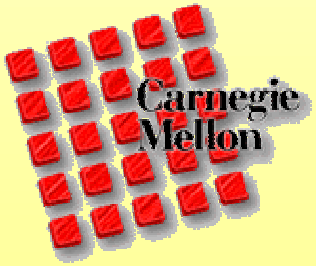
Probabilistic Model Structure from Data

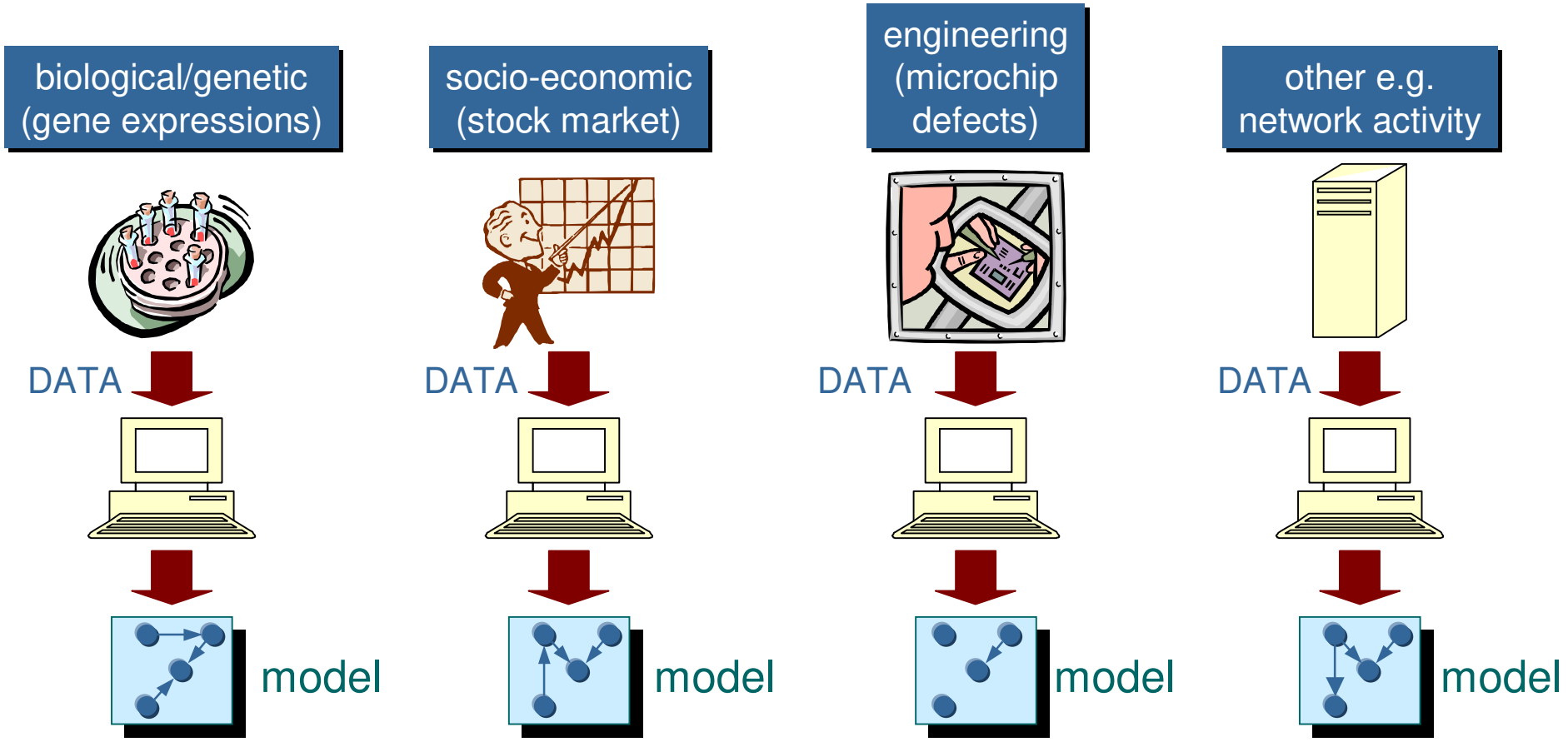
Dimitris Margaritis



School of Computer Science

The Problem I am Addressing

- I want to help in understanding experimental data



Why Do I Care?

Because I can use these models to…

1. …succinctly describe large amounts of data

2. …understand the generative process
   - for comprehension from cause-effect relations
   - for predicting future experimental outcomes
   - for exploiting by manipulation

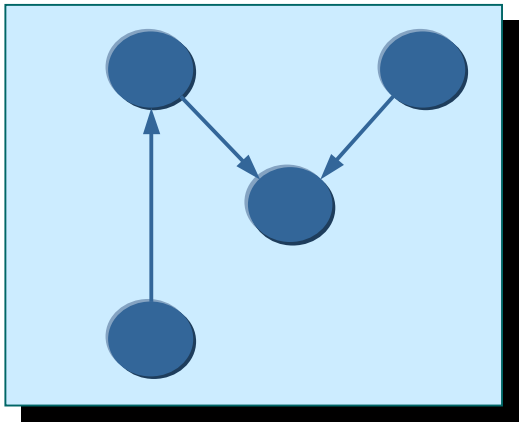
Outline

BN (Bayesian Network) model structure from data

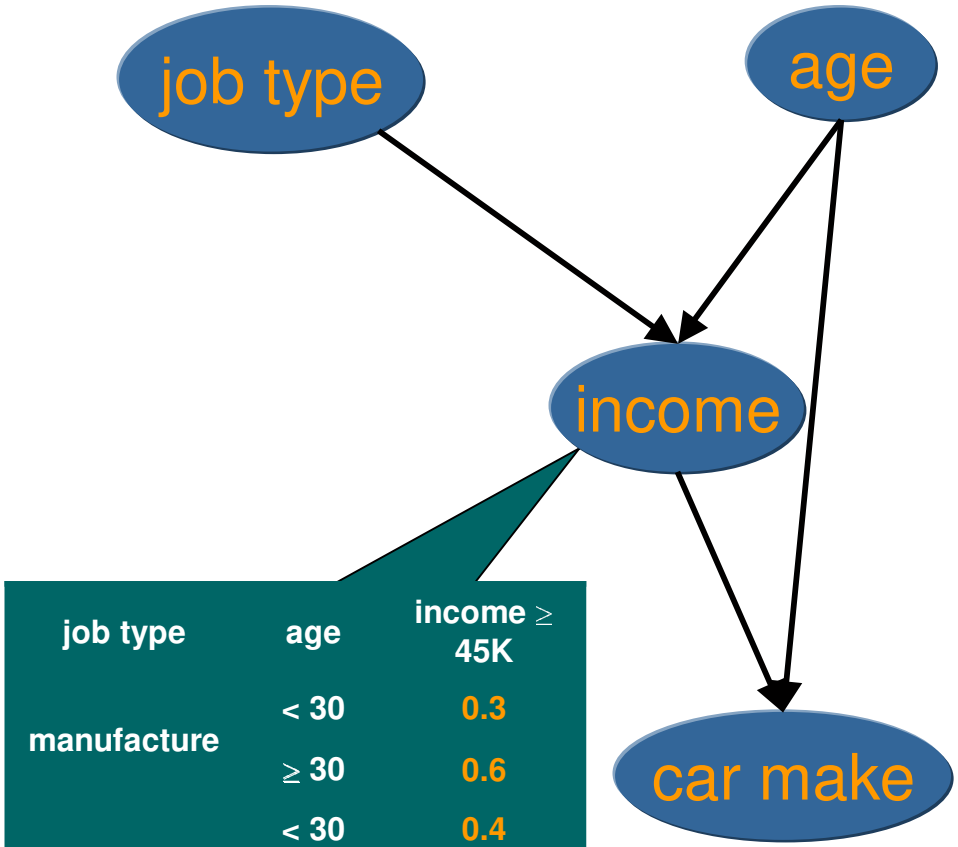
- Background
  - BN models
  - Constructing BN models
  - Constructing BN models using search
- Constructing BN models from independencies
  - GS algorithm
  - Independence test for continuous variables

What Exactly Do I Mean by "Model"?

- Model of a non-deterministic domain
  - Input: current knowledge (variable assignments)
  - Output: Pr(quantity of interest | current knowledge)
- Graphical models may show directed cause-effect relationships



Bayesian Network Models



job type	age	income ≥ 45K
manufacture	< 30	0.3
	≥ 30	0.6
corporate	< 30	0.4
	≥ 30	0.8

- Represents the joint pdf of the domain
- A BN is a directed acyclic graph (DAG)
- Each node is a variable (attribute)
- Tables are conditional probabilities (given the parents)

