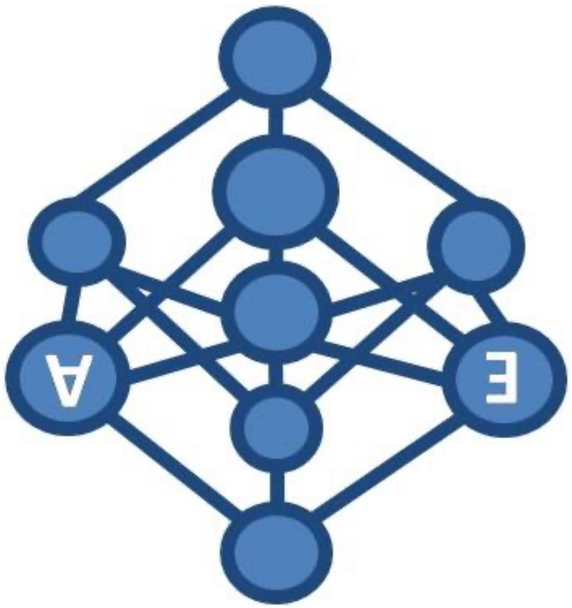
TECHNISCHE UNIVERSITÄT DARMSTADT

Probabilistic Graphical Models*
Bayesian Networks



*Thanks to Carlos Guestrin, Pedro Domingos and many others for making their slides publically available

What you need to know from last class

- Basic definitions of probabilities
- Independence
- Conditional independence
- The chain rule
- Bayes rule

Let‘s start on Bayesian Networks

- One of the most exciting recent advancements in statistical AI

Judea Pearl won the ACM Turing Award 2012 for his fundamental contributions to artificial intelligence through the development of a calculus for probabilistic and causal reasoning

- Compact representation for exponentially-large probability distributions
- Fast marginalization
- Exploit conditional independencies

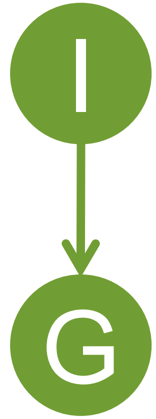
Representing Distributions by Enumeration

- Consider $P(X_i)$
  - Assign a probability to each $x_i \in \mathrm{Val}(X_i)$
  - Number of parameters assuming $|\mathrm{Val}(X_i)| = k$ **?**

    $k-1$
- Now, consider $P(X_1,\ldots,X_n)$
  - How many parameters assuming again $|\mathrm{Val}(X_i)| = k$ **?**
  - Same thing, $k^n - 1$
- **Bayesian networks will often need fewer parameters. What is the trick?**

Conditional Parameterization (2 nodes)

- **This is rarely the case !**
- **Grade is influences by Intelligence**
- Make use of **chain rule**



$$P\left(G,I\right)=P(G\,|\,I)\cdot P(I)$$

$$P(I=VH,G=B)$$

$$=P(I=VH)\cdot P(G=B\,|\,I=VH)$$

$$=0.85\cdot 0.1$$

$$=0.085$$

	VH	H
P(I)	0.85	0.15

	I=	VH	H
P(G|I)	A	0.9	0.5
	B	0.1	0.5

Represent conditional distributions graphically

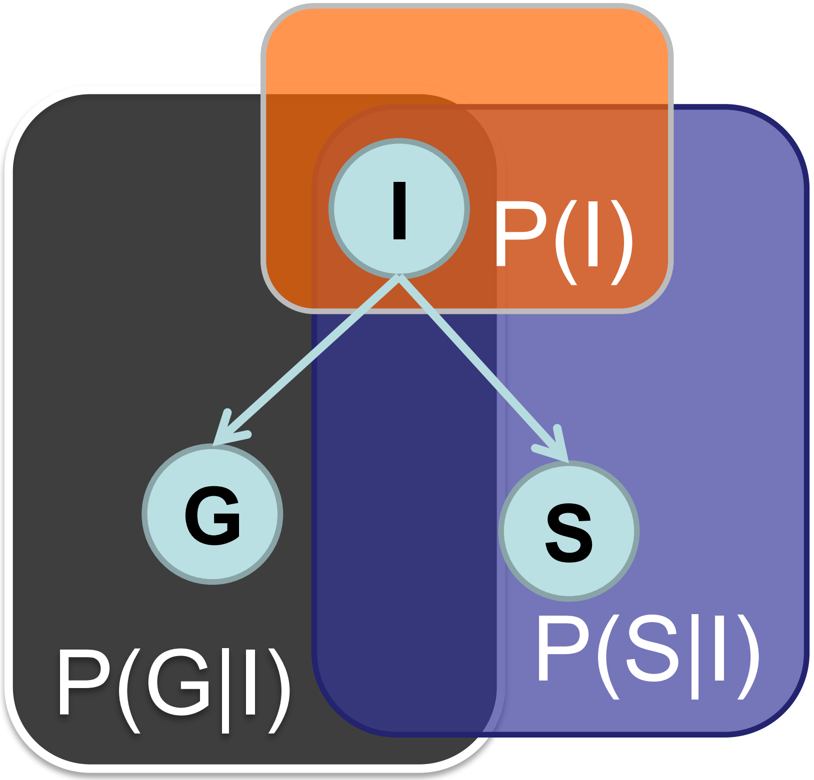
Same # of parameters as by enumeration

Conditional Parameterization (3 nodes) and what if variables are independent?

- **Grade and SAT score are influenced by Intelligence**
- **but** $(G \perp S \mid I)$ **, i.e.,** $P(G \mid S, I) = P(G \mid I)$

$$
\begin{aligned}
& P(G,S,I) \\
& = P(G,S \mid I) \cdot P(I) \\
& = P(G \mid S,I) \cdot P(S \mid I) \cdot P(I) \\
& = P(G \mid I) \cdot P(S \mid I) \cdot P(I)
\end{aligned}
$$

Indpendence can lead to smaller # of parameters as by enumeration



Can we even get a linear complexity?

- $(X_i \perp X_j), \forall i,j$ is not enough
- We must assume that $(\mathbf{X} \perp \mathbf{Y}), \forall \mathbf{X}, \mathbf{Y}$ subsets of $\{X_1, \ldots, X_n\}$

Let *X1 and X2 be drawn from Bernoulli(0.5*) and *X3* = *X1 xor X2*. Now, *P(Xi,Xj)* = *1/4* = *P(Xi)P(Xj) (use a table to show this)*. Since *X3 depends deterministically on X1 and X2 it cannot be the case that X1,X2 are independent of X3.*

- Now, we can write $P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i)$
- How many parameters?

$$n \cdot (k-1) = O(n)$$

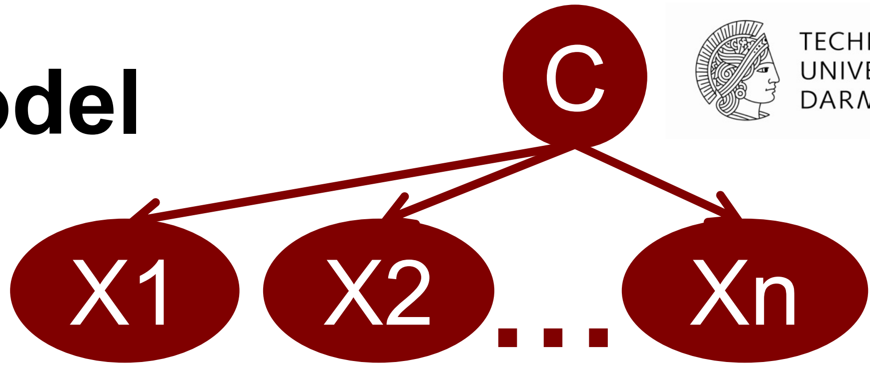
The Naïve Bayes Model

- **Now, your first real Bayesian network !**

- **Class variable:** C
- **Evidence variables:** $\{X_1, \ldots, X_n\}$
- **Assume** that $(\mathbf{X} \perp \mathbf{Y} \mid C), \forall \mathbf{X}, \mathbf{Y}$ subsets of $\{X_1, \ldots, X_n\}$

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

The Naïve Bayes Model



$$P(X_1,\ldots,X_n,C)$$

$$= P(C)\cdot P(X_1 \mid C)\cdot P(X_2 \mid X_1,C)\cdot\ldots\cdot \boxed{P(X_n \mid X_1,\ldots X_{n-1},C)}$$

So far, no assumptions. Now due to $(\mathbf{X} \perp \mathbf{Y} \mid C), \forall \mathbf{X}, \mathbf{Y}$

it becomes the Naïve Bayes model

$$P(X_2 \mid X_1,C) = P(X_2 \mid C) \qquad P(X_n \mid X_1,\ldots X_{n-1},C) = P(X_n \mid C)$$

$$(X_1 \perp X_2 \mid C) \qquad (X_n \perp X_1 X_2 \ldots X_{n-1} \mid C)$$

Is this spam?

