Scaling log-linear analysis to datasets with 1,000+ variables

François Petitjean and Geoff Webb


2015 SIAM International
Conference on **DATA MINING**

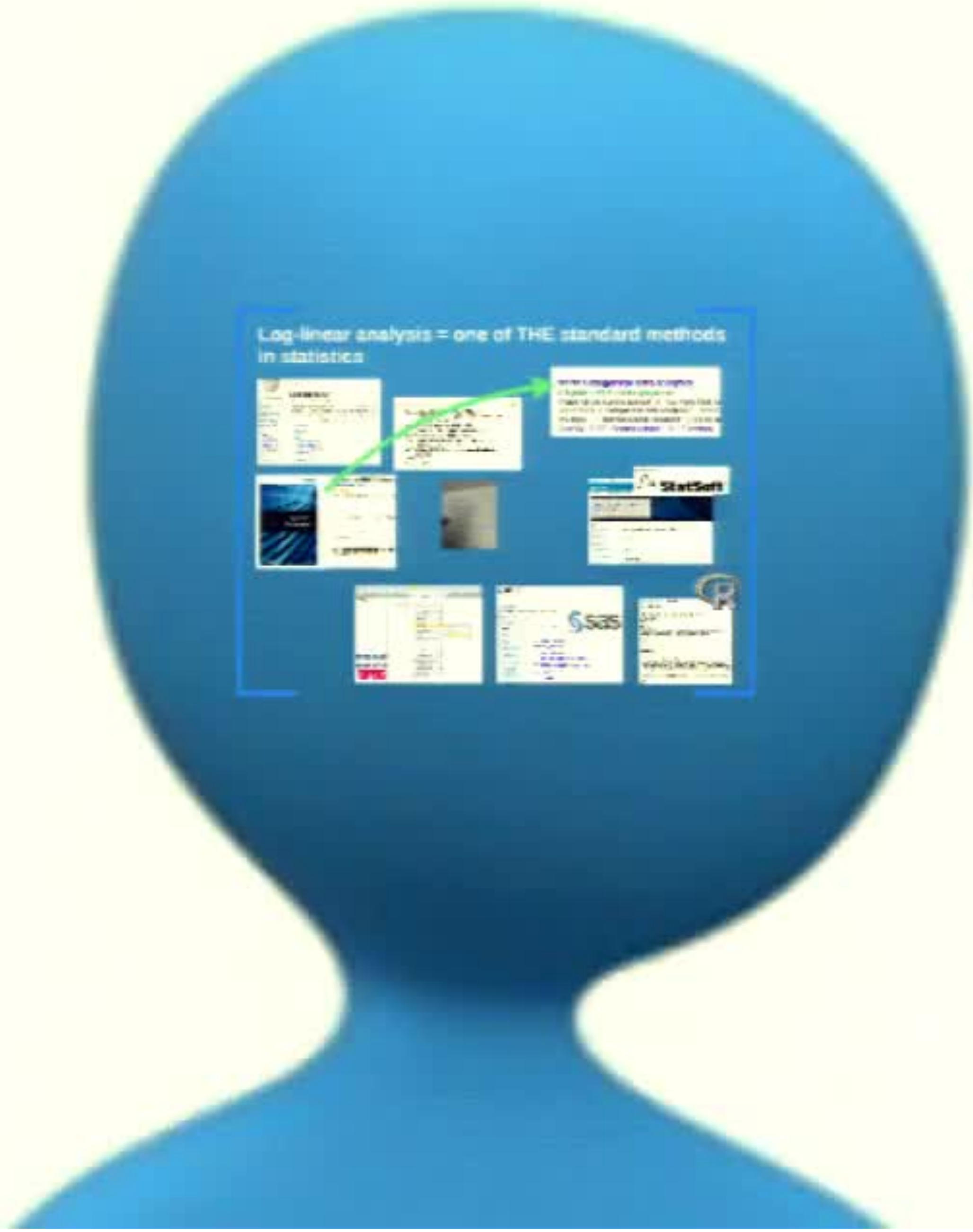
Scaling log-linear analysis to datasets with 1,000+ variables

François Petitjean and Geoff Webb

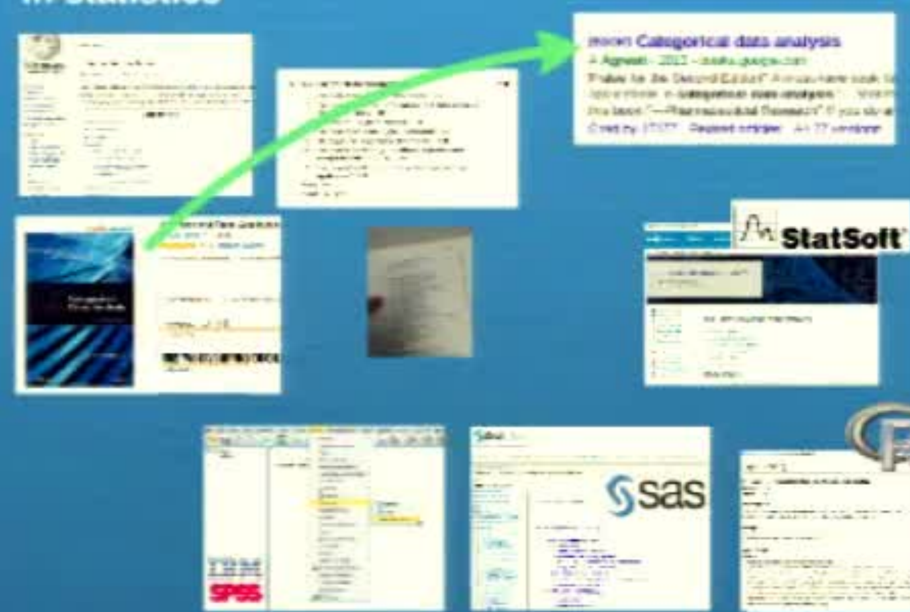
 

2015 SIAM International Conference on **DATA MINING**

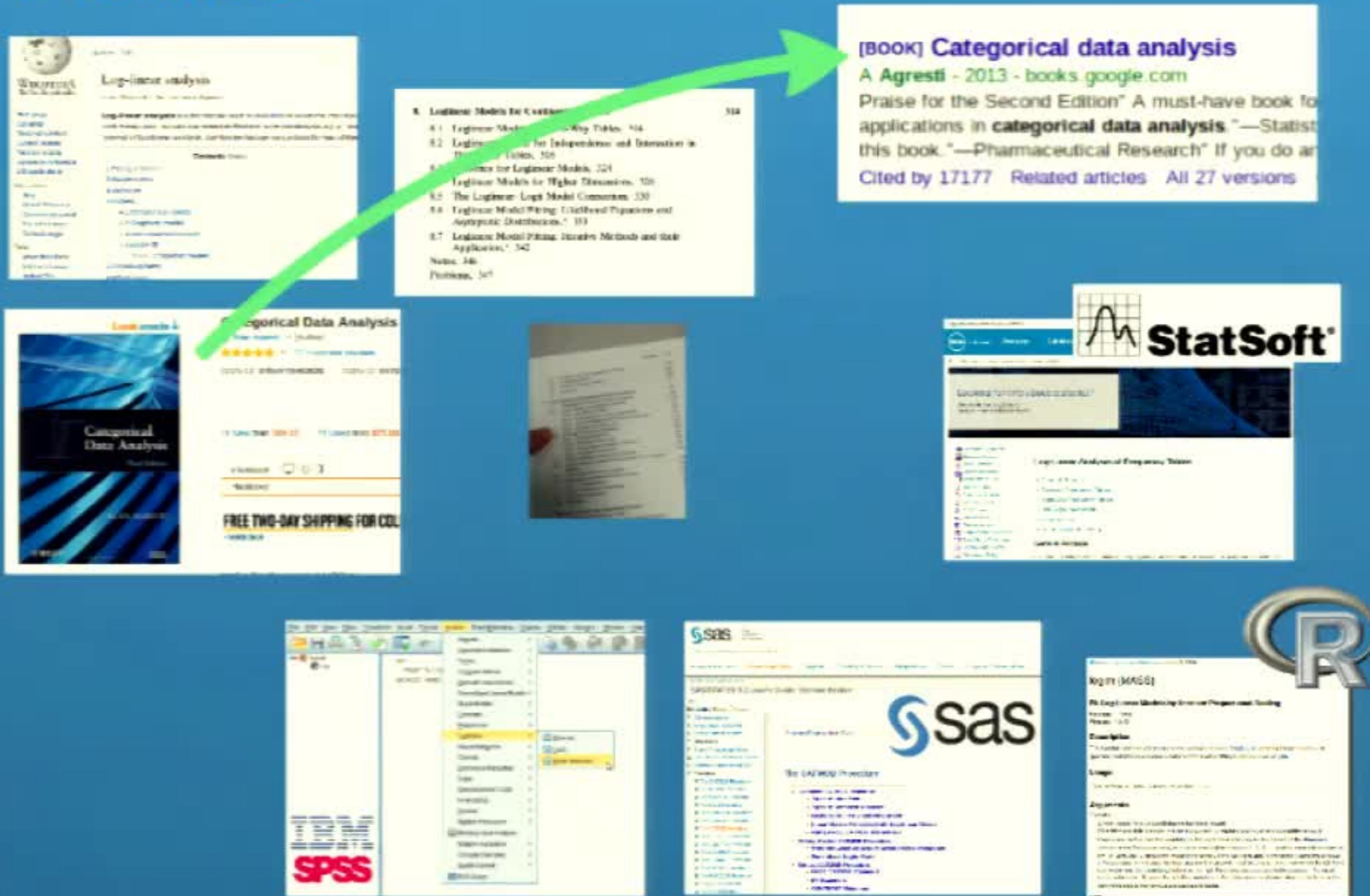
Motivation


Log-linear analysis = one of THE standard methods in statistics
StatSoft
sas

Log-linear analysis = one of THE standard methods in statistics



Log-linear analysis = one of THE standard methods in statistics

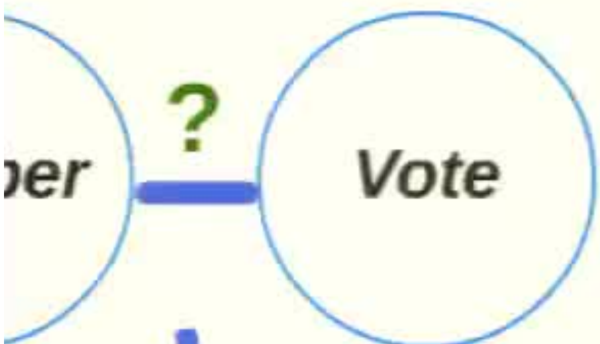


near analysis

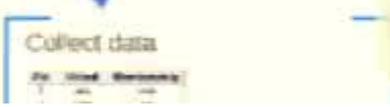
...a very simple example

ng for the presidential election
dent of being a **member of an**
tion?"

. formulate hypotheses



data



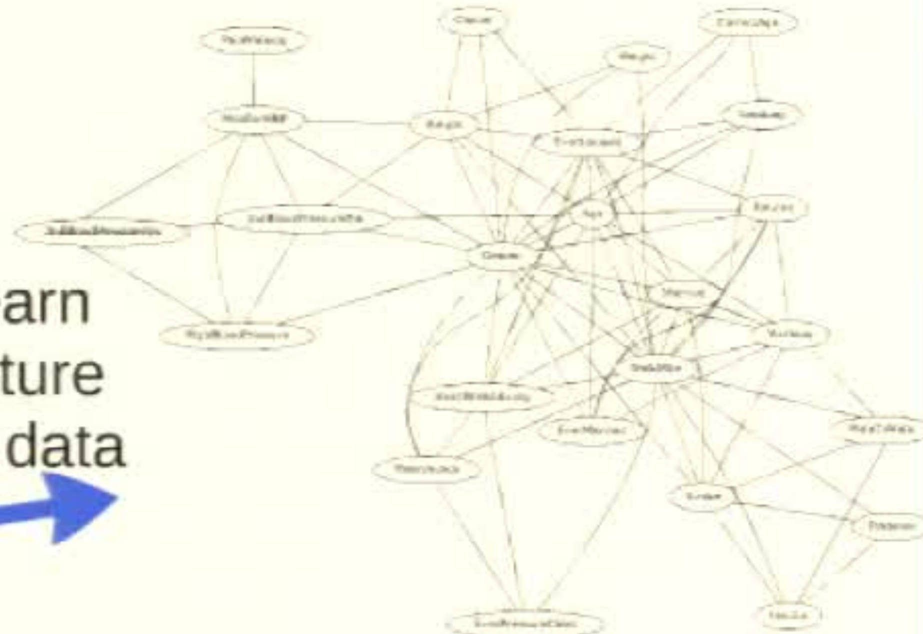
3. Evaluation

Stats

Information theory

Log-linear analysis
Learning graphical models from data

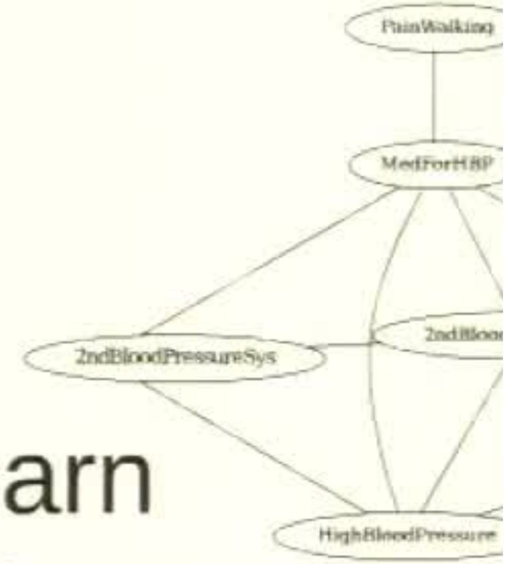
1. Learn structure from data



2. Use the structure

	A	B	E	F	GT	GU	GV	GW
1	Gender	Age	Working	Retire	ke	Cancer	Diabetes	Insulin
2	Male	85over	No	Yes	No	No	No	No
3	Female	85over	No	No	[illegible]	No	No	No
4	Male	85over	?	?	[illegible]	Yes	No	No
5	Male	80-84	No	Yes	No	No	No	No
6	Female	80-84	No	No	No	No	No	No
7	Female	85over	No	Yes	[illegible]	No	No	No
8	Female	80-84	No	No	No	No	No	No
9	Male	80-84	No	Yes	No	No	No	No
10	Female	80-84	No	Yes	[illegible]	No	No	No
11	Male	80-84	No	Yes	No	No	No	No
12	Female	75-79	No	No	No	Yes	No	No
13	Male	80-84	Yes	Yes	[illegible]	No	No	No
14	Female	80-84	No	Yes	No	No	No	No
15	Male	75-79	No	Yes	No	No	No	No
16	Female	80-84	No	Yes	[illegible]	No	No	No
17	Male	75-79	No	Yes	No	No	No	No
18	Male	75-79	Yes	No	Suspect	No	Suspect	No
19	Male	80-84	No	Yes	[illegible]	No	No	No
20	[illegible]	75-79	[illegible]	Yes	No	[illegible]	[illegible]	No
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	No	No	No	[illegible]
[illegible]	[illegible]	[illegible]	No	Yes	No	Yes	No	No
99981	Male	70-74	No	No	No	No	No	No
99982	Female	70-74	Yes	Yes	No	No	No	No
99983	Female	70-74	No	Yes	No	No	No	No
99984	Female	70-74	No	Yes	No	No	No	No
99985	Female	70-74	No	Yes	No	No	No	No
99986	Male	70-74	No	Yes	Suspect	No	Suspect	No
99987	Female	70-74	No	Yes	No	No	No	No
99988	Male	70-74	Yes	No	No	No	No	No