Causal information in edge directions

Some Advantages

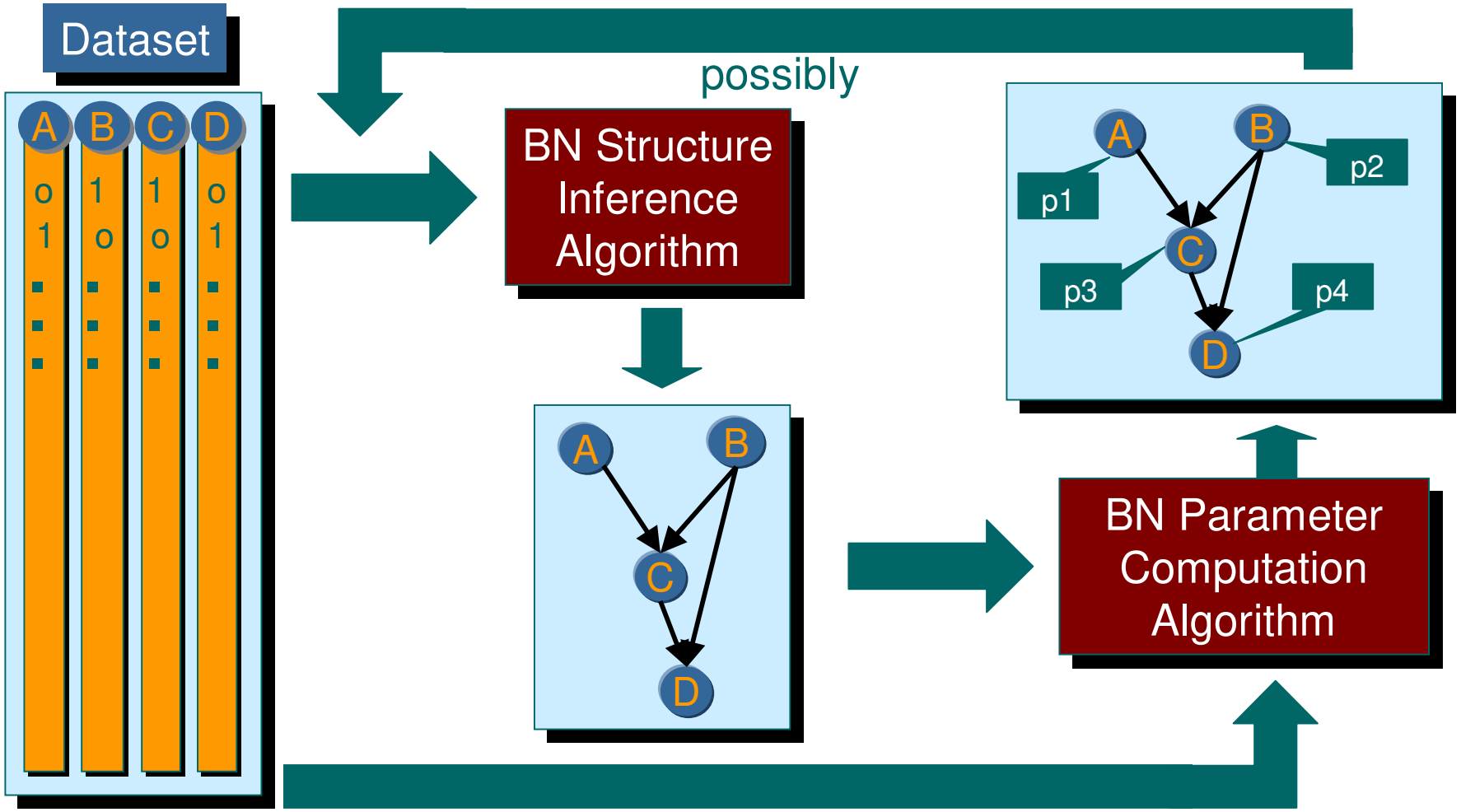
- I chose Bayesian networks to work on because they are...
  - mathematically sound (probability theory)
  - visually expressive
  - structurally meaningful (can represent generative processes)
  - complete (any generative process)

Outline

BN model structure from data

- Background
  - BN models
  - Constructing BN models
  - Constructing BN models using search
- Constructing BN models from independencies
  - GS algorithm
  - Independence test for continuous variables

Generating BN Models from Data



Major Issue

- Biggest problem is determining the model structure
- Incorrect structure can impair…
  - …model accuracy
  - …domain comprehension
  - …model usability (e.g. exponential number of parameters)

Structure Inference Algorithms

Two Broad Classes

1. Score-based
   - Fit data well
   - Do not necessarily learn the correct structure

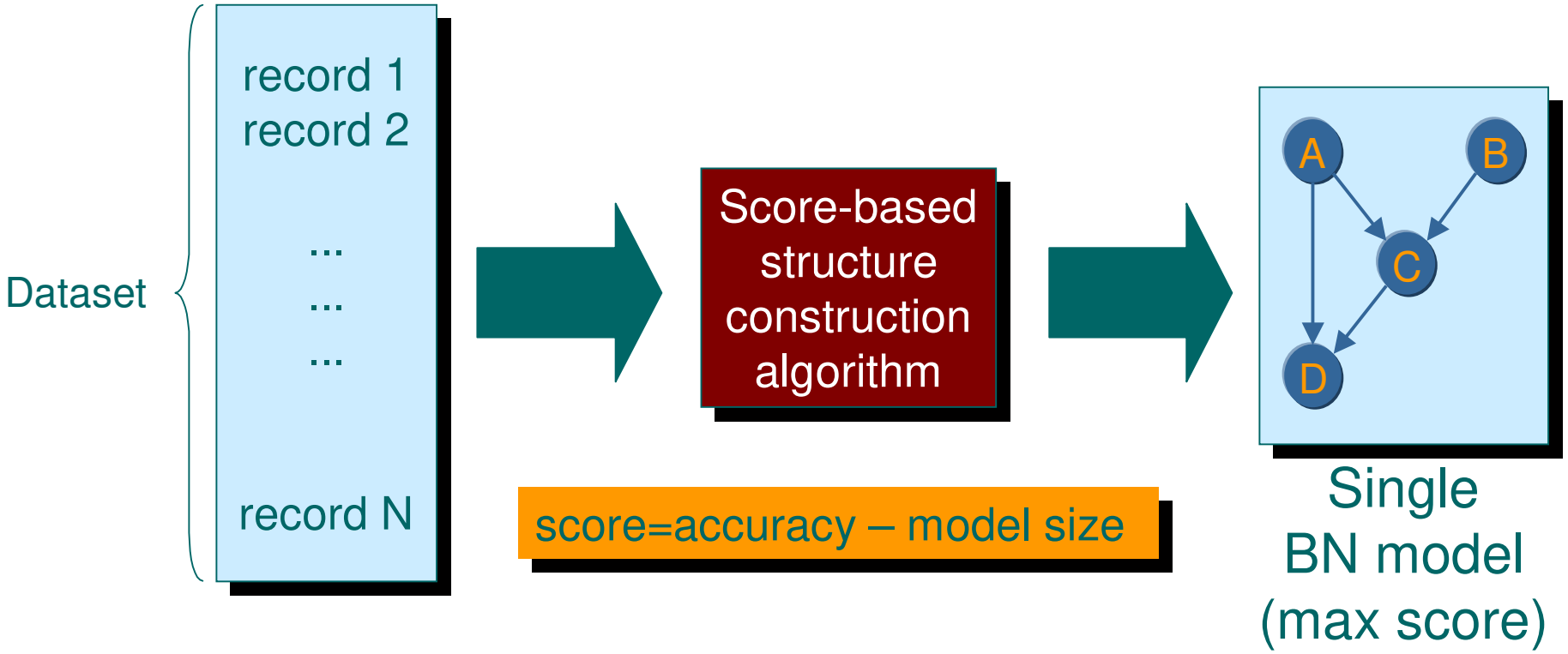
2. Independence-based
   - Guided by independencies in data
   - Learn correct structure (under assumptions)

Outline

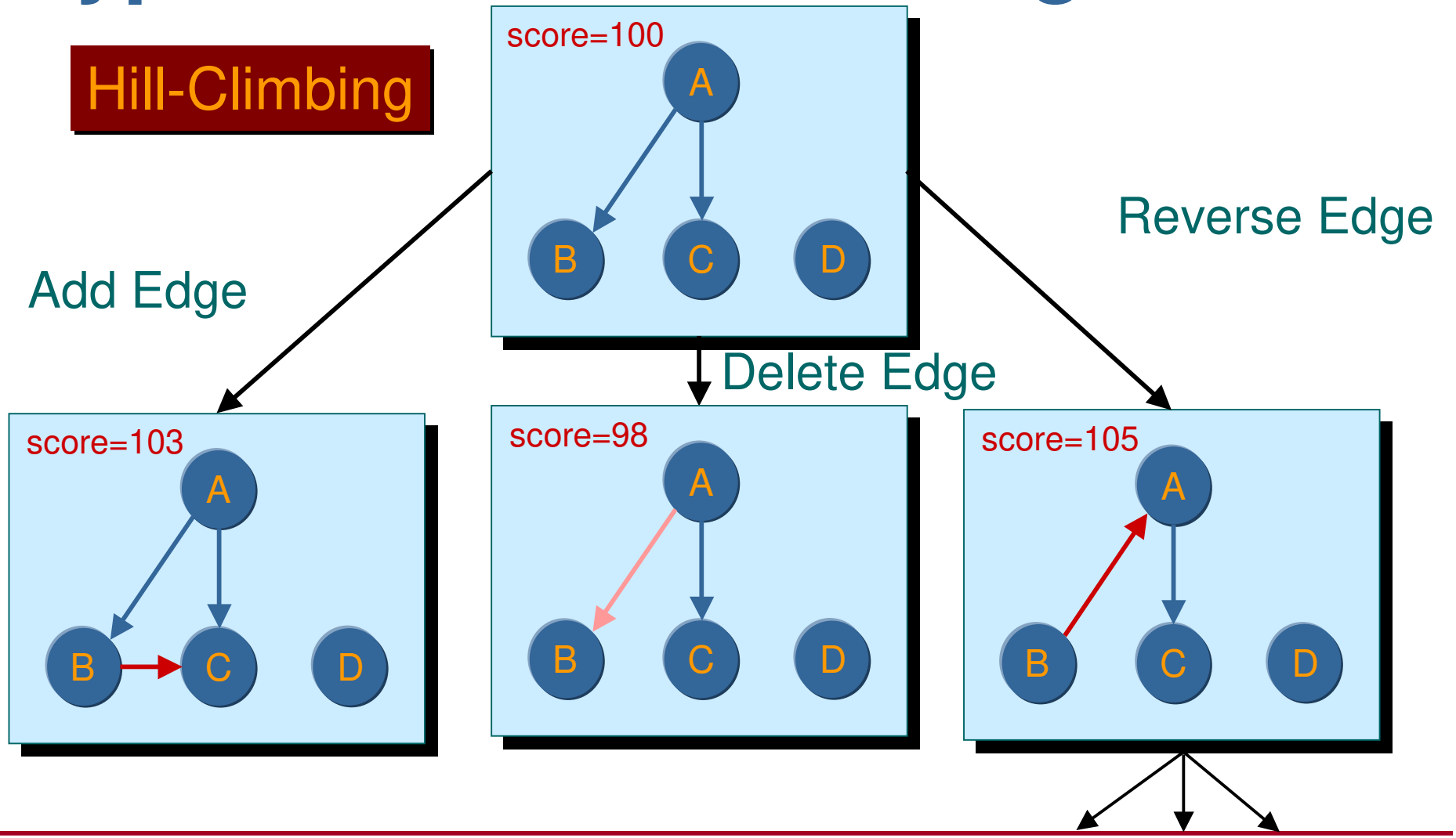
BN model structure from data

- Background
  - BN models
  - Constructing BN models
  - Constructing BN models using search
- Constructing BN models from independencies
  - GS algorithm
  - Independence test for continuous variables