From: "" <takworlld@hotmail.com>
Subject: real estate is the only way... gem oalvgkay

Anyone can buy real estate with no money down

Stop paying rent TODAY !

There is no need to spend hundreds or even thousands for similar courses

I am 22 years old and I have already purchased 6 properties using the methods outlined in this truly INCREDIBLE ebook.

Change your life NOW !

===

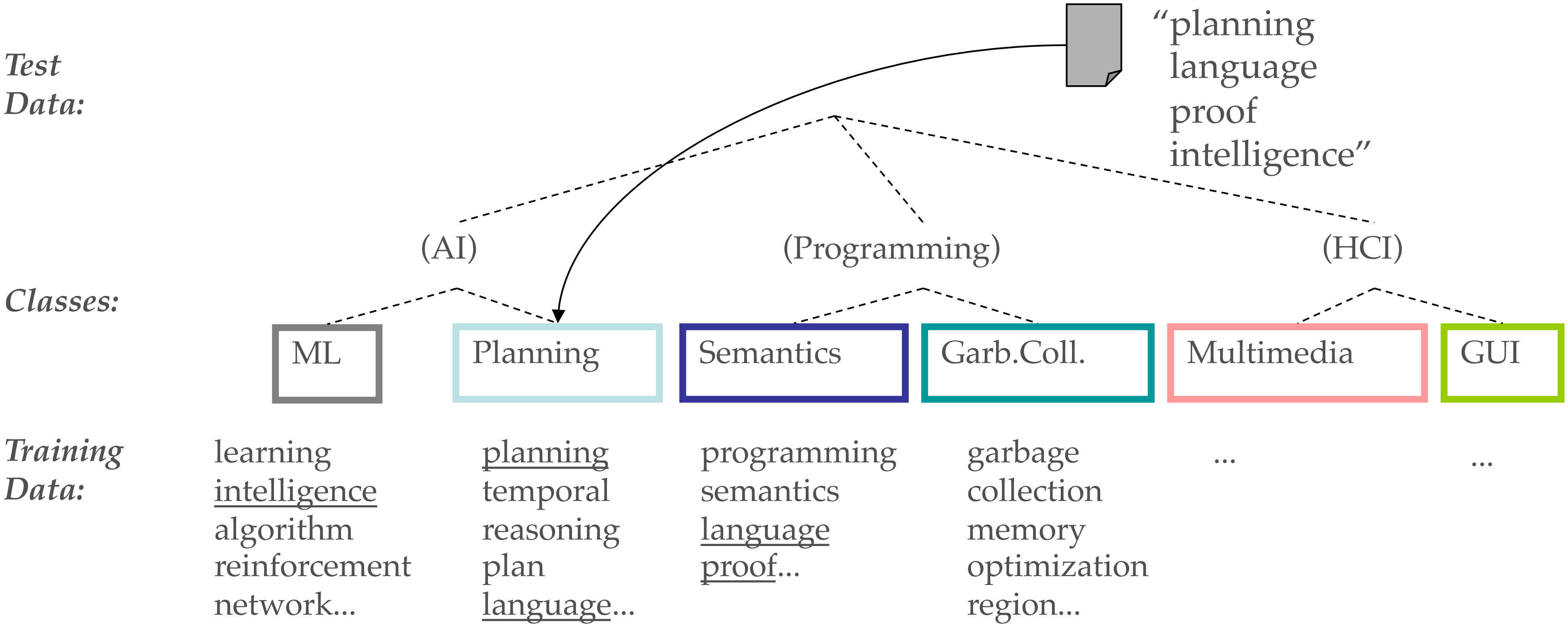
Click Below to order:
http://www.wholesaledaily.com/sales/nmd.htm

===

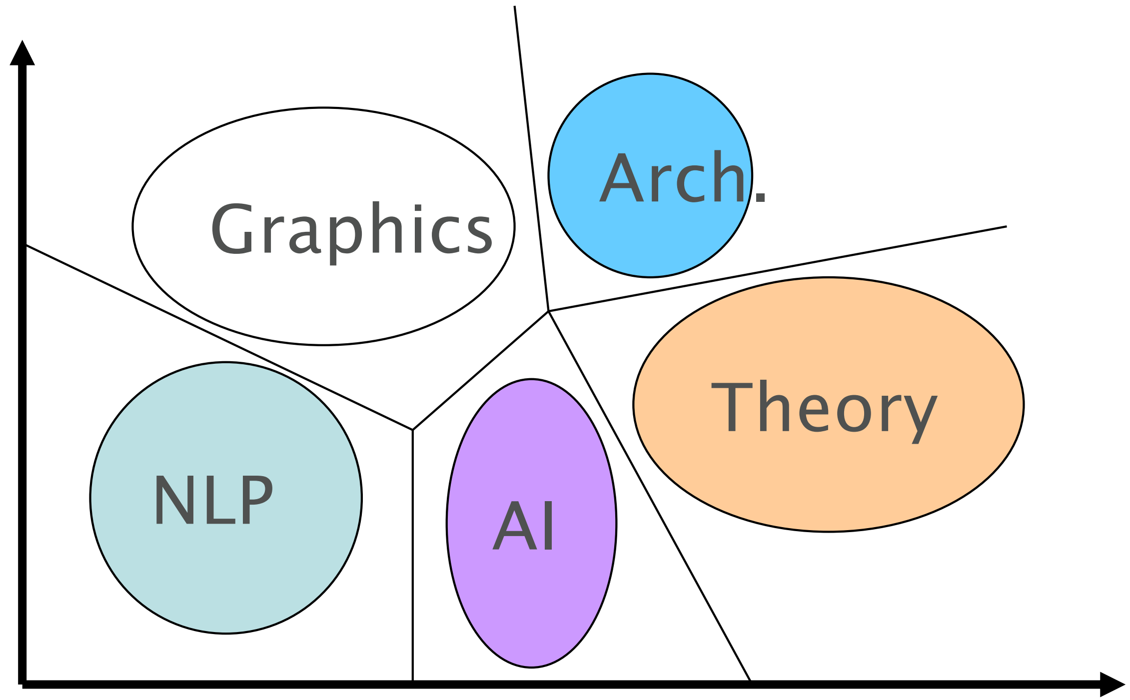
Categorization/Classification

- Given:
  - A description of an instance, $x \in X$, where X is the *instance language* or *instance space*.
    - **Issue: how to represent text documents:**
    - **Bag-Of-Words**
  - A fixed set of categories: $C = \{c_1, c_2, \ldots, c_n\}$
- Determine:
  - The category of x: $c(x) \in C$, where $c(x)$ is a *categorization function* whose domain is X and whose range is C.
    - ***We want to know how to build categorization functions ("classifiers").***

Document Classification



A Graphical View of Text Classification



Examples of Text Clasifications

- LABELS=BINARY
  - “spam” / “not spam”
- LABELS=TOPICS
  - “finance” / “sports” / “asia”
- LABELS=OPINION
  - “like” / “hate” / “neutral”
- LABELS=AUTHOR
  - “Shakespeare” / “Marlowe” / “Ben Jonson”
  - The Federalist papers

Methods (1)