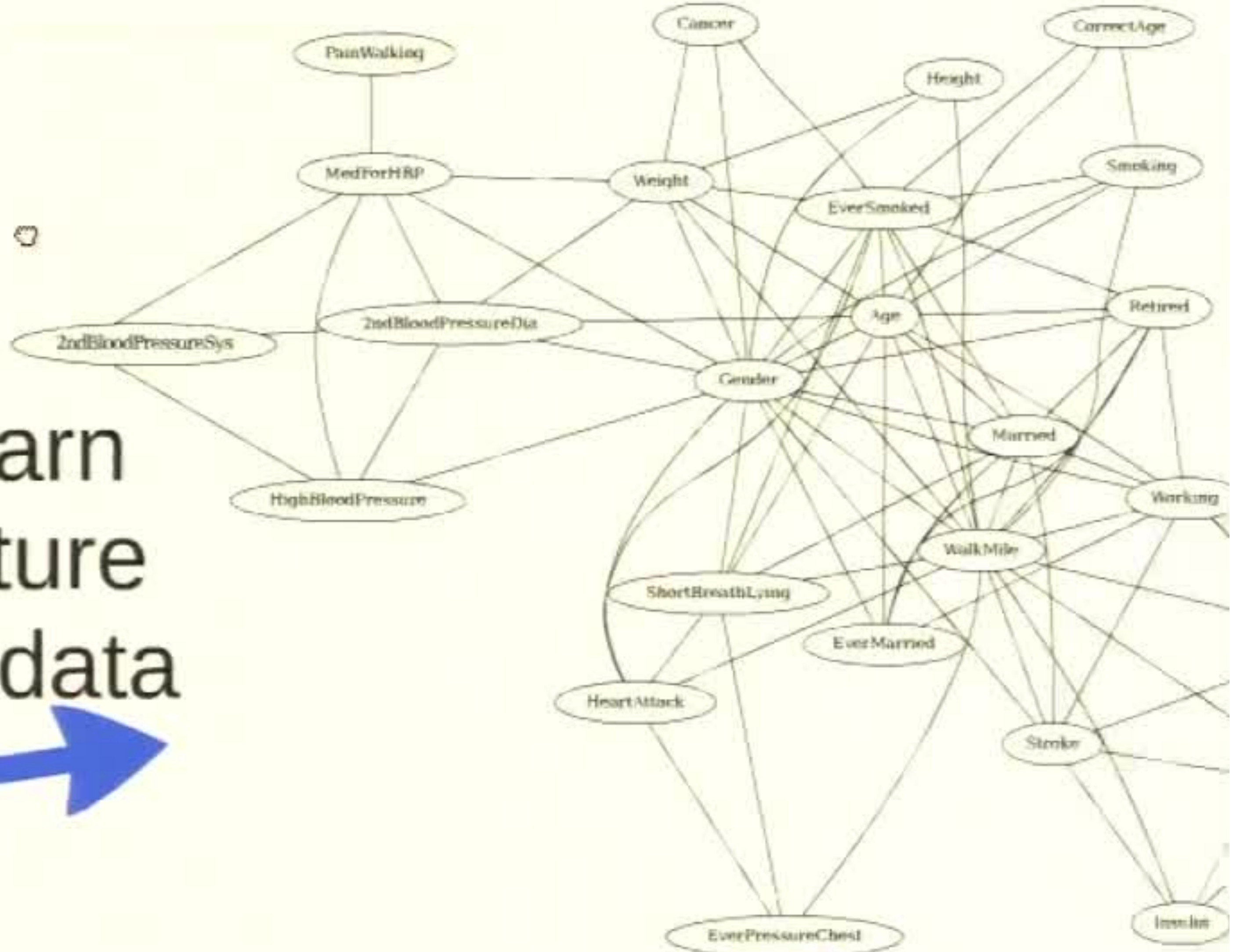
1. Learn structure from data



phical models from data

U	GV	GW
cer	Diabetes	Insulin
	No	No
	No	No
	No	No
	No	No
	No	No
	No	No
	No	No
	No	No
	No	No
	No	No
	No	No
	No	No
	No	No
	No	No
	No	No
	No	No
	Suspect	No
	No	No
		No
	No	
	No	No
	No	No
	No	No
	No	No
	No	No
	No	No
	Suspect	No
	No	No
	No	No

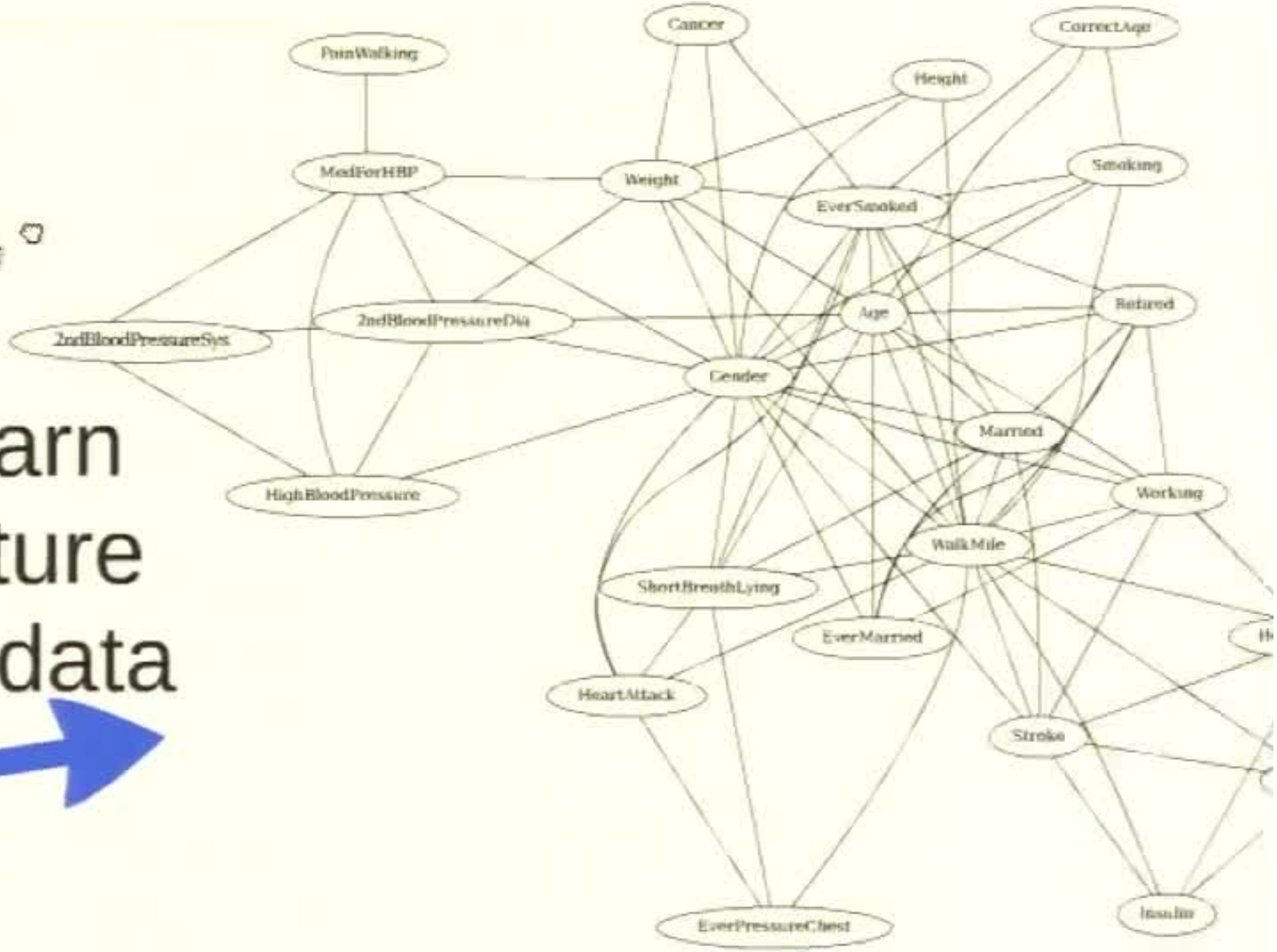
1. Learn structure from data

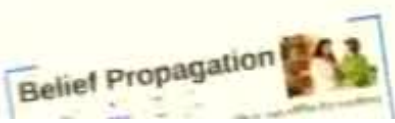


phical models from data

GV	GW
Diabetes	**Insulin**
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
No	No
Suspect	No
No	No
No	No
No	[illegible]
No	No
No	No
No	No
No	No
No	No
No	No
Suspect	No
No	No
No	No