Score-based BN Model Construction



Typical Score-based Algorithm



The Score Function

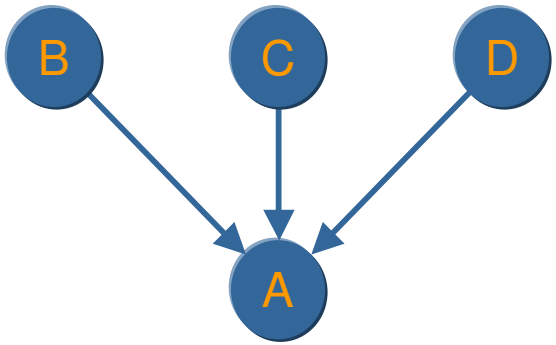
$$\text{Score}(N \text{ Records} \mid \text{BN}) = MDL = \sum_{ijk} n_{ijk} \log \frac{n_{ijk}}{n_{ij}} - \frac{M}{2} \log N$$

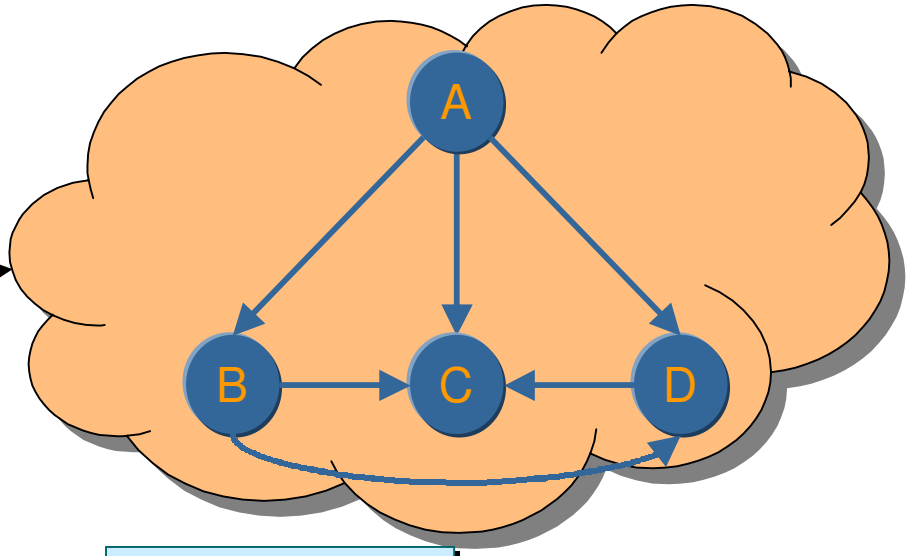
empirical data entropy (accuracy)

penalty term (model size)

Score-based Gone Wrong

True Model





A = B + C + D + noise

A = pdf(A)

B = A + noise(B)

C = A – B – D + noise(C)''

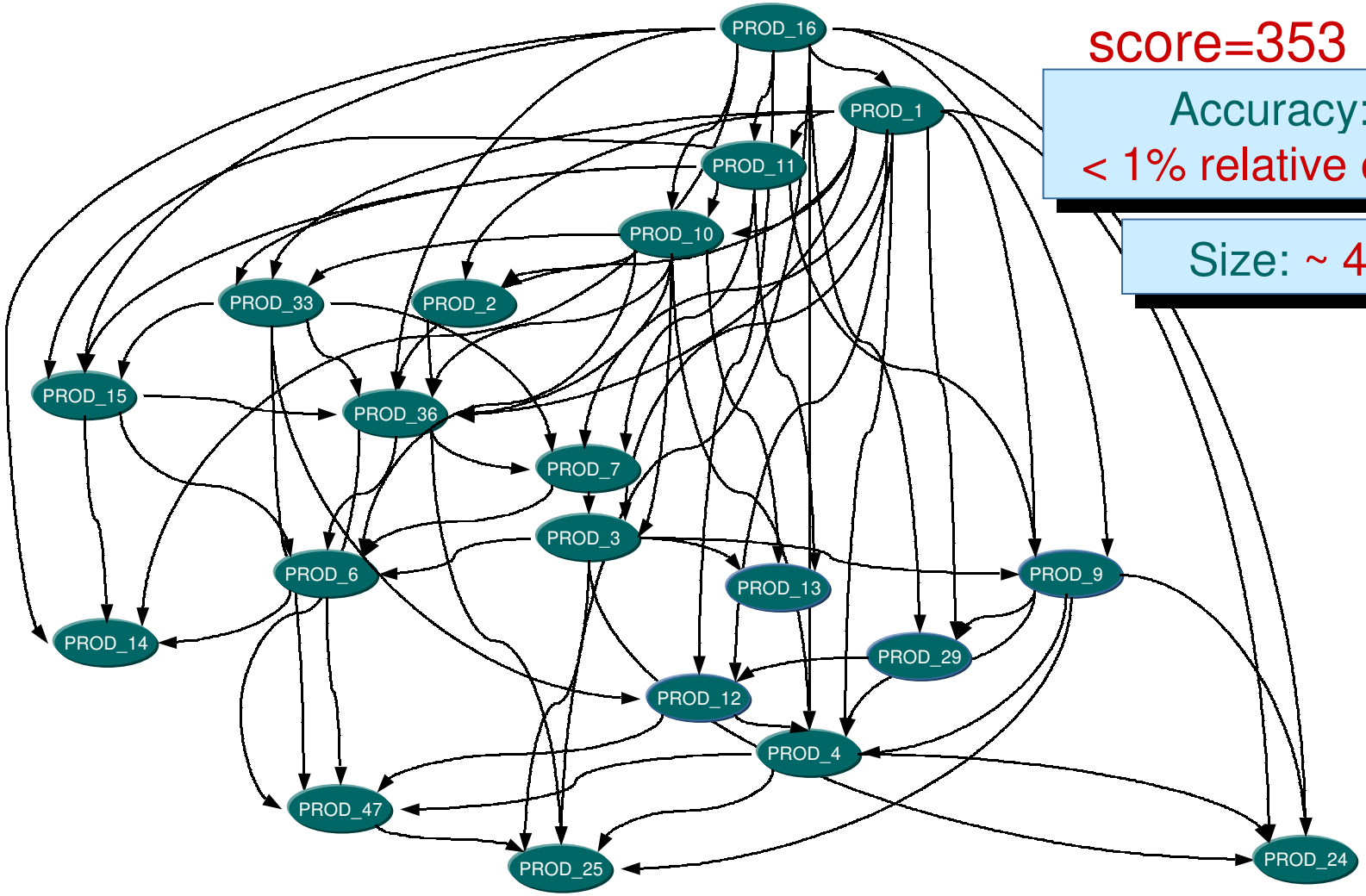
D = A – B + noise(D)'

Perfectly describes the data!
(locally max score)
But it is not the correct model

What is the Problem?

- Heuristic search is inadequate
  - Single-arc addition, deletion and reversal
- Score is inadequate (to lesser extent)
  - Predictive ability is a by-product of correct structure
  - Structure = conditional independence statements

Score-based Example: Retailer Data



score=353

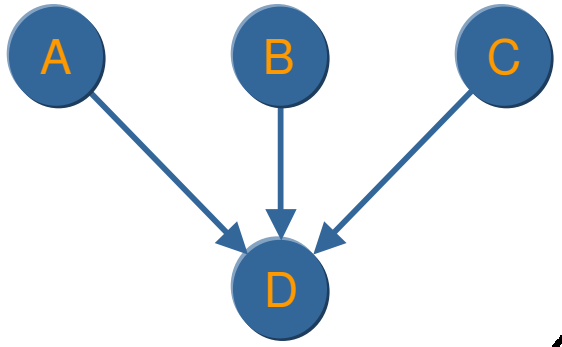
Accuracy:
< 1% relative error

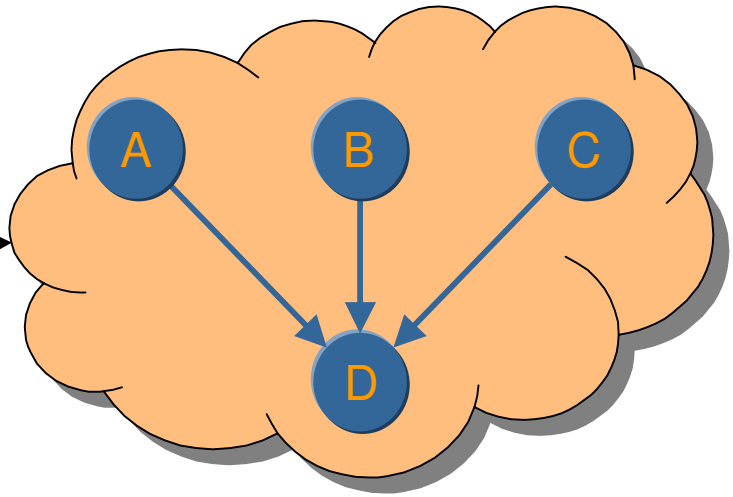
Size: ~ 4KB

In Other Words…

- …current score-based algorithms produce BNs good for summarizing large amounts of data
- …but not necessarily good for producing understandable ones
- …where understandable ≡ providing a verifiable list of independencies during the learning process