1. Manual classification
   - Often used by Yahoo!, Looksmart, about.com, ODP, Medline
   - Very accurate when job is done by experts
   - Consistent when the problem size and team is small
   - Difficult and expensive to scale
2. Automatic document classification
   - Hand-coded rule-based systems
     - Reuters, CIA, Verity, …
     - E.g., assign category if document contains a given Boolean combination of words
     - Commercial systems have complex query languages (everything in IR query languages + accumulators)
     - Accuracy can be high if a rule has been carefully refined over time by a subject expert also expert in the query language
     - Building and maintaining these rules is expensive

Classification Methods (2)

3. Supervised learning

- Many systems partly rely on machine learning (Autonomy, MSN, Verity, Enkata, Yahoo!, …)
  - k-Nearest Neighbors (simple, powerful)
  - *Naive Bayes (simple method)*
  - Support-vector machines (more powerful)
  - Deep Learning (most powerful)
  - … plus many other methods
- **No free lunch**: requires hand-classified training data
- But data can be built up (and refined) by amateurs

Reminder: Bayes' Rule

$$P(C,D) = P(C \mid D)P(D) = P(D \mid C)P(C)$$

$$\boxed{P(C \mid D) = \frac{P(D \mid C)P(C)}{P(D)}}$$

Maximum a posteriori Hypothesis

$$c_{MAP} \equiv \operatorname*{argmax}_{c \in C} P(c \mid D)$$

$$= \operatorname*{argmax}_{c \in C} \frac{P(D \mid c) P(c)}{P(D)}$$

$$= \operatorname*{argmax}_{c \in C} P(D \mid c) P(c)$$

Since P(D) is constant

Maximum Likelihood (ML) Hypothesis

If all hypotheses are a priori equally likely, we only need to consider the $P(D|c)$ term:

$$c_{ML} \equiv \underset{c \in C}{\operatorname{argmax}} P(D \mid c)$$

Naive Bayes Classifiers

Task: Classify a new instance *D* based on a tuple of attribute values $D = \langle x_1, x_2, \ldots, x_n \rangle$ into one of the classes $c_j \in C$

For instance, bag-of-word features when classifying text documents

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}}\, P(c \mid x_1, x_2, \ldots, x_n)$$

$$= \underset{c \in C}{\operatorname{argmax}} \frac{P(x_1, x_2, \ldots, x_n \mid c) P(c)}{P(x_1, x_2, \ldots, x_n)}$$

$$= \underset{c \in C}{\operatorname{argmax}}\, P(x_1, x_2, \ldots, x_n \mid c) P(c)$$

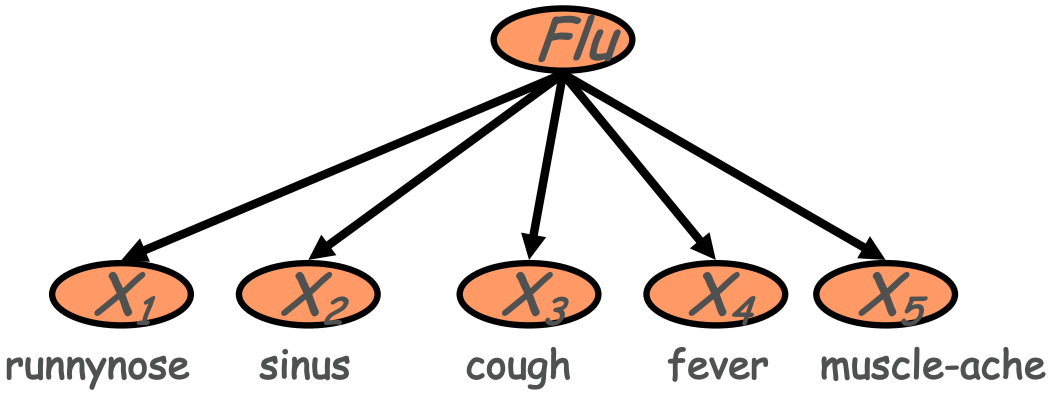
Naïve Bayes (NB) Classifier: Assumption

- $P(c_j)$
  - Can be estimated from the frequency of classes in the training examples.
- $P(x_1,x_2,\ldots,x_n|c_j)$
  - $O(|X|^n \cdot |C|)$ parameters
  - Could only be estimated if a very, very large number of training examples was available.

NB Conditional Independence Assumption to the rescue!

- ***Assume* that the probability of observing the attributes is equal to the product of the individual probabilities $P(x_i|c_j)$.**

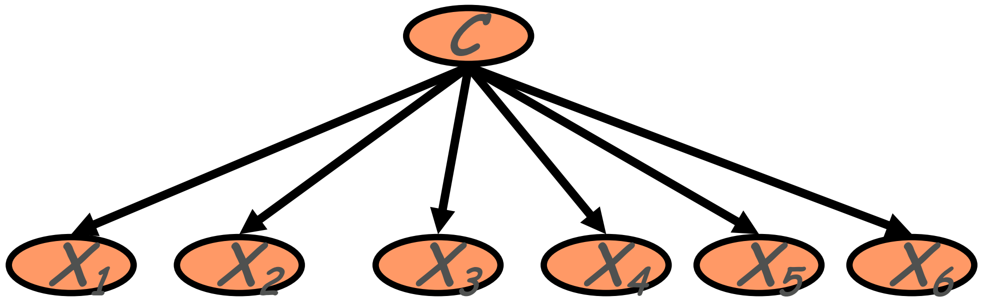
The Naïve Bayes Classifier



- Conditional Independence Assumption: features are independent of each other given the class:

$$P(X_1,\ldots,X_5 \mid C) = P(X_1 \mid C) \bullet P(X_2 \mid C) \bullet \cdots \bullet P(X_5 \mid C)$$

Learning the Model

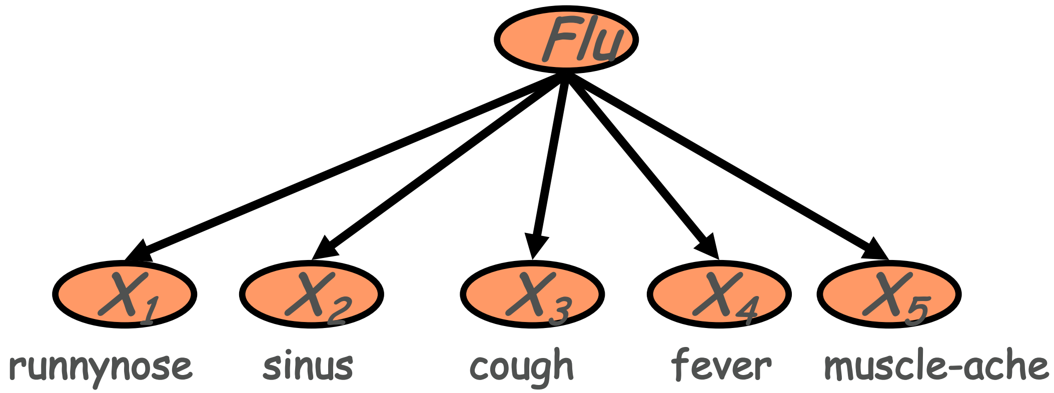


- First attempt: maximum likelihood estimates
  - simply use the frequencies in the data

$$\hat{P}(c_j) = \frac{N(C = c_j)}{N}$$

$$\hat{P}(x_i \mid c_j) = \frac{N(X_i = x_i, C = c_j)}{N(C = c_j)}$$

Problem with Max Likelihood



- What if we have seen no training cases where patient had flu and muscle aches?

$$P(X_1,\ldots,X_5 \mid C) = P(X_1 \mid C) \bullet P(X_2 \mid C) \bullet \cdots \bullet P(X_5 \mid C)$$

$$\hat{P}(X_5 = t \mid C = flu) = \frac{N(X_5 = t, C = flu)}{N(C = flu)} = 0$$

- Zero probabilities cannot be conditioned away, no matter the other evidence!

$$\ell = \arg\max_c \hat{P}(c) \prod_i \hat{P}(x_i \mid c) = 0$$

Smoothing to Avoid Overfitting

$$\hat{P}(x_i \mid c_j) = \frac{N(X_i = x_i, C = c_j) + 1}{N(C = c_j) + k}$$

# of values of X_i

- Somewhat more subtle version

overall fraction in data where X_i=$x_{i,k}$

$$\hat{P}(x_{i,k} \mid c_j) = \frac{N(X_i = x_{i,k}, C = c_j) + mp_{i,k}}{N(C = c_j) + m}$$

extent of "smoothing"