1. Learn structure from data

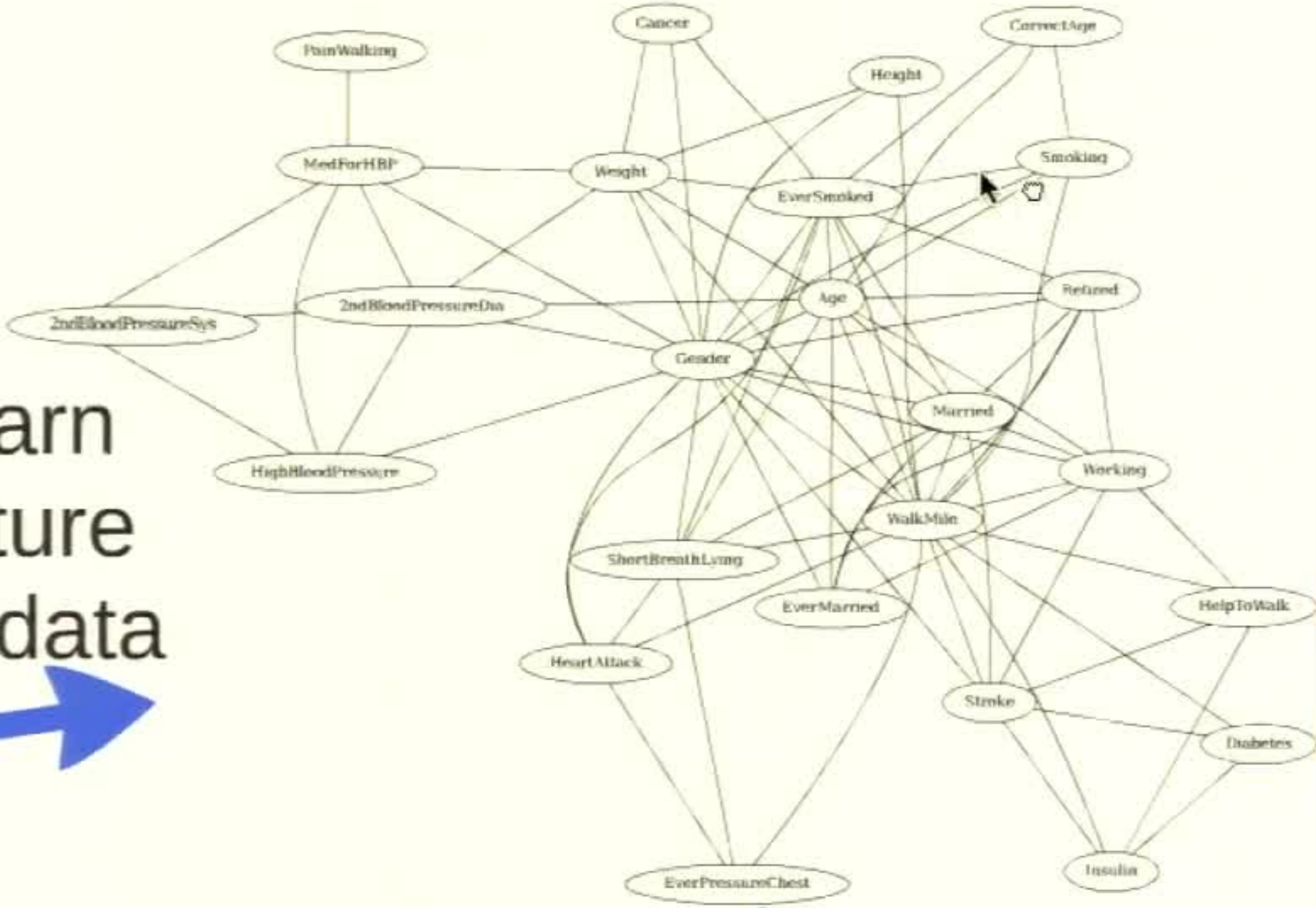




cal models from data

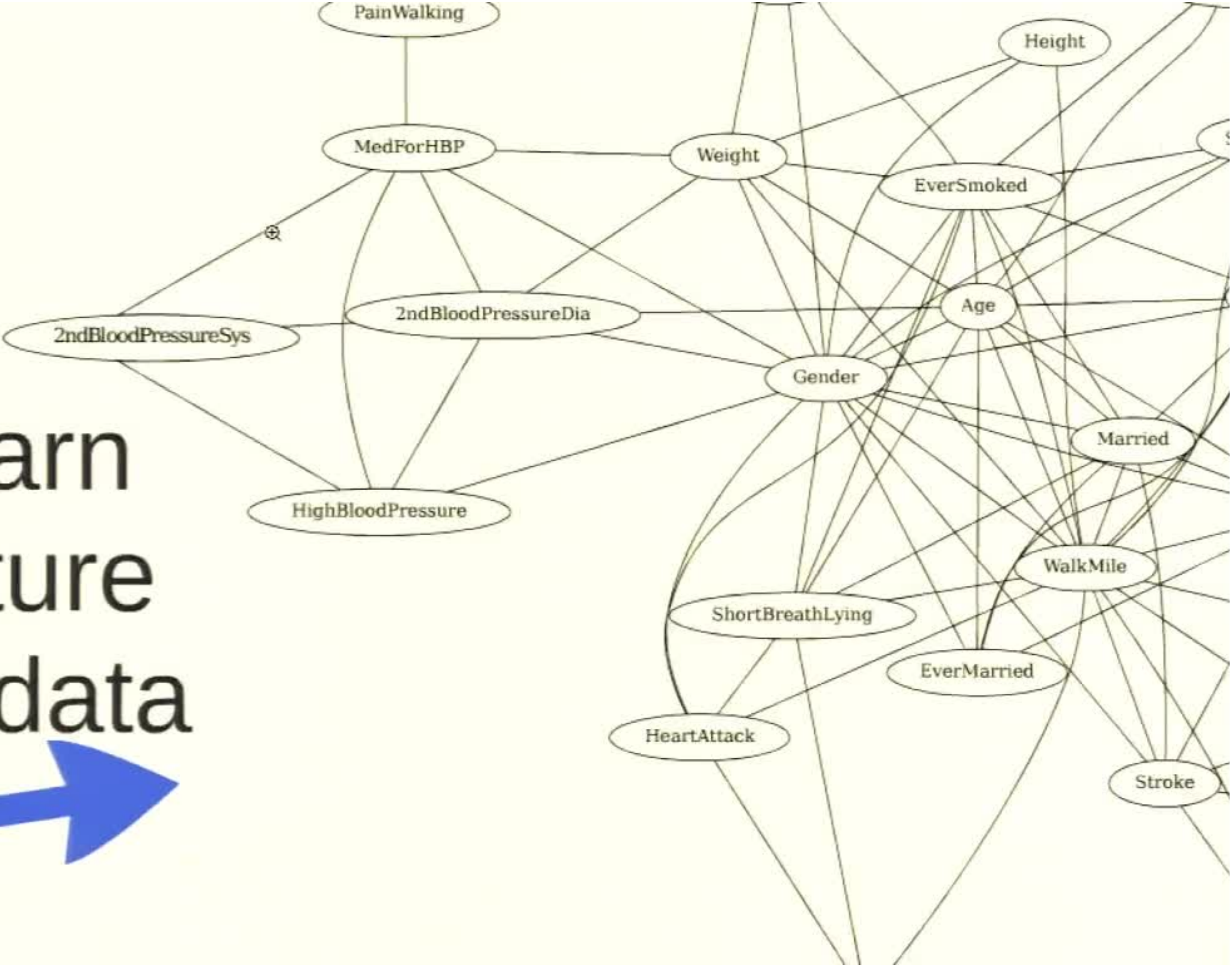
GW
sulin

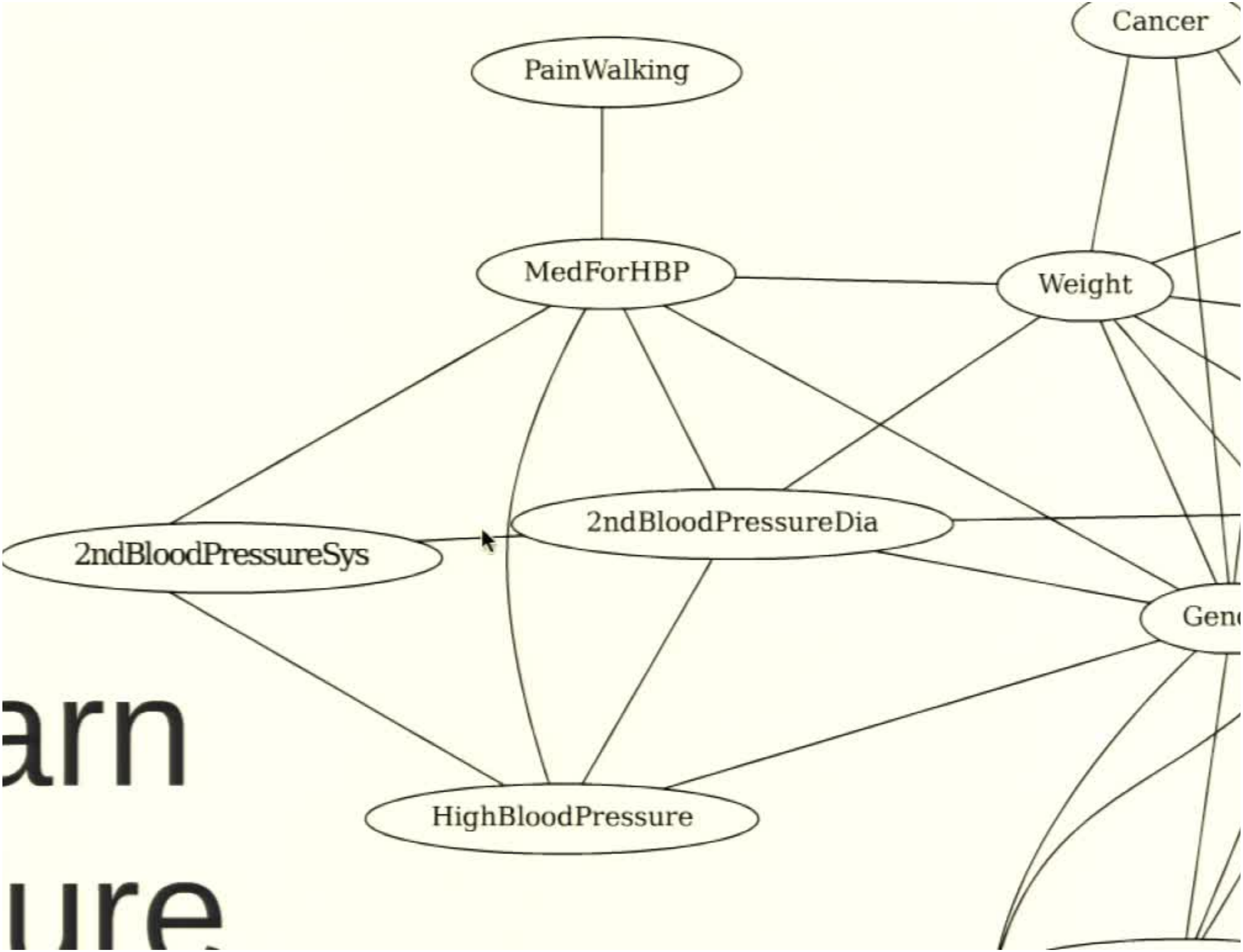
1. Learn structure from data



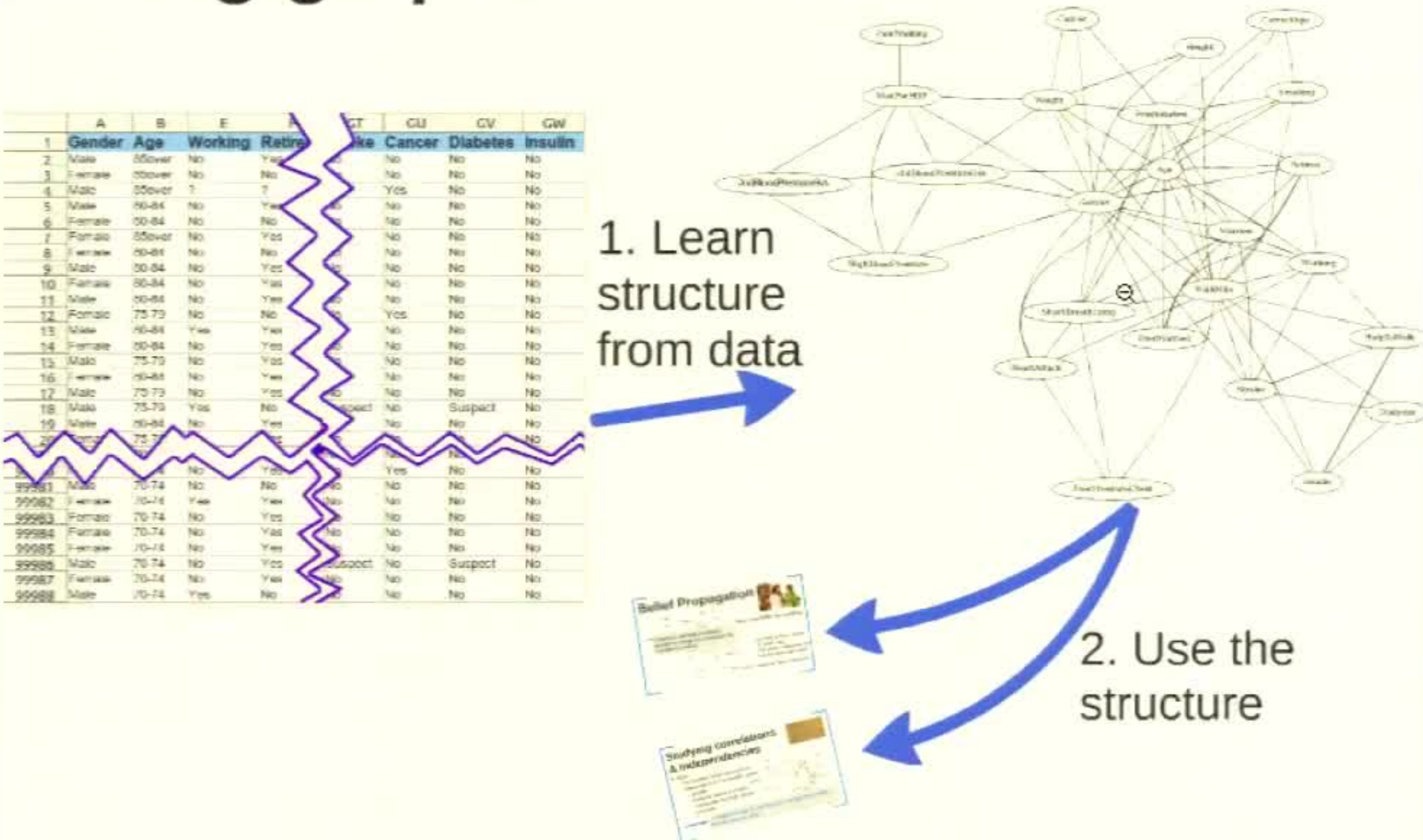
Belief Propagation

earn
ture
data



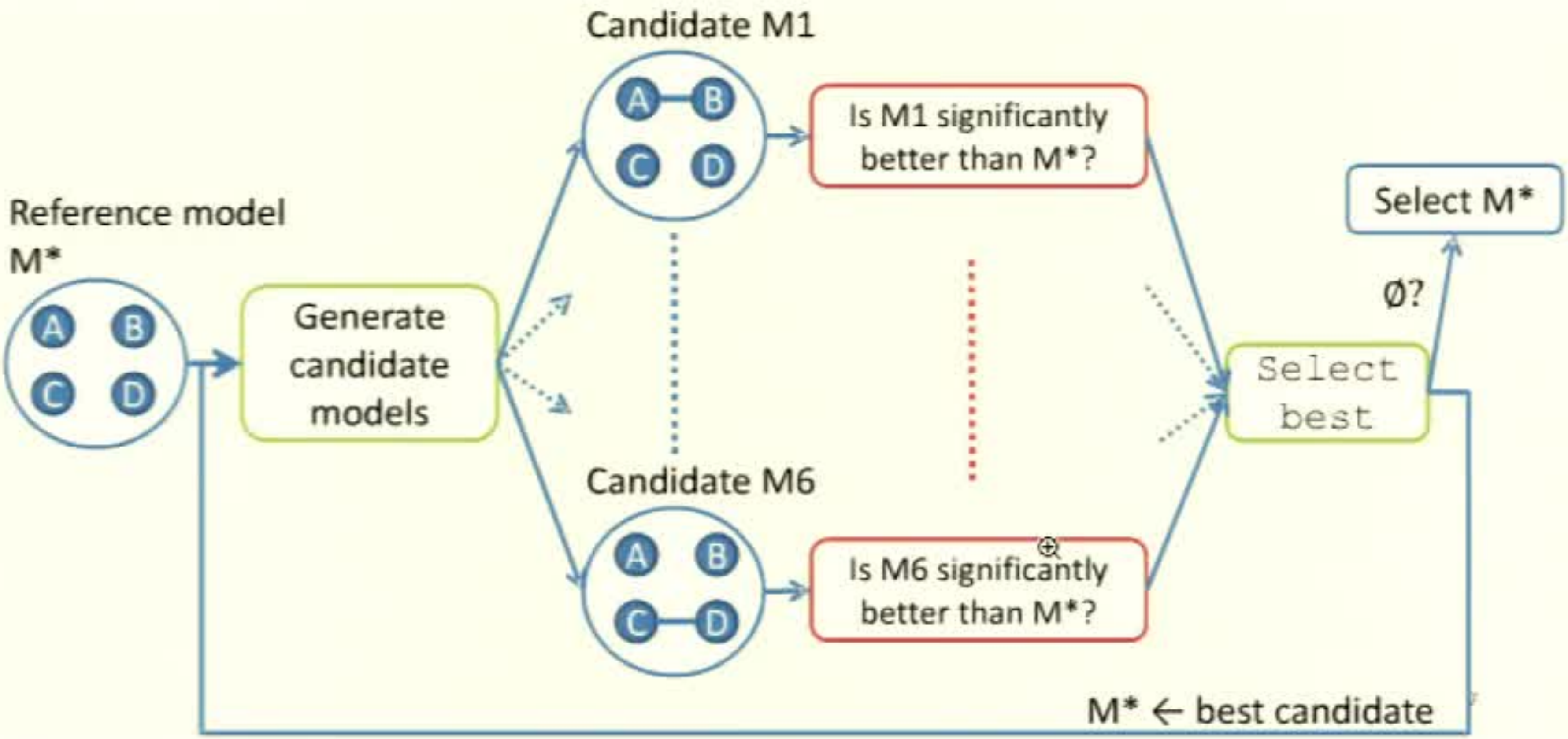

Cancer
PainWalking
MedForHBP
Weight
2ndBloodPressureDia
2ndBloodPressureSys
HighBloodPressure
arn
ure

og-linear analysis
earning graphical models from data



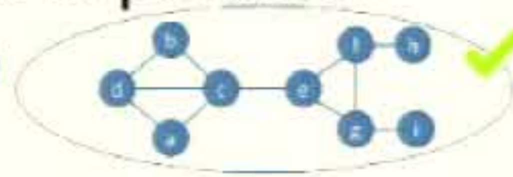
Log-linear analysis for more than 2 variables

General framework: selecting a statistically significant log-linear model = superset of Markov Networks



What have shown that...

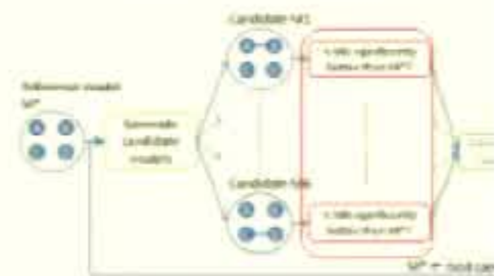
Scalability to datasets with 100 variables is possible for the class of **decomposable models**



- ICDM 2013: statistical tests
- ICDM 2014: minimum description length

Limitation: process quadratic with the number of variable

→ 1,000+ variables => **several days** of computati



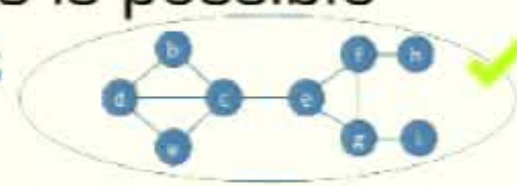
What this paper shows:

We can gain **4 orders of magnitude** *while getting* **exactly** *the same results.*

→ 1,000+ variables => **1 minute** of computation

What have shown that...

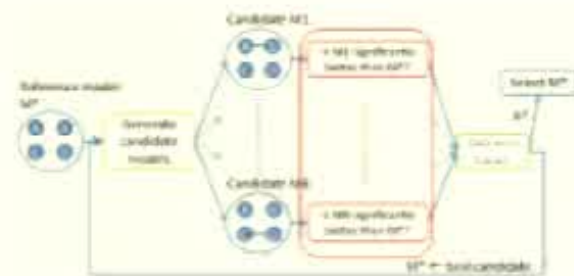
Scalability to datasets with 100 variables is possible for the class of **decomposable models**



- ICDM 2013: statistical tests
- ICDM 2014: minimum description length

Limitation: process quadratic with the number of variables

→ 1,000+ variables => **several days** of computation



What this paper shows:

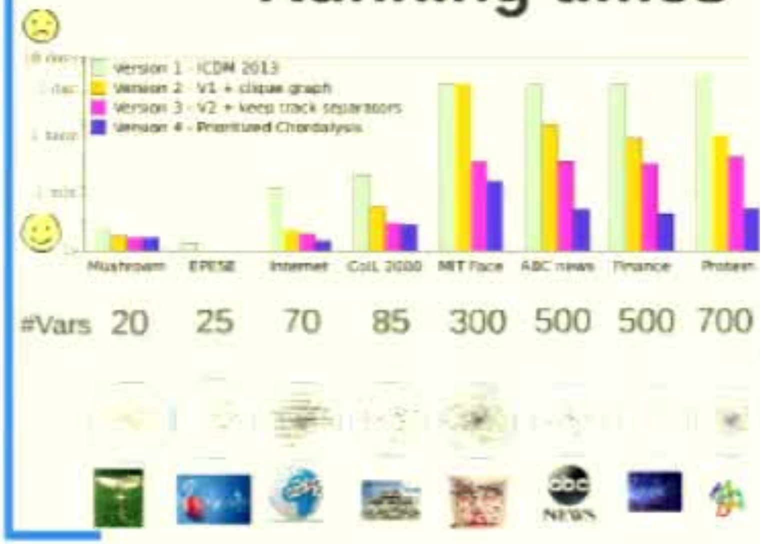
We can gain **4 orders of magnitude** *while getting* **exactly** *the same results.*

→ 1,000+ variables => **1 minute** of computation

zed

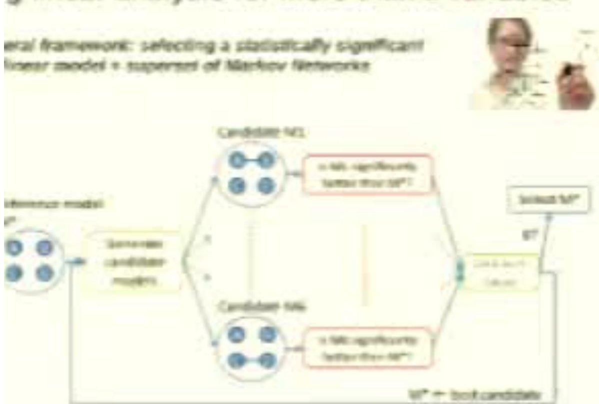
lysis

Running times



g-linear analysis for more than 2 variables

eral framework: selecting a statistically significant
linear model + superset of Markov Networks



What have shown that...