Understandable Models





D = A + B + C + noise

A independent of B
A independent of C
B independent of C
A dependent of B given D
A dependent of C given D
B dependent of C given D

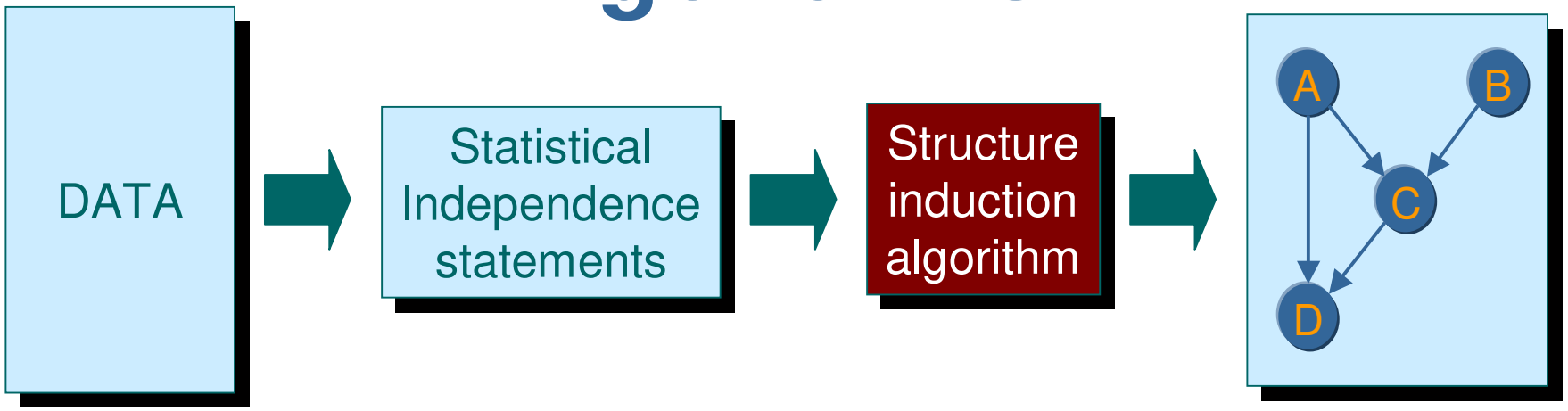
Constructed from independence statements

Outline

BN model structure from data

- Background
  - BN models
  - Constructing BN models
  - Constructing BN models using search
- Constructing BN models from independencies
  - GS algorithm
  - Independence test for continuous variables

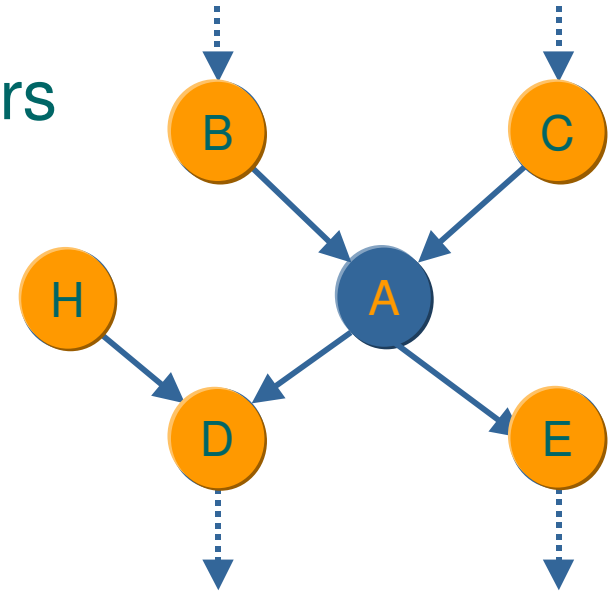
Independence-Based Algorithms



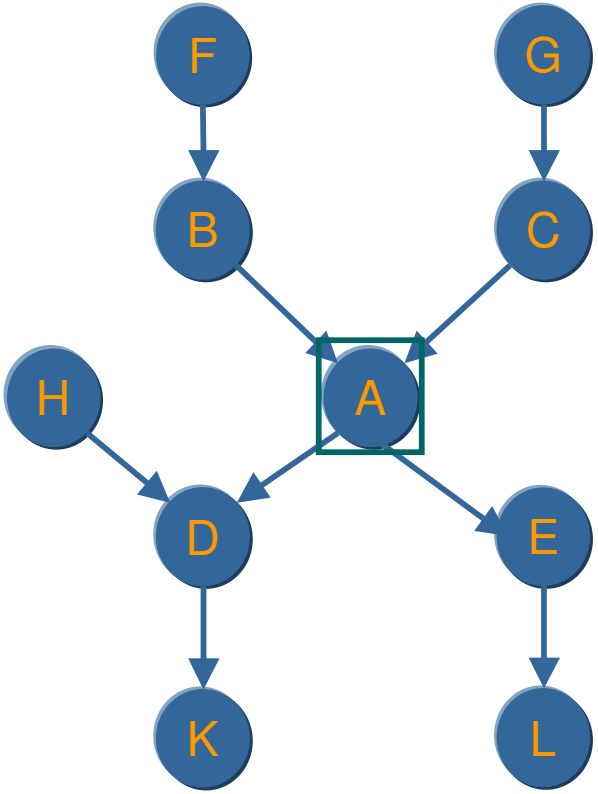
- Provably correct structure assuming
  - no errors in independence tests
  - variables linked are really dependent in dataset
  - no unobserved causes exist

GS algorithm
[Margaritis and Thrun, NIPS 1999]

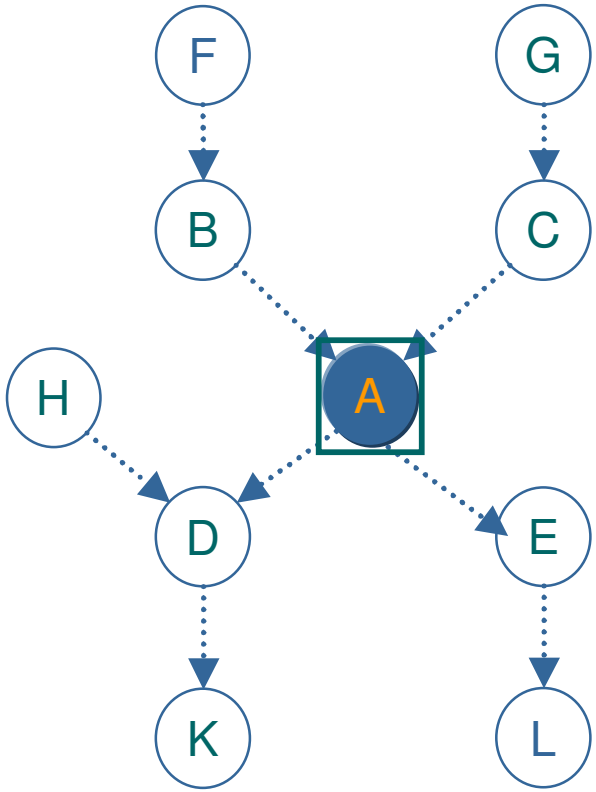
- For every variable
  1. Determines "shielding variables" (minimum Markov blanket)
  2. Eliminates non-direct members
  3. Fills-in directions
- Steps 2 and 3 condition on subsets of min Markov blankets only



Example



Initially No Arcs



Growing phase

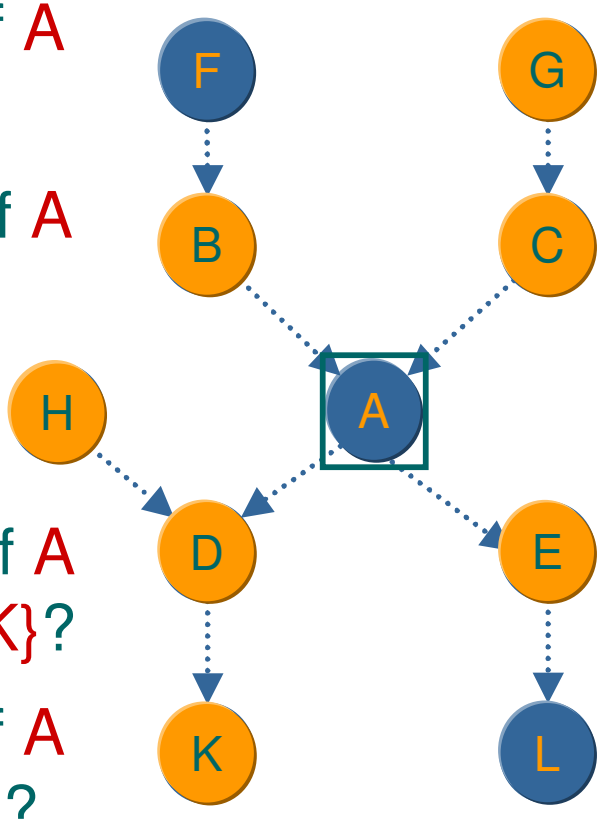
F dependent of A given {B}?

B dependent of A given {}?

H dependent of A given {B,G,C,K,D}?

D dependent of A given {B,G,C,K}?

K dependent of A given {B,G,C}?



G dependent of A given {B}?