Wrap-up: (1) Naïve Bayes Learning

- From training corpus, extract *Vocabulary*
- Calculate required $P(c_j)$ and $P(x_k \mid c_j)$ terms
  - **For each c_j in C do**

    $docs_j \leftarrow$ subset of documents for which the target class is c_j

$$P(c_j) \leftarrow \frac{|docs_j|}{|\text{total \# documents}|}$$

- **For each word x_k in *Vocabulary***

  $n_k \leftarrow$ number of occurrences of x_k in all $docs_j$

$$P(x_k \mid c_j) \leftarrow \frac{n_k + 1}{|docs_j| + |Vocabulary|}$$

Wrap-up: (2) Naïve Bayes Classifying

- Return c_{NB}, where

$$c_{NB} = \operatorname*{argmax}_{c_j \in C} P(c_j) \prod_{i \in vocabulary} P(x_i \mid c_j)$$

- To avoid small numbers, use the sum of logs

SPAMCOP

- Uses a Naïve Bayes classifier
- M is spam if P(Spam|M) > P(NonSpam|M)
- Method
  - Tokenize message using Porter Stemmer
  - Estimate P(W|C) using m-estimate (a form of smoothing)
  - Remove words that do not satisfy certain conditions
  - **Train: 160 spams, 466 non-spams**
  - **Test: 277 spams, 346 non-spams**
- Results: ERROR RATE of 4.33%
  - Worse results using trigrams

Take-Away Message:
Naive Bayes is Not So Naive

- Naïve Bayes: First and Second place in KDD-CUP 97 competition, among 16 (then) state of the art algorithms

  Goal: Financial services industry direct mail response prediction model: Predict if the recipient of mail will actually respond to the advertisement – 750,000 records.
- A good dependable baseline for text classification
  - But not the best!
- Optimal if the Independence Assumptions hold:
  - If assumed independence is correct, then it is the Bayes Optimal Classifier for problem
- Very Fast:
  - Learning with one pass over the data;
  - Testing linear in the number of attributes, and document collection size
- Low Storage requirements

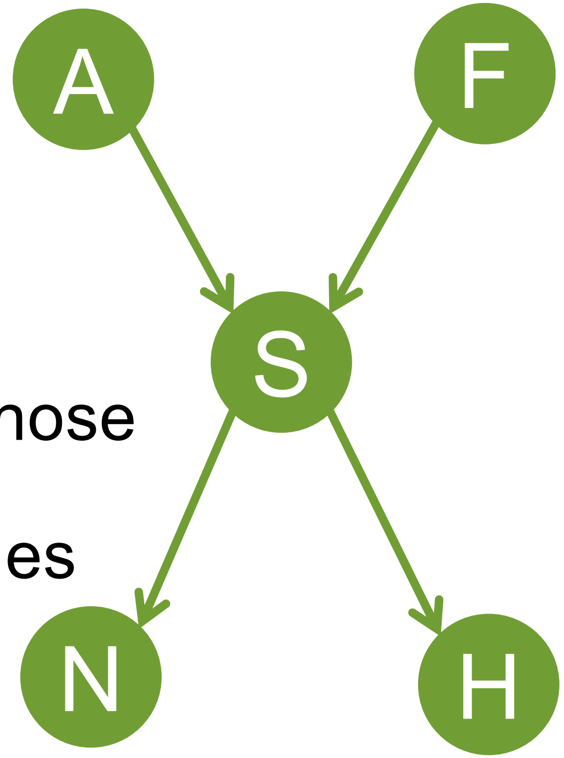
So far

- We’ve heard of Bayes nets
- We’ve played with Bayes nets
- We’ve even used them for classifying spam

- Now, we’ll learn
  - the semantics of BNs and
  - **relate them to independence assumptions encoded by the graph**

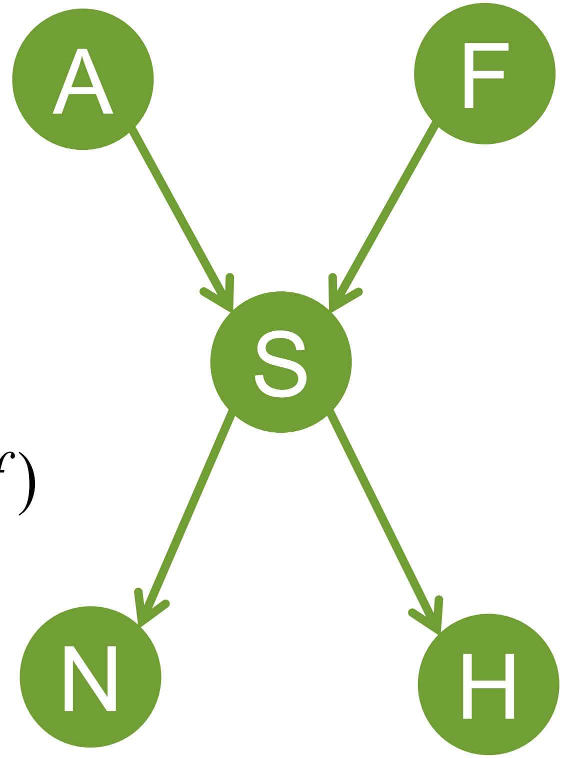
„Causal“ Structure

- 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?