Scalability to datasets with 100 variables is possible for the class of **decomposable models**

- ICDM 2013: statistical tests
- ICDM 2014: minimum description length

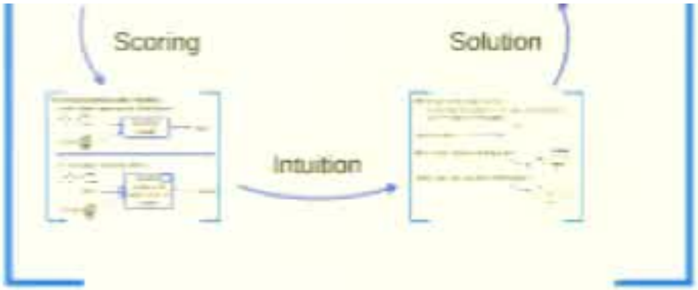
Limitation: process quadratic with the number of variables

→ 1.000+ variables => **several days** of computation

What this paper shows:

We can gain **4 orders of magnitude** *while getting* **exactly** *the same results*

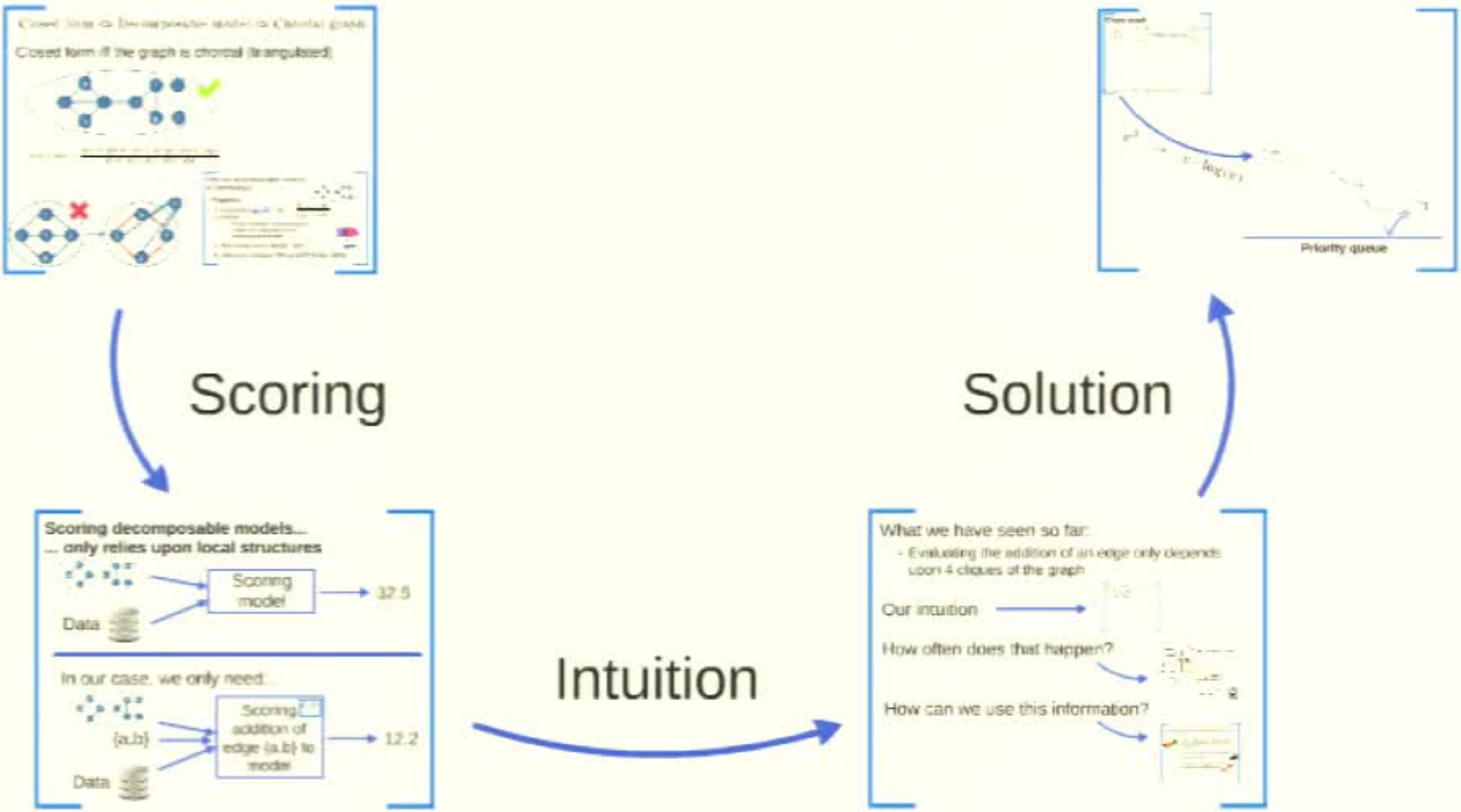
→ 1.000+ variables => **1 minute** of computation



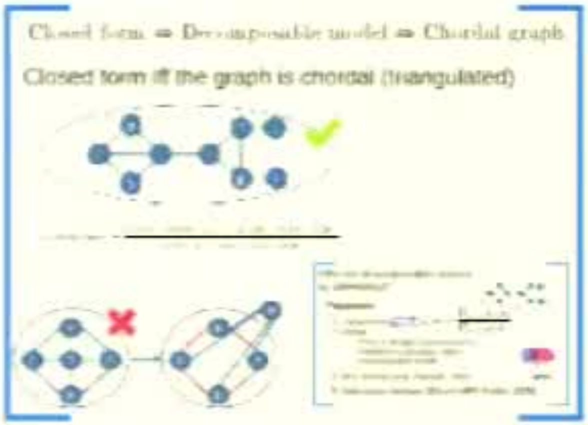
It wo

Prioritized Chordalysis

Prioritized Chordalysis

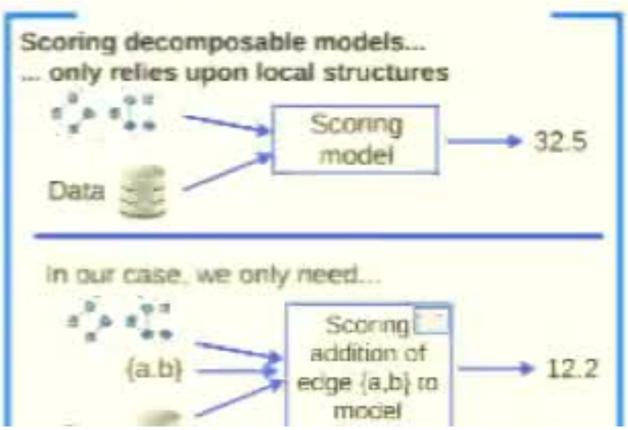


Prioritized Chordalys

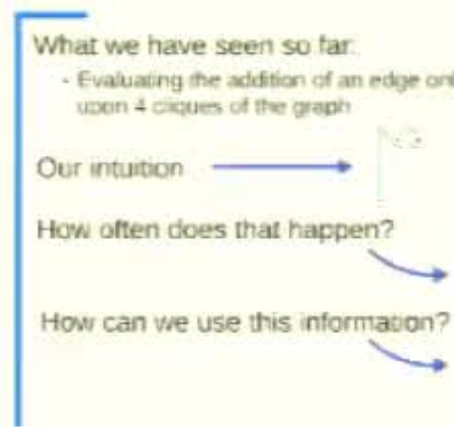


Scoring

Solution

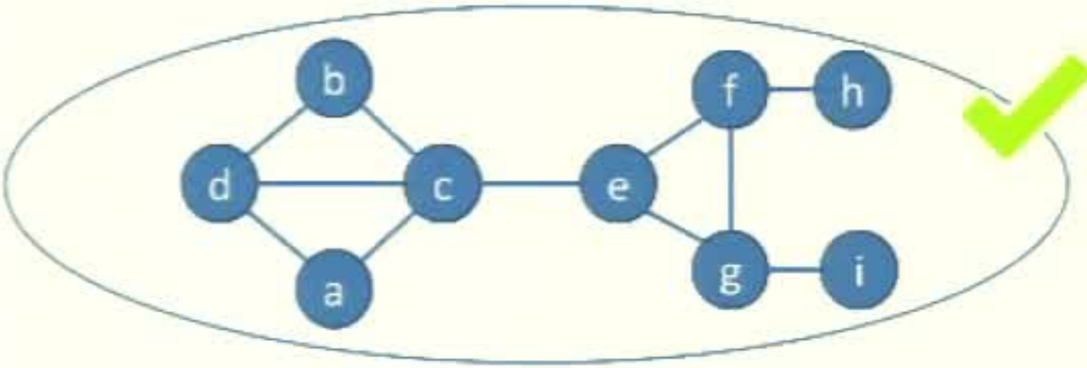


Intuition

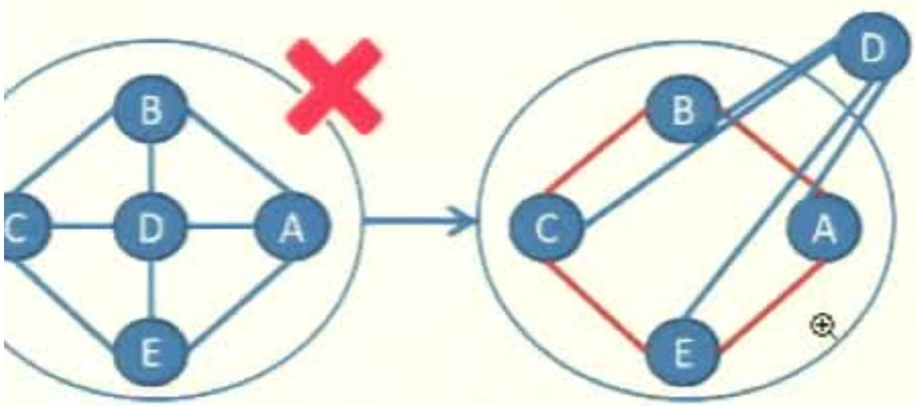


Closed form ⇔ Decomposable model ⇔ Chordal graph

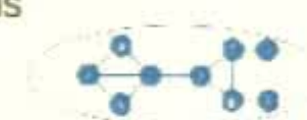
Closed form iff the graph is chordal (triangulated)



$$p(abcdefghi) = \frac{p(acd)\cdot p(bcd)\cdot p(ce)\cdot p(efg)\cdot p(fh)\cdot p(gi)}{p(cd)\cdot p(c)\cdot p(e)\cdot p(f)\cdot p(g)}$$

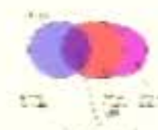


Why are decomposable models so *interesting*?



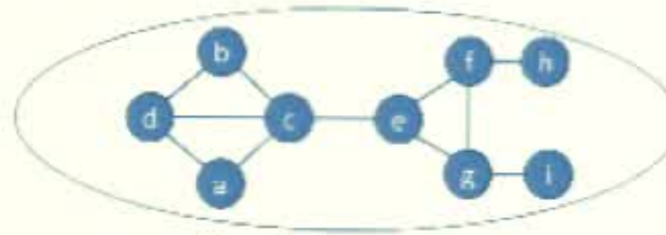
Properties:

1. Closed form $\rightleftarrows$ $p_\mu(\mathbf{x}) = \frac{\prod_{C\in\mathcal{C}} p_C(\mathbf{x})}{\prod_{S\in\mathcal{S}} p_S(\mathbf{x})}$
2. Useful:
   - There is always a decomposable model that subsumes a non-decomposable model



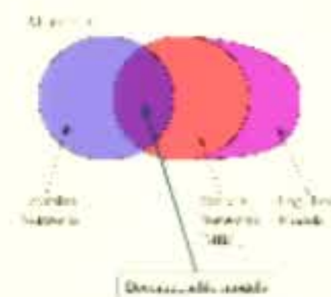
3. MLE always exist [Agresti, 2002]
4. Intersection between BN and MRF [Koller, 2009]

Why are decomposable models so *interesting*?



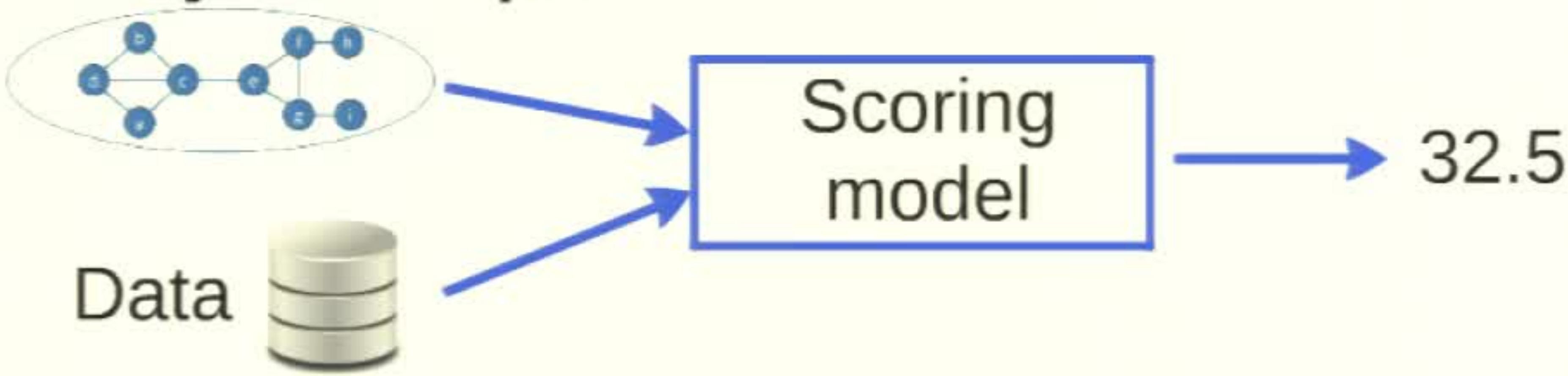
Properties:

1. Closed form $$p_\mu(\mathbf{x}) = \frac{\prod_{C\in\mathcal{C}} p_C(\mathbf{x})}{\prod_{S\in\mathcal{S}} p_S(\mathbf{x})}$$
2. Useful:
   - There is always a decomposable model that subsumes a non-decomposable model

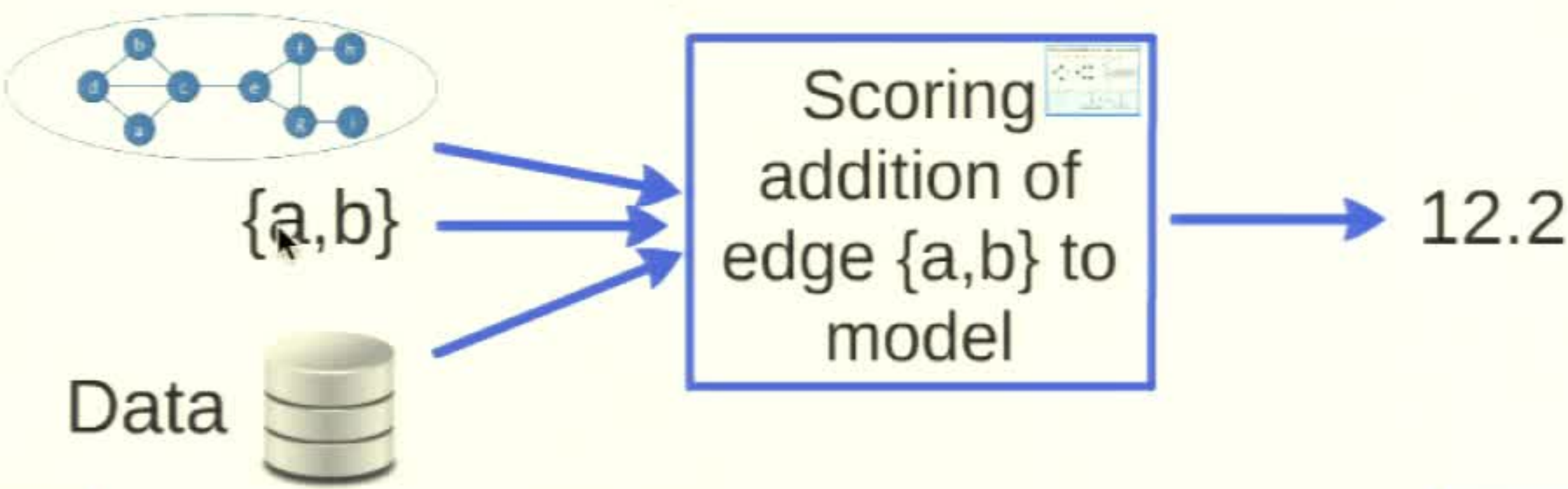


3. MLE always exist [Agresti, 2002]
4. Intersection between BN and MRF [Koller, 2009]

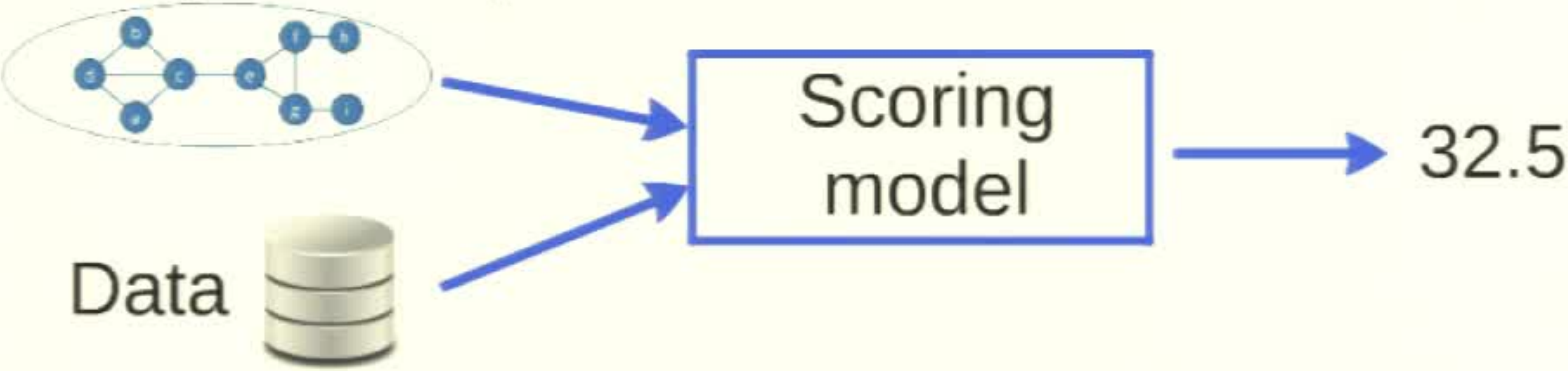
Scoring decomposable models...
... only relies upon local structures



