:

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

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

Defining a BN

- Given a set of variables and conditional independence assumptions
- Choose an ordering on variables, e.g., $X_1, \ldots, X_n$
- For i = 1 to n
  - Add X_i to the network
  - Define parents of X_i, $\mathbf{Pa}_{X_i}$, in graph as the minimal subset of $\{X_1,\ldots,X_{i-1}\}$ such that local Markov assumption holds – X_i independent of rest of $\{X_1,\ldots,X_{i-1}\}$, given parents $\mathbf{Pa}_{Xi}$
  - Define/learn CPT – $P(X_i | \mathbf{Pa}_{Xi})$

How many parameters in a BN?

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

~~[illegible]~~ # vals X_i is $|X_i|$

$$\#\,\text{param}\left[P(X_i \mid Pa_{X_i})\right] = \left[\prod_{X_j \in Pa_{X_i}} |X_j|\right]\left(|X_i| - 1\right)$$

$$\#\,\text{params}(BN) = \sum_i \#\,\text{param}\left[P(X_i \mid Pa(X_i))\right]$$

$$<< \left[\prod_i |X_i|\right] - 1$$

Defining a BN 2

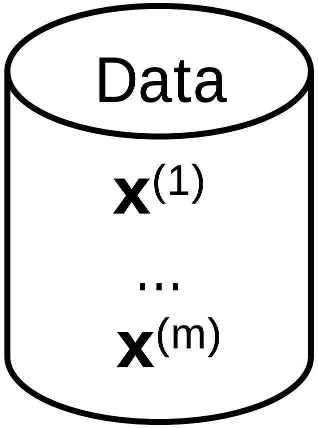
We may not know conditional independence assumptions and even variables

- Given a set of variables and conditional independence assumptions
- Choose an ordering on variables, e.g., X_1, …, X_n
- For i = 1 to n
  - Add X_i to the network
  - Define parents of X_i, **P**
    subset of $\{X_1,\ldots,X_{i-1}\}$ such that local Markov
    assumption holds – X_i independent of rest of $\{X_1,\ldots,X_{i-1}\}$, given parents $\mathbf{Pa}_{Xi}$
  - Define/learn CPT – $P(X_i| \mathbf{Pa}_{Xi})$

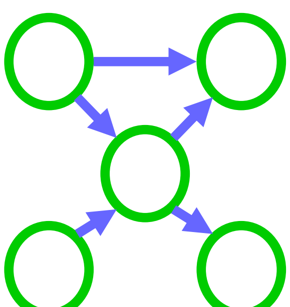
There are good orderings and bad ones – A bad ordering may need more parents per variable $\rightarrow$ must learn more parameters

How???

Learning the CPTs



$x^{(i)} = (x_1^{(i)}, \ldots, x_n^{(i)})$



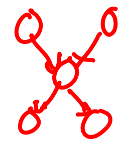
For each discrete variable X_i

$$P(X_i \mid X_j, X_k) = \frac{P(X_i, X_j, X_k)}{P(X_j, X_k)}$$

$$\approx \frac{\text{Count}(X_i, X_j, X_k)}{\text{Count}(X_j, X_k)}$$

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

Learning Bayes nets

	Known structure 	Unknown structure
Fully observable data $X_1^{(i)} = x_1, \ldots X_n^{(i)} = x_n$	counts !	next next lecture
Missing data $X_1^{(i)} = x_1, \ldots, X_n^{(i)} = ?$	later in course	next semester

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