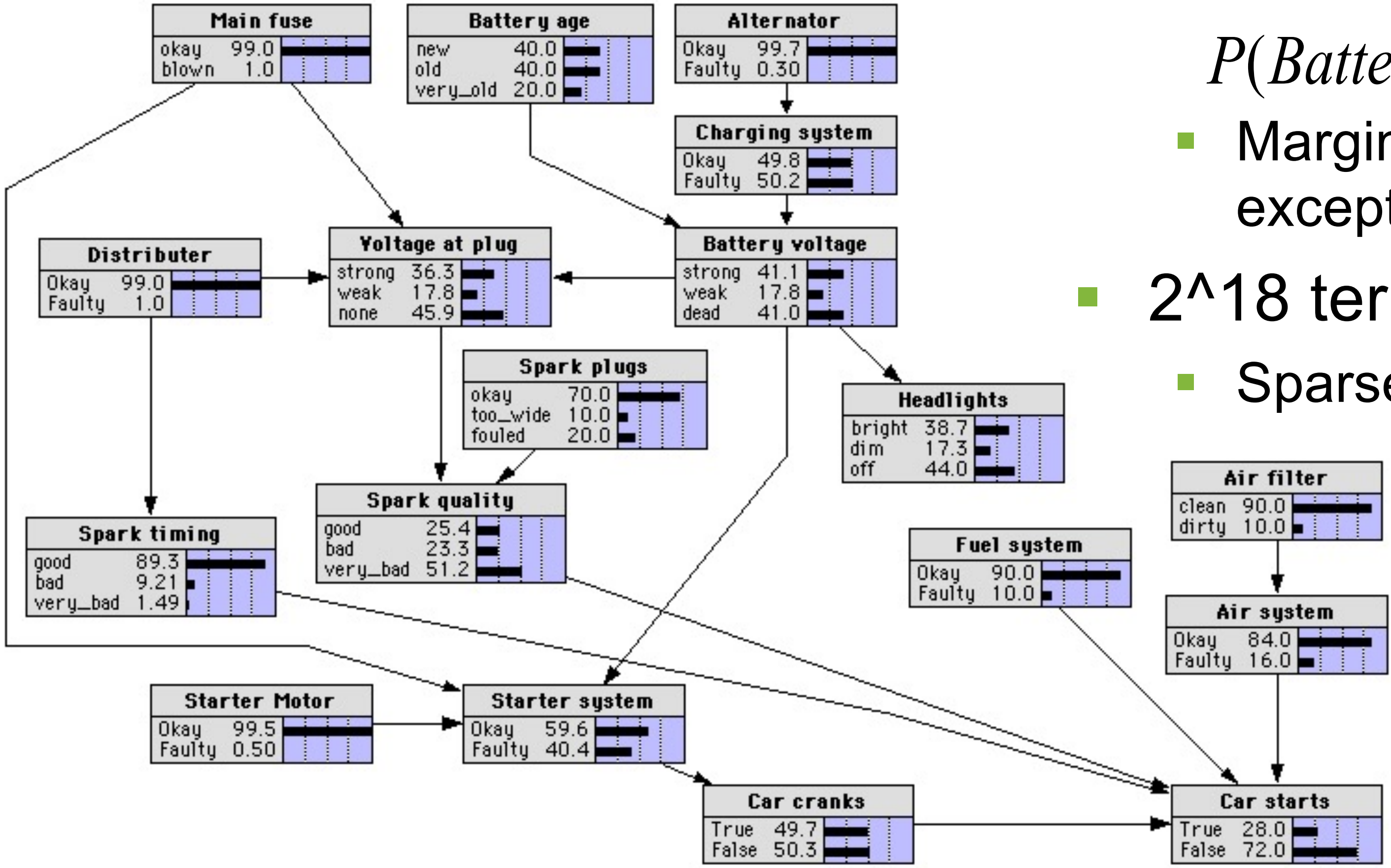


Possible Queries

- **Inference**
  - Assume H=t, N=f. What is the probability P(A=t|H=t,N=f) of an allergic reaction?
- **Most probable explanation**
  - $\max_{f,a,s} P(F=f, A=a, S=s \mid H=t, N=f)$
- **Active data collection**
  - What is next best test, i.e., variable to observe



„Car Diagnosis“ Bayesian Network



- 18 binary attributes
- Inference

  $$P(BatteryAge \mid Starts = f)$$

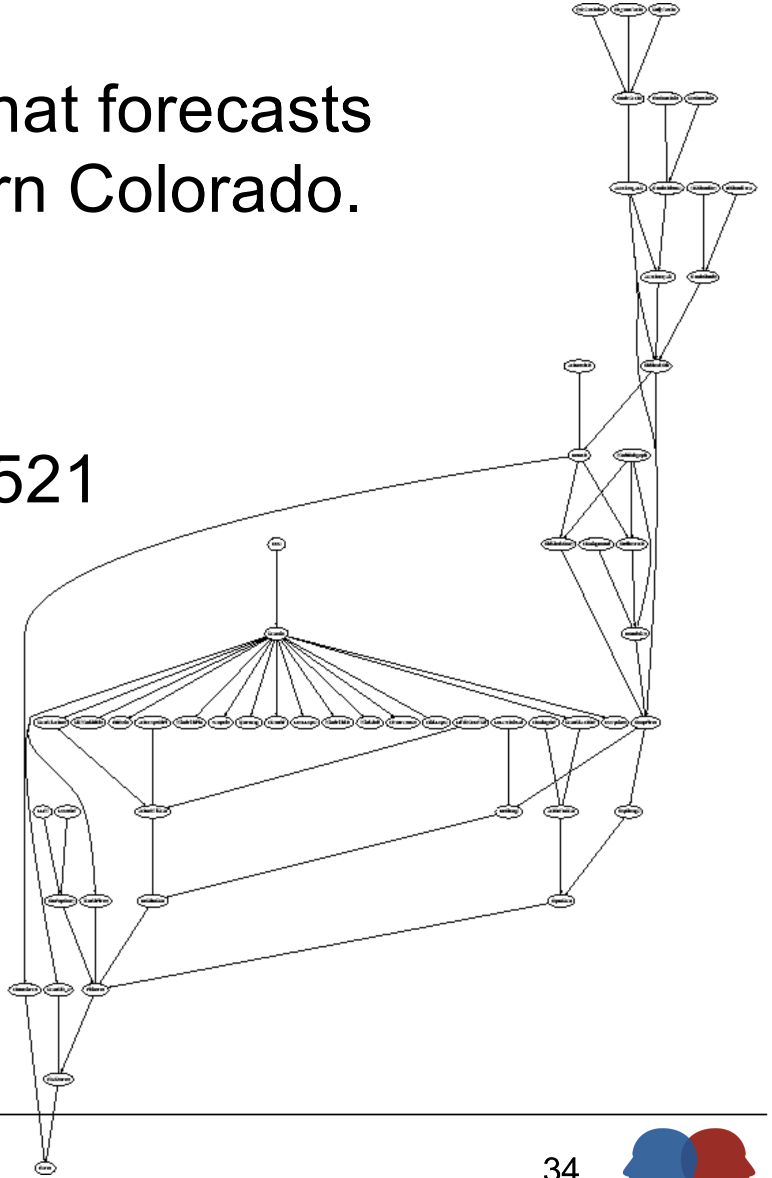
  - Marginalize all variables except BA and S
- 2^18 terms, why so fast?
  - Sparse structure !!!

Not impressed?

- Hailfinder is a normative system that forecasts severe summer hail in northeastern Colorado.
- 56 variables and 66 arcs
- More than 3^54

  = 58149737003040059690390169

  terms



Still not impressed?

- We are currently dealing with models that have far more than 300 variables (nodes)
- > 2^300 ~ 10^{90} terms
- BTW, age of the earth:

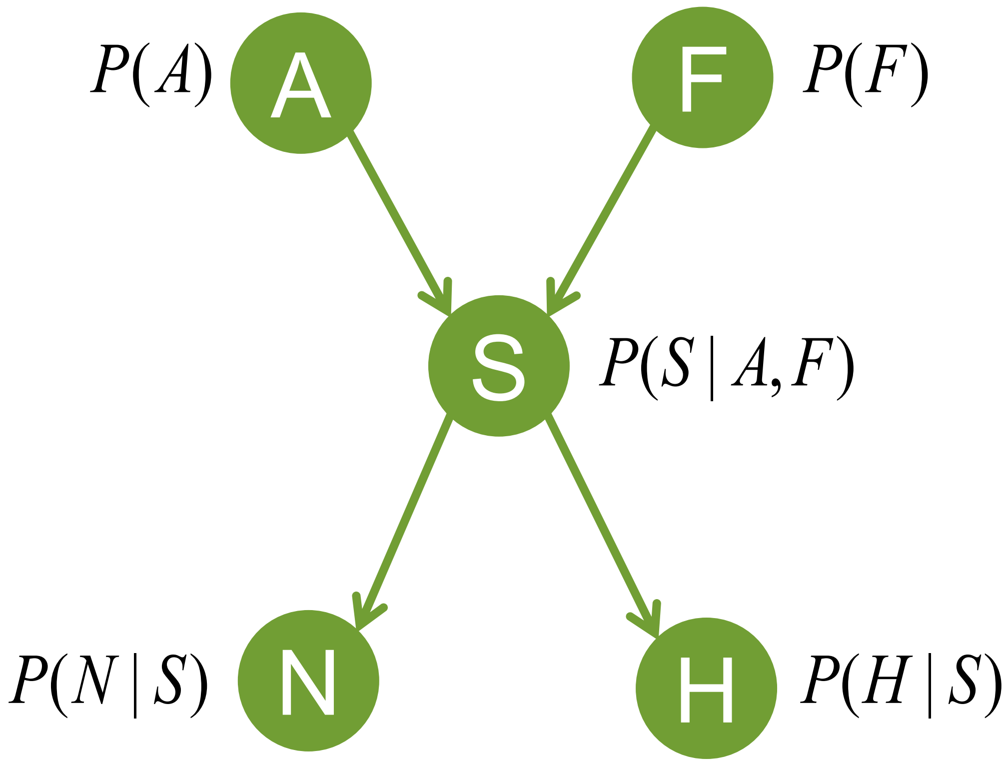
  ~ 4.54 × 10^{9} years ~ 10^{20} seconds