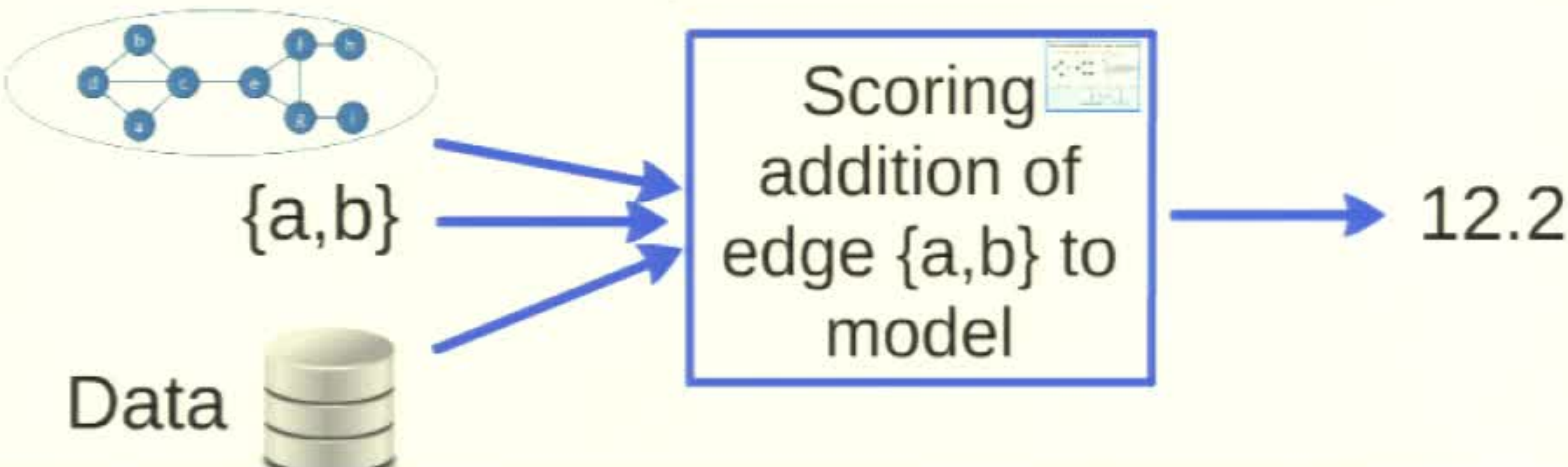
In our case, we only need...



Scoring decomposable models...
... only relies upon local structures



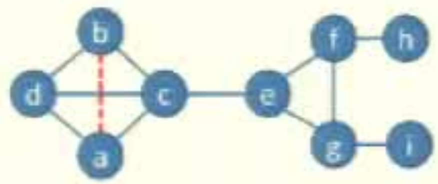
In our case, we only need...



ng

Scoring *the addition of an edge to a model*

$$score(\mathcal{M},(a,b),\mathcal{D}) = score'(a,b,S_{ab},\mathcal{D})$$



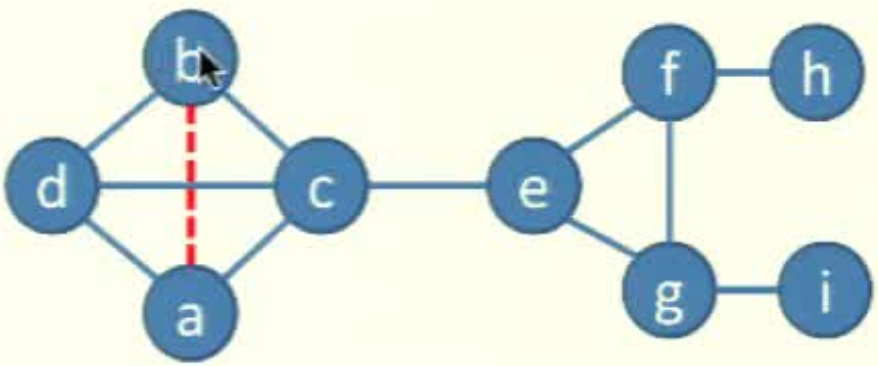
S_{ab}: minimal separator of (a,b)
= **minimal set of vertices** that would **disconnect** a from *b* if removed from the graph
= {c,d}

$$score(\mathcal{M},\{a,b\}) = score'(\{a,b,c,d\}\ ,\ \{a,c,d\}\ ,\ \{b,c,d\}\ ,\ \{c,d\})$$

$S_{ab} \cup a \cup b$ $\quad$ $S_{ab} \cup a$ $\quad$ $S_{ab} \cup b$ $\quad$ S_{ab}

Scoring *the addition of an edge to a model*

$$score(\mathcal{M}, (a,b), \mathcal{D}) = score'(a, b, S_{ab}, \mathcal{D})$$

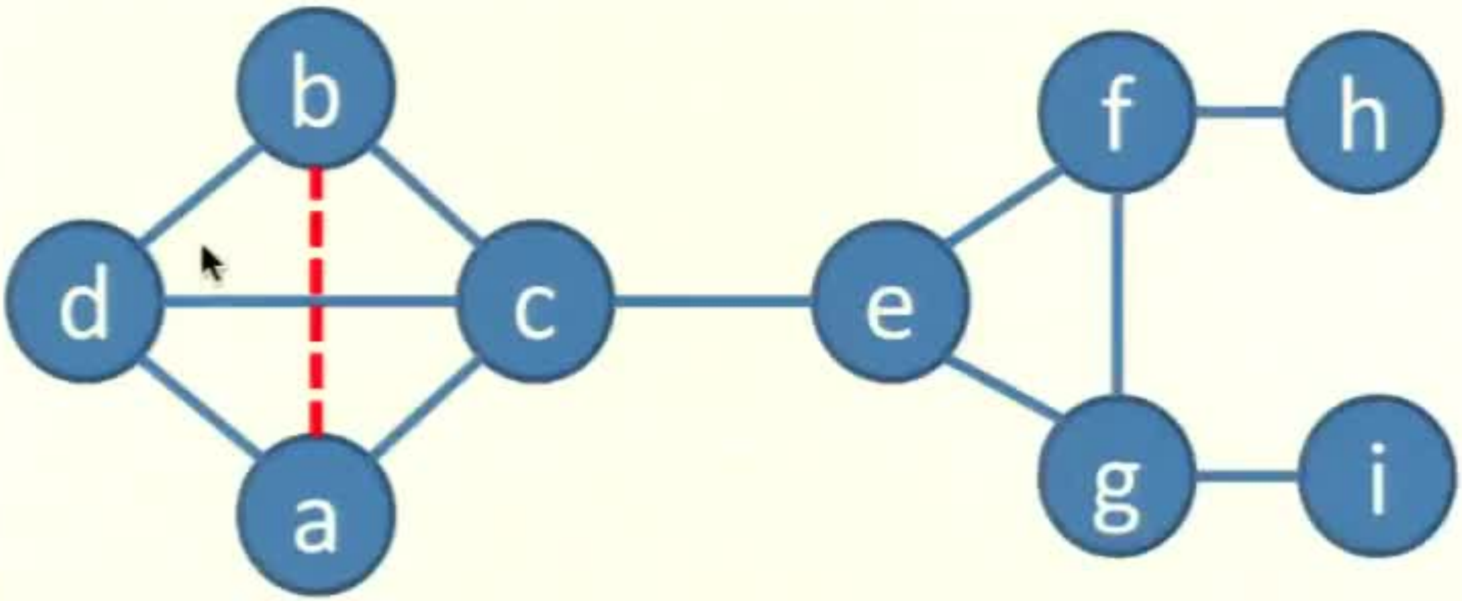


S_{ab} : minimal separator of (a,b)
= **minimal set of vertices** that would **disconnect** *a* from *b* if removed from the graph
= {c,d}

$$score(\mathcal{M}, \{a,b\}) = score'(\{a,b,c,d\}\ ,\ \{a,c,d\}\ ,\ \{b,c,d\}\ ,\ \{c,d\})$$

$S_{ab} \cup a \cup b$ $\quad$ $S_{ab} \cup a$ $\quad$ $S_{ab} \cup b$ $\quad$ S_{ab}

$$score(\mathcal{M},(a,b),\mathcal{D}) = score'(a,$$



S_{ab}: n

=

wo

re

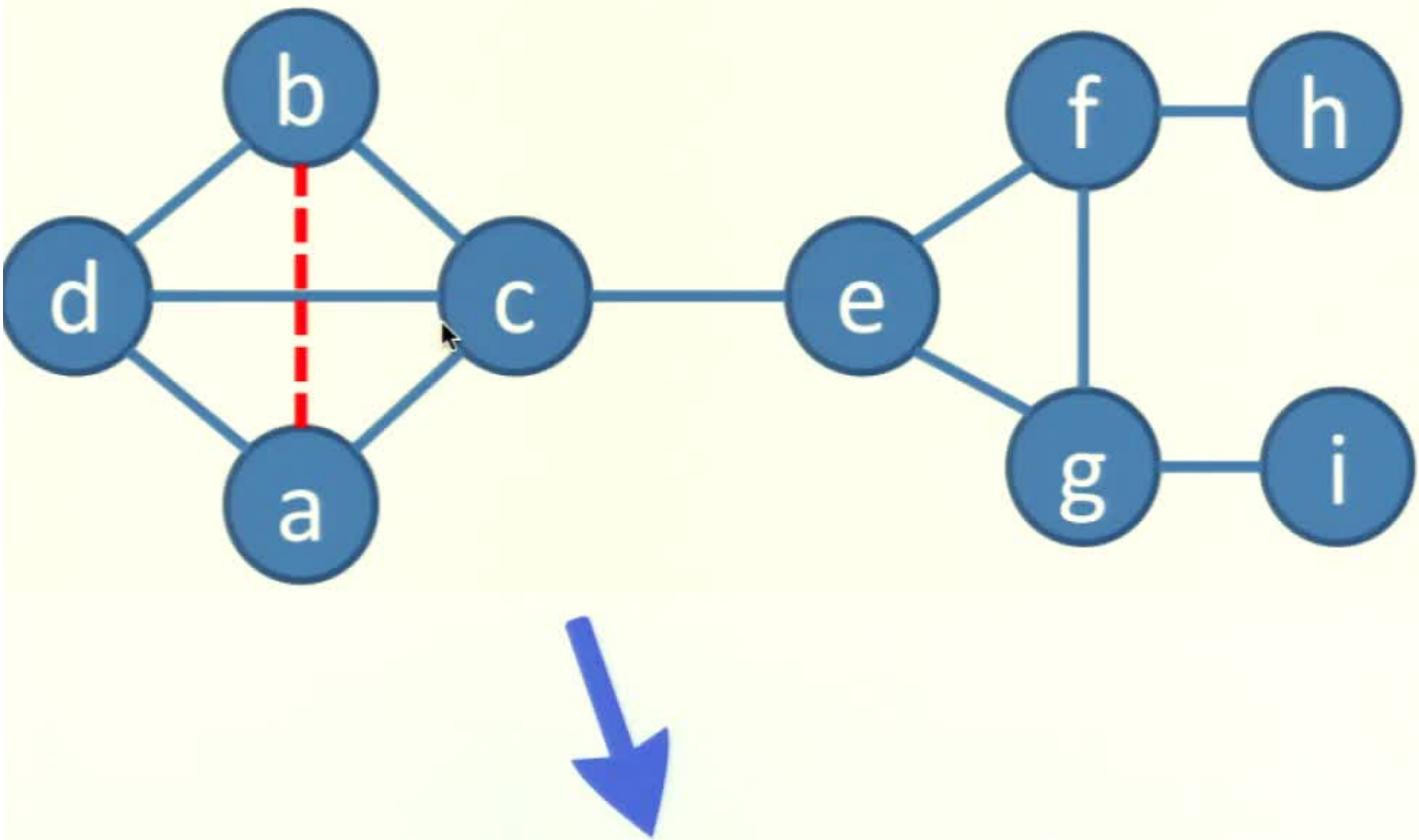
= {

$$score(\mathcal{M},\{a,b\}) = score'(\{a,b,c,d\}\ ,\ \{a,$$

This has been proven for different scorings

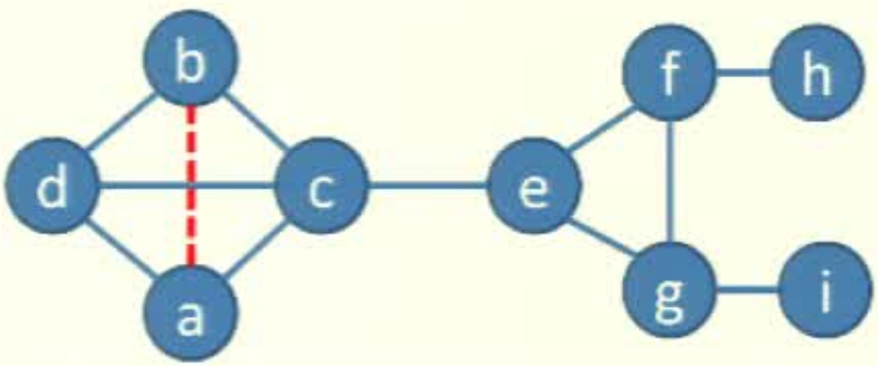
- Statistical tests [1]

$$score(\mathcal{M}, (a, b), D) = sc$$



Scoring *the addition of an edge to a model*

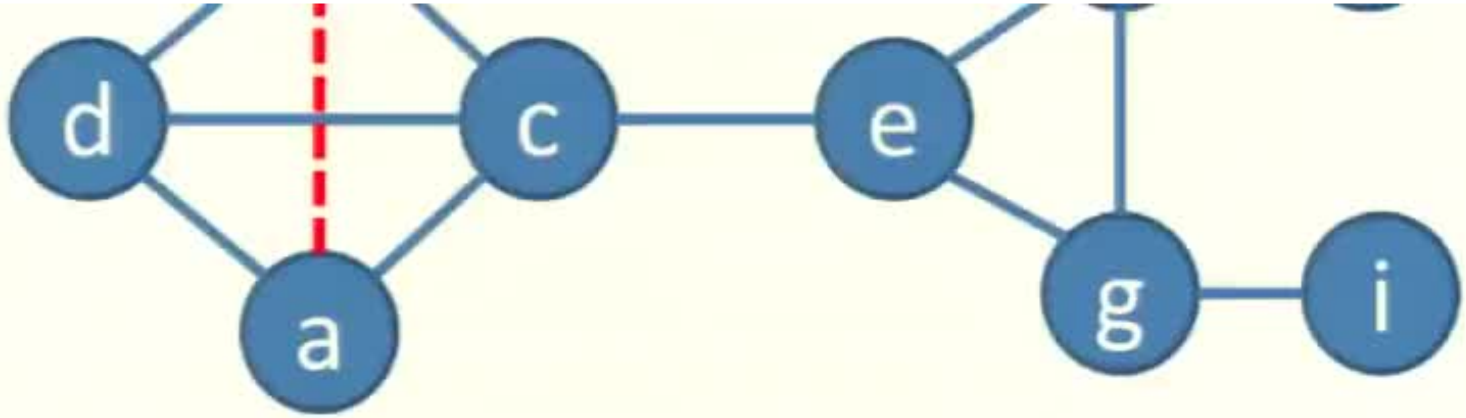
$$score(\mathcal{M}, (a,b), \mathcal{D}) = score'(a, b, S_{ab}, \mathcal{D})$$