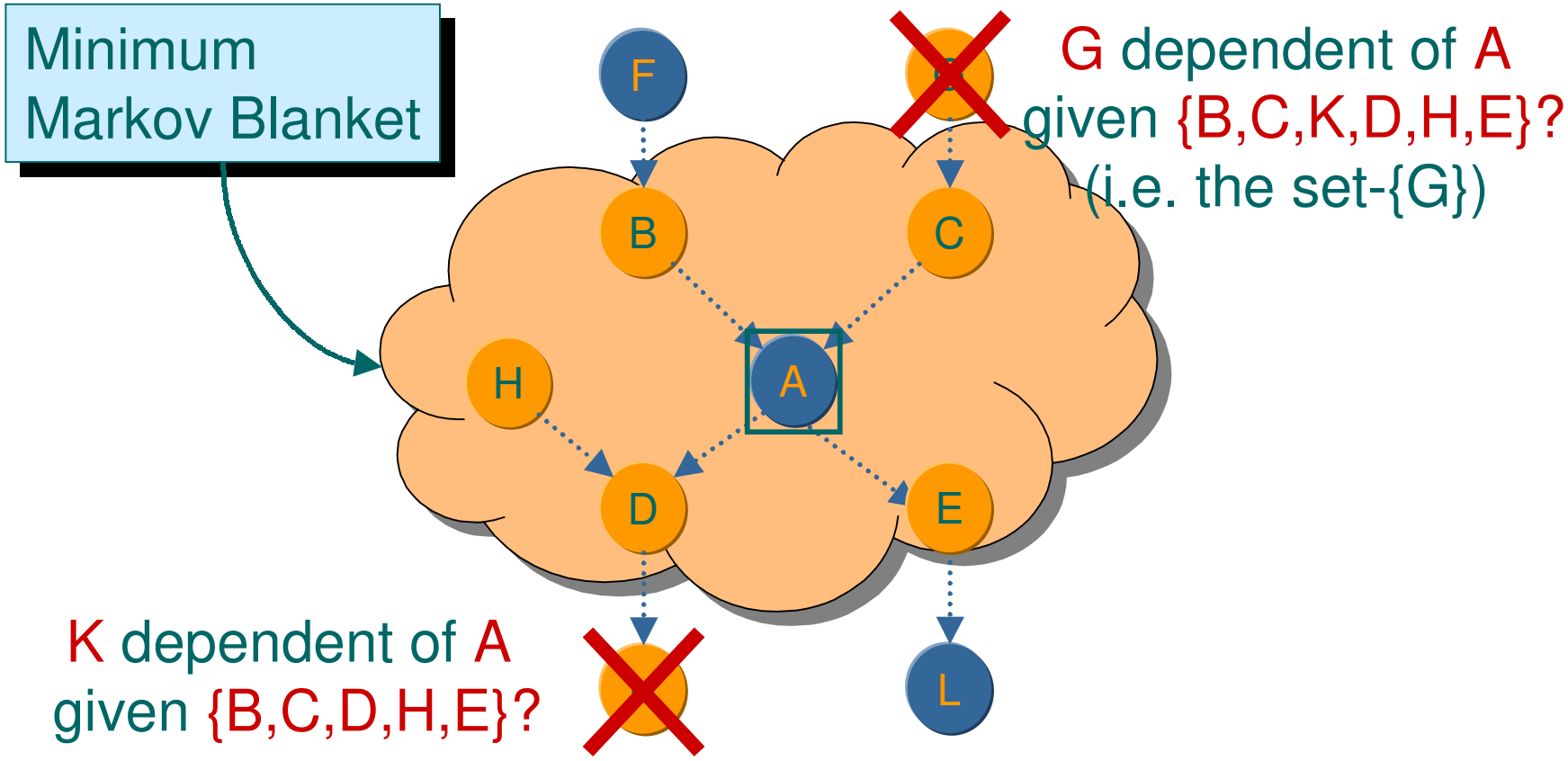
C dependent of A given {B,G}?

E dependent of A given {B,G,C,K,D,H}?

L dependent of A given {B,G,C,K,D,H,E}?

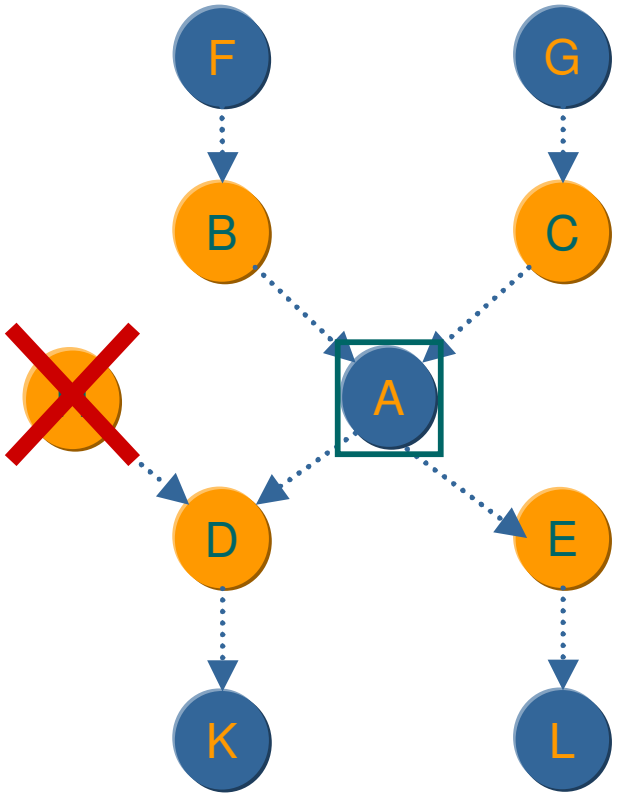
Markov blanket of A = {B,G,C,K,D,H,E}

Shrinking phase



Markov blanket of A = {B,C,D,H,E}

Direct Neighbors

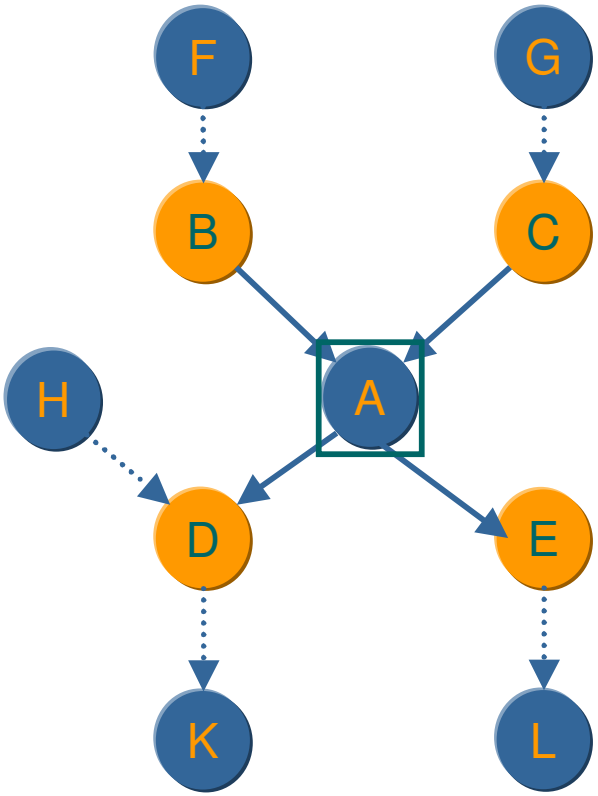


H dependent on A?

All tests conditional on a subset of a Markov blanket only!

Direct neighbors of A = {B,C,D,E}

Directions of Neighbors



All tests conditional on a subset of a Markov blanket only!

Determined by another set of tests

GS Advantages

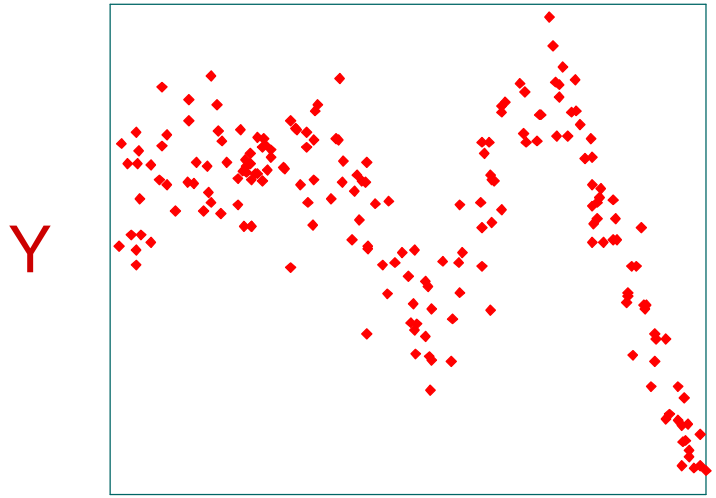
- Conditioning set size the biggest problem with an independence test
  - Large conditioning sets suffer from:
    - exponential time to execute
    - data sparsity
  - The GS algorithm:
    - First recovers the minimum Markov Blanket
    - Then uses subsets of Markov blanket only in tests

Outline

BN model structure from data

- Background
  - BN models
  - Constructing BN models
  - Constructing BN models using search
- Constructing BN models from independencies
  - GS algorithm
  - Independence test for continuous variables

Independence Tests for Continuous Variables



Are X and Y dependent? { YES / NO } (with certainty…)

Existing Non-parametric Tests

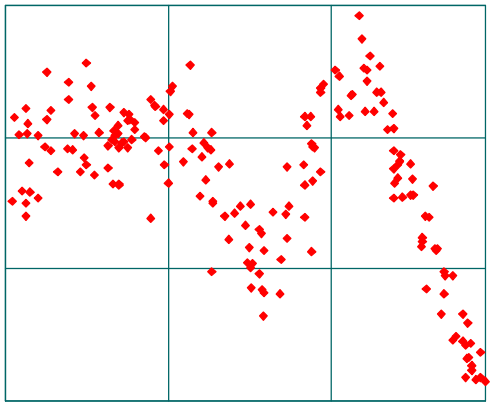
- Easy for categorical variables
  - chi-square
  - Fisher's exact test
- Continuous/ordinal case is difficult
  - Spearman's rank correlation
  - Kendall's tau
  - ...