- **Interested in joining?**

Factored Joint Distributions Preview

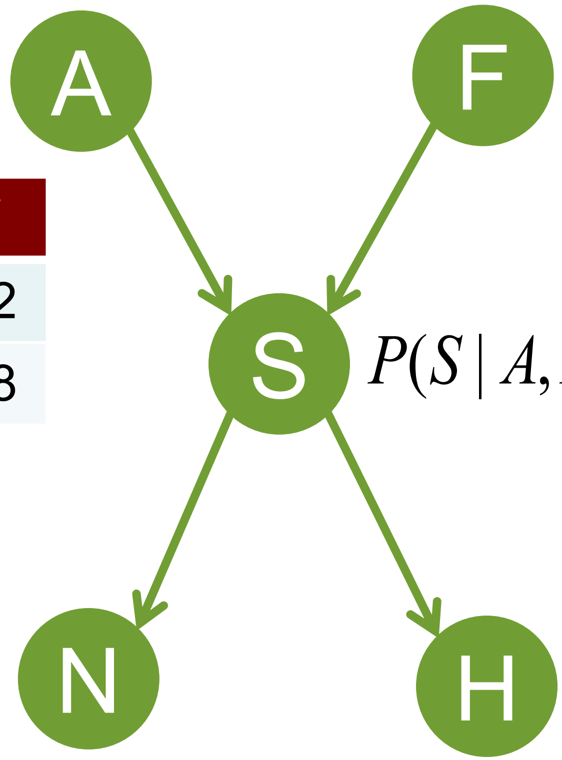
- Intuition: Nodes = Variables, Edges = Influences

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



Number of Parameters

- **Sparse structure**: compact representation for exponentially-large probability distributions
- Enumerative representation of binary variables: **2^5-1=31 parameters**



	A,F	tt	tf	ft	ff
S	t	0.9	0.7	0.8	0.2
	f	0.1	0.3	0.2	0.8

$$1+1+1+1=4$$

- **Using BNs**: **only 10 parameters**

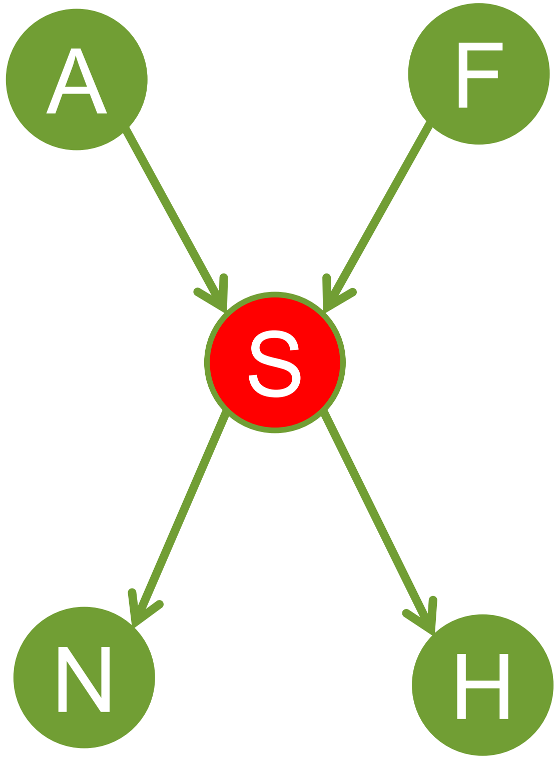
Key: Independence Assumptions

$$\neg(F \perp H)$$

$$(F \perp H \mid S)$$

$$\neg(N \perp A)$$

$$(N \perp A \mid S)$$



Knowing sinus separates symptoms from causes

Recap: (Marginal) Independence

- Flu and Allergy are marginally independent

$$\left(A \perp F\right) \qquad P(A,F) = P(A) \cdot P(F)$$

A	t	f
t	0.3	0.7

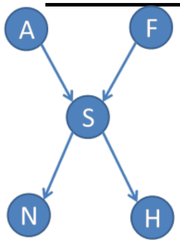
F	t	f
t	0.1	0.9

A,F	t	f
t	0.3*0.1=0.03	0.3*0.9=0.27
f	0.7*0.1=0.07	0.7*0.9=0.63

- General case:

$$\forall \text{ subsets } \mathbf{X}, \mathbf{Y} \subseteq \{X_1, \ldots, X_n\} : \left(\mathbf{X} \perp \mathbf{Y}\right)$$

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



Recap: (Conditional) Independence

- Flu and Headache are (not) marginally independent
- Flu and headache are independent given Sinus infection
- Generally:

$$\neg(F \perp H) \qquad P(F,H) \neq P(F)\cdot P(H)$$

iff

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

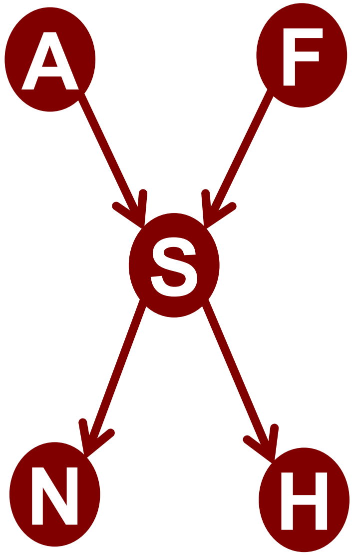
$$P(F \mid H,S) = P(F \mid S)$$

$$(X_1 \perp X_2 \ldots X_n \mid C)$$

$$P(X_1,X_2,\ldots,X_n \mid C) = P(X_1 \mid C)\cdot P(X_2,\ldots,X_n \mid C)$$

$$P(X_1 \mid X_2,\ldots,X_n,C) = P(X_1 \mid C)$$

Local Markov Assumption (Second most important slide!)



A variable X is independent of its non-descendants given its parents and only its parents **($X_i \perp \text{NonDescendants}_{X_i} \mid Pa_{X_i}$)**

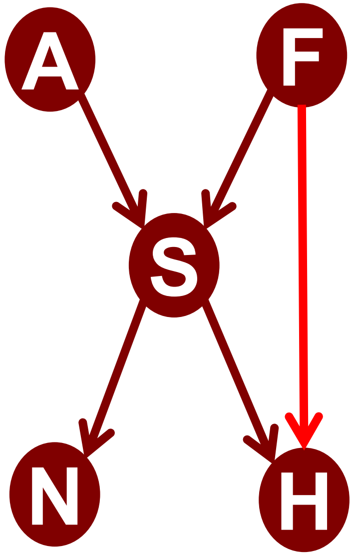
$$\text{Pa}_F = \varnothing$$
$$\text{NonDescendants}_F = \{A\}$$
$$(F \perp A)$$

$$\text{Pa}_S = \{F, A\}$$
$$\text{NonDescendants}_S = \varnothing$$
$$(S \perp ?? \mid F, A)$$

NO ASSUMPTIONS

$$\text{Pa}_N = \{S\}$$
$$\text{NonDescendants}_N = \{F, A, H\}$$
$$(N \perp \{F, A, H\} \mid S)$$

Local Markov Assumption



A variable X is independent of its non-descendants given its parents and only its parents **($X_i \perp \text{NonDescendants}_{X_i} \mid \text{Pa}_{X_i}$)**

before edge included

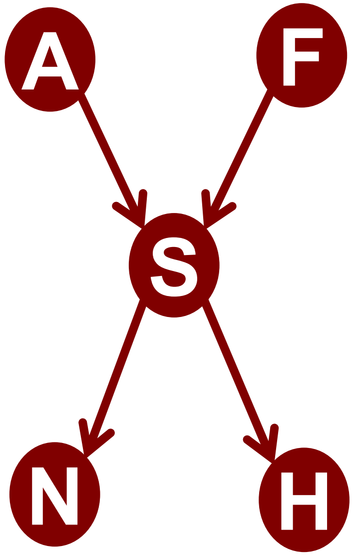
$$\text{Pa}_H = \{S\}$$
$$\text{NonDescendants}_H = \{A, F, N\}$$
$$\left(H \perp \{A, F, N\} \mid S\right)$$

after edge included

$$\text{Pa}_H = \{F, S\}$$
$$\text{NonDescendants}_S = \{A, N\}$$
$$\left(H \perp \{A, N\} \mid F, S\right)$$

Two independent events become conditionally dependent (negatively dependent) given that at least one of them occurs

Explaining Away / Berkson‘s Paradox



A variable X is independent of its non-descendants given its parents and only its parents **($X_i \perp \text{NonDescendants}_{X_i} \mid \text{Pa}_{X_i}$)**