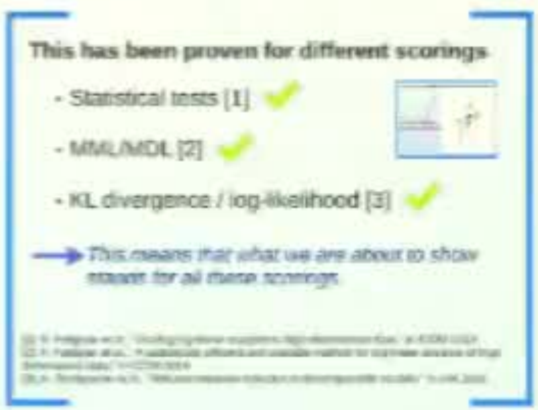
S_{ab}: minimal separator of (a,b)
= **minimal set of vertices** that would **disconnect** *a* from *b* if removed from the graph
= {c,d}

$$score(\mathcal{M}, \{a,b\}) = score'(\{a,b,c,d\}\ ,\ \{a,c,d\}\ ,\ \{b,c,d\}\ ,\ \{c,d\})$$

$S_{ab} \cup a \cup b$ $\quad$ $S_{ab} \cup a$ $\quad$ $S_{ab} \cup b$ $\quad$ S_{ab}



$$score(\mathcal{M}, \{a, b\}) = score'(\{a, b, c, d\} \ , \ \{$$

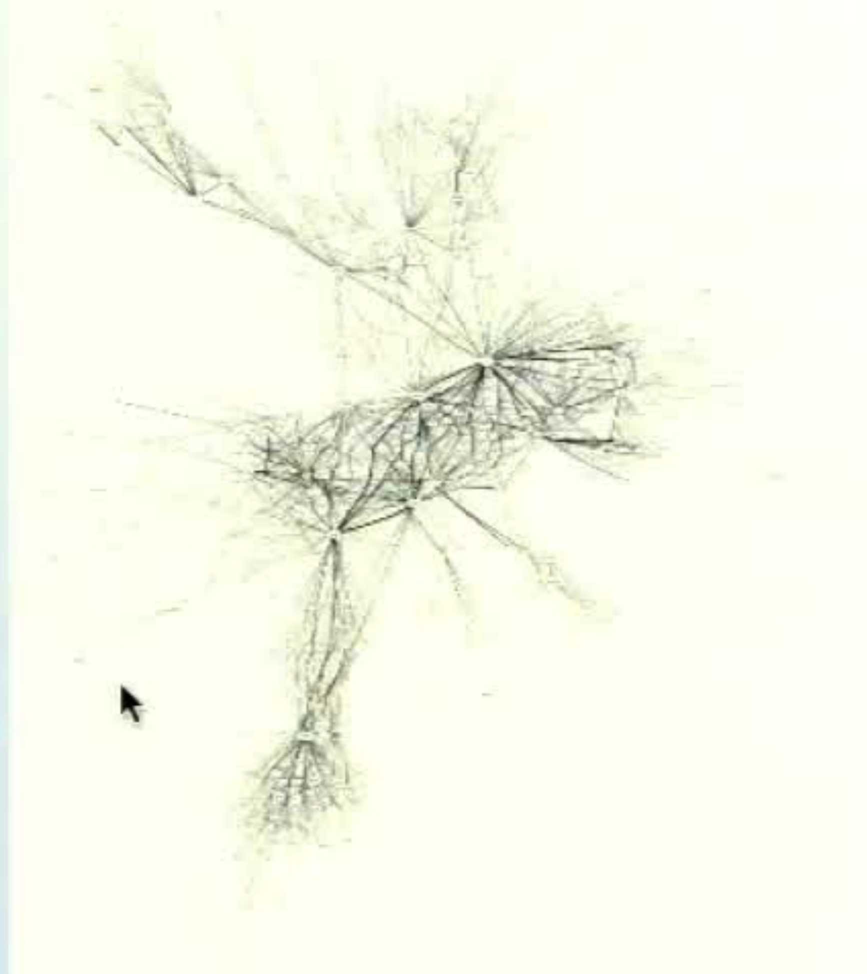


S

$$S_{ab} \cup a \cup b$$

Assessing the addition of one edge to this model?

We only need to consider **4 cliques**



This has been proven for different scorings

- Statistical tests [1]
- MML/MDL [2]
- KL divergence / log-likelihood [3]

This means that what we are about to show stands for all these scorings.

[1]: F. Petitjean *et al.*, "Scaling log-linear analysis to high-dimensional data," in *ICDM 2013*.
[2]: F. Petitjean *et al.*, "A statistically efficient and scalable method for log-linear analysis of high-dimensional data," in *ICDM 2014*.
[3]: A. Deshpande *et al.*, "Efficient stepwise selection in decomposable models," in *UAI 2001*.

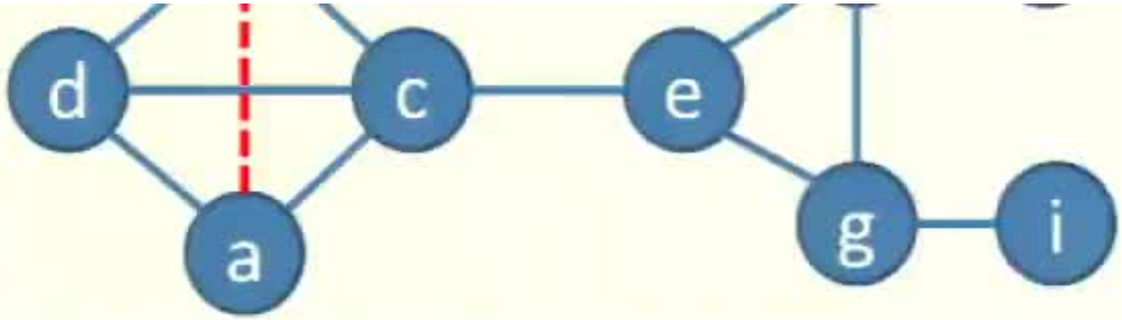
$$score(\mathcal{M}, \{a, b\}) = sc$$

This has been proven for different scorings

- Statistical tests [1] ✓
- MML/MDL [2] ✓
- KL divergence / log-likelihood [3] ✓

→ *This means that what we are about to show stands for all these scorings.*

[1]. F. Petitjean et al., "Scaling log-linear analysis to high-dimensional data," in *ICDM 2013*.
[2]. F. Petitjean et al., "A statistically efficient and scalable method for log-linear analysis of high-dimensional data," in *ICDM 2014*.
[3]. A. Deshpande et al., "Efficient stepwise selection in decomposable models," in *UAI 2001*.



$$score(\mathcal{M}, \{a, b\}) = score'(\{a, b, c, d\} \; ,$$

$$S_{ab} \cup a \cup$$

oring

Sor

dels...
uctures

ring
del → 32.5

...

oring
tion of
{a,b} to
odel → 12.2

Intuition

What we have seen s
- Evaluating the additio
upon 4 cliques of the

Our intuition

How often does that

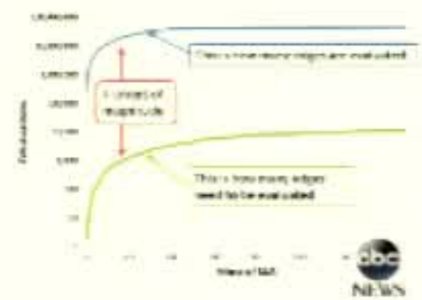
How can we use this

What we have seen so far:

- Evaluating the addition of an edge only depends upon 4 cliques of the graph

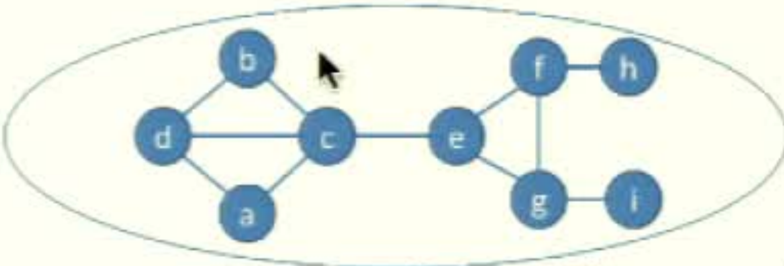
Our intuition

How often does that happen?



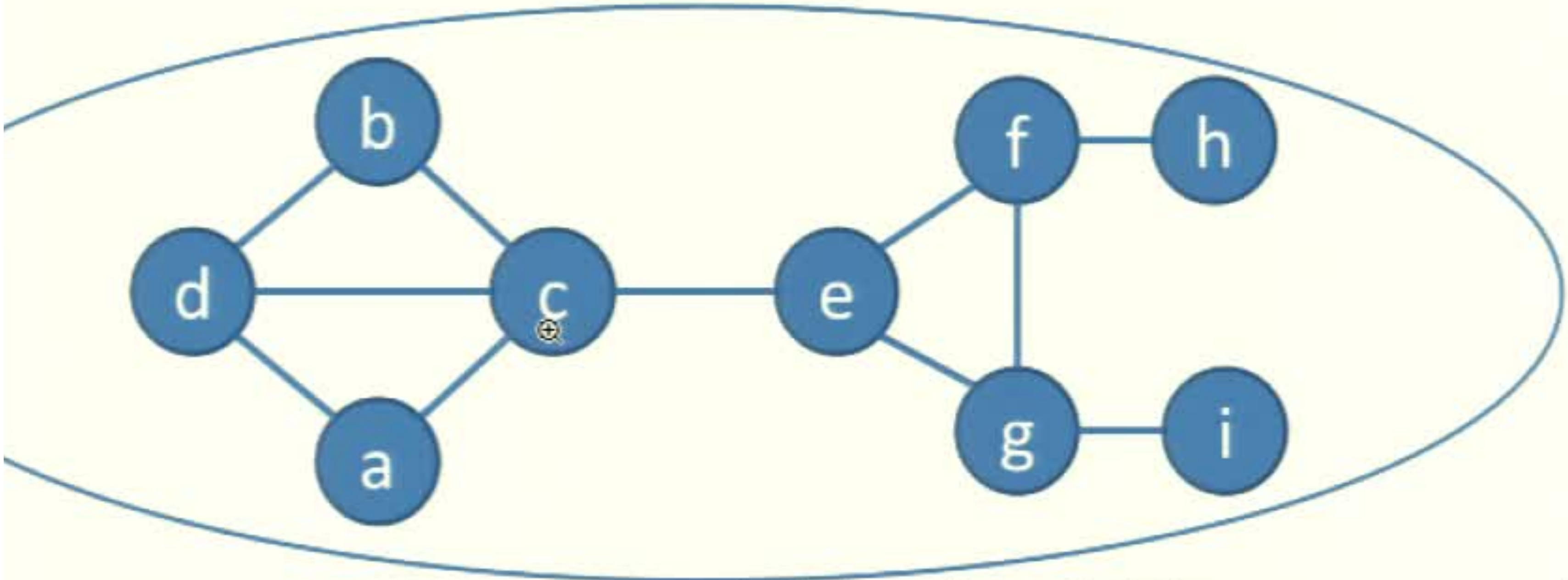
How can we use this information?

Not all edges should be re-examined at every step of the process



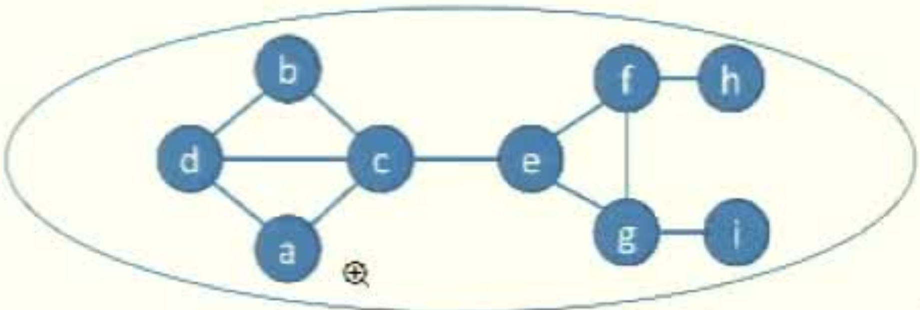
$$score(\mathcal{M}, \{a, b\}) = score'(\{a, b, c, d\}\ ,\ \{a, c, d\}\ ,\ \{b, c, d\}\ ,\ \{c, d\})$$

e process



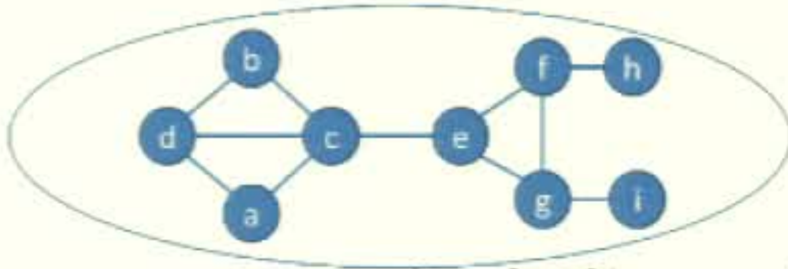
$$score(\mathcal{M}, \{a, b\}) = score'(\{a, b, c$$

Not all edges should be re-examined at every step of the process



$$score(\mathcal{M}, \{a, b\}) = score'(\{a, b, c, d\}\ ,\ \{a, c, d\}\ ,\ \{b, c, d\}\ ,\ \{c, d$$

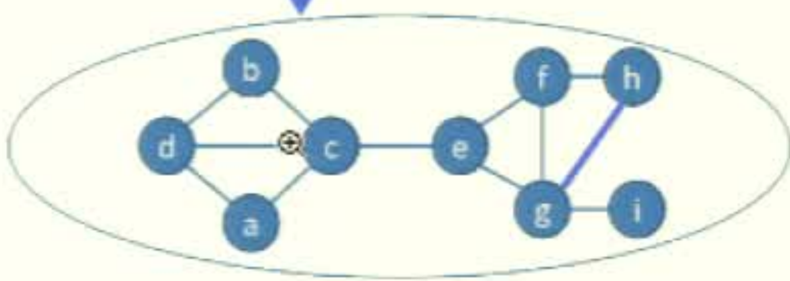
Not all edges should be re-examined at every step of the process



$score(\mathcal{M}, \{a,b\}) = score'(\{a,b,c,d\}\ ,\ \{a,c,d\}\ ,\ \{b,c,d\}\ ,\ \{c,d\})$

Select edge {g,h}

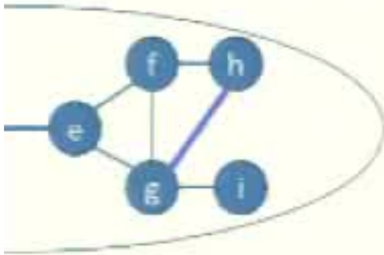
Score({a,b}) did ***not*** change



$score(\mathcal{M}, \{a,b\}) = score'(\{a,b,c,d\}\ ,\ \{a,c,d\}\ ,\ \{b,c,d\}\ ,\ \{c,d\})$