Test monotonicity only

  - Unsolved in generality

Our Approach
[Margaritis and Thrun UAI 2001]

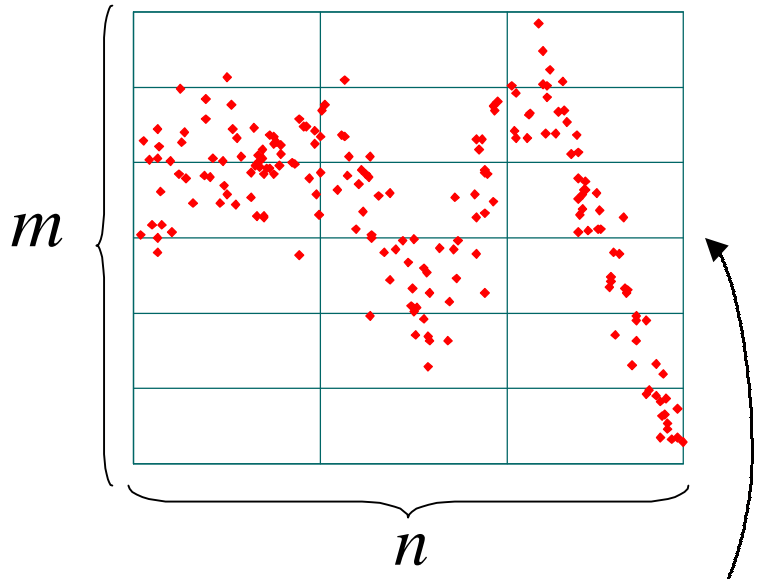
- Addresses the general unconditional non-parametric test
- Discretizes the continuous data to many resolutions (histograms)
  - 2x2
  - 3x3
  - …



- Determines dependence using these histograms

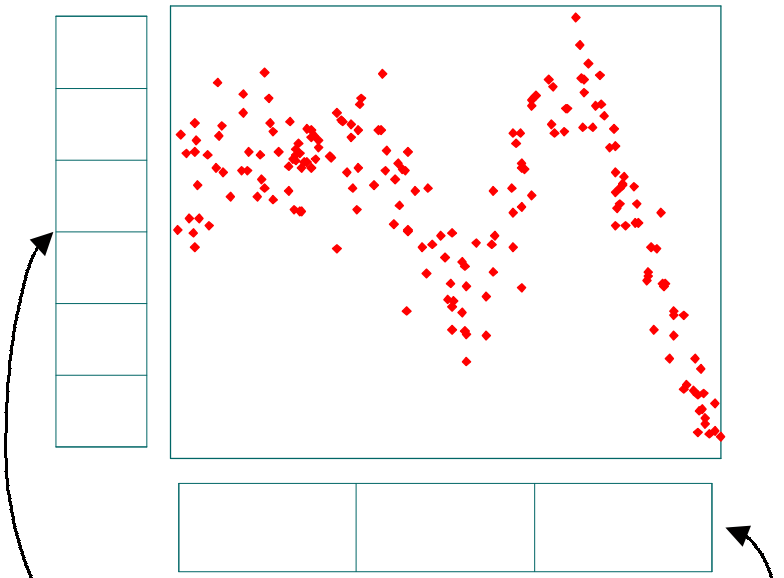
Two model classes

- I compare them on how they fit the data



$n \times m$ parameters

dependent model



$n+m$ parameters

independent model

Model Comparison Metric

- I use the data likelihood Pr(data | model) given each model class

Likelihood ratio: $R = \frac{\Pr(\text{data}|\text{indep.model})}{\Pr(\text{data}|\text{dep.model})}$

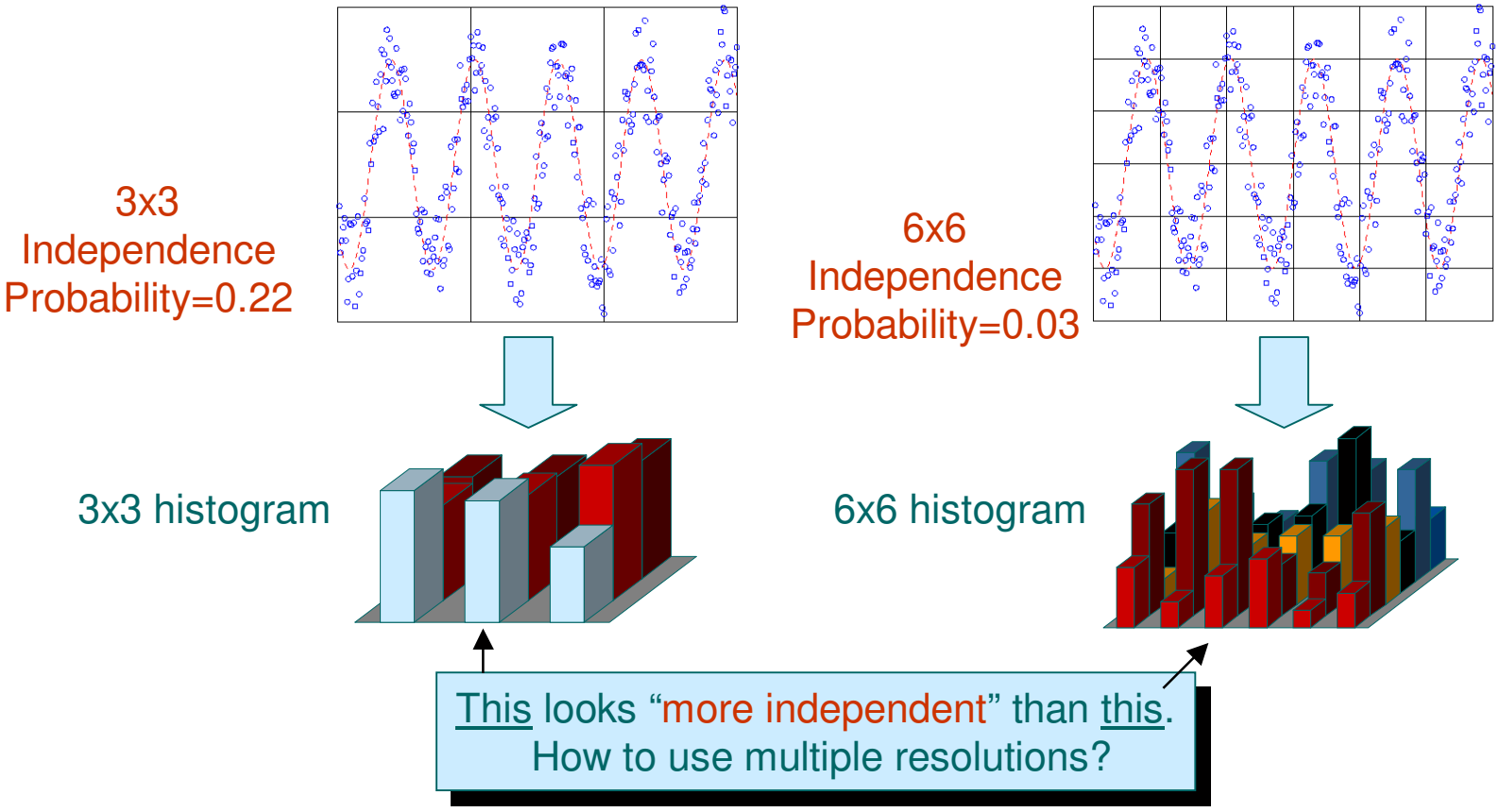
Posterior prob. of independence:

$$\Pr(\text{indep.model} \mid \text{data}) = \frac{1}{1 + \frac{\Pr(\text{indep.model})}{1 - \Pr(\text{indep.model})} R}$$

Two Major Issues

1. At what resolution do we test for dependence?
2. Discrete test ignores the natural ordering in continuous data

Why Resolution Matters



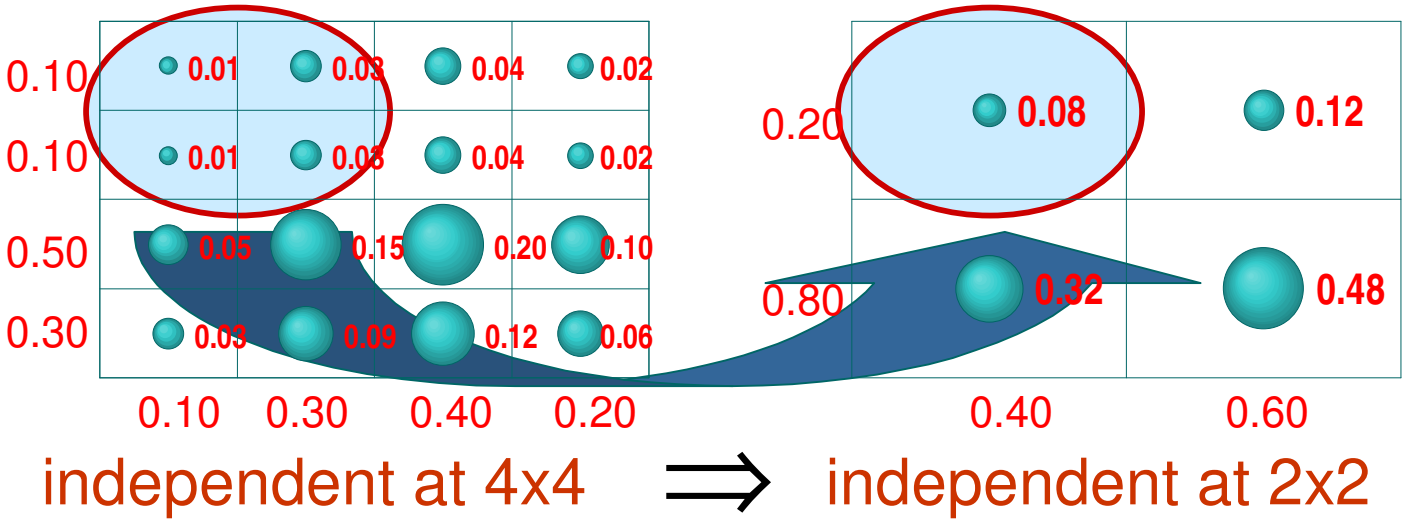
Combining Multiple Resolutions

How? Average? MAX? MIN?