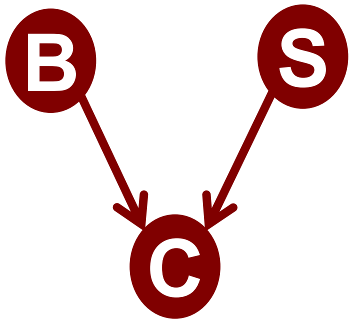
$F \perp A$

S is not a parent but a descendant

$F \perp A \mid S$ is not implied by local Markov assumption

- Two causes „compete“ to „explain“ the observed data
- **Having a flu makes you less likely to have an allergy!!**
  P(A=t) <= P(A=t|S=t,F=T) <= P(A=t|S=t)

Explaining Away / Berkson‘s Paradox

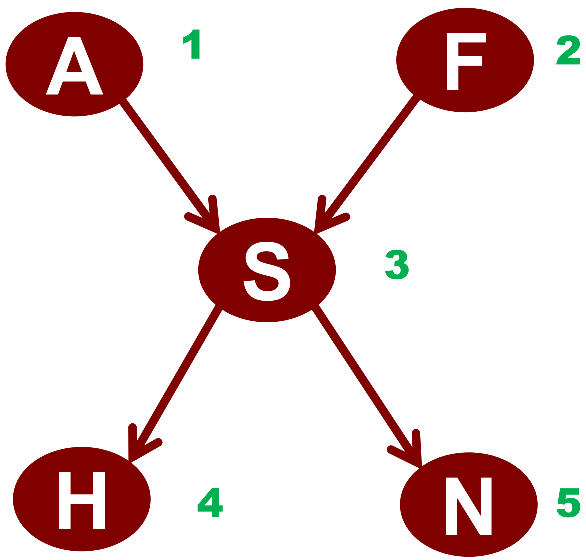


- A college which admits students who are either brainy or sporty (or both!)
- Two causes „compete“ to „explain“ the observed data

- Look at a population of college students
- **Being a brainy makes you less likely to be sportive and vice versa** because either property alone is sufficient to explain the evidence on C

$$P(S=t|C=t,B=T) <= P(S=t|C=t)$$

How to come up with the Joint Distribution



Consider **topological orders**

Now, to interpret a BN

1. Choose particular **chain rule order**
2. Apply independence assumption

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

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

$A \perp F$, $P(A)$

$P(S \mid F,A)$

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

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

We can decompose due to the local Markov assumtion

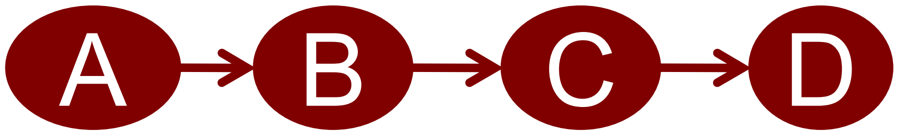
Definition: Bayesian Network
(Most important slide!)

- Set of random variables $\{X_1, \ldots, X_n\}$
- Directed acyclic graph
  - loops are ok but **no directed cycles**
- CPT with each Xi: $P(X_i \mid \mathrm{Pa}(X_i))$
- Joint distribution $P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \mathrm{Pa}(X_i))$
- **Local Makov assumption**

A variable X is independent of its non-descendants given its parents and only ist parents: **($X_i \perp \mathrm{NonDescendants}_{X_i} \mid \mathrm{Pa}_{X_i}$)**

Questions

- What distributions can be represented by a BN?
- What BNs can represent a distribution?
- What are the independence assumptions encoded?
  - Next to the ones due to the local Markov assumption



Local Markov assumption

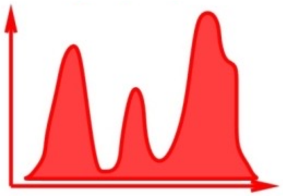
$$\{A, B\} \perp D \mid C$$

derived

$$A \perp D \mid C$$

Independencies in Real Problems

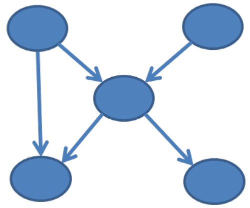
World, Data, Reality



True distribution P contains independence assertions

$$I(P)$$

Model, BN



Graph G encodes local independence assumptions

$$I_l(G)$$

Key representational assumption: $I_l(G) \subseteq I(P)$

The Representation Theorem

If conditional independencies in BN are a subset of conditional independencies in P, i.e., $I_l(G) \subseteq I(P)$

obtain

Then the joint probability distribution factorizes according to BN

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

Important: Every P has at least on BN structure G

If joint probability distribution factorizes according to BN

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

obtain

Then conditional independencies in BN are a subset of conditional independencies in P, i.e., $I_l(G) \subseteq I(P)$

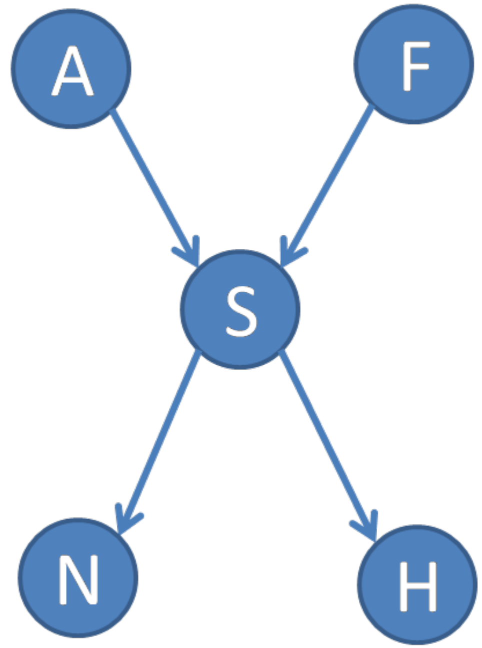
Important: Read independencies of P from BN structure G

Local Markov Assumption & I-maps

A variable X is independent of its non-descendants given its parents and only its parents **($X_i \perp NonDescendants_{X_i} \mid Pa_{X_i}$)**

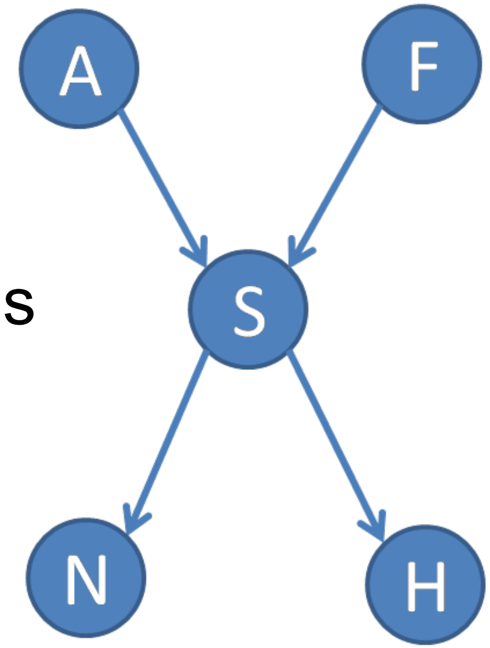
- Local independence assumption in BN structure G: $I_l(G)$
- Independence assertion of P: $I(P)$
- **BN structure G is an I-map (independence map) if:**

$$I_l(G) \subseteq I(P)$$



Factorized Distribution

- Given
  - Random variables $\{X_1, \ldots, X_n\}$
  - P distribution over the same variables
  - BN structure G over the same variables
- P factorizes according to G if



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

The Representation Theorem

TECHNISCHE UNIVERSITÄT DARMSTADT

G is I-map of P

iff

P factorizes according to G

BN Rep. Theorem: I-Map to Factorization

G is I-map of P **obtain** → P factorizes according to G

- Start with a topological ordering, wlog $X_1,\ldots,X_n$
- Apply chain rule
  $$P(X_1,\ldots,X_n) = P(X_1)P(X_2 \mid X_1)\ldots P(X_n \mid X_1,\ldots,X_{n-1})$$
- Consider $P(X_i \mid X_1,\ldots,X_{i-1})$
- We know that $\mathrm{Pa}(X_i) \subseteq \{X_1,\ldots,X_{i-1}\}$, i.e., there are no descendants of Xi in X1,…Xi-1
- Hence, due to local Markov assumption
  $$P(X_i \mid X_1,\ldots,X_{i-1}) = P(X_i \mid \mathrm{Pa}(X_i))$$