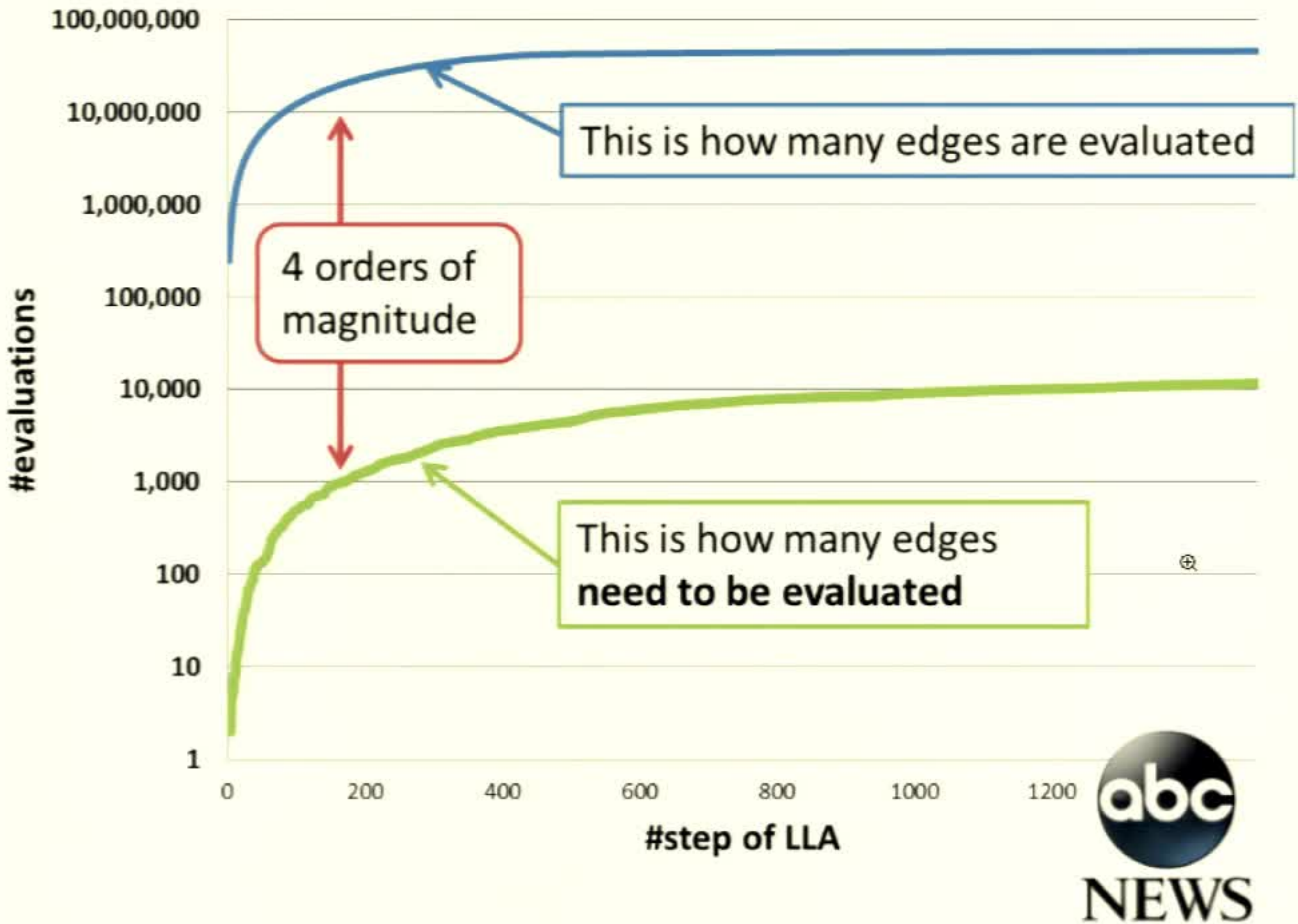
*The addition of edge {a,b} need **not** be re-examined in the new model* n}

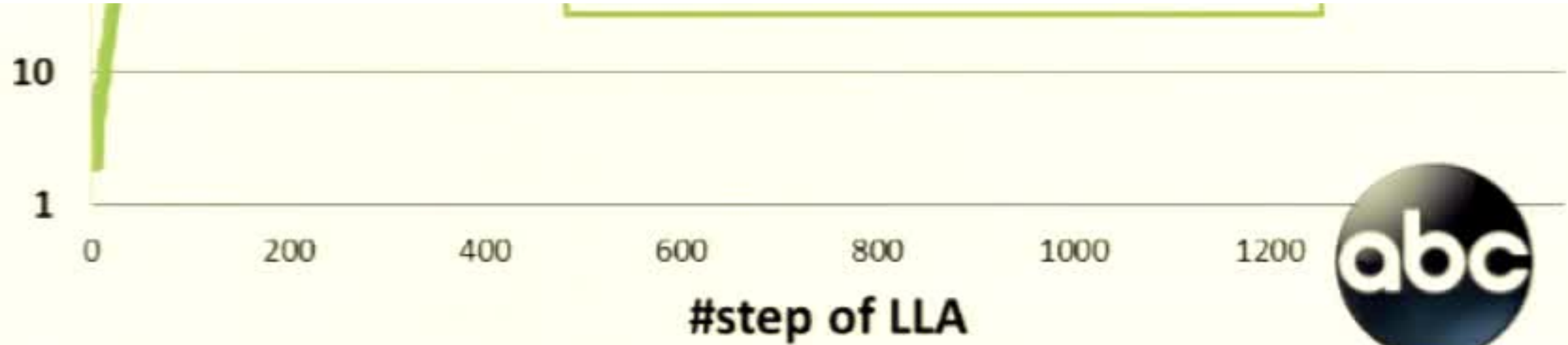
Score({a,b})
did ***not*** change



$$e(\mathcal{M}, \{a,b\}) = score'(\{a,b,c,d\}\ ,\ \{a,c,d\}\ ,\ \{b,c,d\}\ ,\ \{c,d\})$$

ldition of edge {a,b} need ***not*** *be re-*
ned in the new model


100,000,000
10,000,000
1,000,000
100,000
10,000
1,000
100
10
1
#evaluations
0
200
400
600
800
1000
1200
#step of LLA
This is how many edges are evaluated
4 orders of magnitude
This is how many edges
need to be evaluated
NEWS



We know: if S_{ab} does not change between different modifications of the graph, then the addition of {a,b} need not be re-examined

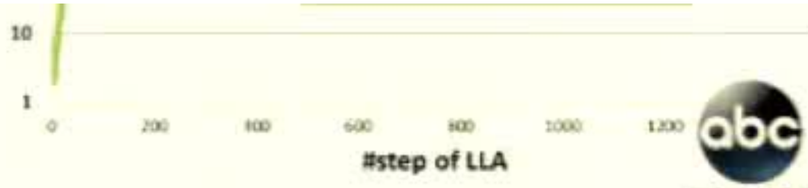
1. Use a data structure that gives direct

We know: if S_{ab} does not change between different modifications of the graph, then the addition of {a,b} need not be re-examined

1. Use a data structure that gives direct access to minimal separators for every potential edge

2. Keep track of the minimal separators for every potential edge

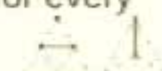
3. Maintain an ordered list of all the potential edges (priority queue)



ion?

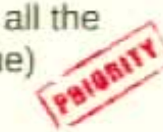
We know: if S_{ab} does not change between different modifications of the graph, then the addition of {a,b} need not be re-examined

1. Use a data structure that gives direct access to minimal separators for every potential edge

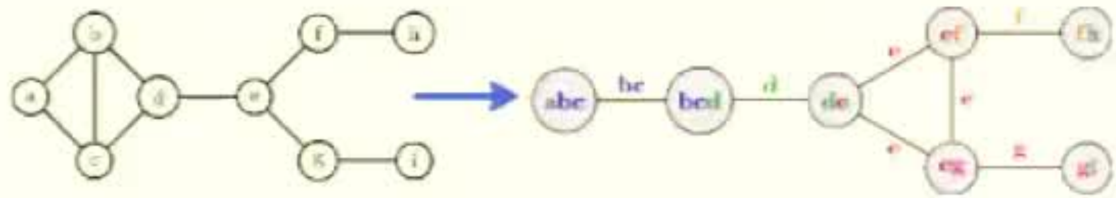


2. Keep track of the minimal separators for every potential edge

3. Maintain an ordered list of all the potential edges (priority queue)

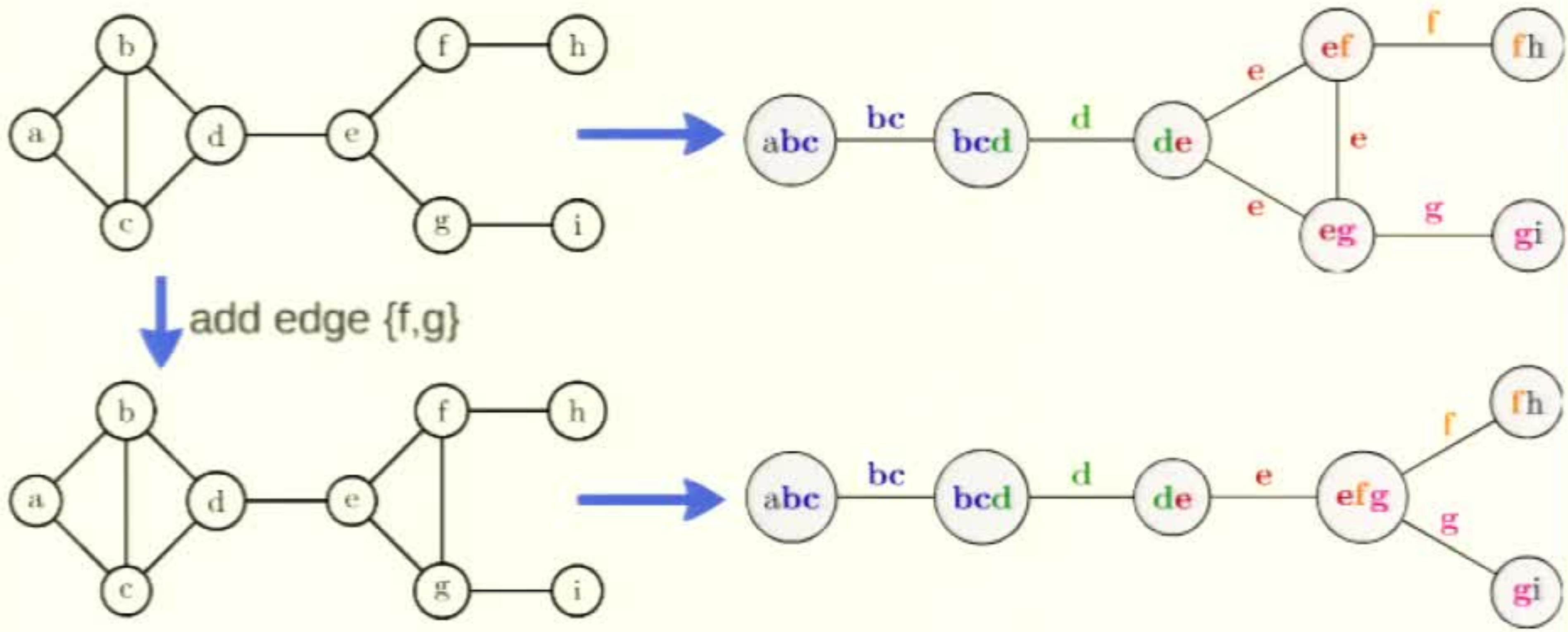


Clique graph



[1]: A. Deshpande *et al.*, "Efficient stepwise selection in decomposable models," in *UAI 2001*.

Clique graph



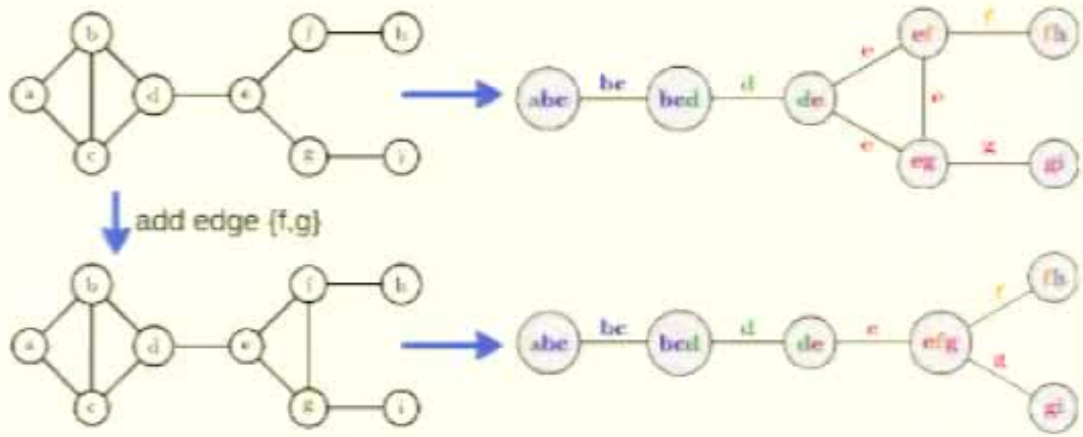
There are algorithms that can directly update the clique-graph [1]

BUT those algorithms cannot track the minimal separators S_{ab}

→ We make this possible - *details in the paper*

[1]: A. Deshpande *et al.*, "Efficient stepwise selection in decomposable models," in *UAI 2001*.

Clique graph

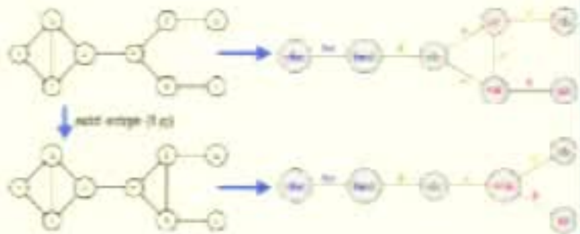


There are algorithms that can directly update the clique-graph [1]

BUT those algorithms cannot track the minimal separators S_{ab}
→ We make this possible - *details in the paper*

[1]: A. Deshpande *et al.*, "Efficient stepwise selection in decomposable models," in *UAI 2001*.

Clique graph



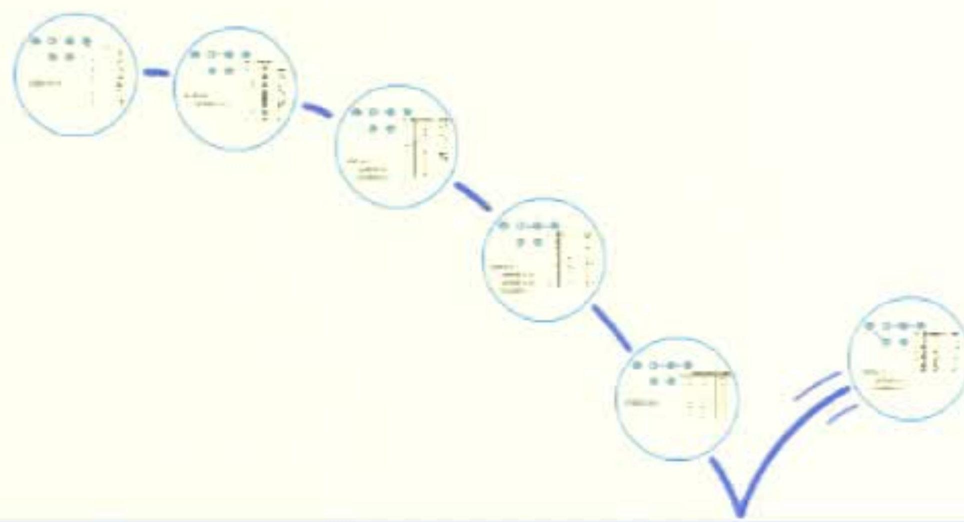
There are algorithms that can directly update the clique graph [1]

BUT those algorithms cannot track the minimal separators S_{ab}

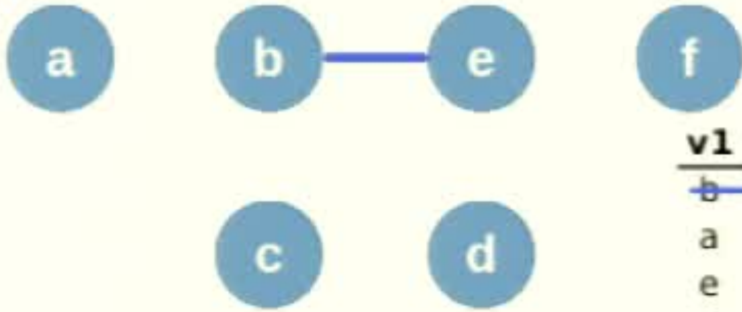
→ *We make this possible - details in the paper*

[1] A. Deshpande et al. "Efficient stepwise selection in decomposable models," in UAI 2001

$$v^2 \rightarrow v \cdot \log(v)$$

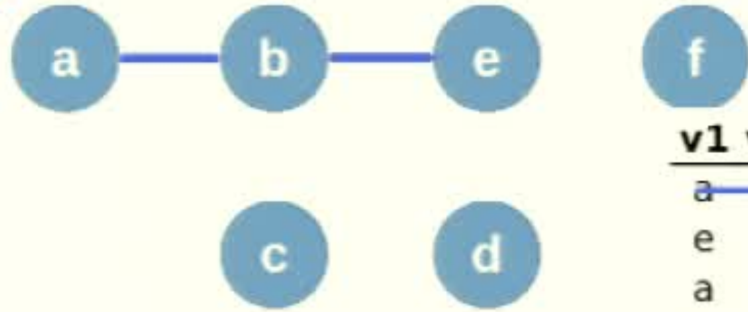


Priority queue



v1	v2	separator	score
~~b~~	~~e~~	~~{}~~	~~96.2~~
a	b	{}	72.8
e	f	{}	60.9
a	e	{}	49.5
a	f	{}	42.8
b	c	{}	31.4
c	e	{}	31.0
a	c	{}	28.8
b	f	{}	17.1
c	d	{}	16.9
b	c	{}	12.7
c	f	{}	8.1
d	e	{}	7.3
d	f	{}	4.8
e	f	{}	4.6