One observation makes our life easier:

IF the data are independent at *2k* x *2k*
THEN they are independent at *k* x *k*

E.g.



independent at 4x4 $\Rightarrow$ independent at 2x2

What Does That Mean?

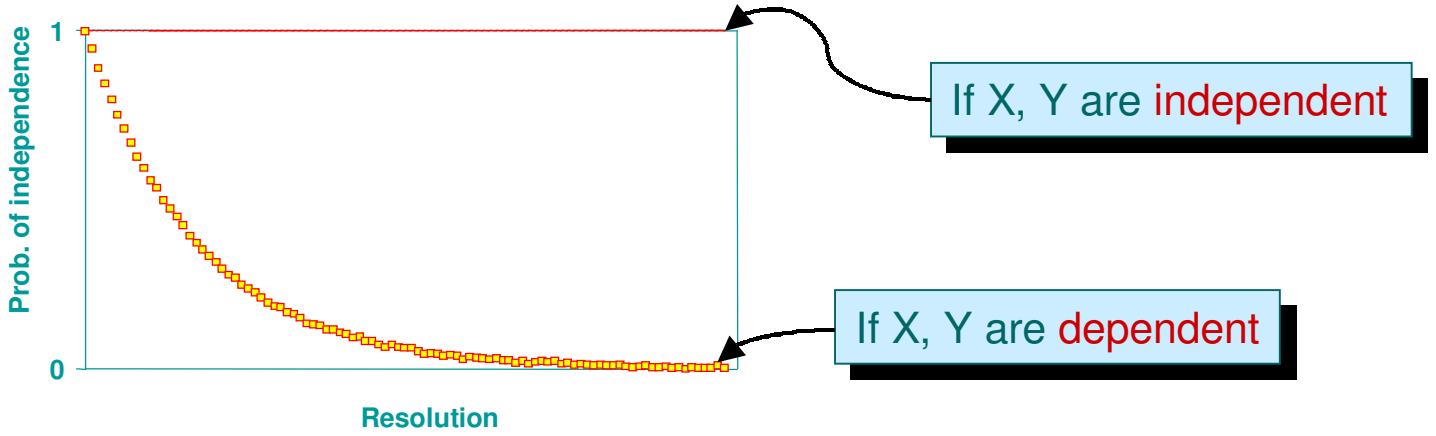
Since indep = indep at (2x2) and
indep at (3x3) and
indep at (4x4) and …

then indep = indep at (∞ x ∞)

- So we just look for independence at the highest resolution possible and we're done!
- Not so fast: we only have a finite amount of data
- What to do?

Use a Consistency Check to Stop

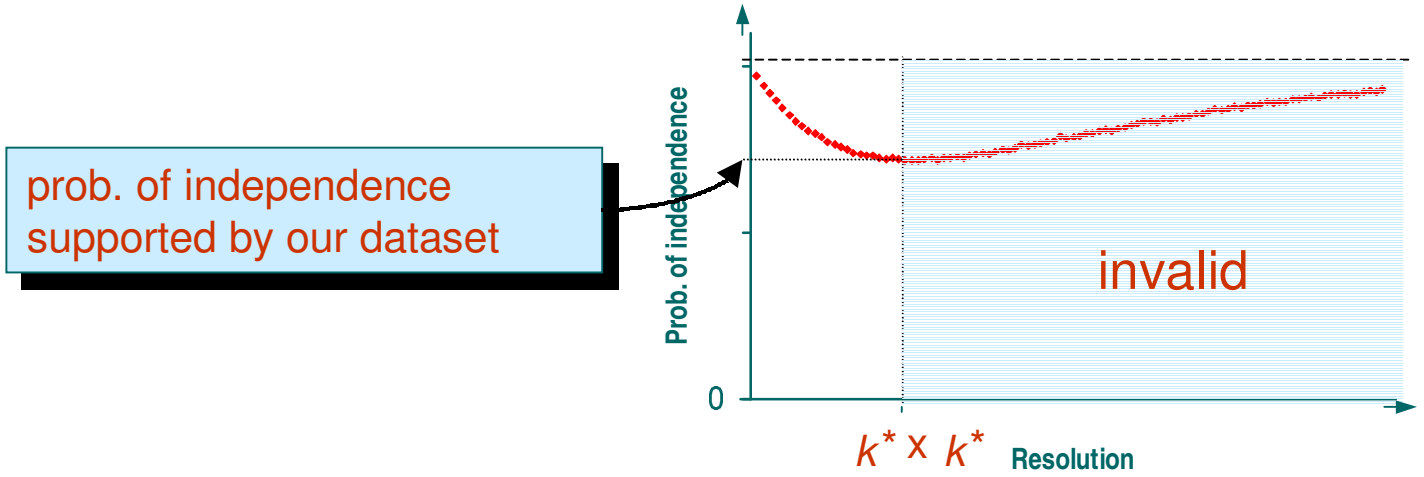
First notice that the prob. of independence is non-increasing



IF we have an infinite data set!

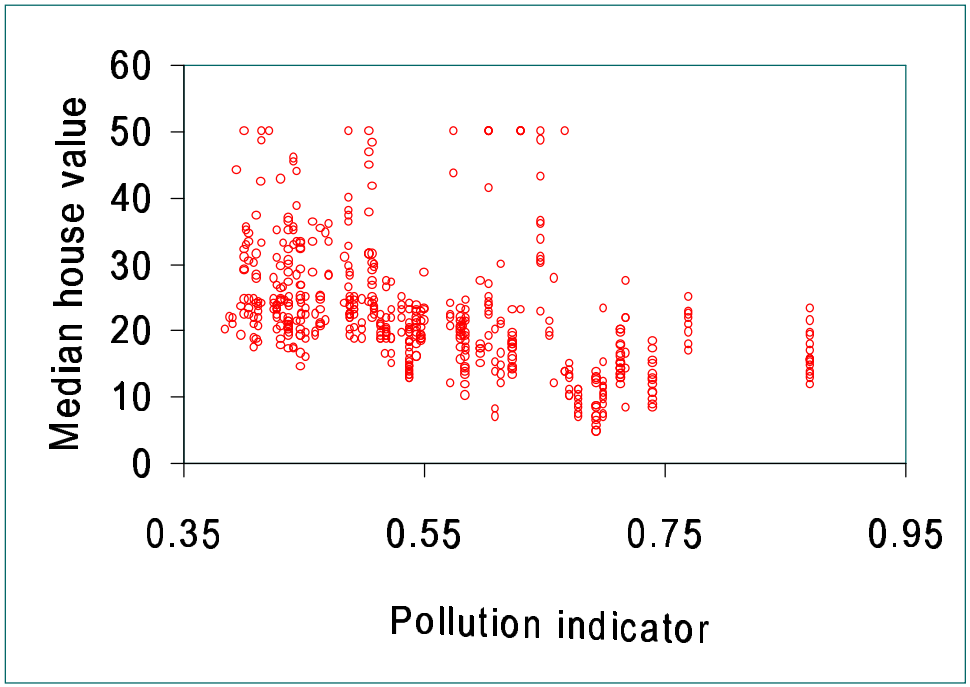
But We Have a Finite Dataset…

The Bayesian approach automatically penalizes complex models:



So we know that beyond k^* x k^* our graph is invalid --- inference on data is unsupported