BN Rep. Theorem: I-Map to Factorization

P factorizes according to G $\xrightarrow{\text{obtain}}$ G is I-map of P

- Most likely done in the exercise session, ☺
- We have to show that the local Markov independence assumptions that hold in G also hold in P:

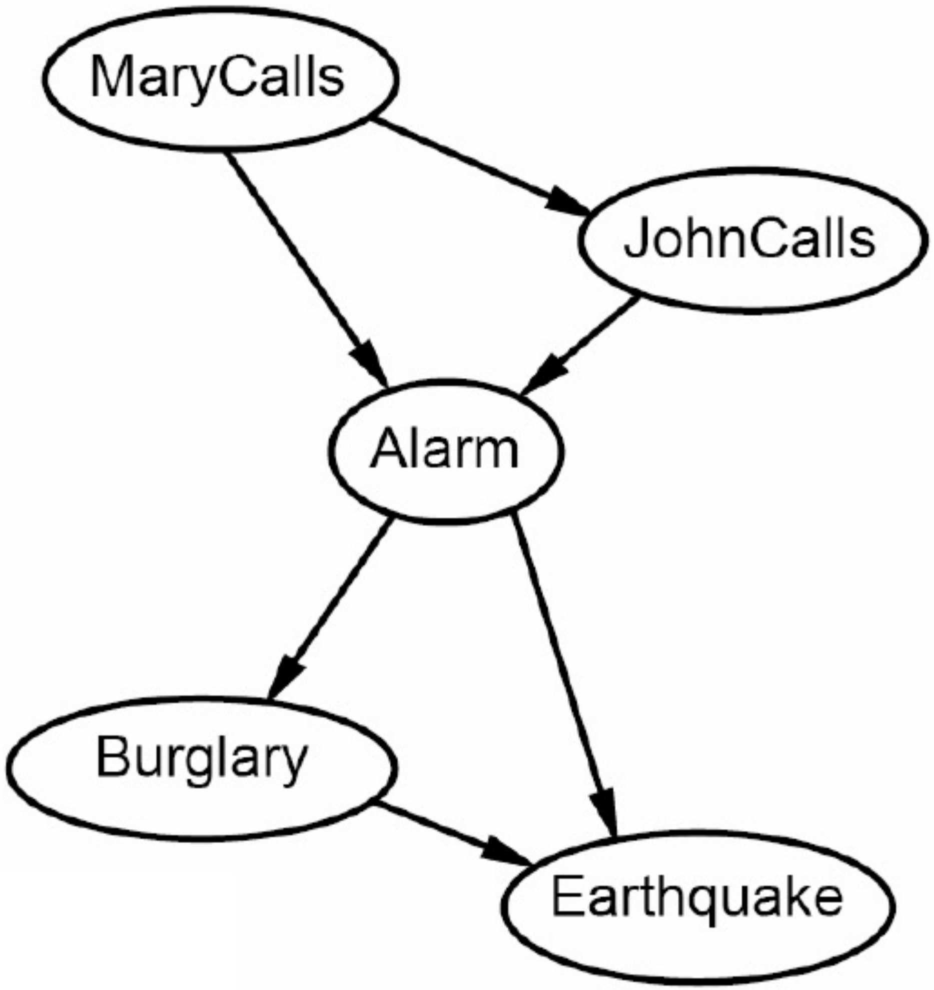
$$P\left(X_i \middle| \mathrm{NonDes}\left(X_i\right), \mathrm{Pa}\left(X_i\right)\right) = P\left(X_i \middle| \mathrm{Pa}\left(X_i\right)\right)$$

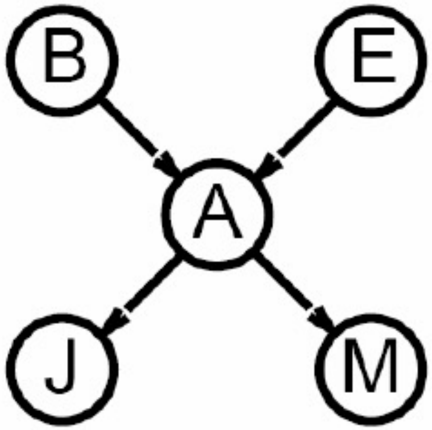
- Then apply definition of cond. prob. and simplify expressions

Coming up with a Bayesian Network

- **Given** a set of variables $\{X_1,\ldots,X_n\}$ and conditional indpendence assertions of P (estimated from data)
- **Choose** an ordering on variables, i.e., $X_1,\ldots,X_n$
- **For i = 1 to n**
  - Add Xi to the network
  - Define parents $\mathrm{Pa}(X_i)$ of Xi in graph as the minimal subset of $\{X_1,\ldots,X_{i-1}\}$ such that local Markov assumption holds
  - Define/learn CPT: $P(X_i \mid \mathrm{Pa}(X_i))$

Alternate Bayesian Networks



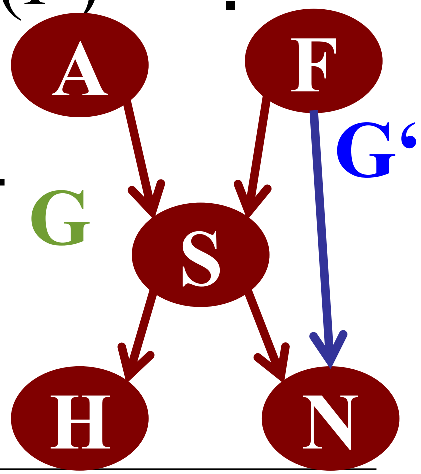


Adding edges does not hurt
(Third most important slide!)

- **Theorem:** Let G be an I-map for P, any DAG G' that includes the same directed edges as G is also an I-map for P.
  - **Corollary 1:** G' is strictly more expressive than G
  - **Corollary 2:** If G is an I-map for P, then adding edges still results in an I-map
- **Proof idea** for if $I_l(G) \subseteq I(P)$ then $I_l(G') \subseteq I(P)$.

  We show $I_l(G) \supset I_l\left(G'\right)$ by induction

  Note that we are free to set the new CPT



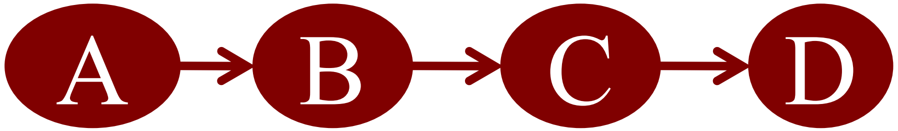
$$\boxed{\begin{aligned} &P\left(X_i \mid Pa\left(X_i\right), X'=t\right) = P\left(X_i \mid Pa\left(X_i\right), X'=f\right) \\ &= P\left(X_i \mid Pa\left(X_i\right)\right) \end{aligned}}$$

What you need to know thus far

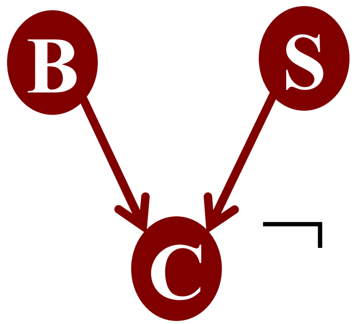
- Independence and conditional independence
- Definition of a Bayesian network
- Local Markov assumption
- The representation theorem
  - G is I-map for P iff P factorizes according to G
  - Interpretation

Independencies encoded in BN

- We said: all you need is the local Markov assumption **($X_i \perp \text{NonDescendants}_{X_i} \mid Pa_{X_i}$)**
- But then we talked about other (in)dependencies such as explaining away



$$A \perp D \mid C$$



$$B \perp S$$

$$\neg \; B \perp S \mid C$$

- So, what are the independencies encoded by a BN?
  - Only assumption is local Markov but many other can be derived using the algebra of conditional independencies!