Add b-e

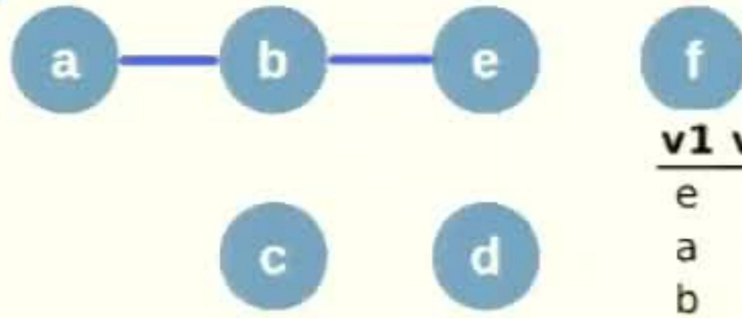


v1	v2	separator	score
~~a~~	~~b~~	~~{}~~	~~72.8~~
e	f	{}	60.9
a	e	{}	~~49.5~~
a	f	{}	42.8
b	e	{}	31.4
c	e	{}	31.0
a	c	{}	28.8
b	f	{}	17.1
c	d	{}	16.9
b	c	{}	12.7
c	f	{}	8.1
d	e	{}	7.3
d	f	{}	4.8
e	f	{}	4.6

b

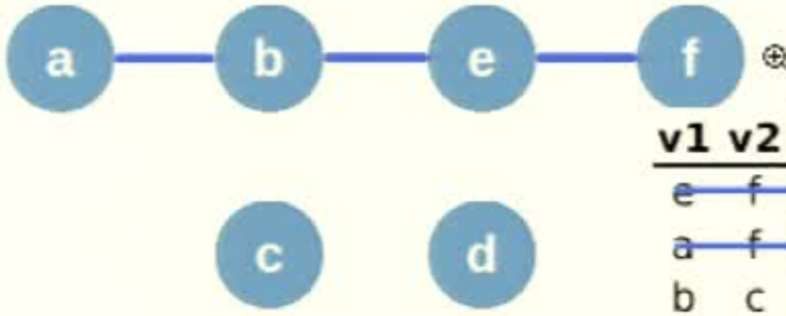
Add a-b

- update a-e



v1	v2	separator	score
e	f	{}	60.9
a	f	{}	42.8
b	c	{}	31.4
c	e	{}	31.0
a	c	{}	28.8
b	f	{}	17.1
c	d	{}	16.9
b	c	{}	12.7
a	e	{b}	12.4
c	f	{}	8.1
d	e	{}	7.3
d	f	{}	4.8
e	f	{}	4.6

Add e-f
- update b-f
- disable a-f



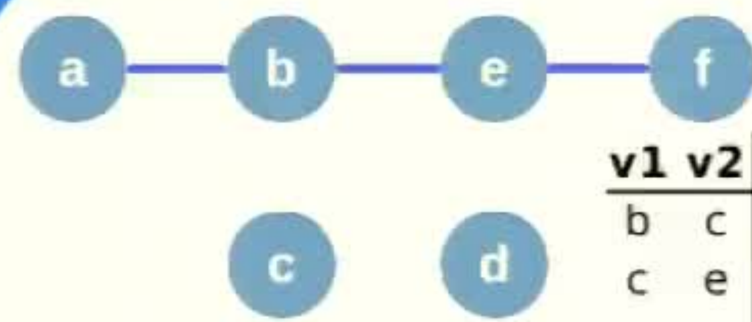
v1	v2	separator	score
~~e~~	~~f~~	~~{}~~	~~60.9~~
~~a~~	~~f~~	~~{}~~	~~42.8~~
b	c	{}	31.4
c	e	{}	31.0
a	c	{}	28.8
b	f	{}	~~17.1~~
c	d	{}	16.9
b	c	{}	12.7
a	e	{b}	12.4
c	f	{}	8.1
d	e	{}	7.3
d	f	{}	4.8
e	f	{}	4.6

e

Add e-f

- update b-f
- disable a-f

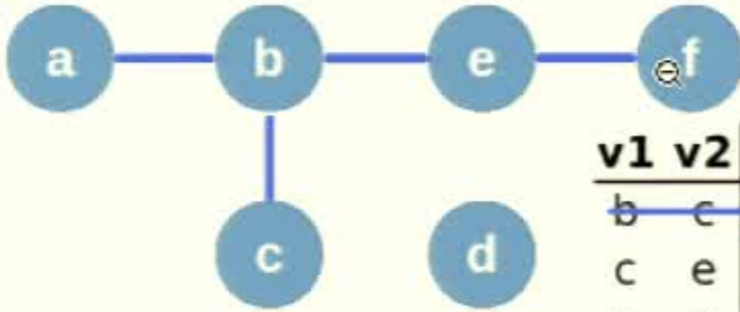
e	{b}	12.4
f	{}	8.1
e	{}	7.3
f	{}	4.8
f	{}	4.6



v1	v2	separato
b	c	{}
c	e	{}
a	c	{}
c	d	{}
b	f	{e}
b	c	{}
a	e	{b}
c	f	{}
d	e	{}
d	f	{}

Add b-c

- update a-c
- update c-e

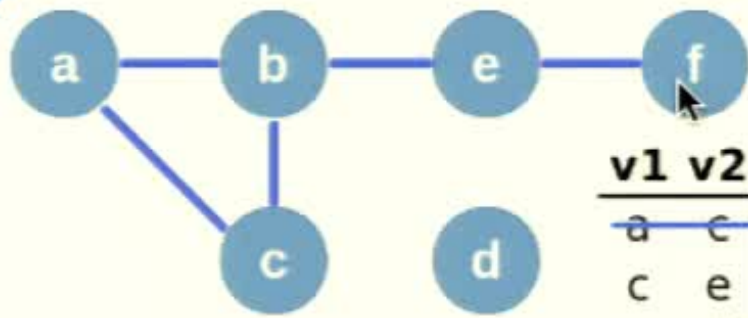


v1	v2	separator	score
~~b~~	~~c~~	~~{}~~	~~31.4~~
c	e	{}	~~31.0~~
a	c	{}	~~28.8~~
c	d	{}	16.9
b	f	{e}	14.0
b	c	{}	12.7
a	e	{b}	12.4
~~c~~	~~f~~	~~{}~~	~~8.1~~
d	e	{}	7.3
d	f	{}	4.8
e	f	{}	4.6

b

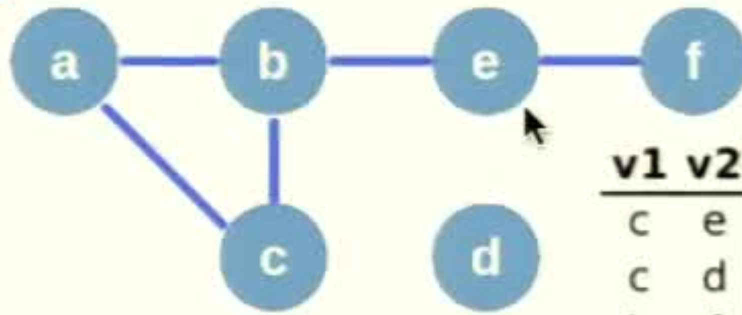
Add b-c

- update a-c
- update c-e
- disable c-f



Add a-c

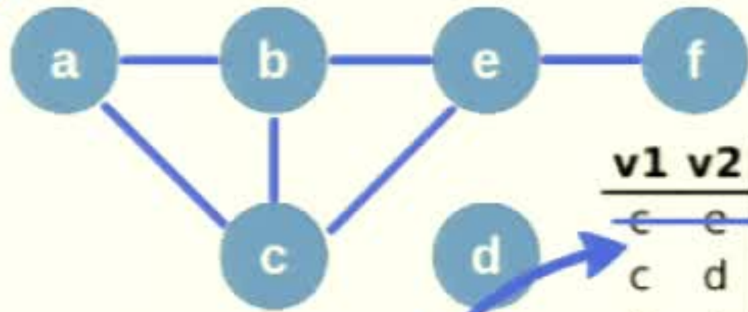
v1	v2	separator	score
~~a~~	~~c~~	~~{b}~~	~~84.5~~
c	e	{b}	24.2
c	d	{}	16.9
b	f	{e}	14.0
b	c	{}	12.7
a	e	{b}	12.4
d	e	{}	7.3
d	f	{}	4.8
e	f	{}	4.6



v1	v2	separator	score
c	e	{b}	24.2
c	d	{}	16.9
b	f	{e}	14.0
b	c	{}	12.7
a	e	{b}	12.4
d	e	{}	7.3
d	f	{}	4.8
e	f	{}	4.6

Add c-e

- update a-e
- enable c-f



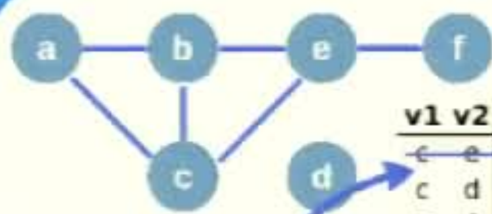
v1	v2	separator	score
~~c~~	~~e~~	~~{b}~~	~~24.2~~
c	d	{}	16.9
b	f	{e}	14.0
b	c	{}	12.7
a	e	{b}	~~12.4~~
d	e	{}	7.3
d	f	{}	4.8
e	f	{}	4.6

c	f	{e}	49.4

c

Add c-e

- update a-e
- enable c-f



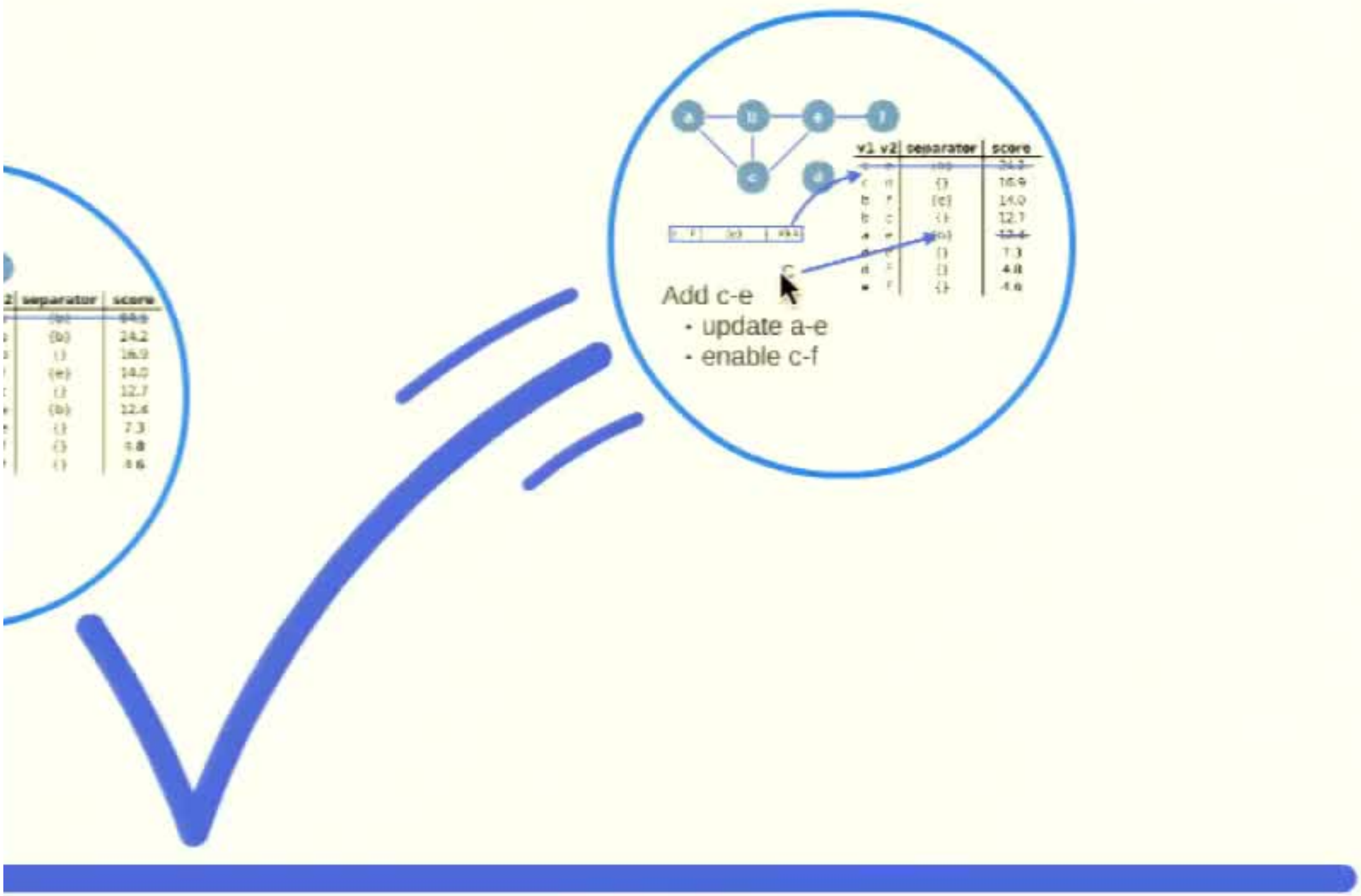
v1	v2	separator	score
~~c~~	~~e~~	~~{b}~~	~~24.2~~
c	d	{}	16.9
b	f	{e}	14.0
b	c	{}	12.7
a	e	{b}	~~12.4~~
d	e	{}	7.3
d	f	{}	4.8
e	f	{}	4.6

c	f	{e}	49.4

c

Add c-e

- update a-e
- enable c-f


v1 v2 separator score
Add c-e
• update a-e
• enable c-f

Solution

on

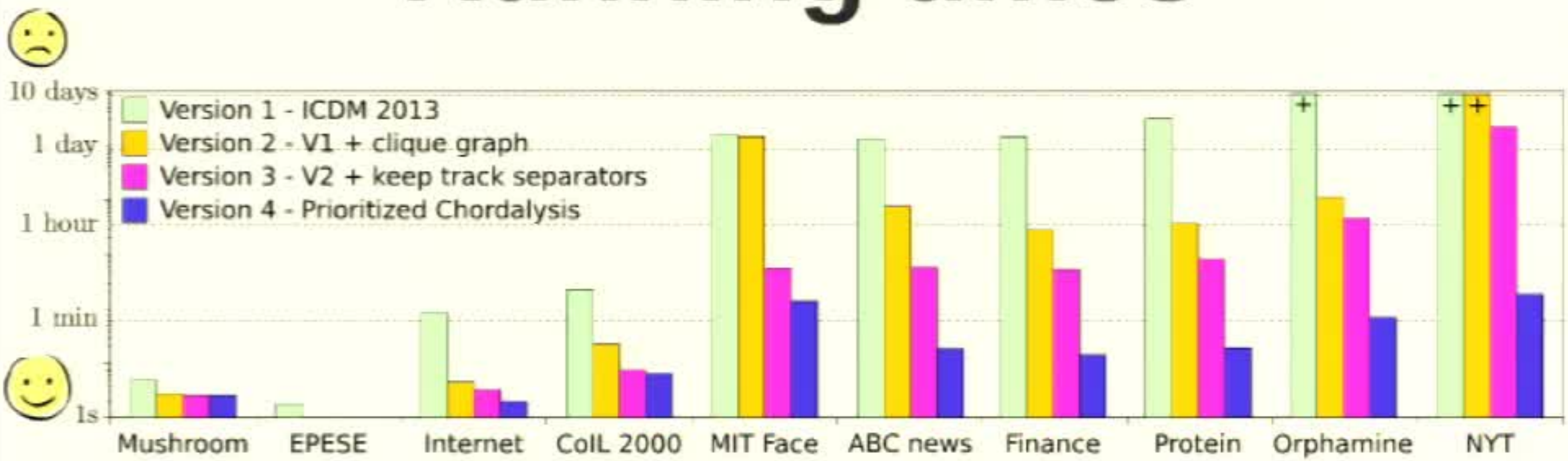
It works!

itized

lalysis

Running times

Running times



#Vars 20 25 70 85 300 500 500 700 1,200 2,000

Take-home message

Prioritized Chordalysis:

1. can analyse data with 1,000+ variables

2. does not sacrifice the soundness

3. is released on GitHub

https://github.com/fpetitjean/

Scaling log-linear analysis to datasets with 1,000+ variables

François Petitjean and Geoff Webb

Thanks for your attention!

http://www.francois-petitjean.com

francois.petitjean@monash.edu

@LeDataMiner

Scaling log-linear analysis to datasets with 1,000+ variables

François Petitjean and Geoff Webb

2015 SIAM International Conference on DATA MINING