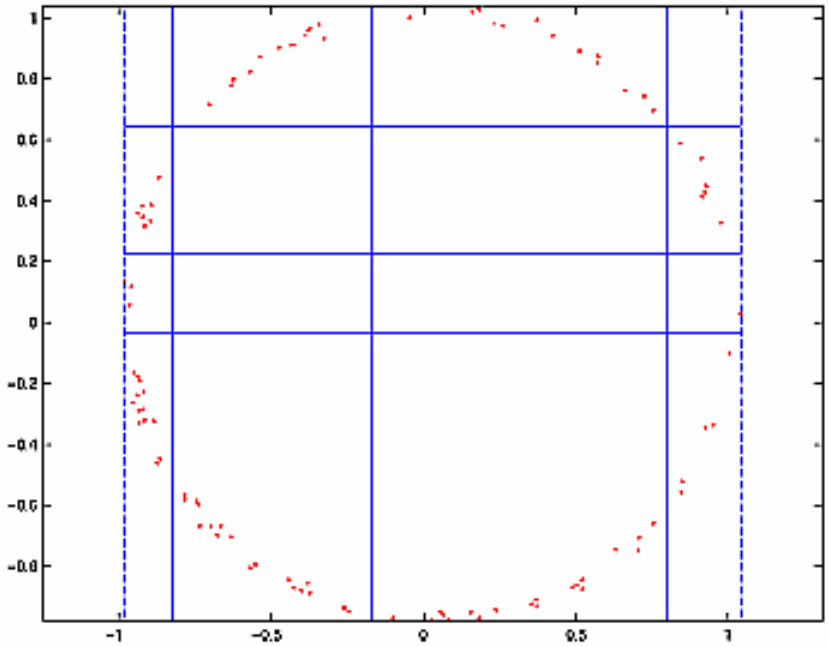
Results: Dependent Quantities



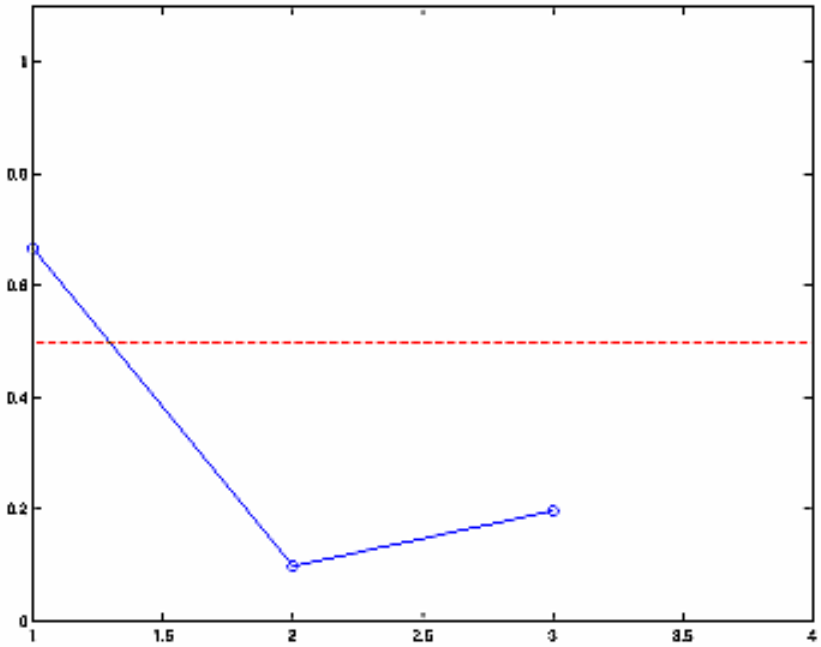
Posterior probability of independence = 0.325
(as in original study)
Rank Correlation = -0.06 (approx independent)

Artificial examples

Max-dependence discretization



prob. of independence vs. grid size

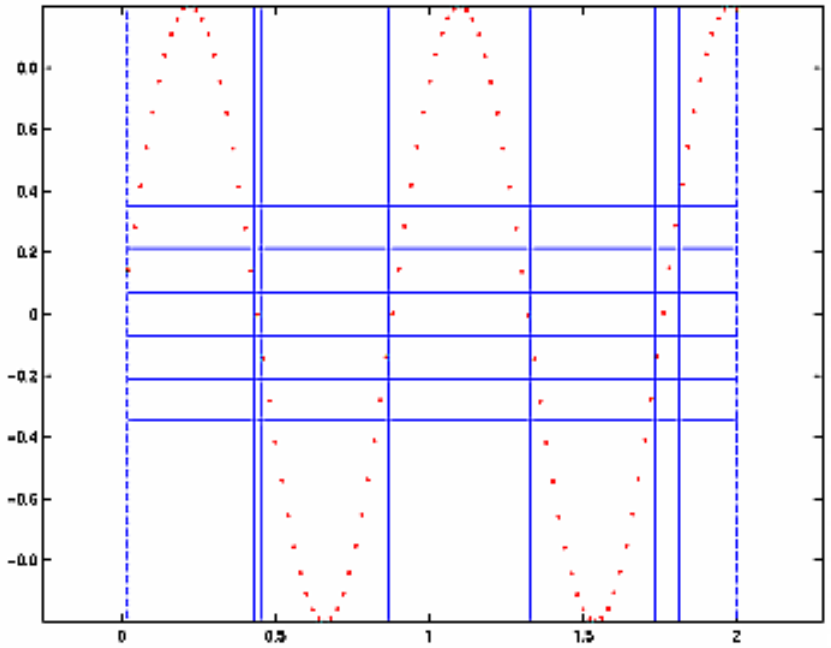


$$\text{independence probability} = 0.0986 \quad (X, Y \text{ dependent})$$

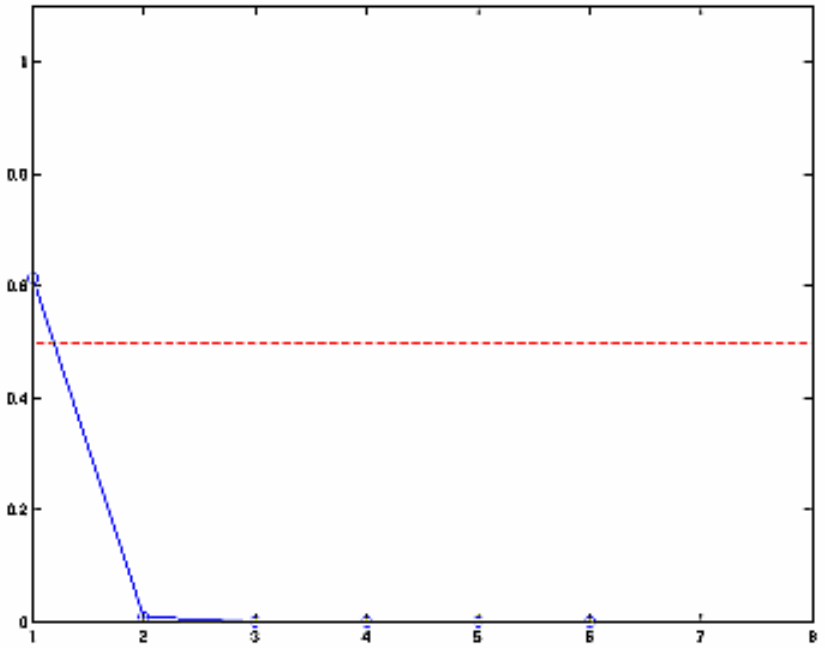
Rank Correlation 0.11 (approx. independent)

Artificial examples (cont'd)

Max-dependence discretization



prob. of independence vs. grid size



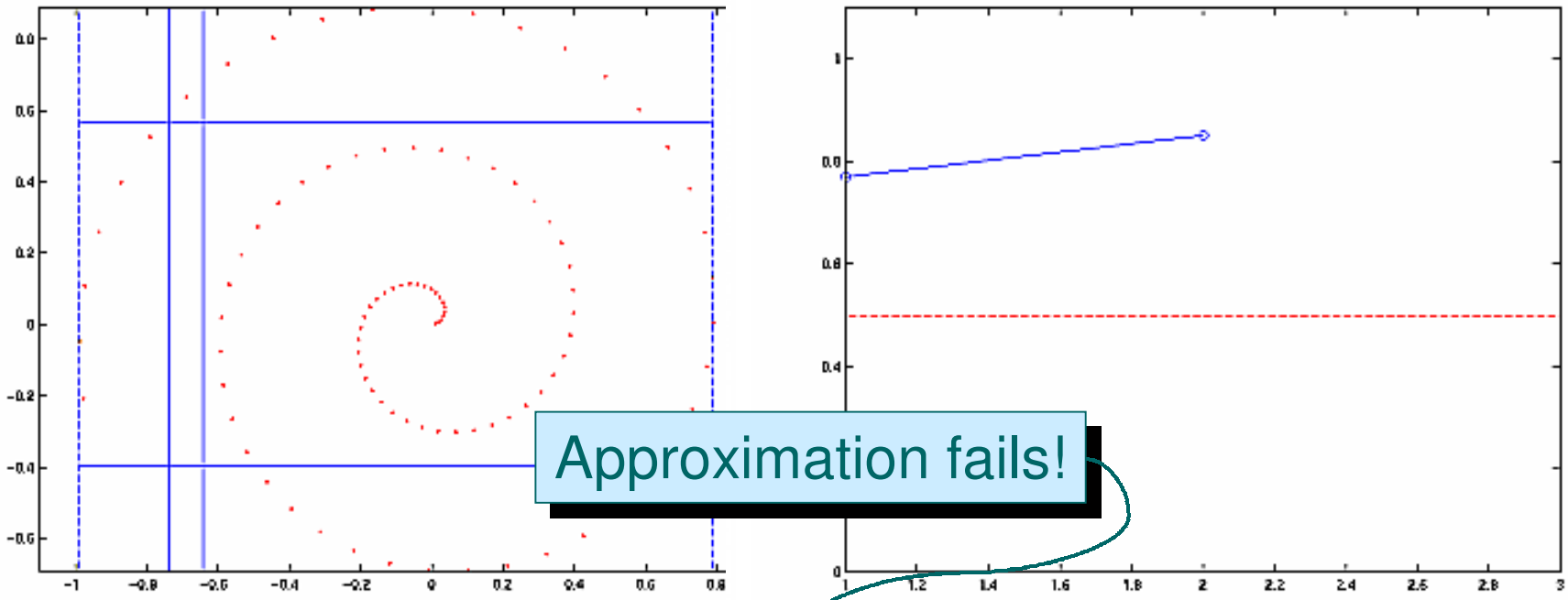
$$\text{independence probability} = 1.4\times10^{-9} \quad (X, Y \text{ dependent})$$

Rank Correlation -0.08 (approx. independent)

Artificial examples (cont'd)

Max-dependence discretization

prob. of independence vs. grid size



$$\text{independence probability} = 0.78 \quad (X, Y \text{ independent})$$

Summary --- Contributions

All my work on BN model structure

1. GS algorithm for inducing BN structure from conditional independence statements [w/Thrun NIPS 1999]
   - Minimizes conditioning set size of tests
2. Statistical independence test for continuous variables [w/Thrun UAI 2001]
   - Previously unsolved in generality
   - Takes into account multiple resolutions

Near-future Work

- Making independence test practical for large datasets---Monte-Carlo approximation
- Using it to develop a conditional test
- Application to BN structure inference with continuous variables

Thank you